Representation learning in Computational Pathology : Application to the Prediction of Molecular Cancer Features

Tristan Lazard

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL
Préparée à Mines Paris-PSL

Apprentissage de Représentations en Pathologie Numérique:
Application à la Prédiction des Caractéristiques Moléculaires du Cancer

Representation Learning in Computational Pathology:
Application to the Prediction of Molecular Cancer Features

Soutenue par
Tristan Lazard
Le 29 nov 2023

Ecole doctorale n° 621
Ingénierie des Systèmes, Matériaux, Mécaniques, Énergétique

Spécialité
Bio-informatique

Composition du jury :

Isabelle BLOCH Professeure, Sorbonne Université	*Présidente*
Jakob Nikolas KATHER Professeur, TU Dresden	*Rapporteur*
Daniel RACOCÉANU Professeur, Sorbonne Université	*Rapporteur*
Jean-Philippe VERT Professeur, Owkin	*Examinateur*
Thomas WALTER Professeur, Mines Paris	*Directeur de thèse*
Étienne DECENCIÈRE Professeur, Mines Paris	*Directeur de thèse*

MINES PARIS | PSL

Tristan Lazard
Representation Learning in Computational Pathology: Application to the Prediction of Molecular Cancer Features
November 29th, 2023

Remerciements

Sacrée aventure que la thèse ! En réalité c'est une aventure plutôt confortable: au chaud dans un bureau, la machine à café toujours à portée de main, la cantine et ses délices juste de l'autre coté de la rue... Je peux difficilement me faire appeler Mike Horn. C'est plutôt une aventure... intérieure ! Une montagne russe émotionnelle, c'est l'image d'une extrême originalité qui nous venait, mon co-thésard Matthieu et moi, lorsque nous en discutions. Sur ses rails, j'ai pu connaître des hauteurs agréables: l'excitation de la découverte, de "quand ça marche", celle des (rares) bonnes nouvelles des reviewers. La réalisation de la chance qui nous est donnée de pouvoir étudier, apprendre et explorer librement - une semaine à lire un livre et des articles : dans quel autre métier cela existe-t-il ?! -.

Mais aussi des bas : quand la majorité du temps "ça ne marche pas", que notre travail se fait rejeter pour la 5ème fois. Quand on doute: de son idée, de la direction qu'on a décidé de donner à un projet, ou même de l'intérêt de la thèse.

Mais heureusement, je ne suis pas monté dans cette attraction seul, et je voudrais donc remercier ici les personnes qui m'y ont accompagné. Et tout d'abord, je voudrais remercier la première ligne ! Les hommes du front : Thomas et Étienne, mes superviseurs.

Thomas, merci pour tout. Tu as été un mentor, un exemple tout au long de ces 4 années. Professionnellement déjà: ton enthousiasme pour la biologie, ton goût pour l'histoire scientifique qui "se raconte bien" et tes qualités d'écriture, ta curiosité. Personnellement ensuite, pour ton optimisme à toute épreuve et ta fidélité qui ont fait que, au sein de ton équipe, on se sent considéré et protégé.

Étienne, c'est toi qui m'as fait découvrir le traitement d'images. Merci pour cela. Merci pour ta supervision bienveillante : tu m'as apporté de la rigueur, mais aussi la capacité à relativiser et à prendre de la distance par rapport à mes projets.

Of course, I would like to extend a large thank you to all members of my jury, starting with my reviewers, for the probing questions and fruitful discussions that followed the defense.

Pr. Jakob Kather, thank you so much for reading my work! Thank you for making the journey from Dresden to Paris, despite the storm. Your work has been an inspiration to me throughout this thesis, and it was an honour to defend it in front of you.

Pr. Daniel Racocéanu, thank you for reading and correcting my work as well as providing a very detailed report. Not only were you present for the defense, but you also followed my entire PhD journey by participating in my follow-up committee: thank you for that.

Pr. Isabelle Bloch, thank you for presiding over my committee and for your insightful questions.

Pr. Jean-Philippe Vert, thank you for taking the time to judge my defense, for your thorough questions and the kind words that followed.

J'aimerai ensuite remercier les gens qui ont partagé mon quotidien: les membres du CBIO! Chloé, j'ai découvert grace à toi l'enseignement, en assurant les TDs de tes cours (plébiscités). En tant que figure du CBIO et des Mines, tu m'intimidais un peu au début. Sans aucune raison: ta bonne humeur et ton accessibilité m'ont vite fait oublié ce sentiment. Véronique, merci pour ton humour, ton franc-parler, ta volonté de tout décortiquer pour en comprendre le fond! Une bonne part de la vie de la V317, c'est toi! Merci au dernier venu des permanents, Florian, pour ton enthousiasme scientifique et ta gentillesse.

Merci aux anciens, qui m'ont accueillis à bras ouverts: Joseph, Hector, Lotfi, Judith et Romain. Romain notamment, car c'est toi qui m'a montré pendant l'été 2020 ce papier, SimCLR, qui m'a ensuite inspiré au long de ma thèse. Et bien sur merci à Peter: tu m'as accompagné dans mes premiers pas au sein du CBIO et de l'histopathologie. Merci pour ta gentillesse, ton ouverture, et l'exemple que tu m'as laissé. Et pour Nextflow.

J'en viens aux membres avec qui j'ai passé la majorité de la thèse, pre, mid et post confinement. Guillaume, ta patience et ta disponibilité ont été clefs dans l'avancement de ce premier projet HRD. Anne, fidèle au poste aux bureaux de Curie, merci pour ta résilience face à pip et conda. Aurélie, toujours dispo pour aider à interpreter un bout de tuile d'anapath, meme avec le pied cassé! Merci Vivien, qui a été parmi les premiers à refaire vivre les bureaux des Mines à la sortie du Covid. Arthur le codeur, sage parmi les sages. J'attends maintenant qu'on se la fasse, cette sortie ski de rando! Mélanie, co-thésarde d'entreprise et grande voyageuse; d'un calme olympien meme au milieu du pire stress. Vincent le toumeur, merci pour les deep discussions scientifiques qu'on a pu avoir en V317. Merci Adeline qui a su revigorer (ressusciter ?) les verres de labo à la montagne. Marvin, le champion du jeu de mot, merci pour tes points de vue souvent (très) tranchés mais intéressants. Asma, merci pour ton accueil, et pour le FOMO que tu as ramené apres ta conf' à Hawaï (!!!). Maguette, merci pour ton humour. Merci aussi de nous avoir donné tout au long de l'année la meilleure des excuses pour allonger le déjeuner. Thomas D, j'ai encore l'impression qu'on a commencé ensemble mais je me mélange souvent les pinceaux. Merci d'avoir toujours la pêche, tu nous as donné la banane tout du long. Ramène ta fraise au béon cette année. Thomas B, merci pour tes services rendus au CBIO. Tu es dans la droite lignée de Thomas W. (c'est une manie pour les Thomas ?!) au niveau de la gentillesse et as participé grandement à l'ambiance accueillante du CBIO. Philippe, on y croit pour Juin ! Que l'esprit de M. Shadow t'habite. Je te lègue mon jambon, prends en soin. Gwenn, partenaire de cantine et de bière après le boulot, j'ai été très heureux de partager du quotidien à tes cotés! Gwenaëlle, bravo pour ta reconversion! Dommage qu'on n'ai jamais eu l'occasion d'une session grimpe ensemble... Julie, avec Aurélie et Florian, tu es vite devenue une des "taulières" indispensables du bureau de Curie. Merci pour ta gentillesse.

Et évidemment, Matthieu, mon partenaire de crime (ou plutôt de montagnes Russes). On peut dire qu'on a passé l'épreuve du feu ensemble, du début à la fin. Depuis la commande de l'ordinateur de thèse, jusqu'à la rédaction même de ces remerciements. Merci pour ces 4 années, pendant lesquels tu as tenu le rôle d'ami mais aussi de papa-administratif pour ton serviteur toujours plus nul avec les démarches et les papiers. J'ajouterai que ta moustache m'aura inspiré à de nombreuses reprises - initiatives régulièrement découragées par Pauline, il est vrai que sur moi ça ressemblait plus à une serpillère qu'un fier plumeau.

A tous les nouveaux, Alice, Paul, Guillaume, Katia, Victor, Julian, je vous souhaite le meilleur des séjours au CBIO et à la cantoche de Curie. Qui sait, peut-être aurez-vous l'occasion de connaitre le val-de-grace ?! Prenez soin des plantes svp, gentikiki est encore à peu près en vie.

Merci aussi à tous ceux dont j'ai croisé la route, aux Mines ou à Curie: Santiago, Jesus, Samy, Anne-clair (fini les pots sur la terrasse, ça me manquera), Loic, Nicolas. Merci à Anne Salomon pour votre disponibilité et votre énergie.

Merci à la cantine de Curie et à ses personnels. Je me suis régalé. C'est peu dire. Merci aussi à tout le personnel des bureaux de Curie comme des Mines, qui m'ont permis de travailler dans de bonnes conditions ces quatre années.

Merci au CRI des Mines, merci à Vincent Brunet pour son aide avec le cluster. Merci beaucoup à l'Idris et à Jean-Zay pour son cluster incroyable, énorme merci à l'équipe support de Jean-zay (mention spéciale à Rémi Lacroix), toujours très sympa et extrêmement réactive.

Merci à Violaine Aubert pour m'avoir donné le goût des maths, et à Amaury Lambert pour l'avoir affermi, au travers de vos enseignements. Merci aussi Amaury de m'avoir conseillé le CBIO en 2017. Merci à l'ENS et à son département de biologie pour m'avoir permis de suivre mon cursus peu commun.

J'ai pu remercié tous mes collègues des Mines ou de Curie, mais en réalité il en manque une partie. Donc à mes amis ET collègues provisoires, avec qui j'ai partagé soit l'un soit l'autre des confinements: Eloi, Jérem, Marion, Mélanie, Jéjé, Drey, Apo, Herby, Matthiou, Stoup, merci de m'avoir supporté! Et pour les pauses clopes, cafés, les déjs à la cantine d'Huez, les calls dans la salle de réunion du haut, bref la vraie vie de TT qu'on a partagé ensemble en montagne. Mention toute spéciale pour Jéjé, mon co-thésard presque comme Matthieu l'a été, avec qui j'ai passé beaucoup de mes journées de TT, à la bibli Jussieu ou ailleurs.

Forcément, je me dois d'ajouter un mot pour mes amis, qui d'une manière ou d'une autre m'ont soutenu, aiguillé, conseillé. Merci aux chefs et aux blondes (big-up Louis, Vicky, Louise, Juliette..!). Merci Albane pour le bout de chemin partagé. Merci aux copains de l'ENS, des exemples de curiosité et de motivation. Théo et Val les deux maitres Dules, avec qui j'ai passé mes aventures les plus polaires. Dariusz HUG, teuffeur geek et grand chercheur à l'humour aiguisé. Tristan F, on s'est suivis depuis le tout début des études sup', confiants dans la "règle des Tristans". Merci François, conteur fascinant et rédacteur d'email de renom. Gracias aux copains de l'agro. Merci aux vags: Robbie, Marine crapule, Clo, PacoNico, Mme Fanène, Stan

Sozper. Comme à l'étable à l'époque, bien qu'on vadrouille chacun deci delà, c'est toujours génial de se retrouver. Arthour couillère, merci pour l'accueil régulier chez toi, t'es vraiment un chic type. Consciencieux et sympa. Sarah Moreira Mirador merci pour les bon tips d'une ancienne thésarde repentie. Grazzie à toute la team Paris, avec qui j'ai passé moultes vacances, aventures sportives et festives au cours de ces années: merci Japo, Gagou, Blanche, Chacha, Gaby, thether, Maxou, Marie et Lou. Merci Kapo, pour ton accueil (abbe)courtois et la statue de chien. Et aussi pour m'avoir fait répéter mon intro! Merci Néné. Watch me. Merci aux vieilles branches, les copains de longue date, dont je vois la moitié moitié moins que je le souhaiterai: Menux, un gored gored svp! Ludo, toujours le bon conseil et l'anecdote qu'il faut; Edo futur papa et gérant de réstau ? Ugo l'autostoppeur esthète, Lukas t'as deux têtes...

Big up Zizi, aka the best coloc', vieux copain s'il en est et débatteur hors pair! Merci pour m'avoir supporté, mon caractère de cochon et moi, et pour tous les bons moments à Aubervilliers, Saint-Maurice, Tourmaline, au Génie, ...

A mes parents, pour votre soutien et votre amour inconditionnel, du plus profond du coeur, merci. Merci Arnaud, un exemple de grand frère, dans le travail mais aussi la littérature, le sport, la musique... Merci Beb, pour ta sensibilité et ton oreille attentive.

Evidemment, le dernier mot est pour toi Pel. Tu as été ma meilleure alliée pendant cet épisode et tout aurait été sacrément plus dur sans toi. Merci pour ton soutien quotidien!

Contents

General Introduction

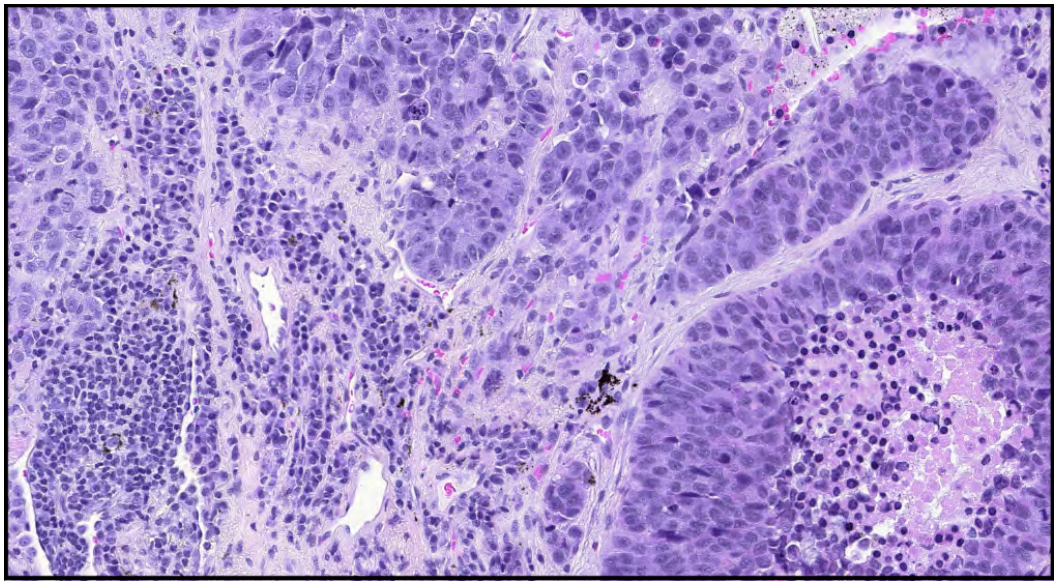


Figure I.1.: Lung cancer tissue, extracted from the TCGA dataset (slide TCGA-05-4245-01Z-00-DX1).

Contents

Summary:

Whole Slide Images (WSIs) are digitalized versions of microscopic images that capture thin layers of stained tissues. These images serve multiple purposes in cancer care, from clinical diagnosis to various stages in the cancer treatment process. The primary focus of this thesis is the development of prediction models based on WSIs. In this chapter, our aim is to provide a comprehensive introduction to these unique objects, which have limited analogs outside of the medical domain. Specifically, we will discuss the acquisition and utilization of WSIs, what we aim to predict from them, and conclude with the broad objectives of this thesis.

Résumé:

Les images de lames entières (WSI) sont des versions numériques de vues microscopiques de fines couches de tissus biologiques teintés. Elles ont diverses applications, allant du diagnostic clinique à une assistance dans diverses étapes du traitement du cancer. Le but principal de cette thèse est de développer des modèles prédictifs basés sur ces WSI. Ce chapitre propose une introduction à ces objets singuliers, qui trouvent peu d'équivalents en dehors du domaine médical. Nous y aborderons le protocole d'acquisition des WSI et listerons les singularités qui font de leur utilisation un défi. Nous verrons ensuite dans quel cadre ces images sont utilisées en cliniques, puis en apprentissage automatique -ce que nous visons à prédire avec elles-, et nous terminerons en énonçant les objectifs clés de cette thèse.

I.1 Context and scope

Histopathology refers to the microscopic examination of diseased tissue, and our emphasis here is on its role in cancer care. The foundations of histopathology can be dated back to 1840 with J. Müller's landmark publication ("On the Nature and Structural Characteristics of Cancer, and of Those Morbid Growths Which May Be Confounded with It" 1840), and to this day, it remains pivotal for cancer diagnosis and prognosis. It involves studying thin layers of fixed and stained tissues from surgical resections or biopsies, mounted on glass slides, and viewed under a microscope. Such slides display hundreds of thousands of cells, a large variety of tissue types, and are thus informative on single cell phenotypes as well as general tissue architecture.

The first virtual microscope, a product of computer science's endeavor in spatial data research, emerged in 1997 (Ferreira et al. 1997). This development, followed by the release of the first commercial slide scanners (Pantanowitz et al. 2011), paved the way for digital pathology. The field witnessed a synergy between evolving scanners and specialized software, leading to the increasing prevalence of Whole Slide Images (WSIs).

These digitalized versions of glass tissue slides offer promising avenues for the development of sophisticated algorithms to assist pathologists in their daily tasks. Another interesting application of computational approaches consists in building predictive models taking these images as input, and predicting a large variety of variables, such as the evolution of the disease, the effect of a treatment or the molecular landscape of the underlying disease. This field of application is generally known as "Computational Pathology".

A significant stride in image processing was made with the advent of deep learning, particularly convolutional neural networks (CNN), which showcased their prowess in pattern recognition for imaging (Krizhevsky, Sutskever, and Hinton 2012). This achievement quickly resonated with adjacent fields, including Computational Pathology, where the number of publication using machine learning *grew* at an unprecedented rate (Asif et al. 2023).

This thesis is nestled within this evolving landscape, focusing on the development of predictive models for WSIs through machine and deep learning techniques. On the application side, I have been involved in several medically driven projects in a variety of cancer types. All projects had in common that we wanted to predict molecular features such as single gene mutations or mutational signatures from WSIs.

The manuscript is organized as follows: this initial chapter introduces the biological subject of focus in this thesis and outlines the objectives pursued. The second chapter introduces literature pertinent to the posed questions and detail the solutions offered by my studies. This will be followed by the articles themselves, either published or under submission. For better readability, I have harmonized their formatting with the general document setup. Each chapter preface outlines the research context

and origins of key ideas. In some instances, unpublished sections that nonetheless contribute to the problem-solving will be included. Lastly, a discussion chapter will discuss the perspectives that this thesis open.

I.2 Whole Slide Images (WSI) acquisitions

I will start by introducing the protagonists of this work, the WSIs, by detailing how they are made and used. Their unique characteristics indeed shaped this whole body of work.

I.2.1 Slide preparation

Biological specimens from which WSIs are derived can be categorized into surgical specimens or biopsies. Regardless of the type of sample, both undergo a standardized process:

1. Fixation: This step is crucial for halting ongoing biological processes within the tissue post-sampling, such as enzymatic reactions, apoptosis, and protein synthesis. Fixation can be achieved either using of a fixative solution, with Formalin being the most prevalent today, or by freezing the samples. Although freezing is simpler, it often results in less clean tissue slides due to the deformation caused by growing ice crystals.
2. Dehydration: This step ensures the prevention of aerobic reactions and therefore preservation of the sample. Typically, this is done by rinsing the sample with ethanol.
3. Embedding: This phase focuses on solidifying the sample. Bathing the specimen in Paraffin is the most common method, imparting a resin-like texture to it.
4. Staining: The samples are stained according to a given protocole.
5. Sectioning: In the final step, samples are sliced into thin layers with the aid of a microtome. These layers usually range from 5 to 10 microns in thickness and are then placed on a microscope slide for subsequent examination.

Several staining protocols can be utilized (see Figure I.2 for examples):

- Non-specific protocols: These primarily highlight the fundamental components of tissues. The Hematoxylin and Eosin (H&E) staining protocol, as detailed by Fischer et al. (Fischer et al. 2008), is recognized internationally as the standard routine staining technique. In this method, hematoxylin distinctly colors nucleic acids (mainly found in the cell nucleus) in a deep purple hue, while eosin non-specifically stains proteins, rendering a predominantly pinkish tone to the slides as in Figure I.1. However, variations in this procedure exist.

For example, in France, the HES staining protocol is more common, wherein saffron is introduced to specifically tint collagen fibers.

- Specific protocols: These are formulated to target particular elements within tissues, such as distinct proteins or cell types. A prime example of this is Immunohistochemistry (IHC), which is achieved through the application of specific antigens that are subsequently paired with fluorophores, as elucidated by Kim et al. (S.-W. Kim, Roh, and Park 2016).

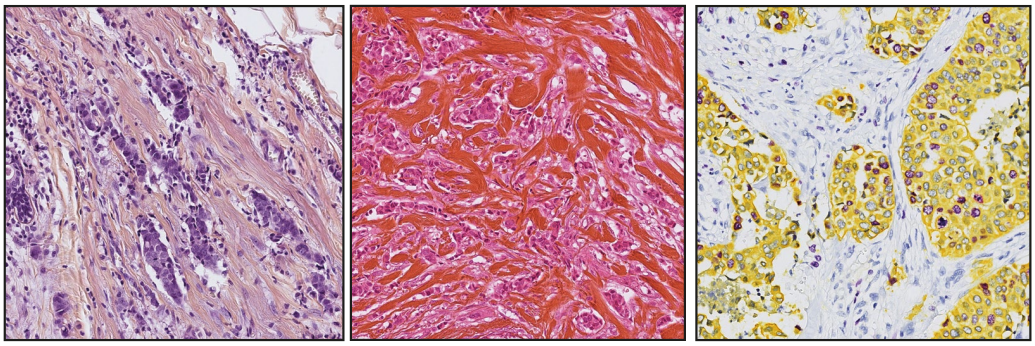


Figure I.2.: Different staining protocoles exemplified.

In addition to crafting glass tissue slides, biological specimens can serve multiple analytical purposes. These range from transcriptomic analysis, single-cell sequencing, to Next-Generation Sequencing and proteomic studies. As a result, it often becomes more practical to divide the original sample into distinct blocks: one dedicated to slide creation and others reserved for diverse biological assessments.

Two crucial implications arise from these practices:

- A single slide provides only a partial view of the tumor. While a wealth of information is encapsulated within this slice of tissue, it represents only a tiny fraction of the tumor's overall complexity.
- Variability can be introduced in the final appearance of slides due to differing protocols and practices across health centers. These differences can often manifest as distinct visual patterns in the images, that are specific identifiers of health centers.

I.2.2 Slide digitalization

Similar to a microscope, WSI scanners maneuver slides in both x and y directions, capturing images at designated magnifications such as 4x, 10x, 20x, and 40x, which correspond to 2, 1, 0.50, and 0.25 microns-per-pixels (mpp), respectively. For instance, most of the images in The Cancer Genome Atlas (TCGA) were acquired at a 40x magnification (0.25 mpp), where a typical cell occupies approximately a 10-20 pixel square, allowing some intracellular features to be discerned. The

scanner utilizes precision mechanics to capture successive fields of view across the entire slide, which are then stitched together to generate a comprehensive digital representation. These digital slides are exceptionally large, containing several billions of pixels, and can span up to 50,000 pixels squared.

To manage these large files, the development of WSI scanners has been paralleled by advances in specialized software and file formats, often borrowed from geospatial imaging. An example is the open-source project QuPath (Bankhead et al. 2017), which provides a WSI viewer.

The resulting digital slides are stored as pyramidal files with extensions such as SVS or TIF. These pyramidal files contain multiple levels of magnification, and at high magnifications, the image is divided into composing tiles, usually sized at 512×512. This pyramidal architecture allows visualization software to zoom seamlessly without loading the entire slide into memory, with tile stitching performed in real-time for x-y navigation. This setup mimics traditional microscopic examination, offering capabilities similar to zooming and panning in platforms like Google Earth. Although the initial cost and complexity of WSI scanners have impeded universal adoption, their use is becoming increasingly prevalent in healthcare, driven by the benefits of digital archiving, remote consultation, and computational analysis : the available dataset are consequently growing rapidly in size.

I.2.3 Specific challenges of WSI processing

WSI processing faces distinct challenges in comparison to natural image processing.

- **Size:** One primary constraint is the sheer size of the WSIs, which makes standard algorithms inapplicable. While viewer software has been developed to manage these large images, most image-processing algorithms still struggle with the computational burden, even on high-performance hardware.
- **Biases:** Additionally, WSI datasets are susceptible to biases stemming from their intricate preparation process, which results in varied visual features. These features can confound WSI processing algorithms and spuriously correlate with variables of interest such as the response to treatment or the grade of the tumour.
- **Not object-centric:** Unlike natural images, WSIs are not object-centric; they lack a singular or limited set of focal points, and relevant information may be present across the image at multiple magnification levels.

These constraints not only differentiate computational pathology from natural image processing but also limit the applicability of algorithms developed in the latter domain, despite its considerably larger research base. As a result, there is a need for the development of algorithms specifically tailored to WSIs.

Although not the primary focus of this thesis, these challenges have been addressed in each of the solutions proposed herein.

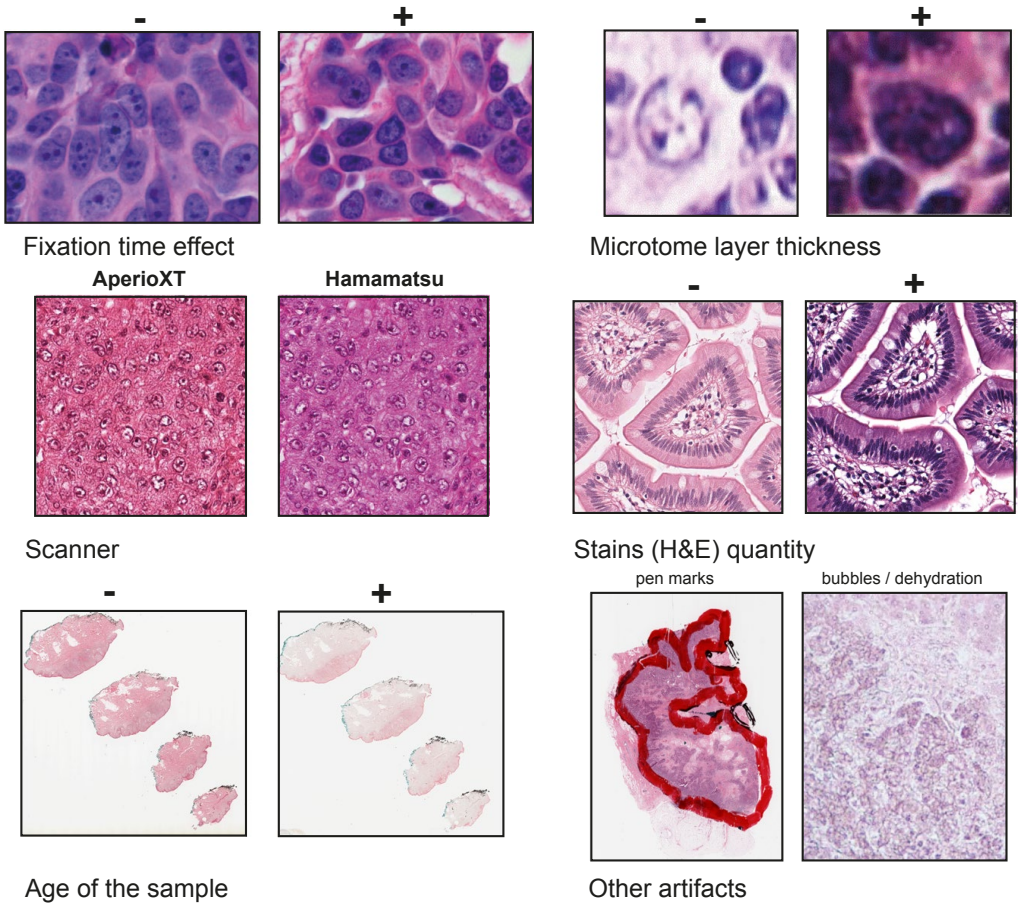


Figure I.3.: **Illustration of possible biases** that may arise during WSI acquisition. Images are adapted from Azevedo Tosta et al. (2019) and Babic et al. (2010).

I.3 Applications of WSIs

I.3.1 Clinical use of WSIs

In clinical practice, these slides are inspected using both their physical form —as glass slides— and in their digital WSI format. The digital format is particularly useful when collaborative discussions on complex cases are needed.

Here, we outline some of the core tasks that pathologists routinely perform using these slides:

- **Diagnosis of Malignancy:** The primary task is to diagnose whether a tumor is benign or malignant. This determination is based on various morphological signs such as atypical cell structures or unusual tissue growth patterns.
- **Staging and Grading:** Once a tumor is identified as malignant, the next steps involve staging and grading it. The stage of cancer ranges from I to IV and gives insights into how much the cancer has spread locally. On the other hand, the grade of the tumor indicates its aggressiveness. Both these classification tasks are carried out following the guidelines described in the WHO classification of tumours. For instance, aspects like lymph node metastasis (Board n.d.b) and mitotic count (Board n.d.a) are considered.
- **Immunohistochemistry and Special Stains:** As elaborated in Section I.2.1, these techniques are employed to identify specific proteins in cancer cells. Knowing the status of hormone receptors in breast cancer, for example, can inform the choice of first-line treatments (Board n.d.a).
- **Margin Assessment:** Pathologists also examine the healthy tissue surrounding a resected tumor to ensure the absence of residual cancer cells. This is crucial for determining whether further surgical intervention is required or not.

The tasks listed here are not exhaustive but serve to illustrate the integral role that diseased tissue slides play throughout the cancer care journey. From initial diagnosis and treatment planning to monitoring treatment efficacy, these slides are central to the practice of oncology.

I.3.2 Machine learning applications

The first goal of WSI processing is to facilitate automated workflows that aid clinicians in their daily practice. One dominant approach targets the automated detection and segmentation of histological primitives such as cells and glands. The endgame is to achieve superhuman accuracy in the quantification of these primitives and to assimilate this information at the slide level for taking informed patient care decisions. For instance, a seminal study by Yuan et al. (Yuan 2015) segmented lymphocytes and tumor regions to construct a spatial model of slides. They computed specific quantities like the number of intra-tumoral and adjacent lymphocytes and

correlated these metrics with patient survival data. This bottom-up methodology has seen burgeoning research interest and benefits from advancements in image processing, ranging from mathematical morphology algorithms to contemporary deep-learning techniques. The approach is known for its rigorous methodology, which facilitates high interpretability; each histological primitive extraction and the associated feature is computed separately and can thus be individually optimized and validated.

This technique mirrors the process by which a pathologist examines a tissue slide, assisting them at various steps such as counting tasks, identifying regions of interest (ROIs), or measuring areas of necrosis. This inherent compatibility allows for swift integration into a pathologist's workflow if adequately validated. This has been the cornerstone of products from companies specializing in computational pathology software, such as Primaa and PathAI, the latter of which relies on a comprehensive set of automatically extracted histological primitives (Diao et al. 2021).

Despite their merits, the bottom-up methods employed in WSI processing are not devoid of limitations. Primarily, these approaches are computationally demanding, often necessitating the execution of multiple specialized and resource-intensive algorithms (e.g., semantic segmentation algorithms for nuclei or ROIs, detection algorithms) across the entire slide. This computational burden extends to the software infrastructure required for the development and maintenance of such pipelines. Furthermore, the outputs of these algorithms can be voluminous, generating data artifacts like comprehensive slide segmentation masks or intricate proximity graphs of cells or ROIs. A further limitation lies in the pre-defined set of extracted features, such as segmented cell types, which imposes constraints on the modeling capacity and could inhibit research flexibility.

Alternatively, emerging methodologies focus on direct, end-to-end predictions of slide-level characteristics or classifications. Since the advent of deep learning, these methods offer several competitive edges (Campanella, Hanna, Geneslaw, Miraflor, Silva, et al. 2019; Coudray et al. 2018). They exploit deep learning for automated feature extraction at both the region and slide levels, thus sidestepping the limitations of our preconceived knowledge about the disease. These techniques also offer computational advantages; they rely on the computation of compact numerical vectors representation for predictions rather than the explicit histological primitives extraction, making the algorithms faster and the outputs more condensed.

Importantly, the optimized task aligns more closely with the ultimate goal of WSI processing: to derive patient- or tumour-level insights for informed treatment decisions. My thesis focuses on these end-to-end prediction methodologies, aiming to develop machine learning models that predict slide-level information directly from the raw WSIs. Specifically, the goal is to utilize a given training dataset of WSIs X with matched slide-level labels y to learn a parametric function f such that $f(X) \simeq y$. In this framework, the nature of the label y delineates the type of supervision and inherently presents its own set of challenges and opportunities.

I.4 Supervision and WSIs

I.4.1 Supervise training with medical knowledge

Firstly, medical doctors themselves can serve as the source of supervisory signals, an approach which is highly intuitive since this supervision emanates directly from tasks embedded in clinical practice. Specifically, the tasks targeted by our algorithms parallel those customarily carried out by clinicians, as outlined in Section I.3.1. These supervised tasks, symbolized by an "eye" in Figure I.4, span a broad spectrum of scales. At the cellular level, pathologists can perform tasks like cell segmentation and classification. Moving to the region or tile level, pathologists can demarcate ROIs and categorize them. For example, in the TissueNet challenge, clinicians were tasked with identifying differently graded lesions in cervical biopsies, thus providing non-exhaustive labelled regional images. Additionally, metrics like mitotic count can be assessed in these regions. Pathologists can also synthesize information across an entire slide to derive similar metrics or perform slide-level segmentation of ROIs such as tumorous areas. They can further characterize the overarching structure of the tumour, evaluating attributes like general architecture, differentiation, and cell atypia. Ultimately, this information can be integrated with other patient-level data, such as lymph node invasion or IHC data, to assign a definitive label to the tumour, such as its stage.

This supervisory information is invaluable for WSI-level prediction algorithms but comes with specific constraints.

- **Cost**. Foremost among these are the costs associated with medical expertise required for annotation. As physicians themselves must undertake these tasks, and given the vastness of WSIs, the process is exceedingly time-consuming. As a consequence, datasets annotated in this manner are very limited in size compared to those in natural image processing, when available at all.
- **Low agreement**. A second significant constraint is the issue of low agreement between annotators, a problem that is especially pronounced in the field of pathology. Studies have shown that the agreement among pathologists ranges from low to moderate on tasks such as mitotic figure recognition (Malon et al. 2012), to tumour staging, grading, and classification (Costantini et al. 2003; Krane et al. 2022). This variability in annotations underscores the necessity of consultative meetings among board-certified pathologists, especially for complex cases. Achieving more consistent supervision would necessitate the concurrent labelling of cases by multiple pathologists, which would further exacerbate the cost issue.

I.4.2 Supervise training with indirect supervision

A tumor is a complex biological system that can be investigated from multiple angles, including but not limited to, genetics, epidemiology, and proteomics. This multidis-

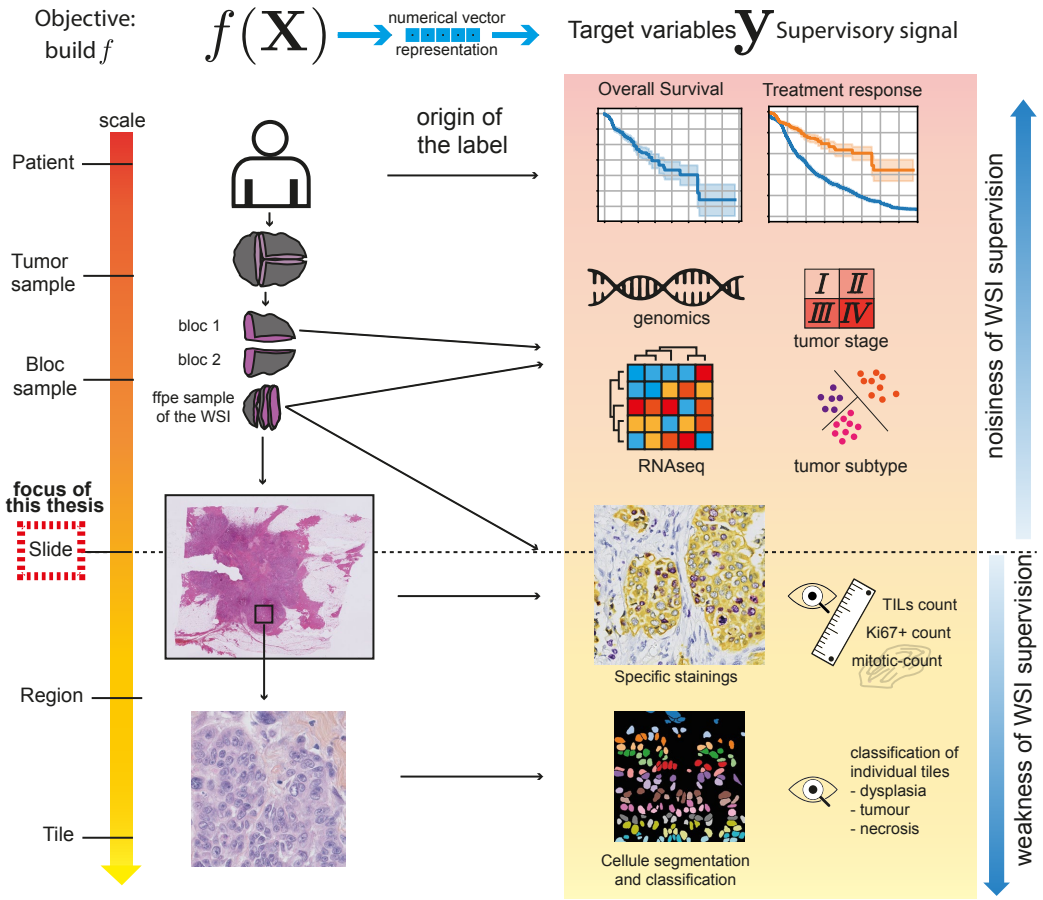


Figure I.4.: **Overview of the different supervisory signal and their origin.** Multiple scales of study are employed, ranging from the patient-level down to individual cellular observations within a tumor sample. The overarching aim is to predict patient-level outcomes, such as overall survival or response to treatment. This can be approached by generating predictions at finer scales, such as the cellular or tile level, and then integrating this information into higher level (tumor subtype) until patient level (survival etc...). The primary focus of this thesis is to develop predictive models based on WSIs. The availability and type of supervisory signals y vary depending on their hierarchical level: Signals derived from a level higher than WSIs, such as tumor samples or the patient, tend to introduce noise into the model as y is not directly linked to the WSI. Conversely, signals originating from a lower hierarchical level, such as regions, tiles, or cells, offer a weak supervisory signal as y is linked to part of the WSI.

ciplinary lens offers a wide array of methods to generate supervisory information, which can subsequently be used to train WSI-based predictive algorithms. Given the rich complexity and the sheer volume of information available in WSIs, it is plausible that information from other modalities, such as Next Generation Sequencing, could manifest discernible traces in these images. Utilizing WSIs to predict such modalities would offer significant advantages, particularly given the relatively lower cost of WSI acquisition.

Additionally, WSIs have the potential to contain markers that are prognostic in nature. This opens another promising avenue—using WSIs to predict survival rates or anticipate responses to treatment. In this context, supervision is determined by future outcomes.

These methodologies share common challenges and opportunities, which we group under the term "indirect supervision": the label acquisition is decorrelated from the WSI itself.

- **Noisiness**. One of the primary challenges is the issue of noisiness. As shown in Figure I.4, other modalities are often measured from a different tissue block than the one used for WSI. For example, in The Cancer Genome Atlas (TCGA), the tumor sample is divided into several blocks; one is formalin-fixed and paraffin-embedded to produce diagnostic WSIs, while others are frozen for genetic and transcriptomic analyses. Tumor heterogeneity may introduce discrepancies between these blocks. Moreover, higher-level supervisory signals like overall patient survival may be influenced by extraneous factors such as socio-economic conditions, adding noise to the supervisory information. As a rule of thumb, we can say that the more global an assessment is, the noisier the labels, e.g. while manual segmentation of nuclei may contain few errors, manual grading is often less consisting.
- **Uncertainty**. Another challenge is the inherent uncertainty about where or how the supervisory signal manifests within the tissue. Unlike more traditional tasks in natural image processing, we lack a priori knowledge about the localization or even the existence of such indirect signals in WSIs. This uncertainty complicates the choice of algorithms, as we cannot be sure whether to focus on cellular features, spatial arrangements, or long-range dependencies. However, the absence of definitive locations for these signals presents a unique research opportunity. If predictive algorithms can find a correlation between the WSIs and indirect supervisory signals, then deciphering the visual features responsible could potentially lead to novel discoveries related to phenotypes.

I.4.3 Objectives of the thesis

We have seen that WSIs play a pivotal role in cancer care, offering a multitude of possibilities for automated processing. Predictive algorithms working on WSIs can assist clinicians in daily tasks, serve as a surrogate for other biological modalities, and even contribute to a deeper understanding of cancer biology. However, distinct

challenges are associated with these predictive tasks, largely due to the nature of the supervisory signals.

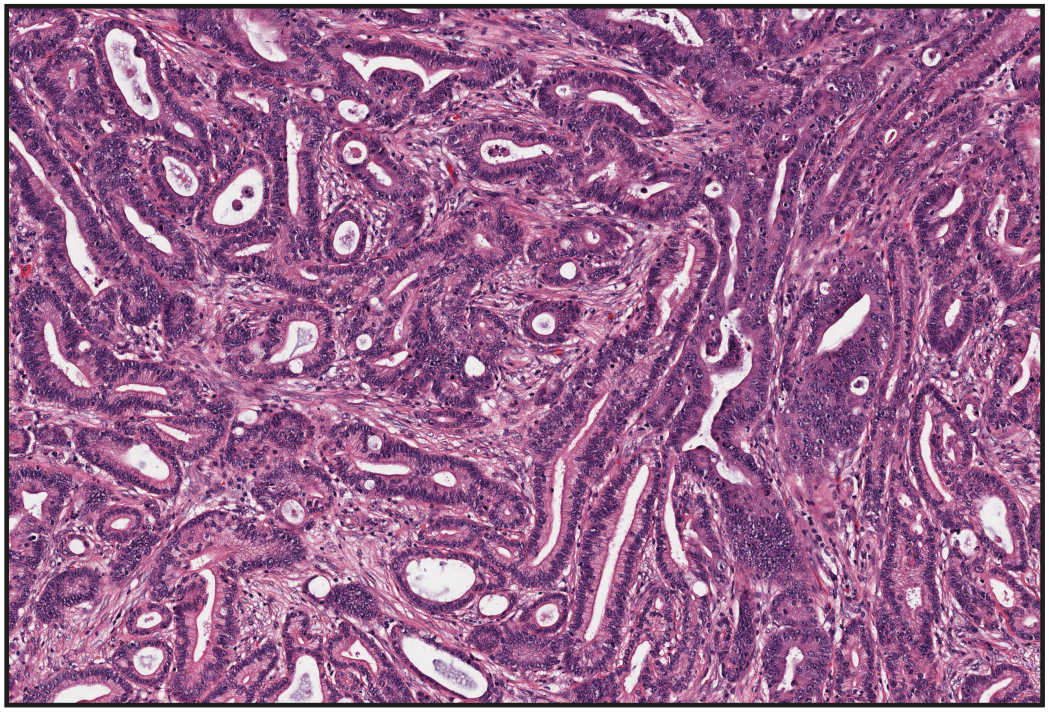
Under the *constraints specific to WSI processing* —such as image size and inherent biases— this thesis aims to address some of the key limitations induced by the supervision signals:

1. How to address the weakness of the slide-level supervision ?
2. How to adapt to the very limited availability of annotation?
3. What is the impact of weak and noisy labels on predictive algorithms, and how can this be mitigated?

Throughout this thesis, we propose potential solutions to these questions. It should be noted that while the individual studies are based on specific biological questions, the primary focus lies on the broader theme of WSI algorithm development. The importance of each unique biological problem solved should not be diminished, although they are independent of the thesis's core focus.

II

Contextualization of the contributions



Contents

Summary:

This chapter serves as a comprehensive guide to the contributions of this thesis by detailing them and contextualizing them within relevant literature. The application of machine learning to Whole Slide Images (WSIs) in histopathology presents a series of unique challenges that distinguish it from other image processing domains. WSIs are characterized by their large size and multi-scale features, and their acquisition process can introduce biases. In terms of supervision, WSIs may be labelled either by medical experts or through in silico experiments. Both types of labels come with specific challenges. They are often weak -concerning only a minute fraction of the WSI-, can be noisy -due to low expert agreements for instance- or uncertain -The question of whether the WSI contains any signal related to the label is often an open question-. We therefore proceed by presenting the solutions related to weak labelling and introducing key training frameworks such as multiple instance learning and self supervised-learning. We further address the challenge of limited labelled data, exploring methods to leverage a small number of regional and global annotations while mitigating associated batch effects. Lastly, we examine the impact of noisy and uncertain labels on model training, presenting both the challenges they pose and the opportunities for machine-teaching they offer.

Résumé:

L'application de l'apprentissage automatique aux images de lames entières (WSI) en histopathologie comporte des défis spécifiques qui la différencient des autres domaine d'application du traitement d'image. Les WSI sont singulières par leur grande taille et car elle peuvent présenter des motifs d'intérêt à plusieurs échelles de grossissement; leur processus d'acquisition peut en outre y ajouter des biais. Les variables cibles des algorithmes d'apprentissages, ou étiquettes, peuvent être annotées soit par des experts médicaux, soit par des expériences et mesures biologiques, référant à un étiquetage *in-silico*. Chaque type d'annotation présente des défis particuliers: elles sont souvent faibles -car ne référant qu'à une petite portion de la WSI, peuvent être imprécises -en raison d'un faible consensus entre les experts annotateurs par exemple-, ou meme incertaines -l'incertitude portant sur l'existence meme d'un signal relatif à l'étiquette au sein de la slide-; et dans tous les cas, elles sont rares. Nous abordons les solutions aux défis d'étiquetage faible en introduisant des algorithmes d'apprentissage clés comme l'apprentissage par instances multiples et l'apprentissage auto-supervisé. Ensuite, nous traitons le problème du nombre limité d'annotations en utilisant des méthodes qui tirent parti de quelques annotations régionales et globales tout en minimisant les effets de biais associés. Enfin, nous étudions l'impact des annotations imprécises ou ambigües sur l'entrainement des modèles, en explorant les défis qu'elles apportent, mais aussi l'opportunité d'*apprentissage par la machine* qu'elles apportent.

Preface

This chapter is structured around the three core questions identified in the thesis's problem statement. The primary objectives are threefold: first, to survey the pertinent literature for each question; second, to delineate how the work conducted in this thesis contributes to resolving these questions; and third, to facilitate a dialogue that interlinks the various research components of the thesis. Contributions of the thesis are framed in gray to improve readability.

II.1 Tackling weak labels

Weak labels in WSI supervision are prevalent and introduce challenges for robust model training. According to Zhou et al. (Zhou 2018), weak labelling or weak supervision can manifest in various ways; and in the context of WSI, the weakness in labels pertains to a specific case of *inexact supervision*, where only coarse-grained information is available.

To elaborate formally, WSIs are inherently large and therefore, unsuitable for direct processing (see Section I.2.3). As a result, these slides are divided into smaller, more manageable images known as tiles. A WSI X can be formally represented as a set of m tiles x_j along with their coordinates c_j, given by $X_i = (x_j, c_j)_{1 \leqslant j \leqslant m} \in \mathcal{X}$, with $\mathcal{X}$ the set of all possible WSI.

We suppose that it exists a surjective function $\mathcal{G}_t : \mathcal{X} \to \mathcal{Y}$, attributing to each slide X a label y, that can be continuous or discrete. The objective is to learn a function $f : \mathcal{X} \mapsto \mathcal{Y}$, using a dataset $D = \mathcal{X}_{\text{train}} \times \mathcal{Y}_{\text{train}} = \{(X_1, y_1), \ldots, (X_m, y_m)\}$, such that on any new slide $X \notin \mathcal{X}_{\text{train}}$, $f(X) = \mathcal{G}_t(X)$.

In this setup, a label y is considered 'weak' if it accurately describes only a minor subset of the tiles in X, that is, $\mathcal{G}_t$ depends on a small portion of X. To illustrate, consider a biopsy X that is labelled as cancerous ($y = 1$) due to the presence of cancerous cells concentrated in a single tile x_j. In this case, y is not representative of all the other tiles $(x_p)_{p \neq j}$.

If **we temporarily disregard the coordinates of the tiles**, this learning framework aligns with the framework of Multiple Instance Learning (MIL) that we describe in the following section. It is therefore unsurprising that the community has readily adopted this framework for the training of WSI predictive models.

II.1.1 The Multiple Instance Learning framework

II.1.1.1. The standard MIL problem

The MIL framework initially gained attention for its application in drug activity prediction, specifically with the musk dataset (Dietterich, Lathrop, and Lozano-Pérez

1997). The dataset consists of various molecules $X_i \in \mathcal{X}$, and the objective is to approximate their activity $y_i \in \mathcal{Y}$ using a function $f : \mathcal{X} \to \mathcal{Y}$. In this case, activity is determined by whether the molecule has a Musk odour or not, 0 or 1, therefore $\mathcal{Y} = \{0, 1\}$. It is important to note that a molecule can have multiple conformations, which are its instances, $x \in \mathbf{x}$ —different shapes it can adopt by rotating around its bonds. If even one conformation binds to a musk receptor, the molecule is considered active, introducing a function $g : \mathbf{x} \to \mathcal{Y}$ that maps a conformation to its activity.

Thus, the dataset consists of molecules and their corresponding activities $(X_i, y_i)_{i \leqslant N}$, with each molecule being a *set* or *bag* of conformations $X_i = \{x_{i1}, \ldots, x_{i,m_i}\}$.

We call a *concept* $\mathcal{P}$ a function that extract a bag statistic from the instance's labels, which is then used to classify the bag with a decision function $\mathcal{C}$. The term $\mathcal{C} \circ \mathcal{P} : \mathcal{Y}^m \to \mathcal{Y}$ represents the MIL problem's assumption, specifically how instance labels aggregate to form the bag's label.

Here, the concept $\mathcal{P}$ asks if at least one instance is positive, and is formally described as:

$$\mathcal{C}(\mathcal{P}(X_i)) = \mathcal{P}(X_i) = \begin{cases} 1, & \text{if } \exists j \text{ such that } g\left(x_{i,j}\right) = 1 \\ 0, & \text{otherwise} \end{cases} \tag{II.1}$$

The challenge lies in identifying g, which is unknown. This constitutes the standard MIL setting, and Equation (II.1) is the **standard MIL assumption**.

The method of breaking down WSI into individual tiles, each potentially containing a range of cell and tissue morphologies, naturally led the community to frame WSI classification as a MIL problem. That's why, for the latter, I may refer to tiles as instances and vice versa.

However, the standard assumption might be insufficient for capturing the complex relationship between instances and the bag label. Consider the problem of tumor grading: one key feature is the mitotic index, or the number of dividing cells within the tumor. A single tile containing a mitotic figure is not sufficient for classifying a WSI as high-grade. In such cases, a broader assumption might be necessary:

Concept: number of instances with mitotic figures

$$\mathcal{C}(\mathcal{P}(X)) = \mathcal{C}\left(\Sigma \mathbf{1}_{\{g(x_i)=1\}}\right) \tag{II.2}$$

$$= \Sigma \mathbf{1}_{\{g(x_i)=1\}} > s \tag{II.3}$$

Assumption: high grade $\leftrightarrow$ more than s mitotic figures

Here, 1 is the label for an instance containing a mitotic figure, and s is a threshold delineating between high and low-grade tumors: the concepts $\mathcal{P}$ counts the number of mitotic figures inside a WSI X and the assumption is to assume that X is high grade if it contains more than s mitotic figures.

Foulds and Frank (2010) review various MIL assumptions that have influenced WSI classification methods. If the assumption linking instance classes to bag labels is known a priori, creating a function to mimic it is advantageous.

However, many WSI classification problems come without such prior knowledge, requiring a more generalized MIL framework.

II.1.1.2. From standard MIL to WSI-MIL

The building blocks To solve the MIL problem using machine learning, we then have to parametrize each of the previously described functions:

1. **A tile-specific function** g that independently processes each instance, here outputting one label per instance.
2. **A pooling function** $\mathcal{P}$, that aggregates instance predictions into a fixed-size *bag concept*.
3. **A classification function** $\mathcal{C}$ that generates a prediction from the bag concept, effectively classifying the entire bag. .

The composition of these three functions defines the WSI classification function:

$$\mathcal{C} \circ \mathcal{P} \circ g(X) = \hat{y} \in [0, 1]$$

.

If $\mathcal{C}$, $\mathcal{P}$, and g are differentiable, a classic classification loss can be computed:

$$L_{\text{classif}}(X) = L_{\text{classif}}(y, \mathcal{C} \circ \mathcal{P} \circ g(X)) = L_{\text{classif}}(y, \hat{y})$$

where y is the label of X. Parameters for each function can then be optimized using Stochastic Gradient Descent (SGD).

The standard assumption given in Equation (II.1) dictates that the function g outputs a single probability for each instance, such as the likelihood of containing a cancerous cell. similarly, $\mathcal{P}$ is typically the *max* function, and $\mathcal{C}$ is the threshold function $\mathbf{1}_{\{\mathbf{x}>\mathbf{0.5}\}}$.

The choice of these functions directly depends on the underlying MIL assumption, which varies according to the problem at hand. For example, when operating under a different assumption as given in Equation Eq. (II.3), it may be necessary to adapt the parametric functions g, $\mathcal{P}$, and $\mathcal{C}$ to align with that specific assumption.

Typically, the MIL assumption at play in WSI classification reflects the architectural choices, specifically the nature and form of g, $\mathcal{P}$, and $\mathcal{C}$. Consequently, most advancements in this area focus on optimizing one or more of these components.

Embedding and classification approaches in MIL The MIL architecture often incorporates an initial step of instance embedding. In domains like molecular activity prediction, this step is essential. Specifically, before classifying a bag of molecule conformations, the first task is to identify a suitable numerical representation for each conformation, denoted as $E(x)$, and taking value in $\mathcal{E}$. Similarly, E can be built to encode WSI's instances in $\mathcal{E}$.

One-step MIL training designates the training of this instance-embedder jointly with the other MIL blocks. Two-step MIL training qualifies the independent training of such a tile-embedder. In this case, it exists a function g' such that $g = g' \circ E$. When MIL operates on these embeddings $E(x)$ the algorithm is said to be trained "on top" of E. g therefore takes input from $\mathcal{E}$.

We can further categorize MIL methods into two types: instance-based and bag-based approaches.

- **Instance-based Methods:** These methods are designed to classify individual instances first, with the subsequent bag-level classification relying on these instance-level results. Such an approach is a straightforward implementation of the fundamental MIL assumptions. In this context, the output of function g serves as a classification result and its dimensionality matches the number of instance classes assumed in the MIL assumption.
- **Bag-based Methods:** Unlike instance-based methods, bag-based approaches focus on constructing a fixed-size representation vector at the bag level, which is then classified. These methods utilize the same overall architecture but diverge slightly from the original MIL assumptions. In bag-based methods, function g projects instances into a new embedding space, and $\mathcal{P}$ aggregates these embedded instances $(g(x_i))_{i \leqslant m}$ into a unified bag representation.

II.1.1.3. A Zoo of MIL Variations

Overall, MIL offers a framework for addressing WSI classification problems, serving as a well-defined pathway for algorithmic advancements. Many WSI classification algorithms evolve from the core MIL framework, with updates typically focused on one or more of the three primary building blocks previously outlined. This section presents a non-exhaustive list of various MIL developments for WSI classification.

Basic MIL I refer to MIL architectures that implement straightforward pooling functions, $\mathcal{P}$, as Basic MIL. These can be further broken down into several categories:

- **Instance-level Max:** Here, the function g serves as a classification network and may be as simple as a single linear layer. The pooling function $\mathcal{P}$ is defined as the *max* function, while $\mathcal{C}$ acts as a threshold function. This setup directly embodies the standard MIL assumption.

- **Instance-level Mean:** This approach is similar to instance-level max, except that the pooling function, $\mathcal{P}$, is the average function defined as

$$\mathcal{P}(X_i) = \frac{\sum_{j=1}^{m_i} g(x_{ij})}{m_i}$$

. The threshold function $\mathcal{C}$ remains unchanged. The underlying assumption is that a bag is classified as positive if the majority of its instances are positive.
- **Bag-level Max:** In this architecture, g can be a more complex encoding function like a Multi-Layer Perceptron (MLP) or Convolutional Neural Network (CNN). The pooling function $\mathcal{P}$ operates feature-wise and calculates their *max* across all instances, resulting in a vector $\mathcal{P}(X_i) \in \mathbb{R}^e$, where e is the size of the instance embeddings. Finally, $\mathcal{C}$ is another classification network, which could be an MLP.
- **Bag-level Mean:** Similar to the bag-level max approach, the pooling function $\mathcal{P}$ calculates the feature-wise average.

Recurrent neural network aggregation A noteworthy adaptation of the standard MIL algorithm is presented by Campanella, Hanna, Geneslaw, Miraflor, Werneck Krauss Silva, et al. (2019). Their work is significant for its application to one of the largest existing WSI database at the time. Their approach employs a two-step MIL process. Initially, they train g, a Resnet50 (He et al. 2015b) using weakly supervised learning, leveraging slide-level labels. They use a simple *instance-level max* model for this first stage. Subsequently, they fine-tune g within another MIL framework: They use the pretrained resnet up to the final linear layer as g. If we name l the final layer of the pretrained resnet, the pooling function $\mathcal{P}$ is $\mathcal{P} = \text{argtop-k}_{x_i}(l(g(x_i)))$, yielding the k best-scoring tiles embeddings with respect to l. They then train $\mathcal{C}$ as a recurrent neural network that aggregates the top-k tiles into the final slide decision. This added layer of complexity is reported to enhance the robustness of their model's predictions.

Positive and negative instance mining Courtiol et al. (Courtiol et al. 2018) focus their MIL algorithmic adjustments primarily on the $\mathcal{P}$ pooling function. With $\updownarrow \circ g$ a ResNet50 mapping to 1 neuron, with g the resnet up to the last linear layer l, their method is similar to the approach by Campanella et al. Specifically, their pooling function is defined as $\mathcal{P} = (\text{argtop-k}_{x_i} l(g(x_i)), \text{arglow-k}_{x_i} l(g(x_i))) \in \mathbf{R}^{2k}$. The bag concept is therefore the concatenation of the k highest and lowest tiles scores. The classification function $\mathcal{C}$ is parameterized as a Multilayer Perceptron (MLP).

The team drew inspiration from Durand et al.'s work (Durand, Thome, and Cord 2016; Durand et al. 2017), which demonstrates the benefit of incorporating *negative* instances into the final prediction. This approach is particularly relevant for natural image classification. For example, the presence of a 'giraffe' instance would logically decrease the bag's probability of being classified as a mountain scene[1].

[1] Not featuring Kilimandjaro, of course.

Attention-based MIL: The attentionMIL algorithm The AttentionMIL algorithm, proposed by Ilse et al. (Ilse, Tomczak, and Welling 2018), has gained considerable attention and often serves as a baseline in numerous studies. Unlike previous approaches that predefine the pooling function $\mathcal{P}$, Ilse et al. introduce a *learnable* pooling function. In their framework, g is a neural network that produces a low-dimensional embedding (of dimension e), rather than a single score.

The pooling function $\mathcal{P}$ is then formulated as a weighted sum of these embeddings. The weights are themselves the output of a neural network parameterized by $\mathbf{V} \in \mathbb{R}^{L\text{e}}$ and $\mathbf{W} \in \mathbb{R}^{L\text{x1}}$. The bag concept therefore writes:

$$\mathcal{P}(g(X)) = \sum_{i=1}^{n} a_i g(x_i)$$

where

$$a_i = \frac{\exp\left\{\mathbf{W}^\top \tanh\left(\mathbf{V}\mathbf{x}_i^\top\right)\right\}}{\sum_{j=1}^{n} \exp\left\{\mathbf{W}^\top \tanh\left(\mathbf{V}\mathbf{x}_j^\top\right)\right\}}$$

Finally, $\mathcal{C}$ is parametrized by a MLP.

This approach adds flexibility to the MIL framework, allowing $\mathcal{P}$ to vary between resembling the `max` or the `mean` function based on the specific problem at hand. Theoretically, this makes it suitable for addressing a broader range of MIL assumptions.

Multi-headed attention-based MIL: the CLAM algorithm The CLAM algorithm extends the attention-based MIL framework to tackle multiclass classification problems. In this approach, g, $\mathcal{P}$, and $\mathcal{C}$ each operate with N parallel attention layers, attention pooling functions, and slide classifier MLP, respectively. As a result, the algorithm outputs N values corresponding to each class, facilitating the computation of a multiclass classification loss.

Additionally, the algorithm incorporates an auto-supervised instance clustering objective, by computation of an instance clustering loss with the instance embeddings $g(x)$. This is designed to optimize g and assist in forming a well-structured instance-level embedding space.

Pooling-based improvements Recent works, particularly those by Oner et al. (Oner et al. 2023) and Schirris et al. (Schirris et al. 2021), have focused on the design of innovative pooling functions, $\mathcal{P}$, capable of capturing more nuanced information from instances.

Oner initially introduced the concept of utilizing discrete estimations of instance-feature distributions as bag representations. Specifically, for a bag of size m and distributions represented with M bins (set as an hyperparameter), the resulting bag

representation would have dimensions $m \times M$. Building upon Ilse's work, Oner extends this idea by incorporating learned attention scores for each instance and uses the discrete estimation of the weighted instance-features distributions-the weights being the attention score-.

On the other hand, Schirris et al. (2021) advocates for employing mean and variance estimations of the marginal distribution of attention-weighted features, thereby resulting in a bag representation with dimensions $m \times 2$.

Expanding on this idea, a possible generalization could involve computing higher moments (M) of the distribution to capture more nuanced details in feature distributions across the bag. We hypothesize that such a generalization would offer a middle ground between Schirris and Oner's methods. Specifically, it would require fewer parameters, similar to Schirris's approach, while preserving maximal distributional information as in Oner's distribution pooling method.

A dynamic field of research I described the earlier algorithms because they served as benchmarks at some point of this thesis: they were once considered state-of-the-art for WSI classification. However, it's important to note that the landscape of research on MIL architectures is both dynamic and expansive: the previous list is not exhaustive at all, and a lot of new MIL architectures and training framework have been since developped (Shao et al. 2021; X. Wang et al. 2023; Xiang and Zhang 2022; Yang et al. 2023; Yu et al. 2023; H. Zhang et al. 2022).

II.1.2 Design of the tile embedder E

While designing a MIL architecture capable of identifying significant tiles in a WSI is a crucial step for solving the weakly supervised problem of WSI classification, it's not the entire solution. Specifically, joint training of the tile-embedding network and the MIL architecture introduces several constraints that must be considered. One primary limitation arises from the computational overhead associated with state-of-the-art image processing networks like ResNets (He et al. 2015b). These architectures are inherently large and resource-intensive. When training a MIL model that has a slide-level objective, each batch would need to include multiple slides. Each of these slides, in turn, comprises a considerable number of individual tiles. Thus, the volume of images that need to be processed by the image network increases substantially, directly proportional to the batch size. To put this into perspective, consider a surgical slide at $10\times$ magnification: on average, it can be divided into around 5000 tiles. This means that joint training would necessitate selecting only a tiny subset of each WSI and generating exceedingly small batches, severely limiting the model's ability to generalize and learn effectively.

II.1.2.1. Transfer learning

Transfer learning has emerged as the predominant approach to mitigate some of these challenges. Initially, transfer learning was conceptualized for applying a model trained on a source data domain to a distinct target domain (Weiss, Khoshgoftaar, and Wang 2016). However, this definition has been broadened to include any model trained for a specific task and later repurposed for a different one.

The advent of CNNs and their subsequent improvements in the ImageNet challenge have been a game-changer for the image processing community (Krizhevsky, Sutskever, and Hinton 2012). ImageNet was groundbreaking as one of the first large-scale datasets of natural images (Deng et al. 2009). It initially comprised around 1.3 million images, spanning a wide array of categories including mammals, insects, man-made structures, plants etc...

Such large, annotated datasets are far less common in the medical imaging realm due to the difficulty in obtaining labelled images. However, we can transfer the models trained on ImageNet to solve poblems in the medical imaging domain.

The study by Kieffer et al. (2017) compared transfer learning strategies of models trained on natural images to models trained from scratch [2] on a histopathology dataset, KimiaPath24, which contains 27,000 images. They found that features[3] extracted from networks pre-trained on ImageNet were highly effective on the histopathology domain. Simply using a Support Vector Machine on top of these pre-trained features—as depicted in Figure II.1 **3**—yielded performance comparable to state-of-the-art networks trained from scratch. Fine-tuning these pre-trained networks further improved performance.

Subsequent studies have confirmed these results across different datasets and tasks (Deniz et al. 2018, 2018; Kensert, Harrison, and Spjuth 2019). Furthermore, transfer learning has also proven effective in a multiple instance learning (MIL) setting: using a model pretrained on ImageNet directly within the MIL architecture as tile-encoder significantly improved performance of WSI classification. (Kanavati and Tsuneki 2021; Sharma et al. 2021)

This suggests that the ImageNet dataset is sufficiently diverse to train networks general enough to extract useful patterns also in histopathology.

II.1.2.2. Self-supervised learning

Leveraging pretrained models on ImageNet has emerged as a common practice in the medical imaging domain. These models offer effective performance and are readily accessible in popular deep-learning frameworks like PyTorch and TensorFlow.

[2] Training from scratch means training a neural networks with randomly initialized weights. For example, PyTorch uses by default the initialization presented in He et al. (2015a).

[3] These features correspond to the activations from a specific layer in a pretrained network, often chosen near the end of the network architecture, as illustrated in Figure II.1.

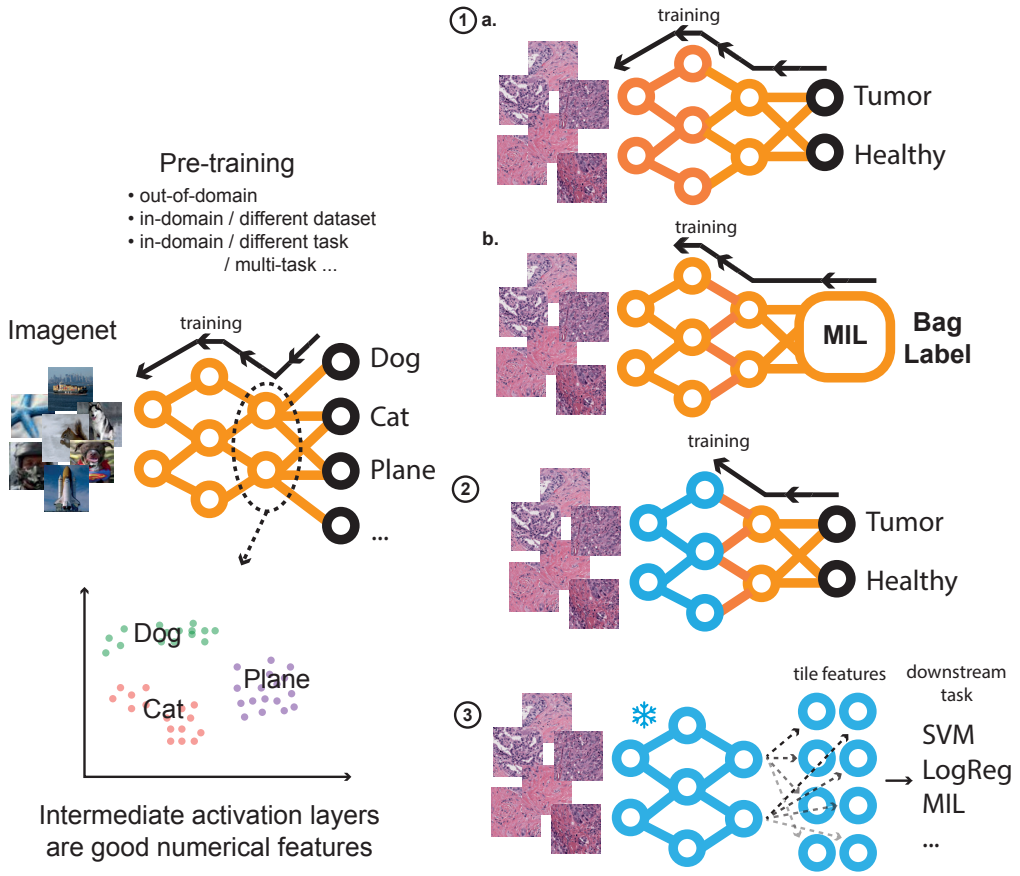


Figure II.1.: **Illustration of Transfer Learning Methods**. A network is first pre-trained on a base task, differing in domain, dataset, or objective from the target task. It can be used in several ways: **1.a-b** Fine-Tuning: The network is retrained on the target task, starting with pre-trained weights. *b* specifies that this network can be part of a larger architecture, a MIL architecture for instance. **2** Partial Fine-Tuning: Only some layers are retrained, reducing computational cost while maintaining performance. Retraining only the last layer is known as *linear probing*. **3** Frozen Networks: The network remains unaltered; intermediate activation weights are used as embeddings.

However, observing the significant domain differences between medical and natural images led to hopes that further improvements could be realized with models pretrained on in-domain data. While the absence of sufficiently large labelled datasets has long hindered the development of in-domain pretrained networks, the medical imaging field is not short on large unannotated image datasets. This is where self-supervised learning comes into play.

At the inception of my Ph.D. journey, spanning from the end of 2019 to the beginning of 2020, self-supervised learning (SSL) witnessed significant advancements. While SSL existed prior to 2020, it achieved major milestones during this period. SSL creates its own supervised task and simultaneously trains a neural network on it—this is what constitutes *self-supervision*. Early examples of SSL tasks include information restoration, learning spatial context, and multi-view invariance, cited in works such as, respectively, Balestriero et al. (2023), R. Zhang, Isola, and Efros (2016), and Noroozi et al. (2018).

Recent advances in SSL such as SimCLR (T. Chen, Kornblith, Norouzi, et al. 2020), MoCo (Xinlei Chen et al. 2020), CPC (Henaff et al. 2019), and BYOL (Grill et al. 2020) have benefited from innovations like the contrastive loss infoNCE and random sampling of negative image pairs. Networks trained using these SSL frameworks have exhibited remarkable performance in linear probing evaluations, coming close to their fully supervised counterparts.

Furthermore, fine-tuning these models has proven to be extremely label-efficient, especially evident when a pre-trained network outperformed a from-scratch model by over 10 accuracy points on 1‰ of ImageNet (T. Chen, Kornblith, Swersky, et al. 2020).

The core idea common to these SSL frameworks, as depicted in Figure II.2, is to train a neural network that maps an image x to an embedding vector $E(x) \in \mathbb{R}^e$ that is invariant to a set of random transformations T. Mathematically, this is expressed as:

$$E(t_1(x)) = E(t_2(x))$$

where $t_1 \sim T$ and $t_2 \sim T$.

Each optimization step in stochastic gradient descent aims to minimize a distance between these embeddings, as illustrated in Figure II.2.

This learning paradigm aims to build robust feature vectors that encapsulate essential characteristics of input images, thus termed *representation learning*. The rationale is that invariant features—those unaffected by random transformations—capture the semantic of the image. For example, both a coloured and grayscale image of a dog, we would easily recognize the dog, and our internal representation of the pictures would be close. This mean that these two images share core features like shape and texture, invariant to colour transformation, and that are used by our brain to build our representations.

However, an obstacle in SSL is the possibility of representation collapse. Indeed, a network E_{collapse} that maps all images x to the same constant c is a trivial solution of this SSL objective:

$$\forall x \; E_{\text{collapse}}(t_1(x)) = c = E_{\text{collapse}}(t_2(x))$$

Each SSL method employs unique strategies to counteract these representation collapse scenarios.

Contrastive learning methods Contrastive learning methods, including SimCLR and MoCo, aim to cluster augmented views of the same image **while pushing views from different images apart**. These methods often employ optimization functions like the Normalized Temperature-scaled Cross-Entropy Loss (NT-Xent Loss) or InfoNCE loss, initially introduced in (Sohn 2016).

Consider T as a distribution of random augmentations, with $t_{ij} \sim T$ for $(i,j) \in \mathbb{N}^2$. Also, let $\mathbf{B}$ denote a mini-batch of pairs of augmented images, denoted as $((t_{i1}(x_i), t_{i2}(x_i)))_{i \leqslant B}$. The encoder and predictor[4] networks generate embeddings $(z_i)_{i \leqslant 2B}$ for the images in $\mathbf{B}$, as illustrated in Figure II.2.

The InfoNCE loss is defined as:

Increases similarity between positive pairs

$$\mathcal{L}_{\text{infoNCE}} = - \sum_{(i,j) \in \mathbf{B}} \log \left(\frac{e^{\text{sim}(z_i, z_j)/\tau}}{\sum_{k=1}^{2B} \mathbf{1}_{\{z_k \neq z_i\}} e^{\text{sim}(z_i, z_k)/\tau}} \right)$$

Over all *oriented* positive pairs

Contrast with any negative pairs

Asymmetric learning methods Methods like SimSiam, BYOL, and DINO use an asymmetric architecture to prevent feature collapse. They optimize a loss that can be schematized by:

$$\mathcal{L}_{\text{asym}} = \frac{1}{B} \sum_{i=1}^{B} \text{dist}(E_\theta(t_1(x_i)), E_\mu(t_2(x_i)))$$

Here, dist is a distance, often the $L2$ distance. E_θ and E_μ are distinct encoding networks. In BYOL, one is the moving average of the other; in SimSiam, they are essentially the same but with an added MLP predictor for one. The asymmetry in the encoding networks is thought to prevent collapse, as discussed in Chaoning Zhang et al. (2022), and the intuition behind it is illustrated in Figure II.2.

[4] A small MLP network, termed the "predictor," is stacked onto the output of the encoder network. The contrastive loss is optimized in the latent space of the predictor's output, but only the encoder network is utilized in downstream tasks. (See T. Chen, Kornblith, Norouzi, et al. (2020) for justification)

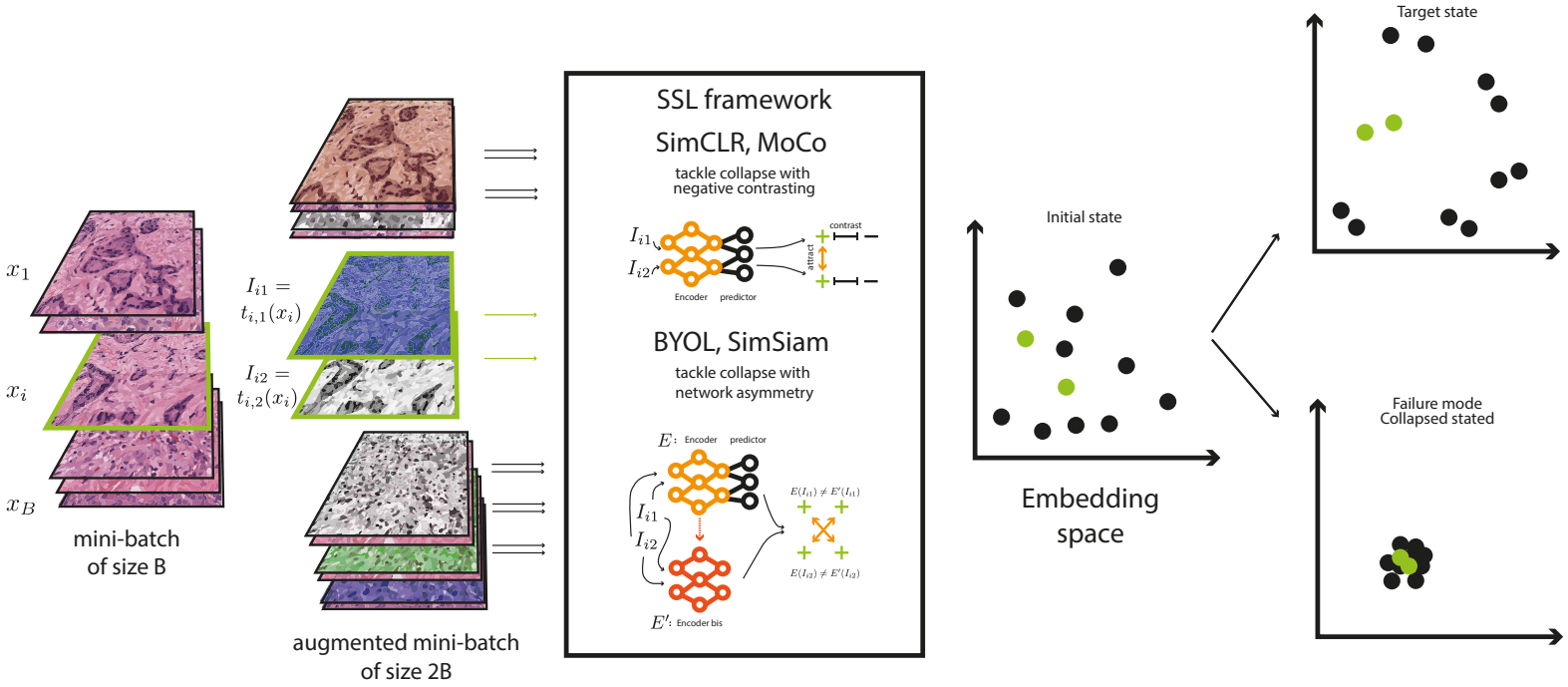


Figure II.2.: Main Frameworks in SSL and Their Rationale Modern SSL tasks rely on random augmentations, or "views," of the same image. The goal is to train a network E that clusters similar views without collapsing them into a single point. Various strategies are employed to achieve this, including contrastive learning methods like SimCLR and MoCo, as well as self-distillation approaches like BYOL and SimSiam. $(x_1, .., x_i, .., x_B)$ a minibatch of images, T a distribution of random transformations, $t_1 \sim T$ and $t_2 \sim T$ two samples of T.

II.1.3 Contributions

SSL is well-suited for histopathology. This domain features small slide-level datasets but large tile-level ones. A single 10x magnification WSI can produce 5,000 to 10,000 tiles, making a tile dataset from 1,000 WSIs much larger than ImageNet. Therefore, I chose to train self-supervised models in this domain early on.

MoCo models I chose MoCo (Xinlei Chen et al. 2020) due to its queue of negative samples, reducing the need for large batch sizes. Initially, SimCLR was thought to require large batches for diverse negative samples, but recent works have debunked this (Balestriero et al. 2023; Bordes, Balestriero, and Vincent 2023).
Training these networks is computationally intensive, therefore limiting my ability to experiment. Key parameters like encoder architecture and data augmentations were carefully chosen but not exhaustively evaluated. Interesting recent research efforts by Kang et al. (2022) have since provided benchmarks and guidelines for SSL in histopathology.
During my PhD, I trained various MoCo models:

Architecture	Organ	dataset	N Slides	magnification	N tiles
ResNet50	Cervix	TissueNet	3062	10x (1 mpp)	1.2 M
ResNet18	Breast	Curie	840	20x (0.5 mpp)	5.3 M
ResNet18	Breast	TCGA	1041	20x (0.5 mpp)	2.2 M
ResNet18	Pan-cancer	TCGA	2000	10x (1 mpp)	1.6 M

The underlined models are made publicly available.

Frozen MoCo embeddings for MIL A key optimization was the direct use of frozen MoCo embeddings as input to MIL architectures in various projects. Bypassing the fine-tuning of the encoder network considerably accelerated the MIL training process. It was crucial, as it allowed me to use more tiles per WSI during training. This increased tile count was especially significant given our MIL model's sensitivity to the number of tiles per slide, as highlighted in Figure B.1.
Evidence across multiple chapters (**Chapters III and IV**) —confirms the advantage of employing an in-domain tile embedder in downstream MIL classification tasks. This improvement was observed across several organs and different classification problems.
Additionally, the findings of **Chapter III** indicate that using a tile embedder trained on a dataset of the histopathological domain, but distinct from the target MIL dataset yields better performance than using ImageNet embeddings. In this case, I trained a MoCo model on the breast cancer slides of TCGA and used it on a breast cancer slide dataset of Institut Curie. Given sufficient training, this approach even surpasses the performance of an embedder trained on the source dataset. However, this transferability appears to be constrained by the similarity between the source and target datasets. Specifically, if the source and target datasets consist in WSI from different organs, the performance gains from transfer learning diminish.

Frozen pre-trained WSI encoder In **Chapter V**, I introduce a pre-training strategy specifically designed for training a MIL architecture without the need for labelled

data. I leverage this strategy to extract generic WSI representations. Downstream tasks can then be solved by training logistic regression models operating on these WSI representations. Remarkably, these logistic regression models, when fed with the pre-trained WSI representations, achieve performance metrics that either match or exceed those obtained from fully supervised MIL methods.
Therefore, since instance aggregation is conducted without supervision in our model, the approach effectively converts a weakly-supervised problem -classification of WSI- into a simple fully supervised one -classification of WSI embeddings-.
To the best of my knowledge, this is the first algorithm capable of generating competitive WSI representations without requiring labelled data.

MIL model used Beyond mere architectural refinements, an ensemble approach provided substantial gains in performance. Specifically, in **Chapter III**, I demonstrated that ensembling multiple MIL models could yield good improvements, achieving up to a 5-point increase in the AUC for the HRD binary classification task. In the same chapter, I also provide a benchmark of some MIL algorithm (mostly these one) and, in accordance with (Ghaffari Laleh et al. 2022), show the suprising efficiency of the most basic MIL pooling functions -bag or instance max or mean-.

Parallel to Chapters III and IV, other research teams have published findings that align with our conclusions on the efficacy of SSL for training tile-embedders (Dehaene et al. 2020; Saillard et al. 2021; Schirris et al. 2021). SSL has rapidly gained traction as a pivotal technique for enhancing MIL training.

Complementing these developments, extensive research has been devoted to understanding the nuances of SSL techniques in histopathology (Kang et al. 2022). Other teams have advanced the state of tile-level SSL through key technological innovations such as Visual Transformers and masked modelling. These innovations have led to substantial performance gains over traditional SSL frameworks (Richard J. Chen, Ding, et al. 2023; Filiot et al. 2023; X. Wang et al., n.d., 2022; Xiang and Zhang 2022).

II.2 Addressing the scarcity of labelled data in histopathology

Collecting a large dataset of labelled histopathological images, whether at the WSI-level or tile-level, poses substantial challenges in terms of cost, time, and feasibility. As elaborated in the introductory section, the rarity of labels in histopathology can be attributed to several factors: the substantial time investment required by experts, the associated costs of molecular profiling technologies, and sometimes the rare incidence of the disease under study.

In front of this observation, our focus here is to survey existing approaches aimed at mitigating this scarcity, particularly in the context of WSI classification.

II.2.1 Combining supervision regimes

One general avenue for improvement for WSI classification is to integrate medical expertise at the more granular level of regions or tiles within the WSIs. Medical experts typically make WSI-level judgments based on a detailed examination of specific regions within the slide. For instance, they may zoom in to assess tissue differentiation in a particular area, evaluate the atypical nature of a cell layer in a tumor region, or identify a unique lesion that can provide a grade for the entire tumor. Their insights can be especially valuable when training end-to-end automatic models like MIL models, as discussed in a zoo of MIL variations.

Given the aforementioned cost constraints, researchers typically face a trade-off: either fully annotate a limited number of WSIs or provide partial, regional annotations for a larger set of WSIs. In either scenario, the quantity of annotations remains low. Consequently, utilizing this expert-level, regional information to enhance WSI classification models becomes a complex task and is an active area of research.

I will here review various approaches aimed at integrating different scales and types of supervision. The idea of enriching the training process through varying levels of supervision is not new. For instance, transfer learning can employ a mix of supervision regimes, utilizing self-supervised learning techniques before transferring the acquired knowledge to a supervised downstream task. Our emphasis here is on the concurrent training of a model under multiple types of supervision operating at different scales. Figure II.3 provides an overview of the various scenarios related to mixed-supervision in this context.

II.2.1.1. Leveraging global labels for regional-level Tasks

A specific set of approaches within mixed-supervision regimes focuses on solving tasks at the regional level. For example, the study by Ciga and Martel (2021) aims to segment histopathological images (not WSIs but region tiles), which is a pixel-level task, by using both pixel-level mask-segmentation ground-truth and weak image-level labels (either cancerous or benign). Similarly, the work by Mlynarski et al. (2019) targets tumor segmentation in magnetic resonance (MR) brain images, using both pixel-level ground-truth and global MR labels for supervision.

These works can be located in quadrant **B.3** of Figure II.3. They have access to a large dataset with global annotations, but only a subset includes detailed pixel-level annotations. This configuration represents a moderately high level of data-labelling cost, and annotation is exhaustive as all pixels in the locally annotated images possess defined labels.

Both studies employ similar strategies involving multi-headed architectures inspired by multi-task machine learning frameworks. In the case of Ciga and Martel (2021), a ResNet18 architecture is used, where the outputs of fourth convolutional block serves as segmentation masks, and a pixel-level segmentation loss is computed with them. Additionally, the framework includes the optimization of a classification loss, derived

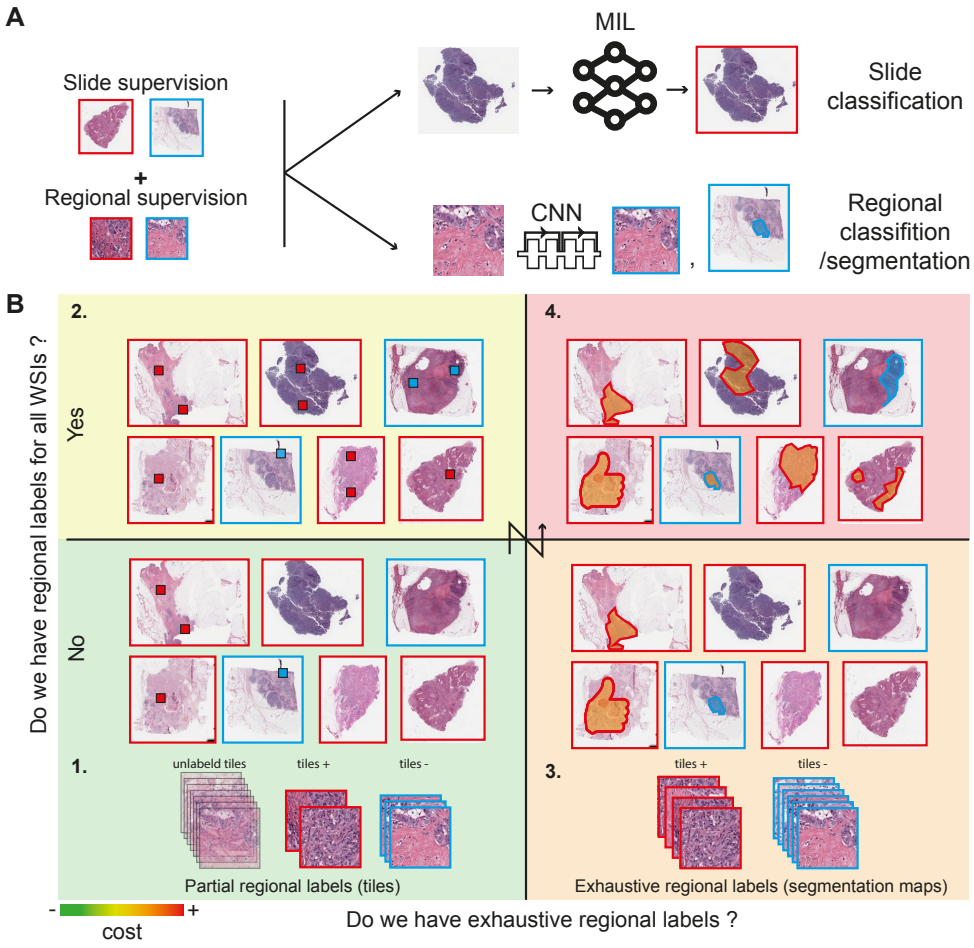


Figure II.3.: Mixed Supervision in Histopathology. **A.** Explains the concept of mixed supervision, which combines regional and slide-level annotations to enhance either slide-level or tile-level classifiers. Here, Weak and strong supervisions are integrated. **B.** Illustrates various dataset configurations for mixed supervision. All WSIs dataset are weakly supervised datasets, i.e. all slides come with a weak global label. The axes represent the availability of regional annotations at both slide and regional levels. Colors indicate the cost of each dataset configuration. Regional annotations may be available for all slides or a subset (rows); they may be partial -i.e. tile annotations- or exhaustive -i.e. segmentation maps- (columns). The arrow indicates the path of increasing cost, as well as the numbering of the quadrants.

from the ResNet18 output. On the other hand, Mlynarski et al. (2019) utilizes a U-Net (Ronneberger, Fischer, and Brox 2015) with an added MLP classification head. Training batches include images both with and without local annotation, and the loss function optimized is a linear combination of classification and segmentation losses.

The shared weights in these architectures are expected to benefit from learning both tasks, thereby improving performance on test data. The results support this expectation, where both classification and segmentation metrics improve when employing dual supervision. The studies also highlight that there exists an optimal ratio for incorporating weak labels. Below this threshold, additional information can still be extracted from weak supervision, whereas exceeding it introduces noise that hampers performance.

II.2.1.2. Enhancing global tasks with regional annotations

Another focus within mixed-supervision research is to enhance global tasks, such as WSI classification, by leveraging a subset of localized annotations. A concrete example is the study by (Tourniaire et al. 2021), which addresses this problem using the Camelyon16 dataset. The Camelyon16 dataset (Ehteshami Bejnordi et al. 2017) comprises 399 WSIs of sentinel lymph nodes, annotated for the presence or absence of metastases. All but 20 of these images are furnished with pixel-level segmentations of all metastases, situating the dataset in quadrant **B.4** of Figure II.3—the most costly but accurate category. The task at hand is to classify WSIs based on the presence of metastases.

Tourniaire et al. (2021) adapted the CLAM algorithm (Lu et al. 2021), which was previously discussed, to incorporate such local annotations. CLAM is an attention-based MIL algorithm with an additional instance clustering objective using tiles pseudo-labels. In the initial algorithm, tiles are pseudo-labelled by the attention mechanism, where high attention scores are expected to correspond to morphological patterns positively associated with the class label, and low attention scores inversely (I question this hypothesis in Section II.3.1.2).
In their adaptation, Tourniaire et al. (2021) use hard labels from the local annotations to randomly sample positive and negative tiles to optimize the clustering objective. This enables CLAM to simultaneously optimize for both tile and slide level objectives. Training proceeds in two steps: an initial phase utilizing slides with local annotations, followed by training on the entire dataset in a weakly supervised manner.

The approach demonstrates efficacy in incorporating local annotations into the MIL framework, resulting in improved slide-level classification. However, this strategy necessitates exhaustive local annotations, meaning segmentations, to facilitate robust sampling of both positive and negative tiles during training.

II.2.1.3. Contributions

Mixed-supervision strategy using inexpensive regional annotations In **Chapter IV**, I introduce a mixed-supervision approach suitable for quadrant 1-2 of Figure II.3. The study's goal, based on the TissueNet dataset from the SFP's first data challenge, was to predict the grade of epithelial lesions in cervical biopsies. The grade of a biopsy is determined by its highest-graded lesion, which can be directly expressed as a MIL assumption. During labelling, expert pathologists identified 3 to 5 regions with bounding boxes and the grade of the lesion present in it, with at least one region containing a lesion that dictates the slide's grade. We thus had both slide labels and a small annotated tile dataset (~5000 images).
Chapter IV propose a multi-head strategy that combines tile-encoder fully-supervised training with self-supervised contrastive learning. This encoder is later integrated into a MIL setup for slide classification.
Table IV.3 demonstrates that pre-training the tile encoder with a small annotated tile dataset significantly enhances performance. However, this is only effective if the tile encoder has undergone prior self-supervised learning. An encoder pre-trained solely on the supervised task did not outperform one with ImageNet pre-trained weights, despite good tile-level classification performance (see Table D.1).

A tile-classifier filter as an alternative to manual segmentation I demonstrate in **Chapter VI** that for the tasks benefiting from preprocessing steps like tumor segmentation (Jakob Nikolas Kather et al. 2020), employing a tile-classifier offers an effective compromise. Specifically, this tile-classifier is trained on a small, randomly chosen subset of tiles' embeddings and utilizes binary labels (keep/don't keep). This strategy effectively balances expert annotation time with the performance of the subsequent classification task.

II.2.2 Dealing with batch-effects

WSI classification datasets often suffer from a limited number of WSI-level labels, generally containing fewer than 1,000 slides. This limitation significantly amplifies the influence of *batch-effects* on algorithms trained on such datasets. Batch-effects refer to variations that occur due to differences in data acquisition, handling, or other experimental conditions. Unlike larger datasets where the impact of spurious correlations is diluted, the problem is exacerbated in smaller datasets.

Compounding this issue is the practice of amalgamating datasets acquired from different healthcare centers to increase dataset size. As previously discussed in the introduction, data acquisition protocols can introduce visual variables that may spuriously correlate with the target variable. The greater the number of contributing centers, the higher the probability of such spurious correlations arising. For instance, Figure II.4 shows that the center of origin often correlates with various classification targets.

In such cases, the risk of predicting the spurious correlate rather than the true target variable is heightened. Furthermore, it has been reported that classification algorithms can even *amplify* existing biases in the data, leading to predictions more

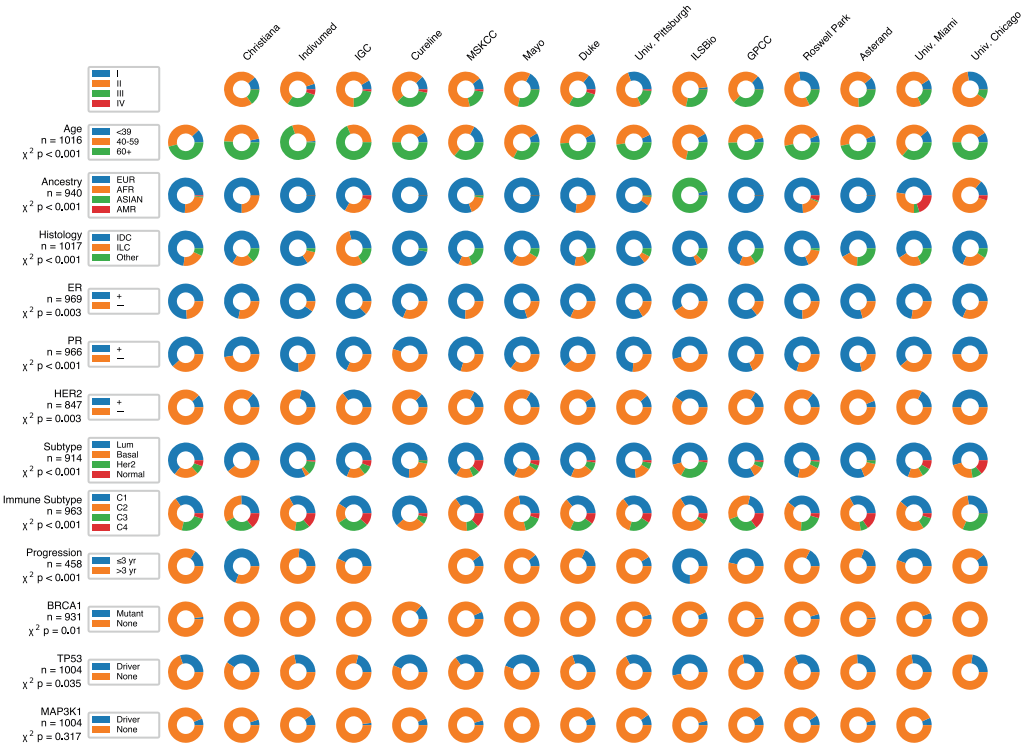


Figure II.4.: Tumor characteristics of breast cancer across sites with 20 or more slides in TCGA. Adapted from (Howard et al. 2021) Each rows is a tumor characteristic, potentially a target variable that one would want to predict from WSI. Columns are different centers. It appears clear that the repartition of the target variable candidate is highly dependant from the center of origin, which could lead to spurious correlation between visual features related to WSI acquisition protocols and the target variables.

correlated with the confounding variable than the actual label (J. Zhao et al. 2017). This phenomenon, known in the machine learning community, is addressed in fair-AI research, where the aim is to make predictions independent of certain *protected classes*, such as ethnic origin or sex.

For histopathology datasets, Howard et al. (Howard et al. 2021) showed that the site of origin for WSIs could be accurately predicted in TCGA. They further proposed a dataset stratification strategy to ensure unbiased evaluation of a model's generalization capabilities, by making sure that the training and testing sets are composed of samples from different centers. This approach led to a significant drop in classification performance in 51 out of 56 tasks across the TCGA dataset when adopting site-aware dataset splits. Some variables even became entirely unpredictable, thereby confirming that the model had learned features that were partially or entirely tied to the center of origin.

Various strategies have been employed to mitigate these challenges outside of the computational pathology field, including of course constructing intentionally unbiased datasets (Richard J. Chen, Wang, et al. 2023). Other methods range from adversarial training that aims to eliminate confounder information (Adeli et al. 2020; Ganin et al. 2016; Q. Zhao, Adeli, and Pohl 2020) to dataset alignment via feature disentanglement (Dwork et al. 2011; Tartaglione, Barbano, and Grangetto 2021). Image normalization techniques, such as grayscale and Macenko normalization (Macenko et al. 2009), have been explored but show limited impact on site-predictability (Hari et al. 2021; Howard et al. 2021; Zanjani, Zinger, and Bejnordi, n.d.).

Z. Wang et al. (2020) benchmarked key bias-mitigation strategies in natural image datasets, revealing that "strategic sampling" is surprisingly competitive, although it requires sufficient samples per attribute of the predicted classes (health centers for instance). Yet, mitigating biases usually comes at the cost of reduced model performance (Richard J. Chen, Wang, et al. 2023).

Figure II.5.: Illustration of the use of a spurious correlation by a classification algorithm. Adapted from (Ribeiro, Singh, and Guestrin 2016). The *explanation* shows portions of the image on which a classification agorithm based its incorrect prediction -wolf instead of husky-.

Biases also complicate model interpretation. Techniques like LIME (Ribeiro, Singh, and Guestrin 2016) can highlight spurious correlations, as demonstrated by the

well-known example where a model mistakingly identified a husky as a wolf based on the snowy background (Figure II.5). In histopathology, biases may manifest in ways not easily discernible by the human eye, making it challenging to separate "good" model interpretations from misleading ones.

II.2.2.1. Contributions

Bias-mitigation strategy In **Chapter III**, I address dataset biases and introduce a strategy for bias mitigation along with an easily interpretable measure for bias in the model's prediction. This study utilizes a dataset of WSIs from the Institut Curie hospital, collected within a period spanning over 15 years. The objective is to predict the status of homologous recombination (HR) in breast cancer patients based on these WSIs. Importantly, the cohort was specifically enriched with homologous recombination deficient (HRD) patients after a particular date—following the demonstration of the clinical significance of HRD status. Coinciding with this date, a modification in the WSI acquisition protocol occurred at the hospital, most notably affecting the fixation step, as illustrated in Figure I.3. This change was found to be a technical confounder, biasing the predictions made by the developed MIL model.
In addition to these technical confounders, the study also identifies biological confounders, specifically the molecular subtype of breast cancer, which likewise biased the HR status prediction. This biological confounder was also present when predicting HR status in the public TCGA dataset. I propose a strategy to mitigate both types of biases through strategic sampling of mini-batches, while aiming to minimize the divergence of this new sampling distribution from a uniform distribution. Until the TCGA study by Howard et al. (2021), dataset biases had not received adequate attention in the field, in my opinion. Indeed, very recently, this question has been ignored even in high-impact studies, thereby raising discussions in the scientific community about the generalizability of reported results (Richard J. Chen et al. 2022; Howard, Kather, and Pearson 2022). My work on bias identification and correction therefore contributed to the awareness of the computational pathology domain to this problem and demonstrates that strategic sampling is an efficient strategy to mitigate bias in digital pathology studies.

Moreover, the most effective bias-mitigation strategy has been found to be the training of separate, independent classifiers for each protected attribute (Z. Wang et al. 2020). However, sub-datasets sharing the same protected attribute are inevitably smaller than the original dataset. In this context, the work presented in **Chapter V** indirectly contributes to this issue by the development of a model that enables the training of robust, generalizable predictors even on tiny datasets.

Using a curated dataset for phenotype-related discoveries In **Chapter III**, I discuss how we utilized deep-learning models and their interpretations to explore the phenotypic consequences of the HR status. I argue that when the goal is to discover new visual correlates, models should be trained on smaller but carefully curated datasets. Given the presence of technical and biological confounders, we opted to use a model trained on a carefully curated, bias-free dataset for subsequent interpretation. This approach allowed us to isolate the un-altered signal related to HR status in the cohort.

II.2.3 Dealing with the size of WSI datasets

At the risk of repeating myself: WSI datasets are often small compared to other domain datasets. First, acquiring labels for these datasets can be a costly endeavour, involving technologies such as Next-Generation Sequencing, RNA sequencing data, or manual annotations by medical doctors. Second, the scale of the healthcare center contributing to the dataset may be limited. Third, the rarity of specific diseases could necessitate prolonged data collection efforts. The datasets used in the early phases of clinical trials often comprise also a small number of patients. Finally, training models on carefully curated sub-populations to mitigate biases could further reduce the size of the dataset.

Consequently, mastering the art of training on small datasets is imperative, as this is a common challenge in WSI processing.

II.2.3.1. Overfitting of MIL models

Traditional MIL models have shown a tendency to overfit when trained on small datasets. Campanella, Hanna, Geneslaw, Miraflor, Silva, et al. (2019) trained their models on a large dataset comprising 44,732 WSIs, roughly equivalent to ∼88 ImageNet datasets in terms of tiles. Their work indicated that the error rate did not saturate even with such a large dataset, and a minimum of 10,000 slides was required to achieve near-clinical classification performance. Below this threshold, the error rates increased disproportionately, pointing to the issue of underperforming models through overfitting. Indeed, a WSI comprises a large matrix of tiles, each containing hundreds of features. However, only a fraction of these tiles is usually relevant for classification, rendering the rest as noise. This high signal over noise ratio may promote overfitting.

II.2.3.2. Improving label-efficiency

CLAM (Lu et al. 2021) aims to enhance label efficiency by adding a clustering objective to the learning process. Apart from optimizing for WSI classification, the model also focuses on clustering the tile embeddings, pseudo-labelled by its attention head. When trained on a small subset of the data, this additional clustering objective led to significant improvements, as evidenced by an increase of 10 AUC points in lymph node metastasis detection tasks and approximately 2 AUC points in lung and renal cancer subtyping.

SSL models have also shown excellent label-efficiency (T. Chen, Kornblith, Swersky, et al. 2020; Azizi et al. 2023). Using embeddings from SSL models pretrained on large datasets can result in good performance even on smaller datasets. This observation holds true at the tile-level (X. Wang et al. 2022; Kang et al. 2022; Ciga, Xu, and Martel 2021). However, the question is still pending on whether this label-efficiency transfers to the slide level when using an SSL pretrained encoder within an MIL architecture.

HIPT (Richard J. Chen et al., n.d.) utilizes hierarchical Vision Transformers (ViT) (Dosovitskiy et al. 2020) and trains them sequentially via the DINO framework (Caron et al. 2021). The approach successfully pre-trains embeddings for regions as big as 4096 square pixels without supervision. A final ViT trained on top of these embeddings with WSI weak labels exhibits improved label-efficiency at the WSI level. Despite these advancements, the technique has yet to achieve success in pre-training complete WSI representations.

II.2.3.3. Contributions

Giga-SSL for increased label-efficiency In chapter **Sections V.1 and V.2**, I introduce Giga-SSL, the first self-supervised learning framework designed to generate embeddings for full WSIs. Utilizing tile encodings from a conventional SSL pre-trained network, this framework focuses on training the aggregation component or MIL component. This framework, which requires approximately 10 hours to train on the entire TCGA dataset, can easily scale to datasets that are orders of magnitude larger.
In terms of label-efficiency, logistic regression models trained on top of WSI embeddings generated by Giga-SSL show remarkable results. Specifically, when utilizing just 50 slides for downstream classification, Giga-SSL outperforms traditional MIL frameworks like AttentionMIL (**Section V.1**) and CLAM (**Section V.2**) by an average of 6.3 and 7 AUC points across five classification tasks.
Lastly, Giga-SSL's effectiveness extends to predicting point mutations in the TCGA dataset, as demonstrated in **Section V.2**. We hypothesize that this success can be attributed to two factors: first, the use of an SSL pre-trained tile encoder, which is not employed in the MIL model used for comparison (Jakob Nikolas Kather et al. 2020); and second, the cumulative label-efficiency advantages of Giga-SSL. This is particularly relevant for mutation prediction tasks that frequently involve highly imbalanced classifications; the mutant class is significantly less common than the wild type. Such a situation can closely resemble a scarce data regime: the minority class consists of very few individuals. Supporting this, it has been shown that SSL pre-training enhances performance in highly imbalanced tasks (Chuyan Zhang, Zheng, and Gu 2023).

II.3 Noise and uncertainty of labels

Noise in labels refers to incorrect label assignments within a dataset. Such noise may stem from a variety of sources. Human error during the labelling process is a common origin. Additionally, the 'distance' between the WSI and the label acquisition can introduce noise (see Figure I.4). For example, a label might be determined based on measurements from a different tissue block than the one displayed in the WSI. This could result in the label being influenced by a different clonal population of cancerous cells. Or again, if the label is linked to patient-level outcomes such as survival data, which could be influenced by other factors like environmental or societal variables. Generally, the further the label and the WSI

are from each other in terms of data acquisition, the greater the likelihood for noise since additional causal factors apart from the WSI can influence the label.

Label uncertainty is more ambiguous than label noise and generally applies to the entire dataset. This uncertainty falls into two categories:

1. Strong uncertainty arises when it's unclear whether a discernible signal exists in the WSI.
2. Weak uncertainty arises when the visual features supporting the label are not known a priori.

The primary source of this uncertainty is our incomplete knowledge of the phenotypic markers for the variables we aim to predict. This gap in understanding leads to challenges in algorithm selection for these ill-defined problems and complicates the task of establishing fair benchmarks.

On the positive side, the prediction of uncertain labels can reveal complex and previously unknown statistical links between input and output variables, and thus help formulating hypotheses about causal relationships. In this scenario, AI acts like a hypothesis generator potentially revealing new morphological biomarkers, and ultimately contributing to our understanding of disease mechanisms.

II.3.1 Uncertainty: deep learning as a machine-teaching tool

To extend our understanding of trained predictive models, the field of model interpretability offers a robust toolkit.

II.3.1.1. Interpretability in machine learning

Interpretability serves multiple purposes. One primary use is to refine the architecture of a model by focusing on its failure modes, thereby aiding in debugging and pointing to limitations of the tested method. It can also uncover biases in the data, exemplified by the wolf/snow bias, as discussed in Figure II.5. Further, interpretability ensures that the machine learning models align with societal values, particularly in the context of fairness and protected attributes. Given its broad applications, interpretability has become a subject of extensive research across all machine learning disciplines. The techniques within this field can generally be categorized into two types: local explanations and global explanations.

Local Explanations Local explanations aim to explain the reasoning behind a model's specific prediction on an individual data point. This involves analyzing the model's behaviour in the proximity of the data point. Techniques such as SHAP (Lundberg and Lee 2017) and LIME (Ribeiro, Singh, and Guestrin 2016) assess how altering a data point influences the model's decision. Specifically, LIME masks random sections of images to observe model behaviour, whereas SHAP examines

feature importance with feature mixing. Methods like GradCAM (Selvaraju et al. 2020) and its derivatives (Shrikumar et al. 2017; Smilkov et al. 2017; Sundararajan, Taly, and Yan 2017) backpropagates the gradient of the predicted logits on the input image - illuminating the significant regions in an image related to a specific prediction. However, these approaches have limitations (Binder et al. 2022) and should not be used exclusively. Counterfactual explanations also offer a way to scrutinize a model's decision boundary (Zemni et al. 2023). The objective is to identify an image that is semantically closest to the original but would result in a different prediction from the model.

Employing local explanations is particularly beneficial for understanding a model's failure modes. In medical applications, such as histopathology, these explanations can guide clinicians or pathologists by offering model-derived insights that can be compared to human interpretations.

Global Explanations Global explanations extend beyond the scope of individual predictions to provide insights at the dataset level. The focus here is not on explaining a model's choice for a specific sample but understanding the features in the dataset that influence the model's general decision-making. Often, deriving global explanations requires starting with local explanations. The next step is to aggregate these local explanations across the dataset, which presents its own challenges. The difficulty lies in designing local explanation methods allowing for full dataset integration - for example, there is not straighforward way to integrate GradCAM heatmaps at the dataset scale. Here, both the local explanation and the aggregation method need to be carefully chosen in order to provides meaningful results.

A prominent approach for global explanations is TCAV (Testing with Concept Activation Vectors) (B. Kim et al. 2018). In this method, the user defines a "concept" by grouping images that represent it. For instance, the concept of 'stripes' could be represented by a set of images featuring striped objects or animals. A designated intermediate layer l in the classification neural network is selected, and its activations are computed for the concept images as well as a random set of images. A linear classifier is then trained on these activation vectors to define a concept activation vector v_C.

Subsequently, for each image-prediction pair, the gradient of the activations at layer l concerning the prediction is computed Δ_l. This gradient indicates the direction along which the prediction varies. The dot-product $E_C = \Delta_l \cdot v_C$ measures the influence of that concept on the prediction and is suitable for aggregation to create global explanations. For example, the average measure for all images classified as 'zebra' would likely indicate a high influence of the 'stripe' concept.

Further developments have improved upon these concept-based methods. Recent research aims to define 'complete' concepts that are sufficient to explain a decision (Yeh et al. 2020). Other work integrates the concept design directly into the network architecture (Koh et al. 2020).

II.3.1.2. Interpretability in WSI processing

In the field of histopathology, interpretability also has the potential to guide new scientific discoveries. One contributing factor to the widespread use of MIL in WSI processing is its architecture, which allows for relatively straightforward interpretation. Specifically, the functions g or $\mathcal{P}$ (see this section) can incorporate tile-specific information such as classification scores.

As a practical example, Courtiol et al. (Courtiol et al. 2019a) employed a similar approach to predict survival in mesothelioma and used tile classification scores to isolate tiles with a correlation to survival. Likewise, models like the AttentionMIL compute attention scores for individual tiles. These scores can serve as an indicator of the tile's relative importance in making the overall MIL prediction.

A prevalent approach for interpreting these scores involves generating heatmaps directly on the WSIs (see Figure II.6). This method is common in most WSI classification studies using deep learning (Ehteshami Bejnordi et al. 2017; Lu et al. 2021; Qu et al. 2021; Schmauch et al. 2020; B. Xu et al. 2019) and provides a **local explanation**, allowing users to identify regions that played a pivotal role in model predictions.

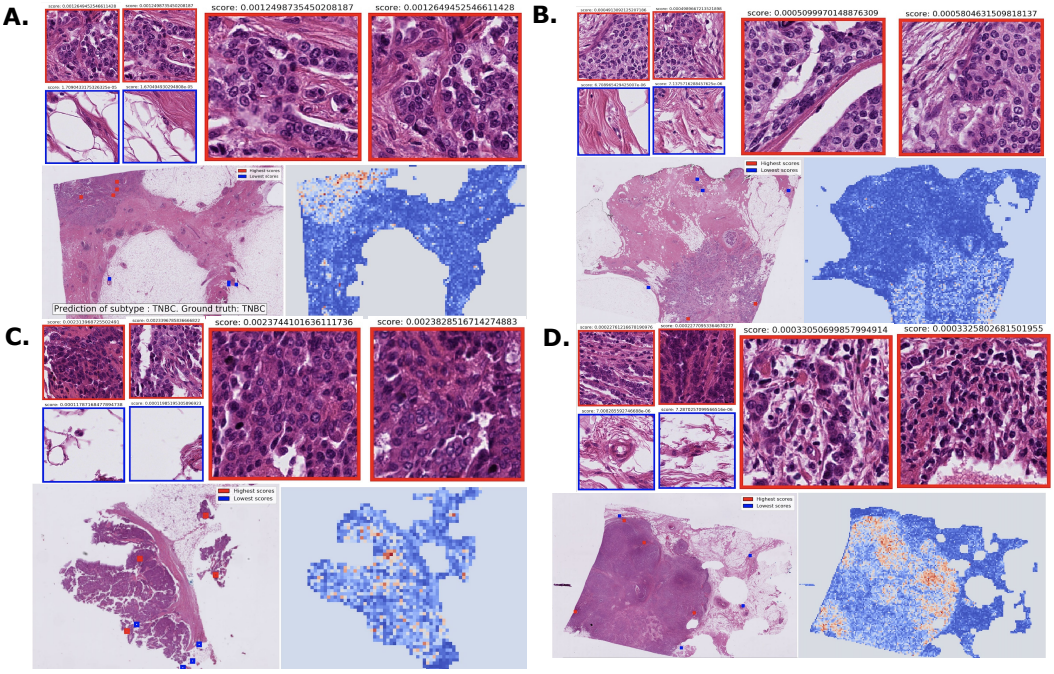


Figure II.6.: Local attention-based visualization of the best (in red) and worst (in blue) scoring tiles, as well as the heatmaps for four successful predictions of the molecular class and HRD. **A.** Visualization for a triple-negative breast cancer (TNBC) WSI. **B.** Visualization for a triple-negative breast cancer (TNBC) WSI. **C.** Visualization for an HRP WSI. **D.** Visualization for an HRD WSI. Outputs can be obtained with my open-source repository, wsi_mil

Moreover, these MIL-derived tile scores can facilitate the localization of ROI. For example, attention scores have been used for weakly-supervised segmentation tasks, such as localizing metastasis in the Camelyon dataset. These segmentation performances can sometimes also serve as an indirect measure of the overall MIL algorithm's effectiveness.

Global explanations often rely on these tile scores as well. One can pool together tile scores from an entire dataset and identify the highest and lowest-scoring tiles. While interpreting these scores is generally straightforward for models that directly implement MIL assumptions (Courtiol et al. 2019b), it becomes less clear when dealing with attention scores. In the case of CLAM (Lu et al. 2021), high attention scores are interpreted as positive evidence for a class prediction, and low scores as negative evidence, an assumption we consider to be generally misleading, because nothing in the model construction assures that low attention tiles will negatively participate to the prediction.

Indeed, following the assumption about attention scores, low-attention scores translates to useless tiles. However, we also find this "importance" interpretation of the attention score quite misleading.

In models like AttentionMIL, the WSI representation vector V_X is computed as a weighted sum of the tile embeddings: $V_X = a_X \cdot g(\mathbf{x}_X) = \sum_i a_{\mathbf{x_i}} g(\mathbf{x_i})$, where a_X is the vector of attention scores for each tiles of X. The influence of a tile $\mathbf{x_i}$ in the norm of the bag representation V_X is thus proportional to its attention score a_i:

$$\|\Delta V_X(\mathbf{x_i})\| = a_i \|g(\mathbf{x_i})\|$$

However, this influence is also dependent on the norm of the tile embedding $\|g(\mathbf{x_i})\|$. When g is trained jointly with $\mathcal{C}$ (the attention network), and there is no normalization of embeddings, interpreting the attention score becomes nontrivial and must be considered alongside the norms of the embeddings. This complexity could potentially explain the subpar performance of AttentionMIL in weakly supervised segmentation tasks.

II.3.1.3. Contributions

A Global Explanation Algorithm for AttentionMIL To address the interpretability challenges in AttentionMIL, I first removed the tile embedding network g as described in **Chapter III**. Therefore, $\mathcal{P}$ acts directly on the pre-trained tiles embeddings $E(\mathbf{x})$. This step was essential for resolving ambiguities in the interpretation of attention scores, highlighted in **Chapter III**. Specifically, both negative and positive evidence for a prediction can exhibit high attention scores, a phenomenon that can mislead interpretation efforts.
In the same chapter, I introduce a new algorithm aimed at providing a global explanation for AttentionMIL. This algorithm leverages attention scores to filter out less relevant tiles and thus to focus the subsequent steps on highly relevant tiles only. It uses the fact that the bag representation V_X and the tile embeddings $E(\mathbf{x})$ live in the same vector space. This enables the decision

network, initially trained to classify V_X, to be repurposed for classifying individual tile embeddings. We interpret these tile classification scores as their "signed" contribution to the overall WSI prediction: tiles with positive scores provide positive evidence, while those with negative scores serve as negative evidence. This procedure of scoring an individual tile using the decision network $\mathcal{C}$ can be interpreted as classifying a WSI that predominantly consists of this specific tile.

New morphological patterns associated to HRD The resulting interpretations themselves stand as contributions to the field. In **Chapter III**, the interpretation approach not only corroborated the association of HRD with well-known morphological features like necrosis and high nuclear atypia but also uncovered a new relationship with a specific kind of fibrosis surrounding the tumor, termed laminar fibrosis.

WSI latent space interpretation In **Section V.3**, we put forth flexible methodology for interpreting the latent space of WSI embeddings -outputs of the Giga-SSL models- using linear models. Our method involves projecting the classifier's hyperplane onto a series of user-defined concepts, or interpretable WSI labels, such as the number of lymphocytes or the size of the tumor. This process aids in identifying specific morphological profiles that are descriptive of a given task, thereby providing a "global explanation" for it.
We demonstrate the scalability of this approach and its applicability to datasets that were not part of either the Giga-SSL model's training or the interpretation method's parametrization.

Recently, a notable advancement in the area of morphological prognostic factors has been made through a two-part study. In the initial phase, Wulczyn et al. (2021) utilized a deep learning system to predict overall survival in patients with colorectal cancer. Alongside the predictive model, the researchers developed an interpretation method that identified a novel histo-prognostic factor, Tumor Associated Fat (TAF).

Subsequently, L'Imperio et al. (2023) validated this newly discovered feature with the help of human pathologists. The pathologists were trained to recognize TAF using the data patches extracted from the initial study (Wulczyn et al. 2021). After the training, these pathologists independently graded slides from a new colorectal dataset, with respect to the TAF pattern. Their assessments not only displayed significant prognostic value but also exhibited reasonable inter-pathologist agreement.

This sequential approach is a convincing proof-of-concept for *machine-teaching* in the medical field. Adopting a similar validation protocol for the morphological pattern that we found associated to HRD (laminated fibrosis for instance) would offer a promising perspective.

II.3.2 Decrypting and mitigating label noise

Noise in the labelling process is a common issue in the field of medical imaging, and it can affect both tile-level and WSI-level tasks in computational pathology. The sources and types of this label noise can vary significantly. For example, one form of

noise stems from expert annotation errors. This can be evidenced by the relatively low agreements of human annotators on challenging cases (Costantini et al. 2003; Krane et al. 2022).

To explore the impact of such expert-induced label noise, a study by Hekler et al. (2020) focused on a skin cancer classification dataset derived from dermatoscopic images. This dataset contains both expert-provided labels and biopsy-confirmed labels, the latter being considered noise-free. Their findings revealed that CNNs trained on expert labels exhibited a 10% drop in accuracy when tested on a noise-free dataset, compared to their performance on an expert-labelled test dataset.

It is commonly assumed that difficult or borderline cases are the one responsible for label noise. These ambiguous cases, that can exhibit characteristics of several classes simultaneously, often demonstrate low inter-observer agreement among experts. But does this uncertainty observed within expert assessments translates to the trained algorithm ?

A growing body of research is aimed at developing reliable uncertainty measures for neural networks (Lubrano et al., n.d.; Mehrtens et al. 2023). Utilizing such confidence scores may help mitigate the impact of label noise during network inference, particularly by excluding predictions with high uncertainty (Mehrtens et al. 2023).

Another facet of label noise relates to cases where the provided labels may not accurately capture the biological state in the image. For example, the binary labels used for HRD prediction, as discussed in Chapter III, are derived from a continuous measure. Such arbitrary binarization can introduce noise and provoke a loss of information, particularly for borderline cases. In these instances, Nahhas et al. (2023) suggests that regression-based deep learning models, trained on continuous labels, perform better than their classification-based counterparts.

However, it's worth noting that the existing solutions for mitigating label noise primarily address issues with borderline cases. Mehrtens et al. (2023) specifically identifies different types of label noise in their work, which benchmarks various confidence scores for machine learning predictions. They explore a tile-classification task using the Camelyon17 dataset, where the labels are either 'tumor' or 'healthy.' To study the impact of noise, they create two distinct datasets with controlled levels of label noise.

- The *border dataset* is created by flipping the labels of tiles at the border of the tumor area. These tiles are often borderline cases containing both cancerous and healthy tissue.
- The *uniform dataset* involves flipping labels randomly across the entire slide. Unlike the *border dataset*, the *uniform dataset* includes blatant mislabelling of clearly tumorous or healthy tiles. This type of noise is considered unrealistic as it doesn't mirror the variability commonly seen in real-world expert annotations.

While existing confidence measures effectively identify mislabelled samples in the *border dataset*, they fall short in detecting erroneous labels in the *uniform dataset*.

Therefore, current approaches are limited in addressing label noise that is not associated with borderline cases.

II.3.2.1. Contributions

Label noise due to tumoral heterogeneity In the study outlined in **Chapter VI**, I aimed to predict the molecular subtypes of cholangiocarcinomas using RNA-seq analysis on tumor samples. These labels were attributed at the patient level. For each patient, we had access to multiple WSIs, some of which originated from the same sample used for RNA-seq analysis, termed *paired WSIs*, and others from different samples, termed *unpaired WSIs*.
Interestingly, the results showed that models trained on all available WSIs performed worse than those trained exclusively on *paired* WSIs when subsequently tested on *paired* WSIs. I hypothesize that this reduction in accuracy is attributable to noise introduced by the *unpaired* WSIs. Specifically, I suspect that intratumoral heterogeneity is the source of this noise.
This conjecture is supported by the nature of intratumoral heterogeneity itself. Because of this heterogeneity, certain *unpaired* WSIs may exhibit characteristics that are vastly different from their corresponding *paired* WSIs, even though they share the same label. This phenomenon would introduces the type of noise that was previously considered unrealistic by the *uniform dataset*.
We thus hypothesize that such noise exists in WSI classification tasks and that it may be a widespread issue.
For example, I speculate that most TCGA-based classification tasks, such as those explored in **Section V.2**, are likely subject to this specific kind of label noise. This speculation is further substantiated by the study conducted by Jakob Nikolas Kather et al. (2020). Surprisingly, they found that many genetic signatures and mutation tasks were more accurately predicted using the frozen WSIs from TCGA, despite their ostensibly lower quality compared to FFPE slides. This observation, which remained unexplained, may be accounted for by the influence of intratumoral heterogeneity. Indeed, it's important to note that the biological assessments used to produce genetic signature labels in TCGA are based on these same frozen samples, making them *paired* slides, while the FFPE slides are *unpaired*.

III

Deep-learning identifies morphological patterns of homologous recombination deficiency

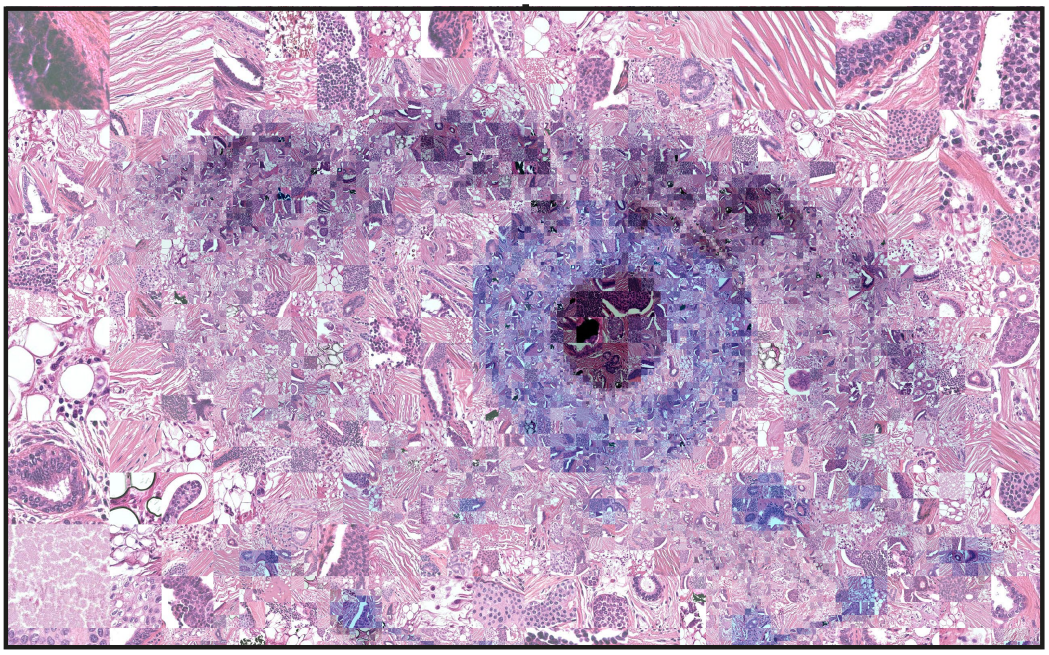


Figure III.1.: Mosaic of breast cancer tiles from the Curie Institute. Made thanks to github.com/xstc55/ImageMosaicBVH.

Contents

Preface

This work, my first PhD project, spanned almost the entire duration of my doctoral studies and resulted in a published article. It underwent many twists and turns, and its structure- which includes several focus points- reflects our research journey.

The idea of Guillaume Bataillon and Anne Vincent-Salomon initiated the project: predict Homologous Recombination (HR) status in breast tumors using only H&E stained slides. HR-deficient patients have been shown to be susceptible to specific medications like PARP-I or platinum salts, but current molecular tests for HR status are costly. The clinical implications of such a prediction tool were evident, although the goal was ambitious: There were indeed no known markers for HR status in WSIs, and it was uncertain if WSIs even contained this information.

Despite these challenges, our classification network yielded very good results, but they seemed too good to be true. Further investigation into the model's predictions revealed a staining inconsistency within our dataset; some slides were stained with HES, while others were stained with H&E. After re-staining the entire cohort, thanks to the insights of the clinicians, we identified three additional variables that could confound HR status prediction. This revelation prompted a deeper investigation into the potential biases affecting predictive models and the means to prevent it.

Regarding the algorithm developpement,the recent success of self-supervised learning methods (SSL) in natural image processing inspired us to explore their applicability to WSIs. This led to the development of the two-step multiple instance learning algorithm detailed in the paper.

Lastly, motivated by the capabilities of deep neural networks for disease analysis, I addressed shortcomings in current attention-based visualization techniques. This led to a new method centered on the decision scores of the network. The method facilitated important morphological findings, which, when medically interpreted, became a crucial aspect of the paper.

In summary, the tight collaboration between computational biologists and clinicians from the Institut Curie was instrumental at every phase of this project. Each decision point was influenced by this interdisciplinary approach and concerns, shaping the work into its final form. Additionally, this work catalyzed further research at the Institut Curie. Open-sourcing the code for WSI encoding, model training, and interpretation spurred its use in a series of related projects, including predicting prognostic factors in adenocortical tumors, studying the phenotypic effects of ATM mutations in breast cancer, and utilizing the classification models in a clinical trial related to HRD.

Contributions

Publications - communications

- **Lazard, T.***, Bataillon, G. * et al. (2022). Deep learning identifies morphological patterns of homologous recombination deficiency in luminal breast cancers from whole slide images. Cell Rep Med 3, 100872. 10.1016/j.xcrm.2022.100872.
- Jouinot, A., Violon, F., **Lazard, T.**, et al. (2023). Analyse d'images histologiques par intelligence artificielle (deep learning) pour l'évaluation pronostique des corticosurrénalomes. Annales d'Endocrinologie 84, 556–557. 10.1016/j.ando.2023.07.136.

Open-source repository

- WSI-MIL: software for the training and prediction of MIL models for WSI analysis and extraction of morphological patterns.

Summary:

Homologous recombination DNA-repair deficiency (HRD) is becoming a well-recognized marker of platinum salt and polyADP-ribose polymerase inhibitor chemotherapies in ovarian and breast cancers. While large- scale screening for HRD using genomic markers is logistically and economically challenging, stained tissue slides are routinely acquired in clinical practice. With the objectives of providing a robust deep-learning method for HRD prediction from tissue slides and identifying related morphological phenotypes, we first show that digital pathology workflows are sensitive to potential biases in the training set, then we propose a method to overcome the influence of these biases, and we develop an interpretation method capable of identifying complex phenotypes. Application to our carefully curated in-house dataset allows us to predict HRD with high accuracy (area under the receiver-operator characteristics curve 0.86) and to identify morpho- logical phenotypes related to HRD. In particular, the presence of laminated fibrosis and clear tumor cells associated with HRD open new hypotheses regarding its phenotypic impact.

Résumé:

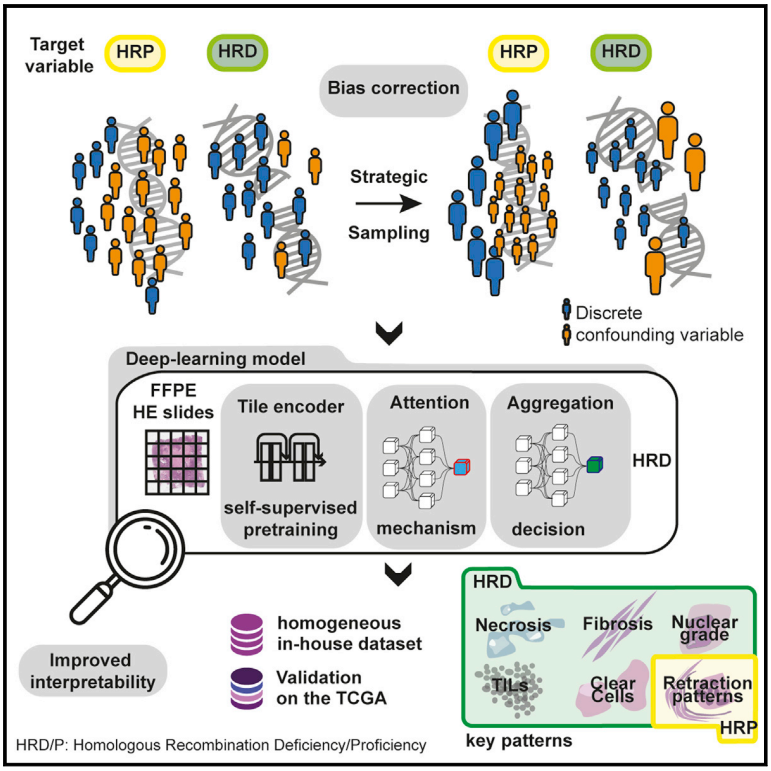
Le déficit en réparation de l'ADN par recombinaison homologue (HRD) est un marqueur clé pour certaines chimiothérapies dans les cancers de l'ovaire et du sein. Le dépistage à grande échelle de la HRD est complexe et coûteux, mais au contraire les lames de tissus sont couramment utilisées en clinique. Nous proposons ici une méthode d'apprentissage profond pour prédire la HRD à partir de ces lames et identifier les phénotypes qui y sont associés. Nos résultats montrent que les méthodes de pathologie numérique peuvent être biaisées et nous offrons donc une solution basée sur l'échantillonnage des *mini-batch* d'apprentissage pour y remédier. En utilisant notre méthode sur des données internes, nous pouvons prédire la HRD avec une bonne précision (AUC : 0,86) et identifier les phénotypes morphologiques spécifiques associés.En particulier, la présence de fibrose stratifiée et de cellules tumorales claires associée à la HRD ouvre de nouvelles hypothèses concernant son impact phénotypique.

Cell Reports
Medicine

Article

Deep learning identifies morphological patterns of homologous recombination deficiency in luminal breast cancers from whole slide images

Graphical abstract



Highlights

- Homologous recombination deficiency is predictable from H&E slides with high accuracy
- Biases in computational pathology data can be alleviated by strategic sampling
- We present a method to identify morphological patterns of complex phenotypes
- We identified five HRD- and two HRP-related morphological patterns

Authors

Tristan Lazard, Guillaume Bataillon, Peter Naylor, ..., Etienne Decencière, Thomas Walter, Anne Vincent-Salomon

Correspondence

thomas.walter@mines-paristech.fr (T.W.), anne.salomon@curie.fr (A.V.-S.)

In brief

Deep-learning models predict homologous recombination deficiency (HRD) from H&E-stained pathology slides. Dataset biases, either biological or technical, can be alleviated by strategic sampling. Interpretation of the predictive models reveals several morphological patterns related to HRD and opens new hypotheses about its phenotypic impact.

Lazard et al., 2022, Cell Reports Medicine *3*, 100872
December 20, 2022
https://doi.org/10.1016/j.xcrm.2022.100872

CellPress

This chapter has been made in collaboration with G. Bataillon (with whom I share first autorship), P. Naylor, T. Popova, F-C. Bidard, D. Stoppa-Lyonnet, M-H. Stern, E. Decencière, T. Walter and A. Vincent-Salomon. It has been published in Cell Reports Medicine.

III.1 Introduction

Worldwide, 2.1 million women are newly diagnosed per year with breast cancer (BC), which is a leading cause of cancer-related death. Improvement of metastatic BC treatment is therefore of highest priority. BC is a heterogeneous disease with four major molecular classes (luminal A and B, HER2 enriched, and triple-negative breast cancer [TNBC]) benefiting from different therapeutic approaches. If early BC patients have an overall survival of 70%–80%, metastatic disease is incurable with a short duration of survival (Deluche et al. 2020) . Homologous recombination (HR) is a major and high-fidelity repair pathway of DNA double-strand breaks. Its deficiency, HRD, results in high genomic instability (Miller et al. 2020) and occurs through diverse mechanisms, including germline or acquired somatic mutations in DNA-repair genes, most frequently BRCA1, BRCA2, or PALB2, or through epigenetic alterations of BRCA1 or RAD51C. Importantly, HRD leads to high sensitivity to polyADP-ribose polymerase inhibitors (PARPi) in vitro, (Bryant et al. 2005; Farmer et al. 2005) a treatment that has been shown to improve metastatic BC progression-free survival (Tung et al. 2020; A. N. J. Tutt et al. 2021). HRDs induced by BRCA1 and BRCA2 mutations are known predictive markers for response to PARPi (Miller et al. 2020; A. N. J. Tutt et al. 2021) and platinum salt (A. Tutt et al. 2018), and somatic HRD has been more recently recognized as a predictive marker for PARPi in ovarian cancer (Miller et al. 2020) and BC (Chopra et al. 2020).

Several methods have been developed to detect HRD, including genomic instability profiling, mutational signatures, or integrating structural and mutational signatures (Popova et al. 2012; Birkbak et al. 2012; Abkevich et al. 2012; Polak et al. 2017; Davies et al. 2017). Today, HRD is diagnosed in clinical practice by DNA-repair gene sequencing, germinal in BCs and somatic in ovarian cancers, respectively. For ovarian cancers, HRD is also assessed by genomic instability tests such as the HRD MyChoice CDx test (Myriad Genetics).

The majority of hereditary BRCA1 cancers are TNBC and up to 60%–69% of sporadic TNBCs harbor a genomic profile of HRD.(Chopra et al. 2020; Popova et al. 2012; Alexandrov et al. 2013) In contrast, the majority of hereditary BRCA2 cancers are luminal (Lakhani et al. 2002), and HRD also exists in sporadic luminal B (Manié et al. 2016), or in HER2 tumors (Ferrari et al. 2016; Turner 2017). Of note, germline or sporadic alterations of BRCA harbor indistinguishable genomic alterations in triplenegative or luminal tumors (Manié et al. 2016; Holstege et al. 2010). Also, the recent results of the Olympia trial emphasize the need for an efficient method of screening for BRCA1 and BRCA2 mutations across all BC phenotypes (A. N. J. Tutt et al. 2021).

In this context, it seems appropriate to systematically screen for HRD induced by BRCA1 and BRCA2 mutations not only for TNBC (18% of all BCs), but also for

luminal B tumors (35% of all BCs). This, however, would represent a real challenge in clinical practice, both economically and logistically. To overcome these challenges, we hypothesized that HRD might be predictable from its phenotypic consequences visible in stained tissue slides acquired in clinical practice. On the other hand, no specific routinely assessed phenotype has been reported to indicate the presence of HRD. For this reason, we set out to predict HRD from whole slide images (WSIs) by deep learning and to identify the underlying morphological patterns.

Deep learning has revolutionized biomedical image analysis and in particular digital pathology. Traditionally, the majority of methods developed in this field were dedicated to computeraided diagnosis, whereby the objective is to partially automatize human interpretation of slides in order to help pathologists in their diagnostic task, e.g., the detection of mitoses(Veta et al. 2015; Ehteshami Bejnordi et al. 2017; Campanella, Hanna, Geneslaw, Miraflor, Silva, et al. 2019). Beyond the automatization of manual inspection, deep learning has also been successfully applied to prediction of patient variables, such as outcome (Mobadersany et al. 2018), and molecular features, such as gene mutations (Jakob Nikolas Kather et al. 2020; Coudray et al. 2018), expression levels (Schmauch et al. 2020), or genetic signatures (Jakob Nikolas Kather et al. 2020; Diao et al. 2021). However, one of the major drawbacks of deep-learning algorithms is their black-box character: because deep learning relies on automatically generated rather than predefined features with a clear biological interpretation, it is difficult to know how a decision was made. This has two major consequences: first, it is difficult to identify potential confounders, i.e., variables that correlate with the output because of the composition of the dataset and that are predicted instead of the intended output variable. Second, even in the absence of statistical artifacts, understanding how the decision was generated in the first place can point to interesting mechanistic hypotheses and to patterns in the image that have so far been overlooked.

One way to overcome the latter problem is to use hand-crafted biologically meaningful features (Diao et al. 2021). This, however, requires an extraordinary effort in terms of annotation. Here, we take a conceptually different approach. Instead of working in a pan-cancer setting on a large number of signatures, we concentrate on one single medically highly relevant signature in one cancer type in a controlled dataset, where we can investigate and correct for potential biases. To understand how the deep-learning decision is generated and which morphological patterns are related to the output variable, we propose a visualization technique that overcomes limitations of current approaches in the presence of complex phenotypes. This paves the way to "machine teaching,'' i.e., a data-driven approach to identify phenotypic patterns related to genomic signatures that is capable of pointing to new mechanistic hypotheses.

In this study, we present an image-based approach to predict HR status from WSIs stained with hematoxylin and eosin (H&E) using deep learning from a large retrospective series of luminal and triple-negative breast carcinomas with a genomically defined HR status from a single cancer center. Furthermore, we identify the morphological patterns associated with HRD. For this we have to tackle two important methodological challenges: the identification and correction of biases in the training

data and the identification of morphological patterns linked to the output variable in the presence of complex pleiotropic phenotypes. Application of these methods to our curated dataset allows us to predict HRD with high accuracy and allows the discovery of decisive, previously unknown morphological patterns related to HRD, leading to new hypotheses on disease-relevant genotype-phenotype relationships.

III.2 Results

III.2.1 A deep-learning Architecture to Predict HRD from Whole Slide Images

We scanned the most representative H&E-stained tissue section of the surgical resection specimens of BC from 714 patients with known HR status. The series was composed of 309 homologous recombination proficient (HRP) tumors and 406 HRD tumors (Table C.4).

Because of their enormous size, analysis of WSIs typically relies on the multiple instance learning (MIL) paradigm (Ilse, Tomczak, and Welling 2018; Maron and Lozano-Perez, n.d.; Amores 2013; Courtiol et al. 2018). MIL techniques only require slide-level annotations and share the overall architecture (Figure III.2), consisting of four main steps: tiling and encoding, tile scoring, aggregation, and decision.

The WSI is divided into tile images (dimensions: 224 $\times$ 224 pixels) arranged in a grid. Background tiles are removed and tissue tiles are encoded into a feature vector. Instead of using representations trained on natural image databases and unlike most studies in this domain, we used the self-supervised technique momentum contrast (MoCo (He et al. 2020); see STAR Methods). This method consists in training a neural network (NN) to recognize images after transformations, such as geometric transformations, noise addition, and color changes. By choosing the type and strength of transformations, we can impose invariance classes, i.e., variations in the input that do not result in significantly different representations. After tile encoding, the feature vector of each tile is then mapped to an attention score by an NN. The slide representation is obtained by the sum of the individual tile representations, weighted by the learned attention scores (Ilse, Tomczak, and Welling 2018). Finally, the slide representation is classified by the decision module (Figure III.2). We optimized hyperparameters by a systematic random search strategy (see STAR Methods). For hyperparameter setting and performance estimation, we used nested 5-fold cross-validation, which allowed us to obtain realistic performance estimations. All reported performance results are averaged over five independent test folds (see STAR Methods).

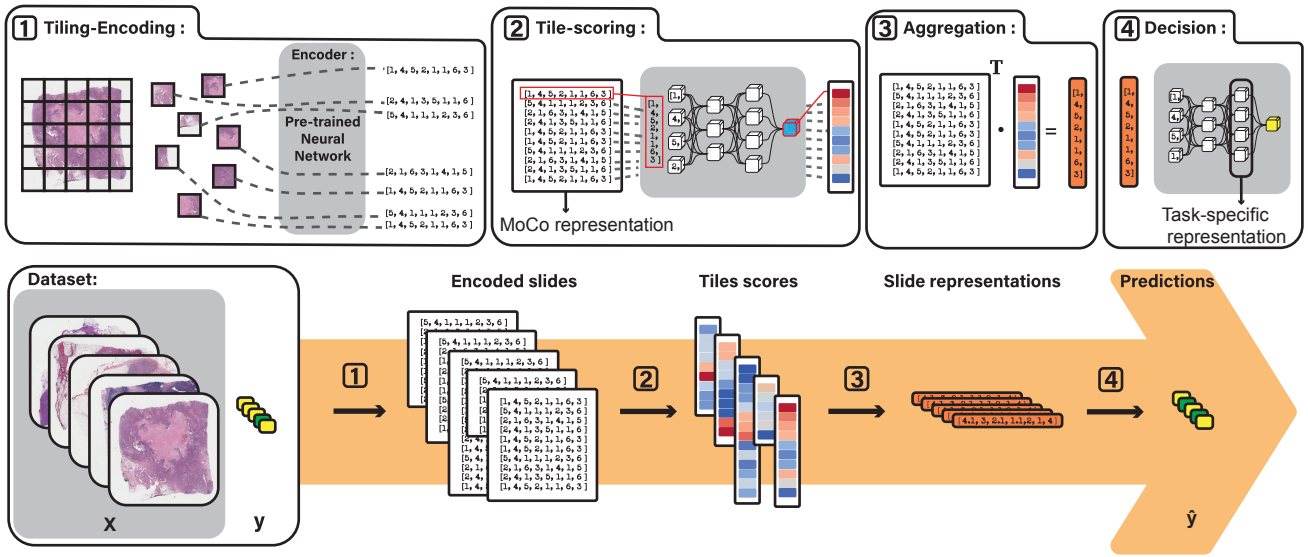


Figure III.2.: **From WSI to prediction.** Four major components are used in this end-to-end pipeline. First, the WSIs (x) are tiled, the tissue parts are automatically selected, and the resulting tiles are embedded into a low-dimensional space (block 1). The embedded tiles are then scored through the attention module (2). An aggregation module outputs the slide-level vector representative (3) that is finally fed to a decision module (4), which outputs the final prediction. When training, the binary cross-entropy loss between the ground truth y and the prediction $\hat{y}$ is computed and back-propagated to update the parameters of the modules. Both the decision module and the attention module are multilayer perceptrons, the encoder is a ResNet18, and the aggregation module consists of a weighted sum of the tiles, the weights being the attention scores.

III.2.2 HRD prediction with correction for potential biases

III.2.2.1. Prediction results obtained without bias correction

We applied this method to predict HRD from the WSI in The Cancer Genome Atlas (TCGA) cohort and obtained results (area under the receiver-operator characteristics curve [AUC] = 0.71, Figure III.3) in line with previous reports (Diao et al. 2021; Jakob Nikolas Kather et al. 2020; Schirris et al. 2021; Valieris et al. 2020). While TCGA is an invaluable resource for pan-cancer studies in genomics and histopathology, it is often seen rather as a starting point whose results need to be corroborated by other cohorts.36 Furthermore, TCGA contains images from many centers around the world with potentially different sample preparation and image-acquisition protocols. While this technical variability might reflect to some degree what could be expected in clinical practice for multiple institutions, we hypothesized that to prove the predictability of HRD independently of potential technical and biological biases, as well as in an in-depth study of morphological patterns related to HRD, it might be advantageous to work on a more homogeneous dataset where we can carefully control for potential technical and biological confounders. We thus turned to our in-house dataset, hereafter referred to as the Curie dataset (see STAR Methods), with data from 714 patients.

We trained an NN to predict HRD on this carefully curated dataset, and we observed a prediction performance largely superior to the best reported to date, trained and tested on TCGA (AUC = 0.88, Figure III.3).

III.2.2.2. Identification and Correction of Biases

As the cohort was generated over 25 years, two experimental variables representing changes in experimental protocols have been identified as potential confounders (c_1 corresponding to the fixation protocol and c_2 to the impregnation protocol, see STAR Methods).

To measure the confounding effects of these variables on the model predictions, we developed a bias score (see STAR Methods). This score is close to zero in the unbiased case and increases with increasing bias. We found that model predictions were indeed biased by these two confounders (Figure III.3**A**).

We then devised a sampling strategy that mitigates biasing during training. Bias mitigation is an increasingly important line of research in machine learning. For instance, it is a well-known problem in training predictive models for functional MRI data, where the age of the patient has been shown to be an important confounder (Varoquaux et al. 2017). While several techniques for bias mitigation exist (Adeli et al. 2020; T. Wang et al. 2019; J. Zhao et al. 2017; Q. Zhao, Adeli, and Pohl 2020), a recent comparison Z. Wang et al. (2020) indicates that strategic sampling is the method of choice if the distribution is not too imbalanced. Strategic sampling aims at ensuring that irrespective of the composition of the training set, each batch presented to the NN is composed of roughly the same number of samples for each

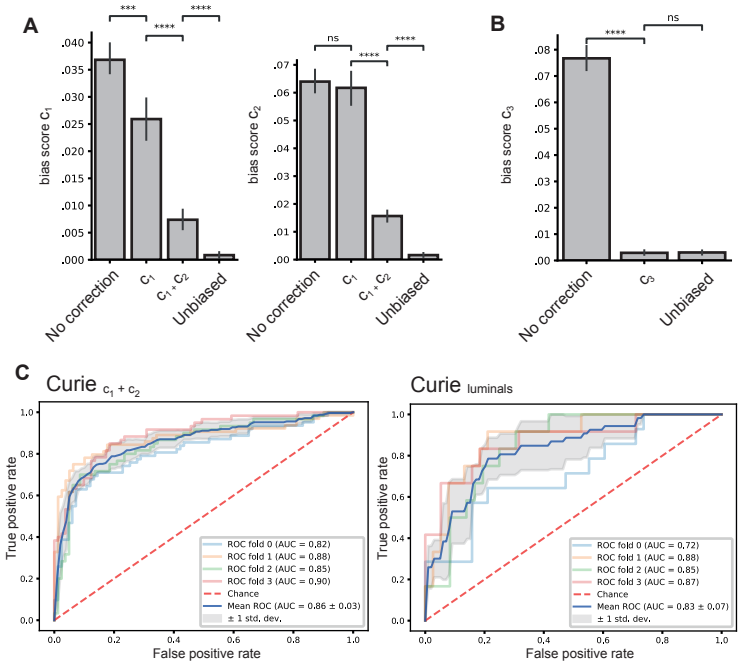


Figure III.3.: Bias corrections and prediction performances. (A and B) Estimation of the bias score of two technical confounders (c_1; c_2) and one biological confounder (c_3) for the Curie dataset (A) and the bias score of the confounder c_3 for TCGA dataset (B) for different correction strategies. A Mann-Whitney-Wilcoxon test, two-sided with Bonferroni correction, is performed for each pair of correction strategies. As detailed in STAR Methods, for each correction strategy a series of 30 unbiased subtest sets are sampled on which the model's bias is evaluated. Error bars indicate standard deviations over the subtest sets. The significance test is performed on this distribution of 30 estimations. The bias score of a model is the average of this distribution. ns, not significant ($p > 0.05$); $p < 0.05$, $p < 0.01$, $p < 1 \times 10^3$, ****$p < 1 \times 10^4$. (C) Receiver-operating characteristic curves. The name of each model indicates the origin of its training set. Indices indicate the correction applied through strategic sampling (Curie$_{c_1}$ has been debiased with respect to c_1). Curie$_{\text{luminals}}$ corresponds to the model trained on a subset containing only luminal tumors.

value combination of output and confounding variable. Correcting for c_1 and c_2 resulted in a 4-fold reduction of the bias score in comparison with the uncorrected model and a slightly lower accuracy (AUC = 0.86, Figure III.3**C**). These results are corroborated using the bias-amplification (BA) measure, a metric widely used in the machine learning fairness literature (Z. Wang et al. 2020; J. Zhao et al. 2017): on the in-house dataset, correcting for c_1 and c_2 lowers the BA from -0.02 to -0.05; on TCGA dataset, the subtype correction lowers the BA from -0.06 to -0.15.

In addition to these technical confounders, we identified the molecular subtype of the tumor to be a potential biological confounder. Successful correction of this biological confounder in TCGA (Figure III.3**B**) led, however, to a dramatic drop in performance (AUC = 0.63). This result suggests that NN trained on the entire BC subset of TCGA for HRD prediction without stratification or bias correction might actually predict to a large extent the molecular subtype, which is also a predictable variable (AUC = 0.89). This shows that the molecular subtype is indeed a biological confounder. In our in-house dataset, we decided to build a subtype-specific NN that specifically predicts HRD for luminal BC instead of applying bias mitigation. The reason for this decision was 3-fold: first, we argued that a dataset focusing on only one molecular subtype was more likely to reveal the underlying patterns exclusively related to HRD; second, HRD prediction in luminal BC is of particular importance for clinical practice, as very few morphological patterns are known to be related to HRD in luminal BC, the most frequent BC phenotype; and third, the relatively low number of TNBCs in our dataset made strategic sampling on three confounding variables challenging. Therefore, we composed a dataset containing only luminal BC and setting both technical confounders, leading us to keep 251 BC WSIs (188 HRD tumors and 63 HRP tumors). We obtained a good, albeit slightly lower performance of this bias-corrected NN (AUC = 0.83; Figure III.3 and Tab. III.1). The trained model carefully freed from both technical and biological biases and validated with respect to cross-dataset performance (Table C.3) was then used for the identification of morpho logical patterns described in the next section. We additionally performed benchmarking experiments to evaluate the influence of the tile encoder network and the MIL algorithm on the classification performances (Tables C.1 and C.2).

III.2.3 Visualization reveals HRD-specific tissue patterns

III.2.3.1. Visualization of attention scores can be misleading

To understand which phenotypic patterns are related to HRD on the WSI, we turned to visualization techniques for NNs. The used MIL framework is equipped with an inherent visualization mechanism: the second module of the algorithm, the tile-scoring module, is in fact an attention module that assigns to each tile an attention score that determines how much a given tile will contribute to the slide representation (and thus to the decision). Attention scores are often used for visualization in the field of digital pathology (Dehaene et al. 2020; Lu et al. 2021; Mobadersany et al. 2018), in the form of either heatmaps to localize the origin of the relevant signals or galleries of tiles of interest (tiles with highest attention scores).

	AUC		B_{acc}	
	Mean	SD	Mean	SD
$TCGA_{raw}$	0.71	0.10	0.59	0.08
$TCGA_{c_3}$	0.63	0.08	0.54	0.02
$Curie_{raw}$	0.88	0.03	0.81	0.02
$Curie_{c_1+2}$	0.86	0.03	0.78	0.04
$Curie_{luminals}$	0.83	0.07	0.72	0.06

Table III.1.: Classification performances Summary of performance metrics. Mean and standard deviation (SD) are computed over the five test sets of the cross-validation. The name of each model indicates the origin of its training set. Indices indicate the correction applied through strategic sampling ($Curie_{c_1}$ has been debiased with respect to c_1). Curieluminals corresponds to the model trained on a subset containing only luminal tumors. We provide an in-depth benchmark of the algorithm in Tables C.1 and C.2 and cross-dataset experiments in Table C.3. AUC, area under the (receiver-operating characteristics) curve; B_{Acc}, balanced accuracy.

However, attention scores do not per se extract the tiles that are related to a certain output variable; they simply reflect that the tile has been taken into consideration in the decision. In particular, in the case of genetic signatures, where we would expect that the output variables can be related to several morphological patterns, analyzing only the attention scores might thus be limited. Figure C.2 illustrates the results obtained by attention-based explanation: while we observe one specific cluster for HRP, most attended tiles seem to be present in both HRD and HRP slides. A possible explanation is that the HRD/HRP decision might be related to the frequency of certain tissue phenotypes rather than to their mere presence.

III.2.3.2. The decision-based visualization technique provides a global explanation of the model

Given these limitations, we propose a visualization protocol that allows us to extract the tiles that are directly associated with a particular slide-level label. As the slide representation is the weighted sum of the tile representations, we applied the decision module, specifically trained to classify slide representations between HRD and HRP, to the individual tile representations. This gives us a score for each tile that can be interpreted as the (tile) probability of being HRD or HRP (see STAR Methods for details). Selecting the tiles with the highest posterior probability for HRD and HRP, respectively, and projecting the tile representations of this selection to a low-dimensional space leads to the emergence of distinct clusters corresponding to different tumor tissue patterns with a clear relation to HRD or HRP and therefore providing a morphological map of HRD (Figure III.5).

Two expert pathologists labeled these clusters. The HRD signal relied on several clusters: HRD tumors present a high tu-mor cell density, with a high nucleus/cytoplasm ratio and conspicuous nucleoli. They also show regions of hemorrhagic suffusion

associated with necrotic tissue. In the stroma, the HRD signal revealed the presence of striking laminated fibrosis and, as expected, high content of tumor-infiltrating lymphocytes (TILs). Lastly, one large cluster contained a continuum of several phenotypes, namely adipose tissue intermingled with scattered and clear tumor cells, histiocytes, and plasma cells. In contrast, the HRP signal was mostly carried by one cluster characterized by low tumor cell density, the cells being moderately atypical, and tumor cell nests separated from the stroma by clear spaces. Notably, it included a few invasive lobular carcinomas (all of the tiles per cluster are available in Figures C.6 to C.8).

III.2.3.3. Validation of the morphological patterns

Some of these patterns, namely high-grade and TIL, had been previously associated with phenotypic hallmarks of HRD in TNBCs (Rakha et al. 2009). To validate these results in the luminal BC cohort, TIL density and nuclear grade were evaluated for each luminal tumor of the in-house dataset by an expert pathologist. As predicted by our algorithm, TILs and nuclear grade were positively associated with the HR status of the tumor in the luminal subset (mean TIL HRD, 29; mean TIL HRP, 17; t-test p value, 0.017; mean nuclear grade HRD, 2.7; mean nuclear grade HRP, 2.3; c_2 p value, 1.2 $\times 10^{-6}$). Moreover, a logistic regression trained on the components of the grade (architecture grade, atypia grade, and mitosis grade) and on the TIL count estimation has an average AUC of 0.76 (5-fold cross-validation).

To further validate the association of these morphological patterns with HRD, we turned to the independent TCGA cohort. Despite the modest prediction accuracy after bias correction, we found that a NN trained on TCGA-extracted morphological patterns strikingly similar to those obtained from our in-house dataset (Figure C.3), with the exception of cluster 4 (Figure III.5). Regarding HRP, we were able to validate all patterns related to HRP, but artifact classes were also identified, which is unsurprising given the limited slide quality and heterogeneity of TCGA dataset and may explain the poor classification performance.

To test the subtype specificity of the morphological patterns, we trained a network on the small TNBC subset of TCGA (129 slides). While classification performances remain poor (AUC = 0.62), because of the small size and large heterogeneity of the dataset, the extracted patterns explaining the predictions are in line with the literature (Figure C.4), suggesting that HRD for TNBC is characterized by high content of TILs and necrosis, while the retraction figures are still an explanation of the HRP signal. This result further confirms the specificity of our extracted morphological patterns and suggests that there are indeed HRD-related morphological patterns specific to the luminal subtype.

Our NN works with different internal representations. While the tile representations provided by MoCo permit the emergence of phenotypic similarity clusters (Figure III.5), internal representations closer to the decision module encode information relevant for HRD. The representation in the penultimate layer can therefore

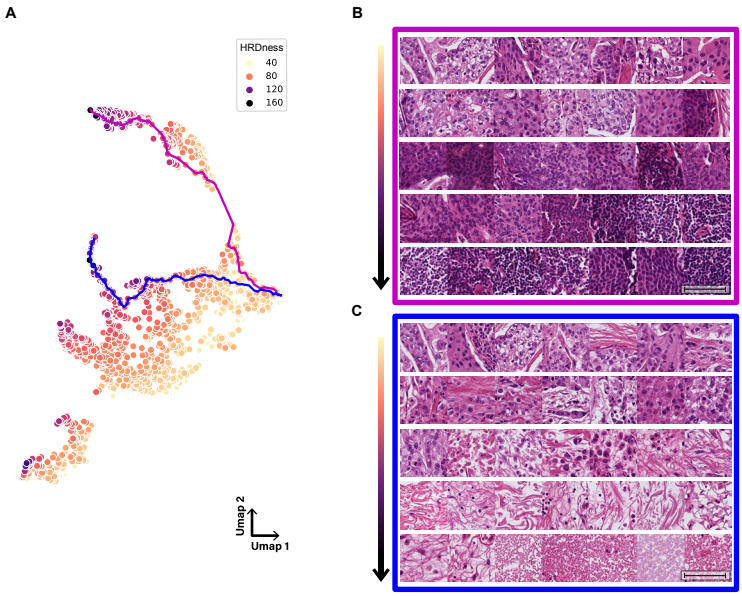


Figure III.4.: llustration of two phenotypic HRDness trajectories (A) UMAP projection of the HR status-specific representation of the meaningful tiles relative to the HRD. HRD-ness is the score given to each tile by the HRD output neuron. Two tile trajectories have been extracted (blue and magenta) starting from the same low HRD-ness region, each leading to a different high HRD-ness region. (B and C) Tiles sampled along each of the trajectories. These are ordered from low HRD-ness to high HRD-ness and read from left to right and from top to bottom. Scale bars, 100 μm. (B) Magenta trajectory, toward densely cellular tumors or inflammatory cells. (C) Blue trajectory, toward fibroinflammatory tumor changes and hemorrhagic suffusions.

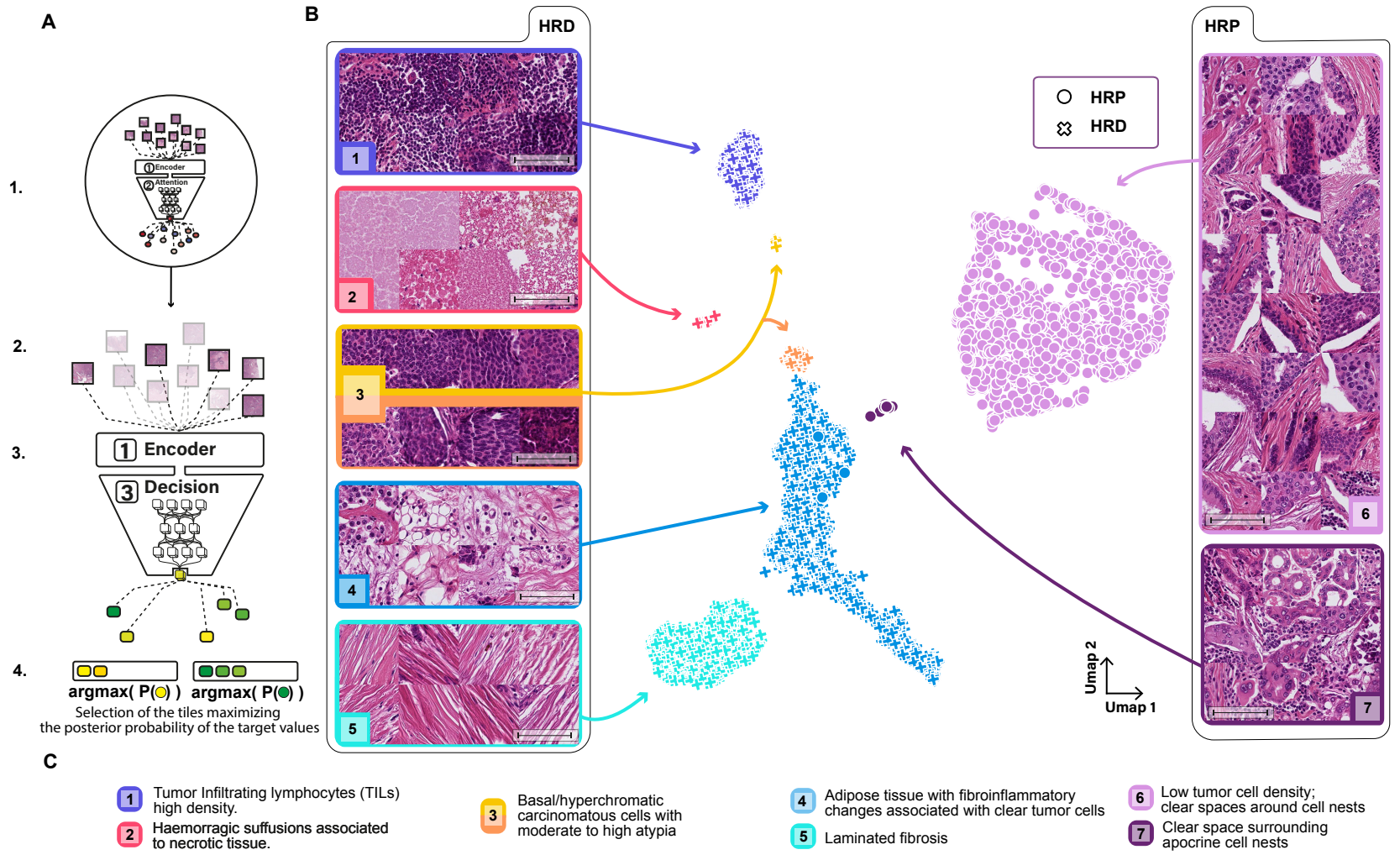


Figure III.5.: Decision-based visualization. (A) Mechanism of the decision-based visualization. 1: each tile in the whole dataset is scored by the attention module. 2: per slides, the 300 best scoring tiles are selected as candidate tiles. 3: the selected tiles are presented to the decision module, and the logit of the probability of each of these tiles being HRD or HRP (yellow or green) is kept. 4: finally, the K tiles with maximal probability for either HRD/HRP are selected. (B) Morphological map of the HR status in the luminal BC cohort. Each dot is the uniform manifold approximation and projection (UMAP) of a tile extracted by the decision-based visualization method. Crosses (circles) are tiles with high HRD (HRP) logit. Each cluster has been linked to a morphological phenotype by two expert pathologists. We identified six different morphological phenotypes associated with the HRD and two associated with the HRP. The exhibited tiles have been randomly sampled among each cluster. 228 slides contributed to the HRP clusters and 232 to the HRD cluster. In total, 249 among 251 slides contributed to the whole figure. The same protocol has been applied to the public datasets TCGA breast invasive carcinoma (BRCA), TCGA BRCA-TNBC, and TCGA ovarian cancer (see Figures C.3 to C.5, respectively). Scale bars, 100 μm. (C) Pathological interpretation of the clusters presented in (B).

be interpreted as encoding "HRD-ness'' of the tiles. Figure III.4 illustrates a low-dimensional representation of this HRD-ness for the same tiles as those present in Figure III.5, where point color represents the HRD score (tile probability to be classified as HRD). From there, we extracted two tile trajectories going from low HRD-ness to high HRD-ness. The magenta trajectory illustrates the successive visual changes corresponding to an increase in tumor cells or inflammatory cell density (from low-density tiles to high-density tiles with large nuclei, nuclear atypia, and infiltrative lymphocytes). The blue trajectory shows, conversely, a decrease in tumor cell density replaced successively by an inflammatory reaction and apoptotic cells, loose fibrosis, and hemorrhagic suffusion associated with necrosis. These different trajectories illustrate the manifestations of HRD and show the pleiotropic character of the induced phenotypes. Moreover, the highlighted gradation of these phenotypes opens the path to a possible reading grid of WSIs for pathologists.

III.3 Discussion

In this study, we set out to predict the HR status in BC from H&E-stained WSIs and to analyze the phenotypic patterns related to HRD. The prediction of HRD is an important challenge in clinical practice. The use of PARPi for BC patients was initiated for metastatic TNBC patients with germline mutations of BRCA1 or BRCA2. However, BRCA2, as well as PALB2 and a minority of BRCA1 cancer patients, develop luminal tumors. The necessity of predicting HRD is therefore not limited to TNBC but extends also to luminal BC. On the other hand, luminal BCs represent a far more frequent group than TNBC. For this reason, systematic screening of HR gene alterations for luminal cancers will be problematic and, in many countries, even unfeasible due to both economic and logistic issues. Therefore, preselection of patients with a high probability of being HR deficient by analysis of WSIs is a cost-efficient strategy that has so far only been hampered by the lack of knowledge about HRD-specific morphological patterns in luminals. Indeed, only high grades and to a lower extent pushing margins have previously been reported to be associated with HRD. In this context, the identification of HRD from WSIs by deep learning and the identification of related morphological patterns could both facilitate the preselection of BCs for molecular determination of HRD, which is particularly important for luminal cancers.

TCGA provides a precious dataset from which to train models for the prediction of genetic signatures from H&E data (Diao et al. 2021; Jakob Nikolas Kather et al. 2020). While we obtained promising results for the prediction of HRD on TCGA dataset in line with previous reports, we found that this result was partly due to the fact that the molecular subtype acts as a biological confounder. This was particularly problematic, as we wanted to investigate the morphological signature of HRD. Of note, the existence of biological and technical confounders is presumably not limited to HRD prediction but may concern many genetic signatures. The use of carefully curated datasets where technical and biological confounders can be controlled for is,

thus, an important step in investigating the predictability of genetic signatures as well as the identification of their morphological counterparts.

In most cases, such in-house datasets also contain technical and biological biases due to the long period during which the dataset is acquired. This motivated us to propose a method to mitigate bias in computational pathology workflows, based on strategic sampling. Such strategies are already used in other fields of medical imaging but have so far, to the best of our knowledge, not been used in computational pathology. We have shown that this approach can successfully mitigate or even eliminate bias. In a larger perspective, it is essential to investigate potential confounding variables in the dataset when applying deep-learning based methods for the prediction of slide-level variables. Biased datasets can lead to false expectations and misinterpretation. For this reason, we expect proper treatment of such variables to become a standard in the field.

While bias correction on TCGA led to a drop in AUC to 0.63, we found that HRD was predictable in our in-house dataset of 251 luminal BC patients with an AUC of 0.83. While homogeneous datasets do not reflect the variability between centers and thus limit direct applicability of the trained networks, they allow for controlled feasibility studies, which now need to be complemented by multicenter studies. In addition, we will validate this algorithm in a prospective neoadjuvant clinical trial for which patients' HRD status will be assessed with the MyChoice CDx test (Myriad Genetics).

Homogeneous datasets are well suited for the identification of underlying phenotypic patterns, even in cases where no or few such patterns are known a priori, such as in the case for HRD. To identify a phenotypic signature related to an output variable (here HRD), either we can use biologically meaningful encodings, also known as human interpretable features (HIF), and infer the most relevant features by analyzing the weights in the predictive model (Diao et al. 2021), or we can turn to network introspection. The HIF approach relies on detailed and exhaustive annotations of a large number of WSIs, for instance (Diao et al. 2021), leverage annotations provided by hundreds of pathologists consisting of hundreds of thousands of manual cell and tissue classifications. Here, we provide a new network introspection scheme relying on the powerful MoCo encodings, trained without supervision directly on histopathology data, and a decision-based tile selection that allows us to automatically cluster tiles and to relate these clusters to the output variable. Interestingly, while our approach confirms the recently published finding that necrosis is a hallmark of HRD (Diao et al. 2021) and identifies morphological features common to HRD in TNBC and luminal BC, such as necrosis, high density in TILs, and high nuclear anisokaryosis (Rakha et al. 2009), it also points to more specific patterns that have so far been overlooked. For instance, we found tiles enriched in carcinomatous cells with clear cytoplasm, suggesting activation of specific metabolic processes in these cells. Moreover, we found intratumoral laminated fibrosis as an HRD-related pattern. Also, we were able to validate most of these patterns on TCGA. This leads to the hypothesis that cancer-associated fibroblasts (CAFs) within the stroma of HRD luminal tumors may play a role in the viability and fate of tumor cells. Furthermore, the presence of adipose tissue within the

tumor suggests first, a different tumor cell density and second, a specific balance between CAFs and adipocytes in the context of a luminal HRD tumor. The molecular mechanisms achieving these patterns remain to be determined by *in vitro* models.

Similar to what we have shown here with respect to HRD, the visualization framework we have developed is versatile and can in principle be applied in the context of other genetic signatures. In particular, our visualization scheme overcomes the limitations of the thus far predominating technique of visualizing attention scores alone. Indeed, attention scores were used previously to identify tumor regions under weak supervision. However, if the output variable depends on the quantity of several morphological patterns in contrast to the presence/absence of a single tissue phenotype, attention scores might not provide a suitable tile selection and visualization tool and might thus be ill suited to investigate the underlying morphological phenotypes. Because the algorithm is fully automated, using the MIL algorithm and the proposed visualization method can constitute a useful tool for the discovery of morphological features related to the predicted genetic signatures. This has the potential to generate new biological hypotheses about the phenotypic impact of these genetic disorders. To maximize the benefit for the scientific community, we release the code to train MIL models on WSIs and create morphological maps as well as tile trajectories publicly and free of charge, and provide detailed documentation.

Altogether, this study provides new and versatile tools for the prediction and phenotypic dissection of genetic signatures from histopathology data. Application to luminal BCs allowed us to show that HRD is predictable from WSIs and to shed light on the phenotypic consequences of HRD. These tools have the potential to impact BC patient care.

III.3.1 Limitations of the study

Our study involves a homogeneous, carefully controlled cohort that allowed us to train a network for HRD prediction with high accuracy and correction for technical and biological confounders. We could thus convincingly show that HRD is predictable from WSIs. However, the study was not designed for the demonstration of clinical applicability. To use HRD prediction in clinical practice, we will need to validate the workflow on larger, multicenter cohorts.

Furthermore, we have identified morphological patterns related to HRD. While our validation results obtained from TCGA suggest that the method works robustly and that these patterns are truly linked to HRD, we will need to validate these findings in a larger independent cohort. In addition, the development and demonstration of a mechanistic model explaining these morphological phenotypes will be a challenging and exciting perspective. Finally, it will be important to further explore the variability of the morphological patterns in different cancer types.

At a methodological level, we have proposed strategic sampling as a method to mitigate biases in digital pathology datasets. While we were able to show that this method is highly effective, it must be noted that it is limited by the number of

variables we can correct for as well as by the class imbalance it can handle. In some cases, stratification might therefore be preferable. Furthermore, we have proposed a method to improve the interpretability of the MIL approach for HRD prediction. However, it is still difficult to precisely understand how the identified tiles impact the prediction. For instance, the method does not give information on a potential hierarchical relation between the morphological clusters. Also, the current strategy does not allow us to assess whether the tiles of a given cluster influence the decision by their proportion on the slide, their mere presence, or the simultaneous presence of tiles from other clusters. A promising methodological perspective is therefore the improvement of these visualization techniques.

Acknowledgments The authors thank Helene Guenon, Saida Sahiri, Martial Caly, and Laure Annette for their help in retrieving the H&E slides and their technical expertise. The authors thank AstraZeneca for the funding of technical time essential for the preparation of the material for the pathology case series. G.B. was supported by a Fondation Curie grant. T.L. was supported by a Q-Life PhD fellowship (Q-life ANR-17-CONV-0005). Furthermore, this work was supported by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir'' program, reference ANR19-P3IA-0001 (PRAIRIE 3IA Institute).

Author contributions A.V.-S. and G.B. initiated the project. A.V.-S., G.B., F.-C.B., and D.S.-L. generated the patient cohort. G.B. reviewed all the slides. T.P. and M.-H.S. performed the genomic analyses. T.L., E.D., and T.W. designed the AI and statistical methods. T.L. and P.N. developed the software. T.L. performed the analysis under the supervision of T.W. and E.D. A.V.S. and G.B. interpreted the morphological patterns. A.V.-S., G.B., T.L., E.D., and T.W. discussed methods, results, and design choices. T.L. prepared the figures. T.L., T.W., and A.V.-S. wrote the manuscript and its revisions.

III.4 STAR Methods

III.4.0.1. Method Details

In-house dataset (Institut Curie) We retrospectively retrieved a series of 715 patients with HE slides of surgical resections specimens of untreated breast cancer and a genomically known HR status (Table C.4). The series is composed of 309 Homologous Recombination Proficient tumors (HRP) and 406 Homologous Recombination Deficient tumors (HRD). The HRD status was either identified by the presence of a germline BRCA1/ 2 (gBRCA1/2) mutation or assessed by LST genomic signature according to Popova et al. (2012) for the sporadic triple-negative and luminal cancers.

All patients have been treated and followed at the Institut Curie between 1995 and 2020. The patient agreed for the use of tumor samples from their surgical resection specimens for research according to the law. Ethical approval from the Institutional Review Board (Institut Curie breast cancer study group N-DATA190031) was obtained for the use of all specimens. Clinical data have been retrieved from the Institut Curie electronic medical records and saved using Research electronic data capture (REDCap) tools hosted at the Institut Curie.

Public dataset (TCGA) This public dataset is composed of 815 WSI of breast cancer fixed in formalin (FFPE) and stained in H&E. They are available at https:// portal.gdc.cancer.gov/. Low-resolution WSI, WSI containing artifacts such as large pen marks, tissue-folds and blurred WSI were removed. The final dataset encompasses 673 WSIs. The HR status of the corresponding tumors was obtained using the LST genomic signature

Architecture and optimization parameters Hyperparameters have been set thanks to a random search evaluated through 5-fold nested cross-validation. The benchmark task is the prediction of the molecular class of the TCGA WSIs. Both the decision module and the tile-scoring module are multi-layer perceptrons with batch normalization (Ioffe and Szegedy 2015) after each hidden layer. The decision module has 3 hidden layers of 512 neurons, the tile-scoring module has 1 hidden layer of 256 neurons.

Dropout has been fixed at 0.4, the optimizer is ADAM (Kingma and Ba 2014) with a learning rate of 3e-3. A batch consists of 16 samples of WSI. A sample of WSI corresponds to a uniform sampling of 300 of its composing tiles. In fact, we observed that this uniform subsampling of the WSIs regularized training as well as diminishes its computational workload. Finally, training is performed during 200 epochs. Training and performance evaluation are done in a 5-fold nested cross-validation framework.

Each dataset is split into 5 independent folds. For each of these folds, a validation set is randomly sampled in the complementary 4/5th. A model is trained on the remaining dataset (= 4/5 * 4/5 th of the total dataset). This process is repeated 10 times for each test fold, then the 3 best models are selected according to their validation performances, ensembled and finally tested on their test set. This process of model selection and ensembling drives itself a net improvement of the performances (see Figure C.1).

Each test and validation set preserves the stratification of the whole dataset with respect to the target variable as well as the confounding variables in case we correct for them. The final performance estimation of the model is the performance averaged over the 5 test performances. During inference time, all the tiles of each WSI are processed.

Strategic sampling Strategic sampling is used both for balancing the training dataset with respect to the output variable ($T(X) \in \{t_1, t_2, \ldots, t_m\}$) and to correct for biases ($B(X) \in \{b_1, b_2, \ldots, b_n\}$).

If X is a given WSI sampled from the dataset, then $T(X)$ and $B(X)$ are respectively the target value and the bias value of X. We note $|t_k|$ the total number of slides in the dataset labeled with t_k, and $|b_i|$ the total number of slides for which the bias variable takes the value i. $|t_k \& b_i|$ is the total number of slides with label value t_k and bias value b_i.

For achieving both balancing with respect to the output and correcting for biases, we sample the WSIs X in each batch in a distribution P under which $P\left(T(X) = t_k\right) = P\left(T(X) = t_{k'}\right)$ for all $k \neq k'$. And. $P\left(\{T(X) = t_k\} \cap \{B(X) = b_i\}\right) = P\left(\{T(X) = t_{k'}\} \cap \{B(X) = b_i\}\right)$ for all i and $k \neq k'$. That is, we sample the slide X depending on its target and bias value with probability: $P\left(X \,|\{T(X) = t_k\} \cap \{B(X) = b_i\}|\right) \propto \frac{|b_i|}{|t_k \& b_i|}$ for each $i \leq n, k \leq n$ Strategic sampling is performed on the fly when building the batches. When correcting for several confounders simultaneously, $B \in \{b_1, b_2, \ldots, b_{n_1}\}$ and $C \in \{c_1, c_2, \ldots, c_{n_2}\}$, we simply correct for a new confounder variable that takes values in all combinations of b_i and c_j.

Bias score We introduce the following notation: for a WSI X_D, sampled in a dataset D under the distribution $P_D, T\left(X_D\right)$ is the label of X_D and $B\left(X_D\right)$ is the candidate confounder value of X_D (for instance bouin).

We want to measure the bias of a predictive algorithm m that outputs, for each X_D, a prediction $m\left(X_D\right)$. We moreover define the accuracy $\mathrm{Acc_m}$ of m as: $Acc_m = \boldsymbol{E}\left(1_{\{\boldsymbol{m}(\boldsymbol{X_D}) = \boldsymbol{T}(\boldsymbol{X_D})\}}\right)$

The mutual information $\mathrm{MI}\left(B\left(X_D\right), m\left(X_D\right)\right)$ between $B\left(X_D\right)$ and $m\left(X_D\right)$ measures the mutual dependence between B and m and highlights the bias of a model. The idea of the bias score is to compute how far away the predictions of a model are from a perfectly unbiased case. To simulate this perfectly unbiased case, we subsample (with strategic sampling) a dataset D_i such that $\mathrm{Mi}\left(B\left(X_{D_i}\right), T\left(X_{D_i}\right)\right) = 0$, i.e. such that the target variable and the confounder variable are statistically independent in this dataset. If m is unbiased, then we should observe that $\mathrm{MI}\left(B\left(X_{D_i}\right), m\left(X_{D_i}\right)\right) = 0$ too. In contrast, the more m is biased, the more $\mathrm{MI}\left(B\left(X_{D_i}\right), m\left(X_{D_i}\right)\right) \geq 0$ will be far away from 0 . In order to obtain a more accurate estimation of the bias score, we iterate this measure over several unbiased datasets $\{D_i\}_{i \leq 30}$. The bias score $BS(B, m)$ is then the average of $\mathrm{MI}\left(B\left(X_{D_i}\right), m\left(X_{D_i}\right)\right)$ over i.

Because by construction, $BS(B, m)$ is non-negative, we build an unbiased reference m^* such that $P\left(\mathrm{m}^*(\mathrm{X}) = T(X)\right) = Acc_m$, and compute its bias-score as a reference value.

Learning MocCo representations For learning MoCo-v2 (Xinlei Chen et al. 2020) representation we used the MoCo repository available at https://github.com/facebookresearch/moco. We randomly used the following transformations: Gaussian blur, crop and resize, color jitter, grayscale, horizontal and vertical symmetries, and a color augmentation in the Hematoxylin and Eosin specific space. (Ruifrok, n.d.) The training dataset is composed of 5.3e6 images of size 224x 224 pixels, or half the Curie dataset at magnification 20x (0.46 μm.px) We used a Resnet18 and trained it from scratch for 60 epochs on 4 GPU Nvidia Tesla V100 SXM2 32 Go. We used the SGD optimizer with a momentum of 0.9, a weight decay of 1e-4, a learning rate of 3e-3 and a batch size of 512. We used a cosine scheduler with a warm restart on the learning rate.

III.4.0.2. Visualization methods

The model used to extract the visualizations has been trained on the luminal subset of the Curie dataset (251 WSI). To benefit from the biggest dataset possible, the model has been trained on the whole dataset, without using early stopping nor testing, during 200 epochs.

To generate the attention-based visualization, the highest ranked tile with respect to the attention score is extracted, for each WSI. The selected tiles are then labeled according to the label of their WSI of origin. Concerning the decision-based visualization, for each WSI the 300 highest ranked tiles with respect to the attention score are selected. Among this pool of tiles, the 2000 highest ranking tiles with respect to the logit of the posterior probability for HRD and HRP are selected. In order to promote diversity in the extracted images, no more than 20 tiles per slide can be selected.

III.4.0.3. Quantification and statistical analysis

Technical biases in the Curie dataset Both technical confounders are related to technical protocols that were modified over time with an unbalanced representation between the HRD and HRP cohorts: $-c_2$ corresponds to a change of fixative agent. $c_2 \in \{$ Bouin, AFA $\}$ - c_1 corresponds to a change of impregnation technique. $c_1 \in \{$ Ethanol, Ethylene $\}$.

We performed the exact Fisher test to test for a correlation between:

1. HRD - c_1(impregnation): test-statistic $12; \mathrm{p}$ value $3.9\mathrm{e}-30$
2. HRD - c_2 (fixation): test-statistic $31; p$ value $2.8\mathrm{e}-78$

Showing the statistical relationship between both confounders and our target variable, the HR status. Fisher test was performed with the scipy package. (Virtanen et al. 2020)

Manual validation of the morphological patterns The t-test and the Xi2 test performed respectively to test the difference of TILs count and nuclear grade between HRD and HRP tumors were done with the scipy package. The logistic regression used to predict HRD from the grade and TILs count was implemented with scikit-learn (Pedregosa et al., n.d.) package, with a parameter C = 10, all other parameters set to their default values.

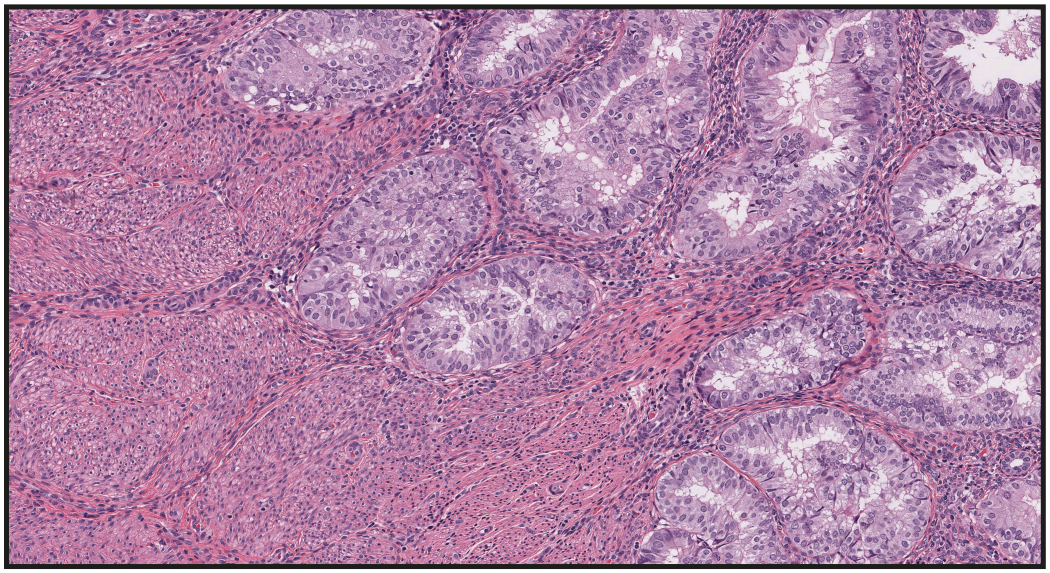
Bias metric significance test The Mann-Whitney-Wilcoxon test two-sided with Bonferroni correction appearing in the legend of Figure III.3 has been performed using the scipy package. The two compared distribution correspond to the mutual information measure iterated over the 30 sub-datasets, as described in the bias score method subsection.

III.4.0.4. Key Resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
TCGA	GDC Portal	N/A
In-house Dataset	Curie Hospital, Paris	N/A
Model predictions	mendeley-dataset	zenodo DOI
Software and algorithms		
wsi-mil	wsi-mil	zenodo versioning
scikit-learn	sklearn	RRID: SCR-019053
openslide-python	OpenSlide	N/A
MoCo	MoCo	N/A
SciPy	SciPy	RRID: SCR-008058

IV

Mixing local and weak supervision



Contents

Preface

This work maturated from our involvement in the inaugural DataChallenge of the Société Français de Pathologie (SFP), titled “Tissuenet”. These time-constrained data challenges present a unique platform to devise new algorithms and evaluate them against competitors in real-time. Crucially, these challenges address pertinent clinical problems, selected by pathologists, on datasets that replicate actual use-case scenarios (sourced from various centers across France).

This work was done in collaboration with another PhD student in the lab, Mélanie Lubrano. The challenge took place before she joined the CBIO, and so we participated independently. It allowed me to apply the self-supervised learning framework to histopathological images for the first time and to experiment with cost-sensitive losses. After the challenge, we decided to work together on the data of the challenge in order to to address a specific aspect we hadn’t explored during the challenge: leveraging the provided regional annotations to enhance slide-level classification tasks

Disclaimer: as the second author, my contribution to this work was less predominant compared to other projects in this thesis. While I was involved in designing the training framework, planning experiments, and contributing to the writing, I did not execute the experiments.

Contribution

Publications - communications

- M. Lubrano, **T. Lazard**, et al. Automatic Grading of Cervical Biopsies by Combining Full and Self-supervision. 13807, Springer Nature Switzerland, pp.408-423, 2023, Lecture Notes in Computer Science, 10.1007/978-3-031-25082-8_27.

Summary:

In computational pathology, the application of Deep Learning to the analysis of Whole Slide Images (WSI) has provided results of unprecedented quality. Due to their enormous size, WSIs have to be split into small images (tiles) which are first encoded and whose representations are then agglomerated in order to solve prediction tasks, such as prognosis or treatment response. The choice of the encoding strategy plays a key role in such algorithms. Current approaches include the use of encodings trained on unrelated data sources, full supervision or self-supervision. In particular, self-supervised learning (SSL) offers a great opportunity to exploit all the unlabelled data available. However, it often requires large computational resources and can be challenging to train. On the other end of the spectrum, fully-supervised methods make use of valuable prior knowledge about the data but involve a costly amount of expert time. This paper proposes a framework to reconcile SSL and full supervision and measures the trade-off between long SSL training and annotation effort, showing that a combination of both has the potential to substantially increase performance. On a recently organized challenge on grading Cervical Biopsies, we show that our mixed supervision scheme reaches high performance (weighted accuracy (WA): 0.945), outperforming both SSL (WA: 0.927) and transfer learning from ImageNet (WA: 0.877). We further provide insights and guidelines to train a clinically impactful classifier with a limited expert and/or computational workload budget. We expect that the combination of full and self-supervision is an interesting strategy for many tasks in computational pathology and will be widely adopted by the field.

Résumé:

En pathologie computationnelle, l'utilisation de l'apprentissage profond pour analyser les images de lames entières (WSI) a donné d'excellents résultats. Les WSI, en raison de leur grande taille, sont divisées en petites tuiles. Ces tuiles sont d'abord encodées, puis leurs représentations sont agglomérées par des modèles d'apprentissage par instance multiple (MIL) pour résoudre des tâches de prédiction à l'échelle de la lame ou du patient, comme le pronostic ou la réponse au traitement. La méthode d'encodage des tuiles est cruciale. Les approches actuelles utilisent des représentations pré-entraînées sur des images naturelles, des images histopathologiques étiquetées ou non étiquetées via auto-supervision. L'apprentissage auto-supervisé (SSL) permet d'utiliser des données non étiquetées mais est gourmand en ressources. Les méthodes entièrement supervisées sont moins coûteuses en calcul mais nécessitent un étiquetage laborieux. Cet article propose une méthode combinant SSL et supervision standard, et évalue le compromis entre le temps d'entraînement SSL et l'effort d'annotation. Nous démontrons que cette combinaison peut améliorer la performance de classification des WSI. Dans un data-challenge sur le classement des biopsies du col de l'utérus, notre méthode mixte atteint une précision pondérée (WA) de 0,945, surpassant une méthode basée uniquement sur SSL (WA : 0,927) ou un pré-entraînement sur ImageNet (WA : 0,877). Nous offrons en outre des conseils pour entraîner un classificateur efficace avec un budget limité en termes d'annotations ou de ressources de calcul. Nous pensons que la combinaison de supervision complète et d'auto-supervision pourrait être bénéfique pour diverses tâches en pathologie computationnelle.

This chapter has been done in collaboration with M. Lubrano, G. Balezo, Y. Bellahsen-Harrar, C. Badoual, S. Berlemont and T. Walter. It was published in the proceedings of the ECCV22 workshop on Medical Computer Vision.

IV.1 Introduction

Recent advances in slide digitization have led to increased interest in Artificial Intelligence (AI) applications for histopathology. The development of AI models could help reduce pathologists' workloads, limit subjectivity and help contributing to medical discoveries. Deep learning models can now match pathologist performance for many tasks: diagnostic, detection of mitoses (Veta et al. 2015), prediction of gene mutations (Coudray et al. 2018; Jakob Nikolas Kather et al. 2020) or genetic signatures (Diao et al. 2021; Jakob Nikolas Kather et al. 2020; Lazard et al. 2022), cancer subtyping (Coudray et al. 2018) and more.

One of the applications, automated diagnosis from Whole Slide Images (WSIs), induces two main challenges: first, WSIs are very high-resolution and, because of memory constraint, cannot be fed directly into traditional neural networks. Second, expert annotations are laborious to attain, costly and prone to subjectivity. The most popular methods today rely on Multiple Instance Learning (MIL), which frames the problem as a bag classification task. WSIs are split into small workable images (tiles), which are processed separately. Features from each of the individual tiles are extracted and then aggregated to classify the WSI.

The extraction of these tiles' specific representation is crucial to the downstream WSI classification task. One common approach consists of initializing the feature extractor with pre-trained weights on ImageNet, a natural image dataset. This technique allows one to extract generic features that are powerful, but that do not lie within the histopathological domain. Different strategies have been developed to extract these tile encodings taking advantage of the available data and their respective level of supervision.

A first strategy aims to learn tile features with full supervision (Ehteshami Bejnordi et al. 2017). To create a supervised dataset, one or several experts manually review tiles and sort them into meaningful classes (preferably related to the downstream task of classifying the WSIs). Even though experts' annotations can bring powerful prior knowledge to the model, this technique often requires large quantities of annotations.

A second strategy consists of learning tile representations through self-supervision. It leverages the unannotated data by training a convolutional neural network on a pretext task. It has proven its efficacy (Saillard et al. 2021) and even its superiority to the fully supervised scheme (Dehaene et al. 2020). However, this approach has a non-negligible computational cost, as training necessitates around 1000 hours of computation on a standard GPU (Dehaene et al. 2020). Moreover, it is not guaranteed that the obtained encodings are most relevant for the prediction task we are trying to solve.

Techniques from both sides of the supervision spectrum have proven to bring important benefits for relevant feature extraction. Combining them could allow us to benefit from the best of both worlds. In this work, in addition to proposing a joint-optimization process mixing self, full and weak supervision (Figure IV.1), we measure the trade-off in performance between the number of annotations and the computational cost of training a self-supervised model. We thus provide guidelines to train a clinically impactful classifier with a limited budget in expert and/or computational workload.

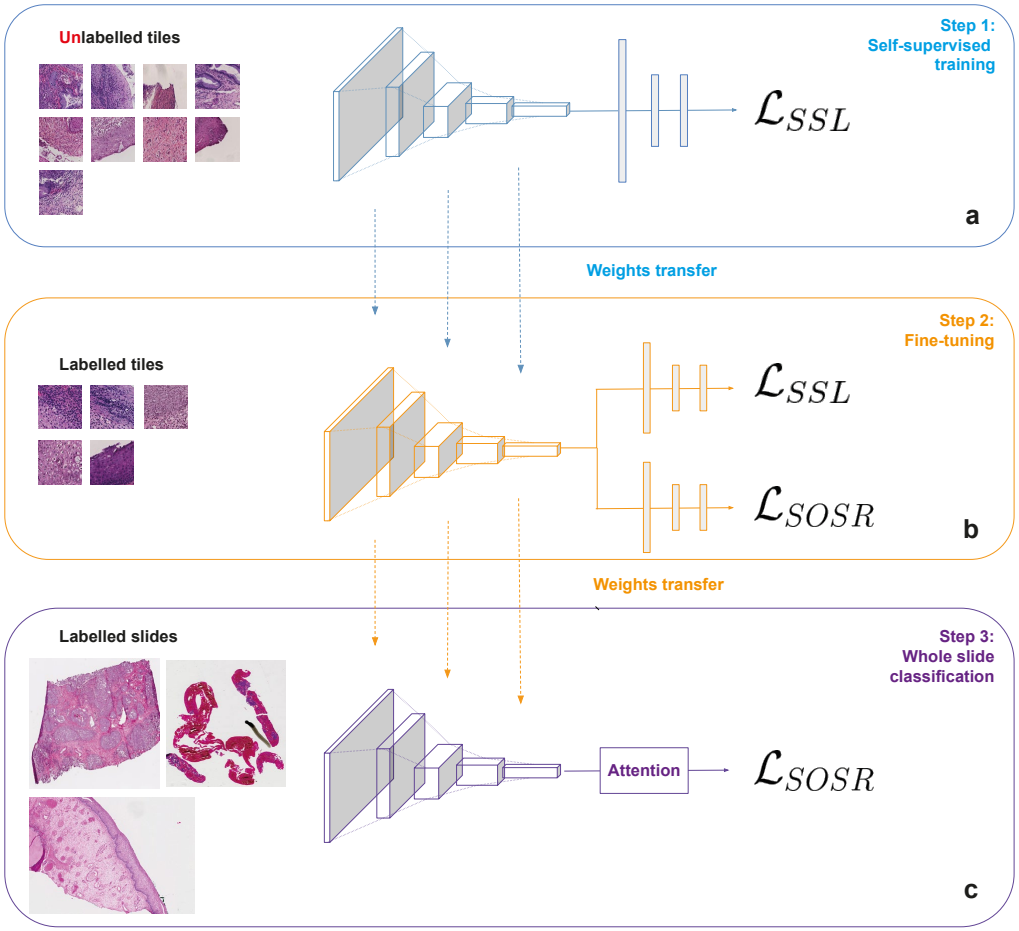


Figure IV.1.: Mixed Supervision Process: a)A self-supervised model (SimCLR) is trained on unlabelled tiles extracted from the slides. Feature extractor and contrastive layer weights are transferred to the joint-optimization architecture b) Joint-optimization model is trained on the labeled tiles of the dataset. The feature extractor weights are transferred to the WS classification model. c) WS classification model is trained on the 1015 whole slide images.

IV.2 Related Work

Mixed Supervision Medical data is often limited. For this reason, one might want to take advantage of all the available data even if annotations might not be homogeneous and even though they might be difficult to exploit because multiple levels of supervision are available. For instance, whole slide images are often associated with one global label (weak supervision), they can contain millions of unlabelled tiles (no supervision), but, as a pathologist reviews the slides and performs a diagnostic, it is almost effortless for them to mark the region of interest that signs the corresponding diagnostic (strong supervision). AI applications have usually been dichotomized between supervised and unsupervised methods, spoiling the potential of combining several types of annotations. For this reason, mixing supervision for medical images analysis has gained interest in past years (Y.-J. Huang et al. 2020; J. Li et al. 2021; Z. Li et al. 2018).

For instance, in Mlynarski et al. (2019) the author showed that combining global labels and local annotations by training in a multi-task setting, the capacities of the model to segment brain tumors on Magnetic Resonance Images were improved.

In Tourniaire et al. (2021), the author introduced a mixed supervision framework for metastasis detection building on the CLAM (Lu et al. 2021) architecture. CLAM is a variant of the popular attention based MIL (Ilse, Tomczak, and Welling 2018) with 2 extensions: first, in order to make the method applicable in a multi-class setting, class-specific attention scores are learned and applied. Second, the last layer of the tile encoding network is trained to also predict the top and bottom attention scores, thus mimicking tile-level annotations. In Tourniaire et al. (2021), the authors highlight the limitations of this instance-classification approach and propose to leverage a low number of fully annotated slides to train the attention mechanism. In a second step, they propose to turn to a standard MIL training (using only slide-level annotations). Even with few annotated slides, this approach allows to boost classification performance. However, there are also some limitations. First, the method relies on exhaustive annotation of selected slides: for the annotated slides, all the key regions are annotated pixel-wise. Second, due to the CLAM architecture, the approach only fine-tunes a single dense layer downstream the pre-trained feature extractor. Third, the algorithm has been designed for an application case in which the slide and tile labels coincide (tumour presence). This however is not always the case: when predicting genetic signatures, grades or treatment responses, it is unclear how tile and slide level annotations relate to each other. In this article, we propose to overcome these limitations. We propose to combine self-supervised learning with supervision prior to training the MIL network. We thus start from more powerful encodings, that are not only capable of solving the pretext task of self-supervised learning, but also the medical classification task that comes with the annotated tiles. Consequently, this method does not require full-slide annotations, optimizes the full tile encoding network and does not come with any constraint regarding the relationship between tile and slide level annotations.

IV.3 Materials and Method

IV.3.0.1. Dataset and Problem Setting

The Tissue Net Challenge DrivenData organized in 2020, the Société Française de Pathologie (SFP) and the Health Data Hub aimed at developing methods to automatically grade lesions of the uterine cervix in four classes according to their severity. The training dataset for the challenge was made up of biopsy samples from female uterine cervix, focusing on squamous lesions (Figure IV.2). These lesions are often benign but can also be qualified as low grade or high grade depending on the risk of invasion of the underlying conjunctive tissue and evolution into carcinomas. The grade of the lesions depends on the proportion of squamous epithelium affected by dysplastic criteria. Lowgrade squamous intraepithelial lesions (LSIL) are defined as having a dysplastic criteria involving less than one third of the thickness of the epithelium. High-grade squamous intraepithelial lesions (HSIL) indicate a greater proportion of the epithelium composed of undifferentiated basal cells with abnormalities. Carcinoma is diagnosed when abnormal epithelial cells invade the underlying conjunctive tissue. The class of a WSI was determined by the highest lesion's grade present on it.

IV.3.0.2. Fully Supervised Dataset

5926 annotated Regions of Interest (ROIs) of fixed size 300x300 micrometers were provided. Each ROI had roughly the same size as a tile at 10x magnification and were labeled by the severity of the lesion it contained: "Normal" (0) if tissue was normal, (1) LSIL or (2) HSIL if it presented precancerous lesions that could have malignant potential and (3) invasive squamous carcinoma (Table IV.1).

Classes	Number of Slides	Number of Tiles
0 (Normal)	270	1923
1 (Low Grade)	288	1405
2 (High Grade)	238	1368
3 (Carcinoma)	219	1230
Total	1015	5926

Table IV.1.: Dataset Summary

IV.3.0.3. Weakly Supervised Dataset

The dataset was composed of 1015 WSIs acquired from 20 different centers in France at an average resolution of $0.234 +/- 0.0086 \mathrm{mpp}(40\mathrm{X})$. The slide resolution varied slightly due to the multicentric provenance of the data. The class of the WSI corresponded to the class of the most severe lesions it contained (grade from 0 to 3 also). All the native WSI formats were converted to pyramidal TIFF (Tagged Image

File Format). Both the WSI-level and tile-level labels have been attributed by a consortium of expert pathologists (Table IV.1).

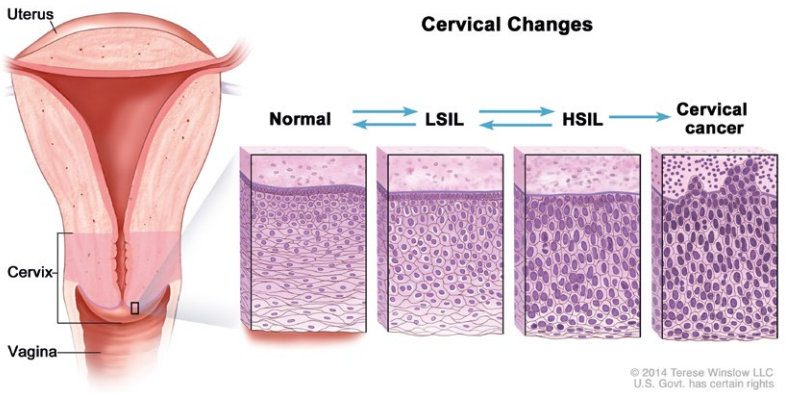


Figure IV.2.: Illustration of Uterine cervix dysplasia - National Cancer Institute, 2011 provenance of the data. The class of the WSI corresponded to the class of the most severe lesions it contained (grade from 0 to 3 also). All the native WSI formats were converted to pyramidal TIFF (Tagged Image File Format). Both the WSI-level and tile-level labels have been attributed by a consortium of expert pathologists (Table IV.1)

IV.3.0.4. Misclassification Costs

Misclassification errors do not lead to equally serious consequences (i.e predicting a benign lesion if it is cancerous is more serious than predicting a LSIL instead of a HSIL). Accordingly, a panel of pathologists established a grading of each of these errors i.e they attributed to each pair of possible outcome $(i, j) \in \{0, 1, 2, 3\}^2$ a severity score $0 \leqslant C_{i,j} \leqslant 1$ (Table IV.2)

The metric used in the challenge to evaluate and rank the submissions is computed from the average of these misclassification costs.

More precisely, if we name $P(S)$ the prediction of a slide S labelled $l(S)$, the challenge metric M_{WA} is:

$$M_{WA} = \frac{1}{N} \sum_{S} \left(1 - C_{l(S),P(S)}\right)$$

with N the number of samples.

The problem is thus framed as a cost-sensitive classification problem, and, to our knowledge, all the winning solutions took awareness of this cost in their training procedure.

Ground Truth/pred	Benign	Low-grade	High-grade	Carcinoma
Benign	0.0	0.1	0.7	1.0
Low-grade	0.1	0.0	0.3	0.7
High-grade	0.7	0.3	0.0	0.3
Carcinoma	1.0	0.7	0.3	0.0

Table IV.2.: Weighted Accuracy Error Table - Error table to ponderate misclassification according to their gap with the ground truth.

IV.4 Proposed Architecture

IV.4.0.1. Multiple Instance Learning and Attention

In Multiple Instance Learning, we are given sets of samples $B_k = \{x_i \mid i = 1 \dots N_k\}$, also called bags. The annotation y_k we are given refers only to the bags and not the individual samples. We assume however, that such tile-level labels exist in principle, but that we just do not have access to them. The strategy is to first map each tile x_i to its encoding z_i, which is then mapped to a scalar value a_i, often referred to as attention score. The tile representations z_i and attention scores a_i are then agglomerated to build the slide representation s_k which is then further processed by a neural network. The agglomeration can be based on tile selection (Campanella, Hanna, Geneslaw, Miraflor, Silva, et al. 2019; Courtiol et al. 2018), or on an attention mechanism (Ilse, Tomczak, and Welling 2018), which is today the most widely used strategy.

IV.4.0.2. Self-Supervised Learning

Self-supervised learning provides a framework to train neural networks without human supervision. The main goal of self-supervised learning is to learn to extract efficient features with inputs and labels derived from the data itself using a pretext task. Many self-supervised approaches are based on contrastive learning in the feature space. SimCLR, a simple framework relying on data augmentation was introduced in T. Chen, Kornblith, Norouzi, et al. (2020). Powerful feature representations are learned by maximizing agreement between differently augmented views of the same data point via a contrastive loss applied in the feature space.

An image is transformed through random data augmentations into two new images. They are then embedded using the feature extractor. The two features vectors (z_i and z_j) are mapped with a projection head (dense layers) to obtain final vectors h_i and h_j. The feature extractor and projection head are trained to maximize agreement using the contrastive loss. Positive pairs consist of the two augmented views of the same image, the other $2(n-1)$ views play the role of negative samples. The loss function (NT-Xent) for a positive pair (i, j) is defined as:

$$\mathcal{L}_{SSL} = -\log \frac{\exp\left(\text{sim}\left(h_i, h_j\right)/\tau\right.}{\sum_{k=1}^{2n} \mathbf{1}_{k\neq i} \exp\left(\text{sim}\left(h_i, h_j\right)\tau\right.}$$

Where $\text{sim}(u, v) = \frac{u^T v}{\|u\| \cdot \|v\|}$, the cosine similarity, $\mathbf{1}_{k\neq i\in(0,1)}$ determines if $k \neq i$ and τ is a parameter. After convergence, the projection head is discarded and the pretrained feature extractor can be used for subsequent tasks.

IV.4.0.3. Cost-Sensitive Training}

Instead of the traditional cross-entropy loss we used a cost-aware classification loss, the SmoothOne-Sided Regression Loss $\mathcal{L}_{\text{SOSR}}$. First introduced to train SVMs in Tu and Lin (n.d.), this objective function was smoothed and adapted for backpropagation in deep networks in Chung, Lin, and Yang (2016). When using this loss, the network is trained to predict the class-specific risk rather than a posterior probability; the decision function chooses the class minimizing this risk.

The SOSR loss is defined as follows:

$$\mathcal{L}_{SOSR} = \sum_i \sum_j \ln\left(1 + \exp\left(\mathbf{2}_{i,j} \cdot (\hat{c}_i - \mathcal{C}_{i,j})\right)\right) \tag{IV.1}$$

With $\mathbf{2}_{i,j} = -\mathbf{1}_{i\neq j} + \mathbf{1}_{i=j}, \hat{c}_i$ the i-th coordinate of the network output and $\mathcal{C}$ the error table.

IV.4.0.4. Mixed Supervision

To be tractable, training of attention-MIL architectures requires freezing the feature extractor weights. While SSL allows the feature extractor to build meaningful representations (Dehaene et al. 2020; Saillard et al. 2021), they are not specialized to the actual classification problems we try to solve. Several studies have shown that such SSL models benefit from fine-tuning specific to the downstream task (T. Chen, Kornblith, Swersky, et al. 2020)

We therefore added a training step to leverage the tile-level annotation and fine-tune the selfsupervised model. However, as the final WSI classification task is not identical to the tile classification task, we suspect that fine-tuning solely on the tile classification task may over-specialize the feature extractor and thus sacrifice the generalizability of SSL (and for this reason ultimately also degrading the WSI classification performances). To avoid this, we developed a training process that optimizes the self-supervised and tile-classification objectives jointly.

Two different heads, plugged before the final classification layer, are used to compute both loss functions $\mathcal{L}_{SSL}$ and $\mathcal{L}_{SOSR}$ The final objective $\mathcal{L}$ is then:

$$\mathcal{L} = \beta\mathcal{L}_{SSL} + (1-\beta)\mathcal{L}_{SOSR} \tag{IV.2}$$

where β is a hyperparameter that has to be tuned. Here, we found $\beta = 0.3$ (see Supplementary).

IV.5 Understanding the Feature Extractor with Activation Maximization

To further understand the features learned by the different pre-training policies (ImageNet, supervised, SSL and mixed), we used Activation Maximization (AM) to visualize extracted features and provide an explicit illustration of the specificity learned.

Methods to generate pseudo-images maximizing a feature activation have been introduced in Erhan et al. (2009). This technique consists in synthesizing the images that will maximize one feature activation. It is summarized as follow (Nguyen, Yosinski, and Clune 2019):

If we consider a trained classifier with set of parameters θ that map an input image $x \in \mathbb{R}^{h \times w \times c}$, (h and w are the height and width and c the number of channels) to a probability distribution over the classes, we can formulate the following optimization problem:

$$x^* = \arg\max_x \left(\sigma_i^l(\theta, x)\right)$$

where $\sigma_i^l(\theta, x)$ is the activation of the neuron i in a given layer l of the classifier. This formulation being a non-convex problem, local maximum can be found by gradient ascent, using the following update step:

$$x_{t+1} = x_t + \epsilon \frac{\partial \sigma_i^l(\theta, x)}{\partial x_t}$$

The optimization process starts with a randomly initialized image. After a few steps, it generates an image which can help to understand what information is being captured by the feature. As we try to visualize meaningful representations of the features, some regularization steps are applied to the random noise input (random crop and rotations to generate more stable visualization, details can be found in Supplementary Materials). To generate filter visualization within the HE space, we transformed the RGB random image to HE input thanks to color deconvolution (Ruifrok, n.d.). This preprocessing allowed to generate images with histology-like colors when converted back to the RGB space.

To select the most meaningful features for each class, we trained a Lasso classifier without bias to classify the extracted feature vectors into the four classes of the dataset for the four pre-training policies. The feature vectors for each tile were first normalized and divided element-wise by the vector of features' standard deviation

across all the tiles. The L1 regularization factor λ was set to 0.01. Details about Lasso training can be found in Supplementary Materials. Contribution scores for each feature were therefore derived from the weights of the Lasso linear classifier: negative weights were removed and remaining positive weights were divided by their sum to obtain contribution scores $[0, 1]$. By filtering out the negative weights, the contribution score corresponds to the proportion of attribution among the features positively correlated to a class, and allows to select feature capturing semantic information related to the class, leaving out those containing information for other classes.

IV.6 Experimental Setting

IV.6.0.1. WSI Preprocessing

Preprocessing on a downsampled version of the WSIs was applied to select only tissue area and non-overlapping tiles of 224×224 pixels were extracted at a resolution of 1mpp. (Details in Supplementary Materials)

IV.6.0.2. Data Splits for Cross-Validation

To measure the performances of our models we performed 3-fold cross-validation for all our training settings. Because the annotated tiles used in our joint-optimization step were directly extracted from the slides themselves, we carefully split the tiles such that tiles in different folds were guaranteed to originate from different slides. The split divided the slides and tiles into a training set, a validation set and a test set.

All subsequent performance results are then reported as the average and standard deviation of the performance results on each of these 3 test folds.

IV.6.0.3. Feature Extractor Pre-Training

The feature extractor is initialized with pre-trained weights obtained with three distinct supervision policies: fully supervised, self-supervised or a mix of supervision. These three policies rely on the fine-tuning of a DenseNet121 (G. Huang et al. 2018), pretrained on ImageNet. The fully-supervised archicture is fine-tuned solely on the tile classification task. The SSL architecture is derived from SimCLR framework and is trained on an unlabeled dataset of 1 million tiles extracted from the slides. Finally for the mixed-supervised architecture, a supervised branch is added to the previous SSL network and trained using the mixed objective function (see Figure IV.1 and Equation (IV.2)) on the fully supervised dataset. Technical details of these three training settings are available in the supplementary material.

IV.6.0.4. Whole Slide Classification

After tiling the slides, the frozen feature extractor (DenseNet121) was applied to extract meaningful representations from the tiles. This feature extractor was initialized sequentially with the pre-trained weights mentioned above and generated as many sets of features. These bags of features were then used to train the Attention-MIL model with SOSR loss applied slide-wise. (Supplementary Materials).

IV.6.0.5. Feature Visualization

To select the most relevant features, we trained an unbiased linear model on the feature vectors extracted from the annotated tiles. The feature vectors were standardised. The weights of the linear model were used to determine which features were the most impactful for each class. Feature visualizations were generated for the selected features and for each set of pre-trained weights. We extracted the tiles expressing the most of these features by selecting the feature vectors with the higher activation for the concerned feature. Implementation details are provided in Supplementary Materials.

IV.7 Results

IV.7.0.1. Self-Supervised Fine-Tuning

We saved the checkpoints of the self-supervised feature extraction model at each epoch of training, allowing us to investigate the amount of time needed to reach good WSI classification performances. We computed the embeddings of the whole dataset with each of the checkpoints and trained a WSI classifier from them. Figure IV.3 reports the performances of WSI classification models for each of these checkpoints. SSL training led to a higher Weighted Accuracy than using ImageNet weights after 3 epochs and resulted in a gain of $+4.8\%$ after 100 epochs. Interestingly, as little as 6 epochs of training are enough to gain 4% of Weighted Accuracy: a significant boost in performance is possible with 50 GPU-hours of training. We then observe a small increase in performance until the 100th epochs.

IV.7.0.2. Pre-Training Policy Comparison

To compare the weights obtained with the various supervision levels, we ran a 3-fold cross-validation on the WS classification task and summarized the results in Table IV.3. The results indicate that the SSL pre-training substantially improves the WSI classification performance. In contrast, we see that initializing the feature extractor with fully-supervised weights gives an equivalent or poorer performance than any other initialization. SSL pre-training allows us to extract rich features that are generic, yet still relevant to the dataset (unlike ImageNet). On the other

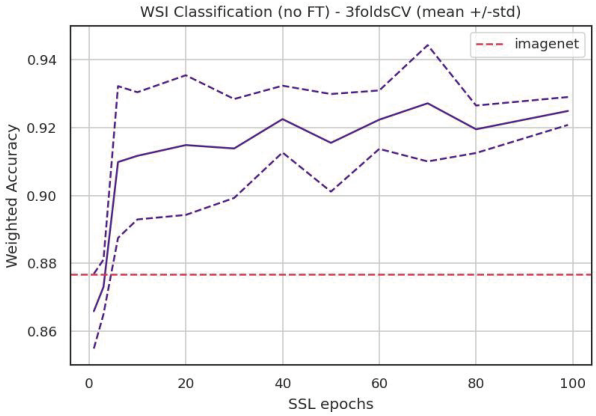


Figure IV.3.: Weighted Accuracy evolution - Weighted Accuracy evolution on WS classification task with respect to the number of epochs of SSL training

hand, fully supervised features are probably too specific and seem to not represent the full diversity of the image data. The joint-optimization process manages to balance out generic and specialized features without neutralizing them: mixing the supervision levels brings significant improvements $(+2\%)$ to the performance, leading to a Weighted Accuracy of 0.945 .

We additionally compared the benefits introduced by the cost-sensitive loss (Equation (IV.1)) with the crossentropy loss. Our results show that with ImageNet weights the SOSR loss improves the Weighted Accuracy by 1% and the accuracy by 3%.

In conclusion, the combination of the SSL pre-trained model, its fully supervised fine-tuning, and the cost-sensitive loss leads to a notable improvement of 8 Weighted Accuracy points over the baseline MIL-imagenet model.

	Accuracy	SFP-metric
imagenet+ce	0.758 ± 0.034	0.865 ± 0.023
imagenet+sors	0.787 ± 0.032	0.877 ± 0.029
supervised+sors	0.772 ± 0.055	0.874 ± 0.027
ssl+sors	0.803 ± 0.016	0.925 ± 0.006
mixed+sors	$\mathbf{0.845 \pm 0.028}$	$\mathbf{0.945 \pm 0.005}$

Table IV.3.: Pre-training policies - Performances summary

IV.7.0.3. Number of Annotations vs Number of Epochs

We have seen that both SSL and supervised pre-training bring together an improvement in the WSI classification task. To further investigate the relationship between these two supervision regimes, we trained models with only some of the fully supervised annotations $(15, 65, 100\%)$ on top of intermediate SSL checkpoints. Results are reported in Table IV.4.

It appears that without SSL pre-training (or with too few epochs of training), the supervised finetuning does not bring additional improvement in WSI classification. This is in line with the work of Chen et al. (Chen, Kornblith, Swersky, et al. 2020) that showed that an SSL model is up to 10x more label efficient than a supervised one.

However, for the 100-epoch checkpoint, we observe an improvement of 2 points of the Weighted Accuracy when using 100% of the tile annotations. Moreover, fine-tuning the models by mixed supervision with too few annotations (15%) leads to a slight drop in WSI classification performances. Finally, we see a diminution of the standard deviations across splits for the different pre-training policies, showing better stability for longer SSL training and more annotations.

We draw different conclusions from these observations:

- In this context, it is always better to pre-train the feature extractor with SSL rather than only invest in annotations.
- The supervised fine-tuning needs enough annotations to bring an improvement to the WSI classification task. We can note however that even when considering the 100% annotation settings, the supervised dataset (approx. 5000 images) is still rather small in comparison to traditional image datasets.
- A full SSL training is mandatory to leverage this small amount of supervised data.

IV.7.0.4. Features Visualisations

We generated the pseudo-images of the most important features for each class and each pre-training policy and extracted the related tiles. The Figure IV.5 displays the most important features along with

	0 Annot.	$\sim$ 1 Annot. / slide (1015 tiles)	$\sim$ 4 Annot. / slide (3901 tiles)	$\sim$ 6 Annot. / slide (5926 tiles)
ImageNet (no SSL)	$0,877 \pm 0.029$	0.872 ± 0.024	0.872 ± 0.023	$0,874 \pm 0.027$
SSL-epoch10	$0,912 \pm 0.019$	$0,907 \pm 0.024$	$0,903 \pm 0.029$	$0,916 \pm 0.019$
SSL-epoch50	$0,915 \pm 0.014$	$0,913 \pm 0.024$	$0,916 \pm 0.014$	$0,914 \pm 0.022$
SLL-epoch100	$0,925 \pm 0.006$	$0,916 \pm 0.010$	$0,921 \pm 0.010$	$\mathbf{0,945 \pm 0.005}$

Table IV.4.: Relationship between self-supervision and full-supervision - Study on the performance improvement on WS classification for different proportion of labelled data versus different training time of SSL

the tiles activating each feature the most for the class “Normal” (0). Although interpretation of such pseudo images must be treated carefully, we notice that the features obtained with SSL, supervised and mixed training are indubitably more specialized to histological data than those obtained with ImageNet. Some histological patterns, such as nuclei, squamous cells or basal layers are clearly identifiable in the generated images. The extracted tiles are strongly correlated with class-specific biomarkers. Feature e represents a normal squamous maturation, i.e. a layer of uniform and rounded basal cells, with slightly larger and bluer

nuclei than mature cells. We can also observe several layers of mature cells with small nuclei and moderately abundant cytoplasm (pink halo around), equidistant from each other. Features $\mathbf{c}$ and $\mathbf{d}$ highlight clouds of small regular and rounded nuclei (benign cytological signs). Feature $\mathbf{g}$ and $\mathbf{h}$ are characteristic of squamous cells (polygonal shapes, stratified organization lying on a straight basal layer). Interestingly, features extracted with the supervised method $(\mathbf{g}, \mathbf{h})$ manage to sketch a normal epithelium with high resemblance, the features are more precise. On the other hand, features extracted with SSL (c, d) highlight true benign criteria but do not entirely summarize a normal epithelium (no basal maturation). The mixed model displays both, suggesting that mixed supervision highlights pathologically relevant patterns to a larger extent than the other regimes (Sellors and Sankaranarayanan 2003).

We can also note by looking at the real tiles that while features from ImageNet $(\mathbf{a}, \mathbf{b})$, SSL $(\mathbf{c}, \mathbf{d})$ and the supervised model $(\mathbf{g}, \mathbf{h})$ focus on the upper half of the cervix epithelium, it appears that features from the mixed supervision model $(\mathbf{e}, \mathbf{f})$ are focusing on the lower half which is known to be the relevant region for discrimination between class Normal (0) and Low Grade (1) (abnormal cells are constricted to the lower third of the epithelium).

In Figure IV.4 we can further identify class-related biomarkers for dysplasia and carcinoma grade. Tiles with visible koilocytes (cells with a white halo around the nucleus) have been extracted from the top features for Low Grade class. Koilocytes are symptomatic of infection by Human Papillomavirus and are a key element for this diagnosis (almost always responsible for precancerous lesions in the cervix, (Sellors and Sankaranarayanan 2003)). High Grade (2) generated image represents disorganised cells with a high nuclear-to-cytoplasmic ratio, marked variations in size and shape and loss of polarity. For the class "Carcinoma" (3), we observe irregular clusters of cohesive cells with very atypical nuclei, separated by a fibrous texture that can be identified as stroma reaction. All these criteria have been identified in Sellors and Sankaranarayanan (2003) as key elements for diagnosis of dysplasia and invasive carcinoma. In Figure IV.6, we observe that features extracted from ImageNet and SSL models are diverse, in particular, features extracted from SSL reflect rich tissue phenotypes which correlates to their generic capacities of image representations. On the other hand, features extracted with supervised and mixed methods are more redundant. We additionally observe in Figure IV.6 that feature visualisation from the mixed model picture realistic histopathological patterns specific to the class. Visualisation for other classes are available in Supplementary Materials.

IV.8 Discussion

In pathology, expert annotations are usually hard to obtain. However, we are often in a situation where a small proportion of labeled annotation exists but not in sufficient quantities to support fully supervised techniques. Yet, even in small quantities, expert annotations carry meaningful information that one could use to enforce biological

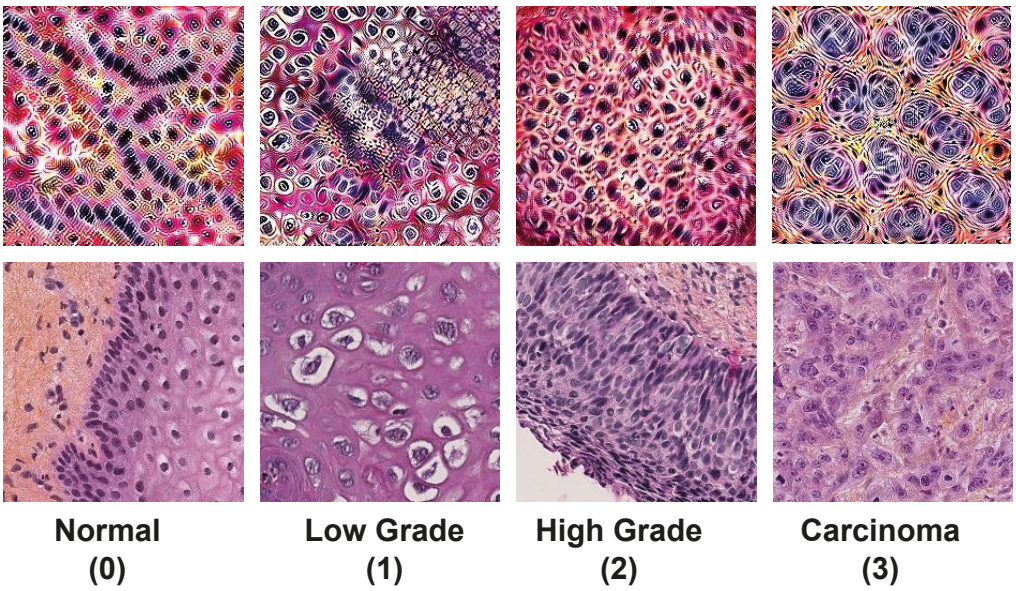


Figure IV.4.: Feature comparison per class - The top row displays the top filter for the Mixed Supervised model for each class. The bottom row displays the tile expressing the feature the most.

context to deep learning models and make sure that networks learn appropriate patterns. On the other hand, self-supervised methods have proven their efficacy to extract generic features in the histopathological domain and their usefulness for downstream supervision tasks, even in the absence of massive ground truth data. Methods capable of reconciling self-supervision with strong supervision can therefore be useful and open the door to better performances.

In this paper, we presented a way to inject the fine-grained tile level information by fine-tuning the feature extractor with a joint optimization process. This process allowed to mix self-supervised learning features with tile classification ones and helped the downstream WSI classification task.

We applied our method to the TissueNet Challenge, a challenge for the automatic grading of cervix cancer, that provided annotations at the slide and tile level, thus representing an appropriate use case to validate our method of mixed supervision. We also propose in this study insights and guidelines for the training of a WSI classifier in the presence of tile annotations.

First, we showed that SSL is always beneficial to our downstream WSI classification tasks. Fine-tuning pre-trained weights with SSL for only 50 hours brings a 4% improvement over WSI classification weighted accuracy, and near to 5% when fine-tuning for longer (100 epochs).

Second, a small set of annotated tiles can bring benefit to the WSI classification task, up to 2% of weighted accuracy for a supervised dataset of around 5000 images.

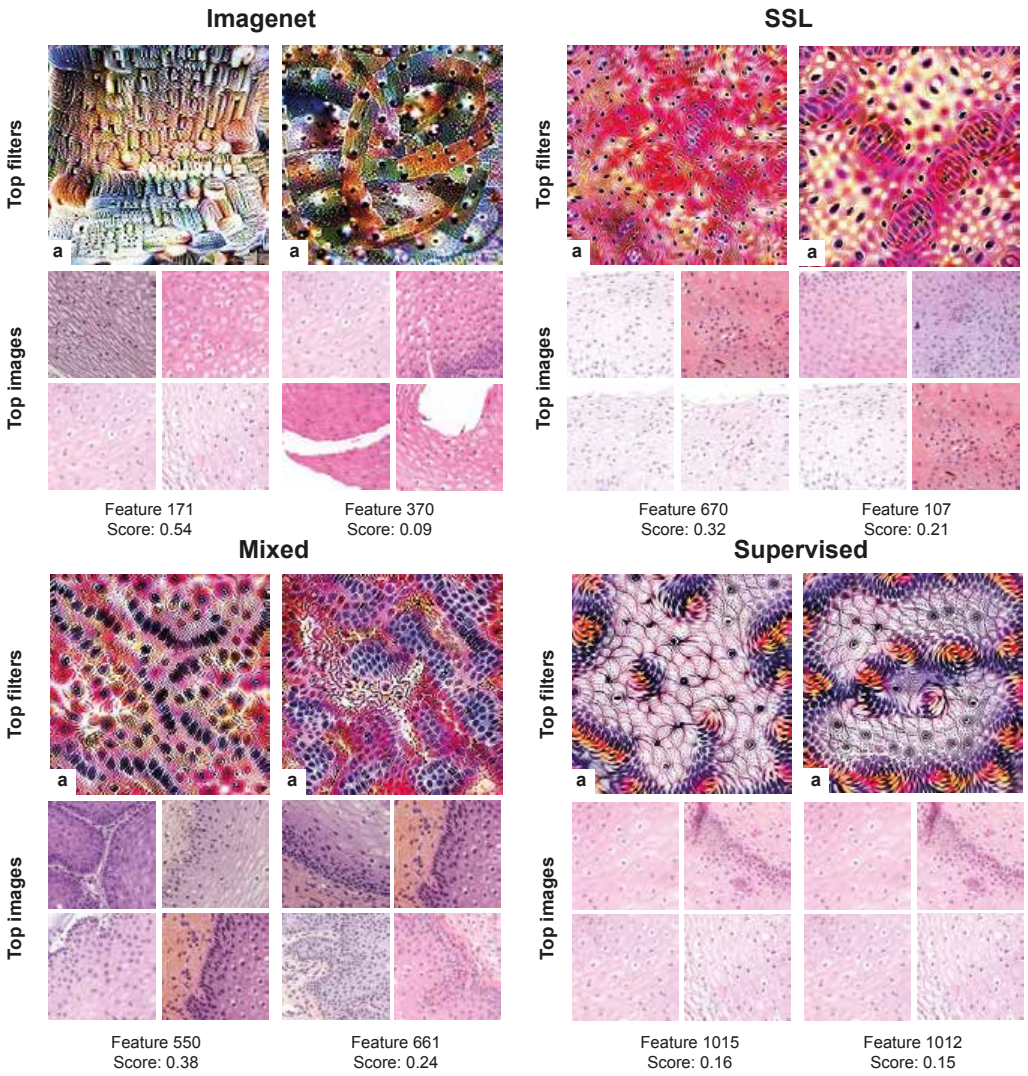


Figure IV.5.: **Feature Visualization** - Top Features for class “Normal” (0) and associated tiles.

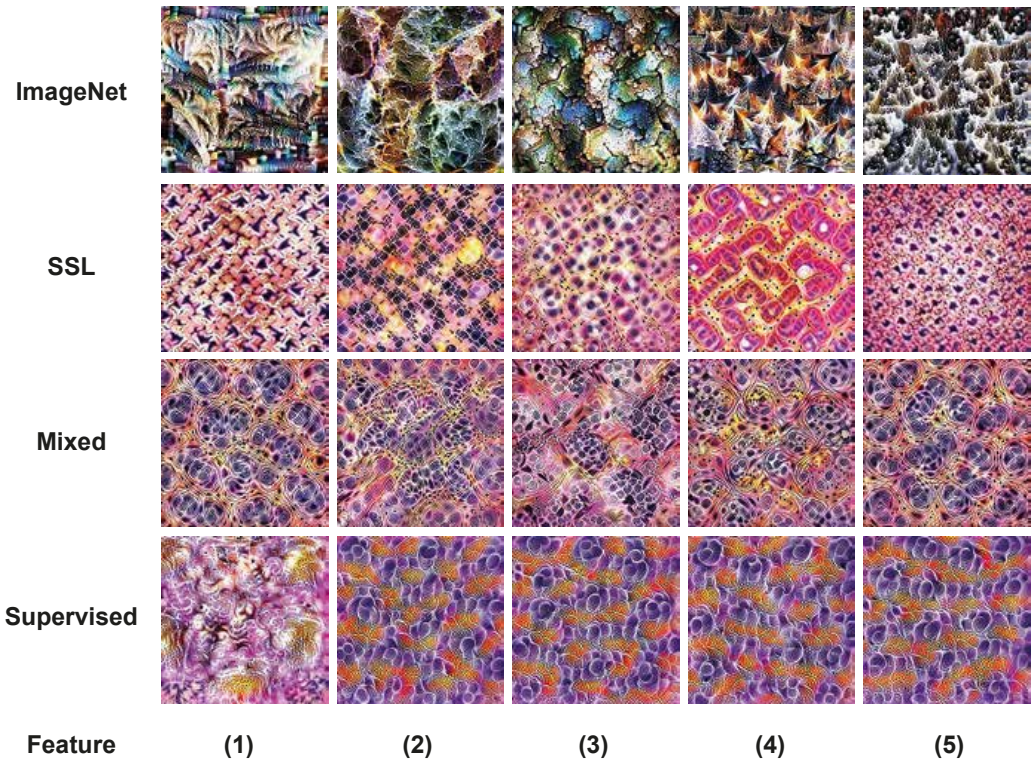


Figure IV.6.: Feature Diversity for the class "Carcinoma" (3) (top 5 features) - Class "Normal" (0) and top 10 features in Supplementary Materials

Such a set of tiles can be obtained easily by asking the pathologist to select a few ROIs that guided his decision while labeling the WSIs, which can be achieved without a strong time commitment. However this boost in performance can be reached only if the feature extractor is pre-trained with SSL, and for sufficiently long: SSL unlocks the supervised fine-tuning benefits.

To further understand the differences between the range of supervision used to extract tile features, we conducted qualitative analysis on features visualizations by activation maximization and observed that features obtained from SSL, supervised or mixed trainings were more relevant for histological tasks and that class-discriminative patterns were indeed identified by the model. We also observed that supervised training on the tiles alone led to much less diversity in the features extracted by the model than the ones obtained with SSL.

The scope of this study contains by design three limitations. First, SSL models were trained by fine-tuning already pre-trained weights on imagenet. This may explain the rapid convergence and boost in performance observed; however it may also underestimate this boost if the SSL models were trained from scratch. We did not compare SSL trained from scratch and fine-tuned SSL, and left it to future work.

Second, all the conclusions reached are conditioned by the fact that we do not fine-tune the feature extractor network during the WSI classification training. Keeping these weights frozen, and even pre-computing the tiles representations brings a large computational benefit (both in memory and speed of computations), but prevents the feature extractor from specializing during the WSI classification training.

Third, the tendency observed in table 4 of better performances correlated with larger numbers of annotations is modest and would require more annotations to validate it.

Finally, our method can be improved in several ways. First, SimCLR, was a pioneer method in self-supervised learning architecture and has proven to be efficient but it suffers from high performance drop when decreasing the batch size (T. Chen, Kornblith, Norouzi, et al. 2020). Other SSL models have been developed to alleviate this limitation. MoCo (He et al. 2020) actually propose a momentum mechanism allowing optimal performances even without large batch size and therefore, numerous available parallel GPUs. Other models like VICReg (Bardes, Ponce, and LeCun 2022) proposed techniques to maximize the variance between the features and therefore limit their redundancies. It will be interesting to benchmark these SSL variants with respect to their impact on WSI classification accuracy and feature interpretability.

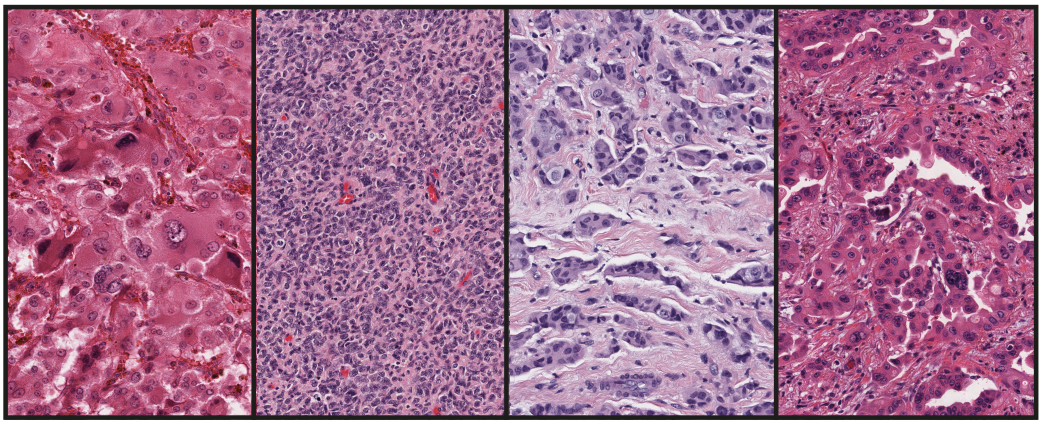
To conclude, we present a method that provides an interesting alternative to using full supervision, pre-training on unrelated data sets or self-supervision. We convincingly show that the learned feature representations are both leading to higher performance and providing more intermediate features that are more adapted to the problem and point to relevant cell and tissue phenotypes. We expect that the mixed supervision will be adopted by the field and lead to better models.

IV.8.0.1. Acknowledgments

The authors thank Etienne Decencière for the thoughful discussions that helped the project. ML was supported by a CIFRE PhD fellowship founded by KEEN EYE, Paris, France and ANRT (CIFRE 2019/1905). TL was supported by a Q-Life PhD fellowship (Q-life ANR-17-CONV-0005). Furthermore, this work was supported by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19P3IA-0001 (PRAIRIE 3IA Institute).

V

Learning WSI representations without supervision



Contents

Preface

This chapter centers on the Giga-SSL framework and explores its various applications across different sections. The framework employs a self-supervised learning approach specifically designed to learn WSI representations.

Applying SSL to WSIs posed a challenge. While WSIs are essentially large images, they haven't traditionally been treated as such in computational pathology. Instead, they are typically divided into tiles and treated as collections of these tiles. Until recently, the spatial relationship between the tiles was largely ignored, with the community focusing on MIL algorithms to solve WSI classification tasks.

This discrepancy led to questions about how to adapt self-supervised learning methods for WSIs. During the development of the MIL algorithms presented in Chapter III, I observed that sampling only a small fraction of WSI tiles during training was sufficient for effective classification on a variety of tasks. This performance reached a saturation point after a certain number of tiles were included, suggesting that crucial classification information resides in only a subset of the WSI.

Besides, a vital element of successful self-supervised training is the transformation applied to the image. This transformation should alter the image's numerical properties while preserving its semantic content. Given that WSIs could be strongly subsampled without losing key classification information, it became clear that self-supervised methods could be adapted to extract this critical biological information, focusing specifically on the tile subsampling transformation.

In collaboration with Marvin Lerousseau, we worked to identify the key components of a WSI-level self-supervised learning framework. These included a hybrid architecture of standard and sparse convolutions, the significance of shared tile augmentations—a visual interpretation of which can be found in Figure E.1 —and the finite approximation of tile augmentation, which allowed for scalability and ablation studies.

The Section V.1 of this chapter details the overall framework, training specificities, evaluation against a limited set of downstream tasks as well as ablation studies providing a detailed understanding of design choices.
The second Section V.2 demonstrates its applicability for the prediction of single-gene mutations and mutational signatures in a pan-cancer context. It also proves that resource-intensive experiments can now be carried out on a regular laptop using Giga-SSL. The final, shorter Section V.3 aims to interpret the latent spaces shaped by the training process with Giga-SSL. While the first two sections have either been published or are under review, the third section is not intended for publication but offers avenues for future research. The first section has been published at the CVPR workshop on Computer Vision for Microscopy Images (CVMI); the second section corresponds to a preprint and is currently under review. The third section has not been published.

As a postscript, a simpler variant of Giga-SSL (without convolutions) also achieved good results in the VisioMel DataChallenge, securing a third-place ranking based on negative log-likelihood loss metrics and second place in terms of the AUC metric.

Contributions

Publications - communications

- **Lazard, T.**, Lerousseau, M., Decencière, E., and Walter, T. (2023). Giga-SSL: Self-Supervised Learning for Gigapixel Images. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 2023 pp. 4305-4314. 10.1109/CVPRW59228.2023.00453
- **Lazard, T.**, et al. (2023). Democratizing Whole Slide Images: optimized representations for The Cancer Genome Atlas. *preprint*.

Open-source repository

- Package to perform Giga-SSL training.
- Reproduce the VisioMel results of Giga-SSL
- Repository to esily encode WSI using pre-trained Giga-SSL models.
- Optimized representations of public WSI datasets.

Other achievements

- Third place at the VisioMel challenge (cash prize: 5000$).

Summary:

Whole slide images (WSI) are microscopy images of stained tissue slides routinely prepared for diagnosis and treatment selection in medical practice. WSI are very large (gigapixel size) and complex (made of up to millions of cells). The current state-of-the-art (SoTA) approach to classify WSI subdivides them into tiles, encodes them by pre-trained networks and applies Multiple Instance Learning (MIL) to train for specific downstream tasks. However, annotated datasets are often small, typically a few hundred to a few thousand WSI, which may cause overfitting and underperforming models. Conversely, the number of unannotated WSI is ever increasing, with datasets of tens of thousands (soon to be millions) of images available. While it has been previously proposed to use these unannotated data to identify suitable tile representations by self-supervised learning (SSL), downstream classification tasks still require full supervision because parts of the MIL architecture is not trained during tile level SSL pre-training. Here, we propose a strategy of slide level SSL to leverage the large number of WSI without annotations to infer powerful slide representations. We show that a linear classifier trained on top of these embeddings maintains or improves previous SOTA performances on various benchmark WSI classification tasks. We also show the high label-efficiency of these linear models compared to MIL models. Then, we showcase the abilities of Giga-SSL representations on a large set of classification tasks (1288) across various cancer types in the TCGA (14). The number of mutations predictable using Giga-SSL roughly doubled compared to the previous MIL method, and we also observed an improvement in the classification performance. Finally, we present a simple yet flexible framework to interpret the latent space of the Giga-SSL embeddings and therefore lead phenotypic studies at the scale of the whole TCGA.

Résumé:

Les WSIs sont des images microscopiques de lames de tissus colorées, préparées en routine pour le diagnostic et le choix du traitement dans la pratique médicale. Les images de lames entières sont très grandes (de l'ordre du gigapixel) et complexes (composées de millions de cellules). L'approche actuelle de l'état de l'art (SOTA) pour classer les WSI consiste à les découper en tuiles, à les encoder à l'aide de réseaux pré-entraînés, et à entraîner des réseaux via l'apprentissage par instances multiples (MIL) pour résoudre des tâches de classification spécifiques en aval. Toutefois, les ensembles de données annotées sont souvent de petite taille, généralement de quelques centaines à quelques milliers de WSI, ce qui peut entraîner un risque de surapprentissage. En revanche, le nombre de WSI non annotés ne cesse d'augmenter, avec des ensembles de données comptant des dizaines de milliers de WSI. Nous proposons ici une stratégie de SSL au niveau des WSIs qui exploite le grand nombre de WSI non annotés afin de construire des représentations vectorielles puissantes. Nous montrons qu'un classificateur linéaire entraîné sur ces représentations numériques maintient ou améliore les performances SOTA sur diverses tâches de classification de WSI de référence. Nous montrons également que l'utilisation de ces représentations apporte un avantage d'autant plus compétitif que les jeux de données utilisés sont petits. Ensuite, nous présentons les capacités des représentations Giga-SSL sur un grand ensemble de tâches de classification (1288) à travers divers types de cancer dans le TCGA (14). Le nombre de mutations prédictibles à l'aide de Giga-SSL est environ doublé par rapport aux méthodes de MIL, et nous avons également constaté une amélioration des performances de classification dans ce contexte pan-cancer. Enfin, nous présentons une méthode simple mais flexible pour interpréter l'espace latent des représentations Giga-SSL, ce qui nous permet de mener des études phénotypiques à l'échelle de l'ensemble du TCGA.

This chapter has been made in collaboration with M. Lerousseau, E. Decencière and T. Walter. It has been published in the CVPR "workshop for computer vision for microscopy" proceedings.

V.1 Giga-SSL: Self-Supervised Learning for Gigapixel Images

V.1.1 Introduction

Whole slide images (WSI) are microscopy images of stained tissue sections. They are enormous (billions of pixels) and complex, often containing millions of individual cells, their environments, and the overall tissue structure. They are routinely used in cancer treatment centers for diagnosis, patient stratification, and treatment selection. Computational pathology is the field concerned with the automatic analysis of WSI. The most clinically impactful task in computational pathology is to make predictions directly from the WSI, such as predicting cancer subtype, survival of the patient, or response to treatment. The major challenges in building predictive models operating on WSI are:

- Prohibitive memory requirements (typically 15GB uncompressed per WSI);
- Signal/noise: The high amount of biological material, not necessarily related to the output variable, is making models: (i) fail to identify the region of interests; (ii) prone to overfitting.
- Technical complexity: WSI are technically demanding to deal with given their large size, which presents a considerable barrier for multi-modal analyses of genomic and pathology data.

Today, the leading methods for WSI classification rely on Multiple Instance Learning (MIL): WSI are tessellated into small images, called tiles, which are encoded by an embedder. Tile embedders are usually pre-trained, either on natural images or - more recently and with great effect - by self-supervised learning (SSL). WSI are then seen as bags of tiles, and the slide representation is obtained by combining the tile embeddings, which are then used as input for the slide classification network. The agglomeration strategy comes in different flavors and usually relies on tile selection or weighted averaging of tile embeddings (Courtiol et al. 2019a; Ilse, Tomczak, and Welling 2018; B. Li, Li, and Eliceiri 2021; Lu et al. 2020; Rymarczyk, Tabor, and Zieliński 2020). The slide classification network is usually trained from scratch on the specific classification task.

While these methods successfully predict a large variety of output variables, such as grade, cancer subtype, gene signatures, mutations or response to treatment (Campanella, Hanna, Geneslaw, Miraflor, Silva, et al. 2019; Coudray et al. 2018; Echle et al. 2021; Jakob Nikolas Kather et al. 2020; Lazard et al. 2022; Naylor et al. 2022; Qu et al. 2021) , the performances remain highly dependent on the size of the training dataset (Campanella, Hanna, Geneslaw, Miraflor, Silva, et al.

2019). Indeed, MIL performance reaches saturation when using thousands of slides with associated ground truth for training (Campanella, Hanna, Geneslaw, Miraflor, Silva, et al. 2019). This might be realistic for the most frequent cancer types and routinely acquired output variables, but in most real-world projects only a few tens or hundreds of WSI with corresponding ground truth are available. However, with the digitalization of many pathology facilities, there is an increasing access to WSI without ground truth which are digitalized in clinical routine. Following the SSL paradigm that has been successfully applied at the tile level (Ciga, Xu, and Martel 2021; Dehaene et al. 2020; Lazard et al. 2022; Saillard et al. 2021) , there is a challenging opportunity to make use of these unannotated data at the slide level to derive meaningful slide representations. These would be particularly useful for small cohorts and non-standard output variables, such as prognosis for rare cancer types or prediction of treatment response in clinical trials.

However, learning representations at the WSI level is difficult since WSI cannot be manipulated as one image object due to their size, impeding the straightforward use of self-supervised learning frameworks developed on natural images. The community needs to innovate to translate SSL at the WSI level regarding the design of pertinent augmentations. For instance, the crop augmentation plays a central role for learning good representations with SSL on natural images (T. Chen, Kornblith, Norouzi, et al. 2020; Misra and van der Maaten 2019). However, randomly cropping one memory-fittable image from a WSI can lead to a complete loss of the cells and tissues that determine its ground-truth, due to the inherent heterogeneity of tissues. Further developments should also be done on the architecture of a SSL framework for WSI representations, as was done in the only paper tackling SSL at the WSI level (Richard J. Chen et al., n.d.).

Here, we propose Giga-SSL, a strategy to perform SSL for gigapixel images. Designed for pathology data, our method is capable of leveraging large datasets, such as The Cancer Genome Atlas (TCGA) (Weinstein et al. 2013), to learn representations at the WSI level without using any ground truth data – but only whole slide images. Our main contributions are:

- Giga-SSL, an efficient self-supervised learning framework for learning discriminative WSI representations.
- Extensive experiments show that a linear classifier that uses these embeddings outperforms the current state-of-the-art performance on several clinically impactful classification tasks. The gains are especially significant for small datasets.

We expect that this method will have an important impact in the field of computational pathology in two ways: (1) Our method specifically boosts performance for small datasets, which are very common in practice. We therefore address a major bottleneck in computational pathology. (2) Having light and discriminative WSI representations would alleviate the use of the image modality for a larger community of researchers in cancer bioinformatics, in order to investigate the complex relationships between genetic, transcriptomic and phenotypic data. Currently, WSIs are mostly used by computer vision experts. To facilitate reproducibility and the

broad use of Giga-SSL, the complete source code of this work will be available upon publication.

V.1.2 Background

V.1.2.1. Multiple instance learning for gigapixel images

In the MIL paradigm, objects (called bags) comprise other objects (called instances). For gigapixel images, the bag is a gigapixel image, and its instances are subimages (also called tiles or patches) extracted throughout the gigapixel image. While traditional MIL assumes independent and identically distributed (i.i.d.) instances within each bag (Ilse, Tomczak, and Welling 2018), this assumption is relaxed for gigapixel images because instances are extracted from the same image, and are therefore not independent. Given a gigapixel image X made of n_x instances $(x_1, \ldots, x_{n_x})$, MIL is implemented as a combination of three modules: (i) an instance embedder $e_{\theta_1}(\cdot)$, (ii) a pooling operator $p_{\theta_2}(\cdot)$ and (iii) a classifier $c_{\theta_3}(\cdot)$ such that a decision $\hat{y}$ is obtained with

$$\hat{y} = c_{\theta_3}\Big(p_{\theta_2}(\{e_{\theta_1}(x_1), \ldots, e_{\theta_1}(x_n)\})\Big).$$

Most MIL architectures differ in the design of the pooling operator p_{θ_2}. There are two families of operators: (i) those that consider instances as i.i.d. and (ii) those that exploit the relationship between instances of a bag. Architectures that consider instances as i.i.d. are either parameterless (using the operators average, maximum, a concatenation of both (Lerousseau et al. 2021), or a noisy-OR function (Srinivas 2013)), or trainable, such as an attention-based neural network (Ilse, Tomczak, and Welling 2018). While these architectures obtain good performances, instances of gigapixel images are dependent and contain information that can be leveraged to produce accurate predictions. Modern MIL architecture for gigapixel images have been designed to exploit the spatial relationship of instances. For instance, transformer-based MIL approaches (Shao et al. 2021) extend the attention mechanism of Ilse, Tomczak, and Welling (2018) by incorporating the positions of instances for decision prediction. Of particular interest in this work, the SparseConvMIL (Lerousseau et al. 2021) architecture leverages spatial information by building a sparse map from both the instance embeddings and their sampled locations. This map is further processed by a sparse-input convolutional neural network that outputs a latent vector to be further classified by a generic classifier.

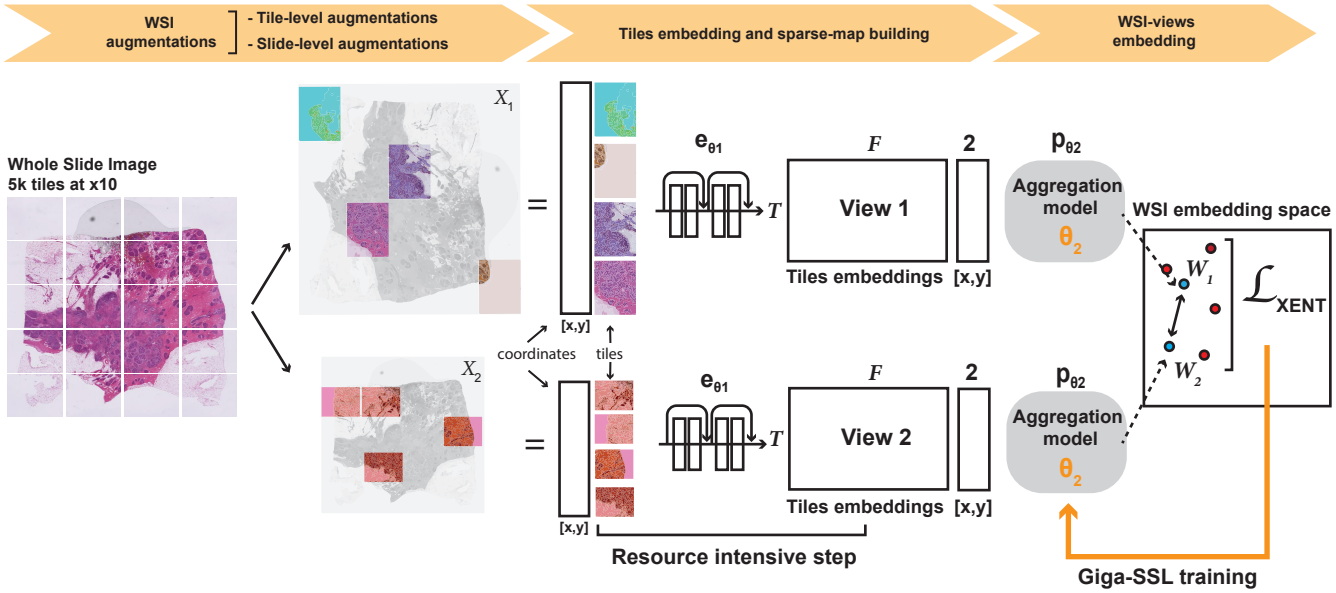


Figure V.1.: Overview of the Giga-SSL method. First, random augmentations of a WSI are used to create two different views X_1 and X_2 of the same WSI. Next, T tiles randomly extracted from each view are embedded using a tile-embedder network e_{θ_1}, resulting in T embeddings in $\mathbb{R}^F$. These embeddings and their associated tile coordinates are fed into a sparse-input CNN model p_{θ_2}, producing two WSI representations W_1 and W_2. A contrastive loss is applied on a minibatch of several whole slide images in order to update p_{θ_2}.

V.1.3 Self-supervised learning for gigapixel images

Self-supervised learning has been investigated in computational pathology at the tile level, for patches extracted from whole slide images (Ciga, Xu, and Martel 2021; Dehaene et al. 2020; Lazard et al. 2022; Saillard et al. 2021). The findings suggest that SSL indeed improved the performance on WSI classification tasks by using the SSL pre-trained tile level model as a frozen tile encoder. Because patches extracted from WSI are of size similar to datasets of natural images, the majority of the work successfully used off-the-shelf frameworks developed on natural images such as SimCLR (T. Chen, Kornblith, Norouzi, et al. 2020) or MoCo (He et al. 2020).

To the best of our knowledge, only one prior work has proposed a self-supervised learning framework for learning representations at the Giga-pixel scale (Richard J. Chen et al., n.d.). To do so, the authors design a new architecture made of 3 hierarchically stacked visual transformers (Dosovitskiy et al. 2020) which is trained on unlabelled WSI regions with the DINO framework (Caron et al. 2021), notably by enforcing consistency between two perturbed views of the same image. As stated by the authors (Richard J. Chen et al., n.d.), their approach cannot be trained end-to-end due to memory issues and needs to be trained in stages, starting from the visual transformer at higher magnification. A major bottleneck of this approach is the necessity to train the last transformer from scratch in order to perform downstream tasks at the WSI level, implying that (i) the whole system does not benefit fully of SSL pretraining, and that (ii) general and discriminative WSI representations are not directly available (Richard J. Chen et al., n.d.). Conversely, we designed an efficient method for learning WSI representations that obtained state-of-the-art performance with a linear classifier without the need to fine-tune any part of our system.

V.1.4 Methods

V.1.4.1. Algorithmic design

Notations and algorithmic background Giga-SSL training comprises 6 sequential steps to extract WSI representations which we detail here, and which is illustrated in Figure V.1. Lets us consider a WSI X. We introduce here an extension of the SparseConvMIL architecture for WSI classification (Lerousseau et al. 2021) by considering a ResNet network f_θ (ResNet18) (He et al. 2015b), which is cut at the beginning of the fourth residual block into two sequential parts:

1. the first part, acting as the tile embedder e_{θ_1}, is made of all layers of f_θ up to the first layer of the fourth block,
2. the second part, acting as the pooling function p_{θ_2}, is made of all layers after and including the fourth block of f_θ up to the fully connected layer. It has been converted into a submanifold convolutional network (Graham and van der Maaten 2017) such that it can process sparse data.

such that for any image i, the ResNet embedding is:

$$f_\theta(i) = p_{\theta_2}(e_{\theta_1}(i)) \in \mathbb{R}^{512}$$

Step 1: Augmentation of the WSI at the tile-level Two augmentation functions t_1 and t_2 are sampled from an image augmentation domain A made of color augmentations (color jitter, grayscale) and geometric augmentations (flips, rotations, scaling, blurring). First, T tiles are subsampled from X for each augmentation function t_1 and t_2, yielding two sets of patches $\{X_1\}$ and $\{X_2\}$. The coordinates of the top-left pixel of the tiles are stored for further processing. Finally, t_1 is applied to all patches of $\{X_1\}$, yielding a set of augmented patches denoted as $t_1(\{X_1\})$, and similarly a set $t_2(\{X_2\})$ for the second set of patches $\{X_2\}$.

Step 2: Embedding of tiles Each tile of both $t_1(\{X_1\})$ and $t_2(\{X_2\})$ are concurrently and independently forwarded through the tile embedder network e_{θ_1}. Each image is thus converted into a feature map which is averaged across all pixels, yielding a tile embedding of size F (256 for ResNet18) for each tile of $t_1(\{X_1\})$ and $t_2(\{X_2\})$

Step 3: Building of the sparse maps Following the framework of SparseConvMIL (Lerousseau et al. 2021), a sparse map S_1 is built by assigning each produced embedding of $t_1(\{X_1\})$ at the location where each of its original tiles was sampled in Step 1 3.1.0.2 but downsampled by a factor $d = 224$. Similarly, a sparse map S_2 is built from the embeddings $t_2(\{X_2\})$.

Step 4: Augmentations of the WSI at the slide-level While WSI are difficult to manipulate due to their huge size, a sparse map can be augmented with geometric transformations, enabling our framework to perform slide-level transformations in real-time. S_1 and S_2 are randomly flipped, rotated, and scaled with a factor uniformly sampled in $[0.5, 2]$ independently for the x and y axis.

Step 5: Embedding of the sparse maps into two augmented WSI representations To compute representations, we apply p_{θ_2} on both augmented sparse maps S_1 and S_2. At this stage, the two augmented views of the input WSI X (augmented at the tile-level and at the slide-level) are vector representations of the WSI.

Step 6: Loss optimization As is done in SimCLR, augmented views are finally fed to a projector, giving two augmented projections with which the loss will be computed. We train the weights of the pooling function p_{θ_2} by optimizing the contrastive loss

NT-XENT loss (T. Chen, Kornblith, Norouzi, et al. 2020). Given a minibatch B of augmented WSI $(X_1^i, X_2^i)_{i \in B}$, we set the loss function for a positive pair of WSI as

$$\ell_i = -\log \frac{\exp\left(\text{sim}\left(X_1^i, X_2^i\right)/\tau\right)}{\sum_{x \in B} \mathbf{1}_{\{x \neq X_1^i\}} \exp\left(\text{sim}\left(X_1^i, x\right)/\tau\right)}$$

where τ is the temperature parameter and $\mathbf{1}_{\{.\}}$ the indicator function. The final loss is computed as the average of these terms across all views.

V.1.4.2. Design choices

Selection of the underlying CNN architecture Giga-SSL does not theoretically rely on a ResNet architecture. There are many choices of good architectures that could be used for the tile encoder and pooling function, including two parts of different architectures. However, the pooling function must be implemented such that it can handle sparse data since it processes the augmented sparse maps (see Step 5 3.1.0.6).

Off-line augmentation strategy A key computational bottleneck of Giga-SSL training is the online computation of tile embeddings for a batch of B WSI, each composed of T tiles. GPU memory limitations put constraints on B and N_t, which effectively limits the number of total tiles per batch that can be used. Besides, it has been shown in SSL for natural images that a large batch size is required to yield representations with good downstream classification performances (T. Chen, Kornblith, Norouzi, et al. 2020; T. Chen, Luo, and Li 2021; Xinlei Chen and He 2020). A strategy for overcoming these issues is to freeze the tile encoder e_{θ_1} and pre-compute the embeddings of randomly sampled and augmented tiles for each WSI, essentially bypassing steps 1 and 2](#methods_step2){reference-type="ref" reference="methods_step2"}. For encoding a WSI, this is implemented by: (i) sampling 50 tile-level augmentation functions (both color and geometric augmentations) $(t_k)_{k \leqslant 50}$, (ii) for each k, randomly subsampling 256 tiles from the WSI and augment them with t_k, and (iii) concurrently and independently forwarding each augmented tile into e_{θ_1} and storing them. This process leads to N*50*256 tile embeddings where N is the total number of WSI of the Giga-SSL training dataset.

Giga-SSL is then trained, starting from step 3) by performing the following to sample a view of a WSI: (i) sample one of the 50 tile-level augmentations, (ii) sample a subset T of the 256 embeddings obtained from this augmentation, (iii) build the sparse map, and (iv) carry on from step 4 .

V.1.5 Experimental validation

V.1.5.1. Step 1: self-supervised pre-training

Self-supervised pre-training of Giga-SSL is done using The Cancer Genome Atlas (TCGA) (Weinstein et al. 2013), a public dataset that comprises 11754 whole slide images containing tissue from virtually all types of solid cancers. This dataset is the result of an international data-collecting effort and therefore features a high variety of participant centers (190). Such slides are crucial for patient care since they are the basis of diagnosis and treatment selection. On average, images have a width of 93000 pixels and a height 67500 pixels, for an average of 6.5 billion pixels per image. Fully compressed, TCGA weighs more than 16 Terabytes, orders of magnitude more than ImageNet (Deng et al. 2009). We tessellated non-overlapping square patches of size 256 pixels from all diagnostic slides of the TCGA at 10x magnification.

e_{θ_1} **pre-training** We choose to pre-train e_{θ_1} using MoCo (He et al. 2020). We trained a full ResNet18 on a subset of 6 millions of these tiles extracted from a random set of 3000 slides from the TCGA for 200 epochs. e_{θ_1} is then extracted from this network as described in 3.1.0.1.

Giga-SSL pretraining: we trained Giga-SSL on the full TCGA dataset, using the augmented embeddings extracted with the previously described pre-trained tile embedder (see 3.2.0.2), with Adam (Kingma and Ba 2014) for 1000 epochs.

V.1.5.2. Step 2: learning from linear embeddings

Training design For Giga-SSL, similarly to the works on natural images (Caron et al. 2021; T. Chen, Kornblith, Norouzi, et al. 2020; He et al. 2020), we measured the quality of the learned representations by performing linear probing either with all the labels available for a given task or by artificially reducing the number of labels to simulate a semi-supervised setting. To do so, one representation was extracted for each WSI after SSL pretraining. These representations were then used as input data to train a logistic regression for each considered downstream task.

Datasets This protocol was applied to six diagnostic WSI classification tasks highly pertinent for clinical practice:

- 3 tasks performed by Chen (Richard J. Chen et al., n.d.) aiming at automating the routine diagnosis of Non-Small Scell Lung Cancer (NSCLC), Breast Cancer (BRCA), and Kidney Cancer (RCC);
- 3 tasks aiming at inferring molecular properties from tissue slides towards faster, cheaper and more accessible molecular testing for cancer therapy selection.

For each of these 6 tasks, 1 reports the number of training WSI of the corresponding dataset, and their class distribution. All the datasets for these tasks are subsets of the TCGA (Weinstein et al. 2013). Results were computed on 10 bootstrapped splits of the data for each experiment, as was done in Chen (Richard J. Chen et al., n.d.), and we also used their train/test splits to ensure fairness of performance comparisons.

Task	# samples	# labels per class
BRCA subtyping	1041	831 - 210
Kidney subtyping	924	510 - 294 - 120
NSCLC subtyping	1033	528 - 505
BRCA Molecular	595	129 - 466
BRCA mHRD	912	447 - 465
BRCA tHRD	634	318 - 316

Table V.1.: Total number of samples and number of samples per class for all of the 6 benchmarked tasks in this paper.

Default settings The number T of tiles sampled per slide to 5. For a slide X, we bootstrap $R = 50$ views without tile augmentation (differing only in the sampled tiles), compute their embedding $\{W_r\}_{1,\dots,50}$ and consider the WSI representation as the elementwise average of the $\{W_r\}_{1,\dots,50}$. Average embeddings are normalized using a standard scale, while the Giga-SSL embeddings are normalized using the L2 unit.

V.1.5.3. Results

Classification results on benchmarked tasks Table V.2 synthesizes the results on all tasks for 5 models average, an attention-based MIL (Ilse, Tomczak, and Welling 2018) on top of a ResNet18 pretrained with MoCo , DeepSMILE (Schirris et al. 2021) and HIPT (Richard J. Chen et al., n.d.). Results from HIPT and DeepSMILE are taken from their respective articles and constitute the SoTA on the task on which they are cited.

Our proposed approach, Giga-SSL, outperforms the state-of-the-art on two out of three tasks benchmarked in (Richard J. Chen et al., n.d.) when using 100% of the available training labels NSCLC and BRCA subtyping. For BRCA subtyping, the AUC is increased by 3 points. Our proposed approach also achieves superior performances for all the other remaining tasks (mHRD, tHRD and BRCA molecular profiling). However, the power of the proposed approach seems to be in the low data regime. This is evident by the results obtained by using only 25% of the available labels. In this semi-supervised regime, the proposed approach obtained the best results on all tasks. While this finding may be expected when comparing Giga-SSL to methods without pretraining, Giga-SSL obtained superior results compared to the other SSL-based approach HIPT. For example, there is a gain of 6.9 AUC points for BRCA subtyping.

	Method	Giga-SSL (proposed)	AverageMIL	DeepMIL	HIPT	DeepSMILE
	Linear	✓	✓	✗	✗	✗
Task	% data					
$NSCLC_{subtyping}$	100	**0.952** ± 0.020	0.913 ± 0.023	0.948 ± 0.017	**0.952** ± 0.021	-
	25	**0.939** ± 0.017	0.885 ± 0.036	0.922 ± 0.034	0.923 ± 0.020	-
$BRCA_{subtyping}$	100	**0.905** ± 0.032	0.859 ± 0.038	0.874 ± 0.050	0.874 ± 0.060	-
	25	**0.890** ± 0.058	0.822 ± 0.072	0.860 ± 0.042	0.821 ± 0.069	-
$RCC_{subtyping}$	100	0.982 ± 0.007	0.973 ± 0.011	**0.986** ± 0.008	0.980 ± 0.013	-
	25	**0.975** ± 0.012	0.959 ± 0.015	0.970 ± 0.016	0.974 ± 0.012	-
$BRCA_{molecular}$	100	**0.938** ± 0.035	0.920 ± 0.037	0.924 ± 0.042	-	-
	25	**0.853** ± 0.075	0.799 ± 0.068	0.810 ± 0.093	-	-
BRCA mHRD	100	**0.756** ± 0.028	0.706 ± 0.030	0.736 ± 0.047	-	0.727 ± 0.010
	25	**0.743** ± 0.039	0.643 ± 0.050	0.660 ± 0.046	-	-
BRCA tHRD	100	**0.855** ± 0.023	0.799 ± 0.034	0.836 ± 0.052	-	0.838 ± 0.012
	25	**0.781** ± 0.050	0.698 ± 0.078	0.721 ± 0.075	-	-

Table V.2.: Benchmark study reporting the 10-fold cross-validated AUC performances of a logistic regression trained with Giga-SSL WSI representations or AverageMIL WSI representations, and retrained from scratch for other benchmarked approaches. For each task, we evaluate the methods with two data budgets with either 100% or 25% of the available training data.

Compared to attention-based MIL and HIPT, the proposed approach (Giga-SSL) provides an overall gain in performance while working in a linear regime. This is in contrast to HIPT and attention-based methods, which require fine-tuning and learning from scratch, respectively. Consequently, the downstream training pipeline for Giga-SSL is extremely efficient in comparison to the other two approaches. For instance, training for BRCA subtyping with 100% of the training data on 10 bootstrapped splits took 1.25 CPU-seconds for the proposed approach versus 150 GPU-minutes for attention-based MIL. This is a difference of 7200 times in favor of Giga-SSL – while also obtaining superior performances.

Tiny datasets In practice, pathological datasets can be tiny for the prediction of treatment response. For instance, phase II clinical trials typically involve 50 patients. Training a model to identify responding and non-responding patients is therefore challenging due to the low number of available labels.

We measured the performance of Giga-SSL in such a context by artificially reducing the size of all 6 datasets to 250, 100 and 50 samples. We compare Giga-SSL to the DeepAttnMIL model, which performances are on par with all other benchmarked algorithms (see Table V.2).

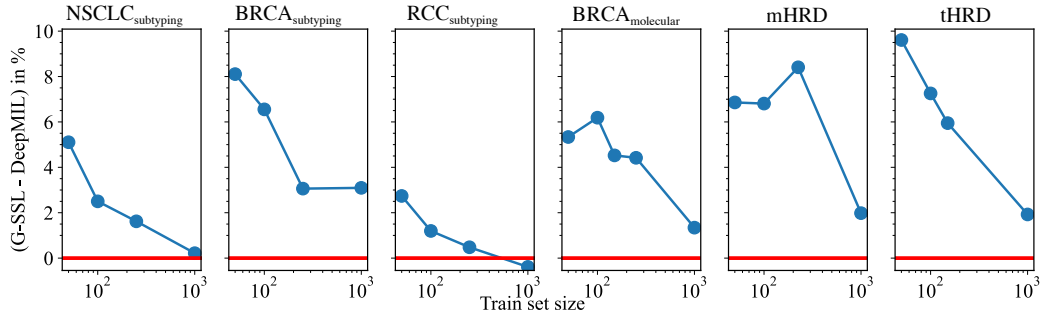


Figure V.2.: Difference between the average AUC performances of Giga-SSL and DeepMIL (in %) as a function of the training set size. The red line represents equal performance. Above the red line, the advantage is given to Giga-SSL.

Figure V.2 shows that the performance gap between the proposed approach and the standard WSI classification method strengthens as the number of samples decreases. The average improvement over all tasks brought by Giga-SSL features is of 5.1 AUC points when using 100 WSI and up to 6.3 AUC points when using only 50 WSI.

V.1.6 Ablation study and sensitivity analyses

In this section, we aim to understand the impact of some of Giga-SSL design choices over the predictive power of the learned representations. All subsequent experiments were conducted with the same conditions (including hyperparameters, epochs, and training dataset) as in the previous experiments, unless otherwise stated.

V.1.6.1. Sharing tile augmentations within views improves performance

Table V.3 reports the performance of Giga-SSL when removing one component at a time, (i) with a tile embedder pre-trained on ImageNet rather than pre-trained with MoCo on histopathological data (Giga-SSL$_{\text{im}}$), (ii) without slide-level augmentation during the WSI-level SSL pretraining; (iii) without shared augmentations across all tiles of a view, each tile is transformed by a randomly and independently sampled augmentation.

	100% data			50 WSI		
	NSCLC	CRC	BRCA	NSCLC	CRC	BRCA
Giga-SSL	0.952	0.982	0.905	0.894	0.960	0.793
w/o slide-aug	0.935	0.973	0.894	0.86	0.951	0.80
NS	0.933	0.971	0.875	0.847	0.939	0.774
Giga-SSL$_{\text{im}}$	0.922	0.978	0.888	0.813	0.952	0.751
Giga-SSL$_{\text{im}}$ NS	0.897	0.975	0.853	0.777	0.935	0.707

Table V.3.: 10-fold cross-validated AUC performances of ablated Giga-SSL models. w/o slide-aug is a Giga-SSL model trained without slide-level augmentations. NS (Not Shared) is a Giga-SSL model trained without sharing the tile-level augmentation among views. Giga-SSL$_{\text{im}}$ stands for a Giga-SSL model trained with tiles embeddings transferred from an ImageNet pretraining.

Using a tile-level SSL algorithm to pretrain the tile encoder e_{θ_1} brings improvement to the WSI-level representations: the Giga-SSL trained with MoCo features outperforms its ImageNet (Giga-SSL$_{\text{im}}$) counterpart on all tasks. On the contrary, the slide-level augmentation does not seem to be extremely important for the SSL task, as removing it has a small to no impact on performances.

However, applying independent transformations to each tile (*not shared*) degrades substantially the performances with an average decrease of 1.9 AUC points using 100% of the data down to 2.8 AUC points when using only 50 WSI, over the classification tasks. When ablating the shared transformations from a Giga-SSL model trained with tile features pretrained with ImageNet, the drop of performances compared to a Giga-SSL$_{\text{im}}$ is even more important: 2.1 AUC points with 100% of the data, 3.2 AUC points with 50 WSI.

Using shared augmentation thus allows the learning of useful features in abundant and scarce data regimes. We hypothesize key features linked to the slide preparation and shared by all the tiles on the slide are still available for shortcut learning if the tile-level augmentations are not shared. It seems that these shortcut features may be more prevalent with ImageNet than with MoCo tile representations. Highlighting such features and finding even more stringent ways to suppress them when learning Giga-SSL should further improve its performance.

The fewer tiles, the better Figure V.3.A presents the performances of 4 Giga-SSL models trained with different numbers of sampled tiles per view. The fewer tiles we sample, the better the resulting WSI representations. This behaviour strengthens when the downstream problem has a smaller training set and is comparable among

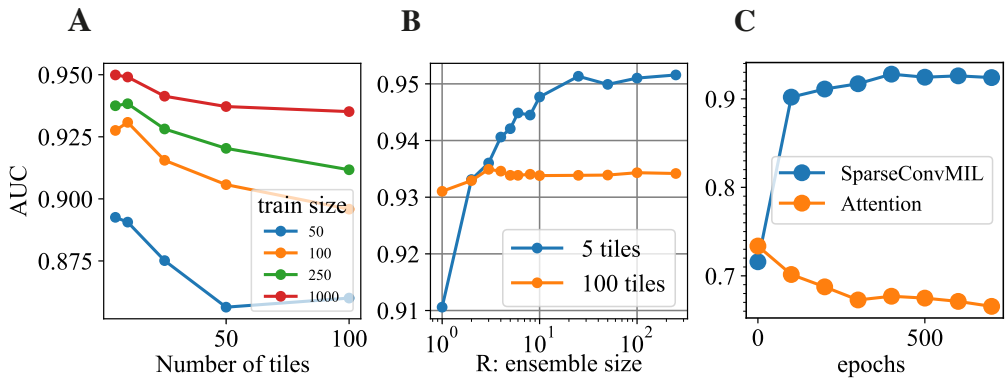


Figure V.3.: Experiments on key parameters of Giga-SSL. Each point is a 10-fold cross-validated AUC performance of a logistic regression fed with Giga-SSL features. The classification task is NSCLC subtyping for the three experiments. **A.** Effect of the number of sampled tiles T per WSI during training. **B.** Effect of the number R of bootstrapped non-augmented views of WSI to feed Giga-SSL at inference time, using a model trained with either 5 or 100 tiles per WSI. **C.** Evolution of the performances of a Giga-SSL with a SparseConvMIL (blue line, normal situation) or an attention-MIL network (orange line) as an aggregator.

all the downstream classification tasks. Interestingly, we can observe the opposite effect when using a DeepMIL model to classify a WSI: the fewer tiles used at training time, the worse the performances (Lerousseau et al. 2021). A very small number T of sampled tiles per view when training Giga-SSL can be seen as an aggressive augmentation. It has been reported (T. Chen, Kornblith, Norouzi, et al. 2020) that SSL benefits from stronger augmentations more than classification tasks, and Tian et al. (2020) have shown that there is an optimal strength of augmentation for each downstream task. This optimum results from a trade-off between keeping enough information to solve the downstream task and minimizing irrelevant features.

As sampling 5 tiles per WSI is enough to learn useful information to solve all the proposed downstream tasks, we can deduce that the signal relative to these problems is distributed among most of the tiles of the WSI. It would be interesting to test the performances of Giga-SSL on a classification task for which we know that the signal is highly concentrated on a few instances.

Ensembling representations brings improvement A constraint of the Giga-SSL model with a SparseConvMIL aggregation module is that it must use the same number of tiles per WSI at inference and training. We therefore decided to bootstrap R views of a WSI at inference time before averaging the Giga-SSL embeddings of these R views. Figure V.3.B investigates the effect of R on the downstream performances of the Giga-SSL representations. If the training uses 100 tiles per view, ensembling WSI representations does not improve their discriminative power. However, when training uses 5 tiles per view, it helps a lot (+4 AUC points on NSCLC subtyping). This performance gain saturates around $R = 50$. Two conditions are therefore required for an efficient training:

data regime	100% data	50 WSI
Full dataset	0.952 ± 0.020	0.894 ± 0.045
Independent training set	0.948 ± 0.017	0.885 ± 0.045

Table V.4.: Linear classification performances (AUC) on NSCLC subtyping of embeddings trained on either the full TCGA or a subset of the TCGA independent from the downstream task dataset.

- A small number of sampled tiles per view at training time, which makes the contrastive task difficult
- The ensembling of WSI views at inference, which helps integrating information from discriminative but incomplete views.

Generalization Giga-SSL has been trained on the full TCGA dataset, and downstream classification dataset also comes from the TCGA. In order to investigate the extent to which Giga-SSL could transfer to other datasets, we extracted from the TCGA all slides coming from the 41 centers that contributed to the NSCLC dataset, leading to an independent set of 6840 WSI. We trained Giga-SSL for 1000 epochs on this training set and reports the results in Table V.4.

Interestingly, Giga-SSL performs almost as good when trained on a set of WSI totally independent from the downstream task set. This suggests that Giga-SSL would generalize well on a different dataset.

Attention-deep-MIL unlearns when trained with SSL Instead of using a sparse-CNN as a tiles features aggregator, one could choose any other MIL model. We trained a Giga-SSL model with a DeepMIL aggregation module and evaluated its downstream linear performances on the NSCLC dataset. Figure V.3.C shows that the performances of such a model decrease while the SSL training is in progress. Although the DeepMIL shows very good classification performances (see Table V.2) when trained from scratch, this architecture seems not suitable for Giga-SSL pretraining. We suspect that the DeepMIL architecture has too easily access to shortcuts features to learn the WSI identity. Understanding what causes its collapse may highlight key pitfall for Giga-SSL training and therefore allow to improve it.

V.1.7 Conclusion

Limitations While Giga-SSL has been shown to generalize well outside of its training data distribution, the tile-embedder is not pre-trained on a dataset that is entirely independent from the downstream tasks' datasets. It would be interesting to conduct the same experiment as Table V.4 but excluding the WSI from the tile-embedder pre-training dataset too. In addition, a drawback of working with frozen embeddings of WSI is that it removes any possibility of building explainable models.

Finally, we have explored self-supervised learning for whole slide images with a versatile design based on specific data augmentation tailored for the multiple instance learning framework. Our proposed approach achieved or beat state-of-the-art performance over a wide range of clinically impactful tasks in both high and low data regimes. In particular, for small datasets (slides), our approach achieved a performance improvement of 6.3 AUC points on average compared to competing methods. Ablation studies and sensitivity analyses highlighted the key components of our approach – including tile encoder pretraining and how to apply augmentations to tiles – to better understand the pitfalls of self-supervised whole slide image representation learning.

This chapter has been made in collaboration with M. Lerousseau, A.V. Salomon, M-H. Stern, M. Rodrigues, S. Gardrat, E. Decencière and T. Walter. It is currently under review.

V.2 Democratizing computational pathology: optimized WSI representations for TCGA

V.2.1 Introduction

Cancer diagnosis heavily relies on the examination of H&E-stained tissue slides, which offer crucial insights into the disease and potential treatment options and which are routinely acquired in pathology labs. Digitizing these slides into Whole Slide Images (WSI) enables automated analysis, aiming at assisting clinicians in executing tedious tasks, such as counting mitoses(Veta et al. 2015), identification of metastases(Ehteshami Bejnordi et al. 2017) and grading(Lubrano Di Scandalea et al. 2022). Furthermore, the availability of large data repositories, such as The Cancer Genome Atlas (TCGA) provides us with the challenging opportunity to identify morphological biomarkers related to survival (Courtiol et al. 2019b; L'Imperio et al. 2023) or treatment response(Naylor et al. 2022), and to unravel the complex genotype-phenotype relationships by building predictive models for molecular features, such as single gene mutations(Jakob Nikolas Kather et al. 2020), and mutational signatures(Lazard et al. 2022).

However, WSI are not yet extensively used outside the pathology community for two primary reasons. First, the size and complexity of WSI require special skills and equipment for their analysis. A single WSI may contain billions of pixels, complicating storage, processing, and analysis. Second, while pathology labs are generating ever increasing WSI datasets, annotated WSI datasets are often scarce, in particular for rare diseases, specific molecular subtypes or in the context of clinical trials. Training current deep learning models on such datasets often leads to underperforming models with poor generalization capability.

Self-supervised learning (SSL) offers a promising approach for addressing these chalenges. This training paradigm leverages unlabeled datasets to pretrain neural networks which then demonstrate exceptional performance when fine-tuned on smaller, annotated datasets. Numerous studies in the computational pathology field have already adopted SSL, but only at the tile level (Dehaene et al. 2020; Lazard et al. 2022; Lubrano Di Scandalea et al. 2022; Saillard et al. 2021; Schirris et al. 2021).

They used such pretrained networks to encode the small images that compose the WSI-the tiles-, effectively reducing them from billions of pixels to a few thousand feature vectors. These feature vectors then serve as the basis for training multiple instance learning (MIL) algorithms (Ilse, Tomczak, and Welling 2018; B. Li, Li, and Eliceiri 2021; Lu et al. 2021; Shao et al. 2021). Nonetheless, the sheer volume of tiles per WSIs still makes MIL models both computationally intensive to

train and prone to overfitting. While (Richard J. Chen et al., n.d.) is a fair effort toward training without supervision wide histopathological images -up to 4096 pixel squared-, they do not succeed in training WSI representations.

Here, we introduce Giga-SSL, a novel self-supervised method that utilizes large, unlabeled WSI datasets to learn compact and highly discriminative WSI-level features. It can encode a WSI into a single vector of 512 values, and we show that a simple logistic regression operating on these representations achieves equal or better performance than fully-supervised MIL architectures across several tasks and datasets. Furthermore, we open-source these representations for the entire TCGA-formalin-fixed, paraffin-embedded (FFPE), reducing its size from 12 TB to 23 MB without loss of predictive power.

Giga-SSL aims to use the principle of contrastive learning (CL) at the slide level, i.e., on an array of tile representations. CL is a SSL framework whose primary task is to draw closer the representations of two randomly transformed versions of the same object, while pushing away-*contrasting*- the representations of different objects.

In order to optimize this objective at the scale of WSIs, we devised a specialized approach that involves both tile-scale and slide-scale transformations. This design is executed through a two-step architecture, as illustrated in Figure V.4A. The architecture consists of two distinct neural networks; the first network, which is pretrained on histopathological images, is responsible for encoding the transformed tiles, and only the second network, comprising sparse convolutional layers, undergoes optimization during the giga-SSL training process.

The output of this second block, the WSI representations, are used in all the downstream analysis tasks by training L2-regularized logistic regressions. For the sake of brevity, we refer to these models as *Giga-SSL classification models*. Their corresponding performance metrics are designated as *Giga-SSL performances*.

V.2.2 Results

Giga-SSL outperforms fully-supervised methods on several classification tasks and across cancer types. We first compared Giga-SSL models to state-of-the-art WSI classification algorithms across five TCGA benchmark tasks. These include breast cancer subtyping (lobular/ductal), lung cancer subtyping (lung adenocarcinoma (LUAD)/ lung squamous cell carcinoma (LUSC)), kidney subtyping (clear/papillary/chromophobe cells), and two breast cancer-related tasks: homologous recombination deficiency / proficiency (HRD/HRP) and molecular profiling (Triple Negative Breast Carcinoma (TNBC)/luminal). We gradually reduced the training dataset size through stratified subsampling down to 50 WSIs and assessed the performance of models trained on these subsets (see Fig. V.4B). We compared Giga-SSL to the CLAM-SB algorithm (Lu et al. 2021) operating on the same tile representations than the one used by the Giga-SSL model. Fig. V.5 **A.** shows the absolute performance of Giga-SSL and CLAM models and their difference as a function of the training set size. Across all tasks, using 100% of the training data, Giga-SSL consistently achieves

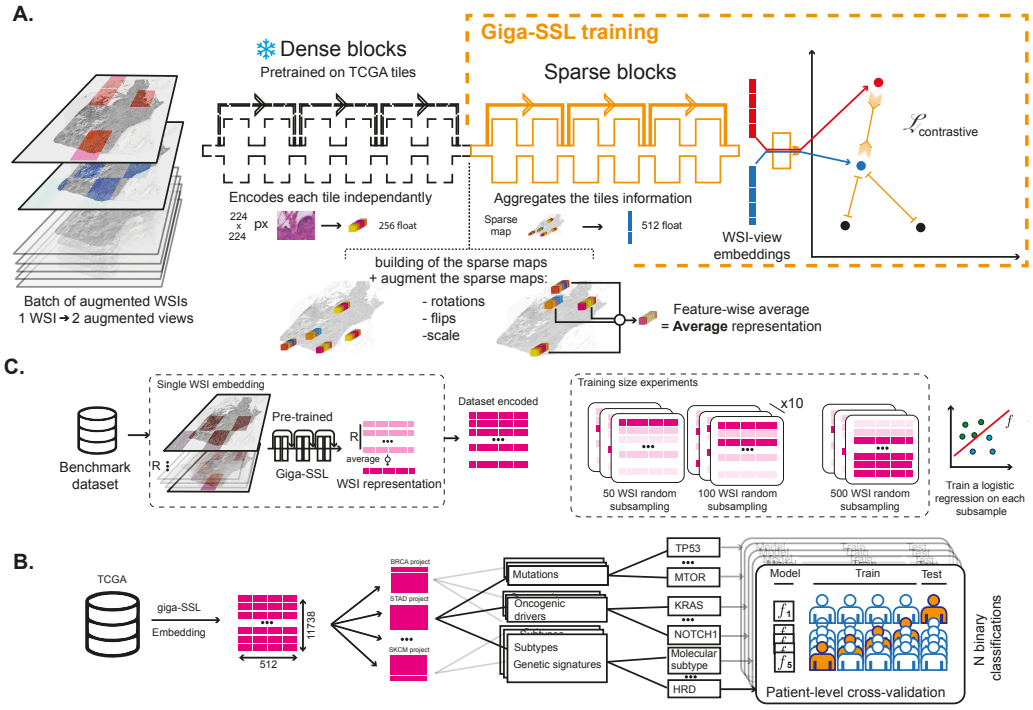


Figure V.4.: **A.** Overview of the Giga-SSL architecture and training procedure: Initially, two distinct views of the same Whole Slide Image (WSI) are created using tile and slide level augmentations. Each augmented tile is encoded using a pre-trained convolutional tile-encoder. The objective of Giga-SSL training is to optimize the second sparse convolutional block to minimize a contrastive loss at the slide level. **B.** Benchmark tasks experiments workflow, detailing the computation of a WSI representation. **C.** Workflow of the pan-cancer classifications experiment.

state-of-the-art performance. The advantage of Giga-SSL over CLAM grows as the training set size shrinks. With only 50 WSIs, Giga-SSL provides an average gain of 7 AUC points over CLAM.

Giga-SSL transfers well to other datasets.

To confirm these findings, we performed external validation experiments to explore the generalization capabilities of our representations. For this we applied Giga-SSL to two in-house WSI datasets for two different cancer types:

- Breast cancer (BC): 788 in-house H&E stained WSIs from BC patients with known Homologous Recombination s
- atus. We conducted two classification tasks: HRD prediction and subtype prediction (TNBC/luminal) (Lazard et al. 2022).
- Uveal Melanoma (UM): 516 in-house H&E stained WSIs from UM patients. The objective here was to predict chromosome 3 status (disomy 3 or monosomy 3, which represent two major UM subtypes with contrasted prognosis) (Cassoux et al. 2014).

Fig. V.5.B shows the improved performance of logistic regression with Giga-SSL representations compared to CLAM. The average AUC increase is 3.8% using 100% of the data and 5.7% using 50 WSIs: Giga-SSL weights maintain their performance and label efficiency when applied to unseen datasets.

Giga-SSL can predict more mutations and genetic signatures than MIL.

We next turned to the prediction of mutations and genetic signatures across cancer types. Kather et al. showed in a seminal study that many mutations and genetic signatures are predictable from H&E stained tissue slides (Jakob Nikolas Kather et al. 2020). Our objective was to assess whether logistic regression could effectively predict mutations and signatures using Giga-SSL representations, as an alternative to employing a full MIL. The data workflow for these experiments is delineated in Fig. V.4D. In total, the prediction tasks encompass (1) 830 point mutation predictions across 14 tumor types (2) 376 known oncogenic driver mutations and (3) 182 subtyping and genetic signature tasks. Of note, genetic variant prediction tasks are usually imbalanced, with the minority class comprising 8% of the dataset on average. This contrasts with subtyping and genetic signature prediction tasks, where the minority class represents on average 33% of the dataset.

The Venn diagram in Fig. V.5.D illustrates the number of mutations that can be predicted by the Giga-SSL model (represented by the blue areas), the Average model -see methods- (shown as the red area), and the MIL model (Jakob Nikolas Kather et al. 2020) (depicted by the green area).

We observe that Giga-SSL is able to predict most of the mutations that are predictable with the previously published method (30/45 for the point mutations, 26/34 for the oncogenic driver mutations). Furthermore, the Giga-SSL representation allow the prediction of an additional 64 point mutations and 30 driver mutations, which roughly doubles the number of predictable mutations from WSIs across these 14

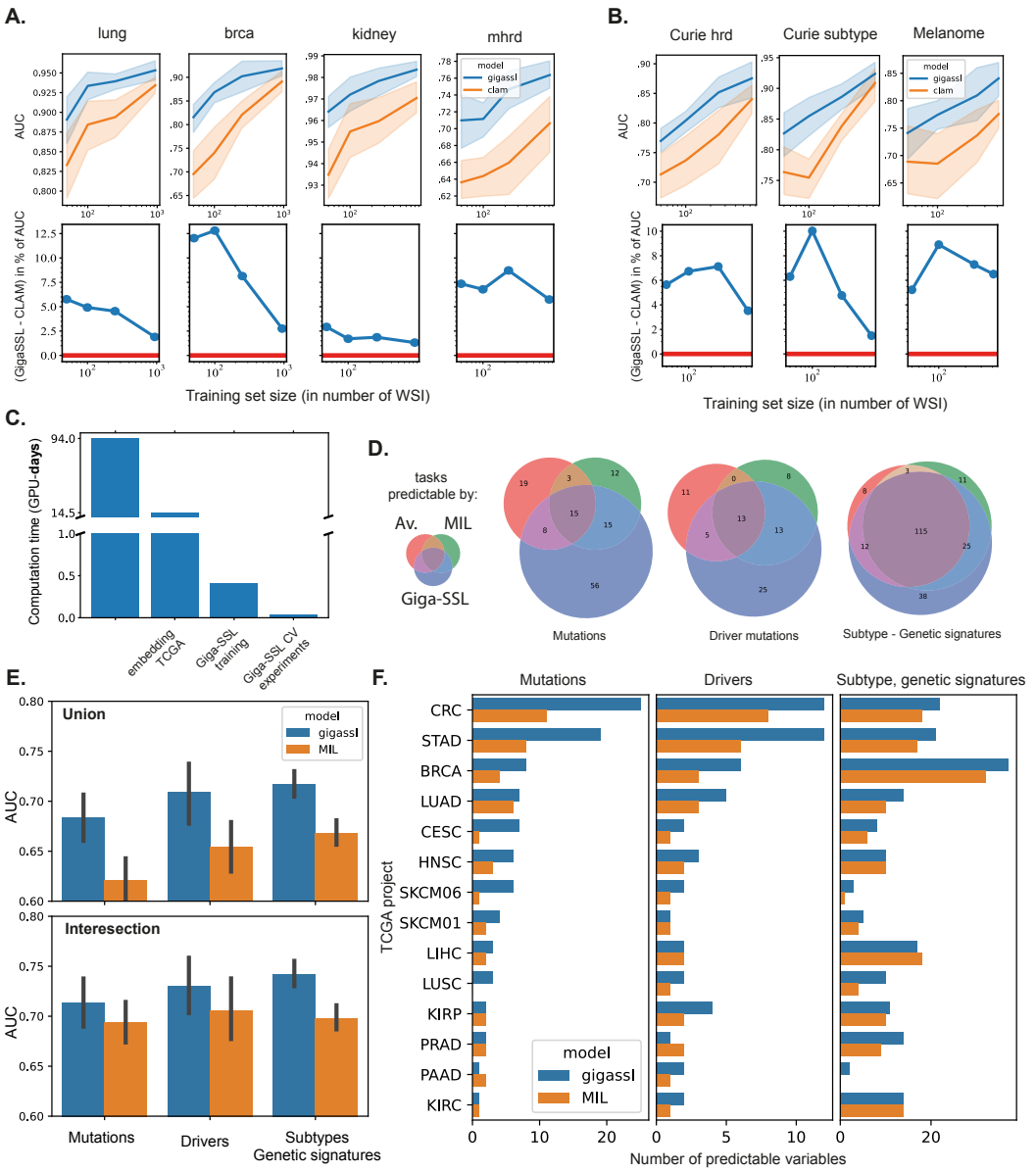


Figure V.5.: **Evaluation Results.** **A.** Displays the results for the four benchmark tasks. The first row of graphs presents the ROC-AUC scores of Giga-SSL and CLAM as a function of the training set size. The second row demonstrates the AUC improvement of Giga-SSL over CLAM as a function of the training set size. **B.** Presents the results on the external validation datasets for both Giga-SSL and CLAM. **C.** Indicates the time requirements for key computation steps, measured in GPU-days. **D.** The number of predictable tasks for each model and each task type (mutations, driver mutations, subtypes, and genetic signatures) in a pancancer setting. **E.** Displays the average cross-validated AUC for the three different types of classification tasks- mutation, oncogenic drivers, subtypes and genetic signatures-for Giga-SSL and MIL models. In the upper panel, the results are averaged across all tasks that are predictable by the Giga-SSL **or** MIL model (union). The lower one shows averages across the tasks predictable by both models. **F.** Provides a detailed breakdown of item D., focusing on the granularity of the TCGA projects.

cancer types. Details about the predictable mutations are available in Supplementary Tables F.3 and F.6.

A similar result is obtained for genetic signatures: out of 374 binary classification tasks (see methods) the majority of signatures predictable by the MIL model (Jakob Nikolas Kather et al. 2020) can also be predicted by Giga-SSL (140 out of 154) and the average model (127 out of 154). Compared to the MIL model, the Giga-SSL model can predict an additional 50 signatures. Additionally, as illustrated in Fig. V.5E, Giga-SSL provides enhanced classification accuracy across all tasks. Details about the performances of Giga-SSL on these tasks are available in Supplementary Figures F.2 and F.4

Finally, the same experiments have also been led by training Giga-SSL-based logistic regressions on site-corrected folds (Howard et al. 2021) leading to similar conclusions (see Figure F.1).

Giga-SSL is modular. Contrary to monolithic frameworks and algorithms, Giga-SSL relies on the correct adjustment of different training elements: the tile-encoder pre-training, tile-level and slide-level augmentations, and the aggregation module. Each module can be updated independently, allowing the entire framework to continuously benefit from advancements in their respective domains. We illustrate this by comparing various pre-trained tile-encoders as shown in Supplementary Table F.1, notably the recent ctranspath network, and demonstrate that the WSI Giga-SSL representation benefits from an improved tile-encoder; a research effort that already receives much attention (Filiot et al. 2023; X. Wang et al. 2022; Xiang and Zhang 2022).

Giga-SSL is computationally efficient. To highlight Giga-SSL's computational efficiency, Fig. V.5.C shows the GPU-days required for our experiments. Training a Giga-SSL model for 1000 epochs on the complete TCGA dataset takes only 10 GPU-hours on a single GPU. Once we have obtained the representations, downstream prediction tasks become highly efficient. This is because the classification algorithm fits only logistic regression, and the cross-validated experiments take approximately 1 hour on a laptop CPU — a sharp contrast to the 92 GPU-days reported in (Jakob Nikolas Kather et al. 2020). When applied to external datasets, Giga-SSL showcases good scalability: on average, a WSI can be encoded by Giga-SSL in roughly 5 seconds using a single GPU. This implies that encoding a typical dataset of 1000 WSIs can be completed in less than 1.5 hours.

V.2.3 Discussion

In this article, we aimed to develop generic representations for H&E stained WSI, targeting robust solutions for small datasets and simplified training, thus lowering barriers for morpho-molecular cancer analyses. Addressing these challenges, we introduce Giga-SSL, the first SSL framework able to derive concise yet highly discriminative WSI-level embeddings. Using logistic regression, we highlight the advantages of these features, achieving state-of-the-art classification performances

on small datasets, exhibiting robust generalization to external datasets and doubling the number of predictable gene mutations in the TCGA across cancer types.

A major advantage of Giga-SSL is its computational efficiency, at training and inference time. This efficiency allows for the potential training of Giga-SSL models on datasets significantly larger than the TCGA at minimal cost and with much lower environmental impact. Moreover, even scientists without detailed knowledge of image analysis and deep learning can easily utilize this modality, facilitating tests on new outcome variables or experiments with different datasets stratifications. We are releasing the entire encoded TCGA-FFPE dataset to the public, reducing its size from 12 TB to 23 MB. Such readily available embeddings can be integrated into TCGA's bioinformatics evaluations without any need for prior image-processing know-how or specialized equipment. We are optimistic that this initiative will spark interest within the bioinformatics sector, encouraging comprehensive integration of pathology and molecular data, and fostering joint exploration of cancer's molecular and morphological landscape.

V.2.4 Methods

V.2.4.1. Training and inference details

Tile embeddings We obtain tile embeddings using contrastive learning. Specifically, we employ MoCo (Xinlei Chen et al. 2020), training a ResNet18 on 6 million tiles extracted from a random set of 3000 FFPE slides from the TCGA over 200 epochs using MoCoV2's standard transformation. The tile embeddings are obtained through spatial pooling of the activations of the third block of this network. The tile representations are kept static, meaning they are not further optimized during the Giga-SSL training phase.

Giga-SSL architecture The architecture of Giga-SSL combines the ResNet18 framework with SparseConvMIL (Lerousseau et al. 2021). The first four residual convolution blocks independently encode each tile. The resulting tile encodings are aggregated within a SparseMap and processed by 4 sparse convolution blocks (Lerousseau et al. 2021). Together, these components achieve functionality akin to a full ResNet18, as detailed in (Lazard et al. 2023), and the final layer of this architectural blocks are the WSI embeddings used in the downstream analysis. A final multi-layer perceptron called *projector* further encodes the embeddings of the WSI, following (T. Chen, Kornblith, Norouzi, et al. 2020). The CL loss is computed using the output of the projector network.

Slide-level transformations aim to generate different views that are to be pulled closer or pushed apart, depending on whether they originate from the same WSI, while maintaining part of the biological information of the WSI. To generate these views, we use the following transformations:

- Subsampling: Tiles are randomly sampled among all the tissue-tiles of a WSI. The harshness of the transformation is given by the number T of tiles subsampled per view. Notably, when $T = 1$, the Giga-SSL training framework becomes equivalent to HiDisc-Slide (Jiang et al. 2023).
- Tile transformations: The tiles are randomly augmented with classical image augmentations (hue, rotation, Gaussian blur, flips, crops) before being encoded by the tile encoder. Training Giga-SSL requires this transformation to be *shared* among views; that is, when building a transformed view of a WSI, the same augmentation must be applied to all sampled tiles.
- Sparse-map transformations: The sparse-maps undergo scale, rotation and flips transformation, augmenting the geometries of the WSIs.

Giga-SSL Training Training is performed in 2 steps. First, the tile encoder is pre-trained on histopathology data using MoCo (Xinlei Chen et al. 2020) and frozen, like in (Lazard et al. 2023). This allows to compare Giga-SSL to competing methods based on the same tile encodings. In a second step, the sparse units (in orange in Figure V.4) are trained on the full TCGA with the WSI-level contrastive pretext task under the slide-level transformations before. Training is performed for 100 GPU-hours on a single V100 GPU. We use the Adam optimizer with a starting learning rate of 0.003. We use a cosine annealing learning rate scheduler with 10 warming epochs and a final learning rate to $3e^{-6}$. As described in the methodological publication (Lazard et al. 2023), we approximate the tile-level augmentations by randomly sampling 25 augmentations per WSI. We then uniformly sample 64 tiles per WSI per augmentation, augment and embed them using the tile-encoder described previously.

WSI representation computation To regularize the embeddings, the Giga-SSL network is applied to $R = 50$ different views of each slide, each view being composed of 5 tiles. The Giga-SSL representations are the average of these 50 runs. The Average representations are the feature-wise average of the representations of all the tiles of a WSI. We call *Average models* the logistic regressions operating on these representations. Both the Average and Giga-SSL representation are then unit-normalized using scikit-learn (Pedregosa et al., n.d.) Normalizer.

Label acquisition Labels for various datasets are extracted from different sources:

- Lung (LUAD/LUSC): GDC Portal
- Breast Cancer (Ductal/Lobular): GDC Portal
- Kidney (clear/papillary/chromophobe cells): GDC Portal
- mhrd Breast Cancer (HRP/HRD): GerkeLab Repository, as in Knijnenburg et al. (2018). HRD score are binarized using mean as threshold.
- All pancancer experiments: Supplementaries of Jakob Nikolas Kather et al. (2020). The continuous variables are binarized using mean as threshold.

V.2.4.2. Statistical procedure

To address downstream classification challenges, we deploy an L2-regularized logistic regression (parameters: C=10, class_weight='balanced'). We report the performance of these logistic regressions over 10 random dataset splits for benchmarking and generalization experiments, and 3 splits for mutation prediction experiments (to account for the very imbalanced nature of the mutation prediction task). The same splits are used in compared methods.

For mutation prediction, the primary criterion is task predictability. As outlined in (Jakob Nikolas Kather et al. 2020), we accumulate the model's posterior probabilities predictions throughout the cross-validation folds. We then conduct a t-test between the predictions of the samples belonging to one class and the predictions from the remainder of the dataset. We adjusted the resulting p-values for multiple testing, accounting for a total of 1388 classification tasks, using the Benjamini-Hochberg correction. We set the significance threshold, $p$thres, to 0.01 and compare the predictability of tasks between Giga-SSL and the MIL model implemented in the original publication. Each task of the pan-cancer experiments employs a patient-level three-fold cross validation strategy.

Acknowledgements This work was performed using HPC resources from GENCI-IDRIS (Grant 2023-AD011011863R3). The results in this article are based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. TL was supported by a Q-Life PhD fellowship (Q-life ANR-17-CONV-0005). Furthermore, this work was supported by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir'' program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and by the ITMO Cancer (20CM107-00).

V.3 Interpretation: Morphological Profiles

One of the primary limitations of using Giga-SSL representations lies in the lack of a straightforward interpretation algorithm. While MIL models allow for the investigation of attention scores to identify ROIs and generate explanations, interpreting frozen Giga-SSL representations can be challenging. To address this, we propose a method for interpreting the Giga-SSL embedding space, drawing inspiration from the TCAV approach (see Section II.3.1.1). This work has not yet been submitted in the form an article, and I describe the method therefore in more detail. I would like to mention that the method makes use of individual cell classifications that have been kindly provided by Marvin Lerousseau, Postdoc at the CBIO.

V.3.1 Method Description

Our methodology is inspired by TCAV (B. Kim et al. 2018), where we employ user-defined *concepts* to construct morphological profiles for a specific binary task. Unlike local explanations that focus on individual WSI, our method offers a global explanation for the classification task at hand.

In TCAV, specific directions in a model's latent space are attributed to a predefined concept. We adopt a similar strategy by defining *concepts* and associating them with specific directions in the Giga-SSL embedding space. These *concepts* are human interpretable morphological features, either continuous or categorical, such as the number of tumor-infiltrating lymphocytes (TILs), tumor size, or cell atypia. The representation of each *concept* is built from a WSI dataset annotated with relevant, interpretable labels.

As an example, for a specific subset of WSIs, one might assess the area covered by necrotic tissue. This measurement serves as an interpretable concept.

A concept c then acts as a target for a linear model trained on Giga-SSL embeddings. Ridge regression is used for continuous labels, while L2-regularized logistic regression is employed for binary labels. These models yield weights $\mathbf{W}_c$, defining directions in the latent space on which the concept varies. Figure V.6**A.** illustrates that whether the concept is continuous or binary, moving along $\mathbf{W}_c$ either increases the value of c (if c is continuous) or approaches its decision boundary (if c is discrete).

We assemble a set of p concepts $(c_i)_{i\leqslant p}$ and their corresponding directional vectors $(\mathbf{W}_{c_i})_{i\leqslant p}$. For a binary classification task T, we train a logistic regression model and obtain weights $\mathbf{W}_T$. The cosine similarity between $\mathbf{W}_T$ and each $\mathbf{W}_{c_i}$ is calculated as:

$$P_i^T = \frac{\mathbf{W}_T \cdot \mathbf{W}_{c_i}}{\|\mathbf{W}_T\|\|\mathbf{W}_{c_i}\|}$$

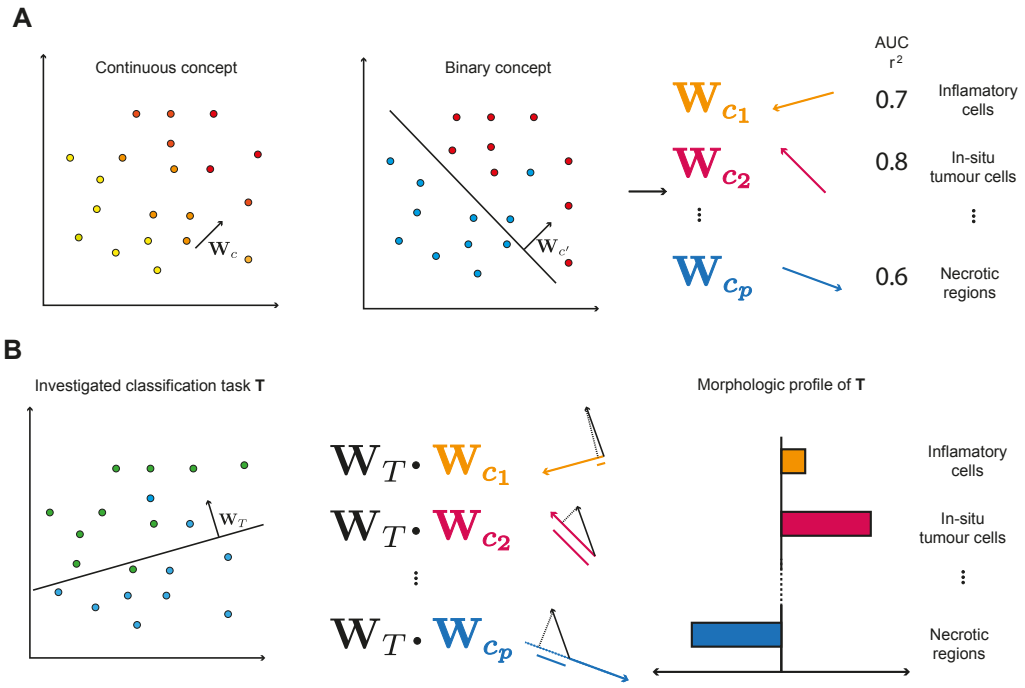


Figure V.6.: **Method for Constructing Morphological Profiles**. **A.** For each predefined concept c_i, a linear model is trained. The model's weights serve as the preferential direction $\mathbf{W}_{c_i}$ for concept c_i. **B.** Subsequently, the target classification task T is addressed using logistic regression. The cosine similarity between the task-specific logistic regression's weights $\mathbf{W}_T$ and the preferential direction of concept c_i $\mathbf{W}_{c_i}$ is interpreted as the contribution of c_i to the decision-making process for T.

This similarity, P_i^T, signifies the contribution of concept c_i to the decision process in task T. Iterating across all concepts, we build a **morphological profile**:

$$\mathbf{P}^T = (P_i^T)_{i \leqslant p}$$

V.3.2 Applications

V.3.2.1. Concepts

To demonstrate the method's effectiveness, we utilized an in-house, unpublished detection and segmentation tool. This tool quantified various features on 981 slides from the TCGA-BRCA dataset, such as the number of invasive tumor cells and the area of necrotic regions. These quantities serve as concept labels c_i:

- glands: number of healthy glands in the WSI.
- connective: number of cells composing connecting tissues.
- tissue: area of tissue free of detected nucleus.
- epithelial: number of epithelial cells.
- necrosis: area of necrotic tissue.
- inflammatory: number of inflammatory cells.
- TAStroma: Tumor Associated Stroma
- in-situ tumor: number of in-situ tumor cells.
- invasive-tumor: number of invasive tumor cells.

In our example, all features were continuous, although binary variables can also be used. We work in the feature space of the Giga-SSL model used in Section V.2.

V.3.2.2. Morphological profiles

We then used these concepts to devise the morphological profiles of a serie of tasks.

Transfer to Other Datasets During the work presented in Chapter III, an expert pathologist, Guillaume Bataillon, evaluated the number of Tumor Infiltrating Lymphocytes (TILs) on each slide from our Curie breast cancer WSI dataset. We binarized these counts using the average as a threshold and computed the morphological profile for this binary classification task.

Figure V.7 shows the morphological profiles for three different tasks on this Curie dataset. Apart from TILs, tumor grading and a molecular classification task were also examined. For tumor grading, grade 3 was used as the binarization threshold.

Both the Triple Negative and High-Grade profiles are consistent with literature findings, showing spikes in inflammation and necrosis (Rakha et al. 2009; Livasy et al. 2006). Importantly, the TILs profile strongly aligns with the inflammation concept. This finding validates the cross-dataset applicability of our interpretation

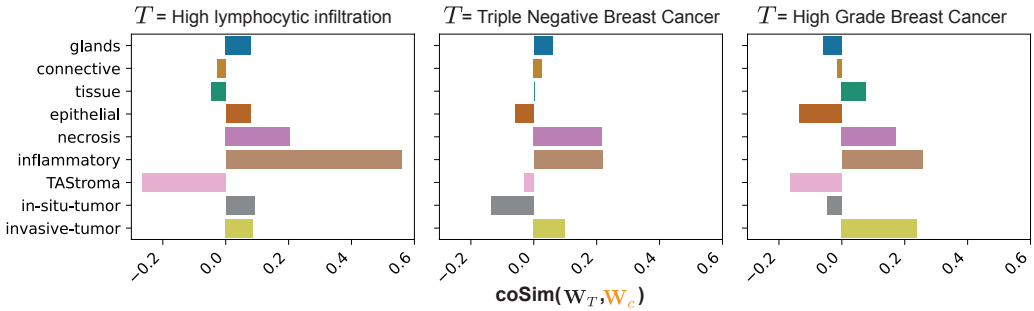


Figure V.7.: Morphological profiles of three classification tasks. Classification is performed on the Curie dataset. Glands referes to the number of healthy glands, tissue to the area of tissue without cells and TAStroma to the Tumor Associated Stromal cells.

method: it is possible to train from concepts on one dataset and use them to interpret a classification on another dataset.

Phenotypic study at the TCGA scale The computational efficiency of this global explanation method enables scalability, allowing us to extend our analysis to large datasets like TCGA. Utilizing a consistent set of concepts, described in Section V.3.2.1, we computed morphological profiles for all predictable binary classification tasks $\mathcal{T}$ that we tackled in Section V.2.

These tasks contain binary mutation predictions as well as genetic signatures and subtyping tasks, for 14 different TCGA projects. It means that a given classification task may be present in several occurrences. For instance, the TGF $-\beta$ response was found to be predictable among almost all of the TCGA projects.

To capture the relationships among these tasks, we applied KMeans clustering to their profiles $(\mathbf{P}^T)_{T\in\mathcal{T}}$, setting the number of clusters (K) to 4.

Figure V.8 **A** shows the average profile for each of the four clusters. Wordclouds facilitate the visual representation of the cluster composition. A noteworthy observation is the similarity in morphological profiles for certain classification tasks across different types of cancer. Specifically, the tasks related to the immune signatures used in Thorsson et al. (2018) for constructing immune subtypes seem to be predictable thanks to the same morphological profile among many cancer types.

For example, tasks dominated by TGF $-\beta$ signals are mainly grouped in cluster 4. Similarly, tasks featuring IFN $-\gamma$ are predominantly found in cluster 2, while tasks related to Wound Healing appear in cluster 1.

Figure V.8 further breaks down the average profiles for specific tasks across multiple cancer types.

These are overall coherent findings.

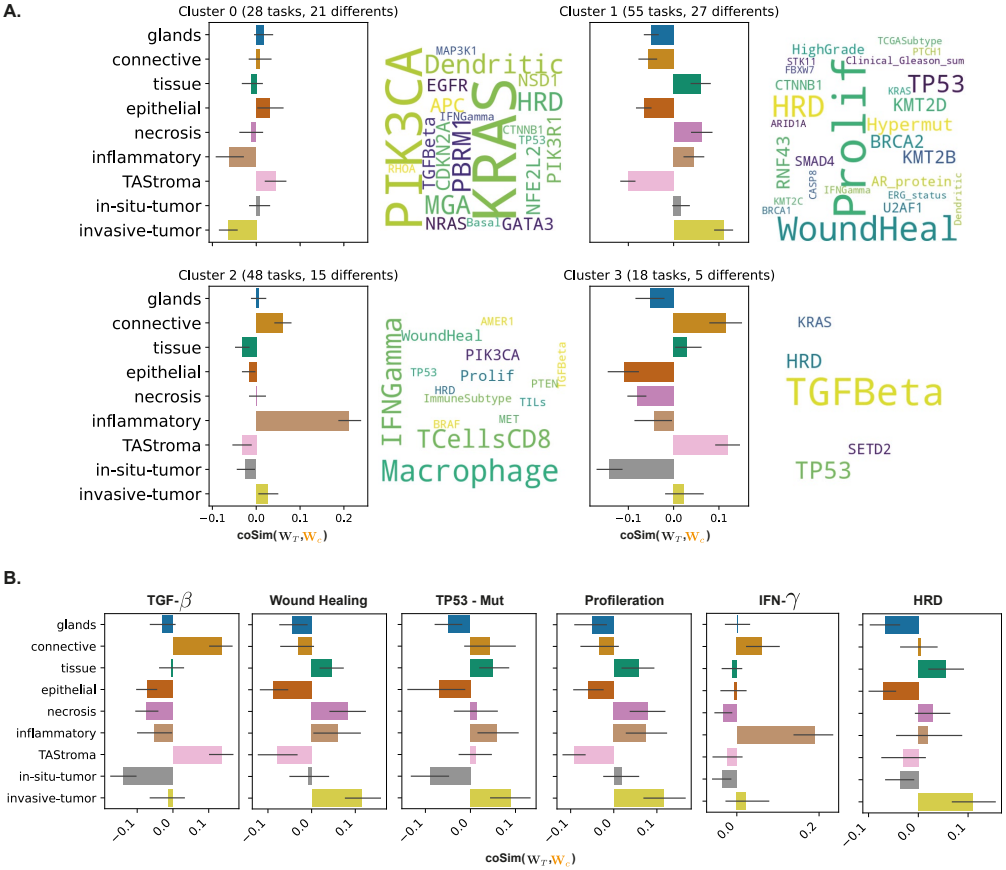


Figure V.8.: **Average profiles of the 4 clusters. A.** The average profile are accompany with wordcloud describing the composition of the clusters: the bigger the name of the task is, the more prevalent the task is in the cluster, meaning that this task shares a common profile across TCGA project. **B.** Average profiles of 6 tasks accross the 14 TCGA projects: 5 genetic signatures (TGF-β, Wound-healing, Proliferation, IFN-γ and HRD) and one point mutation, TP53.

V.3.3 Limitations and Perspectives

While the interpretation method presented here offers valuable insights, caution is essential in its application. The risk of confirmation bias is always present, making it imperative to validate any findings through alternative means before giving them credits. This method should ideally be employed alongside other interpretability techniques for robust analysis.

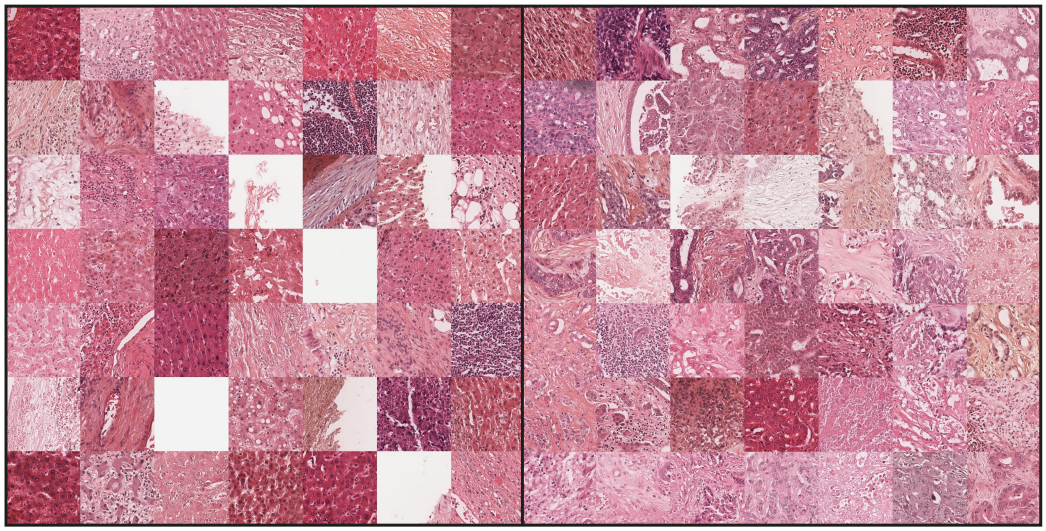
In addition, the morphological profiles generated are subject to noise and their reliability correlates with the predictability of the task and the individual concepts involved. Nevertheless, this approach serves as a cost-effective way to formulate hypotheses about phenotypes and transfer knowledge across datasets.

The method's scope is restricted by the set of *concepts* initially defined. However, it can be augmented by integrating it with other interpretability tools, such as those outlined in Chapter III. These complementary tools can identify morphological patterns without prior knowledge. Once identified, these new patterns could be used to train human observers, who could then select WSIs where these patterns are prevalent. The creation of a new concept dataset could then be used to validate initial findings and extend their application to broader datasets, such as different types of cancer.

In summary, this interpretability method complements the GigaSSL model effectively. Its ease of use, speed, and adaptability make it a useful addition to the WSI interpretation toolbox. It significantly broadens the scope for investigating cancer phenotypes at the scale of large datasets like TCGA, and across different cancer types.

VI

Predicting transcriptomic classes on whole slides images in intrahepatic cholangiocarcinoma



Contents

Preface

This research concludes my PhD thesis and was executed in close partnership with Aurélie Beaufrère, an anatomopathologist with expertise in cholangiocarcinoma. Intrigued by the potential therapeutic benefits of novel molecular subtyping, Aurélie saw an opportunity to apply a cost-effective deep-learning technique. The primary objective was to predict these molecular subtypes using raw Hematoxylin, Eosin and Saffron (HES, see Figure I.2) stained Whole Slide Images (WSIs), for which we had morphological clues.

I opted to utilize the recently developed Giga-SSL encoder, seizing the chance to evaluate its representations on both an external cohort and a tissue stain different from H&E. The ease and speed of classification using these embeddings further enabled a more in-depth exploration of the dataset. Notably, the comprehensive dataset assembled by Aurélie provided additional opportunities for investigation.

It included multiple types of samples, such as biopsies and surgical resections, and multiple samples per patient. Furthermore, these samples originated from different tumor blocks, as illustrated in Figure I.4. Importantly, not all samples were derived from the blocks used for molecular analysis. This provided us with a unique opportunity to investigate the influence of WSI origin, especially its correlation with the molecular sample, on classification performance.

In addition to demonstrating the feasibility of this molecular subtyping using only HES-stained WSIs, this study also highlights the impact of tumor heterogeneity at the resection level. Most existing studies focus on tumor heterogeneity at the slide level, making our research distinctive. This work, just like Chapter III testifies the value of interdisciplinary collaboration within our laboratory.

Contributions

Publications - communications

- A. Beaufrère*, **T. Lazard***, et. al. Self-supervised learning for predicting transcriptomic classes on whole slides images in intrahepatic cholangiocarcinoma, *preprint*.

Summary:

Transcriptomic classification of intrahepatic cholangiocarcinoma (iCCA) has been recently improved from two classes to five classes, associated with pathological features, targetable genetic alterations and survival. Despite its prognostic and therapeutic value, the classification is not routinely performed due to technical limits such as insufficient material or the cost of molecular analyses. Our aim was to predict iCCA transcriptomic classes on whole-slide digital histological images (WSI) using a self-supervised learning (SSL) model. We trained logistic regressions on top of representations extracted with a self-supervised model, Giga-SSL, and show that each transcriptomic class is predictable with good performances (AUC: 0.55-0.81) in a cross-validation setting, particularly for the hepatic stem-like class (AUC=0.81). These models generalized well on two external datasets-the TCGA and an external French test set-. In addition, we studied the role of the training set composition and highlight the potential deleterious effect of tumoral heterogeneity on the trained models.

Résumé:

La classification transcriptomique du cholangiocarcinome intrahépatique (iCCA) a récemment évolué, passant de deux à cinq classes, associées à des caractéristiques pathologiques, des altérations génétiques ciblées, ainsi qu'à la survie globale. Malgré sa valeur pronostique et thérapeutique, cette classification n'est pas couramment réalisée en routine en raison de contraintes techniques, telles que l'insuffisance de matériel biologique prélevé ou le coût élevé des analyses moléculaires. Notre objectif est de prédire les classes transcriptomiques de l'iCCA à partir d'images histologiques numériques de lames entières (WSI) en utilisant un modèle d'apprentissage auto-supervisé (SSL). Nous avons entraîné des régressions logistiques sur les représentations extraites à l'aide d'un modèle de représentation de WSI entraîné sans supervision, appelé Giga-SSL. Nous avons démontré que chaque classe transcriptomique est prédictible avec de bonnes performances (AUC : 0.55-0.81) lors de la validation croisée, en particulier pour la classe de type tige hépatique (AUC=0.81). De plus, nous avons confirmé la validité de ces modèles en les appliquant avec succès à deux ensembles de données externes : le TCGA et un jeu de test externe provenant de l'hôpital Beaujon. En outre, nous avons mené une analyse approfondie du rôle de la composition de l'ensemble d'entraînement et avons mis en évidence l'effet potentiellement néfaste de l'hétérogénéité tumorale sur les modèles entraînés.

This chapter has been made in collaboration with A. Beaufrère (with whom I share first autorship) R. Nicolle, G. Lubuela, J. Augustin, M. Albuquerque, B. Pichon, C. Pignolet, V. Priori, N. Théou-Anton, M. Lesurtel, M. Bouattour, J. Cros, K. Mondet, J. Calderaro, T. Walter and V. Paradis. It is currently under review.

VI.1 Introduction

Intrahepatic cholangiocarcinoma (iCCA) is the second most common primary malignant liver tumour with an increasing incidence worldwide (Global Burden of Disease Liver Cancer Collaboration 2017; Rahnemai-Azar et al. 2017). Despite the absence of chronic liver diseases, the prognosis of patients with iCCA remains very poor. Currently, the only curative treatment for iCCA is surgery; however, only 20-40% of patients can benefit from it due to a diagnosis at an advanced stage of the disease with an overall five-year survival after surgical resection < 40% (Endo et al. 2008; Mavros et al. 2014). In unresectable patients, locoregional treatment in purely intrahepatic cases or systemic treatment (gemcitabine and cisplatin +/- durvalumab in first line) in metastatic cases are proposed with a median overall survival of respectively 15 and 12 months (Bridgewater et al. 2014; European Association for the Study of the Liver. Electronic address: easloffice@easloffice.eu and European Association for the Study of the Liver 2023; Oh et al. 2022).

Recent advances in the pathobiological and molecular understanding of iCCA have provided prognostic and theranostic factors for a better clinical management of patients with iCCA (European Association for the Study of the Liver. Electronic address: easloffice@easloffice.eu and European Association for the Study of the Liver 2023). From a transcriptomic point of view, for a long time, only two distinct groups of iCCA have been identified: including an inflammatory class (40% of cases) characterised by activation of inflammatory signalling pathways, and a proliferation class (60% of cases) characterised by activation of oncogenic signalling pathways and associated with a worse prognosis (Sia et al. 2013). Recently, this classification has been improved by passing from two groups into five groups. The inflammatory class has been divided into two sub-classes (inflammatory stroma and immune classical) while the proliferative class has been divided into three subclasses (Hepatic stem-like, tumour classical and desert-like). Interestingly, this classification has been associated with tumour microenvironment composition, genetic alterations and prognosis (Martin-Serrano et al. 2022). In particular, Hepatic stem-like group, which represents the most frequent transcriptomic group, has been associated with a better prognosis and with targetable genetic alterations such as IDH1-2 mutations and FGFR2 fusions, particularly relevant clinically, since specific inhibitors (e.g. Pemigatinib or Ivosidenib) have been approved by the FDA as second-line treatments for locally advanced or metastatic iCCA (Abou-Alfa et al. 2020; Moeini et al. 2016). However, this transcriptomic classification is not used in routine practise since it is currently based on sophisticated molecular biology techniques (expensive, accessible in expert centres) and requires histological samples rich in tumour cells.

Artificial intelligence (AI) models and in particular deep neural networks (Deep Learning) are rapidly emerging in the medical field, particularly in imaging Calderaro and Kather (2021). With the development of digital pathology and wide access to digitised whole slide images (WSI), AI approaches have shown first their performance for classification tasks, as for example the distinction of cholangiocarcinoma from secondary forms of liver adenocarcinoma (Albrecht et al. 2023). Interestingly, AI approaches have also shown their performance for identifying prognostic microscopic features and transcriptomic classification for example in liver pathology in hepatocellular carcinoma Cheng et al. (2022). Despite these successes, deep learning techniques require large datasets (over a thousand slides (Campanella, Hanna, Geneslaw, Miraflor, Silva, et al. 2019)) and are heavy computational machinery that limit in-depth studies on stratified datasets. We recently introduced Giga-SSL, a self-supervised learning (SSL) algorithm designed to generate generalist low-dimensional feature vectors of WSI, which offer both computational efficiency and label-efficiency. The aim of the present study was to predict iCCA transcriptomic classes on WSI using the Giga SSL model, with a specific focus on identifying the hepatic stem-like class.

VI.2 Methods

VI.2.1 Patient and samples

The workflow of the study is summarised in Figure VI.1. For the discovery set, we selected 246 formalin-fixed paraffin-embedded (FFPE) iCCA cases (109 surgical specimens and 137 biopsies) archived between 2000 and 2021 in the Pathology department of Beaujon Hospital (Clichy, France) representing 769 Hematein eosin saffron (HES) slides, divided into 5 folds at the patient level to perform cross-validation. All available slides for the surgical specimens (including preoperative biopsy when available, n=25) were selected for the study (median of WSI per case: 5 [1-12]). The slides were scanned at 20x magnification with an Aperio scanner (ScanScope AT Turbo).

For the validation sets, we selected 32 iCCA surgical FFPE samples (32 WSI corresponding to the one most representative slide for each case) from the Pathology department of Henri Mondor Hospital (Créteil, France) (French external validation set) and 29 iCCA cases (surgical FFPE samples, 29 WSI) from The Cancer Genome Atlas cholangiocarcinoma (TCGA-CHOL) public dataset. The selection criteria included 1) iCCA diagnosis after reviewing by an expert pathologist, 2) $\leqslant 1$ available WSI from FFPE material, 3) tissue material in sufficient quantity and quality for molecular analysis or molecular analysis already performed. Written consent was obtained from all patients as required by French legislation. This study was approved by the local ethics committee (IRB 00006477 N° CER-2022-168). The clinical and biological data recorded were age at surgery, sex, risk factors of iCCA, tumour size, number of tumours and overall survival (OS).

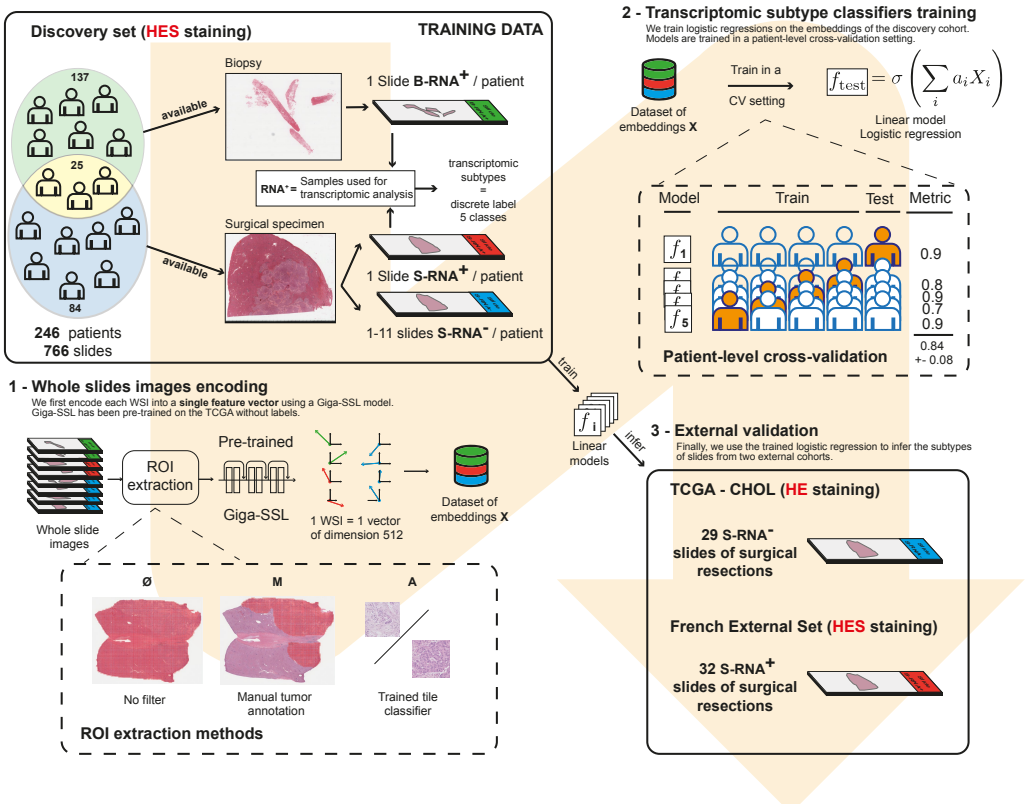


Figure VI.1.: Flow-chart of the study. The model was first trained for the prediction of the five transcriptomic classes in a discovery set of FFPE iCCA biopsy and surgical samples (n=246 patients, Beaujon Hospital, Clichy, France) in a 5-fold cross-validation scheme according to three different ROI extraction methods. Finally, it was validated in a French external validation set (n=32 patients, Henri Mondor Hospital, Créteil, France) and in a set of slides from TCGA (n=29 patients). *Formalin-fixed paraffin-embedded, FFPE; Hematein eosin, HE; Hematein eosin saffron, HES; cross-validation, CV; Region of interest, ROI; Slide from biopsy sample, slide B; Slide from surgical sample corresponding to the sample used for the transcriptomic analysis, slide S + ; Slide from surgical sample, not corresponding to the sample used for the transcriptomic analysis, slide S-; Self-supervised learning, SSL; The cancer genome atlas, TCGA.*

VI.2.2 Pathology reviewing

All histological slides were reviewed by an expert liver pathologist (AB) and the assessed features in the tumour were listed in Table G.1 and Fig. G.1. Stage of fibrosis in the non-tumoral liver when available was evaluated according to the METAVIR staging system (Bedossa and Poynard 1996).

VI.2.3 RNA sequencing

VI.2.3.1. RNA extraction

RNA sequencing was performed on the FFPE block selected for surgical specimens corresponding to the most representative slide in the discovery and the French external sets. These slides directly associated with transcriptomic analysis (consecutive slides), have been labelled as surgical slides S+ whereas slides from other blocks indirectly associated with transcriptomic analysis in the discovery set and in the TCGA set have been labelled S-. For biopsy, the FFPE block used for RNA sequencing corresponded directly to the slide selected (labelled as slide B) (Figure VI.1).

Briefly, five μm-thick sections with macrodissection when needed were cut from FFPE blocks. Total RNAs were further isolated using the Qiagen FFPE RNA extraction kit (RNeasy FFPE kit, Qiagen) for the discovery set and the Recover AllTM Total Nucleic Acid Isolation Kit for the French external validation cohort (Invitrogen, Thermo Fisher Scientific).

VI.2.4 Gene expression analysis

Gene expression was analysed using SMARTer Stranded Total RNA-Seq Kit for the discovery set and QuantSeq 3' mRNA-seq Kit for the French external validation set. Only genes quantified in at least 50% of samples were kept for the analysis. Gene expression profiles were quantile-normalized. The average expression of each gene set defined gene signature was computed following a gene-wise centering in each dataset (without variance scaling). The transcriptomic class with the highest gene-set averaged expression was assigned to each sample. The same process was applied to the TCGA dataset.

VI.2.4.1. Slide preprocessing and tessellation

Slides from the discovery set were stained with HES and encoded in svs format. Slides from the external French validation set were stained with HES and encoded in ndpi format. Slides from the TCGA validation set were stained with hematoxylin-eosin (HE) and encoded in svs format. Tissue regions automatically extracted using Otsu thresholding were then exhaustively split into 2899811 patches of 224×224 pixels (without overlapping) at 10x using the OpenSlide library in Python. We

present the results in the discovery set according to three different pre-processing protocols with or without extraction of region of interest (ROI), each requiring varying levels of expert pathologist involvement (Fig. VI.1):

- No-Filter (∅): All tiles including tumour and non-tumour are processed as they are, encompassing both tumour and non-tumour regions.
- Manual-Filter (M): An expert pathologist (AB) extensively annotates tumour regions using ImageScope software, from which patches are extracted.
- Learning-Filter (A): Tiles are filtered using a logistic regression trained on a dataset of 3000 tile embeddings, randomly extracted and labelled by an expert pathologist (AB).

A detailed illustration of these various ROI extractino methods is available in Figure G.3.

VI.2.5 Machine Learning algorithms

VI.2.5.1. Data-split

Training was done using a 5-fold cross-validation framework in the discovery set. Splits were stratified according to the output variable, at the patient level. They were shared among all training to ensure the fairness of comparison.

VI.2.5.2. Giga-SSL representations

The Giga-SSL model was trained on a single V100 GPU on the TCGA-FFPE dataset following the training framework provided in Lazard et al. (2023) at the exception of the following details: Training was performed for 100 hours, or 7800 epochs WSI embeddings are ensemble over 100 views, then L2-Normalized. Finally, we used L2-regularised logistic regressions (C=7, max_iter = 10000, and class_weight set as 'balanced') as end classification models.

VI.2.5.3. MIL baseline algorithms

Beside the giga-SSL based classifications, we provide some baseline classification algorithms for comparison. They are based on the deep attention multiple instance learning (MIL) algorithm introduced in the work of Ilse et al30 and slightly modified in Lazard et al (Lazard et al. 2022). The results are achieved using a ResNet18 tile encoder (He et al. 2015b) pre-trained on imagenet or on the TCGA (i.e the one used in the giga-SSL model).

VI.2.5.4. External dataset inference

The probabilities predicted by the 5 trained logistic regression on the training set (each corresponding to a training fold) are averaged, performances are computed using these pooled probabilities.

VI.2.5.5. Statistical analysis

Continuous variables were compared by the use of Student's t-test, and categorical variables were compared by use of chi-square or Fisher's exact tests. Survival curves were represented by using the Kaplan-Meier method compared with log-rank statistics. $p \leqslant 0.05$ was considered statistically significant (SPSS software). The performance of AI models were assessed thanks to area under the curve (AUC) score, balanced accuracy score and F1 score (macro-average).

VI.3 Results

VI.3.1 Patient characteristics

Clinical features	Discovery set N=246 (%)	French external validation set N= 32 (%)	TCGA validation set N= 29 (%)	p
Age (mean)	63 [27-88]	64 [25-85]	63 [29-82]	0.810
Sex (Male/Female)	141 (57) / 105 (43)	23 (72) / 9 (28)	13 (45) / 16 (55)	0.099
HBV	28 (11)	3(9)	NA	0.734
HCV	16 (7)	0 (0)	NA	0.231
MS	72 (30)	5 (16)	NA	0.141
Chronic alcohol intake	43 (18)	6 (19)	**0.035**	0.859
PSC	4 (2)	3 (9)	NA	
Other	11 (4)	1 (3)	NA	1.000
No risk factor	72 (29)	17 (53)	NA	**0.003**
Pathological features				
Cirrhosis (F4 Metavir stage)	24 (10)	4 (12)	NA	0.544
Multinodularity	67 (27)	10 (31)	NA	0.633
Size (mean, cm)	7 [1-22]	7 [1-16]	NA	0.604
Small duct type	214 (87)	22 (69)	27 (93)	**0.010**
Large duct type	18 (7)	7 (22)	1 (3)	**0.018**
Well differentiated tumour	82 (33)	15 (47)	12 (41)	0.253
Moderately differentiated tumour	135 (55)	14 (44)	14 (48)	0.426
Poorly differentiated tumour	29 (12)	3 (9)	3 (10)	1.000
Fibrosis (no or mild / moderate or intense)	49 (20) /197 (80)	4 (12) /28 (88)	6 (21) / 23 (79)	0.723
Immune infiltration (no or low / moderate or high)	166 (67) / 80 (33)	13 (41) / 19 (59)	18 (62) / 11 (38)	**0.011**
TLS	8 (3)	5 (16)	8 (28)	**<0.001**
Necrosis (median, %)	18 [0-90]	0 [0-60]	0 [0-30]	**<0.001**

Table VI.1.: Clinical and pathological features of the different datasets of the study. Data not available, NA; hepatitis virus B, HBV; Hepatitis virus C, HCV; Metabolic syndrome, MS; Primary sclerosing cholangitis, PSC; Tertiary lymphoid structures, TLS. In case of not available in the TCGA set, the statistical analyses were performed only between the discovery and the French external validation sets.

The main clinical and pathological features of the patients and tumours for each dataset are presented in Table VI.1. The three sets were similar for most clinical and pathological features in particular for the age (63 years for the discovery and the French external validation sets and 64 for the TCGA set, p=0.810) and the sex distribution (male sex, 57%, 72% and 45%, p=0.99). In the discovery and the French external validation sets, the main risk factors were chronic alcohol intake (18% and 19%, respectively, p=0.859) and metabolic syndrome (30 and 16%, p=0.141). At the pathology level, higher proportions of large duct tumours and intense immune tumour infiltration were observed in the French external validation set compared to the two other sets (22% vs 7% vs 3%, p=0.018; 59% vs 38% vs 33%, respectively). A higher proportion of tumours with TLS was observed in theTCGA set (28% vs 16% vs 3%, p<0.001).

VI.3.1.1. Transcriptomic classes

The proportion of each transcriptomic class in each dataset is represented in Figure VI.2. The most frequent transcriptomic class was the Hepatic stem-like class observed respectively in 37% of cases in the discovery set, 43% of cases in the French external validation set and 59% of cases in the TCGA validation set. Interestingly, in the discovery set, the repartition of the five transcriptomic classes was different between surgical samples and biopsy samples (Table G.2). The hepatic stem-like class was more represented in surgical samples compared to biopsy samples (49% vs 27%, p<0.001) whereas the immune classical class was more represented in biopsy samples (31% vs 14%, p<0.002).

As expected, transcriptomic classes in all cohorts were associated with some pathological features (Fig. VI.2B-C). An intense tumour fibrosis was mainly observed in the inflammatory stroma group (n=38/66, 58%). A high tumour immune infiltration was mainly observed in inflammatory stroma and immune classical groups (n=19/66, 29% and n=8/69, 16%, respectively). A low tumour immune infiltration was observed in hepatic stem-like and desert-like groups (n=67/120, 56% and n=7/15, 47%, respectively). No significant difference was observed between the five transcriptomic groups according to the tumour histological type (Figure G.2).

At the clinical level, three transcriptomic classes in all cohorts were significantly associated with OS. Hepatic stem-like group had an improved OS compared to other transcriptomic groups (OS median: 49 vs 35 months, HR 0.58; 95%CI 0.44-0.75; p<0.001) whereas tumour classical and inflammatory stroma groups presented an altered OS compared to other groups (OS median: 21 vs 43 months, HR 1.76; 95%CI 1.09-2.82; p=0.003 and OS median: 31 vs 43 months, HR 1.50; 95%CI 1.04-2.17; p=0.011, respectively) (figure Fig. VI.2D-F).

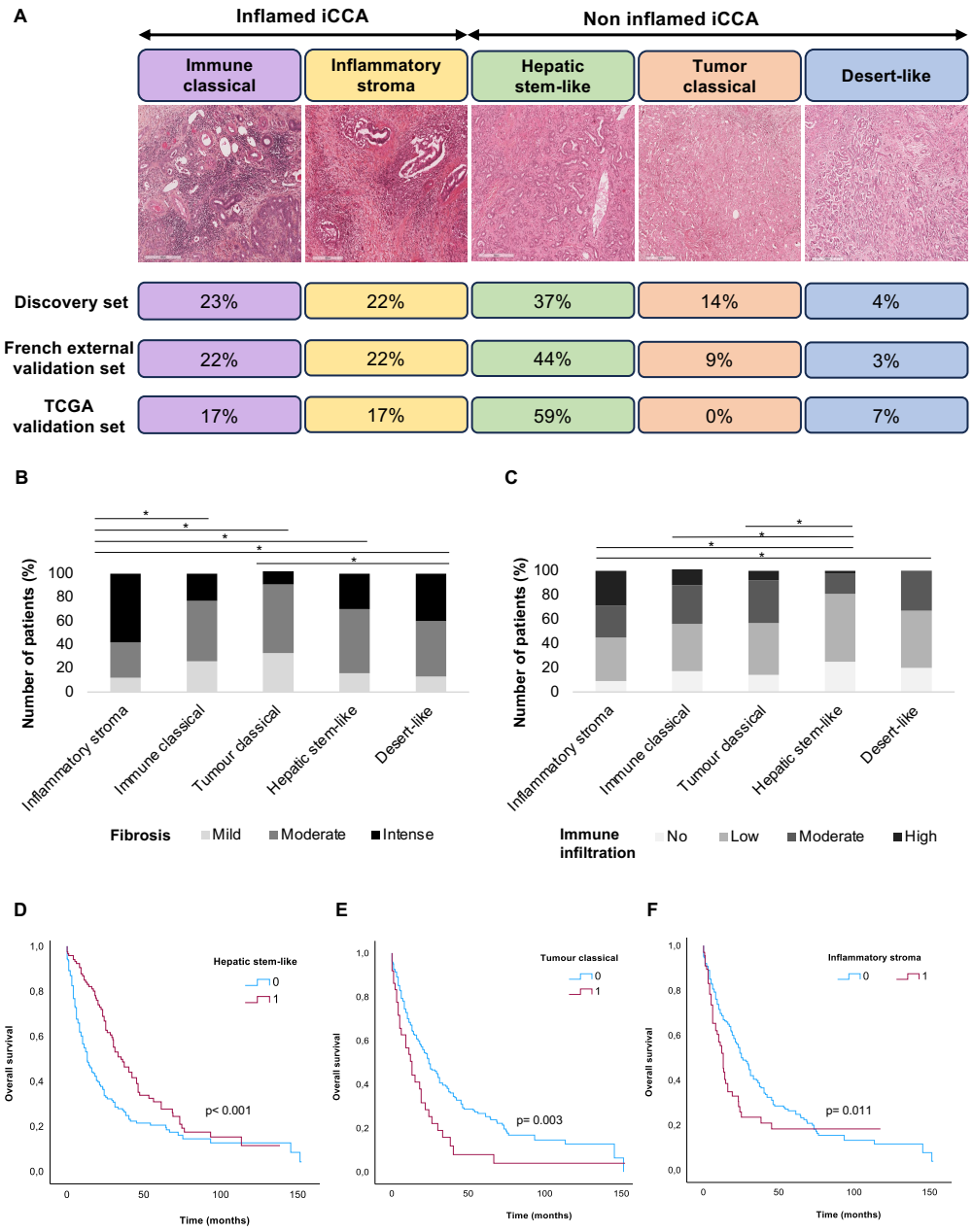


Figure VI.2.: Repartition and characterisation according to histological features and overall survival of the five transcriptomic classes. (A) Repartition of each transcriptomic class according to the different sets used in the study and representative histological images of each transcriptomic class (HES), (B) Semi quantitative assessment of the abundance of tumour fibrosis in each transcriptomic class (*p<0.005), (C) Semi quantitative assessment of the abundance of tumour immune infiltration in each transcriptomic class (*p<0.005), kaplan-Meier Curves for OS according to (D) Hepatic stem-like, (E) Tumour classical and (F) Inflammatory stroma transcriptomic class.

VI.3.2 Utilising self-supervised WSI representations for transcriptomic class prediction

We initially focused on the binary classification task of Hepatic stem-like, the most frequent class, before expanding our analysis to other transcriptomic classes.

VI.3.2.1. Prediction of the Hepatic stem-like class

Model	Tile Filter	AUC score	Balanced accuracy score	F1 score
Giga-SSL	A	0.82	0.74	0.75
	M	**0.84**	**0.76**	**0.76**
	∅	0.8	0.72	0.72
MoCo + MIL	A	**0.82**	**0.75**	**0.74**
	M	0.82	0.75	0.75
	∅	0.74	0.67	0.67

Table VI.2.: Cross-validated performances of both the Giga-SSL and MIL models on the discovery cohort for the Hepatic stem-like binary classification task according to three different pre-processing protocols : No-Filter (∅): All tiles are processed as they are, encompassing both tumour and non-tumour regions, manual-Filter (M): Tiles are extracted from pathologist annotations of tumour regions, learning-Filter (L): A small dataset of randomly extracted tiles is labelled as either tumour or non-tumour by a pathologist. These labels are then used to train a logistic regression model on the tile embeddings, which subsequently filters the tiles across all WSIs. Area under the curve, AUC; Learning-Filter, L; Manual-Filter, M; No-Filter, ∅; Multiple intance learning, MIL; Self-supervised learning, SSL

Table VI.2 showcases the cross-validated performances of both the Giga-SSL and MIL models on the discovery cohort for the Hepatic stem-like binary classification task. The Giga-SSL model peaks in performance when combined with a manual tumour annotation, achieving an average AUC of 0.84.

Indeed, performance improves when WSI are refined, regardless of whether this refinement is manual or learned. This improvement is particularly noticeable when using the classic MIL models, where the absence of WSI filtering leads to an 8-point drop in the AUC. For the Giga-SSL models, the absence of WSI filtering results in a 4-point decline in AUC.

VI.3.3 External validation of the model for Hepatic-stem like class prediction

Table VI.3 presents the results of the external validation of models trained on all the slides of the discovery cohort, with a manual filter applied to the Whole Slide Images (WSI). The logistic regression trained on the giga-ssl embeddings of the discovery cohort demonstrates strong transferability to both the French external (AUC=0.86) and TCGA sets (AUC=0.76). Notably, the TCGA cohort slides were stained with HE,

Validation dataset	AUC score	Balanced accuracy score	F1 score
TCGA set	0.76	0.73	0.74
French external set	0.86	0.73	0.71

Table VI.3.: Main external validation results: The models trained on the discovery cohort shows good generalisations when applied to the external cohort. Area under the curve, AUC; The cancer genome atlas, TCGA

which differs from the staining protocol used for the discovery cohort's slides (HES) which further emphasises the generalizability of the models.

VI.3.3.1. Influence of the pre-processing protocol

We detailed in Table G.3 the external validation results when models are trained with different ROI extraction methods. As observed in the cross-validated experiments, using an ROI extraction method is advantageous for both external datasets.

VI.3.3.2. Impact of the composition of the training set

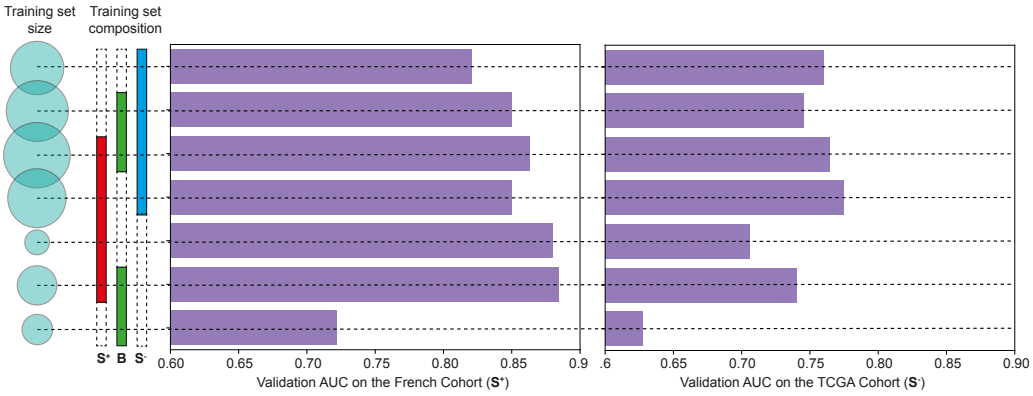


Figure VI.3.: Effects of the composition of the training set on the performances of the Giga-SSL model for the Hepatic stem-like binary classification task. We set the manual method for extracting the Region of Interest (ROI). Each row represents the performance of models trained on a specific subset of the discovery cohort. The characteristics of these subsets can be understood by looking at where the dotted line intersects with the items "training set size" (on the right) and "training set composition" (on the left). For example, the third row represents models trained on the complete training dataset (S+, B, and S-), which signifies it as the largest training set. *Area under the curve, AUC; Self supervised learning, Slide from biopsy sample, B; Self-supervised learning, SSL; Slide from surgical sample corresponding to the sample used for the transcriptomic analysis,S+; Slide from surgical sample, not corresponding to the sample used for the transcriptomic analysis, S-; Region of interest, ROI.*

In Figure VI.1, the discovery dataset is illustrated to contain various types of slides, including both surgical slides and biopsy slides. We can categorise these slides into two groups:

1. Slides on which transcriptomic analysis has been conducted, labelled as surgical slides S+ and biopsies B.
2. Slides that have not undergone transcriptomic analysis (S-).

Our objective was to determine how the composition of the training set influenced the generalisation performance of classification models. For this, we trained models using training sets with different compositions regarding S+, S- and B, and monitored the prediction performance on the two external validation sets. The results are shown in Figure VI.3.

For the French external validation set, which comprised solely S+ slides, incorporating S- slides into the training set appeared detrimental to performance. Remarkably, the highest performance was achieved when the training set was limited to slides that were directly associated to the transcriptomic analysis (S+ and B). Furthermore, training solely with S+ slides, despite them being the smallest possible training dataset, yielded a performance very close to the combined set (S+ and B). Moreover, we noted that the addition of biopsy slides B to surgical specimen slides slightly improved the validation performance.

Next, we turned to the validation on the TCGA dataset, which is exclusively composed of S- slides, i.e. slides for which the transcriptome was analysed on a different block. On this dataset with a putatively noisier ground truth, the model's performance seems to be closely related to the size of the training data rather than to its composition (Figure VI.3).

VI.3.4 Prediction of the four other transcriptomic classes

Transcriptomic classes	CV	French external set	TCGA set
Hepatic stem-like	**0.84+0.06**	0.86	0.76
Desert like	0.52+0.08	0.0†	0.85†
Tumour classical	0.77+0.09	0.88	na
Inflammatory stroma	0.72+0.10	**0.92**	**0.80**
Immune classical	0.63+0.08	0.62	0.78

Table VI.4.: Predictions for the five transcriptomic classes. This table presents cross-validated (CV) and generalisation outcomes on both the French external and TCGA sets for the classification tasks related to other subtypes. The TCGA cohort lacked any tumour classical samples. We trained the models using all slides (biopsies and surgical resections) and applied both automatic and manual filters on the training and validation sets. † is present when the metric is computed using less than two samples. Cross-validation, CV; Not applicable, na; The cancer genome atlas, TCGA.

We conducted analogous experiments for predicting other transcriptomic classes: Inflammatory Stroma, Desert-like, Tumour Classical, and Immune Classical. As with the Hepatic Stem-like class, we trained binary classifiers for each class using a patient-level 5-Fold CV setting and then applied them to the validation sets. We manually set the ROI method and trained using the complete dataset (S+, B, S-). Tab. VI.4 presents the results of these experiments. With the exception of the Desert-like class, all other classes can be predicted in a CV setting and demonstrate generalisation capabilities on the validation sets. The classification task for the Inflammatory Stroma class outperforms the others on the validation set, achieving an AUC of 0.92 for the French set and 0.80 for the TCGA set. Despite the absence of Tumour Classical slides in the TCGA dataset, this class still appears predictable and shows good generalisation with an AUC of 0.88 on the French set. On the other hand, while there is some discernible signal for the Immune Classical class, the models find it the most challenging to classify. As for the Desert-like class, no clear signal is observed in a CV setting. Performance on the French and TCGA external sets cannot be determined with reasonable confidence, as they contain respectively one and two cases

VI.4 Discussion

We show in this study that our SSL method applied to routine WSI has the ability to predict iCCA transcriptomic classes. Transcriptomic classes are particularly interesting in iCCA because of its association with the tumour microenvironment composition, the prognosis and its probable impact on the treatment response such as immunotherapy and targeted therapies (Martin-Serrano et al. 2022; Sia et al. 2013). As previously described, the most frequent transcriptomic group in our three datasets was the hepatic stem-like class (representing 37% to 59% of cases), which was associated with a better OS. Moreover, this class was interestingly described as associated with targeted molecular alterations in the study of Martin-Serrano et al. (Martin-Serrano et al. 2022) We confirmed also in our cohorts the association between the tumour microenvironment composition (inflammation and fibrosis) and the transcriptomic groups in particular for the two inflammatory groups, groups which may benefit from immune checkpoint inhibitor treatments (Martin-Serrano et al. 2022).

Our model showed good performance for predicting the transcriptomic groups in particular the hepatic stem-like, the tumour classical and the inflammatory stroma classes with AUC around 80% in the two external validation sets. The good predictions of these three transcriptomic classes are particularly interesting because they are all associated with OS in our cohorts and in most of the cohorts used in the Martin-Serrano study (Martin-Serrano et al. 2022).

Currently, the leading methods for WSI classification rely on MIL (Ilse, Tomczak, and Welling 2018; B. Li, Li, and Eliceiri 2021; Lu et al. 2020). However, annotated datasets are often small, typically a few hundred to a few thousand WSI, which may cause overfitting and underperforming models, whereas large unannotated

datasets of tens of thousands WSI are available. Here, we used a slide level SSL model, called Giga-SSL, allowing us to leverage the large number of WSI without annotations to infer powerful slide representations (Lazard et al. 2023). Our model surpassed the performance of the standard MIL model in the binary classification task for the hepatic stem-like subtype. Depending on the availability of a tumor region annotation, we observed a gain of 2 to 6 points in AUC. Besides a slight improvement in classification performance, this model significantly improves efficiency through increased speed and reduced use of computational resources. After the WSI embedding, all analyses used logistic regression and were seamlessly processed on a laptop CPU.

Our model has demonstrated better prediction results when applied to tumour tiles, rather than to the integrality of tiles (including non-tumour tiles) suggesting that the essential information regarding transcriptomic subclasses is contained in the tumour itself, rather than in its environment and other parts of the tissue. To bypass the time-intensive process of manual annotations by a pathologist, we suggest an automatic learning filter given its close performance to manual filter, which represents a favourable balance between time invested and classification performance.

Furthermore, we provided a comparative analysis that explores the influence of training set composition on prediction accuracy. Our findings suggest that intratumoral heterogeneity can negatively impact training when non-consecutive slides are employed for molecular profiling and pathological assessment. This discrepancy introduces label noise, as the molecular class we aim to predict may not align with the tissue captured in the image. While it is a well-known requirement to have large datasets for effective neural network training, our results suggest that datasets with high-confidence labels outperform larger, noise-prone datasets.

Moreover, Jakob Nikolas Kather et al. (2020) found that flash-frozen slides yielded better performance in molecular prediction tasks within the TCGA dataset, despite their poorer morphological quality compared to FFPE slides. We propose that this anomaly could also be attributed to label noise arising from tumor heterogeneity, as the molecular labels in TCGA are extracted from flash-frozen samples. These insights underscore the importance of using consecutive slides for molecular class prediction and could potentially inform the design of future studies. Finally, though more research is needed to validate the clinical usage of such predictive models, we conjecture that patient stratification would benefit from ensembling the prediction of several WSIs sampled in different blocks, which would help capture the main iCCA class of the tumour.

We included both biopsy and surgical samples in the discovery set because we believed it was essential for our model to handle both sample types. Even though using biopsies might have reduced our model's performance during cross-validation, it improved performance on the French external validation set. This suggests that biopsies provide complementary information to surgical specimen WSIs (see Fig. VI.3). Currently, most AI studies of primary liver cancers have focused on surgical samples (Cheng et al. 2022; Jakob N. Kather and Calderaro 2020; Saillard et al. 2020; Zeng et al. 2022), but most patients do not have such samples during

their entire cancer history introducing a selection bias. In our study, noticeably, we found that transcriptomic groups were differently represented between surgical and biopsy cases highlighting the importance of work in both samples. Few studies, mainly focused on diagnosis tasks, have laid the groundwork for using biopsies and have demonstrated that encouraging deep-learning–based results can be obtained in this type of sample despite their size (Albrecht et al. 2023; Pantanowitz et al. 2020; F. Xu et al. 2021).

Our study has some limitations. The proportion of each transcriptomic group was very different and in particular the desert-like was infrequent representing less than 10% of cases in all sets making model learning for this class more difficult and leading to poorer prediction performance. In addition, we were unable to evaluate the predictive performance of our model for the tumour classical class in the TCGA set, as there were no cases of this subtype in this set. Learning from a larger number of cases could be beneficial, nevertheless it is difficult to find very complete datasets containing survival and transcriptomic data, and histological slides of FFPE iCCA samples, as evidenced by the low number of cases available in the TCGA dataset. Finally, the giga-SSL models are not visually interpretable. However, the histological reviewing carried out beforehand and the results of Martin-Serrano et al. (2022) have enabled to highlight different histological characteristics between transcriptomic groups, particularly in the tumour microenvironment composition.

VI.5 Conclusion

We have developed and validated a SSL model able to predict iCCA transcriptomic classes on routine WSI from biopsy and surgical samples. This model has shown good performance for the classification of hepatic stem-like class, tumour classical and inflammatory stroma. Our model surpassed the performance of the standard MIL model and our results suggest that datasets with high-confidence labels outperform larger, noise-prone datasets.
The ability to predict transcriptomic iCCA classes on routine WSI could thus have an impact on the management of patients by predicting their prognosis and guiding the treatment strategy.

Discussion

VII

VII.1 Conclusions

In this thesis, I focused on developing predictive models operating on WSI. Although these models may have significant potential for clinical applications—a point elaborated in the introduction of each chapter—the central questions of this thesis are geared towards the methodology of *how to train and use* these models, rather than delving into the biomedical context, which was brought by our collaborators.

A distinctive challenge in applying machine learning to computational pathology is the unique nature of supervision. In this field, labels are not only scarce but also costly to obtain, weak, and noisy. Therefore, a pivotal question that arised is how to adapt machine learning algorithms to operate effectively under these constraints.

VII.1.1 Weakness of the slide-level supervision

The limitations of slide-level supervision in WSI classification are intrinsically tied to the signal-to-noise ratio. Specifically, when only a few tiles carry the classification signal, the remaining tiles introduce noise rather than useful information. This challenge led the community to employ MIL frameworks as a primary approach for addressing WSI classification issues. In this thesis, we introduce several strategies to enhance weakly supervised classification problems:

1. **Improved Instance Representations**: focusing on improving the representations of individual tiles has a direct impact on WSI-level classifications. Self-supervised learning techniques for training the tile-embedder emerged as a central component in this approach. We employed self-supervised learning in all studies presented in this thesis with notable success. Moreover, regional annotations, when available, can further refine the tile-embedder, provided that the model has undergone prior self-supervised training (Chapter IV).

2. **Pre-trained WSI Representations**: Our work in Chapter V shows that using pre-trained WSI representations effectively transforms a weakly supervised problem into a strongly supervised one. The framework, presented in this chapter, autonomously aggregates tile-level information without requiring strong or weak supervision. These aggregated representations demonstrated robust discriminative power across a diverse array of classification tasks.

VII.1.2 Scarcity of labels

The issue of label scarcity is another area where SSL proves beneficial. Our WSI-level SSL framework, as discussed in Sections V.1 and V.2, constructs WSI representations that enable highly label-efficient classifier training. Indeed, these classifiers maintain robust performance even when trained on a dataset containing as little as 50 WSIs. This makes them particularly well-suited for applications with limited labeled data, such as clinical trials.

The paucity of labels can also results in strong batch-effect in merged datasets. As suggested in Chapter III, we offer a method to counteract this effect during WSI classification training.

VII.1.3 Label uncertainty - label noise

Highly clinically relevant classification tasks, such as those outlined in Chapters III and VI, often use biological measurements like RNA and DNA sequencing obtained from tissue samples. The correlation between these measurements and the WSI content is uncertain, posing questions about whether a relationship exists and, if so, its nature. This uncertainty increases the importance of interpretability; a well-understood algorithm could bridge gaps in current knowledge.

We introduce two methods for algorithmic interpretation in Chapter III and Section V.3. The first is an unsupervised[1] method that extracts visual explanations as sets of representative tiles. While powerful for discovering new patterns, as in the case of identifying laminated fibrosis in HR-deficient tumors, this approach is qualitative, outputting visual explanations in the form selected tile images. This makes it difficult to combine with other datasets or explanations for a more general understanding, such as comparing HRD classifications between luminal and general breast cancer cohorts. The second method is based on predefined interpretation concepts and delivers quantitative explanations, thereby facilitating broader insights. Combining these two methods could provide an interesting avenue. For instance, the unsupervised approach could define morphological concepts that could then be queried in a shared WSI latent space, enhancing our understanding of cancer phenotypes.

Another problem of datasets in computational pathology is label noise. This issue can arise when the measured tumor sample is spatially separated from the WSI sample, with tumor heterogeneity possibly leading to inconsistencies between the two. We demonstrate the significance of this effect in Chapter VI. In scenarios like this, classification models benefit more from training on a smaller, accurate dataset than a larger, potentially noisy one.

[1] This method is unsupervised because it doesn't require predefined prototypes of morphological interpretations. However, it interprets a classification model trained with supervision.

VII.2 Perspectives

This section outlines research directions opened by this study. I'll begin with short-term opportunities, followed by an insight on long-term prospects.

VII.2.1 Short-term opportunities

VII.2.1.1. Improving the Giga-SSL framework

Firstly, there is room for improvement by using more advanced architecture for the Giga-SSL aggregation network. Visual transformers have shown performance gains and integrating them into the Giga-SSL aggregation block could boost effectiveness. Also, the bottleneck in Giga-SSL training is the pre-computation of augmented tile embeddings. Zaffar et al. (2022) recently suggested a generative model to generate augmentations in the embedding space. This could lower computational costs of Giga-SSL training. If used to augment Giga-SSL WSI embeddings directly -and not tile embeddings-, it could be used to regularize the logistic regression trained on top of the Giga-SSL representations and hopefully improve their performances, for example on small datasets.

VII.2.1.2. Studying label noise at scale

We made the hypothesis in Chapter VI that the mismatch in performance accuracy between paired and unpaired WSIs is attributable to intra-tumoral heterogeneity (ITH). Indeed, a model trained on WSIs paired with their label's acquisition process will yield poorer results on un-paired WSIs simply because of the mis-labeling of some. We name this measure mismatch-ITH for clarity. This hypothesis would first need a proper validation using a dataset with multiple samples per patient for both RNA-seq and WSI acquisition. This approach would allow us to measure true RNA expression heterogeneity and correlate it with mismatch-ITH.

This surrogate measure of morphological ITH could be useful in two ways: first, it could aid in identifying WSI classification tasks with minimal label noise, thus facilitating the establishment of a robust set of benchmark tasks as outlined in Section VII.2.2.1. The measure could also provide insights about evolutionary processes that either encourage or inhibit different levels of ITH associated with specific genetic signatures. Finally, it could also inform the development of effective inference protocols in clinical settings where predictive algorithms are used. For instance, if WSI-derived HRD predictions are used in a clinical setting, it could help determine the optimal number of WSIs to sample per patient in order to minimize the rate of false negatives -in case of high heterogeneity-.

VII.2.1.3. Improving the morphological interpretation

As outlined in Section VII.1.3, the two interpretability algorithms developed in this thesis can function in a complementary manner. For a given classification task, we could train both a MIL model and logistic regressions in the Giga-SSL space. The unsupervised visualization method described in Chapter III would then operate on the trained MIL model and provide a qualitative approach for identifying unknown morphological patterns linked to the task. Once these patterns are discovered, they can serve for creating new morphological *concepts* to be integrated into the morphological profiles discussed in Section V.3. Subsequently, these newly identified patterns can be evaluated for their relevance across different types of cancer, in relation to different classification tasks, which could in-turn yield insights into the causes of these phenotypes.

VII.2.2 Broader perspectives

VII.2.2.1. Collective choice of benchmark tasks

The research practices in computational pathology, especially the integration of ML techniques, inheritate much to established practices in ML research. This approach is model-centric: new algorithms are developed to compete on a predefined set of benchmark tasks. This has significant implications; the specific challenges presented by these tasks have the potential to steer future developments. I argue here that our limited understanding of the link between the WSI and the labels, i.e. the ground-truth function $\mathcal{G}_t$, may be slowing the field's progress.

A notable example underscoring this limitation is the issue of signal localization within WSIs. MIL algorithms assume by design a certain localization of the signal, i.e. that the slide level is driven by specific tiles, even though their number might be very low, as detailed in Section II.1.1.1. MIL's initial adoption was largely due to its capacity to mimic pathological diagnosis criteria[2] and its success in early data challenges like Camelyon[3]. However, I argue that most of the classification tasks we confront likely depend on signals that are diffused across the WSI rather than being highly localized.

Support for this claim comes from the Giga-SSL algorithm discussed in Section Section V.1. This method employs a strong tile sub-sampling transformation, which have the effect of diluting localized information within WSI embeddings. Despite this, Giga-SSL has exhibited strong performance across a range of tasks. This suggests that the ground-truth function $\mathcal{G}_t$ for these tasks may be far more reliant on diffuse signals across the WSI than on isolated, localized signals. Further evidence comes from the nature of the tasks tackled in Chapter III and Secs. V.1 and V.2, which predominantly

[2] Answering question such as *"does the WSI contains tumourous cells ?"* or *"Does this WSI contains at least one high grade lesion?"*

[3] The aim of Camelyon was to detect WSIs of lymph nodes containing metastatic tumor, which is likened in some cases to *a needle in a haystack* problem.

involve bulk measurements on tumor samples. These measurements inherently average out any localized features, potentially explaining the acceptable performance achieved even by basic MIL models like instance-mean and the improvements seen with Giga-SSL.

Given these observations, we must question whether MIL is the most appropriate framework for WSI classification. Moreover, if we continue to benchmark algorithms exclusively on tasks with diffuse signals, the ability of these algorithms to effectively handle localized information may remain unexplored and unoptimized. This highlights the need for a critical reassessment of the benchmark tasks commonly used for algorithm development in computational pathology.

The issue extends beyond signal localization. Other potential pitfalls include confounding variables causing batch-effects. As highlighted by Howard et al. (2021), site-specific information can confound the prediction outcomes for various classification tasks. Despite this, numerous studies still utilize these flawed benchmarks, raising questions about the validity of algorithmic improvements.

Similarly, as indicated in Chapter VI, labels based on RNA/DNA sequencing can be noisy. This is particularly relevant for TCGA datasets, where biological measures are obtained from frozen samples, different from the FFPE samples used to prepare the slides. Such inconsistencies may mask real algorithmic improvements in current benchmark tasks.

Considering all these factors—signal localization, potential confounding variables, and an unknown level of noise—I advocate for the careful design of a new set of benchmark tasks. These tasks should have a minimized level of label-noise and known and controled biases. They also should span a wide array of problems, allowing the development of specialized algorithms based on the specifics of each task's ground-truth function $\mathcal{G}_t$.

Thus, a concerted collective effort should be made to focus on the dataset side of the equation, rather than solely on model development, to genuinely propel the field of computational pathology forward.

VII.2.2.2. Toward multimodal foundation models

SSL has significantly advanced various disciplines, including medical imaging and pathology, and it has been at the basis of the improvements discussed in this thesis. WSI becomes increasingly prevalent in healthcare settings, and we are on the cusp of having access to datasets comprising tens of thousands of slides. Effectively utilizing this wealth of data will be critical for enhancing the performance of future WSI algorithms, and SSL offers this opportunity.

In recent months, key actors of the domain such as academic researchers and industrial R&D teams, have developed various SSL methods for training tile-encoder networks (Richard J. Chen, Ding, et al. 2023; Filiot et al. 2023; X. Wang et al. 2022; Lin et al. 2023). Some have even utilized tiles from as many as a million WSIs in their training sets (Vorontsov et al. 2023). The overarching aim is to create versatile

"foundation models" that can be fine-tuned for specialized tasks. These foundation models not only serve as the basis for improved downstream classifiers but also offer a better understanding of the input data. Their learned latent spaces act as manipulable interfaces, which can be explored using methods like those detailed in Section V.3, for example.

These SSL-derived latent spaces are not limited to image modalities. For example, textual data, such as medical descriptions of histopathological features, can also be embedded within these spaces. The CLIP framework (Radford et al. 2021) provides a successful algorithm to learn joint visual and textual embeddings using only image and text pairing. This concept has been recently applied to histopathology through works like PLIP, Conche, and Quilt (Z. Huang et al. 2023; Ikezogwo et al. 2023; Lu et al. 2023). The success reported by these joint embeddings, exhibiting zero-shot classification capabilities that rival state-of-the-art supervised methods, thereby expand the realm of what is achievable.

The next frontier lies in extending this multimodal framework beyond traditional image and text data. For instance, recent progress in developing foundation models for genomic data (Fishman et al. 2023) signals an opportunity for further integration. Similarly, radiology data can be informative about another scale and about other properties of tumors and metastases.

Aligning the manifold embeddings of various tumor modalities, similar to what has been done with CLIP, could yield groundbreaking insights. Such an approach would promote inter-modality dialogue and represent a significant step toward a more comprehensive understanding of cancer.

References

Abkevich, V., K. M. Timms, B. T. Hennessy, J. Potter, M. S. Carey, L. A. Meyer, K. Smith-McCune, et al. 2012. "Patterns of Genomic Loss of Heterozygosity Predict Homologous Recombination Repair Defects in Epithelial Ovarian Cancer." *British Journal of Cancer* 107 (10): 1776–82. https://doi.org/10.1038/bjc.2012.451.

Abou-Alfa, Ghassan K., Teresa Macarulla, Milind M. Javle, Robin K. Kelley, Sam J. Lubner, Jorge Adeva, James M. Cleary, et al. 2020. "Ivosidenib in IDH1-mutant, Chemotherapy-Refractory Cholangiocarcinoma (ClarIDHy): A Multicentre, Randomised, Double-Blind, Placebo-Controlled, Phase 3 Study." *The Lancet. Oncology* 21 (6): 796–807. https://doi.org/10.1016/S1470-2045(20)30157-1.

Adeli, Ehsan, Qingyu Zhao, Adolf Pfefferbaum, Edith V. Sullivan, Li Fei-Fei, Juan Carlos Niebles, and Kilian M. Pohl. 2020. "Representation Learning with Statistical Independence to Mitigate Bias." *arXiv:1910.03676 [Cs]*, November. https://arxiv.org/abs/1910.03676.

Albrecht, Thomas, Annik Rossberg, Jana Dorothea Albrecht, Jan Peter Nicolay, Beate Katharina Straub, Tiemo Sven Gerber, Michael Albrecht, et al. 2023. "Deep Learning-Enabled Diagnosis of Liver Adenocarcinoma." *Gastroenterology*, August, S0016-5085(23)04883-7. https://doi.org/10.1053/j.gastro.2023.07.026.

Alexandrov, Ludmil B., Serena Nik-Zainal, David C. Wedge, Samuel A. J. R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, et al. 2013. "Signatures of Mutational Processes in Human Cancer." *Nature* 500 (7463): 415–21. https://doi.org/10.1038/nature12477.

Amores, Jaume. 2013. "Multiple Instance Classification: Review, Taxonomy and Comparative Study." *Artif. Intell.* https://doi.org/10.1016/j.artint.2013.06.003.

Asif, Amina, Kashif Rajpoot, Simon Graham, David Snead, Fayyaz Minhas, and Nasir Rajpoot. 2023. "Unleashing the Potential of AI for Pathology: Challenges and Recommendations." *The Journal of Pathology*, August, path.6168. https://doi.org/10.1002/path.6168.

Azevedo Tosta, Thaína A., Paulo Rogério de Faria, Leandro Alves Neves, and Marcelo Zanchetta do Nascimento. 2019. "Computational Normalization of H&E-Stained Histological Images: Progress, Challenges and Future Potential." *Artificial Intelligence in Medicine* 95 (April): 118–32. https://doi.org/10.1016/j.artmed.2018.10.004.

Azizi, Shekoofeh, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, et al. 2023. "Robust and Data-Efficient Generalization of Self-Supervised Machine Learning for Diagnostic Imaging." *Nature Biomedical Engineering* 7 (6): 756–79. https://doi.org/10.1038/s41551-023-01049-7.

Babic, Andrea, Isabell R. Loftin, Stacey Stanislaw, Maria Wang, Rachel Miller, Stephanie M. Warren, Wenjun Zhang, et al. 2010. "The Impact of Pre-Analytical Processing on Staining Quality for H&E, Dual Hapten, Dual Color in Situ Hy-

bridization and Fluorescent in Situ Hybridization Assays." *Methods (San Diego, Calif.)* 52 (4): 287–300. https://doi.org/10.1016/j.ymeth.2010.08.012.
Balestriero, Randall, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, et al. 2023. "A Cookbook of Self-Supervised Learning." arXiv. https://doi.org/10.48550/arXiv.2304.12210.
Bankhead, Peter, Maurice B. Loughrey, José A. Fernández, Yvonne Dombrowski, Darragh G. McArt, Philip D. Dunne, Stephen McQuaid, et al. 2017. "QuPath: Open Source Software for Digital Pathology Image Analysis." *Scientific Reports* 7 (1): 16878. https://doi.org/10.1038/s41598-017-17204-5.
Bardes, Adrien, Jean Ponce, and Yann LeCun. 2022. "VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning." arXiv. https://doi.org/10.48550/arXiv.2105.04906.
Bedossa, P., and T. Poynard. 1996. "An Algorithm for the Grading of Activity in Chronic Hepatitis C. The METAVIR Cooperative Study Group." *Hepatology (Baltimore, Md.)* 24 (2): 289–93. https://doi.org/10.1002/hep.510240201.
Binder, Alexander, Leander Weber, Sebastian Lapuschkin, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2022. "Shortcomings of Top-Down Randomization-Based Sanity Checks for Evaluations of Deep Neural Network Explanations." arXiv. https://doi.org/10.48550/arXiv.2211.12486.
Birkbak, Nicolai J., Zhigang C. Wang, Ji-Young Kim, Aron C. Eklund, Qiyuan Li, Ruiyang Tian, Christian Bowman-Colin, et al. 2012. "Telomeric Allelic Imbalance Indicates Defective DNA Repair and Sensitivity to DNA-damaging Agents." *Cancer Discovery* 2 (4): 366–75. https://doi.org/10.1158/2159-8290.CD-11-0206.
Board, WHO Classification of Tumours Editorial. n.d.a. *Breast Tumours*. Accessed August 29, 2023.
———. n.d.b. *Digestive System Tumours*. Accessed August 29, 2023.
Bordes, Florian, Randall Balestriero, and Pascal Vincent. 2023. "Towards Democratizing Joint-Embedding Self-Supervised Learning." arXiv. https://doi.org/10.48550/arXiv.2303.01986.
Bridgewater, John, Peter R. Galle, Shahid A. Khan, Josep M. Llovet, Joong-Won Park, Tushar Patel, Timothy M. Pawlik, and Gregory J. Gores. 2014. "Guidelines for the Diagnosis and Management of Intrahepatic Cholangiocarcinoma." *Journal of Hepatology* 60 (6): 1268–89. https://doi.org/10.1016/j.jhep.2014.01.021.
Bryant, Helen E., Niklas Schultz, Huw D. Thomas, Kayan M. Parker, Dan Flower, Elena Lopez, Suzanne Kyle, Mark Meuth, Nicola J. Curtin, and Thomas Helleday. 2005. "Specific Killing of BRCA2-deficient Tumours with Inhibitors of Poly(ADP-Ribose) Polymerase." *Nature* 434 (7035): 913–17. https://doi.org/10.1038/nature03443.
Calderaro, Julien, and Jakob Nikolas Kather. 2021. "Artificial Intelligence-Based Pathology for Gastrointestinal and Hepatobiliary Cancers." *Gut* 70 (6): 1183–93. https://doi.org/10.1136/gutjnl-2020-322880.
Calderaro, Julien, Tobias Paul Seraphin, Tom Luedde, and Tracey G. Simon. 2022. "Artificial Intelligence for the Prevention and Clinical Management of Hepatocellular Carcinoma." *Journal of Hepatology* 76 (6): 1348–61. https://doi.org/10.1016/j.jhep.2022.01.014.
Campanella, Gabriele, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S.

Klimstra, and Thomas J. Fuchs. 2019. “Clinical-Grade Computational Pathology Using Weakly Supervised Deep Learning on Whole Slide Images.” *Nature Medicine* 25 (8): 1301–9. https://doi.org/10.1038/s41591-019-0508-1.

Campanella, Gabriele, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. 2019. “Clinical-Grade Computational Pathology Using Weakly Supervised Deep Learning on Whole Slide Images.” *Nature Medicine* 25 (8): 1301–9. https://doi.org/10.1038/s41591-019-0508-1.

Caron, Mathilde, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. “Emerging Properties in Self-Supervised Vision Transformers.” arXiv. https://doi.org/10.48550/arXiv.2104.14294.

Cassoux, Nathalie, Manuel Jorge Rodrigues, Corine Plancher, Bernard Asselain, Christine Levy-Gabriel, Livia Lumbroso-Le Rouic, Sophie Piperno-Neumann, et al. 2014. “Genome-Wide Profiling Is a Clinically Relevant and Affordable Prognostic Test in Posterior Uveal Melanoma.” *The British Journal of Ophthalmology* 98 (6): 769–74. https://doi.org/10.1136/bjophthalmol-2013-303867.

Chen, Richard J, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. n.d. “Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning,” 12.

Chen, Richard J., Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, et al. 2023. “A General-Purpose Self-Supervised Model for Computational Pathology.” arXiv. https://doi.org/10.48550/arXiv.2308.15474.

Chen, Richard J., Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, et al. 2022. “Pan-Cancer Integrative Histology-Genomic Analysis via Multimodal Deep Learning.” *Cancer Cell* 40 (8): 865–878.e6. https://doi.org/10.1016/j.ccell.2022.07.004.

Chen, Richard J., Judy J. Wang, Drew F. K. Williamson, Tiffany Y. Chen, Jana Lipkova, Ming Y. Lu, Sharifa Sahai, and Faisal Mahmood. 2023. “Algorithmic Fairness in Artificial Intelligence for Medicine and Healthcare.” *Nature Biomedical Engineering* 7 (6): 719–42. https://doi.org/10.1038/s41551-023-01056-8.

Chen, Ting, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. “A Simple Framework for Contrastive Learning of Visual Representations.” *arXiv:2002.05709 [Cs, Stat]*, February. https://arxiv.org/abs/2002.05709.

Chen, Ting, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. 2020. “Big Self-Supervised Models Are Strong Semi-Supervised Learners.” *arXiv:2006.10029 [Cs, Stat]*, October. https://arxiv.org/abs/2006.10029.

Chen, Ting, Calvin Luo, and Lala Li. 2021. “Intriguing Properties of Contrastive Losses.” *arXiv:2011.02803 [Cs, Stat]*, October. https://arxiv.org/abs/2011.02803.

Chen, Xinlei, Haoqi Fan, Ross Girshick, and Kaiming He. 2020. “Improved Baselines with Momentum Contrastive Learning.” *arXiv:2003.04297 [Cs]*, March. https://arxiv.org/abs/2003.04297.

Chen, Xinlei, and Kaiming He. 2020. “Exploring Simple Siamese Representation Learning.” *arXiv:2011.10566 [Cs]*, November. https://arxiv.org/abs/2011.10566.

Chen, Xuxin, Ximin Wang, Ke Zhang, Kar-Ming Fung, Theresa C. Thai, Kathleen Moore, Robert S. Mannel, Hong Liu, Bin Zheng, and Yuchen Qiu. 2022. "Recent Advances and Clinical Applications of Deep Learning in Medical Image Analysis." *Medical Image Analysis* 79 (July): 102444. https://doi.org/10.1016/j.media.2022.102444.

Cheng, Na, Yong Ren, Jing Zhou, Yiwang Zhang, Deyu Wang, Xiaofang Zhang, Bing Chen, et al. 2022. "Deep Learning-Based Classification of Hepatocellular Nodular Lesions on Whole-Slide Histopathologic Images." *Gastroenterology* 162 (7): 1948–1961.e7. https://doi.org/10.1053/j.gastro.2022.02.025.

Chopra, Neha, Holly Tovey, Alex Pearson, Ros Cutts, Christy Toms, Paula Proszek, Michael Hubank, et al. 2020. "Homologous Recombination DNA Repair Deficiency and PARP Inhibition Activity in Primary Triple Negative Breast Cancer." *Nature Communications* 11 (1): 2662. https://doi.org/10.1038/s41467-020-16142-7.

Chung, Yu-An, Hsuan-Tien Lin, and Shao-Wen Yang. 2016. "Cost-Aware Pre-training for Multiclass Cost-sensitive Deep Learning." *arXiv:1511.09337 [Cs]*, May. https://arxiv.org/abs/1511.09337.

Ciga, Ozan, and Anne L. Martel. 2021. "Learning to Segment Images with Classification Labels." *Medical Image Analysis* 68 (February): 101912. https://doi.org/10.1016/j.media.2020.101912.

Ciga, Ozan, Tony Xu, and Anne L. Martel. 2021. "Self Supervised Contrastive Learning for Digital Histopathology." *arXiv:2011.13971 [Cs, Eess]*, September. https://arxiv.org/abs/2011.13971.

Costantini, Massimo, Stefania Sciallero, Augusto Giannini, Beatrice Gatteschi, Paolo Rinaldi, Giuseppe Lanzanova, Luigina Bonelli, et al. 2003. "Interobserver Agreement in the Histologic Diagnosis of Colorectal Polyps. The Experience of the Multicenter Adenoma Colorectal Study (SMAC)." *Journal of Clinical Epidemiology* 56 (3): 209–14. https://doi.org/10.1016/s0895-4356(02)00587-5.

Coudray, Nicolas, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L. Moreira, Narges Razavian, and Aristotelis Tsirigos. 2018. "Classification and Mutation Prediction from Non–small Cell Lung Cancer Histopathology Images Using Deep Learning." *Nature Medicine* 24 (10): 1559–67. https://doi.org/10.1038/s41591-018-0177-5.

Courtiol, Pierre, Charles Maussion, Matahi Moarii, Elodie Pronier, Samuel Pilcer, Meriem Sefta, Pierre Manceron, et al. 2019b. "Deep Learning-Based Classification of Mesothelioma Improves Prediction of Patient Outcome." *Nature Medicine* 25 (10): 1519–25. https://doi.org/10.1038/s41591-019-0583-3.

———, et al. 2019a. "Deep Learning-Based Classification of Mesothelioma Improves Prediction of Patient Outcome." *Nature Medicine* 25 (10): 1519–25. https://doi.org/10.1038/s41591-019-0583-3.

Courtiol, Pierre, Eric W. Tramel, Marc Sanselme, and Gilles Wainrib. 2018. "Classification and Disease Localization in Histopathology Using Only Global Labels: A Weakly-Supervised Approach." *arXiv:1802.02212 [Cs, Stat]*, February. https://arxiv.org/abs/1802.02212.

Davies, Helen, Dominik Glodzik, Sandro Morganella, Lucy R. Yates, Johan Staaf, Xueqing Zou, Manasa Ramakrishna, et al. 2017. "HRDetect Is a Predictor of

BRCA1 and BRCA2 Deficiency Based on Mutational Signatures." *Nature Medicine* 23 (4): 517–25. https://doi.org/10.1038/nm.4292.

Dehaene, Olivier, Axel Camara, Olivier Moindrot, Axel de Lavergne, and Pierre Courtiol. 2020. "Self-Supervision Closes the Gap Between Weak and Strong Supervision in Histology." *arXiv:2012.03583 [Cs, Eess]*, December. https://arxiv.org/abs/2012.03583.

Deluche, Elise, Alison Antoine, Thomas Bachelot, Audrey Lardy-Cleaud, Veronique Dieras, Etienne Brain, Marc Debled, et al. 2020. "Contemporary Outcomes of Metastatic Breast Cancer Among 22,000 Women from the Multicentre ESME Cohort 2008–2016." *European Journal of Cancer* 129 (April): 60–70. https://doi.org/10.1016/j.ejca.2020.01.016.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. "ImageNet: A Large-Scale Hierarchical Image Database." In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–55. https://doi.org/10.1109/CVPR.2009.5206848.

Deniz, Erkan, Abdulkadir Şengür, Zehra Kadiroğlu, Yanhui Guo, Varun Bajaj, and Ümit Budak. 2018. "Transfer Learning Based Histopathologic Image Classification for Breast Cancer Detection." *Health Information Science and Systems* 6 (1): 18. https://doi.org/10.1007/s13755-018-0057-x.

Diao, James A., Jason K. Wang, Wan Fung Chui, Victoria Mountain, Sai Chowdary Gullapally, Ramprakash Srinivasan, Richard N. Mitchell, et al. 2021. "Human-Interpretable Image Features Derived from Densely Mapped Cancer Pathology Slides Predict Diverse Molecular Phenotypes." *Nature Communications* 12 (1): 1613. https://doi.org/10.1038/s41467-021-21896-9.

Dietterich, Thomas G., Richard H. Lathrop, and Tomás Lozano-Pérez. 1997. "Solving the Multiple Instance Problem with Axis-Parallel Rectangles." *Artificial Intelligence* 89 (1): 31–71. https://doi.org/10.1016/S0004-3702(96)00034-3.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, et al. 2020. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale." *arXiv:2010.11929 [Cs]*, October. https://arxiv.org/abs/2010.11929.

Durand, Thibaut, Taylor Mordan, Nicolas Thome, and Matthieu Cord. 2017. "WILDCAT: Weakly Supervised Learning of Deep ConvNets for Image Classification, Pointwise Localization and Segmentation." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 642–51.

Durand, Thibaut, Nicolas Thome, and Matthieu Cord. 2016. "WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks." In, 4743–52. https://doi.org/10.1109/CVPR.2016.513.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2011. "Fairness Through Awareness." arXiv. https://doi.org/10.48550/arXiv.1104.3913.

Echle, Amelie, Narmin Ghaffari Laleh, Peter L. Schrammen, Nicholas P. West, Christian Trautwein, Titus J. Brinker, Stephen B. Gruber, et al. 2021. "Deep Learning for the Detection of Microsatellite Instability from Histology Images in Colorectal Cancer: A Systematic Literature Review." *ImmunoInformatics* 3–4 (December): 100008. https://doi.org/10.1016/j.immuno.2021.100008.

Ehteshami Bejnordi, Babak, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, and and the CAMELYON16 Consortium. 2017. "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer." *JAMA* 318 (22): 2199–2210. https://doi.org/10.1001/jama.2017.14585.

Endo, Itaru, Mithat Gonen, Adam C. Yopp, Kimberly M. Dalal, Qin Zhou, David Klimstra, Michael D'Angelica, et al. 2008. "Intrahepatic Cholangiocarcinoma: Rising Frequency, Improved Survival, and Determinants of Outcome After Resection." *Annals of Surgery* 248 (1): 84–96. https://doi.org/10.1097/SLA.0b013e318176c4d3.

Erhan, Dumitru, Y. Bengio, Aaron Courville, and Pascal Vincent. 2009. "Visualizing Higher-Layer Features of a Deep Network." *Technical Report, Univeristé de Montréal*, January.

European Association for the Study of the Liver. Electronic address: easloffice@easloffice.eu, and European Association for the Study of the Liver. 2023. "EASL-ILCA Clinical Practice Guidelines on Intrahepatic Cholangiocarcinoma." *Journal of Hepatology*, March, S0168-8278(23)00185-X. https://doi.org/10.1016/j.jhep.2023.03.010.

Farmer, Hannah, Nuala McCabe, Christopher J. Lord, Andrew N. J. Tutt, Damian A. Johnson, Tobias B. Richardson, Manuela Santarosa, et al. 2005. "Targeting the DNA Repair Defect in BRCA Mutant Cells as a Therapeutic Strategy." *Nature* 434 (7035): 917–21. https://doi.org/10.1038/nature03445.

Ferrari, Anthony, Anne Vincent-Salomon, Xavier Pivot, Anne-Sophie Sertier, Emilie Thomas, Laurie Tonon, Sandrine Boyault, et al. 2016. "A Whole-Genome Sequence and Transcriptome Perspective on HER2-positive Breast Cancers." *Nature Communications* 7 (1): 12222. https://doi.org/10.1038/ncomms12222.

Ferreira, R., B. Moon, J. Humphries, A. Sussman, J. Saltz, R. Miller, and A. Demarzo. 1997. "The Virtual Microscope." *Proceedings: A Conference of the American Medical Informatics Association. AMIA Fall Symposium*, 449–53.

Filiot, Alexandre, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. 2023. "Scaling Self-Supervised Learning for Histopathology with Masked Image Modeling." medRxiv. https://doi.org/10.1101/2023.07.21.23292757.

Fischer, Andrew H., Kenneth A. Jacobson, Jack Rose, and Rolf Zeller. 2008. "Hematoxylin and Eosin Staining of Tissue and Cell Sections." *Cold Spring Harbor Protocols* 2008 (5): pdb.prot4986. https://doi.org/10.1101/pdb.prot4986.

Fishman, Veniamin, Yuri Kuratov, Maxim Petrov, Aleksei Shmelev, Denis Shepelin, Nikolay Chekanov, Olga Kardymon, and Mikhail Burtsev. 2023. "GENA-LM: A Family of Open-Source Foundational Models for Long DNA Sequences." bioRxiv. https://doi.org/10.1101/2023.06.12.544594.

Foulds, James, and Eibe Frank. 2010. "A Review of Multi-Instance Learning Assumptions." *The Knowledge Engineering Review* 25 (1): 1–25. https://doi.org/10.1017/S026988890999035X.

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. "Domain-Adversarial Training of Neural Networks." arXiv. https://doi.org/10.48550/arXiv.1505.07818.

Ghaffari Laleh, Narmin, Hannah Sophie Muti, Chiara Maria Lavinia Loeffler, Amelie Echle, Oliver Lester Saldanha, Faisal Mahmood, Ming Y. Lu, et al. 2022. "Benchmarking Weakly-Supervised Deep Learning Pipelines for Whole Slide Classification in Computational Pathology." *Medical Image Analysis* 79 (July): 102474. https://doi.org/10.1016/j.media.2022.102474.

Global Burden of Disease Liver Cancer Collaboration. 2017. "The Burden of Primary Liver Cancer and Underlying Etiologies From 1990 to 2015 at the Global, Regional, and National Level: Results From the Global Burden of Disease Study 2015." *JAMA Oncology* 3 (12): 1683–91. https://doi.org/10.1001/jamaoncol.2017.3055.

Graham, Benjamin, and Laurens van der Maaten. 2017. "Submanifold Sparse Convolutional Networks." arXiv. https://doi.org/10.48550/arXiv.1706.01307.

Grill, Jean-Bastien, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, et al. 2020. "Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning." arXiv. https://doi.org/10.48550/arXiv.2006.07733.

Hari, Surya Narayanan, Jackson Nyman, Nicita Mehta, Haitham Elmarakeby, Bowen Jiang, Felix Dietlein, Jacob Rosenthal, et al. 2021. "Examining Batch Effect in Histopathology as a Distributionally Robust Optimization Problem." bioRxiv. https://doi.org/10.1101/2021.09.14.460365.

He, Kaiming, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. "Momentum Contrast for Unsupervised Visual Representation Learning." *arXiv:1911.05722 [Cs]*, March. https://arxiv.org/abs/1911.05722.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015a. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification." *arXiv.org*. https://arxiv.org/abs/1502.01852v1.

———. 2015b. "Deep Residual Learning for Image Recognition." *arXiv:1512.03385 [Cs]*, December. https://arxiv.org/abs/1512.03385.

Hekler, Achim, Jakob N. Kather, Eva Krieghoff-Henning, Jochen S. Utikal, Friedegund Meier, Frank F. Gellrich, Julius Upmeier zu Belzen, et al. 2020. "Effects of Label Noise on Deep Learning-Based Skin Cancer Classification." *Frontiers in Medicine* 7.

Henaff, Olivier J., Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. 2019. "Data-Efficient Image Recognition with Contrastive Predictive Coding," September.

Holstege, Henne, Hugo M. Horlings, Arno Velds, Anita Langerød, Anne-Lise Børresen-Dale, Marc J. van de Vijver, Petra M. Nederlof, and Jos Jonkers. 2010. "BRCA1-mutated and Basal-Like Breast Cancers Have Similar aCGH Profiles and a High Incidence of Protein Truncating TP53 Mutations." *BMC Cancer* 10 (1): 654. https://doi.org/10.1186/1471-2407-10-654.

Howard, Frederick M., James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo, et al. 2021. "The Impact of Site-Specific Digital Histology Signatures on Deep Learning Model Accuracy and Bias." *Nature Communications* 12 (1): 4423. https://doi.org/10.1038/s41467-021-24698-1.

Howard, Frederick M., Jakob Nikolas Kather, and Alexander T. Pearson. 2022. "Multimodal Deep Learning: An Improvement in Prognostication or a Reflection

of Batch Effect?" *Cancer Cell*, November, S1535-6108(22)00522-0. https://doi.org/10.1016/j.ccell.2022.10.025.

Huang, Gao, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2018. "Densely Connected Convolutional Networks." arXiv. https://doi.org/10.48550/arXiv.1608.06993.

Huang, Yi-Jie, Weiping Liu, Xiuying Wang, Qu Fang, Renzhen Wang, Yi Wang, Huai Chen, Hao Chen, Deyu Meng, and Lisheng Wang. 2020. "Rectifying Supporting Regions With Mixed and Active Supervision for Rib Fracture Recognition." *IEEE Transactions on Medical Imaging* 39 (12): 3843–54. https://doi.org/10.1109/TMI.2020.3006138.

Huang, Zhi, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Zou. 2023. "A Visual–language Foundation Model for Pathology Image Analysis Using Medical Twitter." *Nature Medicine*, August, 1–10. https://doi.org/10.1038/s41591-023-02504-3.

Ikezogwo, Wisdom Oluchi, Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Stefan Chan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. 2023. "Quilt-1M: One Million Image-Text Pairs for Histopathology." arXiv. https://arxiv.org/abs/2306.11207.

Ilse, Maximilian, Jakub M. Tomczak, and Max Welling. 2018. "Attention-Based Deep Multiple Instance Learning." *arXiv:1802.04712 [Cs, Stat]*, June. https://arxiv.org/abs/1802.04712.

Ioffe, Sergey, and Christian Szegedy. 2015. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." *arXiv:1502.03167 [Cs]*, February. https://arxiv.org/abs/1502.03167.

Jiang, Cheng, Xinhai Hou, Akhil Kondepudi, Asadur Chowdury, Christian W. Freudiger, Daniel A. Orringer, Honglak Lee, and Todd C. Hollon. 2023. "Hierarchical Discriminative Learning Improves Visual Representations of Biomedical Microscopy." arXiv. https://doi.org/10.48550/arXiv.2303.01605.

Kanavati, Fahdi, and Masayuki Tsuneki. 2021. "Breast Invasive Ductal Carcinoma Classification on Whole Slide Images with Weakly-Supervised and Transfer Learning." *Cancers* 13 (21): 5368. https://doi.org/10.3390/cancers13215368.

Kang, Mingu, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. 2022. "Benchmarking Self-Supervised Learning on Diverse Pathology Datasets." arXiv. https://arxiv.org/abs/2212.04690.

Kather, Jakob N., and Julien Calderaro. 2020. "Development of AI-based Pathology Biomarkers in Gastrointestinal and Liver Cancer." *Nature Reviews. Gastroenterology & Hepatology* 17 (10): 591–92. https://doi.org/10.1038/s41575-020-0343-3.

Kather, Jakob Nikolas, Lara R. Heij, Heike I. Grabsch, Chiara Loeffler, Amelie Echle, Hannah Sophie Muti, Jeremias Krause, et al. 2020. "Pan-Cancer Image-Based Detection of Clinically Actionable Genetic Alterations." *Nature Cancer* 1 (8): 789–99. https://doi.org/10.1038/s43018-020-0087-6.

Kensert, Alexander, Philip J. Harrison, and Ola Spjuth. 2019. "Transfer Learning with Deep Convolutional Neural Networks for Classifying Cellular Morphological Changes." *SLAS Discovery* 24 (4): 466–75. https://doi.org/10.1177/2472555218818756.

Kieffer, Brady, Morteza Babaie, Shivam Kalra, and H. R. Tizhoosh. 2017. "Convolutional Neural Networks for Histopathology Image Classification: Training Vs. Using Pre-Trained Networks." In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 1–6. https://doi.org/10.1109/IPTA.2017.8310149.

Kim, Been, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)." *arXiv:1711.11279 [Stat]*, June. https://arxiv.org/abs/1711.11279.

Kim, So-Woon, Jin Roh, and Chan-Sik Park. 2016. "Immunohistochemistry for Pathologists: Protocols, Pitfalls, and Tips." *Journal of Pathology and Translational Medicine* 50 (6): 411–18. https://doi.org/10.4132/jptm.2016.08.08.

Kingma, Diederik P., and Jimmy Ba. 2014. "Adam: A Method for Stochastic Optimization." *arXiv:1412.6980 [Cs]*, December. https://arxiv.org/abs/1412.6980.

Knijnenburg, Theo A., Linghua Wang, Michael T. Zimmermann, Nyasha Chambwe, Galen F. Gao, Andrew D. Cherniack, Huihui Fan, et al. 2018. "Genomic and Molecular Landscape of DNA Damage Repair Deficiency Across The Cancer Genome Atlas." *Cell Reports* 23 (1): 239–254.e6. https://doi.org/10.1016/j.celrep.2018.03.076.

Koh, Pang Wei, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. "Concept Bottleneck Models." *arXiv:2007.04612 [Cs, Stat]*, December. https://arxiv.org/abs/2007.04612.

Krane, Gregory A., Keith R. Shockley, David E. Malarkey, Andrew D. Miller, C. Ryan Miller, Debra A. Tokarz, Heather L. Jensen, Kyathanahalli S. Janardhan, Matthew Breen, and Christopher L. Mariani. 2022. "Inter-Pathologist Agreement on Diagnosis, Classification and Grading of Canine Glioma." *Veterinary and Comparative Oncology* 20 (4): 881–89. https://doi.org/10.1111/vco.12853.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc.

L'Imperio, Vincenzo, Ellery Wulczyn, Markus Plass, Heimo Müller, Nicolò Tamini, Luca Gianotti, Nicola Zucchini, et al. 2023. "Pathologist Validation of a Machine Learning–Derived Feature for Colon Cancer Risk Stratification." *JAMA Network Open* 6 (3): e2254891. https://doi.org/10.1001/jamanetworkopen.2022.54891.

Lakhani, Sunil R., Marc J. Van De Vijver, Jocelyne Jacquemier, Thomas J. Anderson, Peter P. Osin, Lesley McGuffog, and Douglas F. Easton. 2002. "The Pathology of Familial Breast Cancer: Predictive Value of Immunohistochemical Markers Estrogen Receptor, Progesterone Receptor, HER-2, and P53 in Patients with Mutations in BRCA1 and BRCA2." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 20 (9): 2310–18. https://doi.org/10.1200/JCO.2002.09.023.

Lazard, Tristan, Guillaume Bataillon, Peter Naylor, Tatiana Popova, François-Clément Bidard, Dominique Stoppa-Lyonnet, Marc-Henri Stern, Etienne Decencière, Thomas Walter, and Anne Vincent-Salomon. 2022. "Deep Learning Identifies Morphological Patterns of Homologous Recombination Deficiency in Luminal

Breast Cancers from Whole Slide Images." *Cell Reports. Medicine* 3 (12): 100872. https://doi.org/10.1016/j.xcrm.2022.100872.

Lazard, Tristan, Marvin Lerousseau, Etienne Decencière, and Thomas Walter. 2023. "Giga-SSL: Self-Supervised Learning for Gigapixel Images." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4304–13.

Lerousseau, Marvin, Maria Vakalopoulou, Eric Deutsch, and Nikos Paragios. 2021. "SparseConvMIL: Sparse Convolutional Context-Aware Multiple Instance Learning for Whole Slide Image Classification." arXiv. https://doi.org/10.48550/arXiv.2105.02726.

Li, Bin, Yin Li, and Kevin W. Eliceiri. 2021. "Dual-Stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning." In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14313–23. Nashville, TN, USA: IEEE. https://doi.org/10.1109/CVPR46437.2021.01409.

Li, Jiahui, Wen Chen, Xiaodi Huang, Zhiqiang Hu, Qi Duan, Hongsheng Li, Dimitris N. Metaxas, and Shaoting Zhang. 2021. "Hybrid Supervision Learning for Pathology Whole Slide Image Classification." *arXiv:2107.00934 [Cs]*, October. https://arxiv.org/abs/2107.00934.

Li, Zhe, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. 2018. "Thoracic Disease Identification and Localization with Limited Supervision." arXiv. https://doi.org/10.48550/arXiv.1711.06373.

Lin, Tiancheng, Zhimiao Yu, Zengchao Xu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. 2023. "SGCL: Spatial Guided Contrastive Learning on Whole-Slide Pathological Images." *Medical Image Analysis* 89 (October): 102845. https://doi.org/10.1016/j.media.2023.102845.

Livasy, Chad A., Gamze Karaca, Rita Nanda, Maria S. Tretiakova, Olufunmilayo I. Olopade, Dominic T. Moore, and Charles M. Perou. 2006. "Phenotypic Evaluation of the Basal-Like Subtype of Invasive Breast Carcinoma." *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc* 19 (2): 264–71. https://doi.org/10.1038/modpathol.3800528.

Lu, Ming Y., Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, et al. 2023. "Towards a Visual-Language Foundation Model for Computational Pathology." arXiv. https://doi.org/10.48550/arXiv.2307.12914.

Lu, Ming Y., Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. 2020. "Data Efficient and Weakly Supervised Computational Pathology on Whole Slide Images." *arXiv:2004.09666 [Cs, Eess, q-Bio]*, April. https://arxiv.org/abs/2004.09666.

———. 2021. "Data-Efficient and Weakly Supervised Computational Pathology on Whole-Slide Images." *Nature Biomedical Engineering*, March, 1–16. https://doi.org/10.1038/s41551-020-00682-w.

Lubrano Di Scandalea, Melanie, Tristan Lazard, Guillaume Balezo, Yaëlle Bellahsen-Harrar, Cécile Badoual, Sylvain Berlemont, and Thomas Walter. 2022. "Automatic Grading of Cervical Biopsies by Combining Full and Self-Supervision." https://doi.org/10.1101/2022.01.14.476330.

Lubrano, Mélanie, Yaëlle Bellahsen-Harrar, Rutger Fick, Cécile Badoual, and Thomas Walter. n.d. "Simple and Efficient Confidence Score for Grading Whole Slide Images."

Lundberg, Scott, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." arXiv. https://doi.org/10.48550/arXiv.1705.07874.

Macenko, Marc, Marc Niethammer, J. S. Marron, David Borland, John T. Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E. Thomas. 2009. "A Method for Normalizing Histology Slides for Quantitative Analysis." In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 1107–10. Boston, MA, USA: IEEE. https://doi.org/10.1109/ISBI.2009.5193250.

Malon, Christopher, Elena Brachtel, Eric Cosatto, Hans Peter Graf, Atsushi Kurata, Masahiko Kuroda, John S. Meyer, Akira Saito, Shulin Wu, and Yukako Yagi. 2012. "Mitotic Figure Recognition: Agreement Among Pathologists and Computerized Detector." *Analytical Cellular Pathology (Amsterdam)* 35 (2): 97–100. https://doi.org/10.3233/ACP-2011-0029.

Manié, Elodie, Tatiana Popova, Aude Battistella, Julien Tarabeux, Virginie Caux-Moncoutier, Lisa Golmard, Nicholas K. Smith, et al. 2016. "Genomic Hallmarks of Homologous Recombination Deficiency in Invasive Breast Carcinomas." *International Journal of Cancer* 138 (4): 891–900. https://doi.org/10.1002/ijc.29829.

Maron, Oded, and Tomas Lozano-Perez. n.d. "A Framework for Multiple-Instance Learning," 7.

Martin-Serrano, Miguel A., Benjamin Kepecs, Miguel Torres-Martin, Emily R. Bramel, Philipp K. Haber, Elliot Merritt, Alexander Rialdi, et al. 2022. "Novel Microenvironment-Based Classification of Intrahepatic Cholangiocarcinoma with Therapeutic Implications." *Gut*, May, gutjnl-2021-326514. https://doi.org/10.1136/gutjnl-2021-326514.

Mavros, Michael N., Konstantinos P. Economopoulos, Vangelis G. Alexiou, and Timothy M. Pawlik. 2014. "Treatment and Prognosis for Patients With Intrahepatic Cholangiocarcinoma: Systematic Review and Meta-analysis." *JAMA Surgery* 149 (6): 565–74. https://doi.org/10.1001/jamasurg.2013.5137.

Mehrtens, Hendrik A., Alexander Kurz, Tabea-Clara Bucher, and Titus J. Brinker. 2023. "Benchmarking Common Uncertainty Estimation Methods with Histopathological Images Under Domain Shift and Label Noise." *Medical Image Analysis* 89 (October): 102914. https://doi.org/10.1016/j.media.2023.102914.

Miller, R. E., A. Leary, C. L. Scott, V. Serra, C. J. Lord, D. Bowtell, D. K. Chang, et al. 2020. "ESMO Recommendations on Predictive Biomarker Testing for Homologous Recombination Deficiency and PARP Inhibitor Benefit in Ovarian Cancer." *Annals of Oncology: Official Journal of the European Society for Medical Oncology* 31 (12): 1606–22. https://doi.org/10.1016/j.annonc.2020.08.2102.

Misra, Ishan, and Laurens van der Maaten. 2019. "Self-Supervised Learning of Pretext-Invariant Representations." *arXiv:1912.01991 [Cs]*, December. https://arxiv.org/abs/1912.01991.

Mlynarski, Pawel, Hervé Delingette, Antonio Criminisi, and Nicholas Ayache. 2019. "Deep Learning with Mixed Supervision for Brain Tumor Segmentation." *Journal of Medical Imaging* 6 (03): 1. https://doi.org/10.1117/1.JMI.6.3.034002.

Mobadersany, Pooya, Safoora Yousefi, Mohamed Amgad, David A. Gutman, Jill S. Barnholtz-Sloan, José E. Velázquez Vega, Daniel J. Brat, and Lee A. D. Cooper.

2018. “Predicting Cancer Outcomes from Histology and Genomics Using Convolutional Networks.” *Proceedings of the National Academy of Sciences* 115 (13): E2970–79. https://doi.org/10.1073/pnas.1717139115.

Moeini, Agrin, Daniela Sia, Nabeel Bardeesy, Vincenzo Mazzaferro, and Josep M. Llovet. 2016. “Molecular Pathogenesis and Targeted Therapies for Intrahepatic Cholangiocarcinoma.” *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 22 (2): 291–300. https://doi.org/10.1158/1078-0432.CCR-14-3296.

Nahhas, Omar S. M. El, Chiara M. L. Loeffler, Zunamys I. Carrero, Marko van Treeck, Fiona R. Kolbinger, Katherine J. Hewitt, Hannah S. Muti, et al. 2023. “Regression-Based Deep-Learning Predicts Molecular Biomarkers from Pathology Slides.” arXiv. https://doi.org/10.48550/arXiv.2304.05153.

Naylor, Peter, Tristan Lazard, Guillaume Bataillon, Marick Lae, Anne Vincent-Salomon, Anne-Sophie Hamy, Fabien Reyal, and Thomas Walter. 2022. “Neural Network for the Prediction of Treatment Response in Triple Negative Breast Cancer *.” bioRxiv. https://doi.org/10.1101/2022.01.31.478433.

Nguyen, Anh, Jason Yosinski, and Jeff Clune. 2019. “Understanding Neural Networks via Feature Visualization: A Survey.” arXiv. https://doi.org/10.48550/arXiv.1904.08939.

Noroozi, Mehdi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. 2018. “Boosting Self-Supervised Learning via Knowledge Transfer.” arXiv. https://doi.org/10.48550/arXiv.1805.00385.

Oh, Do-Youn, Aiwu Ruth He, Shukui Qin, Li-Tzong Chen, Takuji Okusaka, Arndt Vogel, Jin Won Kim, et al. 2022. “Durvalumab Plus Gemcitabine and Cisplatin in Advanced Biliary Tract Cancer.” *NEJM Evidence* 1 (8): EVIDoa2200015. https://doi.org/10.1056/EVIDoa2200015.

“On the Nature and Structural Characteristics of Cancer, and of Those Morbid Growths Which May Be Confounded with It.” 1840. *The Medico-Chirurgical Review* 33 (65): 119–48.

Oner, Mustafa Umit, Jared Marc Song Kye-Jet, Hwee Kuan Lee, and Wing-Kin Sung. 2023. “Distribution Based MIL Pooling Filters: Experiments on a Lymph Node Metastases Dataset.” *Medical Image Analysis*, April, 102813. https://doi.org/10.1016/j.media.2023.102813.

Pantanowitz, Liron, Gabriela M. Quiroga-Garza, Lilach Bien, Ronen Heled, Daphna Laifenfeld, Chaim Linhart, Judith Sandbank, et al. 2020. “An Artificial Intelligence Algorithm for Prostate Cancer Diagnosis in Whole Slide Images of Core Needle Biopsies: A Blinded Clinical Validation and Deployment Study.” *The Lancet. Digital Health* 2 (8): e407–16. https://doi.org/10.1016/S2589-7500(20)30159-X.

Pantanowitz, Liron, Paul N. Valenstein, Andrew J. Evans, Keith J. Kaplan, John D. Pfeifer, David C. Wilbur, Laura C. Collins, and Terence J. Colgan. 2011. “Review of the Current State of Whole Slide Imaging in Pathology.” *Journal of Pathology Informatics* 2 (1): 36. https://doi.org/10.4103/2153-3539.83746.

Pedregosa, Fabian, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. n.d. “Scikit-Learn: Machine Learning in Python.” *MACHINE LEARNING IN PYTHON*, 6.

Polak, Paz, Jaegil Kim, Lior Z. Braunstein, Rosa Karlic, Nicholas J. Haradhavala, Grace Tiao, Daniel Rosebrock, et al. 2017. "A Mutational Signature Reveals Alterations Underlying Deficient Homologous Recombination Repair in Breast Cancer." *Nature Genetics* 49 (10): 1476–86. https://doi.org/10.1038/ng.3934.

Popova, T., E. Manie, G. Rieunier, V. Caux-Moncoutier, C. Tirapo, T. Dubois, O. Delattre, et al. 2012. "Ploidy and Large-Scale Genomic Instability Consistently Identify Basal-like Breast Carcinomas with BRCA1/2 Inactivation." *Cancer Research* 72 (21): 5454–62. https://doi.org/10.1158/0008-5472.CAN-12-1470.

Qu, Hui, Mu Zhou, Zhennan Yan, He Wang, Vinod K. Rustgi, Shaoting Zhang, Olivier Gevaert, and Dimitris N. Metaxas. 2021. "Genetic Mutation and Biological Pathway Prediction Based on Whole Slide Images in Breast Carcinoma Using Deep Learning." *Npj Precision Oncology* 5 (1): 87. https://doi.org/10.1038/s41698-021-00225-9.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. 2021. "Learning Transferable Visual Models From Natural Language Supervision." arXiv. https://doi.org/10.48550/arXiv.2103.00020.

Rahnemai-Azar, Amir A., Allison Weisbrod, Mary Dillhoff, Carl Schmidt, and Timothy M. Pawlik. 2017. "Intrahepatic Cholangiocarcinoma: Molecular Markers for Diagnosis and Prognosis." *Surgical Oncology* 26 (2): 125–37. https://doi.org/10.1016/j.suronc.2016.12.009.

Rakha, Emad A., Maysa E. El-Sayed, Jorge Reis-Filho, and Ian O. Ellis. 2009. "Patho-Biological Aspects of Basal-Like Breast Cancer." *Breast Cancer Research and Treatment* 113 (3): 411–22. https://doi.org/10.1007/s10549-008-9952-1.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier." arXiv. https://doi.org/10.48550/arXiv.1602.04938.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation." *arXiv:1505.04597 [Cs]*, May. https://arxiv.org/abs/1505.04597.

Ruifrok, Arnout C. n.d. "Quantification of Histochemical Staining by Color Deconvolution," 21.

Rymarczyk, Dawid, Jacek Tabor, and Bartosz Zieliński. 2020. "Kernel Self-Attention in Deep Multiple Instance Learning." *arXiv:2005.12991 [Cs, Stat]*, May. https://arxiv.org/abs/2005.12991.

Saillard, Charlie, Olivier Dehaene, Tanguy Marchand, Olivier Moindrot, Aurélie Kamoun, Benoit Schmauch, and Simon Jegou. 2021. "Self Supervised Learning Improves dMMR/MSI Detection from Histology Slides Across Multiple Cancers." *arXiv:2109.05819 [Cs, Eess]*, September. https://arxiv.org/abs/2109.05819.

Saillard, Charlie, Benoit Schmauch, Oumeima Laifa, Matahi Moarii, Sylvain Toldo, Mikhail Zaslavskiy, Elodie Pronier, et al. 2020. "Predicting Survival After Hepatocellular Carcinoma Resection Using Deep Learning on Histological Slides." *Hepatology (Baltimore, Md.)* 72 (6): 2000–2013. https://doi.org/10.1002/hep.31207.

Schirris, Yoni, Efstratios Gavves, Iris Nederlof, Hugo Mark Horlings, and Jonas Teuwen. 2021. "DeepSMILE: Self-supervised Heterogeneity-Aware Multiple Instance Learning for DNA Damage Response Defect Classification Directly from H&E Whole-Slide Images." arXiv. https://doi.org/10.48550/arXiv.2107.09405.

Schmauch, Benoît, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maillé, Julien Calderaro, Aurélie Kamoun, et al. 2020. "A Deep Learning Model to Predict RNA-Seq Expression of Tumours from Whole Slide Images." *Nature Communications* 11 (1): 3877. https://doi.org/10.1038/s41467-020-17678-4.

Sellors, John W., and R. Sankaranarayanan. 2003. *Colposcopy and Treatment of Cervical Intraepithelial Neoplasia: A Beginners' Manual*. Lyon: Intern. Agency for Research Cancer.

Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2020. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization." *International Journal of Computer Vision* 128 (2): 336–59. https://doi.org/10.1007/s11263-019-01228-7.

Shao, Zhuchen, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. 2021. "TransMIL: Transformer Based Correlated Multiple Instance Learning for Whole Slide Image Classification." arXiv. https://arxiv.org/abs/2106.00908.

Sharma, Yash, Lubaina Ehsan, Sana Syed, and Donald E. Brown. 2021. "Histo-Transfer: Understanding Transfer Learning for Histopathology." arXiv. https://doi.org/10.48550/arXiv.2106.07068.

Shi, Jie-Yi, Xiaodong Wang, Guang-Yu Ding, Zhou Dong, Jing Han, Zehui Guan, Li-Jie Ma, et al. 2021. "Exploring Prognostic Indicators in the Pathological Images of Hepatocellular Carcinoma Based on Deep Learning." *Gut* 70 (5): 951–61. https://doi.org/10.1136/gutjnl-2020-320930.

Shrikumar, Avanti, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2017. "Not Just a Black Box: Learning Important Features Through Propagating Activation Differences." arXiv. https://doi.org/10.48550/arXiv.1605.01713.

Sia, Daniela, Yujin Hoshida, Augusto Villanueva, Sasan Roayaie, Joana Ferrer, Barbara Tabak, Judit Peix, et al. 2013. "Integrative Molecular Analysis of Intrahepatic Cholangiocarcinoma Reveals 2 Classes That Have Different Outcomes." *Gastroenterology* 144 (4): 829–40. https://doi.org/10.1053/j.gastro.2013.01.001.

Smilkov, Daniel, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. "SmoothGrad: Removing Noise by Adding Noise." arXiv. https://doi.org/10.48550/arXiv.1706.03825.

Sohn, Kihyuk. 2016. "Improved Deep Metric Learning with Multi-class N-pair Loss Objective." In *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc.

Srinivas, Sampath. 2013. "A Generalization of the Noisy-Or Model." arXiv. https://doi.org/10.48550/arXiv.1303.1479.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. 2017. "Axiomatic Attribution for Deep Networks." arXiv. https://doi.org/10.48550/arXiv.1703.01365.

Tartaglione, Enzo, Carlo Alberto Barbano, and Marco Grangetto. 2021. "EnD: Entangling and Disentangling Deep Representations for Bias Correction." In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13503–12. https://doi.org/10.1109/CVPR46437.2021.01330.

Thorsson, Vésteinn, David L. Gibbs, Scott D. Brown, Denise Wolf, Dante S. Bortone, Tai-Hsien Ou Yang, Eduard Porta-Pardo, et al. 2018. "The Immune Landscape of Cancer." *Immunity* 48 (4): 812–830.e14. https://doi.org/10.1016/j.immuni.2018.03.023.

Tian, Yonglong, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. "What Makes for Good Views for Contrastive Learning?" *arXiv:2005.10243 [Cs]*, December. https://arxiv.org/abs/2005.10243.

Tourniaire, Paul, Marius Ilie, Paul Hofman, Nicholas Ayache, and Herve Delingette. 2021. "Attention-Based Multiple Instance Learning with Mixed Supervision on the Camelyon16 Dataset," 11.

Tu, Han-Hsing, and Hsuan-Tien Lin. n.d. "One-Sided Support Vector Regression for Multiclass Cost-sensitive Classification," 8.

Tung, Nadine M., Mark E. Robson, Steffen Ventz, Cesar A. Santa-Maria, Rita Nanda, Paul K. Marcom, Payal D. Shah, et al. 2020. "TBCRC 048: Phase II Study of Olaparib for Metastatic Breast Cancer and Mutations in Homologous Recombination-Related Genes." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 38 (36): 4274–82. https://doi.org/10.1200/JCO.20.02151.

Turner, Nicholas C. 2017. "Signatures of DNA-Repair Deficiencies in Breast Cancer." *The New England Journal of Medicine* 377 (25): 2490–92. https://doi.org/10.1056/NEJMcibr1710161.

Tutt, Andrew N. J., Judy E. Garber, Bella Kaufman, Giuseppe Viale, Debora Fumagalli, Priya Rastogi, Richard D. Gelber, et al. 2021. "Adjuvant Olaparib for Patients with BRCA1- or BRCA2-Mutated Breast Cancer." *New England Journal of Medicine*, June. https://doi.org/10.1056/NEJMoa2105215.

Tutt, Andrew, Holly Tovey, Maggie Chon U. Cheang, Sarah Kernaghan, Lucy Kilburn, Patrycja Gazinska, Julie Owen, et al. 2018. "Carboplatin in BRCA1/2-Mutated and Triple-Negative Breast Cancer BRCAness Subgroups: The TNT Trial." *Nature Medicine* 24 (5): 628–37. https://doi.org/10.1038/s41591-018-0009-7.

Valieris, Renan, Lucas Amaro, Cynthia Aparecida Bueno de Toledo Osório, Adriana Passos Bueno, Rafael Andres Rosales Mitrowsky, Dirce Maria Carraro, Diana Noronha Nunes, Emmanuel Dias-Neto, and Israel Tojal da Silva. 2020. "Deep Learning Predicts Underlying Features on Pathology Images with Therapeutic Relevance for Breast and Gastric Cancer." *Cancers* 12 (12): 3687. https://doi.org/10.3390/cancers12123687.

Varoquaux, Gaël, Pradeep Reddy Raamana, Denis Engemann, Andrés Hoyos-Idrobo, Yannick Schwartz, and Bertrand Thirion. 2017. "Assessing and Tuning Brain Decoders: Cross-Validation, Caveats, and Guidelines." *NeuroImage* 145 (January): 166–79. https://doi.org/10.1016/j.neuroimage.2016.10.038.

Veta, Mitko, Paul J. van Diest, Stefan M. Willems, Haibo Wang, Anant Madabhushi, Angel Cruz-Roa, Fabio Gonzalez, et al. 2015. "Assessment of Algorithms for Mitosis Detection in Breast Cancer Histopathology Images." *Medical Image Analysis* 20 (1): 237–48. https://doi.org/10.1016/j.media.2014.11.010.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, et al. 2020. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." *Nature Methods* 17 (3): 261–72. https://doi.org/10.1038/s41592-019-0686-2.

Vorontsov, Eugene, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqi Liu, Philippe Mathieu, et al. 2023. "Virchow: A Million-Slide Digital Pathology Foundation Model." arXiv. https://doi.org/10.48550/arXiv.2309.07778.

Wang, Tianlu, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. "Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations." *arXiv:1811.08489 [Cs]*, October. https://arxiv.org/abs/1811.08489.

Wang, Xiyue, De Cai, Sen Yang, Yiming Cui, Junyou Zhu, Kanran Wang, and Junhan Zhao. 2023. "SAC-Net: Enhancing Spatiotemporal Aggregation in Cervical Histological Image Classification via Label-Efficient Weakly Supervised Learning." *IEEE Transactions on Circuits and Systems for Video Technology*, 1–1. https://doi.org/10.1109/TCSVT.2023.3294938.

Wang, Xiyue, Jinxi Xiang, Jun Zhang, Sen Yang, Zhongyi Yang, Minghui Wang, Wei Yang, Junzhou Huang, and Xiao Han. n.d. "SCL-WC: Cross-Slide Contrastive Learning for Weakly-Supervised Whole-Slide Image Classification."

Wang, Xiyue, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. 2022. "Transformer-Based Unsupervised Contrastive Learning for Histopathological Image Classification." *Medical Image Analysis* 81 (October): 102559. https://doi.org/10.1016/j.media.2022.102559.

Wang, Zeyu, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. 2020. "Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation." *arXiv:1911.11834 [Cs]*, April. https://arxiv.org/abs/1911.11834.

Weinstein, John N., Eric A. Collisson, Gordon B. Mills, Kenna M. Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. 2013. "The Cancer Genome Atlas Pan-Cancer Analysis Project." *Nature Genetics* 45 (10): 1113–20. https://doi.org/10.1038/ng.2764.

Weiss, Karl, Taghi M. Khoshgoftaar, and DingDing Wang. 2016. "A Survey of Transfer Learning." *Journal of Big Data* 3 (1): 9. https://doi.org/10.1186/s40537-016-0043-6.

Wulczyn, Ellery, David F. Steiner, Melissa Moran, Markus Plass, Robert Reihs, Fraser Tan, Isabelle Flament-Auvigne, et al. 2021. "Interpretable Survival Prediction for Colorectal Cancer Using Deep Learning." *Npj Digital Medicine* 4 (1): 1–13. https://doi.org/10.1038/s41746-021-00427-2.

Xiang, Jinxi, and Jun Zhang. 2022. "Exploring Low-Rank Property in Multiple Instance Learning for Whole Slide Image Classification." In *The Eleventh International Conference on Learning Representations*.

Xu, Bolei, Jingxin Liu, Xianxu Hou, Bozhi Liu, Jon Garibaldi, Ian O. Ellis, Andy Green, Linlin Shen, and Guoping Qiu. 2019. "Look, Investigate, and Classify: A Deep Hybrid Attention Method for Breast Cancer Classification." *arXiv:1902.10946 [Cs]*, February. https://arxiv.org/abs/1902.10946.

Xu, Feng, Chuang Zhu, Wenqi Tang, Ying Wang, Yu Zhang, Jie Li, Hongchuan Jiang, Zhongyue Shi, Jun Liu, and Mulan Jin. 2021. "Predicting Axillary Lymph Node Metastasis in Early Breast Cancer Using Deep Learning on Primary Tumor Biopsy Slides." *Frontiers in Oncology* 11: 759007. https://doi.org/10.3389/fonc.2021.759007.

Yamashita, Rikiya, Jin Long, Atif Saleem, Daniel L. Rubin, and Jeanne Shen. 2021. "Deep Learning Predicts Postsurgical Recurrence of Hepatocellular Carcinoma from Digital Histopathologic Images." *Scientific Reports* 11 (1): 2047. https://doi.org/10.1038/s41598-021-81506-y.

Yang, Zhongyi, Xiyue Wang, Jinxi Xiang, Jun Zhang, Sen Yang, Xinran Wang, Wei Yang, Zhongyu Li, Xiao Han, and Yueping Liu. 2023. "The Devil Is in the Details: A Small-Lesion Sensitive Weakly Supervised Learning Framework for Prostate Cancer Detection and Grading." *Virchows Archiv* 482 (3): 525–38. https://doi.org/10.1007/s00428-023-03502-z.

Yeh, Chih-Kuan, Been Kim, Sercan O. Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. 2020. "On Completeness-aware Concept-Based Explanations in Deep Neural Networks." *arXiv:1910.07969 [Cs, Stat]*, June. https://arxiv.org/abs/1910.07969.

Yu, Jin-Gang, Zihao Wu, Yu Ming, Shule Deng, Yuanqing Li, Caifeng Ou, Chunjiang He, Baiye Wang, Pusheng Zhang, and Yu Wang. 2023. "Prototypical Multiple Instance Learning for Predicting Lymph Node Metastasis of Breast Cancer from Whole-Slide Pathological Images." *Medical Image Analysis*, January, 102748. https://doi.org/10.1016/j.media.2023.102748.

Yuan, Yinyin. 2015. "Modelling the Spatial Heterogeneity and Molecular Correlates of Lymphocytic Infiltration in Triple-Negative Breast Cancer." *Journal of The Royal Society Interface* 12 (103): 20141153. https://doi.org/10.1098/rsif.2014.1153.

Zaffar, Imaad, Guillaume Jaume, Nasir Rajpoot, and Faisal Mahmood. 2022. "Embedding Space Augmentation for Weakly Supervised Learning in Whole-Slide Images." arXiv. https://doi.org/10.48550/arXiv.2210.17013.

Zanjani, Farhad Ghazvinian, Svitlana Zinger, and Babak E Bejnordi. n.d. "Histopathology Stain-Color Normalization Using Deep Generative Models," 11.

Zemni, Mehdi, Mickaël Chen, Éloi Zablocki, Hédi Ben-Younes, Patrick Pérez, and Matthieu Cord. 2023. "OCTET: Object-aware Counterfactual Explanations." arXiv. https://doi.org/10.48550/arXiv.2211.12380.

Zeng, Qinghe, Christophe Klein, Stefano Caruso, Pascale Maille, Narmin Ghaffari Laleh, Daniele Sommacale, Alexis Laurent, et al. 2022. "Artificial Intelligence Predicts Immune and Inflammatory Gene Signatures Directly from Hepatocellular Carcinoma Histology." *Journal of Hepatology* 77 (1): 116–27. https://doi.org/10.1016/j.jhep.2022.01.018.

Zhang, Chaoning, Kang Zhang, Chenshuang Zhang, Trung X. Pham, Chang D. Yoo, and In So Kweon. 2022. "How Does SimSiam Avoid Collapse Without Negative Samples? A Unified Understanding with Self-supervised Contrastive Learning." arXiv. https://doi.org/10.48550/arXiv.2203.16262.

Zhang, Chuyan, Hao Zheng, and Yun Gu. 2023. "Dive into the Details of Self-Supervised Learning for Medical Image Analysis." *Medical Image Analysis* 89 (October): 102879. https://doi.org/10.1016/j.media.2023.102879.

Zhang, Hongrun, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E. Coupland, and Yalin Zheng. 2022. "DTFD-MIL: Double-Tier Feature Distillation Multiple Instance Learning for Histopathology Whole Slide Image Classification." arXiv. https://arxiv.org/abs/2203.12081.

Zhang, Richard, Phillip Isola, and Alexei A. Efros. 2016. "Colorful Image Colorization." arXiv. https://doi.org/10.48550/arXiv.1603.08511.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. "Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-level Constraints." *arXiv:1707.09457 [Cs, Stat]*, July. https://arxiv.org/abs/1707.09457.

Zhao, Qingyu, Ehsan Adeli, and Kilian M. Pohl. 2020. "Training Confounder-Free Deep Learning Models for Medical Applications." *Nature Communications* 11 (1): 6010. https://doi.org/10.1038/s41467-020-19784-9.
Zhou, Zhi-Hua. 2018. "A Brief Introduction to Weakly Supervised Learning." *National Science Review* 5 (1): 44–53. https://doi.org/10.1093/nsr/nwx106.

Appendix - List of contributions

A

Publications

- **Lazard, T.** *, Bataillon, G.* et al. (2022). Deep learning identifies morphological patterns of homologous recombination deficiency in luminal breast cancers from whole slide images. Cell Rep Med 3, 100872. 10.1016/j.xcrm.2022.100872.
- M. Lubrano, **T. Lazard**, et al. Automatic Grading of Cervical Biopsies by Combining Full and Self-supervision. 13807, Springer Nature Switzerland, pp.408-423, 2023, Lecture Notes in Computer Science, 10.1007/978-3-031-25082-8_27.
- **Lazard, T.**, Lerousseau, M., Decencière, E., and Walter, T. (2023). Giga-SSL: Self-Supervised Learning for Gigapixel Images. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 2023 pp. 4305-4314. 10.1109/CVPRW59228.2023.00453
- **Lazard, T.**, et al. (2023). Democratizing Whole Slide Images: optimized representations for The Cancer Genome Atlas. *under submission*.
- A. Beaufrère*, **T. Lazard***, et. al. Self-supervised learning for predicting transcriptomic classes on whole slides images in intrahepatic cholangiocarcinoma, *under submission*.

Communications

- "Giga-SSL: self-supervised learning for giga-pixel images", June 2023, Computer Vision for Microscopy Image analysis (CVMI) Workshop, CVPR, Vancouver (Canada).
- "Prediction of Homologous Recombination Deficiency from breast cancer WSI", BioImage Informatics 2023, Pasteur Institute, Paris, France.

Open-source repository

- WSI-MIL: perform MIL classification and interpret them.
- Giga-SSL: package to perform Giga-SSL training.
- Reproduce the VisioMel results of Giga-SSL

Achievements

- Third place at the VisioMel challenge (cash prize: 5000$).

Appendix - Chapt. II.

B

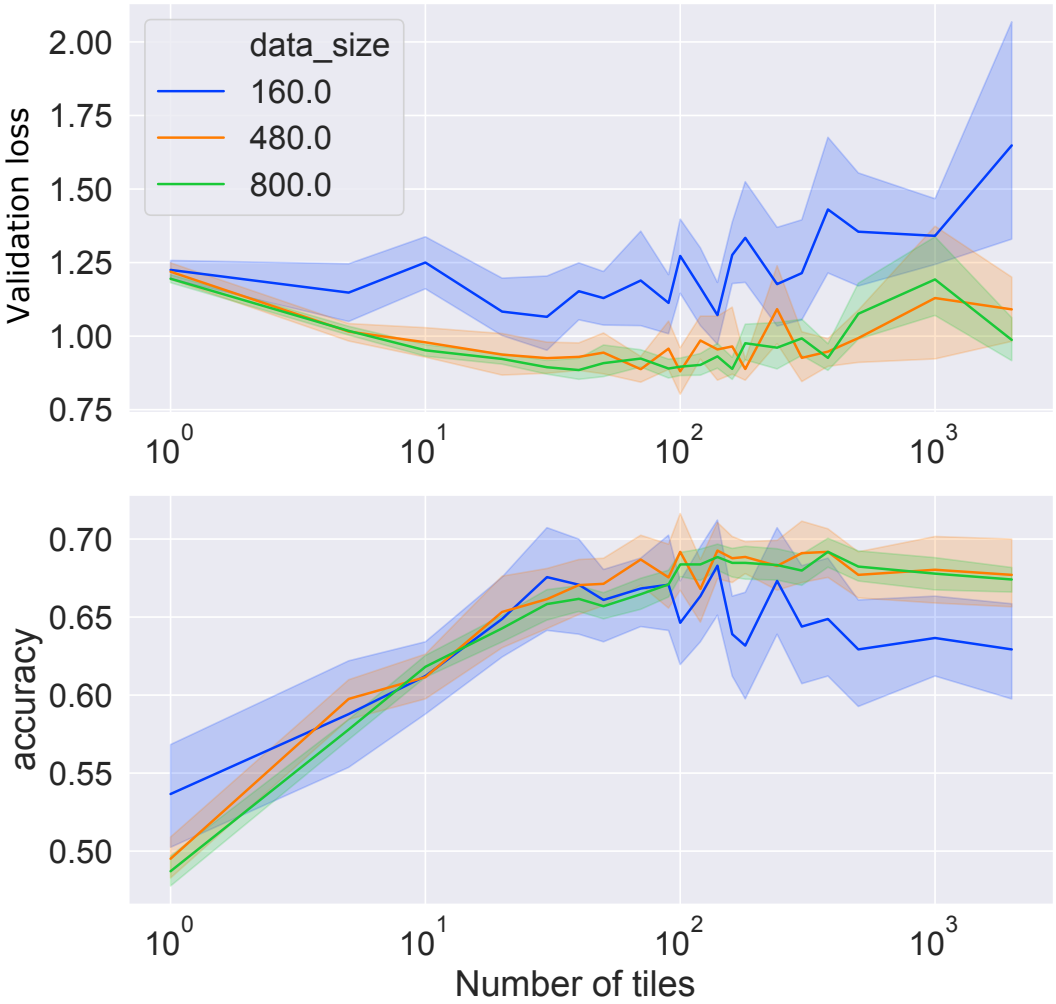


Figure B.1.: Effect of Number of Sampled Tiles per WSI on Training. Using the dataset from Chapter IV, I sampled 160, 480, and 800 WSIs, preserving output variable proportions. I trained attention-MIL models (from Chapter III) on these datasets using 1 to 1000 tiles per WSI. Validation accuracy initially increases with more tiles but plateaus around 100 tiles. On smaller datasets, loss and accuracy decline after this plateau, indicating overfitting. Tile sub-sampling may help regularize MIL networks.

Appendix - Chapt. III.

C

Architecture	AUC		F_1		B_{acc}	
	mean	std	mean	std	mean	std
AVG -IT	0.81	0.07	0.69	0.06	0.69	0.07
AVG +IT	0.79	0.09	0.71	0.07	0.7	0.07
ARGMAX -IT	0.58	0.12	0.58	0.06	0.58	0.04
ARGMAX +IT	0.65	0.07	0.48	0.03	0.52	0.017
MAX -IT	0.58	0.04	0.43	0.005	0.5	0
MAX +IT	0.82	0.06	0.68	0.06	0.67	0.05
K-RANK -IT	0.71	0.13	0.6	0.1	0.60	0.07
K-RANK +IT	0.75	0.06	0.56	0.09	0.58	0.05
ILSE -IT (ours)	0.83	0.07	0.72	0.06	0.72	0.06
ILSE +IT	0.81	0.05	0.74	0.03	0.73	0.05
CLAM	0.80	0.05	0.66	0.12	0.68	0.12

Table C.1.: MIL models benchmark. related to Figure III.3 and Table III.1. Benchmarking of different models for the prediction of the HRD on the Curie luminal dataset. AVG: average of the tile encodings; MAX: element-wise maximum of the file encodings; ARGMAX: selects the most attended tile. K-RANK: selects the top-k most attended tiles and bottom-k least attended tiles and stack them. ILSE: encodings are mapped to an attention score; the slide encoding is the weighted sum of the tile encodings, where the attention scores are the weights [1]. CLAM: the current state of the art weakly supervised WSI classification algorithm [2]. For each architecture, we considered two variants: a version where instance representations are first mapped to a lower dimensional vector (128) with a Multi-Layer-Perceptron (MLP), as proposed by llse et al., and another version where this is omitted. This mapping is referred to as "Instance Transformation" (IT) and we note +IT the versions with instance transformation, and -IT the versions without. Ilse performed best, followed closely by AVG -IT and MAX +IT; but unlike them, is interpretable through attention and decision scores. B_{acc} stands for balanced accuracy. F_1 reported is the average of the F_1 for both classes.

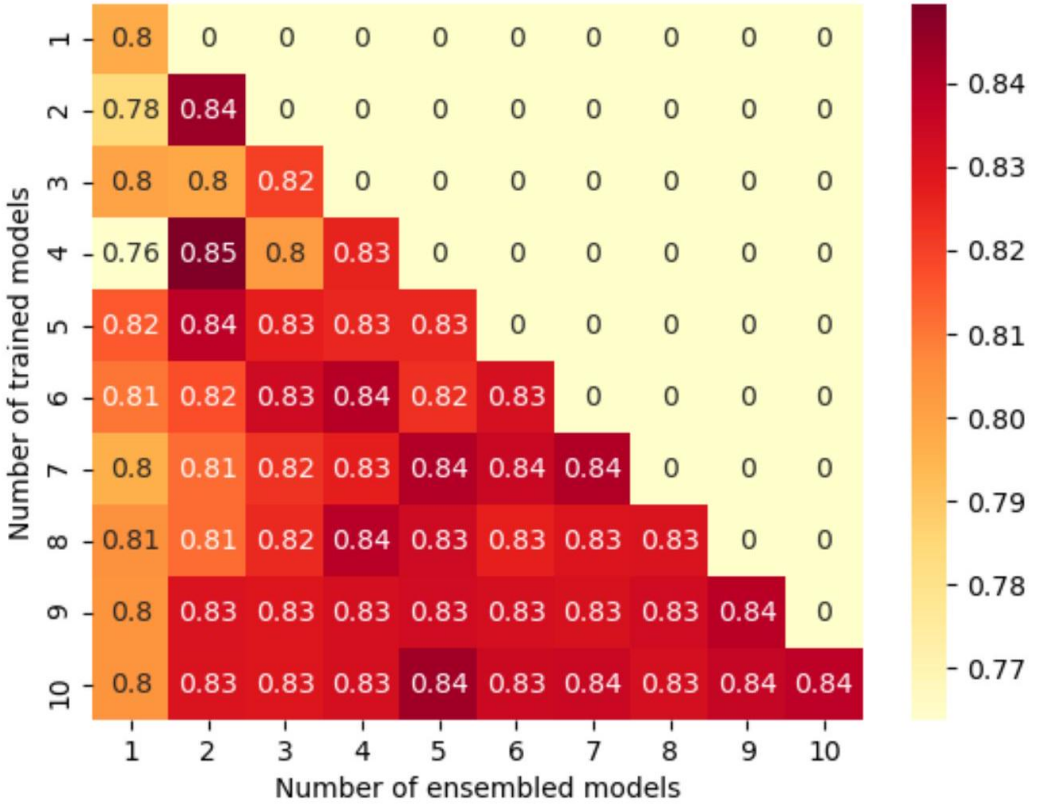


Figure C.1.: Influence of ensembling. Related to the Results section and Table III.1. 10 models are trained for each test fold, on the Curie-Luminal dataset. Among these 10 models, we can randomly sample n of them and compute the performances of the ensembling of the e best models among this selection, according to the validation metrics. Average test-AUC-ROC performances of the pairs (n, e) of trained and ensembled models $(n >= e)$ are reported here: the number at line i and column j correspond to the average test AUC performance of the ensemble of the j among i best performing models. Training several models and ensembling the best performers allows a gain between 3 to 5 AUC points. We fixed $(n, e) = (10, 3)$.

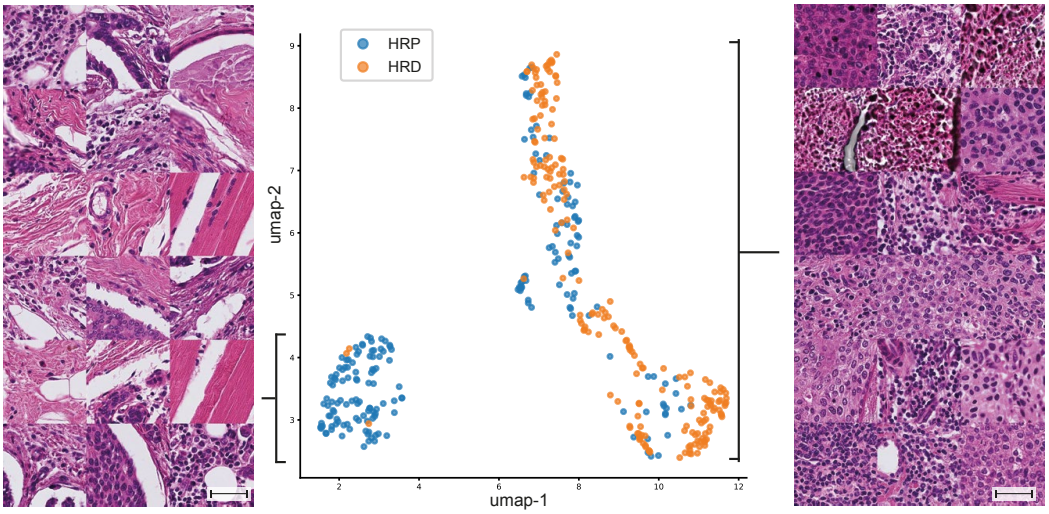


Figure C.2.: Attention based visualization. Related to the visualization section and Figure III.5. The 200 most attended tiles are extracted from the 20 slides predicted with the highest probability as being HRD or HRP. Points are here the umap projection of the embeddings of each of these 200 tiles, their color is the label of their slides of origin. We set the maximum number of extracted tiles per slides at 20 (like in Figure 3). Although a clear morphological cluster associated to HRP emerges, corresponding to cluster 6 of Figure III.5, the tiles associated to HRD do not cluster apart from other tiles associated with HRP. This suggest that these tiles may be predictive for HRD but still be present in HRP slides, and make the ABV method difficult to interpret. Scale bar $= 50\mu\mathrm{m}$.

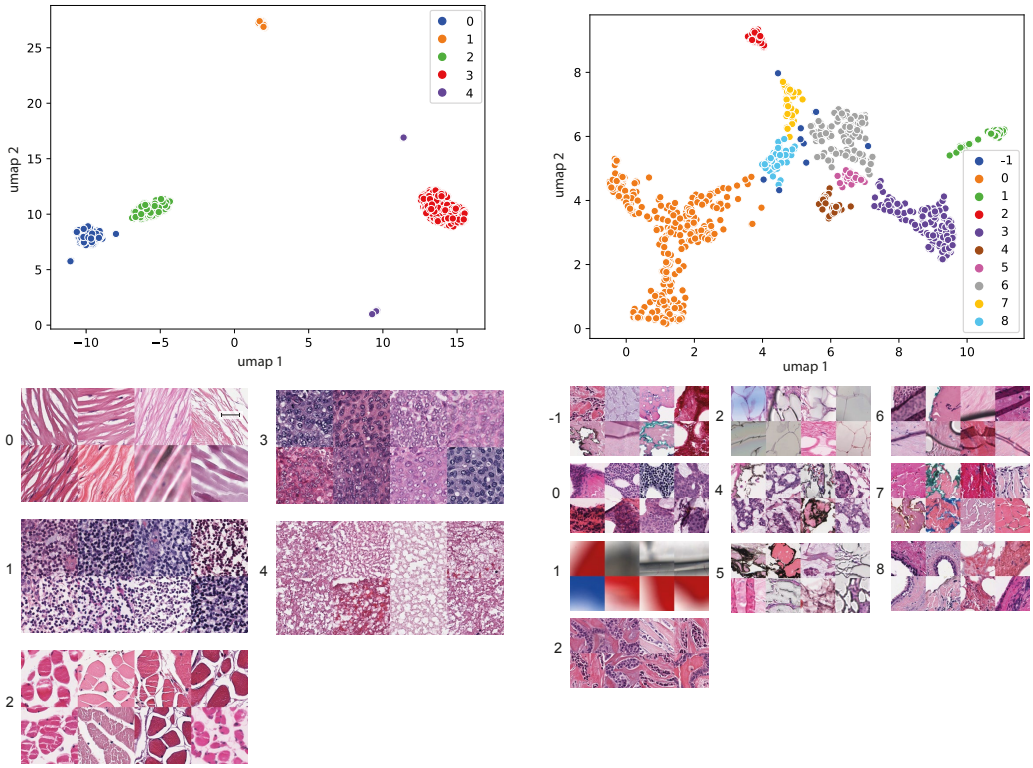


Figure C.3.: Morphological Patterns associated with HRD (left) and HRP (right), extracted from the TCGA breast cancer dataset. Related to Figure III.5. The procedure of extraction is similar to the one used in the core article, using a model trained on the TCGA-BRCA dataset with subtype correction. Interestingly, we observe an important overlap in the previously unknown patterns associated with HRD in luminal tumors that we identified in our breast cancer series. These results indicate that the patterns related to HRD are to a large extent reproducible. The patterns related to HRP contain the retraction pattern (clear space surrounding the tumoral cell nests) also observed in Figure III.5. However, the patterns include artifacts (bubbles under the coverslip, written annotations on the glass slides...) that escaped us during the manual review of the slides. This shows that the quality of the TCGA dataset is less well controlled and this may be one source of errors contributing to the lower AUC. Scale bar $= 50\mu\text{m}$.

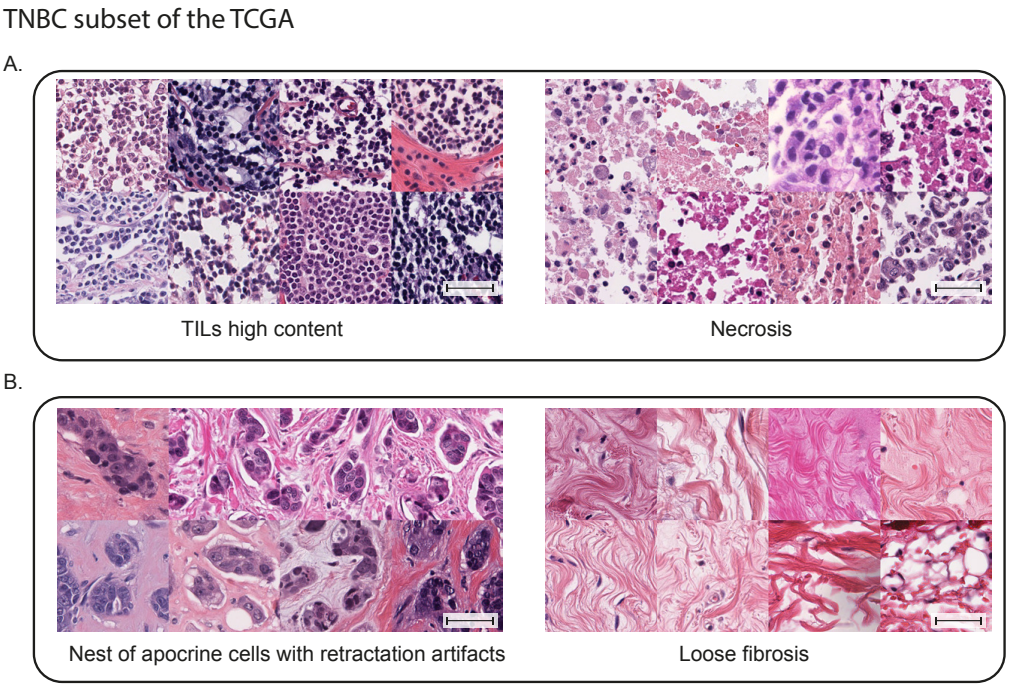


Figure C.4.: Patterns extracted with a model trained on the TNBC subset of the TCGA. We trained a NN on the TCGA TNBC data set (129 slides, test AUC: 0.62) and extracted the patterns as described in the methods. Even though the AUC of the NN is relatively low, the identified patterns related to HRD correspond to patterns that have been previously identified for this cancer subtype (in particular high percentage of TILs and necrosis). This is an additional indication that the association provided by our algorithm reflects a biological reality. Related to Figure III.5. Scale bar $= 50\mu\mathrm{m}$. A. patterns positively associated with HRD. B. Patterns positively associated with HRP.

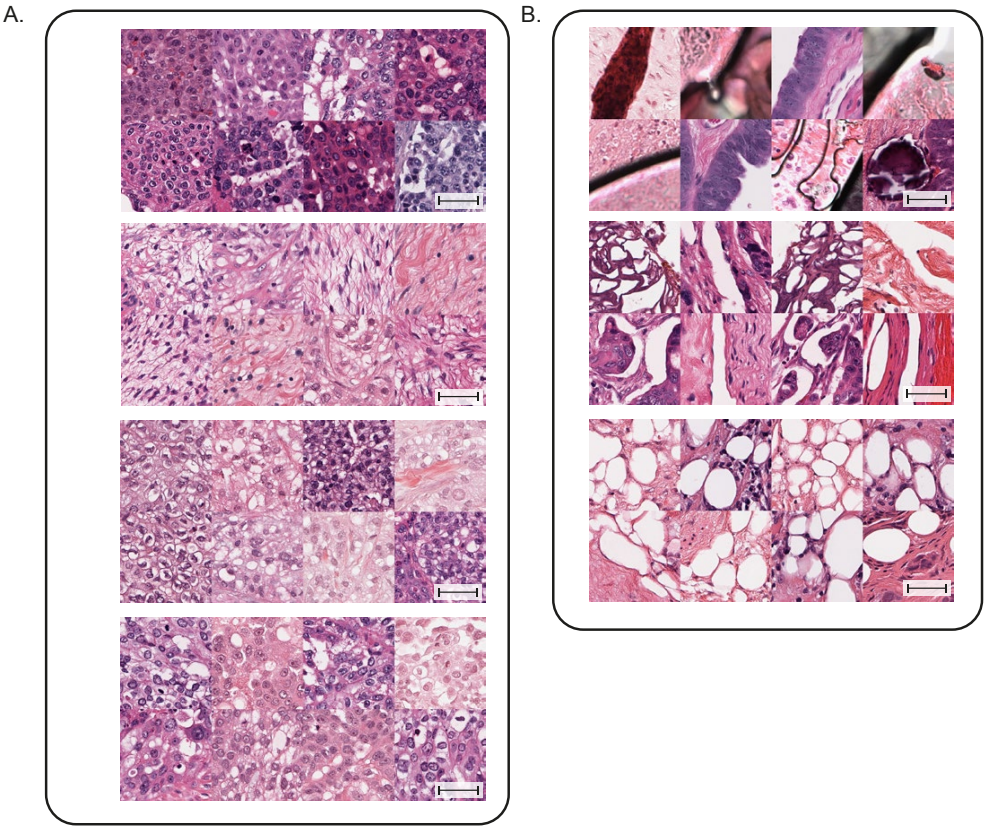
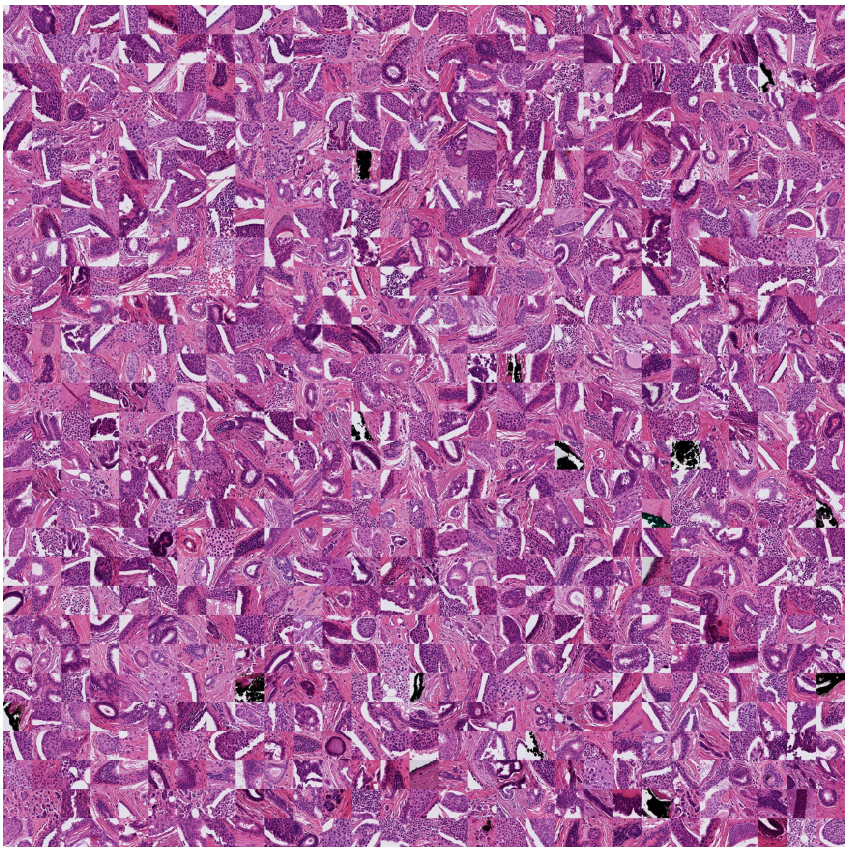


Figure C.5.: HR status related patterns extracted by a model trained on the Ovarian cohort of the TCGA dataset (92 FFPE WSIs). We trained a NN on the ovarian cancers of the TCGA (90 slides, test AUC: 0.73). Our method identified hyperchromatic cells with high atypia and clear cancerous cells, as well as fibrosis rich in TILs as patterns related to HRD. As patterns related to HRP, the algorithm again identified cell nests surrounded by clear space in addition to other patterns. The TCGA ovarian cancer dataset is small, and the results need to be corroborated by the analysis of larger independent and carefully controlled datasets, but even so, we observe some overlap in patterns across cancer types. Related to Figure III.5. Scale bar $= 50\mu\mathrm{m}$. A. Patterns related to the HRD. We can see hyperchromatic cells with high atypia, clear cancerous cells and TILs rich fibrosis. B. Patterns related to the HRP.

A



B

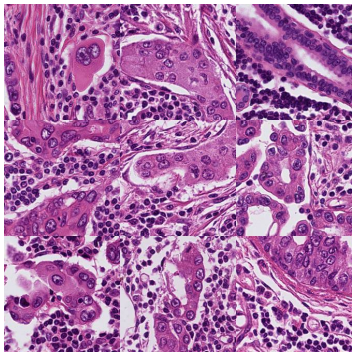


Figure C.6.: A. Tiles of the cluster 6 of Figure III.5. Scale bar $= 200\mu\mathrm{m}$.

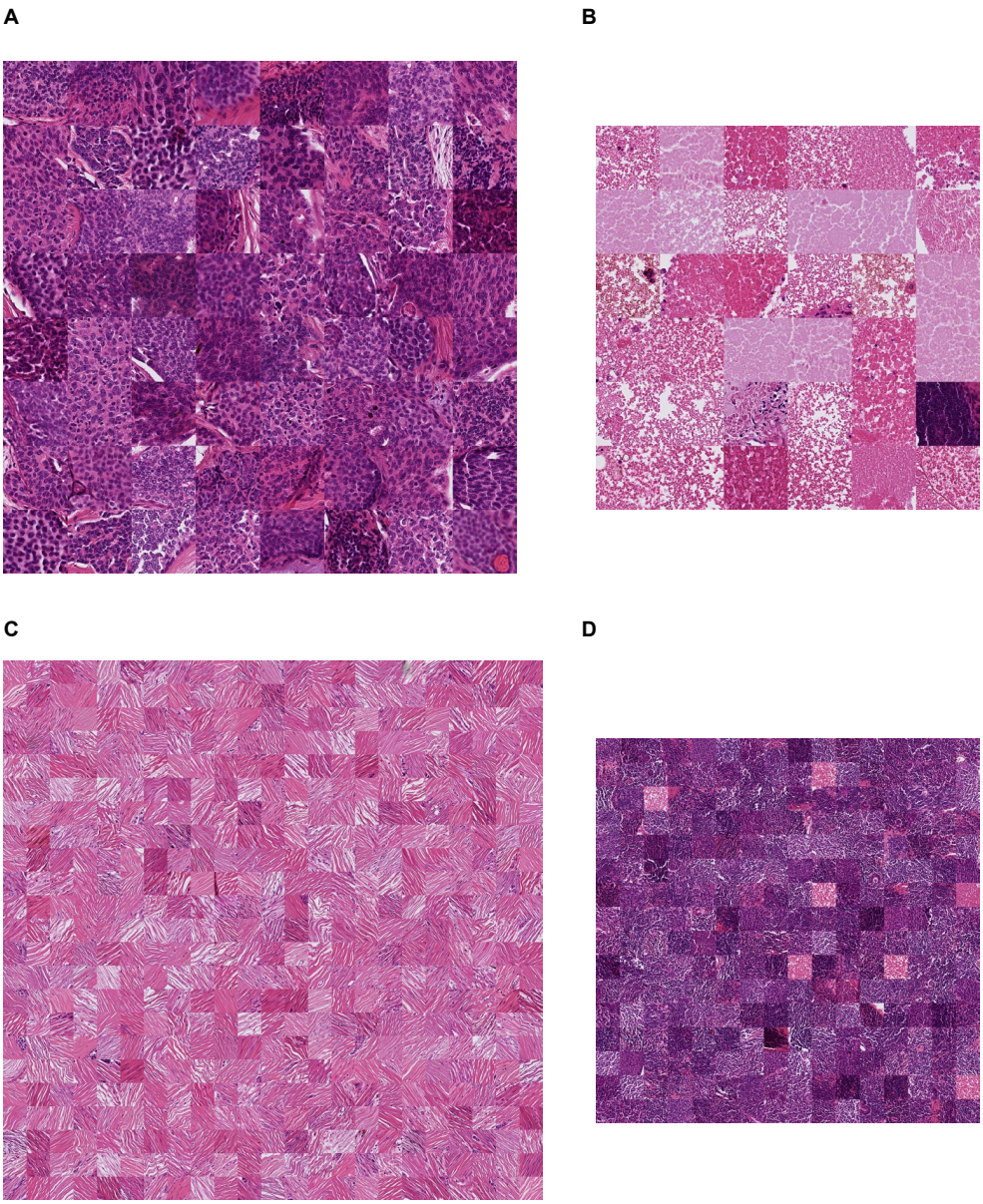


Figure C.7.: Scale bar $= 100\mu$m. A. Tiles of the cluster 3 of Figure III.5.

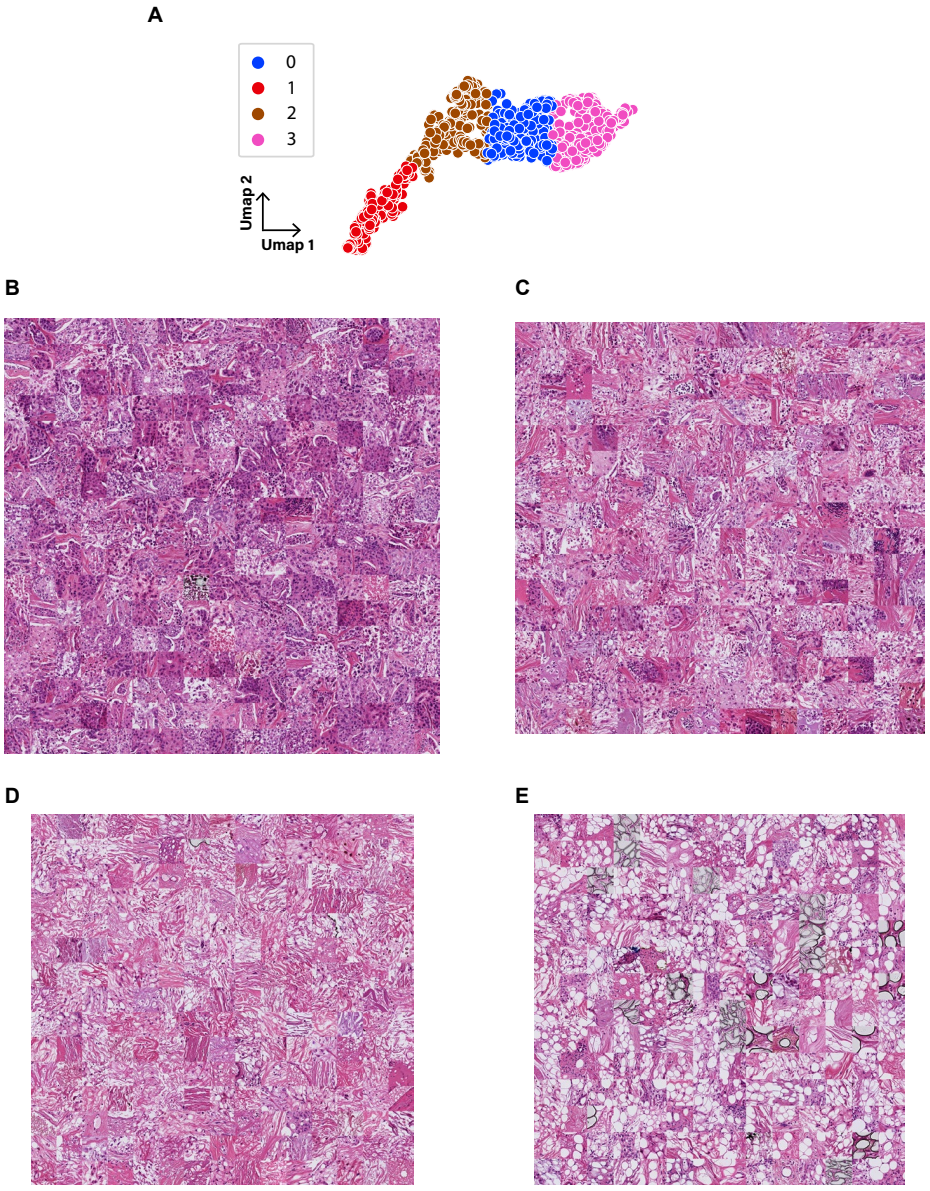


Figure C.8.: Scale bar $= 100\mu$m. A. Sub-clustering of the cluster 4 of Figure III.5. This cluster features tiles from very cellular tiles (subcluster 3) to adipocyte-rich tiles.

Tile-encoder	AUC		F_1		B_{acc}	
	mean	std	mean	std	mean	std
Imagenet No SSL pretraining	0.80	0.12	0.70	0.11	0.69	0.1
MoCo-Curie 600 K steps	0.83	0.06	0.72	0.06	0.72	0.06
TCGA-BRCA 600 K steps	0.81	0.05	0.73	0.05	0.72	0.05
TCGA-BRCA 1600K steps	0.84	0.06	0.76	0.08	0.74	0.1

Table C.2.: Tile-encoder pretraining influence. Related to Figure III.3 and Table III.1. Study of the impact of the feature encoder on the HRD classification performances, for the luminal BC dataset. Except for the Imagenet model, pretrained on Imagenet but without self-supervised pretraining, all models have been pre-trained from scratch, with the MoCo objective. Because the Curie dataset and the TCGA-BRCA dataset are different in size, models are compared in regard to their number of processed steps. Curie dataset $= 5.3$ million tiles, TCGA-BRCA $= 2.2$ million tiles. B_{acc} stands for balanced accuracy. F_1 reported is the average of the F_1 for both classes. All models have been trained and used for inference at the same magnification of $20 \times (0.46 \mu m.px)$. It is to note that all the pre-trained models outperform the ImageNet pre-trained model. In addition, while high performances are more quickly reached for a model trained on its target downstream dataset, similar or higher performances can be reached when training for longer the encoder on a different dataset of the same organ.

model	inference	AUC-ROC
Curie-Whole	TCGA-Whole	0.66
	TCGA-Luminal	0.62
Curie-Luminal	TCGA-Whole	0.62
	TCGA-Luminal	0.65
	TCGA-OV - Whole	0.64

Table C.3.: Results of the cross-dataset experiments. Related to the Results section and Table III.1. The model column indicates the dataset on which the tested model had been trained, either the entirety of the in-house dataset with bias correction (Curie-Whole) or the Luminal subset of the in-house dataset (Curie-Luminals). The 'inference' column indicates on what dataset the model was tested. Prediction performances were relatively low, but we still observe that performances on the TCGA of the models trained on the Curie Dataset are close to performances of bias-corrected models trained directly on the TCGA, thus suggesting that the signal relative to the HRD is not tied to our particular dataset. Moreover, we observe that training on the whole dataset causes a decrease in performance when inferring on the luminal subset of the TCGA, contrary to training only on the luminal slides. This further supports our hypothesis that the subtype information is used when predicting the HR status. We used the model learned on the Curie-luminal dataset with the small ovarian cohort of the TCGA (TCGA-OV) composed of 90 FFPE WSIs balanced with respect to the HR status. The resulting ROC-AUC score is 0.64 . We also trained a NN directly on the TCGA-OV and obtained a test AUC of 0.73 . Even though the number of slides is too low in order to reach a final conclusion, this suggests that probably the HR status signal is partly generalizable across organs, but that there are also tissue specific properties. Whether in the future there will be specialized neural networks for different cancer types or generalist networks for HRD prediction across organs still remains an open question.

	HRP		HRD		Total	
Cases	309		406		715	
Age at diagnosis (year)	52 ± 10		47 ± 11		49 ± 11	
Tumor size (mm)	17 ± 8		19 ± 13		19 ± 12	
Lymph node						
pNO	255	83%	278	68%	533	75%
pNi+	0	0%	2	0%	2	0%
pNmi	1	0%	11	3%	12	2%
pN1	32	10%	69	17%	101	14%
pN2	9	3%	33	8%	42	6%
pN3	4	1%	5	1%	9	1%
pNx	8	3%	8	2%	16	2%
Laterality						
Right	156	50%	193	48%	349	49%
Left	147	48%	206	51%	353	49%
ND	6	2%	7	2%	13	2%
Histological type						
Non special type	277	90%	343	84%	620	87%
Lobular	17	6%	18	4%	35	5%
Other	14	5%	45	11%	59	8%
Grade (EE)						
I	80	26%	13	3%	93	13%
II	161	52%	116	29%	277	39%
III	68	22%	275	68%	343	48%
ND	0	5%	2	1%	2	3%
Mitotic index	$3,5 \pm 5$		10 ± 9		$7,5 \pm 7$	
TILs (%)	16 ± 17		38 ± 24		30 ± 24	
Molecular class						
Triple Negative	25	8%	208	51%	233	33%
Luminal	284	92%	198	49%	482	67%
BRCA status						
gBRCA1	0	0%	188	46%	188	26%
gBRCA2	8	3%	174	43%	182	25%
Somatic mutation in HRD genes	0	0%	42	10%	42	6%
ND	301	97%	7	0%	308	43%

Table C.4.: Supplementary Table 4: Details of the in-house Curie dataset.

Appendix - Chapt. IV.

D

Pre-training Policy	Weighted Accuracy
ImageNet	0.865
Supervised	0.941
SSL	0.897
Mixed	0.965

Table D.1.: Performances of L2-regularized logistic regression trained on top of encoder with different pre-training policies. The task is the tile-level classification grading task. Interestingly, the performances of the *supervised* network is very good, while its usage in a MIL setting does not yield such good results at the WSI scale.

Appendix - Chapt. V.1

E

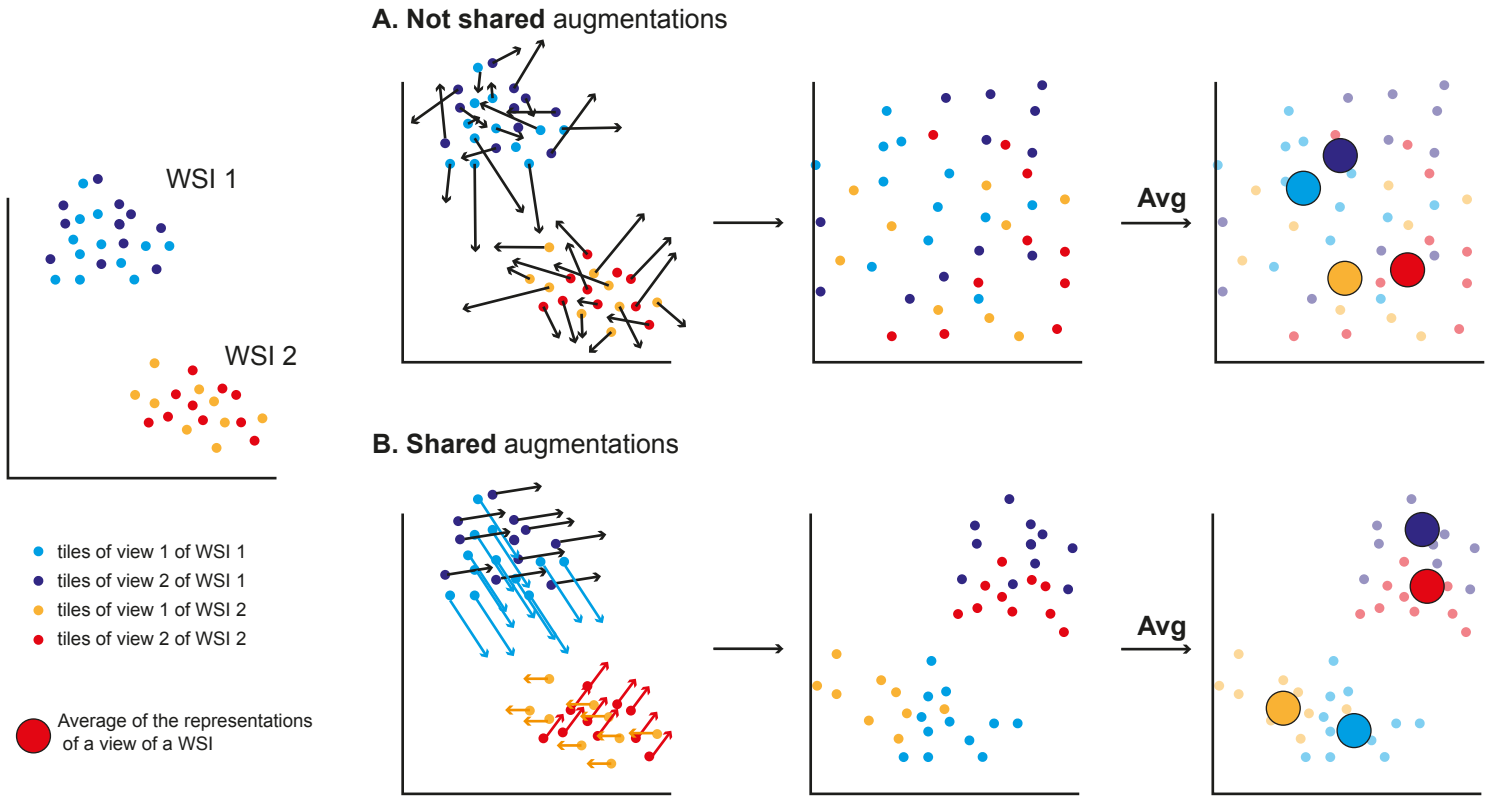


Figure E.1.: **Interpretation of Effectiveness of Shared Augmentations in Giga-SSL.** The figure progresses from left to right. It shows two WSIs and their tiles, projected into a space that captures color variations—consider the blue and red channels for instance. The left side displays the projections of all tiles from these two WSIs. In Giga-SSL training, two views are created for each WSI. First, the WSIs are subsampled, indicated by slight variations in color. Then each tile is transformed, such as through color alteration. Because hue transformations act as centered random augmentations to color channels, averaging independently augmented tiles likely retains information about their original color; i.e. what we aim at removing using the color transformation. In case **A**, where transformations are not shared, tiles of all views get randomly shuffled in the color space. It is likely howevere that averaging these transformed views can restore the slide-specific color, making the SSL task too simple by providing it with short-cuts. In contrast, case **B** involves sharing transformations among views. Simple operations like averaging no longer simplify the SSL task, as the averaged transformed views from the same slide remain distinct.

Appendix - Chapt. V.2

F

Tile-encoder	task	AUC	F1 Score	Balanced Accuracy
MoCo	brca	0.919 ± 0.023	0.792 ± 0.024	0.84 ± 0.031
	kidney	0.982 ± 0.014	0.891 ± 0.033	0.91 ± 0.028
	lung	0.959 ± 0.014	0.897 ± 0.012	0.898 ± 0.012
	mhrd	0.783 ± 0.042	0.734 ± 0.035	0.734 ± 0.035
	thrd	0.843 ± 0.032	0.778 ± 0.036	0.779 ± 0.036
CtransPath	brca	0.95 ± 0.024	0.849 ± 0.019	0.875 ± 0.026
	kidney	0.991 ± 0.005	0.929 ± 0.023	0.943 ± 0.016
	lung	0.968 ± 0.013	0.914 ± 0.023	0.914 ± 0.022
	mhrd	0.814 ± 0.045	0.749 ± 0.032	0.749 ± 0.032
	thrd	0.879 ± 0.026	0.795 ± 0.036	0.796 ± 0.036

Table F.1.: Effect of the tile-encoder on Giga-SSL representation performance. Results show a 5-fold patient-level stratified cross-validated average. Models are logistic regressions (C=10, class_weight='balanced') built on Giga-SSL representations using various tile-encoders. Advances in tile-encoder design enhance the efficacy of Giga-SSL representations. The tasks referenced are the 5 benchmark tasks from the main paper. The same cross-validation folds are used for all the experiment of the same task.

	N Patients	N Slides	Labels repartition
TCGA - brca	1041	977	831 (ductal) / 210 (lobular)
TCGA - lung	1033	936	528 (TCGA-LUAD) / 505 (TCGA-LUSC)
TCGA - mhrd	912	853	465 (0) / 447 (1)
TCGA - kidney	924	882	510 (ClearCell) / 294 (Papillary) / 120 (Chromophobe)
TCGA - thrd	634	586	318 (HRD) / 316 (HRP)
Curie - HRD	787	787	485 (HRD) / 302 (HRP)
Curie - Subtype	787	787	514 (luminal) / 273 (TNBC)
Curie Melanoma	515	515	336 (Monosomy) / 179 (Disomy)

Table F.2.: Detail of the composition of the datasets used for the benchmark tasks.

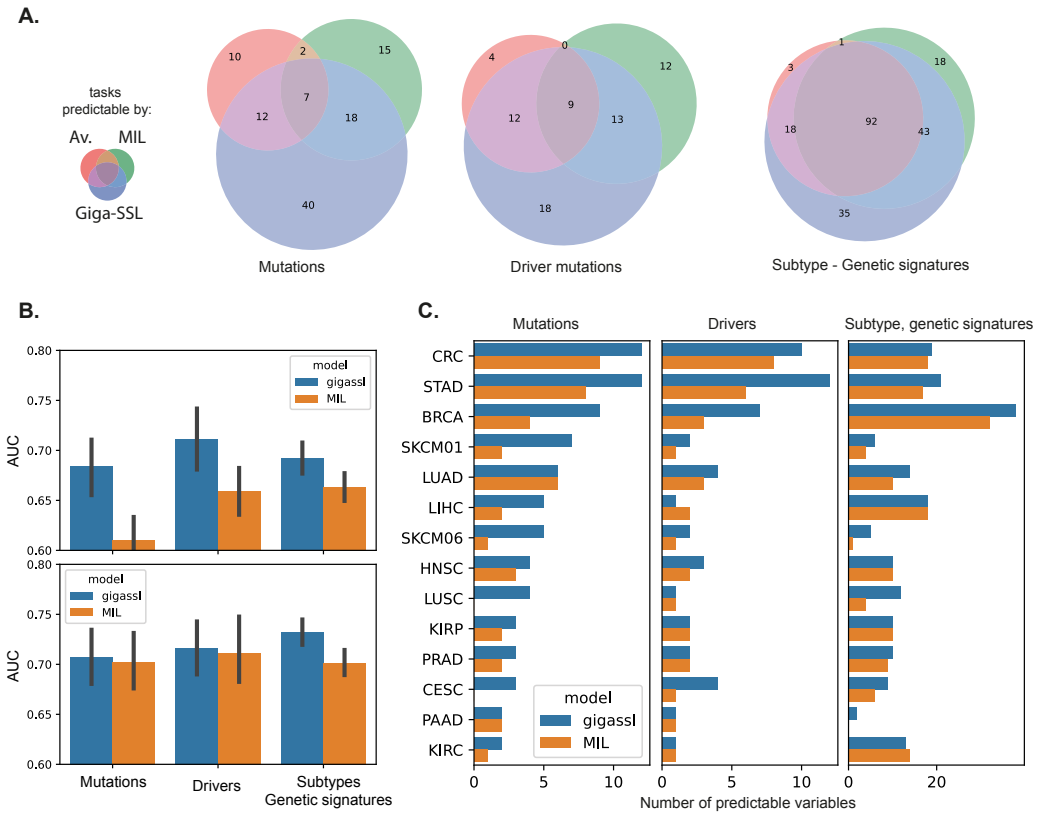


Figure F.1.: Results of logistic regression on top of Giga-SSL representation, trained and evaluated on folds stratified by center, following the method of Howard et al. (2021). The results corresponding to Giga-SSL show results corrected for the health-center bias. However, MIL results are not corrected and are the same as the one used in Figure V.5. The site-corrected evaluation of Giga-SSL methods still shows an improvement in all metrics, compared to the non-corrected results of MIL. **A.** The number of predictable tasks for each model and each task type (mutations, driver mutations, subtypes, and genetic signatures) in a pancancer setting. **B.** Displays the average cross-validated AUC for the three different types of classification tasks- mutation, oncogenic drivers, subtypes and genetic signatures-for Giga-SSL-*corrected* and MIL models. In the upper panel, the results are averaged across all tasks that are predictable by the Giga-SSL-corrected **or** MIL model (union). The lower one shows averages across the tasks predictable by both models. **C.** Provides a detailed breakdown of item A., focusing on the granularity of the TCGA projects.

Project	Mutations predictables by GigaSSL and MIL				
BRCA	MAP3K1	PIK3CA	TP53		
CESC	STK11				
CRC	APC	BRAF	KMT2B	KMT2D	KRAS
	MGA	PIK3CA	PTCH1	RNF43	TP53
HNSC	CASP8	NSD1	TP53		
KIRC	PBRM1				
KIRP	SETD2				
LIHC	CTNNB1				
LUAD	PDGFRB	TP53			
PRAD	TP53				
STAD	FBXW7	KMT2B	KMT2C	KMT2D	MTOR
	PIK3CA	TP53			

Table F.3.: All point mutations predictable by Giga-SSL and the MIL model, sorted by TCGA project.

Project	Mutations predictables by GigaSSL only					
BRCA	BRCA1	ERBB2	FBXW7	GATA3	PBRM1	
CESC	APC	ERBB2	KMT2D	KRAS	TCERG1	TP53
CRC	ACVR2A	ATM	BRCA2	CDC27	FAT1	FBXW7
	JAK2	MIER3	MSH6	PPP6C	PTEN	RB1
	RHOA	TCERG1	TTN			
HNSC	APC	HRAS	JAK2			
KIRP	PBRM1					
LIHC	TP53	TSC2				
LUAD	EGFR	KRAS	MED12	NFE2L2	TTN	
LUSC	KEAP1	RAC1	TP53			
PAAD	TP53					
PRAD	TTN					
STAD	ACVR2A	ARHGAP6	ARID1A	B2M	CDK12	KMT2A
	MAP3K1	MET	MGA	RBM10	RHOA	TCERG1
SKCM01	MGA	PIK3CA	RNF43	TTN		
SKCM06	CDC27	CDKN2A	FBXW7	GNAS	KDM6A	PPP6C

Table F.4.: All point mutations newly predictable with the Giga-SSL features, sorted by TCGA project.

Project	Driver mutations predictable by GigaSSL and MIL				
BRCA	MAP3K1	PIK3CA	TP53		
CRC	APC	BRAF	KMT2B	KMT2D	KRAS
	PIK3CA	RNF43	TP53		
HNSC	NSD1	TP53			
KIRC	PBRM1				
KIRP	KRAS	SETD2			
LIHC	CTNNB1				
LUAD	EGFR	TP53			
PAAD	KRAS				
PRAD	TP53				
STAD	BRCA2	FBXW7	KMT2B	KMT2D	PIK3CA

Table F.5.: Oncogenic driver mutations predictable by Giga-SSL and the MIL model, sorted by TCGA project.

Project	Driver mutations predictable only by GigaSSL			
BRCA	BRCA1	GATA3	SMAD4	
CESC	KRAS	STK11		
CRC	BRCA2	MGA	NRAS	PTEN
HNSC	CASP8			
KIRC	TP53			
KIRP	MET	PBRM1		
LIHC	TP53			
LUAD	KRAS	MGA	U2AF1	
LUSC	NFE2L2	PIK3CA		
PAAD	TP53			
SKCM01	PIK3R1			
SKCM06	CDKN2A	CTNNB1		
STAD	AMER1	ARID1A	KMT2C	PTCH1
	RHOA	RNF43	TP53	

Table F.6.: Oncogenic driver mutations newly predictable with the Giga-SSL features, sorted by TCGA project.

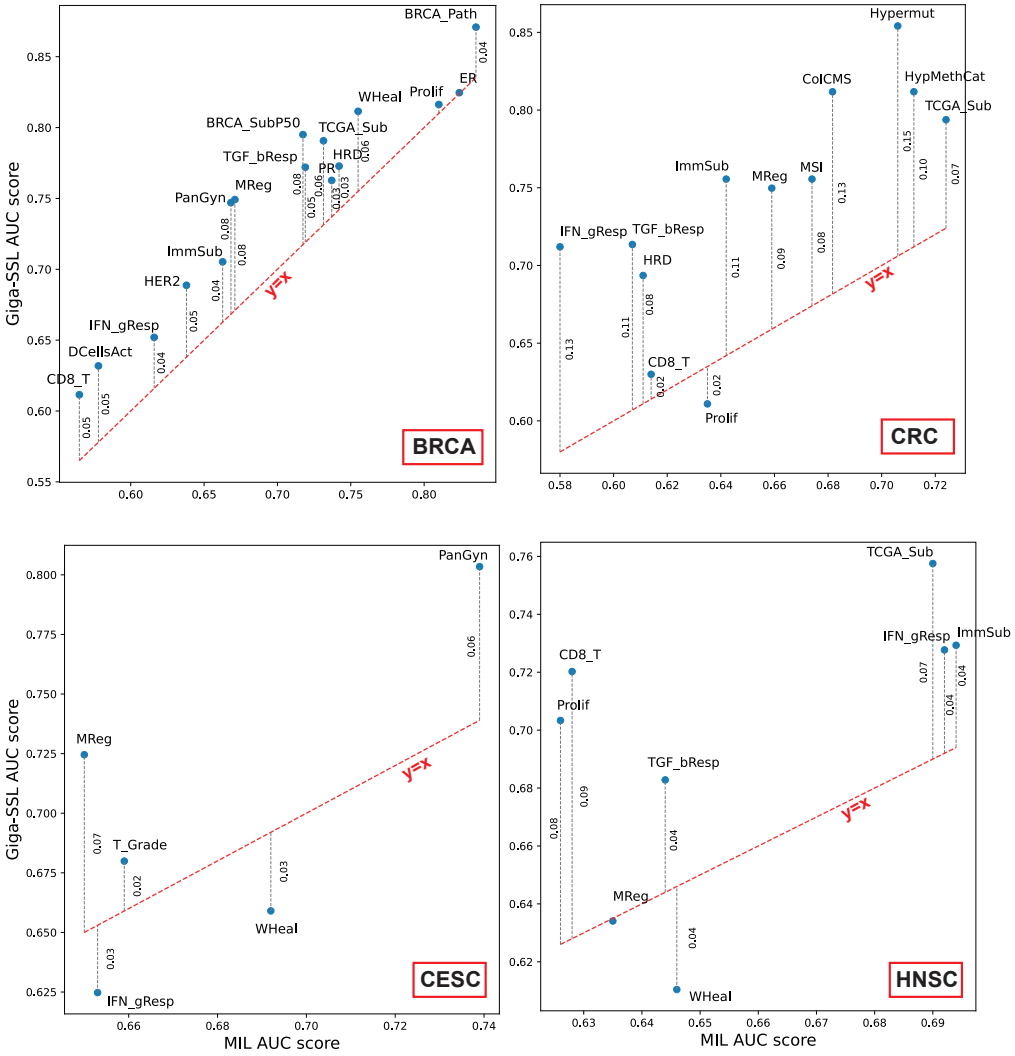


Figure F.2.: Performances of the non-mutation tasks for BRCA, CRC, CESC and HNSC TCGA projects. Plots shows the performances of the logistic regression trained on top of the Giga-SSL features against the performances of the MIL model used in Jakob Nikolas Kather et al. (2020). Red line indicates same performances between the two models. PR - PRStatus (Progesterone Receptor); AR - AR_protein (Androgen Receptor); N_HistGrade - Neoplasm Histologic Grade; T_Grade - Tumor Grade; DCellsAct - Activated Dendritic Cells; SCNA - Somatic Copy Number Alterations; TCGA_Sub - TCGA Subtypes; HER2 - HER2 Final Status; ImmSub - Immune Subtypes; HypMethCat - Hypermethylation Category; WHeal - Wound Healing; CD8_T - T Cells CD8; PanGyn - Pan-Gynecologic Clusters; GHistClass - Gastric Histological Classification; GrowthPat - Major Growth Pattern; Prolif - Proliferation; ClinGleason - Clinical Gleason Sum; PanKidPath - Pan-Kidney Pathology; IFN_gResp - IFN-gamma Response; MReg - Macrophage Regulation; HomRecDef - Homologous Recombination Defects; HCCSub - Hepatocellular Carcinoma Subtypes; ERG - ERG Status; ColCMS - Colorectal Cancer CMS; MSI - Microsatellite Instability Status; ER - Estrogen Receptor Status; BRCA_Path - BRCA Pathology; Hypermut - Hypermutated; TGF_bResp - TGF-beta Response; BRCA_SubP50 - BRCA Subtype (PAM50)

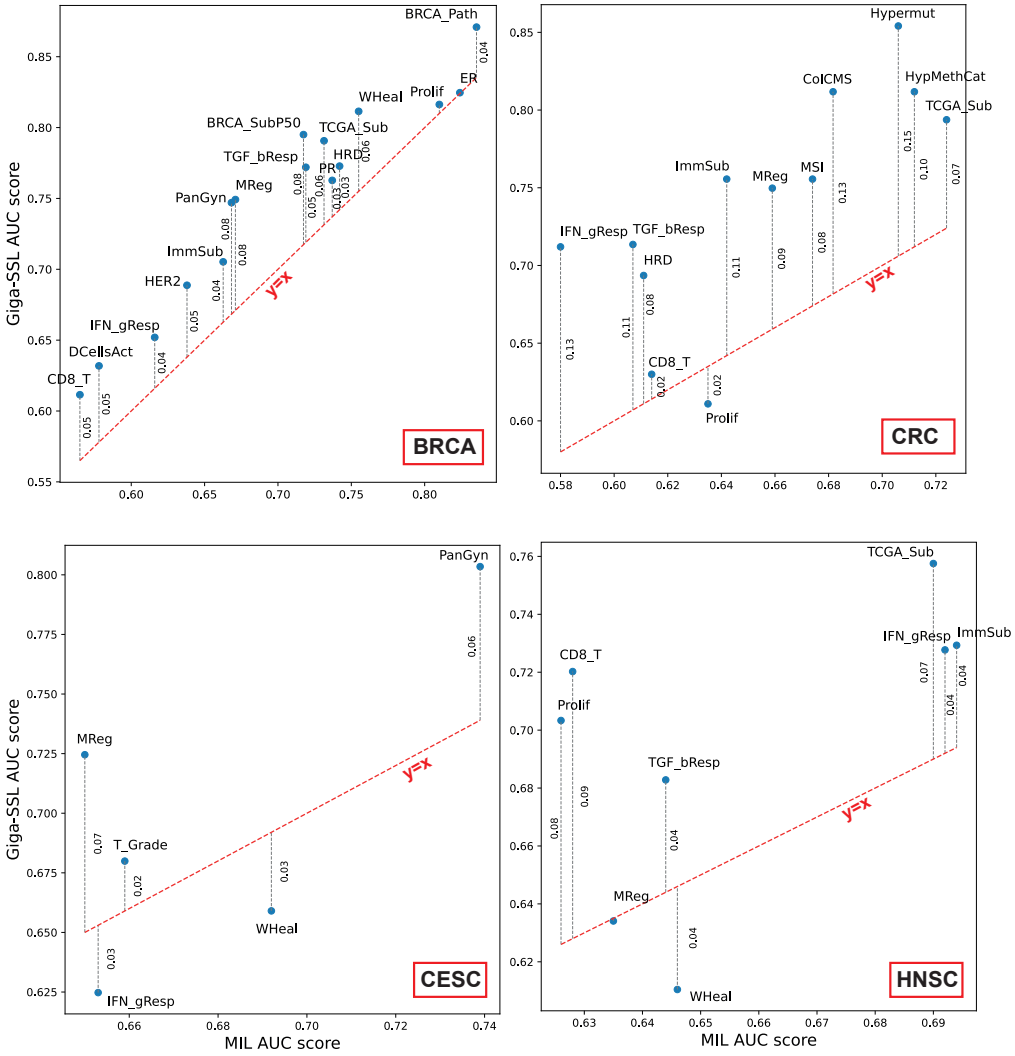


Figure F.3.: Performances of the non-mutation tasks for KICH, KIRC, KIRP and LIHC TCGA projects. Plots shows the performances of the logistic regression trained on top of the Giga-SSL features against the performances of the MIL model used in Jakob Nikolas Kather et al. (2020). Red line indicates same performances between the two models. PR - PRStatus (Progesterone Receptor); AR - AR_protein (Androgen Receptor); N_HistGrade - Neoplasm Histologic Grade; T_Grade - Tumor Grade; DCellsAct - Activated Dendritic Cells; SCNA - Somatic Copy Number Alterations; TCGA_Sub - TCGA Subtypes; HER2 - HER2 Final Status; ImmSub - Immune Subtypes; HypMethCat - Hypermethylation Category; WHeal - Wound Healing; CD8_T - T Cells CD8; PanGyn - Pan-Gynecologic Clusters; GHistClass - Gastric Histological Classification; GrowthPat - Major Growth Pattern; Prolif - Proliferation; ClinGleason - Clinical Gleason Sum; PanKidPath - Pan-Kidney Pathology; IFN_gResp - IFN-gamma Response; MReg - Macrophage Regulation; HomRecDef - Homologous Recombination Defects; HCCSub - Hepatocellular Carcinoma Subtypes; ERG - ERG Status; ColCMS - Colorectal Cancer CMS; MSI - Microsatellite Instability Status; ER - Estrogen Receptor Status; BRCA_Path - BRCA Pathology; Hypermut - Hypermutated; TGF_bResp - TGF-beta Response; BRCA_SubP50 - BRCA Subtype (PAM50)

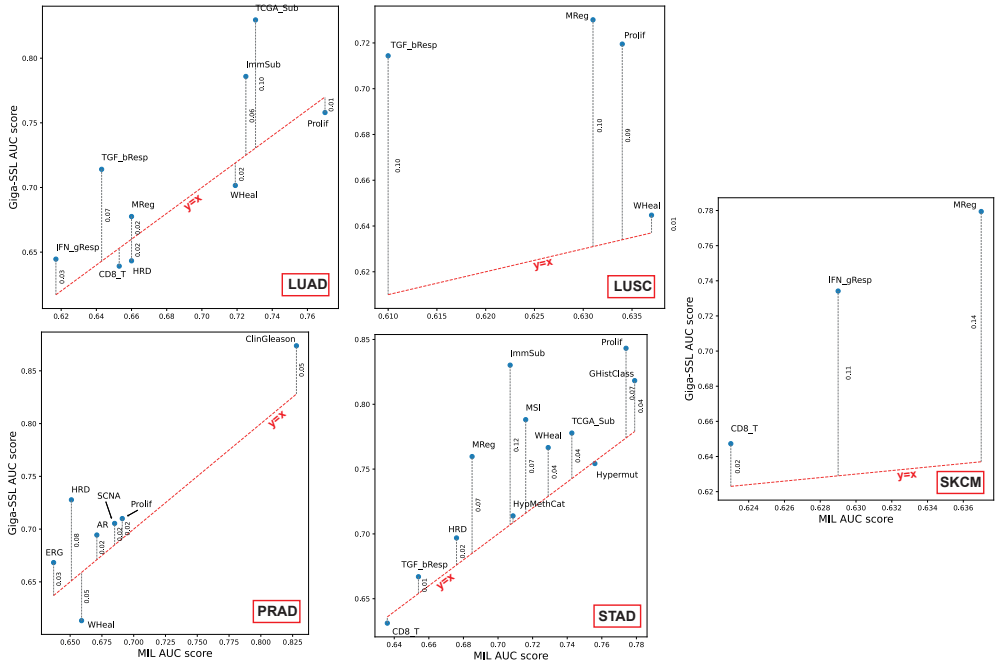


Figure F.4.: Performances of the non-mutation tasks for LUAD, LUSC, PRAD, STAD and SKCM TCGA projects. Plots shows the performances of the logistic regression trained on top of the Giga-SSL features against the performances of the MIL model used in Jakob Nikolas Kather et al. (2020). Red line indicates same performances between the two models. PR - PRStatus (Progesterone Receptor); AR - AR_protein (Androgen Receptor); N_HistGrade - Neoplasm Histologic Grade; T_Grade - Tumor Grade; DCellsAct - Activated Dendritic Cells; SCNA - Somatic Copy Number Alterations; TCGA_Sub - TCGA Subtypes; HER2 - HER2 Final Status; ImmSub - Immune Subtypes; HypMethCat - Hypermethylation Category; WHeal - Wound Healing; CD8_T - T Cells CD8; PanGyn - Pan-Gynecologic Clusters; GHistClass - Gastric Histological Classification; GrowthPat - Major Growth Pattern; Prolif - Proliferation; ClinGleason - Clinical Gleason Sum; PanKidPath - Pan-Kidney Pathology; IFN_gResp - IFN-gamma Response; MReg - Macrophage Regulation; HomRecDef - Homologous Recombination Defects; HCCSub - Hepatocellular Carcinoma Subtypes; ERG - ERG Status; ColCMS - Colorectal Cancer CMS; MSI - Microsatellite Instability Status; ER - Estrogen Receptor Status; BRCA_Path - BRCA Pathology; Hypermut - Hypermutated; TGF_bResp - TGF-beta Response; BRCA_SubP50 - BRCA Subtype (PAM50)

Appendix - Chapt. VI.

G

Histological criteria	Assessment
Tumour grade	Classification into well, moderately or poorly differentiated tumour according to the 5th edition of the WHO classification
Tumour histological type	Small duct, large duct or other subtypes
Necrosis	Percentage
Tumour fibrosis	Semi quantitatively assessed, classified into three classes: no or mild, moderate and intense
Immune tumour infiltration	Semi quantitatively assessed, classified into four classes: no inflammation, low, moderate and high
Tertiary lymphoid structure (TLS)	Presence or absence

Table G.1.: List of morphological criteria assessed by the expert pathologist for all cases of the three datasets

Transcriptomic groups	Total n=246 (%)	Surgical samples n=109 (%)	Biopsies n = 137 (%)	p
Hepatic stem-like	90 (37)	53 (49)	37 (27)	**<0.001**
Desert like	11 (4)	4 (4)	7 (5)	0.759
Tumor classical	34 (14)	16 (15)	18 (13)	0.853
Immune classical	57 (23)	15 (14)	42 (31)	**0.002**
Inflammatory stroma	54 (22)	21 (19)	33 (24)	0.439

Table G.2.: Repartition of the five transcriptomic classes according to the type of samples in the discovery set.

Transcriptomic groups	Total n=246 (%)	Surgical samples n=109 (%)	Biopsies n = 137 (%)	p
Hepatic stem-like	90 (37)	53 (49)	37 (27)	**<0.001**
Desert like	11 (4)	4 (4)	7 (5)	0.759
Tumor classical	34 (14)	16 (15)	18 (13)	0.853
Immune classical	57 (23)	15 (14)	42 (31)	**0.002**
Inflammatory stroma	54 (22)	21 (19)	33 (24)	0.439

Table G.3.: Repartition of the five transcriptomic classes according to the type of samples in the discovery set.

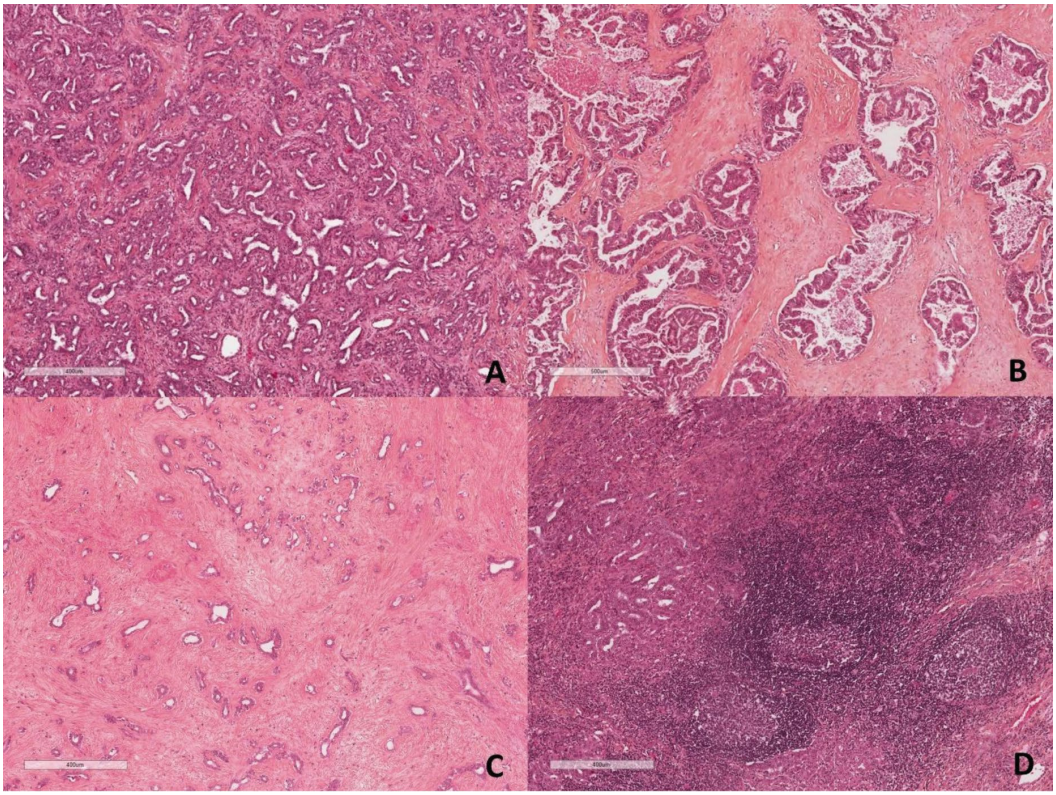


Figure G.1.: **Histological features of iCCA.** A. Example of small duct type iCCA. B. Example of large duct type iCCA. C. Example of an iCCA with an abundant fibrous stroma. D. Example of an iCCA with an abundant inflammatory stroma containing tertiary lymphoid structures.

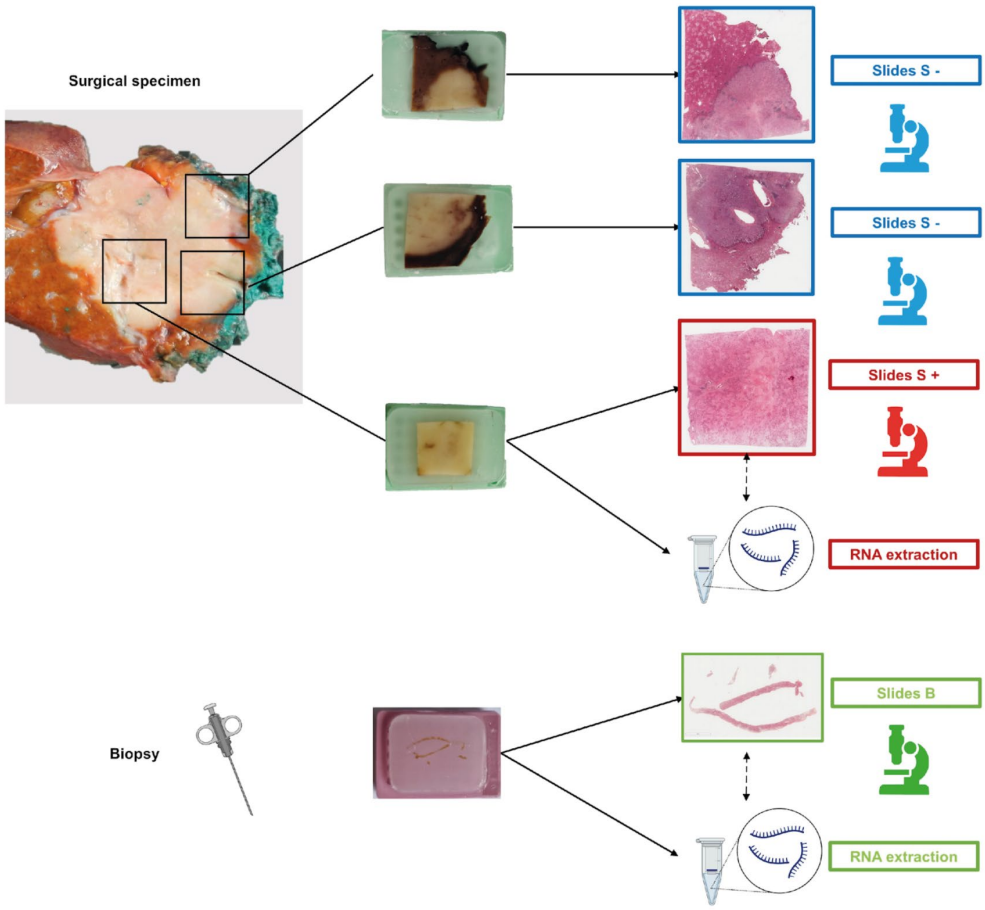


Figure G.2.: **Type of slides including in the study.** The slides directly associated with transcriptomic analysis (consecutive slides), have been labelled as surgical slides S+ whereas slides from other blocks indirectly associated with transcriptomic analysis in the discovery set and in the TCGA set have been labelled S-. For biopsy, the FFPE block used for RNA sequencing corresponded directly to the slide selected (labelled as slide B).

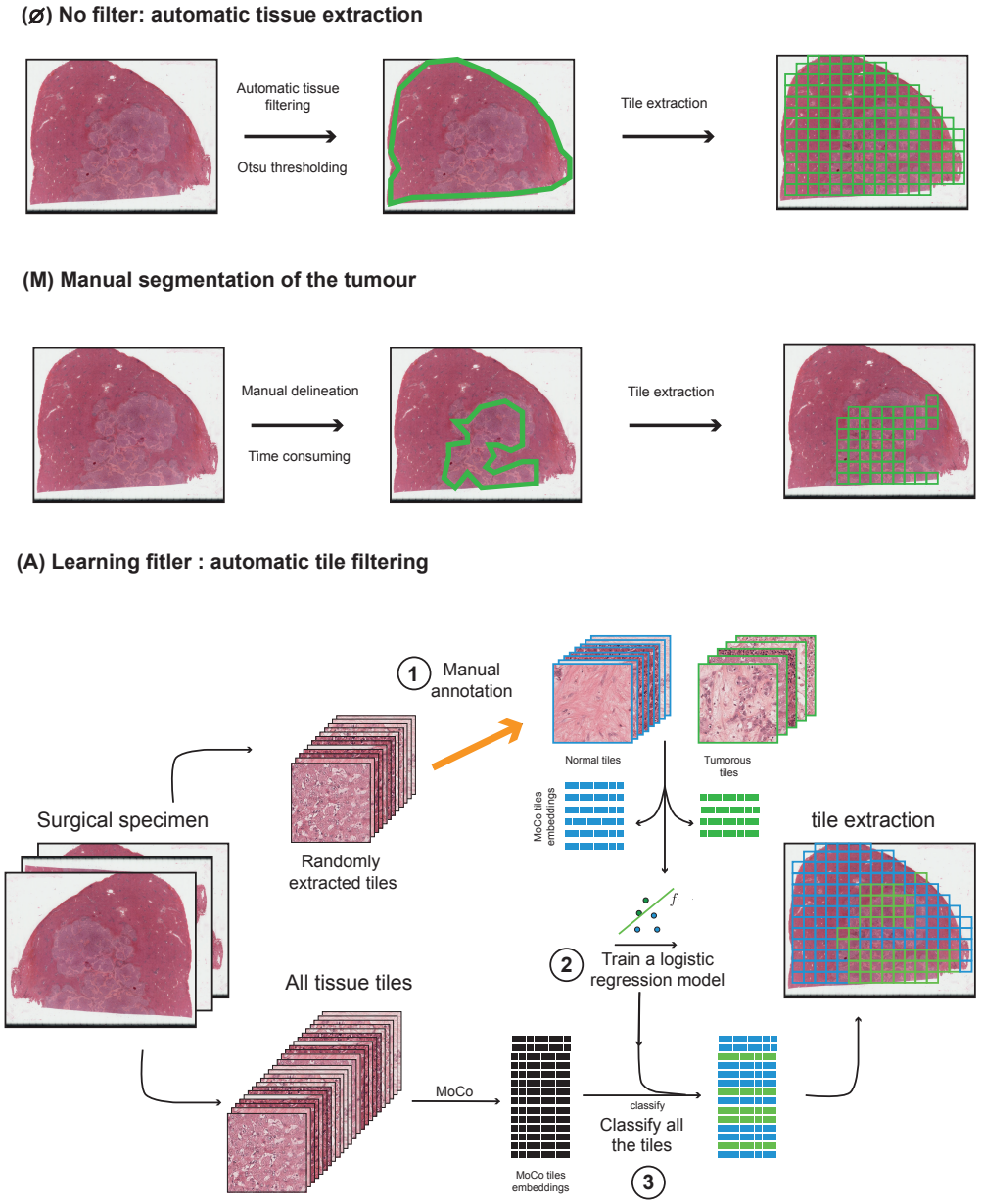


Figure G.3.: Three different pre-processing protocols with or without extraction of region of interest (ROI). -No-Filter (∅): All tiles including tumour and non-tumour are processed as they are, encompassing both tumour and non-tumour regions. -Manual-Filter (M): An expert pathologist (AB) extensively annotates tumour regions using ImageScope software, from which patches are extracted. -Learning-Filter (A): Tiles are filtered using a logistic regression trained on a dataset of 3000 tile embeddings, randomly extracted and labelled by an expert pathologist (AB).

Validation dataset	ROI extraction method	AUC	Balanced accuracy	F1 score
TCGA	L	0.82	0.68	0.68
	M	0.78	0.73	0.74
	ø	0.75	0.61	0.6
mondor	L	0.8	0.73	0.72
	M	0.86	0.76	0.74
	ø	0.81	0.71	0.68

Table G.4.: Effects of the ROI extraction method on the external validation performances for the Hepatic stem-like binary classification task. The same method is applied on both the training and validation datasets. On the TCGA set, no distinct advantage is observed for method A over M or vice versa, as it varies based on the metric under consideration. In the French external set, manually segmenting the tumour seems to be advantageous. Nevertheless, in both datasets, using an ROI extraction method is more effective than not using any. Area under the curve, AUC; Learning-Filter, L; Manual-Filter, M; No-Filter, ∅; Region of interest, ROI

RÉSUMÉ

Les images de lames entières (WSI) sont des versions numérisées de coupes microscopiques de tissus colorés. Ces images remplissent plusieurs fonctions dans la prise en charge du cancer. Elles servent non seulement d'outil diagnostique de référence, mais aussi pour la stratification de patient, le sous-typage de la maladie et l'orientation vers des options de traitement personnalisés. Elles sont également utilisées pour évaluer l'efficacité des traitements et suivre leurs résultats au fil du temps.
En effet, les WSI contiennent des informations biologiques complexes. On peut y trouver des centaines de milliers de cellules à travers différents types de tissus ainsi que des motifs visuels allant de la texture nucléaire à l'architecture des tissus. Cette thèse se concentre sur l'utilisation de l'apprentissage automatique pour extraire l'information importante contenue dans les WSI, un processus connu sous le nom d'apprentissage de représentation. La supervision pour l'apprentissage de représentation d'images d'histopathologie peut prendre plusieurs formes, allant des étiquettes générées par les médecins à des mesures biologiques supplémentaires telles que les données de séquençage de l'ADN et de l'ARN. Cependant, ces signaux de supervision présentent des défis : ils peuvent être faibles, bruités, incertains et surtout, rares, car difficilement accessibles.
L'objectif principal de cette thèse est donc de répondre à ces défis par le développement d'algorithmes d'apprentissage de représentation qui fonctionnent efficacement sous ces contraintes de supervision. Nous y détaillerons des contributions de plusieurs natures, allant du développement de nouveaux algorithmes d'interprétabilité à l'introduction d'un nouveau cadre d'apprentissage auto-supervisé conçu spécifiquement pour l'apprentissage de représentation de WSI. Finalement, chacune de ces avancées seront présentées dans le cadre de la résolution de tâches de classification de WSI, basées sur des données moléculaires et ayant une importance clinique significative.

MOTS CLÉS

Apprentissage Automatique, Apprentissage Profond, Images de Lames Entières, Cancérologie, Auto-Supervision, Interprétabilité, Histopathologie, HRD.

ABSTRACT

Whole-slide images (WSI) are digitized versions of microscopic images that capture thin layers of stained tissue samples. These images serve multiple roles in cancer care, functioning not only as the gold standard for cancer diagnosis but also as a tool for patient stratification, disease subtyping, and guiding personalized treatment plans. They are also used for evaluating treatment efficacy and monitoring outcomes over time. Indeed, WSIs carry complex biological information, capturing data from hundreds of thousands of cells across different tissue types, and features ranging from nuclear texture to tissue architecture.
This thesis focuses on using machine learning and deep learning to automate the extraction of meaningful information from WSIs, a process known as representation learning. Supervision for representation learning can take multiple forms, from human-generated labels to additional biological measurements such as DNA and RNA sequencing data. However, these supervision signals often present challenges; they can be weak, noisy, uncertain, and most critically, scarce in availability. The primary objective of this thesis is to address these challenges by developing robust algorithms for histopathological image representation learning that operate effectively under these supervisory constraints. The contributions of this work are of several natures, spanning from the development of new interpretability algorithms to the introduction of a novel self-supervised framework designed explicitly for WSI-level representation learning. In addition, these advancements are contextualized within WSI classification tasks that rely on molecular data and hold significant clinical importance.

KEYWORDS

Histopathology, Weak Supervision, Self Supervision, Oncology, Interpretability, Machine Teaching, HRD.