



UNIVERSITY OF TRENTO
FONDAZIONE BRUNO KESSLER

DOCTORAL THESIS

AI for Omics and Imaging models in Precision Medicine and Toxicology

Candidate:

Nicole Bussola

Tutor:

Cesare Furlanello

Advisor:

Andrea Lunardi

*International PhD Program in Biomolecular Sciences - Transdisciplinary Program in
Computational Biology*

Department of Cellular, Computational and Integrative Biology – CIBIO

XXXIV cycle

Academic year 2020/2021

Nicole Bussola: *AI for Omics and Imaging models*

in Precision Medicine and Toxicology, DOCTORAL THESIS, © May 2022

DECLARATION

I Nicole Bussola confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Trento, May 2022

Nicole Bussola

ABSTRACT

This thesis develops an Artificial Intelligence (AI) approach intended for accurate patient stratification and precise diagnostics/prognostics in clinical and preclinical applications. The rapid advance in high throughput technologies and bioinformatics tools is still far from linking precisely the genome-phenotype interactions with the biological mechanisms that underlie pathophysiological conditions. In practice, the incomplete knowledge on individual heterogeneity in complex diseases keeps forcing clinicians to settle for surrogate endpoints and therapies based on a generic one-size-fits-all approach. The working hypothesis is that AI can add new tools to elaborate and integrate together in new features or structures the rich information now available from high-throughput omics and bioimaging data, and that such restructured information can be applied through predictive models for the precision medicine paradigm, thus favoring the creation of safer tailored treatments for specific patient subgroups. The computational techniques in this thesis are based on the combination of dimensionality reduction methods with Deep Learning (DL) architectures to learn meaningful transformations between the input and the predictive endpoint space. The rationale is that such transformations can introduce intermediate spaces offering more succinct representations, where data from different sources are summarized. The research goal was attacked at increasing levels of complexity, starting from single input modalities (omics and bioimaging of different types and scales), to their multimodal integration. The approach also deals with the key challenges for machine learning (ML) on biomedical data, i.e. reproducibility, stability, and interpretability of the models. Along this path, the thesis contribution is thus the development of a set of specialized AI models and a core framework of three tools of general applicability:

- i. A Data Analysis Plan (DAP) for model selection and evaluation of classifiers on omics and imaging data to avoid selection bias.

- ii. The histolab Python package that standardizes the reproducible pre-processing of Whole Slide Images (WSIs), supported by automated testing and easily integrable in DL pipelines for Digital Pathology.
- iii. Unsupervised and dimensionality reduction techniques based on the UMAP and TDA frameworks for patient subtyping.

The framework has been successfully applied on public as well as original data in precision oncology and predictive toxicology. In the clinical setting, this thesis has developed¹:

1. (DAPPER) A deep learning framework for evaluation of predictive models in Digital Pathology that controls for selection bias through properly designed data partitioning schemes.
2. (RADLER) A unified deep learning framework that combines radiomics features and imaging on PET-CT images for prognostic biomarker development in head and neck squamous cell carcinoma. The mixed deep learning/radiomics approach is more accurate than using only one feature type.
3. An ML framework for automated quantification tumor infiltrating lymphocytes (TILs) in onco-immunology, validated on original pathology Neuroblastoma data of the Bambino Gesù Children's Hospital, with high agreement with trained pathologists. The network-based INF pipeline, which applies machine learning models over the combination of multiple omics layers, also providing compact biomarker signatures. INF was validated on three TCGA oncogenomic datasets.

In the preclinical setting the framework has been applied for:

1. Deep and machine learning algorithms to predict DILI status from gene expression (GE) data derived from cancer cell lines on the CMap Drug Safety dataset.
2. (ML4TOX) Deep Learning and Support Vector Machine models to predict potential endocrine disruption of environmental chemicals on the CERAPP dataset.

¹ See Table of Acronyms for definitions of Acronyms in this section and in the thesis.

3. (PathologAI) A deep learning pipeline combining generative and convolutional models for preclinical digital pathology. Developed as an internal project within the FDA/NCTR AIRForce initiative and applied to predict necrosis on images from the TG-GATEs project, PathologAI aims to improve accuracy and reduce labor in the identification of lesions in predictive toxicology. Furthermore, GE microarray data were integrated with histology features in a unified multi-modal scheme combining imaging and omics data.

The solutions were developed in collaboration with domain experts and considered promising for application.

CONTENTS

I	INTRODUCTION	1
1	BACKGROUND	1
1.1	Precision medicine and why we need it	1
1.2	Biomarkers for patient stratification	4
1.2.1	Omics data	5
1.2.2	Radiology	8
1.2.3	Digital Pathology	10
2	AI AND BIOMARKER DISCOVERY	13
2.1	Deep Learning	13
2.2	Upscale resources	15
2.3	AI and Radiomics	16
2.4	AI and Digital Pathology	17
2.5	AI and toxicology	18
2.6	Deployment of medical AI: technical challenges	21
2.6.1	Reproducibility of predictive models	23
2.6.2	Data processing in Digital Pathology	27
2.6.3	Deep feature analysis	30
2.6.4	Multi-omics integration	33
3	PROJECT AIMS	37
II	PREDICTIVE MODELS IN DIGITAL PATHOLOGY	39
4	THE DAPPER FRAMEWORK	41
4.1	Abstract	42
4.2	Materials and methods	47
4.3	Results	54
4.4	Discussion	67
4.5	Appendix	69

5	DATA LEAKAGE IN DIGITAL PATHOLOGY	77
5.1	Abstract	77
5.2	Material and methods	80
5.3	Results	86
5.4	Discussion	92
5.5	Appendix	94
6	THE HISTOLAB LIBRARY	97
6.1	Abstract	97
6.2	Software description	100
6.3	Illustrative Example	106
6.4	Appendix	107
7	TILS DETECTION IN NEUROBLASTOMA	114
7.1	Abstract	114
7.2	Materials and Methods	121
7.3	Results	132
7.4	Discussion	144
	III MULTI-MODAL INTEGRATION IN RADIOLOGY	146
8	THE RADLER FRAMEWORK	148
8.1	Abstract	148
8.2	Material and Methods	152
8.3	Results	159
8.4	Discussion	161
	IV PREDICTIVE MODELS ON OMICS FOR TOXICOLOGY	164
9	DILI PREDICTION FROM OMICS DATA	166
9.1	Abstract	166
9.2	Material and Methods	168
9.3	Results	173
9.4	Discussion	177
10	ML4TOX FOR DRUG BINDING ACTIVITY	181

10.1	Abstract	181
10.2	Materials and Methods	183
10.3	Results	189
10.4	Discussion	193
V	MULTI-OMICS INTEGRATION	196
11	THE INF FRAMEWORK	198
11.1	Abstract	198
11.2	Materials and Methods	201
11.3	Results	210
11.4	Discussion	218
11.5	Appendix	221
VI	AI ON OPEN RESEARCH QUESTIONS IN TOXICOLOGY	224
12	CAN AI IMPROVE ON TOXICOLOGICAL PATHOLOGY REPORTS?	225
12.1	The PathologAI weak-label framework	225
12.2	Related works	228
12.2.1	Weakly-supervised approaches for clinical tasks	228
12.2.2	AI in Preclinical Toxicologic Pathology	229
12.3	Imaging data	230
12.3.1	BiGAN training data	231
12.3.2	Classification data	233
12.3.3	Tile extraction and preprocessing	234
12.4	Unsupervised encoding network	235
12.4.1	Compression and packing	236
12.5	CNN architecture	237
12.6	Feature projection and clustering	237
12.7	Lesion mapping	238
12.7.1	HistoMAP calibration with pathologists	239
12.8	Experiments and Results	240
12.8.1	Necrosis Score	240
12.8.2	Data Analysis Plan and Ensemble model	241

12.8.3 External validation	243
13 IMAGING AND OMICS COMBINATION	246
13.1 Gene expression data	247
13.2 Machine Learning pipeline	249
13.3 Experiments and Results	250
13.3.1 Omics data	250
13.3.2 Omics and Imaging combination	252
VII DISCUSSION	254
14 DISCUSSION	255
14.1 Personal contributions	261
A APPENDIX A	263
A.1 UMAP Background	263
A.2 Topological Descriptors	266
A.3 EfficientNets	270
A.4 Intrinsic dimensionality	272
A.5 BiGAN architecture	272
BIBLIOGRAPHY	277

ACRONYMS

ACC	Accuracy	HGP	Human Genome Project
ADR	Adverse Drug Reaction	HNSCC	Head-and-Neck Squamous Cell Carcinoma
BC	Breast Cancer	HapMap	Haplotype map
BiGAN	Bidirectional Generative Adversarial Network	IHC	Immunohistochemistry
BN	Batch Normalization	MCC	Matthews Correlation Coefficient
CNN	Convolutional Neural Network	MLP	Multi Layer Perceptron
CPATH	Computational Pathology	ML	Machine Learning
CT	Computed Tomography	NGS	Next Generation Sequencing
CV	Cross Validation	PD	Persistent Diagram
DAP	Data Analysis Plan	PET	Positron Emission Tomography
DICOM	The Digital Imaging and Communications in Medicine	PH	Persistent Homology
DILI	Drug-induced liver injury	RF	Random Forest
DLR	Deep Learning based Radiomics	ROI	Region of Interest
DL	Deep Learning	SCS	Single-cell Sequencing
DP	Digital Pathology	SVM	Support Vector Machine
FAERS	Adverse Event Reporting System	TCGA	The Cancer Genome Atlas
FCH	Fully Connected Head	TCIA	The Cancer Image Archive
FDA	Food and Drug Administration	TDA	Topological Data Analysis
GTE _x	Genotype-Tissue Expression	TG-GATEs	Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System
GWAS	Genome Wide Association Studies	TILs	Tumor Infiltrating Lymphocytes
HCR	Hand Crafted Radiomics	UMAP	Uniform Manifold Approximation and Projection
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise	WSI	Whole Slide Image
HE	Hematoxylin and Eosin		

Part I

INTRODUCTION

Chapters 1 and 2 introduce the required background and the research topics addressed in this PhD project. The first chapter introduces the precision medicine paradigm and its implications in the clinical and preclinical setting. In particular, the landscape of current high-throughput technologies for biomarker development on Omics and Imaging data is described. Chapter 2 outlines AI approaches for biomarker discovery and patient stratification, with a focus on Radiology, Digital Pathology, and Toxicology. Further, the technical challenges hindering the deployment of AI in clinical practice are detailed, including data preprocessing, model reproducibility and interpretability, and multi-omics integration. The aims and the overall structure of this thesis are summarized in Chapter 3.

BACKGROUND

Highlights

- Precision Medicine aims at patient stratification to improve targeted healthcare and reduce medical malpractice.
- As measurable indicators of physiological and pathological mechanisms, biomarkers are the bases of precision medicine.
- High-throughput technologies have accelerated biomarker discovery through the rapid acquisition of massive data from molecular and imaging sources.
- Radiomics and Digital Pathology are rapidly emerging to improve clinical practice and extract quantitative features from high-resolution medical images.
- **Specific endpoints of the thesis.** Develop AI/ML tools for reproducible stratification over omics, radiology and pathology and their combination. Test in clinical and preclinical applications.

1.1 Precision medicine and why we need it

The COVID-19 outbreak drew worldwide attention to the challenges faced by the healthcare system in dealing with diseases relying on complex biological mechanisms and presenting diverse phenotypes among the affected population. More than 60 subtypes have been proposed for the SARS-CoV-2 patients in 2020, including subgroups based on disease vulnerability (*e.g.*, age, respiratory comorbidities), and response to therapies (due, for example, to genetic variants), to identify more precise biomarkers or at least gain a broader understanding of disease progress, spread, and prognosis [117, 534]. Nonetheless, the plethora of COVID-19 symptoms and manifestations, even among patients with otherwise similar phenotypes, makes it difficult

to predict individual risks or to determine candidates for targeted treatments.

1.1.0.1 *The precision medicine paradigm for targeted healthcare*

The frailty of the *one-size-fits-all* approach is not a surprise; the evidence that individual factors, such as environmental, biological, or genetic conditions drastically affect how patients suffering the same disease would benefit from the same treatment or dosage goes back centuries [490]. The more complex the disease, the more likely the medical intervention based on the average patient population is to fail, resulting in inefficient care. For example, cancer chemotherapy has been reported effective for only 25% of oncology patients [431], and drug ineffectiveness was found as the most frequent Adverse Drug Reaction (ADR) in the US Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS), accounting for 6.4% of all reports [323]. Remarkably, the clinical responsiveness of the top ten selling drugs for widespread disorders in the US, including schizophrenia, heartburn, and arthritis, was estimated at around 25% [351], corresponding to more than 17 million patients with no expected benefit from the prescribed treatment [401].

In contrast, the precision medicine paradigm seeks to improve targeted healthcare by exploiting individual variability to design tailored treatments and increase drug safety (Figure 1). Thus, precision medicine represents an evolution from *reaction* to disease progression, side effects, and ADR, to *prevention*, early intervention, and early detection of disease onset [101]. Other than decreasing trial-and-error prescriptions and directing target therapies, this new treatment strategy has the potential to reduce surgical diagnostic procedures in favour of less invasive techniques (*e.g.*, imaging tests, liquid biopsies), avoiding associated complications and risks.

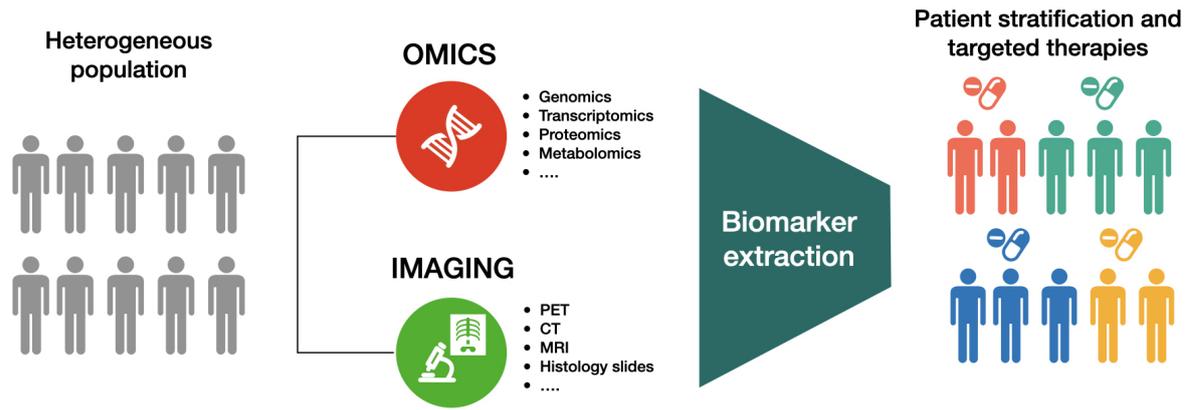


Figure 1: **Precision medicine.** Biomarker extraction from Omics and Imaging data can allow for patient stratification to deliver precise diagnosis/prognosis and avoid drug adverse reactions.¹

1.1.0.2 Controlling the economic waste of medical malpractice

The adoption of precision medicine in clinical practice can also reduce healthcare costs and stem the economic waste resulting from mistaken, excessive diagnoses and failed tests. The concern for overuse in clinical medicine, *i.e.*, the adoption of unnecessary and potentially dangerous procedures [83], is increasing worldwide [55]. In 2012, an analysis of the wasteful spending in the US healthcare system determined that 21% of total expenditures can be categorized as unnecessary, primarily due to overused tests for relatively harmless conditions [39]; an update on these estimates yielded an increase in the evaluated wasteful expenses to 25% in 2019 [423], corresponding to more than \$ 800 billion per year, exceeding the budget of the US federal defense department and almost 15 times higher than the cost of education [38, 144]. Moreover, each pivot trial established by the US FDA to assess the efficacy and safety of novel therapeutic agents requires on average a \$ 19 million investment nearly doubles if the design includes the comparison between placebo and an active drug [328]. The picture is not much more encouraging in Europe, where 10% of hospital expenses is used to amend preventable medical mistakes and treat hospital-acquired infections [344]. For example, in Italy, the estimated cost of medical malpractice for

¹ Original image created with Keynote software.

a single hospital has increased from €2.5 in 2012 to €3.5 million in 2019, with 35% of adverse events are caused by therapeutic or diagnostic errors [302].

Notably, a significant portion of misdiagnosis in clinical practice is due to human cognitive bias [168], as doctors are influenced by their clinical experience and personal biases (e.g. racial prejudice [44, 216]). In particular, one of the leading causes of diagnostic errors is caused by physicians' overconfidence, whether it is unconscious (*therapeutic illusion* [77]) or intentional [258], which results in poor decision-making, distorted evaluations, non-compliance with guidelines, and decreased the adoption of additional tests or decision-supporting resources [78, 258, 285].

Precision medicine is thus expected to have a critical impact on the whole health-care system by (i) increasing the clinical responsiveness of drugs and the patient safety, (ii) containing the overall cost of health, (iii) improving the diagnostic process and reducing the adoption of surrogate endpoints [369], and ultimately promoting patient-centered care through efficient use of physicians' time [340].

1.2 Biomarkers for patient stratification

To fulfill the promises of precision medicine, it is clear that a more comprehensive picture at the molecular and cellular level of disease and human heterogeneity is the *conditio sine qua non* for an highly accurate and precise medical practice [135]. Specifically, the foundation of precision medicine is the discovery and development of novel biomarkers, which can be used to stratify patients into subgroups according to disease susceptibility, pathogenesis, and response to treatment [482].

Biomarker - FDA-NIH Biomarker Working Group definition

A biomarker is a characteristic that is an indicator of normal biological processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions [171].

Biomarkers can be derived from different sources of measurement (e.g. genome, proteins, microbiome, imaging) and serve to varying stages of patient care, includ-

ing diagnostic evaluation (diagnostic biomarkers), assessment of disease progression or relapse (prognostic biomarkers), and planning of therapeutic strategies (predictive biomarkers) [50]. Although such measurable indicators have always been adopted in standard clinical and preclinical practice, the investigation of potential biomarkers has exploded in the last few decades [417], with more than 65 000 papers published each year [507]. Because cancer is one of the utmost complex and heterogeneous diseases with many factors affecting the outcome of patients, starting with early detection [110], the hunt for the ideal biomarker has been especially fruitful in precision oncology. For example, 1582 potential molecular biomarkers, both diagnostic and prognostic, are currently proposed to assess treatment effectiveness and patient survival in Breast Cancer (BC) (MarkerDB database [507]). This rapid evolution of biomarkers for patient and disease stratification has been enabled by technological breakthroughs transforming nearly every aspect of patient care, from diagnostic testing to drug development.

1.2.1 *Omics data*

Two key milestones have paved the way for a digital revolution in healthcare in the genomic field, the Human Genome Project (HGP) and the International Haplotype map (HapMap) project (1990–2003), which have led to the first successful Genome Wide Association Studies (GWAS) in 2005 promoting a new data-driven, hypothesis-free approach in medicine [470]. However, the prohibitive time and costs of the first sequencing techniques (around \$ 3 billion for a single strand of the human genome) initially resulted in a limited number of studies, hindering the excitement for the potential of genome-related analysis [176].

1.2.1.1 *NGS technologies*

The following advent of high-throughput technologies, in particular Next Generation Sequencing (NGS), has allowed for the simultaneously reading of hundreds of thou-

sands of nucleotide sequence reads, slashing the cost of sequencing to less than \$ 1000 in twenty years [231], and thus making genomic sequencing within everyone's reach. The flourishing of NGS led to an unprecedented number of novel discoveries regarding the correlation of DNA with complex traits, confirming the heterogeneity that characterizes individuals and human disease [469]. The increasing evidence in the potential of genetic findings demanded a simultaneous improvement in high-throughput technologies; the standard sequencing approaches estimate the average signal produced by a multitude of different cells (also called bulk sequencing), thus ignoring information about cellular diversity. Single-cell Sequencing (SCS) has enabled the analysis of the genome at cell resolution, needed to identify rare sub-populations of cells or new species, understand their relationship within tissues and organs, or unveil their development and differentiation paths [318]. Single-cell methods have been applied to model organisms and address key medical questions in physiology, pathogenesis and disease progression [452]. In precision oncology, SCS applications include: resolving intratumor heterogeneity, studying invasion during cancer progression, profiling of Circulating Tumor Cells, tracing metastatic dissemination, and investigating adaptive or acquired resistance in chemotherapy [335].

1.2.1.2 NGS for Precision Medicine

The rapid progress of NGS technologies has advanced biomarker discovery and its translation in the routine clinical practice; more than 75000 genetic tests are available on the market for a wide range of medical domains [361], with an increasing number of tests being developed for non-invasive prenatal screening [261], and diagnosis of rare pediatric diseases [136]. Beyond genomics, high-throughput technologies have led to novel insights in other omics fields, including proteomics, metabolomics, epigenomics, and transcriptomics [50, 373]. In particular, the analysis of RNA transcript levels to explore the physiology, activity, and state of cells can complement the purely genetic investigations in providing more accurate biomarkers [298]. NGS techniques for RNA sequencing (RNA-seq) are used to detect and quantify multiple sub-populations of the transcriptome that have different roles in physiological and pathological conditions, such as mRNA sequencing (mRNA-seq) to identify genetic

variations [223] and alternative splicing [530, 532], or miRNA-seq to study small molecules of highly-conserved non-coding RNA (microRNAs or miRNAs) that have a fundamental function in the regulation of targeted genes [67]. RNA-seq is thus instrumental to understand the interconnection between the transcriptome and the phenotype in healthy tissues [289] and organisms [181], as well as biological mechanisms involved in disease occurrence and progression [84, 421, 435, 522]. Transcriptomics is progressively harnessed in cancer research to study tumor heterogeneity and establish biomarkers and therapeutic targets [207]; in particular, SCS techniques for RNA-seq (scRNA-seq) provide a detailed picture of the tumor microenvironment [16, 528] that can help in understanding the role of its immune component in suppressing (*e.g.* Tumor Infiltrating Lymphocytes (TILs)) or promoting tumorigenesis [510].

1.2.1.3 NGS for toxicology

In preclinical research, the high-throughput analysis of the transcriptome has enabled the evaluation of toxicity and ADR resulting from the exposure to xenobiotics [226]. That drugs and environmental chemicals can alter gene expression profiling in target organs through epigenetic mechanisms is well known in toxicology [21, 257, 292]. Several studies have demonstrated the suitability of transcriptomics to detect sensitive biomarkers for early detection of toxicity based on differentially expressed genes [7, 389]. As the liver plays a major role in drug metabolism [477], it is crucial to understand mechanisms of hepatotoxicity of drugs that affect the liver as primary target; a particular class of ADR is Drug-induced liver injury (DILI), encompassing ADRs that cause liver damage. The liver is the most common target of ADR, because of its crucial role in the metabolism of endogenous and exogenous compounds [273]. Predictive markers of DILI able to identify susceptible patients would give an enormous advantage to accelerate safe drug development and to prevent severe reactions after approval [157, 403]. DILI is indeed one of the major concerns in drug development and acute liver failure, accounting for 40% of liver transplants [82], and 32% of drug withdrawals [158]. However, DILI prediction poses particular challenges, as pre-clinical testing for side effects in animals does not automatically transfer to clinical trials and then to post-marketing treatment in the population. Indeed,

individual susceptibility may arise in patients different from those enrolled in trials, or range from clinically serious to worse as a function of interaction with other factors [36].

Further, evaluation of idiosyncratic DILI is one of the most challenging tasks in toxicity evaluation, due to the wide range of manifestations and the lack of specific diagnostic tests [82]. Although animal testing remains the standard way to assess drug safety in DILI studies, there is a constant effort to develop alternative assays that are cheap and non-animal based [226]; as an example, the analysis of global miRNA expression profiling through NGS uncovered four potential biomarkers for DILI prediction associated to a common antifungal medication [268].

The digital revolution in medicine does not only refer to NGS technologies for omics data, but encompasses a broader transformation that includes the advent of high-resolution bioimaging techniques [470]. Modern bioimaging tools are used in research and clinical practice to extract morphological information that results from underlying biological mechanisms, thus linking the molecular nature of a disease with its phenotypic manifestation [196]. Bioimaging biomarkers are a key component of the precision medicine paradigm, with the role of increasing diagnostic accuracy, guiding interventions, and monitoring response to treatments [163].

1.2.2 Radiology

Among radiology imaging techniques, Positron Emission Tomography (PET) and Computed Tomography (CT) scans, often combined together, have a great impact on cancer management and they are now an important part of both oncology clinical practice as well as translational cancer research [142]. PET and CT supply useful and complementary clinical information: PET can recognize functional malignancies, undetectable on CT alone, while CT offers peculiar anatomical information, but can normally identify abnormalities only after structural alterations have occurred. Localizing the exact position of a site of malignant disease often has a significant impact on decisions affecting the diagnosis, prognosis, staging, treatment, and overall pa-

tient management. However, precise localization is difficult in PET first because the spatial resolution with this technique is of some mm (with modern technologies for clinical equipment it is of the order of 4 mm). Second, PET is characterized by poor identification of anatomical structures, since the radiopharmaceuticals are functionally specific and are absorbed differently only in target or target-like tissue; other tissues are not shown in PET image. Although PET has the advantage of being able to sensitively interrogate specific and varied abnormalities in tumor biology, its poorer resolution and variable noise (e.g. time of image acquisition after injection, voxel dimension, respiratory motion or reconstruction algorithm) pose additional technical limitations. Moreover, patho-physiological mechanisms can lead to potentially false-positive and false-negative assessments [372]. In radiology, technological advances have enhanced image reconstruction and scan speed, improving image interpretation and reducing procedure-associated risks such as radiation exposure from CT and PET acquisitions [35].

1.2.2.1 Radiomics

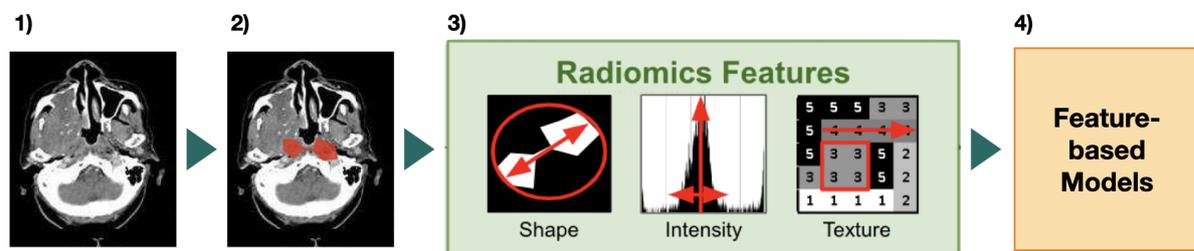


Figure 2: **Radiomics pipeline.** 1) Radiology scan acquisition; 2) lesion segmentation; 3) feature extraction; 4) model building.²

The diffusion of high-throughput scanners and the parallel development of computational tools has implicated the translation of the standard visual assessment of radiology scans into quantitative features (e.g. tissue intensity, shape, or texture) that can serve as imaging biomarkers [164]. In particular, the extraction of minable data from radiology images and their subsequent analysis in terms of computer-vision features,

² Original image created with Keynote software. The sample CT scan is from the TCIA Head-Neck-PET-CT repository.

known as *Radiomics*, aims at providing objective, reproducible measures of a disease evolution and response to treatment [260, 305]. A conventional radiomics pipeline consists of four main tasks: (1) image acquisition/reconstruction; (2) image segmentation; (3) feature extraction and quantification; (4) statistical analysis and model building (Figure 2). Radiomics is rapidly emerging in oncology as a non-invasive way to improve tumor staging, metastasis detection, and prediction of clinical outcomes [425], especially when integrated with genomic profiles (radiogenomics) that complement the pathophysiology description of a tumor [49, 424]. For example, several radiomics and radiogenomics approaches have been adopted to identify imaging biomarkers for diagnostic and prognostic endpoints in Head-and-Neck Squamous Cell Carcinoma (HNSCC), including patient stratification into low or high survival, and toxicity of oncologic therapies [57].

1.2.3 Digital Pathology

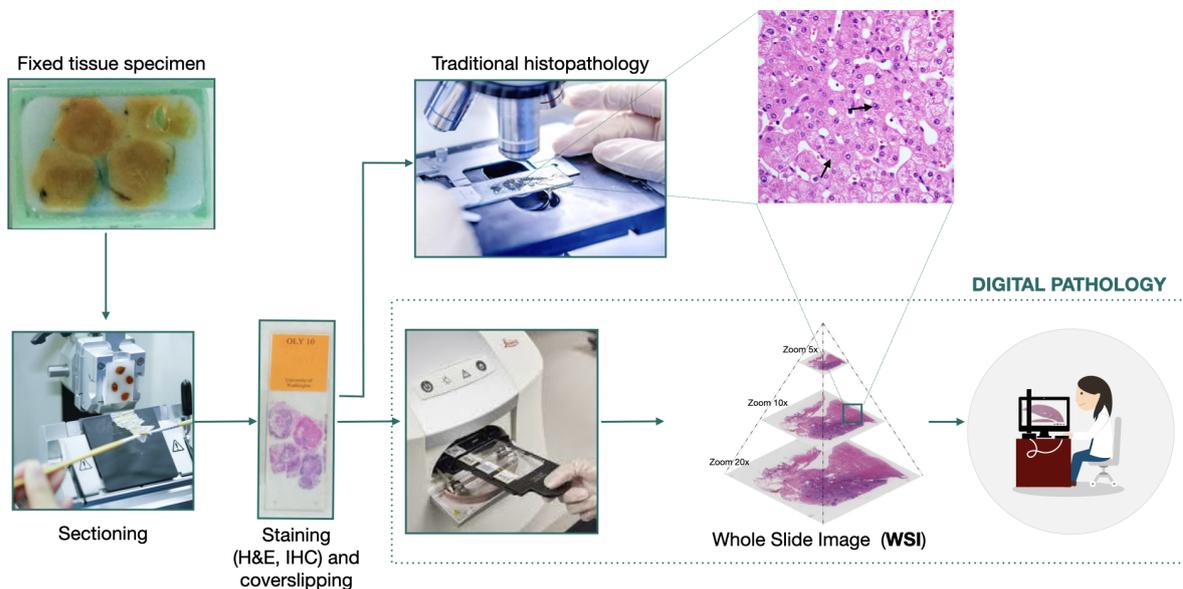


Figure 3: **Digital Pathology workflow.** Glass slides result from several steps to prepare tissue specimens for histopathology assessment; in traditional histopathology, slides are evaluated by pathologists under a microscope. Digital Pathology translates the acquisition and analysis of tissue slides into a digital environment.³

Although technological breakthroughs certainly speed innovation in radiology, possibly the major transformation brought by the digital revolution in medical imaging is occurring in the pathology field. While the use of digitized radiological scans is a long-standing clinical procedure, the histopathology examination of tissue samples has been traditionally performed manually through a sequence of laborious steps: first, organs and tissue fragments are collected during surgery or via biopsy; samples are then processed with chemical substances to prevent damages and putrefaction (*fixation*); samples are cut into thin slices (around 5 μ m), stored on glass supports, and colored with histochemical stains (*e.g.* Hematoxylin and Eosin (**HE**)) to provide contrast; glass slides are finally evaluated by pathologists under optical microscopes. The whole pipeline, from the sample extraction to the delivery of the pathology report, can require up to 10 days [500], and tissue slides of each patient must be stored for 10-20 years after analysis [442]. The morphological characterization of structures and patterns on tissue slides is pivotal to deliver a diagnosis and inform in real time physicians on potential treatment options. In oncology, histopathology is often referred to as the gold standard for tumor diagnosis and grading [220, Chapter 35]. Therefore, the replacement of a centuries-old manual routine [473] with a digital workflow in pathology, namely Digital Pathology (**DP**) (Figure 3), has initially raised quality and efficiency concerns, slowing down its translation into clinical practice [170].

1.2.3.1 *Advantages in clinical and preclinical practice*

The current availability of high-resolution scanners that efficiently replicate glass slides in a digital multi-resolution format, named Whole Slide Image (**WSI**), have established the advantages of **DP** in tracking, automation, and interoperability over the standard manual procedure [349]. In particular, the adoption of a digital environment can facilitate the diagnostic process, allowing pathologists to review the whole tissue section rather than small Region of Interest (**ROI**)s, and improving the ergonomics of traditional settings [493]. Moreover, the use of **WSI** can help reducing laboratory errors or issues resulting from the management and storage of glass slides, such

3 Original image created with Keynote software. Lab images credits: <https://bit.ly/3s7mbYm> (tissue specimen and glass slide), <https://bit.ly/3G5frPu> (sectioning), <https://bit.ly/3rXJdk5> (scanning), and <https://bit.ly/3o9GD9I> (microscope).

as sample mislabelling [333] or stain deterioration [321]. Finally, DP provides an effective and cost-efficient method for remote consultation (telepathology), promoting collaboration and peer-reviewed evaluation among pathologists [415].

In the preclinical space, pathology is a central component in toxicity evaluation and regulatory assessment with animal studies [183]; however, it is a laborious process that requires extensive training; histopathology evaluation is often affected by subjectivity and, unfortunately, discrepancies among pathologists are not uncommon. Ethical aspects are also significant, because intensive use of animal testing is costly and hard to justify in terms of reproducibility and translation from model to human [169]. DP can open the door to the development of alternative assays that have a lesser impact on animals or possibly being non-animal based.

1.2.3.2 *Computational pathology*

The advent of DP represented a paradigm shift in clinical and preclinical pathology, crowned in 2017 by the FDA approval of the first WSI-based device (Philips IntelliSite Pathology Solution) for primary diagnosis of surgical slides. WSI are currently used in clinical practice and research, and serve as building blocks for computational approaches that can enhance image analysis, assist clinical decision, and lead to the identification of sub-visual morphological features [1, 284]. Computational Pathology (CPATH) techniques have been used to automate and refine routine tasks in pathology including nuclei segmentation and count, staining estimation, and metastasis detection as well as to provide new perspective on complex biological mechanisms, such as cell interactions in tumor microenvironment [149, 336].

Chapter 2

AI AND BIOMARKER DISCOVERY

Highlights

- Deep learning (DL) techniques adopt a data-driven learning process that abstract complex relationships in the input.
- The increasing availability of massive public datasets is boosting the applications of Deep and Machine Learning models in the biomedical domain.
- In Radiology, DL can be combined with Radiomics to develop accurate imaging biomarkers.
- In Digital Pathology, DL is applied to automate clinical/preclinical tasks and identify sub-visual histological patterns.
- In toxicology, DL is improving the detection of adverse drug reactions, such as drug-induced hepatotoxicity.
- **General endpoints of the thesis.** Development of (i) a bias free Data Analysis Plan for model development, (ii) standardized procedures for data preprocessing, and (iii) robust mathematical procedures for model development to overcome technical issues in the deployment of medical AI.

2.1 Deep Learning

The flood of outputs generated by the new high-throughput technologies has offered new options for computational tools that can manage the massive amount of data to extract valuable information [1, 60]. Based on an increasing list of success stories, Deep Learning (DL) techniques are already to represent a major breakthrough in diagnosis, therapy decision, prognosis and treatment evaluation [393]. DL refers to a class of machine learning methods that model high-level abstractions in data through the use of modular architectures, typically composed by multiple nonlinear

transformations estimated by data-driven training procedures. DL is now surpassing pattern recognition methods in the most complex medical images challenges such as the ones endorsed by the Medical Image Computing and Computer Assisted Intervention (MICCAI), and it is comparable to expert accuracy in the diagnosis of skin lesions [138], classification of colon polyps [245, 248], ophthalmology [115], radiomics [98] and other areas [275]. In general, deep learning architectures based on Convolutional Neural Network (CNN)s hold state-of-the-art accuracy in numerous image classification tasks without prior feature selection. Further, intermediate steps in the pipeline of transformations implemented by CNNs or other DL architectures can provide a mapping (*embedding*) from the original feature space into a *deep feature* space (Figure 4). Of interest for medical diagnosis, deep features can be used for interpretation of the model and can be directly employed as inputs to other Machine Learning (ML) strategies.

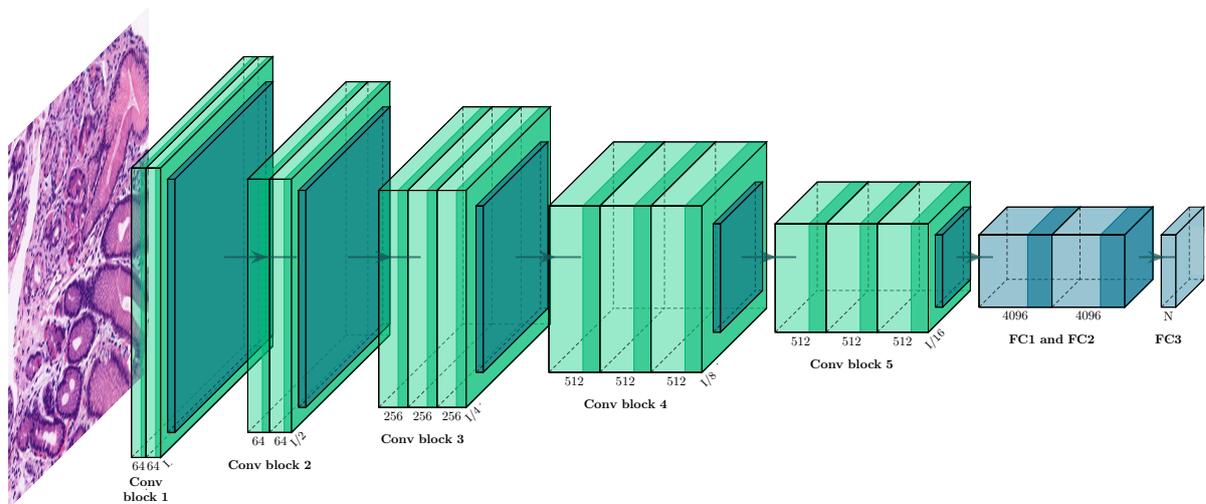


Figure 4: Example of a Convolutional Neural Network (CNN) model for image recognition, composed of five convolutional blocks, and three fully connected layers. A similar CNN is implemented in Chapter 12.¹

¹ Original image created with PlotNeuralNet: <https://bit.ly/3Hna1AE>

2.2 Upscale resources

The increased availability of massive annotated medical data from health systems has led to high expectations about the impact of AI on challenging biomedical problems [393]. Among the available public resources, three massive datasets are particularly noteworthy for clinical and preclinical research:

- **GTEX** [462]. The Genotype-Tissue Expression (**GTEX**) Consortium was established in 2013 to build an extensive public resource to study gene expression and regulation across different tissues and identify risks of disease among healthy individuals. **GTEX** provides a comprehensive collection of genomic, transcriptomics, and histology data; the current release (v8) includes a total of 15,201 **HE**-stained **WSIs**, retrieved with an Aperio scanner (20× native magnification), and 15201 RNA-seq samples, gathered from a cohort of 838 non-diseased donors². To ensure that the collected tissues meet prescribed standard criteria, a Pathology Resource Center validated each sample origin, content, integrity and target tissue³.
- **TCGA** [467]. The Cancer Genome Atlas (**TCGA**) was introduced in 2006 by the National Institutes of Health to explore cancer genomics. **TCGA** has been further extended by The Cancer Image Archive (**TCIA**) to include clinical images (radiology scans and histology slides) matched with genomic profiles [196]. **TCGA** and **TCIA** are currently the most exhaustive public resource for oncology research, providing 2.5 petabytes of molecular and clinical data collected across 33 different cancer types for more than 11K individuals, and 160 imaging collections.
- **TG-GATEs** [214]. The Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System (**TG-GATEs**) collection is the largest public toxicogenomics database to date, developed by the Japanese Toxicogenomics Project consortium (TGP) in 2015 to enhance drug safety assessment in preclinical studies. Toxicology data in **TG-GATEs** includes gene expression profiles, biochemistry, hematology

² <https://bit.ly/3KU0ALi>

³ <https://bit.ly/3r7BS27>

and histopathology findings generated over 10 years from *in vivo* and *in vitro* studies. To evaluate treatment-related toxicity, 170 hepato- or nephro-toxic compounds were administered at multiple dosages and time points. Tissue sections were sampled from the liver and the left kidney of control and treated rats and digitized with an Aperio ScanScope scanner at 20x magnification for a total of 52879 HE-stained WSIs and corresponding genomic profiles (17528 treated and 6183 control livers, 21485 treated and 7683 control kidneys).

2.3 AI and Radiomics

Deep learning frameworks have been successfully applied to address clinical challenges in radiation oncology such as lesion or disease detection, classification, and segmentation, which are crucial for diagnostic purposes and treatment planning [177, 222, 306]. In particular, remarkable improvements in software have been made for PET-CT image processing and analysis [184, 188, 531]. These advancements have been applied in the context of (1) noise filtering and image de-noising; (2) lesion and tumor segmentation; (3) the extraction of quantitative metrics from PET-CT images; and (4) the definition of multi-modal models for disease diagnosis and staging prediction [187]. These four methodological components represent the key building blocks of Radiomics (Section 1.2.2) [491]. Radiomics features are also referred to in the literature as Hand Crafted Radiomics (HCR) [3], as opposed to Deep Learning based Radiomics (DLR), in which features are derived in association to the outcome labels by using DL models. DLR leverages DL models to extract deep features from medical images based on the specification of a predefined task, e.g. disease-diagnostic, or cancer-type prediction [3]. Although hard to describe in biological or morphological terms, the DLR features can outperform the hand-crafted ones [19, 247, 272, 440]. Both the two radiomics approaches have their own advantages and disadvantages, thus it makes room for testing approaches combining HCR and DLR features. Methods to combine HCR and DLR features relies on two separate data fusion approaches: (A) *decision-level* fusion; and (B) *feature-level* fusion [3].

In the decision-level fusion framework, **HCR** and **DLR** features are derived independently, and a voting strategy is then adopted to join the predictions of separately trained models. On the other hand, the feature-level fusion strategy is based on the idea that **HCR** and **DLR** features can be combined into a unique set of features (e.g. using concatenation) that will be further used as input for **DL** models [239].

Notably, combining features from both domains [3, 42], and incorporating different types of features [147] results in significantly improved performance. In particular, the feature-level fusion typically results in more accurate models [89, 342, 355, and Chapter 8].

2.4 AI and Digital Pathology

Similarly to what happened to several other sectors, the novel **DL** paradigm has revolutionized **DP**, leading to a fast growing flow of publications for a wide range of applications in clinical and preclinical setting [2, 28, 132, 178, 215, 430].

Deep learning methods have been applied to analysis of histological images for diagnosis and prognosis [63, 109, 325]. As human assessments of histology are subjective and hard to repeat, **DP** has already allowed pathologists to gain a much more effective diagnosis capability and to dramatically reduce time for information sharing. Starting from the principle that underlying differences in the molecular expressions of the disease may manifest as tissue architecture and nuclear morphological alterations [30], it is clear that automatic evaluation of disease aggressiveness level and patient subtyping has a key role aiding therapy in cancer and other diseases. Digital pathology and the quantification of **TILs** are in particular the cornerstone of the immunotherapy approach [122]. Indeed, quantitative analysis of the immune microenvironment by histology is crucial for personalized treatment of cancer [320, 404]. **TILs** assessment is used for risk prediction models, adjuvant, and neoadjuvant chemotherapy decisions, and for developing the potential of immunotherapy [420, 438]. Digital pathology is also a natural application domain for computer vision machine learning, with the promise of accelerating routine reporting and standardizing

results across trials. Notably, deep learning features learned from digital pathology scans can improve validity and robustness of current clinico-pathological features, up to identifying novel histological patterns, *e.g.*, from [TILs](#).

AI for digital pathology has already demonstrated its utility in clinical applications. The [FDA](#) has cleared: (i) An AI software for detection of potentially cancerous lesions in the lungs. (ii) Marketing of the GI Genius, the first device that uses ML to assist clinicians in detecting lesions (such as polyps or suspected tumors) in the colon in real time during a colonoscopy; (iii) Marketing of Paige Prostate, the first software that uses AI to identify an area of interest on prostate biopsies (FDA Announcement Sept 21, 2021)

The need for a time-consuming and labor-intensive process is however an impeding factor for the actual digitization of existing pathology collections. While AI workflows in Digital Pathology are being implemented for clinical and preclinical pathology for new studies, it is thus hard to profit from precious preclinical data collection, even if already digitized.

2.5 AI and toxicology

AI approaches are accelerating the mechanism of drug development, both by selecting novel molecules [436] and predicting toxicity of existing compounds [161]. To understand pathways of toxicity, AI strategies have been implemented at various stages of drug development (*e.g.* design, biomarker extraction, drug repurposing), using molecular data from *in vivo* and *in silico* studies (predictive toxicology), or imaging data from animal-based experiments such as histological samples (toxicologic pathology). However, the introduction of AI in preclinical safety studies, either performed by pharmaceutical companies or academic research institutions, requires validating new tools and in parallel the development of regulations.

2.5.0.1 Predictive toxicology

Toxicology has traditionally relied on animal models for the assessment of chemical risk. However, ethics and regulatory policies limiting or prohibiting the use of animals, and the need for earlier recognition of toxic molecules are now pushing for replacing *in vivo* assays with *in silico* toxicological methods. Such approaches are developed through different mathematical models [376, 384], possibly integrated with animal-free *in vitro* assays [338]. In particular, algorithmic and technological advances in machine learning have boosted the new paradigm known as predictive toxicology [293]. The relevance of predictive toxicology is supported by initiatives led by public agencies [474] and societies [398], with web platforms offering analytical services (e.g., INSPECT [143]). The main change proposed by predictive toxicology is the shift from detecting adverse effects at the organism level to the identification of biologically significant disruptions of toxicity pathways at the molecular level [398]. Predictive toxicology is a broad term encompassing four main categories of approaches and resources [352]: data analytics (e.g., toxicoinformatics); chemical and toxicity databases (e.g., toxicogenomics and metabolomics); chemoinformatics (e.g., quantum chemical methods for generating molecular descriptors); and, the quantitative structure-activity relationships (QSAR) modeling framework.

In particular, the QSAR framework is based on the assumption that compounds with similar structures are likely to exhibit similar behavior in terms of biological activity or chemical property, including toxicity. Computational models are trained in order to describe such relationships between chemical structures and toxicological processes, and possibly predict the biological activity of additional chemicals out of the training datasets. Known limits of QSAR are issues with validation, model interpretation, and model selection [287]. On the other side, QSAR has made available to machine learning frameworks a critical mass of data. Recently, a hazard database of more than 800K chemical properties for about 81,000 chemicals had been used to train supervised ML models (logistic regression and RF) predictive of hazard labels with previously unseen accuracy for purely *in silico* models [286].

Given the enhanced ability of DL in extracting complex features over traditional ML models, their applications to predictive toxicology is expected to increase. As an example, Stokes et al. [436] fed RNA-seq data to a DL model to predict toxicity of novel compounds for antibiotic discovery. Despite the key importance – both theoretical and applicative – of predictive toxicology, the scientific community is far from having reached an optimal shared solution. Hurdles range from data quality issues to the low specificity (high number of false positives) affecting most of the *in silico* methods [515].

2.5.0.2 Toxicologic pathology

As tissue slides in toxicologic pathology includes a variable number of morphological changes (*e.g.*, necrosis, hypertrophy, cell proliferation), DL models have been explored to triage samples and differentiate between normal and abnormal tissues, or to localize anomalous regions for faster evaluation [208, 313, 364, 399]. Other than decision support systems, DL models may improve translational research by identifying the underlying pathomechanisms of ADRs [327] and linking histological features with molecular signatures [313]. Unlike clinical applications, the adoption of AI solution in toxicologic pathology is still at an early stage due to unique challenges faced in the preclinical setting. In particular, a typical preclinical pathology task is dominated by healthy tissue where the difference between injured and control samples are small [313]; therefore, the impact of expert effort to manually annotate the lesions to be conducted before AI analysis is significant. Recent studies using DL models in toxicologic pathology on manually annotated regions are reviewed in Chapter 12, Section 12.2.2.

2.5.0.3 DILI risk prediction

Given the complexity of DILI status prediction (see Chapter 1, Section 1.2.1.3), a number of groups have developed approaches and strategies to predict DILI from different data types, such as compound chemical structures, gene expression and histopathology collections. Modelling based on chemical structures and molecular descriptors has been broadly used for DILI classification [88, 134, 205, 538]. As genomic infor-

mation can expose adverse reactions not detectable from compound structures only, several studies have focused on the use of gene expression signatures [244, 400, 429]. Few CPATH approaches have been also investigated to evaluate drug-related hepatotoxicity, in particular DL models trained on collections of manually labeled ROIs [364, 371]. However, the scarcity of annotated WSI datasets has limited the application of AI for DILI prediction [478].

2.6 Deployment of medical AI: technical challenges

AI breakthroughs in Digital Pathology, Radiology, and Toxicology are expected to be pervasive over diverse medical domains, and translation to clinical practice is expected to accelerate due to faster regulatory approvals for medical AI [34]. However, multiple technical challenges have to be addressed in order to deploy AI algorithms in healthcare, as suggested in the review of Rajpurkar *et al.* [377]. A structured summary of opportunities and issues summarising the review is organized in the sunburst diagram in Figure 5.

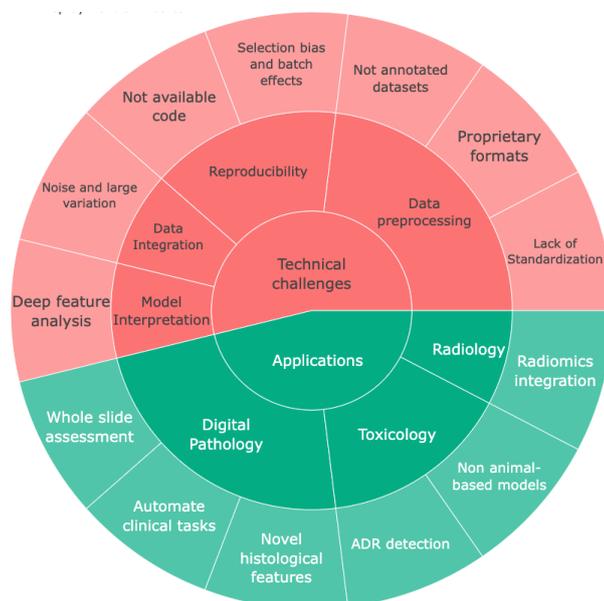


Figure 5: **Deployment of medical AI.** Overview of selected applications and technical challenges of AI models on biomedical imaging and omics data.⁴

⁴ Original sunburst plot realized with the Plotly Python library <https://bit.ly/3rYW59H>.

The challenges range from issues in model development to their validation, while other are specific of the domain of the data modality:

1. **Reproducibility** is possibly the major concern regarding the clinical adoption of AI systems. Predictive models should output comparable results when trained repeatedly on a specific dataset (*reproducibility*), and perform consistently across equivalent data on the same task (*replicability*). Reproducibility issues often arise from engineering aspects that introduce bias in datasets and AI pipelines. Reproducibility and replicability of AI models would be boosted by code availability and the adoption of software engineering guidelines, a practice not yet widely adopted in the research community [277].
2. **Data processing** is particularly challenging in AI pipelines for medical imaging datasets, due to the large size of data, the heterogeneity of image acquisition systems, and the lack of standardized procedures [504]. Above all, Digital Pathology is the most complex medical imaging modality. Differently from radiology that rely on a consensus protocol to store imaging data (The Digital Imaging and Communications in Medicine (DICOM)), WSI can be retrieved in many digital formats [40]. Moreover, a standard procedure to process WSI, usually too large to fit GPUs, for model development is still missing [407]. Finally, the shortage of precise labels on digital slides can require supervised models to work in a *weak-label* mode, possibly reducing their predictive accuracy.
3. **Deep feature analysis.** Understanding the internal decision process of predictive models is still an unresolved issue [133, 350]. As AI aims to evolve its status from exploratory to clinically effective solutions, explainability of predictive models remains a major hindrance [133, 350]. In particular, the analysis of the inner state of DL models (*deep features*) can help both in inferring novel patterns, and improving error detection in AI pipelines.
4. **Multi-omics integration** The problem of data integration in computational biology is far from having a consolidated and shared solution. The rapid extraction of high-dimensional, high-throughput features from different sources provided

by modern technologies requires collaboration between biologists, computer scientists, physicians and other experts. The lack of common methodologies and terminologies can transform this synergy into a further level of complexity in the process of data integration [413]. As observed in [68, 370], specific technological limits, noise levels and variability ranges affect the different omics, thus confounding the underlying biological signals. As a result, really integrative analysis is still rare, as evidenced by the lack of consistency in the published results [307, 509].

In the following sections, these four challenges are detailed, along with the description of available tools that can help mitigate technical issues concerning medical AI.

2.6.1 *Reproducibility of predictive models*

Reproducibility is a paramount concern in biomarker research [217], and in science in general [24], with scientific communities, institutions, industry, and publishers struggling to foster adoption of best practices, with initiatives ranging from enhancing reproducibility of high-throughput technologies [419] to improving the overall reuse of scholarly data and analytics solutions (e.g. the FAIR Data Principles [503]).

Although the landscape seems to have improved [558], and broad efforts have been spent across different biomedical fields [334], computational reproducibility and replicability still fall short of the ideal. Lack of reproducibility has been linked to inaccuracies in managing batch effects [549, 559], small sample sizes [541], or flaws in the experimental design such as data normalization simultaneously performed on development and validation data [539, 551].

*Data leakage*⁵

Among the various types of selection bias that threaten the reproducibility of machine learning algorithms, *data leakage* is possibly the most subtle one [560]. Data

⁵ This section is adapted from Bussola et al 2021 [62].

leakage refers to the use of information from outside the training dataset during model training or selection [555]. A typical leakage occurs when data in the training, validation and/or test sets share indirect information, leading to overly optimistic results. For example, one of the preclinical sub-dataset in the MAQC-II study consisted of microarray data from mice triplets. These triplets were expected to have an almost identical response for each experimental condition, and therefore they had to be kept together in data partitioning to circumvent any possible leakage from training to internal validation data [463].

Despite openness in sharing algorithms and benchmark data is a solid attitude of the machine learning community, the reliable estimation on a given training dataset of predictive accuracy and stability of DL models (in terms of performance range as a function of variations of training data) and the stability of deep features used by external models (as the limited difference of top ranking variables selected by different models) is still a gray area.

2.6.1.1 The Data Analysis Plan

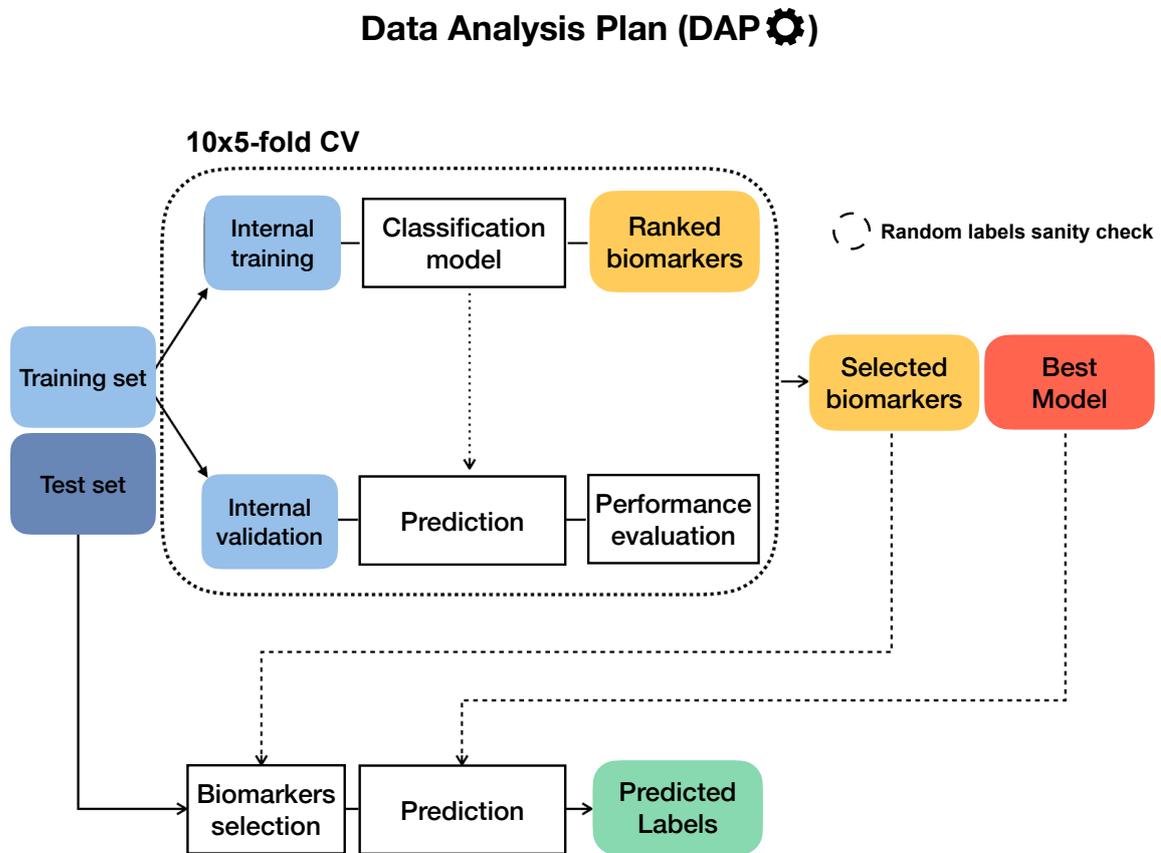


Figure 6: Diagram of the **DAP**, originally developed within the FDA-led MAQC/SEQC-II initiatives [463]. If the training set labels are stochastically shuffled beforehand, the **DAP** runs in *random labels* mode as a sanity check to ensure that the procedure is not affected by systematic bias.

To assess the impact of engineering aspects on reproducibility, the MAQC/SEQC projects adopt a Data Analysis Plan (**DAP**) that forces bioinformatics teams to submit classification models, top features ranked for importance and performance estimates all built on training data only, before testing on unseen external validation data (Figure 6). The **DAP** approach is methodologically more robust than a simple **CV** [463] as the internal Cross Validation (**CV**) and model selection phase is replicated multiple times (e.g., 10 times) to smooth the impact of a single training/test split; the performance metrics is thus evaluated on a much larger statistics. Also, features are analyzed and ranked multiple times, averaging the impact of a small round of par-

titions. The ranked feature lists are fused in a single ranked list using the Borda method [227] and the bootstrap method is applied to compute the confidence intervals. This approach helps mitigate the risk of selection bias in complex learning pipelines [151], where the bias can stem in one of many preprocessing steps as well as in the downstream machine learning model. Further, it clarifies that increasing task difficulty is often linked to a decrease not only in accuracy measures but also of stability of the biomarker lists [227], *i.e.* the consistency in the selection of the top discriminating features across all repeated cross validation runs.

The input dataset is first partitioned in two separate datasets, the *training set* and the *test set*, also referred as *external validation set* as reported in [463, 465]. The external validation set is kept completely unseen to the model, and it is only used in the very last step of the DAP for the final model evaluation. The training set further undergoes a 5-fold CV iterated 10 times, resulting in 50 separated *internal validation sets* used for model evaluation within the DAP.

At each CV iteration, features are ranked by KBest, with ANOVA F-score as the scoring function, and the model is trained on an increasing number of ranked features (namely: 10%, 25%, 50%, 100% of the total number of features). A list of top-ranked features is also obtained by Borda aggregation of the ranked lists

As for model evaluation, the preferred metric is the Matthews Correlation Coefficient (MCC), which in a general multi-class problem is defined as [25, 229, 304]:

$$MCC = \frac{\sum_{k,l,m=1}^N (C_{kk}C_{ml} - C_{lk}C_{km})}{\sqrt{\sum_{k=1}^N \left[\sum_{l=1}^N C_{lk} \sum_{\substack{f,g=1 \\ f \neq k}}^N C_{gf} \right]} \sqrt{\sum_{k=1}^N \left[\sum_{l=1}^N C_{kl} \sum_{\substack{f,g=1 \\ f \neq k}}^N C_{fg} \right]}}, \quad -1 \leq MCC \leq 1 \quad (1)$$

where N is the number of classes and C_{st} is the number of elements of true class s that have been predicted as class t .

MCC is widely used in Machine Learning as a performance metric, especially for unbalanced sets, for which accuracy can be misleading [95]. In particular, MCC gives

an indication of prediction robustness among classes: $MCC=1$ is perfect classification, $MCC=-1$ is extreme misclassification, and $MCC=0$ corresponds to random prediction. Finally, the overall performance of the model is evaluated across all the iterations (*i.e.*, internal validation sets), in terms of average MCC and ACC with 95% Studentized bootstrap confidence intervals (CI) [124], and then on the external validation set. The optimality of MCC with respect to other metrics such as $F1$ has been recently described in [224]. As a sanity check to avoid unwanted selection bias effects, the DAP can be repeated stochastically scrambling the training set labels (*random labels* mode); in this mode, a procedure unaffected by selection bias should achieve an average MCC close to 0. An example of *random labels* schema in Digital Pathology is illustrated in Chapter 5, Section 5.2.

2.6.2 Data processing in Digital Pathology

The application of AI models in Digital Pathology reflects the serious technical challenges faced in the collection of histopathological images, starting from the manual preparation of the tissue specimen to the acquisition of the digital images (Section 1.2.3). The protocol adopted in clinical practice for the slide preparation includes a diversity of decisive steps that can alter the quality of the slide (*e.g.*, tissue artifacts, storage conditions) and encumber the complete standardization of the process [407].

In particular, $WSIs$ are typically retrieved at a resolution of 20x or 40x [520]. While a high magnification is important to study relevant structures in the tissue, it also represents a technical difficulty: a biopsy specimen scanned with magnification factor 40x has a resolution of $\sim 0.25\mu\text{m}/\text{pixel}$ and a color depth of 24 bits. Therefore, approximately 48 MB are needed to represent only 1mm^2 of tissue. As the typical glass slide is much bigger (around 25mm^2), the corresponding WSI file is a $10^5 \times 10^5$ RGB images (called Gigapixel) which typically exceeds the GB [520], making the time required for a single complete human analysis almost prohibitive. Further, even though many compression techniques exist, their adoption is generally not advised, due to the potential introduction of artifacts hiding possibly interesting patterns. Moreover, the availability of different scanning platforms results in a wide range of

proprietary formats for storing the [WSIs](#) [40], further hindering the implementation of an homogeneous processing. Although several efforts have been made towards data standardization in [DP](#) [197, 349], a common [WSI](#) format is still lacking [40].

2.6.2.1 *WSI preprocessing*

In practice, [DL](#) pipelines for [CPATH](#) usually require extensive image preprocessing. The first issue is that often the [WSI](#) manipulation methods are developed *ad hoc* and kept undescribed, thus making data and results hard to reproduce. For example, due to their very large size, [WSIs](#) are first divided into smaller patches (tiles) to fit the GPU memory for model training (Figure 7).

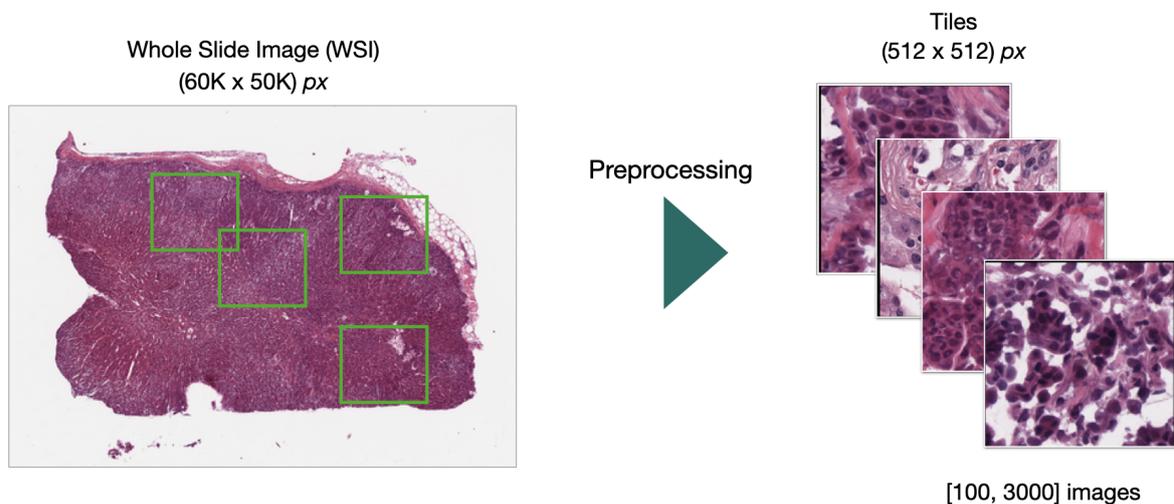


Figure 7: Example of tile extraction in AI pipelines for Digital Pathology. (left) Thumbnail of a Gigapixel WSI (GTEx public repository) (right) a variable number of tiles is extracted to train AI models.⁶

Multiple options are possible in the definition of the tiling procedure, with critical impact on model performance and computational cost [407]. Hence, the control of the tile extraction strategy is a key aspect for reproducibility of the results. Similarly, specific (*e.g.*, tissue and [WSI](#) artifact detection methods) or generic (*e.g.*, image morphology operators) procedures introduce significant variability and may introduce bias in the AI models applied to [WSIs](#). Notably, a limited generalizability of the developed algorithms stays as one of the biggest obstacles for the large-scale

⁶ Original image created with Keynote software.

implementation of computational pathology approaches for routine clinical application [256].

The second problem for AI models as tools for scientific discovery and application in pathology is the potential for flaws in the code hidden in the preprocessing pipelines. Identifying errors or corner cases is especially critical in machine learning workflows that target healthcare-specific data structures. Currently, most DL pipelines for DP include custom preprocessing code that is tailored to the data and task of interest.

Among tools for WSI analysis and processing, there is a growth of solutions that are publicly available and open-source, suggesting an effort towards promoting reproducibility and collaborative research, also in clinical tasks [64, 130, 300, 330, 437]. However, available tools are rarely provided with a testing suite, thus potentially contributing to the lack of reproducibility of AI in medical applications [75, 179]. Risk can be high, as for influential coronavirus simulations sharply criticized for the lack of systematic software testing and quality control [427]. This thesis will introduce the `histolab` Python package as a dedicated solution for reproducible preprocessing of WSI images (Chapter 6).

2.6.2.2 *Weak label*

The annotations of labels available for WSIs, in particular for retrospective collections, are often given as one for to the slide. This is typically an effect of associating the slide to a clinical pathology general evaluation, which rarely includes a specific marking or annotation of the pattern or patterns characterizing the diagnosis. In CPATH applications, given the high complexity of a WSI, the computation is typically performed at the level of tiles, or sub-images, extracted from a WSI. In absence of predefined annotations at tile level, unsupervised DL models, *e.g.*, BiGAN (see Appendix A.5), and dimensionality reduction techniques, *e.g.*, UMAP (see Section 2.6.3.1), can be used to transform tiles into lower-dimensional representations to find patterns of similarity [313]. On the other hand, weakly supervised approaches recover the annotation at slide level to train models on sub-images collections, each tile obtaining the label of the original slide. Weakly supervised strategies can be successfully trained for

clinical tasks that do not require pixel-wise annotations, such as tissue recognition (see Chapter 4), or tumor subtyping [132]. In toxicologic pathology, however, normal histology on WSIs exceeds atypical patterns (see Section 2.5.0.2), which can be found at different locations and with different grades of severity [296]. Therefore, weakly supervised learning is highly likely to perform poorly in toxicologic pathology. In addition, training classification models on tile set requires a careful planning of the model training, not to incur in unwanted biases such as the data (or information) leakage: whenever tiles extracted from the same WSI occur in both the training and the validation set, model results are heavily affected by overfitting (see Chapter 5). This thesis will first introduce a framework for toxicologic pathology using weak labels to train DL models on compressed WSIs, thus avoiding the tiling procedure (Chapter 12).

2.6.3 Deep feature analysis

During the training process, DL models learn high-level representations of the input data; these intermediate features can be extracted at different depths of a neural network, avoiding the need for feature engineering.

Deep feature - DeepAI glossary

A deep feature is the consistent response of a node or layer within a hierarchical model to an input that gives a response that is relevant to the model's final output. [118].

Deep features are usually high-dimensional and hard to interpret, as they abstract complex connections between the data. Several mathematical descriptors have been applied for the unsupervised analysis and clustering of the deep features, in order to investigate the behaviour of predictive models, to infer novel relationships, or to be used as independent predictors.

2.6.3.1 UMAP and HDBSCAN

This thesis makes intensive use of dimensionality reduction (or projection) methods, which is in general a mathematical transformation that reduces the space dimensionality under constraints on conservation of distances or enhancement of statistical properties, e.g. maximization of variance as in the Principal Component Analysis (PCA) method. Modern dimensionality reduction methods have been extensively used with machine learning models for model interpretation or just in the attempt of compressing data complexity. Uniform Manifold Approximation and Projection (UMAP) is a novel dimensionality reduction technique introduced in 2018 by McInnes and colleagues [310, 311] with roots in the fields of algebraic topology and Riemannian geometry (see Appendix A.1). UMAP is a manifold learning algorithm projecting high-dimensional data in lower spaces. The underlying hypothesis is that data lie on one or more manifolds, whose structure UMAP tries to approximate. UMAP can capture the local and global structure of high-dimensional data more accurately than many alternative dimensionality reduction algorithms (e.g. t-SNE [480]) and it is computationally faster [33]. UMAP has also been adopted to investigate artificial neural networks. One example is the activation atlas by Carter and colleagues [76], using UMAP to explore the distribution of activation maps from hidden layers of an Inception V1 network [446], enlightening how different filters of the Artificial Neural Network are correlated. Another example is [402], where UMAP loss is extended to DL thus improving classifier performance by better capturing data structure. Nonetheless, initialisation seems to be critical and deserves special care [242]. The ability to vary the embedding dimensionality allows UMAP to be used for more than just data visualization: for instance UMAP enhance clustering performance of the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) unsupervised algorithm [10].

HDBSCAN

HDBSCAN [66] is an extension of the classic DBSCAN algorithm [137], improved by providing a hierarchical structure of clusters found from density estimation and a

more intuitive approach for cluster selection. The density based approach can identify clusters with arbitrary shapes, thus overcoming limitations of algorithms that are able to work only with convex clusters such as K-Means. The main advantage of [HDBSCAN](#) relies on the simplicity of tuning its key hyperparameters, namely the minimum cluster size, and the number of neighbors used to estimated the density for each point in the dataset. The hierarchical, density based approach is also robust with respect to subsampling. Furthermore, the [HDBSCAN](#) algorithm can count on a really fast implementations [312]. An example of unsupervised clustering with [HDBSCAN](#) is reported in Figure 8 for a subset of the MNIST handwritten digits datasets; clustering performance of partitioning algorithms (*e.g.*, K-Means) are improved when [HDBSCAN](#) is applied on features reduced with the [UMAP](#) algorithm.

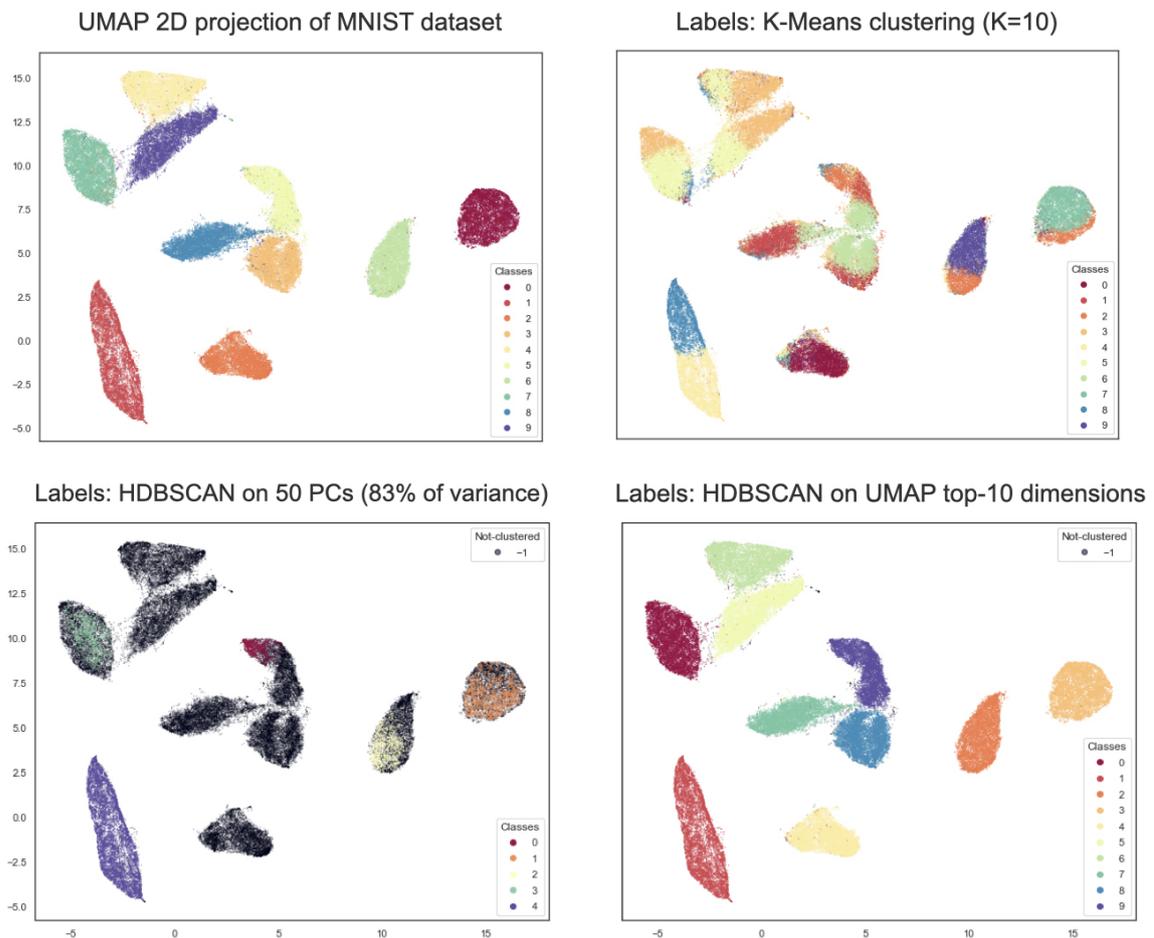


Figure 8: Example of 2D UMAP projection and unsupervised clustering on the MNIST handwritten digits dataset colored by: A) ground truth labels; B) KMeans labels; C) HDBSCAN labels computed on reduced 50 Principal Components; D) HDBSCAN labels computed on reduced 10 UMAP dimensions.⁷

2.6.3.2 *Topological Data Analysis*

Topological Data Analysis (TDA) is a recent approach to data analysis relying on concepts from algebraic topology [72, 86], providing solid qualitative and often also quantitative information about the geometric structure of the considered dataset (see Appendix A.2). In particular, TDA allows the description of topological properties of data as point clouds, time series analysis [359], images [73] or even volumetric and time varying data [387]. From a computational point of view, a great effort has recently been put into building efficient TDA algorithms, data structures and software libraries such as Ripser [31], Mapper [428], and Giotto-TDA [454]). The grounded theoretical framework and the performing implementations make TDA a powerful data science tool, effectively used nowadays by several labs worldwide for a wide range of applications: a non-exhaustive list of recent examples in bioinformatics is [15, 294, 388, 418, 494].

In deep learning pipelines, TDA techniques can overcome standard analytical tools in providing a rigorous characterization of complex relationships within the data, and allowing for the simultaneous collection of informative features, supporting the critical step of model interpretability and explainability [194].

2.6.4 *Multi-omics integration*

The underlying hypothesis of multi-omics integration is that different omics data can provide complementary information [68], although sometimes redundant [80], and thus a broader insight with respect to single-layer analysis, for a better understanding of disease mechanisms [234]. This assumption has been confirmed by multiple studies on diverse diseases, such as cardiovascular disease [266], diabetes [368], liver disease [120], or mitochondrial diseases [237], and also longitudinally [453], suggesting that the more complex the disease the more advantageous the integration. As the co-occurrence of multiple causes and correlated events is a well-known charac-

⁷ Original plots created with Python libraries: UMAP <https://bit.ly/3IJmAGX> for the UMAP projection, scikit-learn [357] for PCA and K-Means clustering, and hdbscan <https://bit.ly/3s1VEvc> for the HDBSCAN clustering.

teristic of tumorigenesis and cancer development, the integration of data generated from multiple sources can thus be particularly useful for the identification of cancer hallmarks [81, 155, 279, 409].

2.6.4.1 Integration strategies

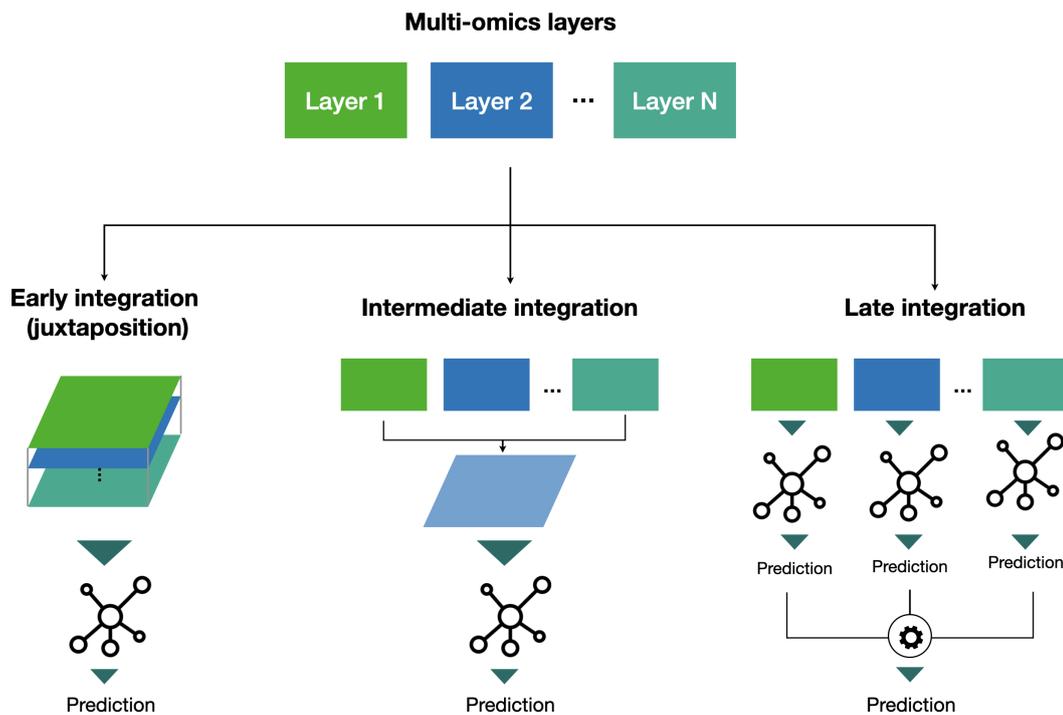


Figure 9: Early, Intermediate, and Late integration strategies for model prediction on an arbitrary number N of multi-omics layers. (left) Early integration (juxtaposition) concatenates all layers into a single matrix for model prediction. (middle) Intermediate integration derives a common representation of the input layers before model learning. (right) Late integration combines predictions independently obtained on each omics layer.⁸

Many computational strategies have been introduced that combine multiple types of data to identify novel biomarkers and thus to predict a phenotype of interest or drive the development of intervention protocols. Given the heterogeneity of data and tasks, these techniques deal with the data integration at different levels of the learning process (Figure 9): (i) by concatenating the features before fitting a model (*early-integration*), (ii) by incorporating the integration step into the model training

⁸ Original image created with Keynote software.

(*intermediate-integration*), or (iii) by combining the outputs of distinct models for the final prediction (*late-integration*) [270, 492].

In the *early-integration* approach, also known as juxtaposition-based, the multi-omics datasets are first concatenated into one matrix. To deal with the high dimensionality of the joint dataset, these methods generally adopt matrix factorization [68, 370, 434, 509], statistical [18, 112, 120, 266, 317, 322, 374, 391, 509, 521], and machine learning tools [322, 432, 509]. Alternatively, data models relying on polyglot approaches can be used especially in bioinformatics applications [97, 385]. Although the dimensionality reduction procedure is necessary and may improve the predictive performance, it can also cause the loss of key information [270]. Moreover, biomarkers identified purely on a computational statistics rationale from meta-omics features often lack biological plausibility [380].

In order to maximize the contribution of the single-omics layer, the *late-integration* methods first model each dataset individually, and then merge or average the results; they are also known as model-driven [492, 563]. Although these techniques avoid the pre-selection of the features, they do not leverage the hidden correlations between the data, posing again the risk of signal loss [154, 380].

The *intermediate-integration* strategies aim at developing a joint model that accounts for the correlation between the omics layers, to boost their combined predictive power [481]. Among these methods, the network-based models refer to the reconstruction of a graph representing the complex biological interactions [322, 536], known or predicted, between the variables to discover novel informative relationships [37]. They have successfully been applied in cancer research for the identification of pan-cancer drug targets [564], the detection of subtype-specific pathways [481, 486] and of genetic aberrations [126], or the stratification of cancer patients [379, 514, 529]. In particular, Koh and colleagues [243] predicted breast cancer subtypes by applying a modified shrunken centroid method in the development of their network-based tool, iOmicsPASS. Further, breast cancer datasets in TCGA represent a benchmark for integrative models [232, 236, 511], as well as AML [314].

More recently, the success of deep learning algorithms in various bioinformatics fields [271] prompted the adoption of deep neural networks for omics-integration

in precision oncology. Autoencoders and convolutional neural networks have been effectively trained for the prediction of prognostic outcomes [80, 366], response to chemotherapeutic drugs [416], and gene targeting [358], by adopting either an *early-integration* [80, 358] or a *late-integration* [366, 416]. Although integrative deep learning models hold the potential to include image-derived features in the integration workflow, they suffer from interpretability and generalization issues [195].

It is clear that no single method is consistently preferable, and that most of the proposed approaches are task and/or data dependent [380]; however, the complexity of tumor analysis suggests that network-based approaches are needed [471, 486].

Omics-integration is one of the most promising and demanding challenge of modern bioinformatics, and that there is an urgent need to prove the reproducibility, interpretability, and generalization capability of the proposed methods [37, 517].

This thesis will explore different methods for variable integration, and in particular for omics in Chapter 11.

A general layout of the path to integration of imaging and omics is described in Chapter 3, which summarizes the thesis project aims.

PROJECT AIMS

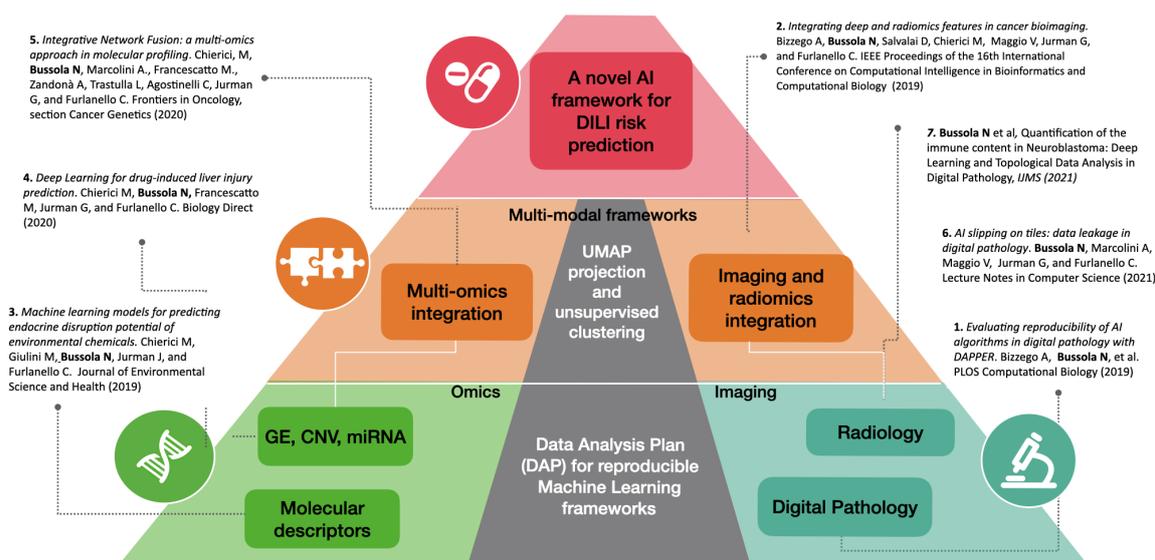


Figure 10: Summary of thesis aims and main results. Three levels were progressively completed moving from single modality to multimodal analysis on integrated data. A central methodological pillar (in gray) was also developed for reproducible analysis and for subgroup analysis after dimensionality reduction.

Thesis Aims

This thesis aims at the development of an AI framework intended for accurate patient stratification and precise diagnostics/prognostics in clinical and preclinical applications. A summary of aims and main results is structured in Figure 10. The thesis aims are depicted as subsequent steps and solved by developing dedicated components that are composed at higher levels. The first level steps aimed at implementing Machine and Deep Learning pipelines for diverse biomedical data modalities, namely Bioimaging (at the radiology and pathology levels) and Omics (different types), in a reproducible environment. Also at the first level step, the thesis aimed to develop

the methodological DAP component. The second level step aimed to evaluate the benefits of multi-omics and multi-modal integration to improve predictive frameworks and extract biologically meaningful features. For methodology, the second step required the adaptation of dimensionality reduction and unsupervised methods. The last step aimed at combining the different methods into reproducible AI framework potentially operating on Imaging, Omics, or combined modalities to extract improved predictive biomarkers for relevant clinical or preclinical questions.

Thesis outline

Novel Deep Learning and Machine learning frameworks are introduced, which operate on single modality data, namely Imaging (Chapters 4–7) and Omics (Chapters 9 and 10), and investigate the combination of multi-modal (Chapter 8) and multi-omics sources (Chapter 11). Strategies to overcome technical challenges that arise during the development of AI pipelines on medical data are systematically adopted, and novel solutions are implemented to address specific issues in Digital Pathology (Chapters 5 and 6). Finally, Chapters 12 and 13 aim to apply the acquired knowledge to open preclinical pathology problems and toxicogenomics. Two fundamental questions are evaluated: (i) Methodological: Can AI better characterize injuries on histopathology samples than a pathologist? (ii) Biological: Can a better safety biomarker in toxicology be derived from imaging, genomics, or their combination? The two topics are investigated for DILI prediction on the TG-GATEs toxicological data collection, as the largest reference data resource for toxicology.

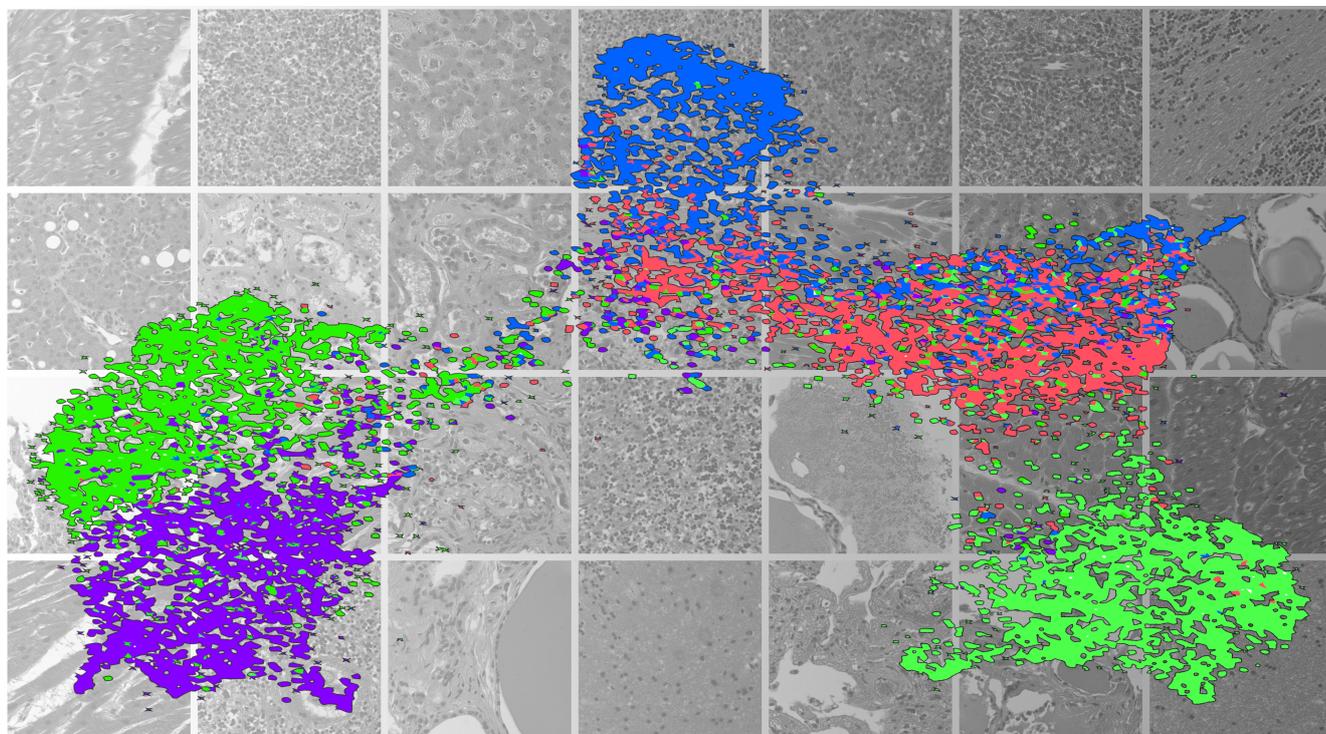
An indication of the scientific publication that is associated to the chapter is indicated in a front page. The striking images in the front pages are all original artwork, unless indicated. All Chapters start with a light blue box including the main insights and results. The last item in the box is usually highlighting the personal contribution to the research.

Part II

PREDICTIVE MODELS IN DIGITAL PATHOLOGY

Chapters 4–7 include DL frameworks in clinical pathology and strategies to overcome technical issues in AI pipelines for Digital Pathology. Chapter 4 introduces the DAPPER framework for reproducibility of predictive models on extensive WSI collections. The impact of selection bias on AI classifiers caused by flawed WSI preprocessing protocols is estimated in Chapter 5. The histolab solution introduced in Chapter 11 is the first open-source library for data preprocessing in Digital Pathology that relies on a comprehensive test suite to guarantee software robustness and modularity. histolab has been adopted to standardize the preprocessing step in the development of complex designs, as the DL framework described in Chapter 7 for automated lymphocyte detection on IHC-stained WSIs. The potential of the proposed approach for TILs assessment in immuno oncology applications is proved on an original collection of Neuroblastoma samples.

THE DAPPER FRAMEWORK



Authors: A. Bizzego, N. Bussola, M. Chierici, V. Maggio, M. Francescato, L. Cima, M. Cristoforetti, G. Jurman, C. Furlanello. *Original title:* Evaluating reproducibility of AI algorithms in digital pathology with DAPPER. *Published in:* PLOS Computational Biology (March 2019)

Chapter 4

THE DAPPER FRAMEWORK

Highlights

- DAPPER is a novel framework to assess reproducibility of AI models in classification pipelines for Digital Pathology.
- DAPPER evaluates stability and predictive power of ML classifiers with a rigorous DAP and unsupervised analysis of the deep features.
- DAPPER is tested on 787 WSIs from the GTEx public repository to identify tissue of origin on HE-stained tiles using an increasing number of classes (5, 10, 20, 30) to train the DL models.
- Classification results indicate a high predictive ability of AI models on normal histology samples, with $MCC > 0.7$ and narrow CIs on external validation data for all the experiments.
- The morphological similarity/dissimilarity of different tissues and organs is reflected in the UMAP visualization of the deep features extracted from the trained classifiers.
- The DAPPER framework and the tile collection are publicly released as a basis for standardisation and validation of AI models in Digital Pathology.

Personal contribution

I contributed to the design of the DAPPER framework and performed several main steps of the experimental pipeline, including data preprocessing and DL/ML model training. I collaborated personally with the expert pathologist (L. Cima) for model accuracy validation. I also significantly contributed to the writing and prepared the figures for the article.

4.1 Abstract

Artificial Intelligence is exponentially increasing its impact on healthcare. As deep learning is dominating computer vision tasks, its application to digital pathology is natural, with the promise of aiding in routine reporting and standardizing results across trials. Deep learning features inferred from digital pathology scans can improve validity and robustness of current clinico-pathological features, up to identifying novel histological patterns, *e.g.*, from tumor infiltrating lymphocytes. In this study, we examine the issue of evaluating accuracy of predictive models from deep learning features in digital pathology, as an hallmark of reproducibility. We introduce the DAPPER framework for validation based on a rigorous Data Analysis Plan derived from the FDA's MAQC project, designed to analyze causes of variability in predictive biomarkers. We apply the framework on models that identify tissue of origin on 787 Whole Slide Images from the GTEX project. We test three different deep learning architectures (VGG, ResNet, Inception) as feature extractors and three classifiers (a fully connected multilayer, Support Vector Machine and Random Forests) and work with four datasets (5, 10, 20 or 30 classes), for a total of 53,000 tiles at 512×512 resolution. We analyze accuracy and feature stability of the machine learning classifiers, also demonstrating the need for diagnostic tests (*e.g.*, random labels) to identify selection bias and risks for reproducibility. Further, we use the deep features from the VGG model from GTEX on the KIMIA24 dataset for identification of slide of origin (24 classes) to train a classifier on 1,060 annotated tiles and validated on 265 unseen ones. The DAPPER software, including its deep learning pipeline and the Histological Imaging – Newsy Tiles (HINT) benchmark dataset derived from GTEX, is released as a basis for standardization and validation initiatives in AI for digital pathology.

This study provides three main practical contributions to controlling for algorithmic bias and improving reproducibility of ML algorithms for digital pathology:

1. A DAP specialized for digital pathology, tuned on the predictive evaluation of deep features, extracted by a network and used by alternative classification heads. To the best of our knowledge, this is the first study where a robust model

validation method is applied in combination with the DL approach. We highlight that this methodology can be adopted in other medical/biology domains in which AI is rapidly emerging, *e.g.*, in the analysis of radiological images.

2. A benchmark dataset (HINT) of 53,727 tiles of histological images from 30 tissue types, derived from GTEx (see Section 2.2) for the recognition of tissue of origin of up to 30 classes. The HINT dataset can be used by other researchers to pretrain the weights of DL architectures that shall be applied on digital pathology tasks (*e.g.*, detection of TILs thus accelerating the training of application-specific models. In the past 5 years, having a shared image dataset (*e.g.*, the ImageNet) allowed the development of a number of DL models for general image classification (*e.g.* VGG, ResNet, AlexNet). Such pretrained networks have then been effectively applied on a variety of different tasks. With the HINT dataset we aim at favouring a similar boost on digital pathology.
3. An end-to-end machine learning framework (DAPPER) as a baseline environment for predictive models in digital pathology, where end-to-end indicates that the DAPPER framework is directly applied to the digital pathology images, with the deep learning component producing features for the machine learning head, without an external procedure (*e.g.*, a handcrafted feature extraction) to preprocess the features. To the best of our knowledge, this is the first example of a DL approach for the classification of up to 30 different tissues, all with the same staining, which represents, *per se*, a valuable contribution to the digital pathology community.

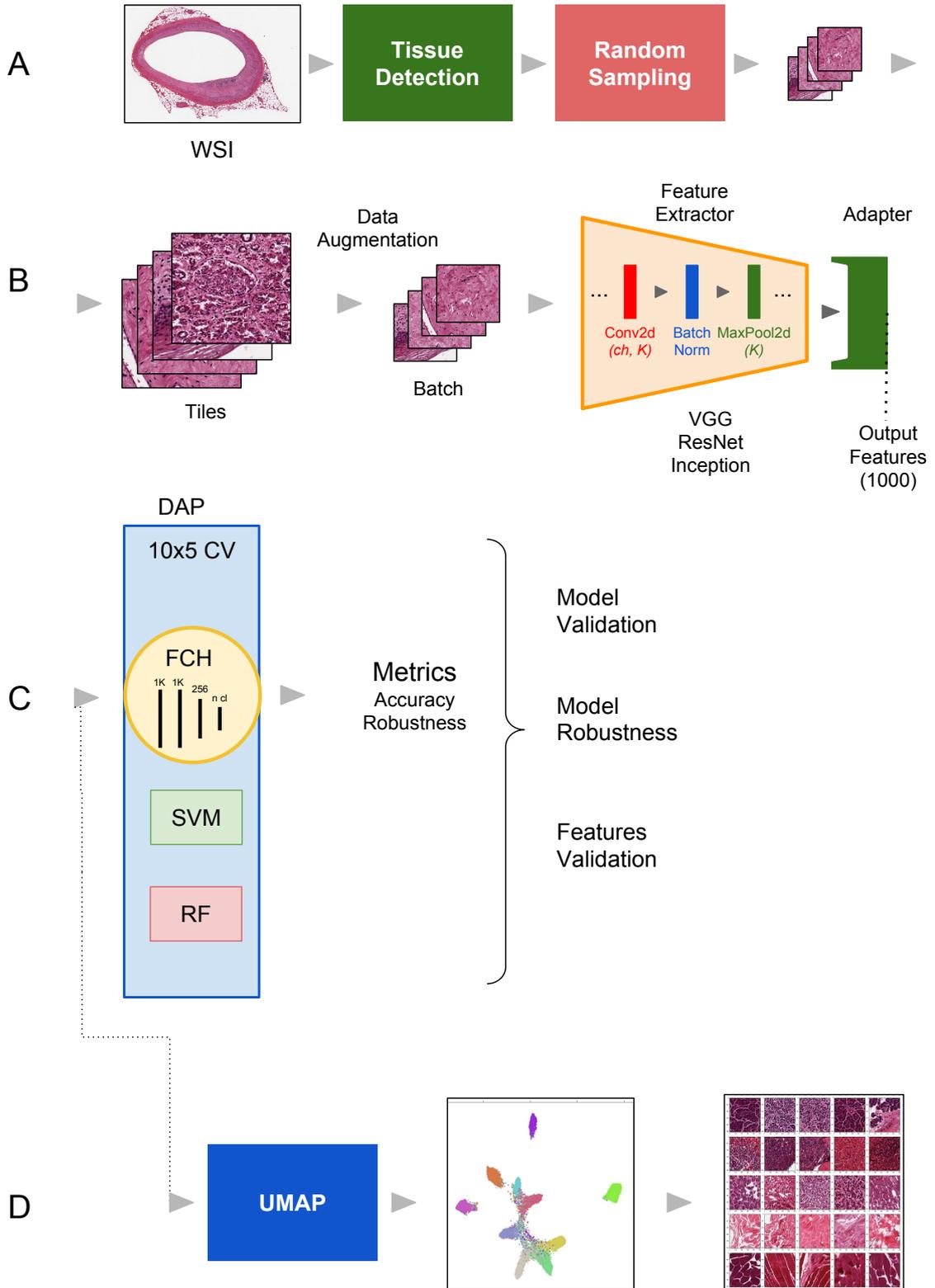


Figure 11: The DAPPER environment. Components: A) The **WSI** preprocessing pipeline; B) the deep learning backend to extract deep features; C) the **DAP** for the machine learning models; and D) the **UMAP** module and other modules for unsupervised analysis.

Summary

We first apply DAPPER to a set of classification experiments on 787 WSIs from GTE_x. The framework (see Figure 11) is composed by (A) a preprocessing component to derive patches from WSIs; (B) a 3-step machine learning pipeline with a data augmentation preprocessor, a backend deep learning model, and an adapter extracting the deep features; (C) a downstream machine learning/deep learning head, *i.e.* the task specific predictor. In our experiments, we evaluate the accuracy and the feature stability in a multiclass setting for the combination of three different deep learning architectures, namely VGG, ResNet and Inception, used as feature extractors, and three classifiers, a fully connected multilayer network, Support Vector Machine (SVM) [108] and Random Forest (RF) [201]. This component is endowed with the DAP, *i.e.*, a 10×5 CV (5-fold cross validation iterated 10 times). The 50 internal validation sets are used to estimate a vector of metrics (with confidence intervals) that are then used for model selection. In the fourth component (D) we finally provide an unsupervised data analysis based on the UMAP projection method (See Section 2.6.3.1), and methods for feature exploration. The DAPPER software is available together with the HINT benchmark dataset as a collection of Jupyter notebooks at <https://bit.ly/3rY0Pw8>.

Notably, the DAP estimates are provided in this paper only for the downstream machine learning/deep learning head in component (C); whenever computational resources are available, the DAP can be expanded also to component (B). Here we kept as a separate problem the model selection exercise on the backend deep learning architecture in order to clarify the change of perspective with respect to optimization of machine learning models in the usual training-validation setting.

As a second experiment, in order to study the DAPPER framework in a transfer learning condition, we use the deep features from the VGG model trained on a subset of HINT on the 1,300 annotated tiles of the KIMIA Path24 dataset [20] to identify in this case the slide of origin (24 classes).

Previous work on classifying *WSIs* by means of neural networks was introduced by [20, 255], also with the purpose of distributing the two original datasets KIMIA Path960 (KIMIA960) and KIMIA Path24 (KIMIA24). KIMIA24 consists of 24 *WSIs* chosen on purely visual distinctions. Babaie and coauthors [20] manually selected a total of 1,325 binary patches with 40% overlap. On this dataset, in addition to two models based on Local Binary Patterns (LBP) and Bag-of-Visual-Words (BoVW), they applied two shallow *CNNs*, achieving at most 41.8% accuracy. On the other hand, KIMIA960 contains 960 histopathological images belonging to 20 different *WSIs* that, again on visual clues, were used to represent different texture/pattern/staining types. The very same experimental settings as the one for KIMIA24, *i.e.*, LBP, BoVW and *CNN*, has been replicated on this dataset by Kumar and coauthors [255]. In particular, the authors applied AlexNet or VGG16, both pretrained on ImageNet, to extract deep features; instead of a classifier, accuracy was established by computing similarity distances between the 4,096 features extracted. Also, Kieffer and coauthors in [238] explored the use of deep features from several pretrained structures on KIMIA24, controlling for the impact of transfer learning and finding an advantage of pretrained networks against training from scratch. Conversely, Alhindi and coworkers [9] analyzed KIMIA960 for slide of origin (20 slides preselected by visual inspection), and similarly to our study they compared alternative classifiers as well as feature extraction models in a 3-fold *CV* setup. Considering the importance of clinical validation of predictive results [235], we finally compared the performance of the DAPPER framework with an expert pathologist. DAPPER outperforms the pathologist in classifying tissues at tile level, while at *WSI* level performance are similar.

DAPPER represents an advancement over previous studies, due to the *DAP* structure and its application to the large HINT dataset free of any visual preselection.

4.2 Materials and methods

Dataset

The images used to train the models were derived from the [GTE_x Study](#) [462].

A custom Python script was used to download 787 [WSIs](#) through the Biospecimen Research Database (total size: 192 GB, average 22 [WSIs](#) for each tissue). The list of the downloaded [WSIs](#) is available in Table 10.

A data preprocessing pipeline was developed to prepare the [WSIs](#) as training data (see Figure 12).



Figure 12: **Tissue detection pipeline.** The identification of the tissue bounding box is performed on the WSI thumbnail in three steps: a) Binarization of the grayscale image by applying Otsu thresholding; b) Binary dilation and filling of the holes; c) Selection of the biggest connected region as tissue region and computation of the vertex of the containing rectangle.

The [WSIs](#) have a resolution of $0.275 \mu\text{m}/\text{pixel}$ (Magnification 40X) and variable dimensions. Further, the region interested by the tissue is only a portion of the [WSI](#) and it varies across the samples. Hence, we first identified the region of the tissue in the image (see Figure 12), then we extracted at most 100 tiles (512×512 pixel) from the [WSIs](#), by randomly sampling the tissue region. We applied the algorithm for the detection of the tissue region on each tile and rejected those where the portion of the tissue was below 85%. A total number of 53,727 tiles was extracted, with a number of tiles per tissue varying between 59 (for Adipose - Visceral (Omentum)) and 2,689 (for Heart - Left Ventricle). Four datasets (HINT₅, HINT₁₀, HINT₂₀, HINT₃₀) have been derived with increasing number of tissues for a total of 52,991 tiles; the full number of tiles per anatomical zone, for each dataset, is available in Table 11 and summarized in Table 1. We refer to the four sets as the HINT collection, or the HINT

dataset in brief. We choose the five tissues composing HINT₅ based on exploratory experiments, while the other three datasets were composed including the tissues with higher number of tiles.

The class imbalance is accounted for by weighting the error on predictions. In detail, the weight w of the class i used in the cross entropy function is computed as: $w_i = n_{\max}/n_i$, where n_{\max} is the number of tiles in the class with more tiles and n_i is the number of tiles in the class i .

Since image orientation should not be relevant for the tissue recognition, the tiles are randomly flipped (horizontally and vertically) and scaled, following a common practice in deep learning known as *data augmentation*. Data augmentation consists of different techniques (such as cropping, flipping, rotating images) performed each time a sample is loaded, so that the resulting input image is different at each epoch. Augmentation has proven effective in multiple problems, increasing the generalization capabilities of the network, preventing overfitting and improving models performance [111, 331, 498].

Such randomized transformations were found to provide more comparable performance between the prognostic accuracy of the deep learning SCNN architecture and that of standard models (*i.e.*, Support Vector Machine, Random Forest) based on combined molecular subtype and histologic grade [325]. In addition, each tile is cropped to a fixed size, which is dependent on the type of network used to extract the deep features.

Table 1: **Summary of the HINT datasets.** Total: total number of tiles composing the dataset; Min: number of tiles in the class with less samples; Max: number of tiles in the class with more samples; Average; average number of tiles for each class.

Name	# tissues	Total	Min	Max	Average
HINT ₅	5	8,218	1,009	2,424	1,643.6
HINT ₁₀	10	22,885	1,890	2,689	2,288.5
HINT ₂₀	20	40,516	1,574	2,689	2,025.8
HINT ₃₀	30	52,991	957	2,689	1,766.4

Deep learning architectures and training strategies

We exploited three backend architectures commonly used in computer vision tasks:

1. VGG, Net-E version (19 layers) with Batch Normalization (BN) layers [426];
2. ResNet, 152-layer model [191];
3. Inception, version 3 [445].

These architectures have reached highest accuracy in multiclass classification problems over the last 4 years [69] and differ in resource utilization (see Table 2).

Table 2: Backend architecture statistics.

Name	Output features	#Parameters	Layers
VGG	25,088	155×10^6	19
ResNet	2,048	95×10^6	152
Inception	2,048	35×10^6	42

The feature extraction layer of each backend network is obtained as the output of an end-to-end pipeline composed of the following main blocks (see panel B in Figure 11):

1. Data augmentation: the input tiles are processed and assembled into batches of size 32;
2. Feature Extractor: series of convolutional layers (Conv2d: with different number of channels and kernel size), normalization layers (BN) and pooling layers (MaxPool2d: with different kernel size) designed to fit with the considered backend architecture (VGG, ResNet, Inception). The number of output features of the Feature Extractor depends on the structure of the backend architecture used;
3. Adapter: as the backend networks have output features of different sizes, we add a linear layer at the end of the Feature Extractor, in order to make the

pipeline uniform. The Adapter takes the features of the backend network as input and output a fixed number of features (1,000).

The 1000 Adapter features are then used as input for a classifier providing predicted tissue labels as output. As predictive models, we used a linear SVM with regularization parameter C set to 1, a RF classifier with 500 trees (both implemented in *scikit-learn*, v0.19.1) and a Fully Connected Head (FCH), namely a series of fully connected layers (see panel C in Figure 11). Inspired by [325] and [200], our FCH consists of four dense layers with 1,000, 1,000, 256 and *number of tissue classes* nodes, respectively. The feature extraction block was initialized with the weights already trained on the ImageNet dataset [121], provided by *PyTorch* (v0.4.0) and frozen. Training also the weights of the feature extraction block improves accuracy (see Table 12). However, these results were not validated rigorously within the DAP and therefore they are not claimed as generalized in this study.

For the optimization of the other weights (Adapter and FCH) we used the Adam algorithm [543] with the learning rate set to 10^{-5} and fixed for the whole training. We used the cross entropy as the loss function, which is appropriate for multiclass models.

The strategy to optimize the learning rate was selected based on results of a preparatory study with the VGG network and HINT5. The strategy approach with fixed learning rate achieved the best results (see Table 13) and was therefore adopted in the rest of the study.

Data Analysis Plan

Following the rigorous model validation techniques proposed by the MAQC projects, we adopted a DAP to assess the validity of the features extracted by the networks (see Chapter 2, Section 2.6.1.1). In our experimental settings, we used 80% of the total samples for the training set, and the remaining 20% for the external validation set. A stratification strategy upon the classes of tiles, *i.e.*, 5, 10, or 20, has been adopted in the

partitioning. We also report the Canberra stability with a computational framework designed for sets of ranked biomarker lists [227].

Beyond *MCC*, we also considered Accuracy (*ACC*) for model evaluation, in its multiclass generalization [25, 229, 304]:

$$ACC = \frac{\sum_{k=1}^N C_{kk}}{\sum_{i,j=1}^N C_{ij}}, \quad 0 \leq ACC \leq 1 \quad (2)$$

where N is the number of classes and C_{st} is the number of elements of true class s that have been predicted as class t .

Experiments on HINT

We designed a set of experiments to provide indications about the optimal architecture for deep feature extraction, while keeping fixed the other hyper-parameters. In particular we set batch size (32) and number of epochs (50), large enough to let the network converge: we explored increasing numbers of epochs (10, 30, 50, 100) and, since the loss stabilizes after about 35 epochs, we set the number of epochs to 50. First, we compared the three backend architectures on the smallest dataset HINT₅, with fixed learning rate. Both VGG and ResNet architectures achieved good results, outperforming Inception as shown in Table 4 and Table 5. In successive analyses we thus restricted to use VGG and ResNet as feature extractors and validated performance and features with the *DAP*. The same process was adopted on HINT₁₀ and HINT₂₀. An experiment with 30 tissues has also been performed. Results are listed in Table 14. A summary of the experiments with considered backend architectures is reported in Table 3.

Table 3: Summary of experiments with the backend architectures.

Experiment	Dataset	Feature extractor	Version/Model
VGG-5	HINT ₅	VGG	Net-E+BN
ResNet-5	HINT ₅	ResNet	152-layer
Inception-5	HINT ₅	Inception	3
VGG-10	HINT ₁₀	VGG	Net-E+BN
ResNet-10	HINT ₁₀	ResNet	152-layer
VGG-20	HINT ₂₀	VGG	Net-E+BN
ResNet-20	HINT ₂₀	ResNet	152-layer

Table 5: Accuracy values for each experiment, and classifier head pairs on HINT dataset. The average cross validation ACC with 95% CI and ACC on the external validation set are reported. Best-performing backend network, and classifier head combination on each dataset are reported in bold.

Experiment	FCH		SVM		RF	
	H-ACC _t	H-ACC _v	H-ACC _t	H-ACC _v	H-ACC _t	H-ACC _v
VGG-5	87.2 (87.0, 87.5)	85.6	82.9 (82.7, 83.1)	82.1	79.9 (79.7, 80.1)	79.7
ResNet-5	90.3 (90.1, 90.5)	90.7	88.1 (88.0, 88.3)	87.2	86.3 (86.1, 86.5)	87.9
Inception-5			79.8 (79.5, 80.0)	78.7	76.2 (75.9, 76.4)	75.9
VGG-10	90.6 (90.5, 90.7)	90.5	87.5 (87.3, 87.6)	88.0	90.0 (89.9, 90.2)	89.7
ResNet-10	87.2 (87.0, 87.3)	87.4	84.3 (84.1, 84.4)	84.9	86.1 (85.9, 86.2)	86.5
VGG-20	78.2 (78.1, 78.4)	78.5	74.1 (74.0, 74.2)	74.4	77.3 (77.2, 77.4)	77.7
ResNet-20	76.7 (76.6, 76.9)	76.9	79.9 (79.8, 80.0)	80.3	75.1 (75.0, 75.2)	75.2

Experiments on KIMIA₂₄

In the second experiment, we used VGG on the KIMIA₂₄ dataset with the deep features extracted by VGG on [GTEx](#); the task is the identification of the slide of origin (24 classes). In the DAPPER framework, classifiers were trained on 1,060 annotated tiles and validated on 265 unseen ones.

Table 4: MCC values for each experiment, and classifier head pairs on HINT dataset. The average cross validation MCC with 95% CI (**H-MCCt**), and MCC on the external validation set (**H-MCCv**) are reported. Best-performing backend network, and classifier head combination on each dataset are reported in bold.

Experiment	FCH		SVM		RF	
	H-MCCt	H-MCCv	H-MCCt	H-MCCv	H-MCCt	H-MCCv
VGG-5	0.841 (0.838, 0.843)	0.820	0.786 (0.783, 0.789)	0.777	0.750 (0.748, 0.753)	0.747
ResNet-5	0.879 (0.877, 0.881)	0.883	0.852 (0.850, 0.854)	0.840	0.829 (0.827, 0.832)	0.849
Inception-5			0.747 (0.744, 0.750)	0.734	0.703 (0.699, 0.706)	0.701
VGG-10	0.896 (0.894, 0.897)	0.894	0.861 (0.859, 0.862)	0.866	0.889 (0.888, 0.891)	0.886
ResNet-10	0.857 (0.856, 0.859)	0.860	0.825 (0.824, 0.827)	0.832	0.845 (0.843, 0.847)	0.850
VGG-20	0.771 (0.770, 0.772)	0.774	0.729 (0.727, 0.730)	0.731	0.761 (0.760, 0.762)	0.766
ResNet-20	0.756 (0.754, 0.757)	0.757	0.788 (0.787, 0.789)	0.792	0.738 (0.737, 0.739)	0.738

4.3 Results

Results of the tissue classification tasks in the DAPPER framework are listed in Table 4 for *MCC* and Table 5 for *ACC*, respectively. See also Figure 13 for a comparison of *MCC* in internal cross validation with external validation.

All backend network-head pairs on HINT have *MCC* > 0.7 with narrow CIs, with estimates from internal validation close to performance on the external validation set (Figure 13). Agreement of internal estimates with values on external validation set is a good indicator of generalization and potential for reproducibility. All models reached their top *MCC* with 1,000 features. On HINT₅ and HINT₁₀, the *FCH* neural network performs better than *SVM* and *RF*. As expected, *MCC* ranged close to 0 for random labels (tested for *SVM*, results not shown).

The most accurate models both for internal and external validation estimates were the ResNet+*FCH* model with *MCC*=0.883 on HINT₅, the VGG+*FCH* model on HINT₁₀, and the ResNet+SVM model on HINT₂₀. In Table 15 we show the results with a lower number of dense layers in the *FCH*, which are comparable with the *FCH* with 4 dense layers. Results on HINT₃₀ are detailed in 14; on the external validation set, the VGG model reaches accuracy *ACC*= 61.8% and *MCC*= 0.61. Performance decreases for more complex multiclass problems. Notably the difficulty of the task is also complicated by tissue classes that are likely to have similar histological patterns, such as misclassification of Esophagus-Muscularis (*ACC*: 72.1%) with Esophagus-Mucosa (*ACC*: 53.2%), or the two Heart tissue subtypes or the 58 Ovary

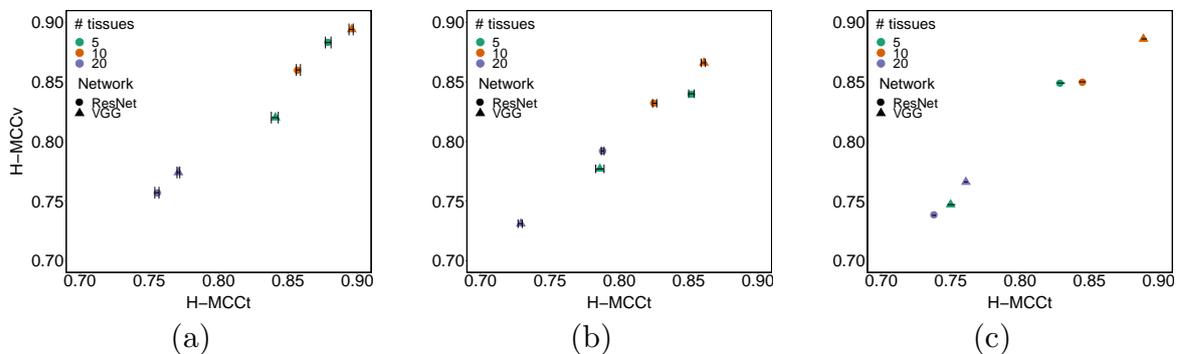


Figure 13: Comparison of DAPPER cross validation *MCC* (*H-MCCT*), vs *MCC* on external validation (*H-MCCv*) performance for each classifier. (a) *FCH*; (b) *SVM*; (c) *RF*.

	Adrenal Gland	Heart-Atrial Appendage	Esophagus-Mucosa	Brain-Cortex	Vagina	Skin-Not Sun Exposed	Brain-Cerebellum	Uterus	Muscle-Skeletal	Thyroid	Pituitary	Esophagus-Muscularis	Spleen	Pancreas	Testis	Prostate	Kidney-Cortex	Ovary	Liver	Heart-Left Ventricle
Adrenal Gland	203	2	8	0	7	0	2	3	0	4	16	6	3	2	8	10	13	3	13	0
Heart-Atrial Appendage	1	238	5	3	0	6	0	0	3	5	0	0	0	0	0	3	2	3	0	12
Esophagus-Mucosa	11	13	171	1	24	11	1	16	3	5	4	27	2	0	1	21	1	6	0	4
Brain-Cortex	0	1	1	314	0	1	13	0	0	0	0	0	0	1	2	0	0	0	0	0
Vagina	4	8	44	2	193	3	0	19	3	4	3	11	1	1	8	9	0	14	0	0
Skin-Not Sun Exposed	2	5	12	1	5	368	0	2	0	2	0	1	0	1	4	1	1	11	0	0
Brain-Cerebellum	2	0	1	18	1	0	351	0	0	0	1	0	0	1	0	0	1	0	0	0
Uterus	1	1	9	0	26	0	0	262	0	2	3	3	1	3	3	13	3	29	1	0
Muscle-Skeletal	0	2	4	0	0	1	0	1	324	2	0	2	0	0	1	0	0	0	0	10
Thyroid	5	4	12	0	9	8	1	1	1	305	5	2	1	0	2	3	10	3	0	0
Pituitary	19	1	8	1	7	3	5	1	1	5	260	2	2	6	6	1	18	2	4	2
Esophagus-Muscularis	2	6	49	1	16	6	1	5	0	2	3	289	2	5	2	7	2	1	1	1
Spleen	3	0	2	0	1	0	0	3	0	1	3	0	422	2	1	7	1	0	0	0
Pancreas	6	1	6	1	9	2	0	2	0	2	7	2	5	407	1	9	3	0	0	0
Testis	7	6	8	4	4	2	2	1	1	1	15	0	1	4	370	10	9	3	0	0
Prostate	11	10	28	0	29	6	3	22	7	7	11	14	2	7	3	298	7	12	1	0
Kidney-Cortex	13	1	7	0	0	1	0	5	0	4	18	1	2	2	2	4	393	2	11	0
Ovary	6	4	18	1	25	13	1	58	0	7	6	1	0	3	11	13	3	354	2	0
Liver	18	1	3	0	1	0	1	0	0	2	6	0	2	3	0	0	6	1	475	0
Heart-Left Ventricle	4	13	2	0	2	1	0	2	9	6	0	3	0	0	1	9	1	1	2	509

Figure 14: Confusion matrix for ResNet+SVM model on HINT20. Red shaded cells indicate the most confused classes.

(ACC: 68.3%) tiles predicted as Uterus (ACC: 72.8%). The full confusion matrix for ResNet with SVM on HINT20 is reported in Figure 14.

In this paper we establish a methodology to evaluate reproducibility and predictive accuracy of machine learning models, in particular of the model selection phase. This is obtained by moving from a single training-test split procedure to an evaluation environment that uses data replicates and averaged statistical indicators, thus enabling to select a model on the basis of statistical indicators derived from the inter-

nal validation loop. In this framework, we can honestly evaluate model performance differences along a set of experiments on a group of tasks. The DAPPER framework cannot by itself identify the reason of such difference, and indeed the emergence of optimal architectures for a specific task may be due to different factors, as revealed by appropriate experimental design. In terms of the experimental design described in this paper, for any model type we expect and find a decrease in accuracy for an increasing number of classes, which requires learning more decision surfaces with less data per class. Notably, the best model in the internal DAPPER validation is confirmed to be the best also on the unseen test data, with a value within the confidence interval or immediately close.

Results on KIMIA24

Table 6: Performance of DAPPER framework for VGG backend network, and classifier heads (FCH, SVM, RF) on KIMIA24 dataset. The average cross validation MCC (**K24-MCCt**), and ACC (**K24-ACCt**) with 95% CI, as well as MCC (**K24-MCCv**), and ACC (**K24-ACCv**) on external validation set are reported.

Model	K24-MCCt	K24-MCCv	K24-ACCt	K24-ACCv
VGG+FCH	0.317 (0.306, 0.327)	0.207	34.4 (33.2, 35.2)	23.8
VGG+SVM	0.446 (0.439, 0.454)	0.409	47.1 (46.4, 47.8)	43.4
VGG+RF	0.457 (0.449, 0.465)	0.409	48.0 (47.3, 48.8)	43.4

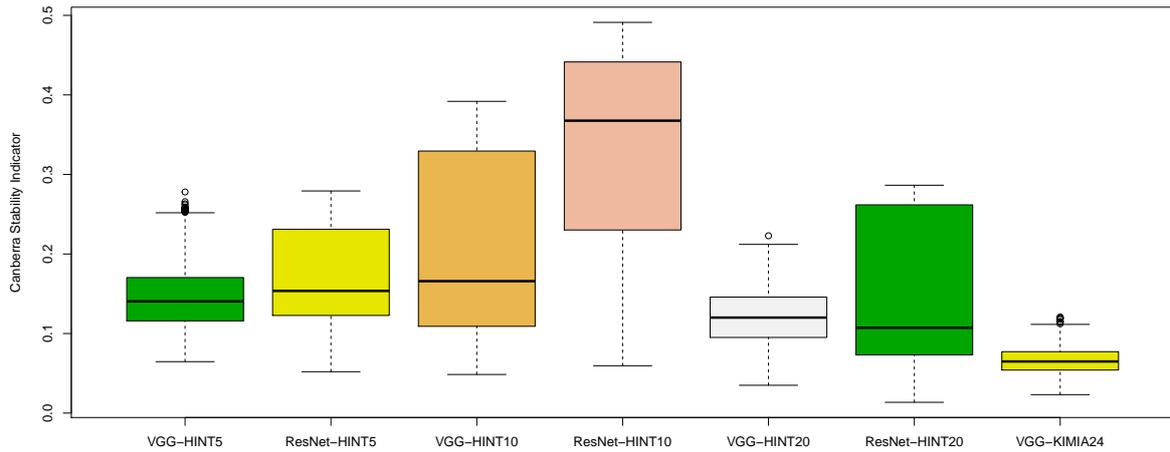


Figure 15: Canberra stability indicator on HINT and KIMIA datasets. For each architecture, a set of deep feature lists is generated, one list for each internal run of training in the nested cross validation schema, each ranked with KBest. Canberra stability is computed as in [227]: lower is better.

Regardless of difference in image types, VGG-KIMIA24 with both RF and SVM heads with $ACC=43.4\%$ (see Table 6), improving on published results ($ACC=41.8\%$ [20]).

It is worth noting that transfer learning from ImageNet to HINT restricts training to the Adapter and FCH blocks. In one-shot experiments, MCC further improves when the whole feature extraction block is retrained (see Table 12). However, the result still needs to be consolidated by extending the DAP also to the training or retraining of the deep learning backend networks to check for actual generalization. The Canberra stability indicator was also computed for all the experiments, with minimal median stability for ResNet-20 (Figure 15).

Results at WSI-level

We evaluated the performance of DAPPER at WSI-level on the HINT20 external validation set, with the ResNet+SVM model. In particular, all the predictions for the tiles are aggregated by WSI, and the resulting tissue will be the most common one among those predicted on the corresponding tiles. However, it is worth noting that the number of tiles per WSI in the HINT20 external validation set varies (min 1, max 31)

due to a stratification strategy only considering the tissues-per-sample distribution (see Section *Data Analysis Plan*). Therefore, we restricted our evaluation to a subset of 15 WSI per class (300 WSI in total), each of which associated to 10 tiles randomly selected. This value represents a reasonable number of Regions of Interest (ROIs) a human pathologist would likely consider in his/her evaluations. In this regard, we further investigate how the DAPPER framework performs on an increasing number of tiles per WSI, namely 3, 5, 7, and 10. As expected, the overall accuracy improves as the number of tiles per WSI increases, reaching 98.3% when considering all 10 tiles per WSI. Notably, the accuracy is high even when reducing to 3 tiles per WSI (see Table 7).

Table 7: Metrics at WSI-level for increasing number of tiles per WSI. Metrics are computed on a subset of HINT20 external validation set, consisting of 15 WSI per class (300 WSI in total). The WSI class is determined by the most frequently predicted class by the *ResNet+SVM* model for the considered tiles.

# Tiles per WSI	MCC	ACC (%)
3	0.86	86.3
5	0.93	93.7
7	0.96	96.0
10	0.98	98.3

Comparison with pathologist

We tested the performance of DAPPER against an expert pathologist on about 25% of the HINT20 external validation set, 2,000 tiles out of 8,103, with 100 randomly selected tiles for each class. We asked the pathologist to classify each tile by choosing among the 20 classes of the HINT20 dataset, without imposing any time constraint. The confusion matrix resulting from the evaluation of tiles as produced by the pathologist is shown in Figure 16. Predictions produced by the DAPPER framework for comparative results are then collected on the same data. The best-performing model on the HINT20 dataset, namely the *ResNet+SVM* model, has been considered for

this experiment. As reported in Table 8, DAPPER outperforms the pathologist in the prediction of tissues at a tile-level.

	Adrenal Gland	Heart-Atrial Appendage	Esophagus-Mucosa	Brain-Cortex	Vagina	Skin-Not Sun Exposed	Brain-Cerebellum	Uterus	Muscle-Skeletal	Thyroid	Pituitary	Esophagus-Muscularis	Spleen	Pancreas	Testis	Prostate	Kidney-Cortex	Ovary	Liver	Heart-Left Ventricle
Adrenal Gland	55	5	0	1	0	4	0	2	0	0	12	0	1	1	6	3	2	1	3	4
Heart-Atrial Appendage	0	15	0	0	0	14	0	0	13	0	0	16	0	0	0	3	0	0	0	39
Esophagus-Mucosa	0	10	9	0	16	14	0	7	6	2	0	9	3	3	1	9	0	6	0	5
Brain-Cortex	0	0	0	96	0	0	1	0	0	0	0	0	0	0	0	0	1	2	0	0
Vagina	0	3	1	0	9	24	0	31	0	1	0	5	0	0	2	17	0	5	0	2
Skin-Not Sun Exposed	1	61	0	0	9	23	0	0	1	0	0	3	0	0	0	0	0	2	0	0
Brain-Cerebellum	0	0	0	33	0	0	65	0	0	0	0	0	1	0	1	0	0	0	0	0
Uterus	0	2	0	0	1	3	0	45	1	0	0	2	2	0	2	2	0	36	0	4
Muscle-Skeletal	0	2	0	0	0	0	0	2	92	0	0	0	0	0	0	0	0	0	0	4
Thyroid	0	2	0	0	1	7	0	0	0	86	0	0	0	0	1	1	1	1	1	0
Pituitary	12	1	0	0	2	3	0	2	0	3	63	1	6	0	2	3	0	2	0	0
Esophagus-Muscularis	0	6	0	0	0	4	0	3	28	3	0	22	3	2	1	4	0	2	0	22
Spleen	0	0	0	0	1	0	1	1	0	0	1	0	90	0	1	2	0	3	0	0
Pancreas	3	0	0	0	0	1	0	0	0	0	46	0	1	46	1	1	1	0	0	0
Testis	1	5	0	0	0	1	0	0	1	0	2	0	0	0	85	0	0	5	0	0
Prostate	0	5	0	0	3	2	0	12	1	0	4	4	1	0	6	56	0	2	0	4
Kidney-Cortex	3	0	0	0	0	0	0	0	0	3	6	0	0	0	3	1	79	3	2	0
Ovary	1	9	0	0	1	4	0	9	0	0	0	0	0	0	3	0	72	0	1	0
Liver	25	0	0	0	0	1	0	0	0	0	6	0	0	0	1	0	0	0	67	0
Heart-Left Ventricle	0	0	0	0	0	1	0	0	46	0	0	1	0	0	0	2	0	0	0	50

Figure 16: Confusion matrix for pathologist classification on a subset of HINT20 external validation set. Red shaded cells indicate the most confused classes.

Table 8: Tissue classification performance of DAPPER vs Pathologist. DAPPER with *ResNet+SVM* model outperforms the pathologist at tile-level. Metrics are computed on a subset of HINT20 external validation set (2,000 tiles).

Classifier	MCC	ACC (%)
Pathologist	0.542	56.3
DAPPER	0.786	79.6

To provide an unbiased estimation of the performance of DAPPER, we repeated the same evaluation on 10 other randomly generated subsets of 2,000 tiles extracted from the HINT20 external validation set. The obtained average *MCC* and *ACC* with 95% CI are 0.786 (0.783, 0.789), and 79.6 (79.3, 79.9), respectively.

Finally, since the classification at tile-level is an unusual task for a pathologist, who is instead trained on examining the whole context of a tissue scan, as a second task we asked the pathologist to classify 200 randomly chosen *WSIs* (10 for each class of HINT20). As expected, the results in this case are better than those at tile-level, *i.e.*, *MCC*=0.788, and *ACC*=79.5%, to be compared with the DAPPER performances reported in Table 7.

The HINT Benchmark Dataset

As a second contribution of this study, we are making available the HINT dataset, generated by the first component of tools in the DAPPER framework, as a benchmark dataset for validating machine learning models in digital pathology. The HINT dataset is currently composed of 53,727 tiles at 512×512 resolution, based on histology from *GTEx*. HINT can be easily expanded to over 78,000 tiles, as for this study we used a fraction of the *GTEx* images and at most 100 tiles from each *WSI* were extracted. Digital pathology still misses a universally adopted dataset to compare deep learning models as already established in vision (*e.g.*, ImageNet for image classification, COCO for image and instance segmentation). Several initiatives for a “BioImageNet” will eventually improve this scenario. Histology data are available in the generalist repository Image Data Resource (IDR) [343, 505]. Further, the In-

ternational Immuno-Oncology Biomarker Working Group in Breast Cancer and the MAQC Society have launched a collaborative project to develop data resources and quality control schemes on Machine Learning algorithms to assess [TILs](#) in Breast Cancer.

HINT is conceptually similar to KIMIA24. However, HINT inherits from [GTEx](#) more variability in terms of sample characteristics, validation of donors and additional access to molecular data. Further, we used a random sampling approach to process tiles excluding background and minimize human intervention in the choice and preparation of the images.

Deep features

We applied an unsupervised projection on all the features extracted by VGG and ResNet networks on all tissues tasks. In the following, we discuss an example for features extracted by VGG on the HINT20 task, displayed as [UMAP](#) projection (Figure 17), points are coloured for 20 tissue labels. The [UMAP](#) displays for the other tasks are available in Fig 20–Figure23.

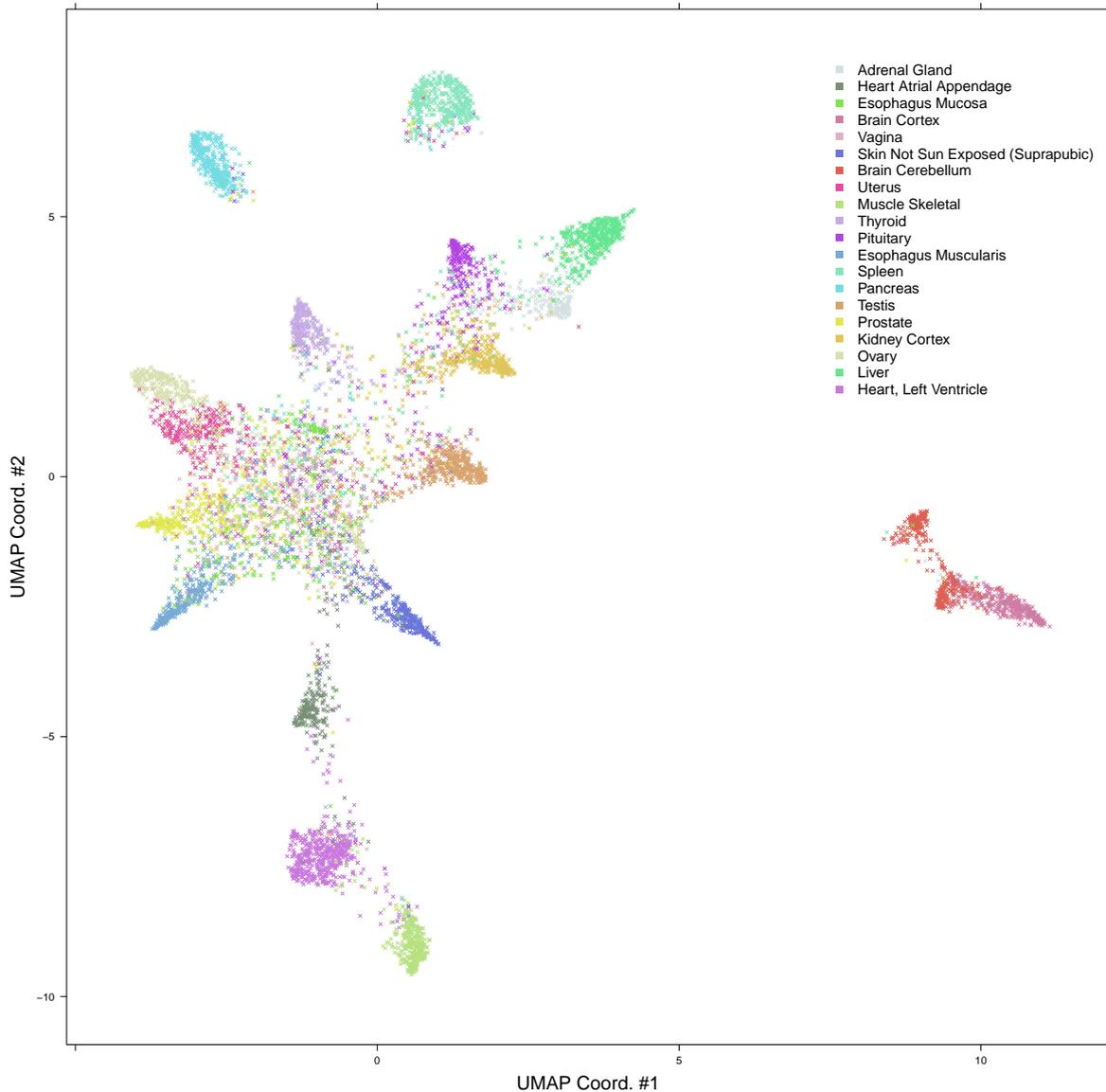


Figure 17: UMAP projection of external validation set for VGG-20 experiment.

The UMAP display is in agreement with the count distributions in the confusion matrix (Figure 14). The deep learning embedding separates well a set of histology types, including Muscle-Skeletal, Spleen, Pancreas, Brain-Cortex and Cerebellum, Heart-Left Ventricle and Atrial Appendage which group into distinct clusters (See Figure 17 and Table 9). The distributions of the activations for the top-3 deep features of the VGG backend network on the HINT₁₀ dataset are displayed in Figure 18; the top ranked deep feature (#668) is clearly selective for Spleen.

Table 9: Histology types well separated by SVM+ResNet model for HINT20. Accuracy is computed with respect to the confusion matrix in Figure 14 and expressed in percentage, together with the total number of samples for each class.

Histology type	ACC(%)	#samples
Spleen	94.6	446
Brain - Cortex	94.3	333
Muscle - Skeletal	93.4	347
Brain - Cerebellum	93.4	376
Heart - Left Ventricle	90.1	565
Pancreas	87.9	463
Heart - Atrial Appendage	84.7	317

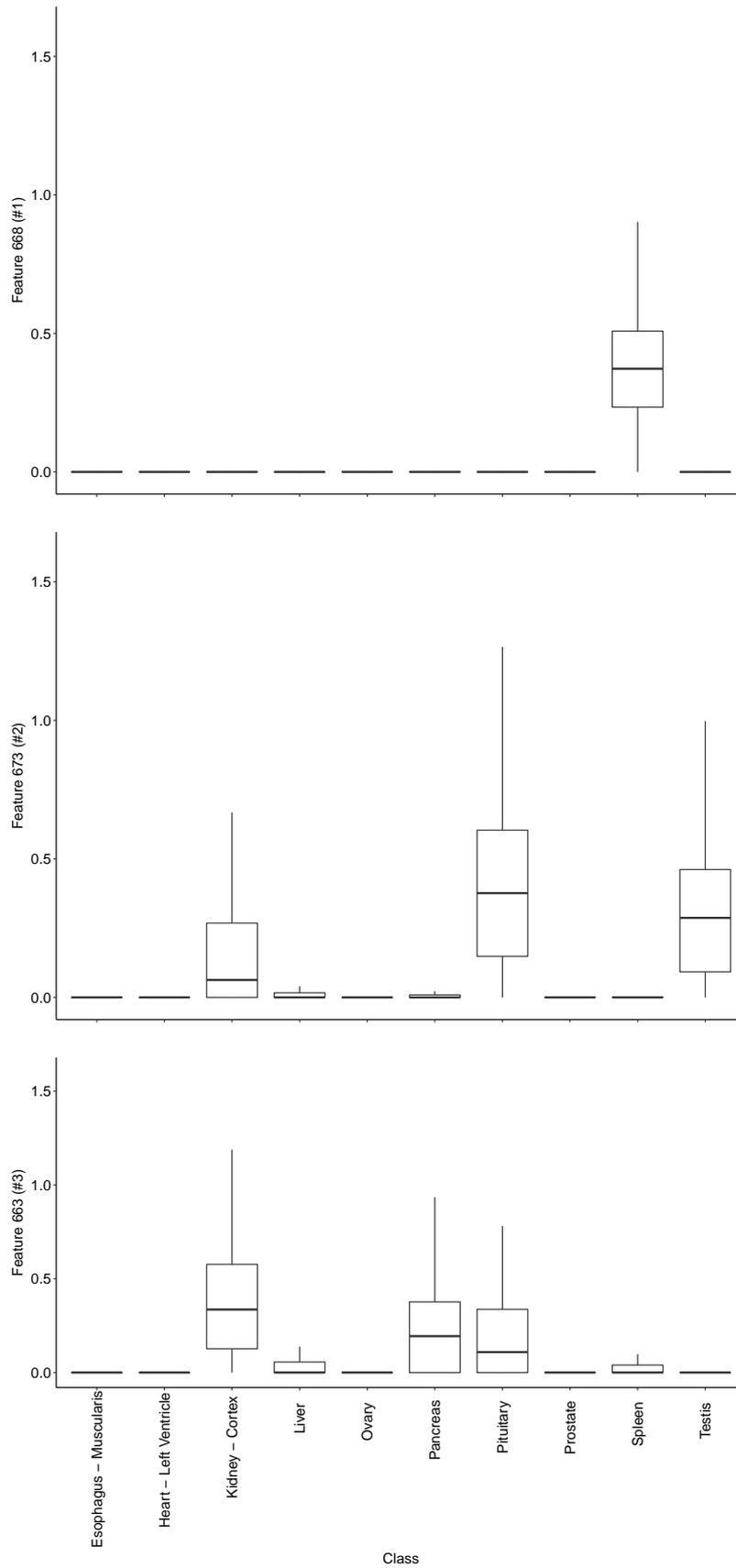


Figure 18: Deep features and tissue of origin. Distributions of the activations for the top-3 deep features computed with the VGG backend architecture for HINT10 dataset.

The UMAP projection also shows an overlapping for tissues such as Ovary and Uterus, or Vagina and Esophagus-Mucosa, or the two Esophagus histotypes, consistently with the confusion matrix (Figure 14).

Examples of five tiles from two well separated clusters, Muscle-Skeletal (ACC: 93.4%) and Spleen (ACC: 94.6%), are displayed in panel A of Figure 19. Tiles from three clusters partially overlapping in the neural embedding and mislabeled in both the VGG-20 and ResNet-20 embeddings with SVM (Esophagus-Mucosa ACC=53.2%, Esophagus-Muscularis ACC=72.1%, Vagina ACC=59.0%) are similarly visualized in Figure 19B. While the aim of this paper is to introduce a framework for honest comparison of models that will be used for clinical purposes rather than fine-tuning accuracy in this experiment, it is evident that these tiles have morphologies that are hard to classify. This challenge requires more complex models (*e.g.* ensembles) and a structured output labeling, already applied in dermatology [138].

Further, we are exploring the combination of DAPPER with image analysis packages, such as HistomicsTK ¹ or CellProfiler [74], to extract features useful for interpretation and feedback from pathologists.

¹ <https://bit.ly/3Hn9NcR>

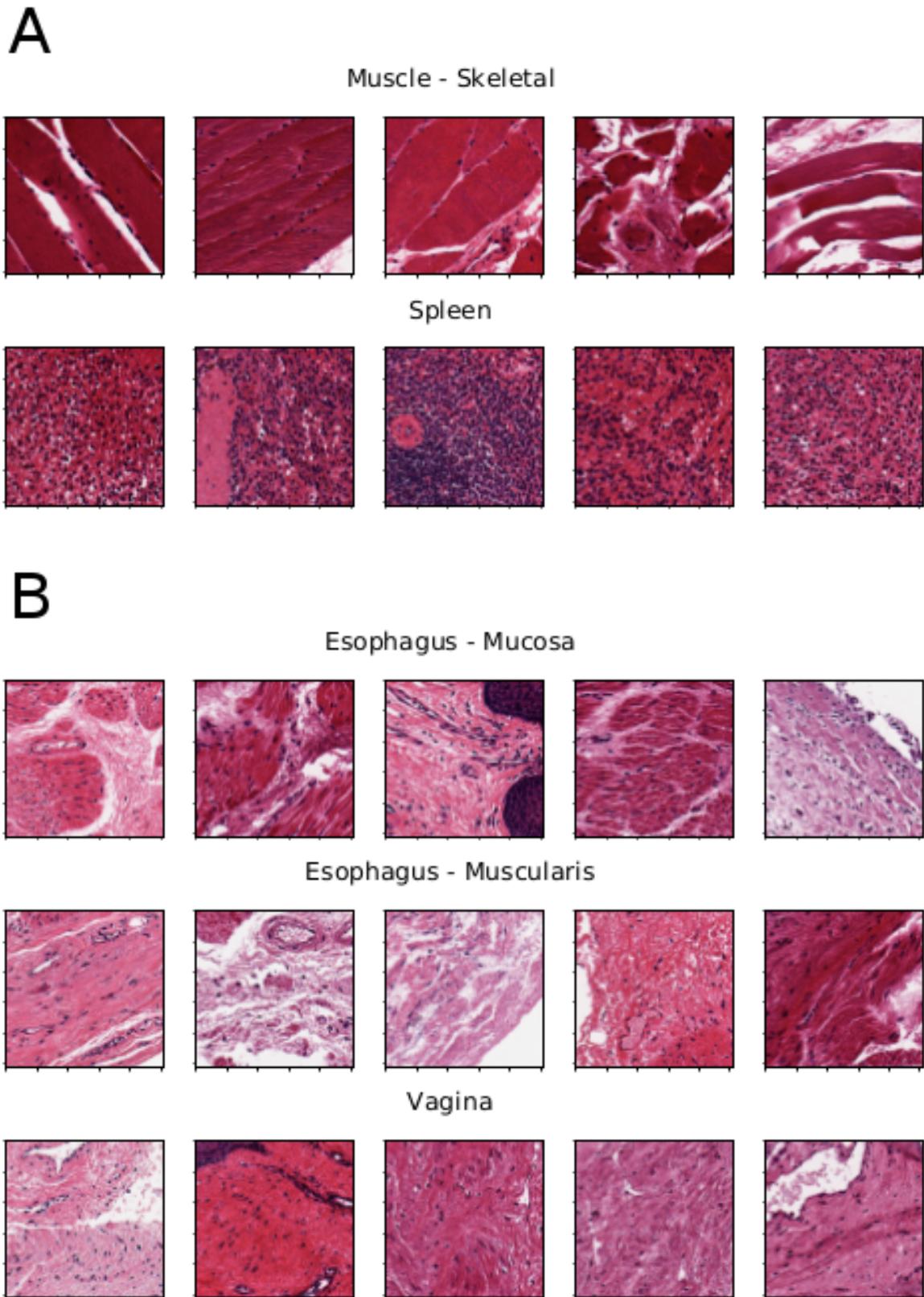


Figure 19: Representative tiles predicted from VGG-20 experiment. A) Examples from two well-separated clusters observed in the UMAP embedding. B) Samples of mislabeled tiles from tissues partially overlapping in the UMAP embedding.

4.4 Discussion

Digital pathology would greatly benefit from the adoption of machine learning, shifting human assessment of histology to higher quality, non-repetitive tasks. Unfortunately, there is no fast, easy route to improve reproducibility of automated analysis. The adoption of the [DAP](#) clearly sets in a computational aggravation not usually considered for image processing exercises. However, this is an established practice with massive omics data [\[419\]](#), and reproducibility by design can handle secondary results useful for diagnostics and for interpretation.

We designed the DAPPER framework as a tool for evaluating accuracy and stability of deep learning models, currently only backend elements in a sequence of processing steps, and possibly in the future end-to-end solutions. We choose as test domain [HE](#) stained [WSIs](#) for prediction of tissue of origin, which is not a primary task for trained pathologists, but a reasonable benchmark for machine learning methods. Also, we are aware that tissue classification is only a step in real digital pathology applications. [Mobadersany and colleagues \[325\]](#) used a deep learning classifier to score and visualize risk on the [WSIs](#). Similarly, deep learning tile classification may be applied to quantify histological differences in association to a genomic pattern, *e.g.*, a specific mutation or a high-dimensional protein expression signature. In this vision, the attention to model selection supported by our framework is a prerequisite for developing novel AI algorithms for digital pathology, *e.g.*, for analytics over [TILs](#).

Although we are building on deep learning architectures known for applications on generic images, they adapted well to [WSIs](#) in combination with established machine learning models; we expect that large scale bioimaging resources will give the chance of improving the characterization of deep features, as already emerged with the [HINT](#) dataset that we are providing as public resource. In this direction, we plan to release the network weights of the backend DAPPER models that are optimized for histopathology as alternative pretrained weights for digital pathology, similarly to those for the ImageNet dataset and available in [torchvision](#).

Implementation and code availability

All the code of the DAPPER framework is written in *Python* (v3.6) and *R* (v3.4.4). In addition to the general scientific libraries for Python, the scripts for the creation and training of the networks are based on *PyTorch*; the backend networks are implemented in *torchvision*. The library for processing histological images (available at gitlab.fbk.eu/mpba-histology/histolib) is based on *OpenSlide* and *scikit-image*. For the **UMAP** analysis we used the *R umap* module with the following parameters: `n_neighbors=40`, `min_dist=0.01`, `n_components=2`, and Euclidean metric. The computations were performed on Microsoft Azure Virtual Machines with 4 NVIDIA K80 GPUs, 24 Intel Xeon E5-2690 cores and 256 GB RAM.

Acknowledgments

Cloud computing was funded by the Azure Research grant “Deep Learning for Precision Medicine”, endowed to CF. We derived the HINT dataset from the Genotype-Tissue Expression (GTEx) Project, supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS (data downloaded from the GTEx Portal on 05/10/18).

The authors thank Intel Italy for technical support and availability of high performance computing resources. We also thank H. Tizhoosh for the availability of KIMIA Path24 dataset, L. Coviello for his help in the networks’ optimization and G. Franch for the realization of the striking image (cover figure of this Chapter).

4.5 Appendix

Table 15: Matthew Correlation Coefficient values for each experiment, and FCH classifier with <4 dense layers on HINT dataset. FCH₃: three dense layers with 1000, 256 and # tissue classes nodes, respectively; FCH₂: two dense layers with 256 and # tissue classes nodes, respectively. The average cross validation MCC with 95% CI (**H-MCCt**), and MCC on the external validation set (**H-MCCv**) are reported. The bold numbers correspond to bolds numbers in Table 4 of the main text.

Experiment	FCH ₃		FCH ₂	
	H-MCCt	H-MCCv	H-MCCt	H-MCCv
VGG-5	0.841 (0.838, 0.843)	0.823	0.832 (0.829, 0.834)	0.822
ResNet-5	0.881 (0.878, 0.883)	0.887	0.871 (0.869, 0.873)	0.877
VGG-10	0.895 (0.894, 0.896)	0.895	0.893 (0.892, 0.895)	0.892
ResNet-10	0.858 (0.856, 0.859)	0.860	0.853 (0.851, 0.854)	0.855
VGG-20	0.773 (0.772, 0.774)	0.775	0.768 (0.766, 0.769)	0.769
ResNet-20	0.757 (0.756, 0.759)	0.756	0.749 (0.748, 0.751)	0.746

Table 10: Summary of available samples, downloaded WSIs and extracted tiles.

Anatomical part	# WSIs	# Tiles
1. Adipose - Subcutaneous	18	174
2. Adipose - Visceral (Omentum)	21	59
3. Adrenal Gland	25	1574
4. Artery - Aorta	19	876
5. Artery - Coronary	23	915
6. Artery - Tibial	22	957
7. Bladder	-	-
8. Brain - Cerebellum	20	1825
9. Brain - Cortex	19	1731
10. Breast - Mammary Tissue	26	1291
11. Cells - EBV-transformed lymphocytes	-	-
12. Cells - Transformed fibroblasts	-	-
13. Cervix - Ectocervix	-	-
14. Cervix - Endocervix	-	-
15. Colon - Sigmoid	21	1401
16. Colon - Transverse	20	1301
17. Esophagus - Gastroesophageal Junction	22	1440
18. Esophagus - Mucosa	23	1700
19. Esophagus - Muscularis	25	1926
20. Fallopian Tube	-	-
21. Heart - Atrial Appendage	20	1622
22. Heart - Left Ventricle	28	2689
23. Kidney - Cortex	28	2424
24. Liver	26	2583
25. Lung	21	1009
26. Minor Salivary Gland	19	896
27. Muscle - Skeletal	23	1875
28. Nerve - Tibial	23	1286
29. Ovary	26	2452
30. Pancreas	26	2193
32. Pituitary	21	1890
32. Prostate	24	2323
33. Skin - Not Sun Exposed (Suprapubic)	25	1811
34. Skin - Sun Exposed (Lower leg)	22	1129
35. Small Intestine - Terminal Ileum	24	1184
36. Spleen	22	2184
37. Stomach	21	1477
38. Testis	26	2221
39. Thyroid	26	1890
40. Uterus	21	1857
41. Vagina	21	1746
42. Whole Blood	-	-

Table 11: Summary of the tissues composing the datasets.

Dataset	Anatomical zone [# Tiles samples]
HINT ₅	Kidney - Cortex [2424], Pancreas [2193], Colon - Transverse [1301], Breast - Mammary Tissue [1291], Lung [1009].
HINT ₁₀	Heart - Left Ventricle [2689], Liver [2583], Ovary [2452], Kidney - Cortex [2424], Prostate [2323], Testis [2221], Pancreas [2193], Spleen [2184], Esophagus - Muscularis [1926], Pituitary [1890].
HINT ₂₀	Heart - Left Ventricle [2689], Liver [2583], Ovary [2452], Kidney - Cortex [2424], Prostate [2323], Testis [2221], Pancreas [2193], Spleen [2184], Esophagus - Muscularis [1926], Pituitary [1890], Thyroid [1890], Muscle - Skeletal [1875], Uterus [1857], Brain - Cerebellum [1825], Skin - Not Sun Exposed (Suprapubic) [1811], Vagina [1746], Brain - Cortex [1731], Esophagus - Mucosa [1700], Heart - Atrial Appendage [1622], Adrenal Gland [1574].
HINT ₃₀	Heart - Left Ventricle [2689], Liver [2583], Ovary [2452], Kidney - Cortex [2424], Prostate [2323], Testis [2221], Pancreas [2193], Spleen [2184], Esophagus - Muscularis [1926], Pituitary [1890], Thyroid [1890], Muscle - Skeletal [1875], Uterus [1857], Brain - Cerebellum [1825], Skin - Not Sun Exposed (Suprapubic) [1811], Vagina [1746], Brain - Cortex [1731], Esophagus - Mucosa [1700], Heart - Atrial Appendage [1622], Adrenal Gland [1574], Stomach [1477], Esophagus - Gastroesophageal Junction [1440], Colon - Sigmoid [1401], Colon - Transverse [1301], Breast - Mammary Tissue [1291], Nerve - Tibial [1286], Small Intestine- Terminal Ileum [1184], Skin - Sun Exposed (Lower leg) [1129], Lung [1009], Artery - Tibial [957].

Table 12: Summary of the FCH performances when the layers of the feature extraction block are trained. When training also the feature extraction block (VGG backbone network, not in DAP) we observe an improvement of the accuracy from 5.5% to 24.8% for the four chosen experiments. A possible interpretation of this phenomenon is that the neural network benefits from adjusting also the initial weights because the layers have to learn characteristics on the images much more different than the ImageNet dataset.

Experiment	Trained		Freezed	
	ACC %	MCC	ACC %	MCC
VGG-5	97.6	0.970	92.1	0.901
VGG-10	97.5	0.972	88.6	0.874
VGG-20	93.9	0.936	76.9	0.759
VGG-30	77.1	0.765	61.8	0.607

Table 13: Comparison of the three optimization methods to set the learning rate. The best method for setting the learning rate was assessed using the VGG as backbone network (not in DAP) on the classification of 5 tissues. Three methods were tested: Fixed (FIX): the learning rate is set to 10^{-5} for the whole training; Step-wise (STEP): the learning rate is initialized at 10^{-3} and updated every 10 epochs with the following rule: $\lambda_{new} = \lambda_{old}/10$; Polynomial (POLY): the learning rate is initialized at 10^{-3} and updated every 10 iterations with a polynomial law: $\lambda_{new} = 10^{-3}(1 - \frac{i}{I_{max}})^{0.9}$, where i is the index of the iteration and I_{max} is the total number of iterations.

Opt. method	FCH	
	ACC	MCC
FIX	93.7	0.921
STEP	88.4	0.854
POLY	91.1	0.888

Table 14: Summary of the performances using the VGG as backbone network (not in DAP) on datasets with increasing number of tissues. We observe that the performance decreases when the number of tissues increases. This result can be explained with the increased difficulty of the task and is also possibly complicated by the introduction of tissues with similar histological patterns.

Experiment	FCH	
	ACC %	MCC
VGG-5	93.7	0.921
VGG-10	88.6	0.874
VGG-20	76.9	0.760
VGG-30	61.8	0.607

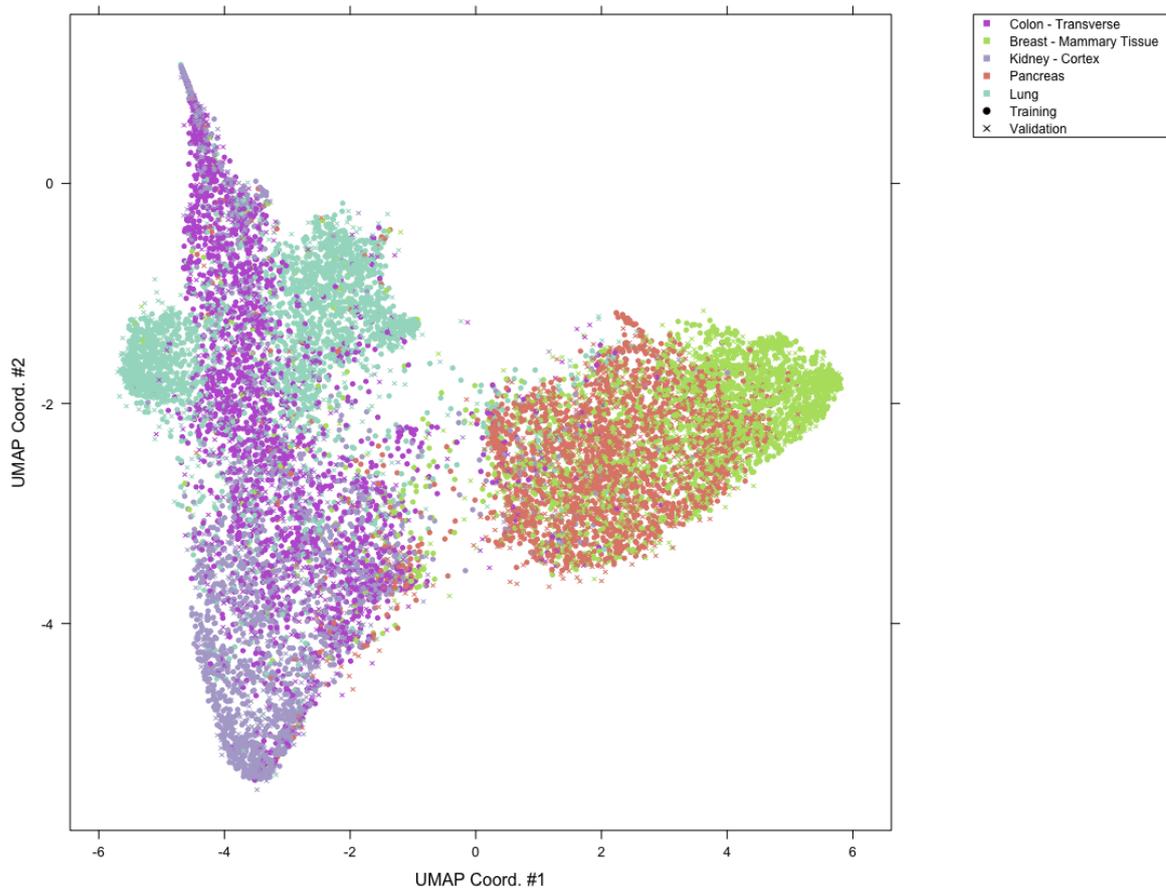


Figure 20: UMAP projection on training and (external) validation set for VGG-5 experiment.

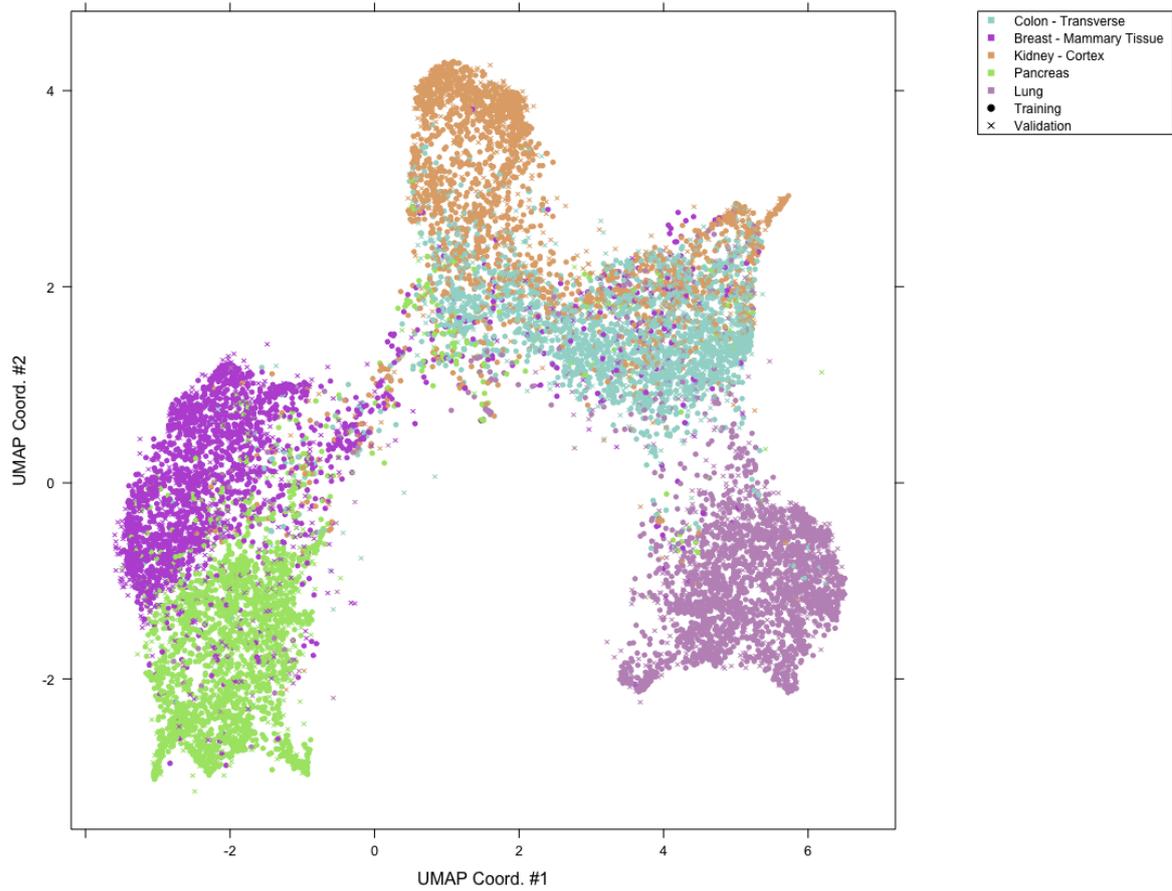


Figure 21: UMAP projection on training and (external) validation set for ResNet-5 experiment.

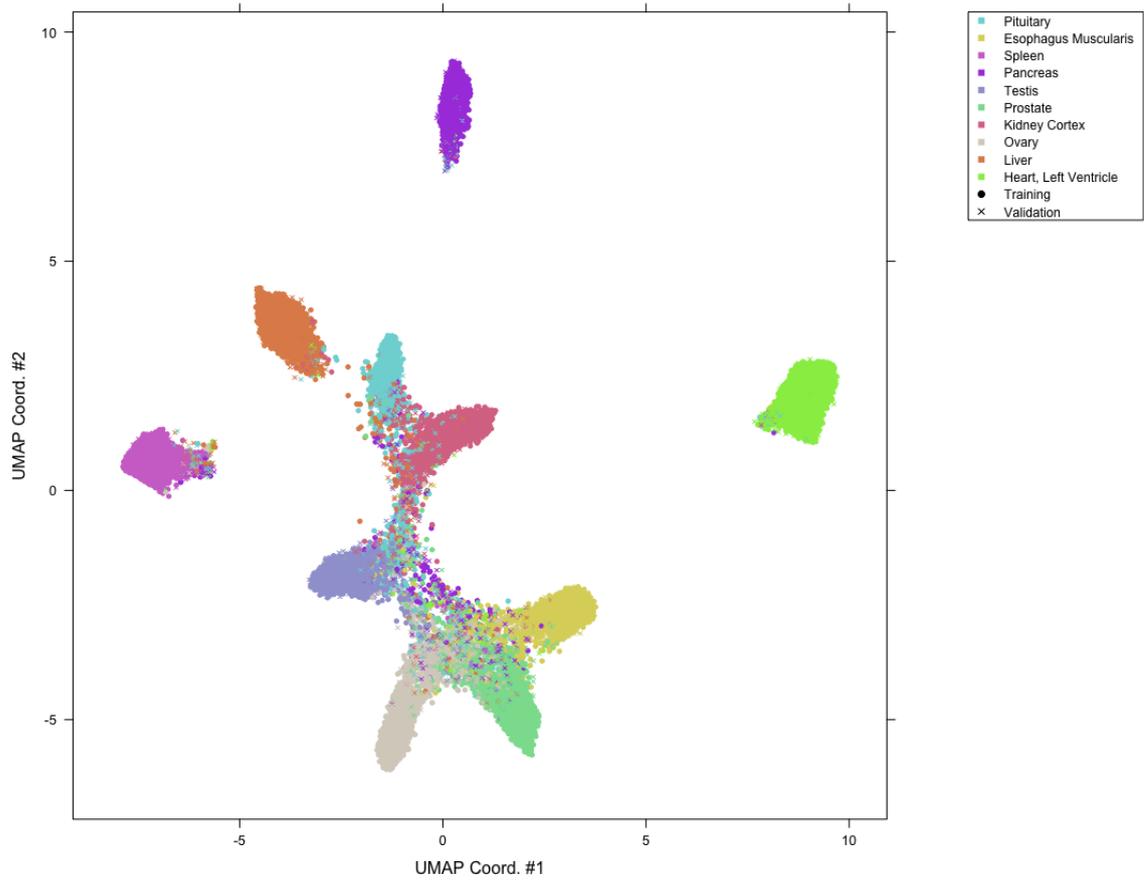


Figure 22: UMAP projection on training and (external) validation for VGG-10 experiment.

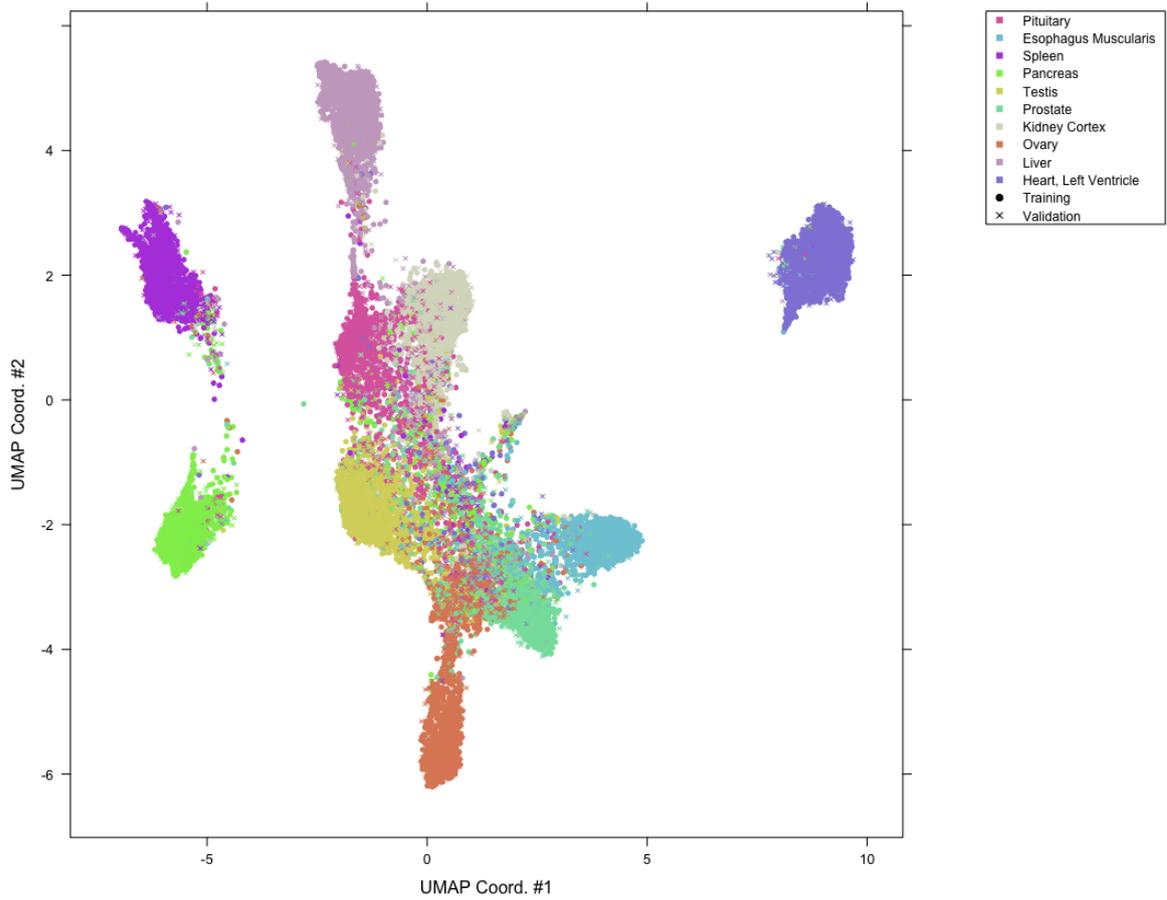
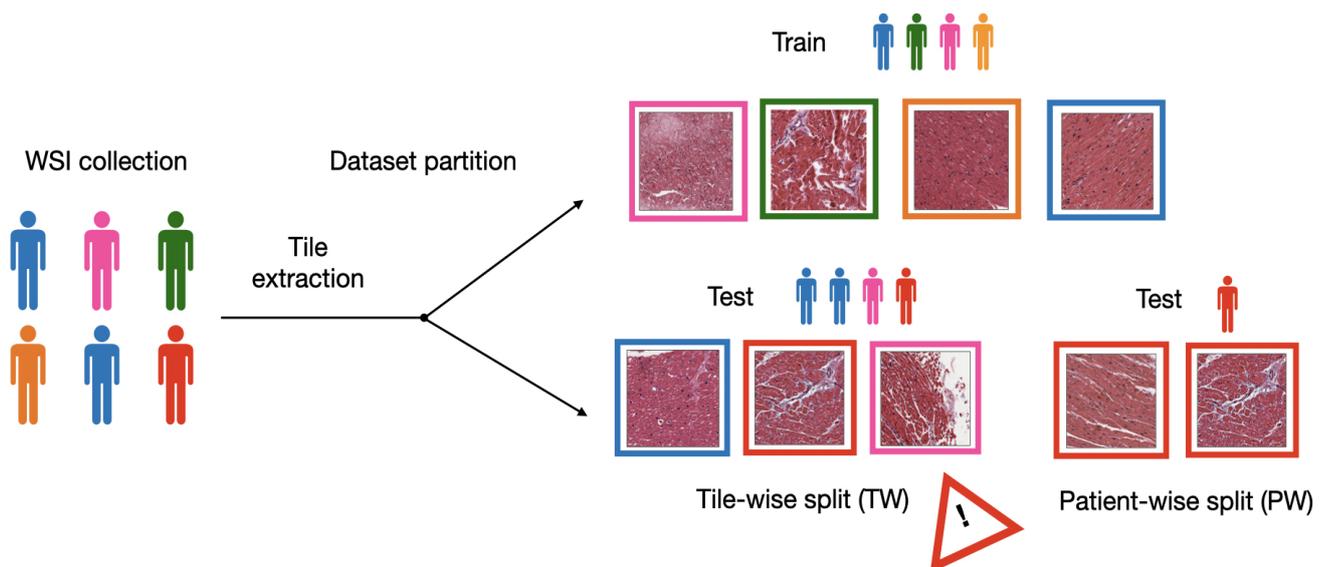


Figure 23: UMAP projection on training and (external) validation for ResNet-10 experiment.

DATA LEAKAGE IN DIGITAL PATHOLOGY



Authors: N. Bussola, A. Marcolini, V. Maggio, G. Jurman, C. Furlanello. *Original title:* AI slipping on tiles: data leakage in digital pathology. *Published in:* ICPR 2021 International Conference on Pattern Recognition. (Feb. 2021)

DATA LEAKAGE IN DIGITAL PATHOLOGY

Highlights

- Protocols for data partitioning in AI pipelines for histopathology tile collections are particularly susceptible to the risk of selection biases among bioimaging applications.
- Results on 4 classification tasks from 3 public repositories indicate that predictive score can be inflated up to 41% when tiles from the same subject are used both in training and validation sets.
- The DAPPER framework is adopted to first quantify the impact of data leakage on reproducibility and reliability of ML classifiers in Digital Pathology.
- We design an original protocol for reproducible preprocessing and leakage-free data partitioning for AI pipelines on massive histopathology collections.

Personal contribution I co-designed the overall pipeline and run the experiments on the considered datasets. I implemented the protocols for data preprocessing on GTEx and TCGA data. I also drafted the article and prepared the figures.

5.1 Abstract

Reproducibility of AI models on biomedical data still stays as a major concern for their acceptance into clinical practice. Initiatives for reproducibility in the development of predictive biomarkers as the MAQC Consortium already underlined the importance of appropriate Data Analysis Plans (DAPs) to control for different types of bias, including data leakage from the training to the test set. In the context of digital pathology, the leakage typically lurks in weakly designed experiments not accounting for the subjects in their data partitioning schemes. This issue is then exacerbated

when fractions or subregions of slides (i.e. *tiles*) are considered. Despite this aspect is largely recognized by the community, we argue that it is often overlooked. In this study, we assess the impact of data leakage on the performance of machine learning models trained and validated on multiple histology data collection. We prove that, even with a properly designed DAP (10×5 repeated cross-validation), predictive scores can be inflated up to 41% when tiles from the same subject are used both in training and validation sets by deep learning models. We replicate the experiments for 4 classification tasks on 3 histopathological datasets, for a total of 374 subjects, 556 slides and more than 27,000 tiles. Also, we discuss the effects of data leakage on transfer learning strategies with models pre-trained on general-purpose datasets or off-task digital pathology collections. Finally, we propose a solution that automates the creation of leakage-free deep learning pipelines for digital pathology based on *histolab*, a novel Python package for histology data preprocessing. We validate the solution on two public datasets (TCGA and GTEx).

Summary

The goal of this study is to provide evidence that reproducibility issues are still lurking in the grey areas of preprocessing, ready to emerge in the everyday practice of machine learning for digital pathology. The BreakHis [545] dataset, one of the most popular histology collections of breast cancer samples, has been used in more than 40 scientific papers to date [546], with reported results spanning a broad range of performance. In a non-negligible number of these studies, overfitting effects due to data leakage are suspected to impact their outcomes.

DL pipelines for histopathological data typically require WSIs to be partitioned into multiple patches tiles to augment the original training data, and to comply with memory constraints imposed by GPU hardware architectures [103]. For example, a single WSI of size $67,727 \times 47,543$ pixels can be partitioned in multiple 512×512 tiles, which are randomly extracted, and verified such that selected subregions preserve enough tissue information. These tiles are then processed by data augmentation op-

erators (e.g. random rotation, flipping, or affine transformation) to reduce the risk of overfitting. As a result, the number of multiple subimages originating from the very same histological specimen is significantly amplified [542, 557], consequently increasing the risk for data leakage. Protocols for data partitioning are not naturally immune against replicates, and so the source originating each tile should be considered to avoid any risk of bias [299].

In this work, we quantify the importance of adopting *Patient-Wise* split procedures with a set of experiments on digital pathology datasets. All experiments are based on DAPPER (see Chapter 4), a reproducible framework for predictive digital pathology composed of a deep learning core ("backbone network") as feature encoder, and multiple task-related classification models, i.e. Random Forest or Multi-Layer Perceptron Network (see Figure 24). We test the impact of various data partitioning strategies on the training of multiple backbone architectures, i.e. DenseNet [547], and ResNet models [191], fine-tuned to the histology domain.

Our experiments confirm that train-test contamination (in terms of modeling) is a serious concern that hinders the development of a dataset-agnostic methodology, with impact similar to the lack of standard protocols in the acquisition and storage of WSIs in digital pathology [28]. Thus, we present a protocol to prevent data leakage during data preprocessing. The solution is based on *histolab*, an open-source Python library designed as a reproducible and robust environment for WSI preprocessing (see Chapter 6). The novel approach is demonstrated on two public large scale datasets: *GTEX* [462] (i.e. non-pathological tissues), and *TCGA* [467] (i.e. cancer tissues).

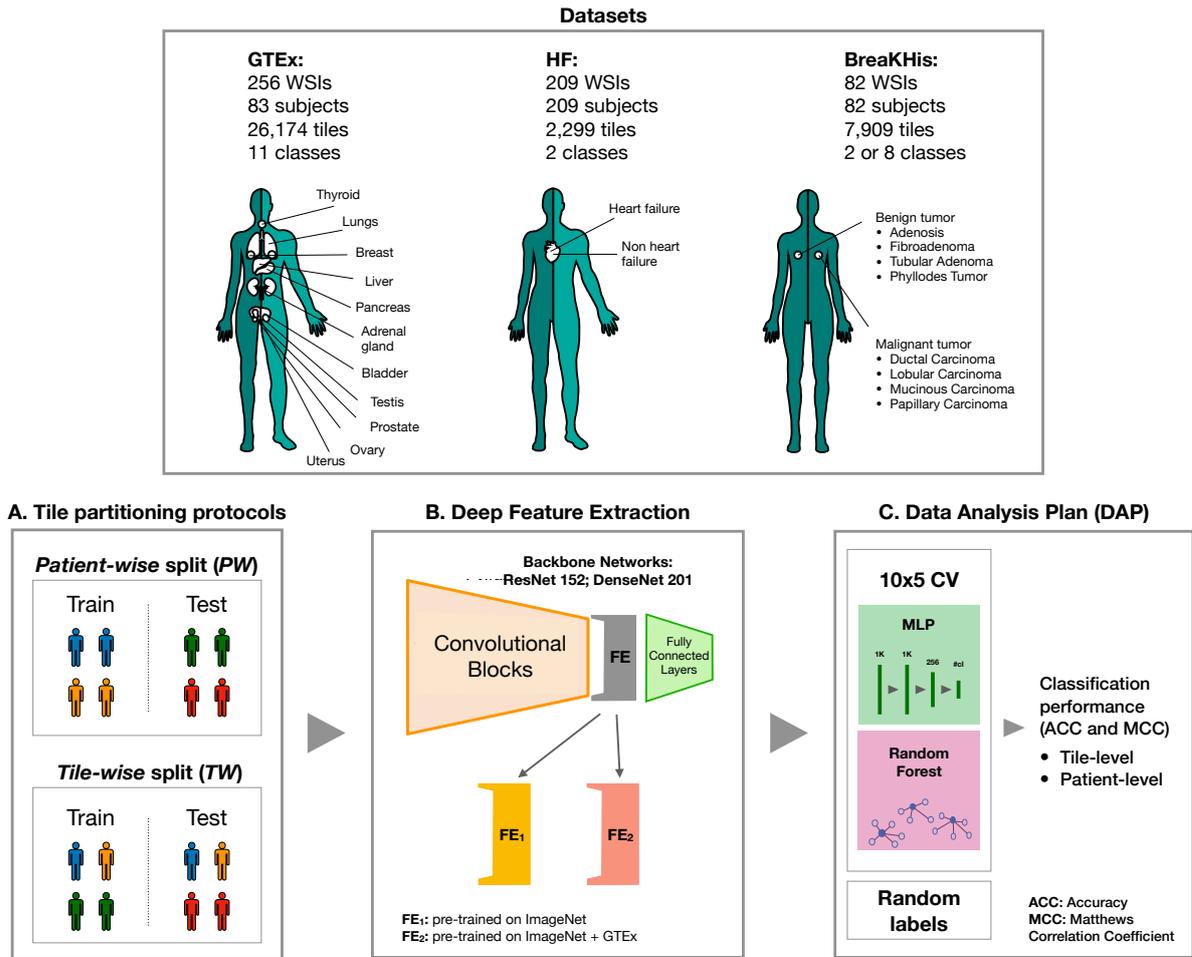


Figure 24: Experimental environment for evaluation of data leakage impact on machine learning models in digital pathology. (A) Tile datasets are split into train/test set following either the *Tile-Wise* or the *Patient-Wise* protocol; (B) the train set is fed to a backbone network for feature extraction, using different transfer learning strategies; (C) machine learning classifiers on the deep features are evaluated within the Data Analysis Plan.

5.2 Material and methods

We tested our experimental pipeline on three public datasets for image classification in digital pathology, namely [GTEx](#) [462], Heart Failure (HF) [343], and [BreaKHis](#) [545]. Descriptive statistics of the datasets are reported in [Table 16](#), and [Figure 24](#).

Table 16: Statistics of the datasets considered in this study.

Dataset	Subjects	WSIs	WSIs per Subject			Tiles	Tiles per Subject		
			Min	Max	Median		Min	Max	Median
GTE _x	83	265	1	7	3	26,174	1	700	300
HF	209	209	1			2,299	11		
BreaKHis	82	82	1			2,013	9	62	21

GTEX DATA In this work, we consider a subset of 265 **WSIs** randomly selected from 11 histological classes of the **GTE_x** repository, for a total of 83 subjects. From this subset, we randomly selected a balanced number of **WSIs** per tissue: adrenal gland ($n = 24$); bladder ($n = 19$); breast ($n = 26$); liver ($n = 26$); lung ($n = 21$); ovary ($n = 26$); pancreas ($n = 26$); prostate ($n = 24$); testis ($n = 26$); thyroid ($n = 26$); uterus ($n = 21$).

We implemented a data preprocessing pipeline to prepare the tile dataset from the **WSIs** collection (see Chapter 4, Section 4.2). A maximum of 100 tiles of size 512×512 is then randomly extracted from each slide. To ensure that only high-informative images are used, tiles with tissue area that accounts for less than 85% of the whole patch are automatically rejected. At the end of this step, a total of 26,174 random tiles is extracted from the **WSIs**, each available at different magnification levels (i.e., $20\times$, $10\times$, $5\times$). In this paper we limit experiments and discussions to tiles at $5\times$ magnification, with no loss of generality.

THE HF DATASET The Heart Failure collection [343] originates from 209 **HE**-stained **WSIs** of the left ventricular tissue, each corresponding to a single subject. The learning task is to distinguish images of *heart failure* ($n = 94$) from those of *non-heart failure* ($n = 115$). Slides in the former class are categorized according to the disease subtype: ischemic cardiomyopathy ($n = 51$); idiopathic dilated cardiomyopathy ($n = 41$); undocumented ($n = 2$). Subjects with no heart failure are further grouped in: normal cardiovascular function ($n = 41$); non-HF and no other pathology ($n = 72$); non-HF and other tissue pathology ($n = 2$). **WSIs** in this dataset have been acquired with an Aperio ScanScope at $20\times$ native magnification, and then downsampled at $5\times$ magnification by authors. From each **WSI**, 11 non-overlapping patches of size

250×250 were randomly extracted. The entire collection of 2,299 tiles is publicly available on the Image Data Resource Repository ¹ (IDR number: idr0042).

THE BREAKHIS DATASET The BreakHis histopathological dataset [545] collects 7,909 HE-stained tiles (size 700×460) of malignant or benign breast tumour biopsies. Tiles correspond to regions of interest manually selected by expert pathologists from a cohort of 82 patients, and made available at different magnification factors, i.e., $40\times$, $100\times$, $200\times$, $400\times$ [545]. To allow for a more extensive comparison with the state of the art, only the $200\times$ magnification factor is considered in this paper. The BreakHis dataset currently contains 4 histological distinct subtypes of benign, and malignant tumours, respectively: Adenosis ($n = 444$); Fibroadenoma ($n = 1,014$); Tubular Adenoma ($n = 453$); Phyllodes Tumor ($n = 569$); Ductal Carcinoma ($n = 3,451$); Lobular Carcinoma ($n = 626$); Mucinous Carcinoma ($n = 792$); Papillary Carcinoma ($n = 560$). This dataset is used for two classification tasks: (BreakHis-2) binary classification of benign and malignant tumour samples; (BreakHis-8) classification of the 8 distinct tumour subtypes.

The pipeline used in this work is based on the DAPPER framework for digital pathology (see Chapter 4), extended by (i) integrating specialised train-test splitting protocols, i.e. *Tile-Wise* and *Patient-Wise*; (ii) extending the feature extractor component with new backbone networks; (iii) applying two transfer learning strategies for feature embedding. Figure 24 shows the three main blocks of the experimental environment defined in this paper: (A) dataset partition in train and test set; (B) feature extraction procedure with different transfer learning strategies; (C) the DAP employed for machine learning models.

A. Dataset partitioning protocols

The tile dataset is partitioned in the *training* set and *test* set, considering 80% and 20% split ratio for the two sets, respectively. We compare two data partitioning protocols to investigate the impact of a train-test contamination (Figure 24A): in the

¹ idr.openmicroscopy.org/

Tile-Wise (TW) protocol, tiles are randomly split between the training and the test sets, regardless of the original [WSI](#). The *Patient-Wise* (PW) protocol splits the tile dataset strictly ensuring that all tiles extracted from the same subject are found either in the training or the test set. To avoid other sources of leakage due to class imbalance [381], the two protocols are both combined with stratification of samples over the corresponding classes, and any class imbalance is accounted for by weighting the error on generated predictions.

B. Deep Learning models and feature extraction

The training set is then used to train a deep neural network for feature extraction (Figure 24B), *i.e.* a “backbone” network whose aim is to learn a vector representation of the data (*features embedding*). In this study, we consider two backbone architectures in the residual network (ResNet) family, namely ResNet-152 [191] and DenseNet-201 [547]. Given that the DenseNet model has almost the double of parameters², and so a higher footprint in computational resources, diagnostic experiments and transfer learning are performed only with the ResNet-152 model. Similarly to [557], and [540], we started from off-the-shelf versions of the models, pre-trained on ImageNet, and then fine-tuned to the digital pathology domain using transfer learning. Specifically, we trained the whole network for 50 epochs with a learning rate $\eta = 1e - 5$, and Adam optimizer [543], in combination with the categorical cross-entropy loss. The β_1 and β_2 parameters of the optimizer are respectively set to 0.9 and 0.999, with no regularization. To reduce the risk of overfitting, we use train-time data augmentation, namely random rotation and random flipping of the input tiles.

The impact of adopting a single or double-step transfer learning strategy in combination with the *Patient-Wise* partitioning protocol is also investigated in this study. Two sets of features embeddings (FE) are generated: FE₁, backbone model fine-tuned from ImageNet; FE₂, backbone model sequentially fine-tuned from ImageNet and [GTEx](#).

² DenseNet-201: ~ 12M parameters; ResNet-152: ~ 6M parameters.

C. Classification and Data Analysis Plan

The classification is finally performed on the feature embedding within a **DAP** (see Chapter 2, Section 2.6.1.1) for machine learning models (Figure 24C). In this work, we compare the performance of two models: **RF** and Multi Layer Perceptron (**MLP**). Results have been reported both at tile-level and at patient-level, in order to assess the ability of machine learning models to generalise on unseen subjects (see section 5.3).

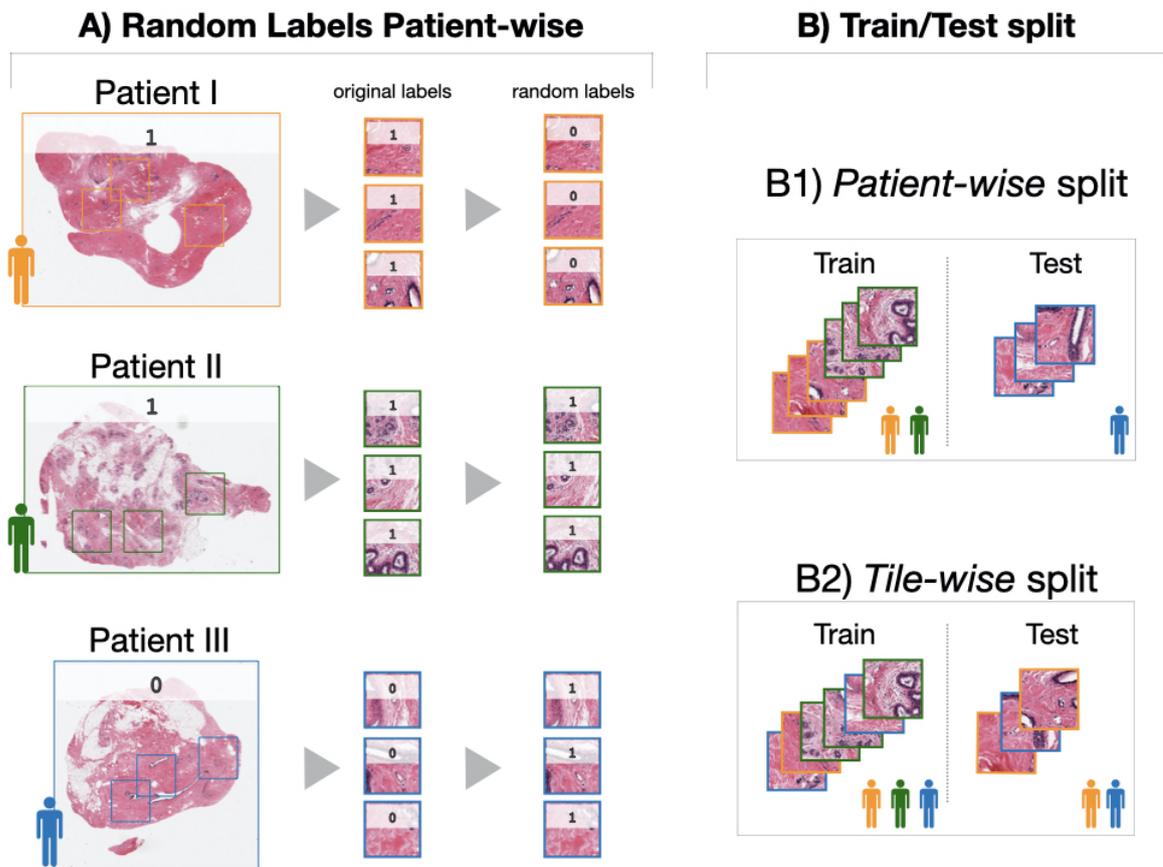


Figure 25: Random Labels experimental settings. A) The labels of the extracted tiles are randomly shuffled consistently with the original patient. B) The train/test split is then performed either *Patient-Wise* or *Tile-Wise*.

We adapted the *random labels* schema (RLab) of **DAP** to assess the impact of the tile partitioning strategies. In particular, we consistently randomize the labels for all the tiles of a single subject, thus they would all share the same random label (Figure 25A); then we alternatively use the *Patient-Wise* (Figure 25B1) or the *Tile-Wise* (Figure 25B2)

splits within the [DAP](#) environment. We focus on the RLab validation to emphasise evidence of data leakage derived from the *TW* and the *PW* protocols.

Performance metrics

Several patient-wise performance metrics have been defined in the literature [343, 545, 554]. Two metrics are considered in this study: (1) *Winner-takes-all* (WA), and (2) *Patient Score* (PS).

In the WA metric, the label associated with each patient corresponds to the majority of the labels predicted for their tiles. With this strategy, standard metrics based on the classification confusion matrix can be used as overall performance indicators. In this paper, [ACC](#) is used for comparability with the PS metric. The PS metric is defined for each patient [545] as the ratio of the N_c correctly classified tiles over the N_p total number of tiles per patient, namely $PS = \frac{N_c}{N_p}$. The overall performance is then calculated using the *global recognition rate* (RR), defined as the average of all the PS scores for all patients:

$$RR = \frac{\sum PS}{|P|} \quad (3)$$

In this paper, the WA metric and the PS metric are used for comparison of patient-level results on the HF dataset and the BreakHis dataset, respectively.

Preventing Data Leakage: the histolab library

As a solution to the data leakage pitfall, we have developed a protocol for image and tile splitting based on *histolab* (see Chapter 6). In order to intercept data leakage conditions, the protocol is designed to create a data-leakage free collection (tile extraction with the *Patient-Wise* split) that can be easily integrated in a deep learning workflow (Figure 26). The protocol is already customized for standardizing [WSI](#) preprocessing

on [GTEx](#) and [TCGA](#), two large scale public repositories that are widely used in computational pathology. The code can be also adapted to rebuild the training and test datasets from [GTEx](#) used in this study, thus extending the HINT collection presented in Chapter 4.

5.3 Results

Data Leakage effects on classification outcome

The results of the four classification tasks using the ResNet-152 pre-trained on ImageNet as backbone model (i.e. feature vectors FE_1) are reported in Figure 27, Table 20 and Table 21, with the *Tile-Wise* and the *Patient-Wise* partitioning protocols, respectively. The average cross validation MCC_V and ACC_V with 95% CI are presented,

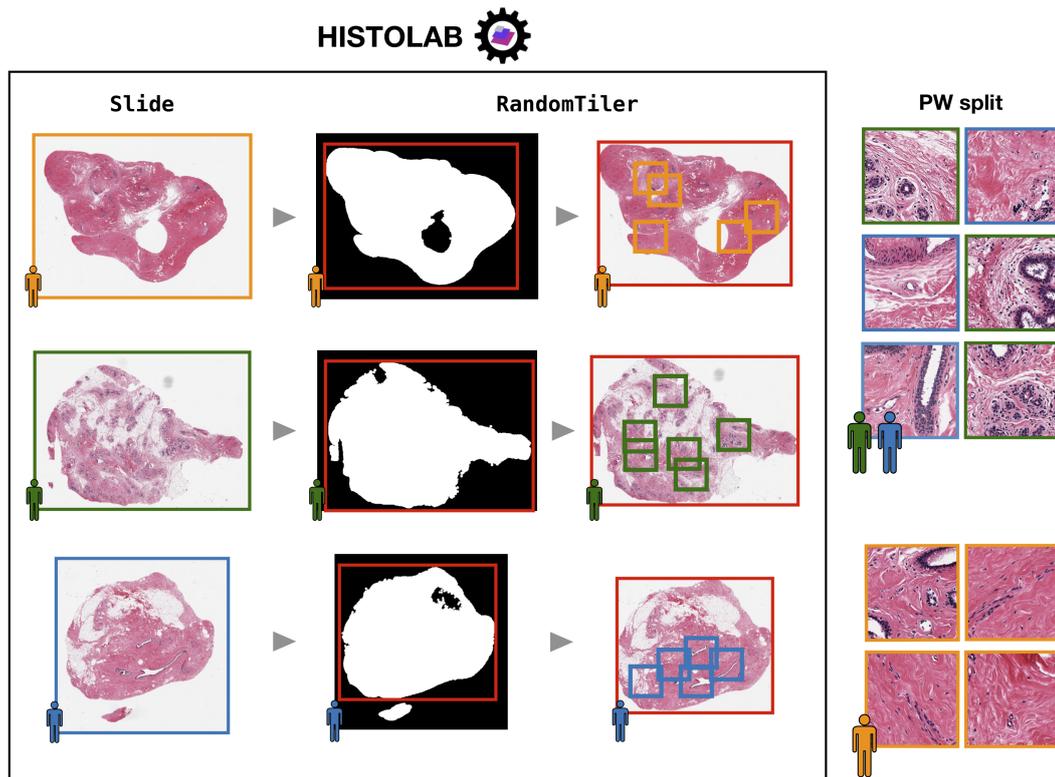


Figure 26: Workflow of the proposed protocol against data leakage in digital pathology, using the *histolab* software.

along with results on the test set (i.e. MCC_t , and ACC_t). State of the art results (i.e. *Others*) are also reported for comparison, whenever available.

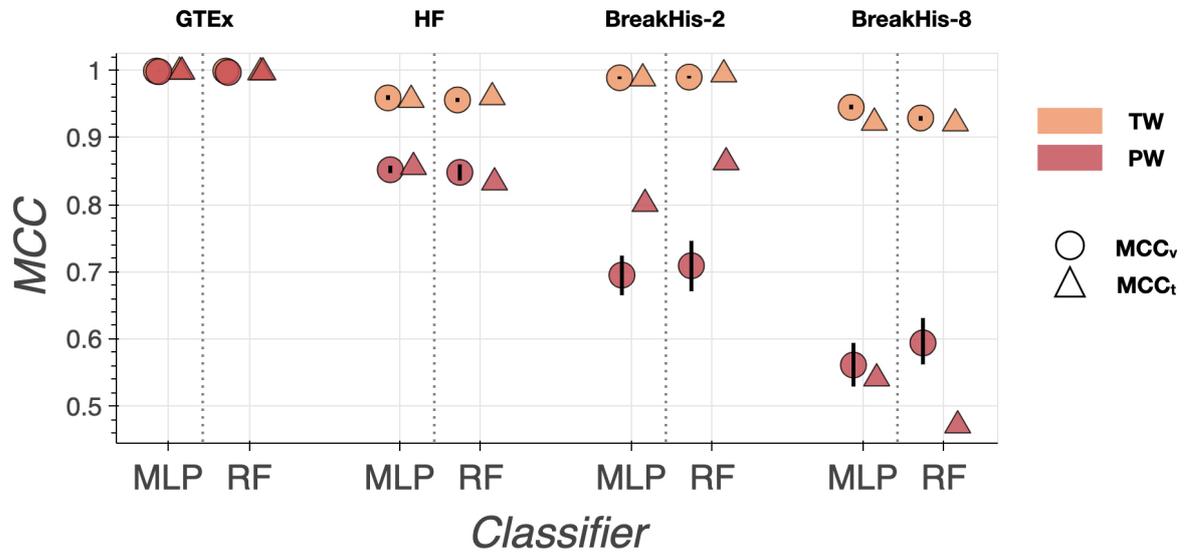


Figure 27: DAP results for each classifier head, using the *Tile-Wise* and the *Patient-Wise* partitioning protocol, and the FE_1 feature embedding with the ResNet-152 as backbone model. The average cross validation MCC_v with 95% CI are reported for each classification task, along with MCC_t on the test set.³

As expected, estimates are more favourable for the *TW* protocol with respect to the *PW* one, both in validation and in test and consistently for all the datasets. Moreover, the inflation of the *Tile-Wise* estimates is amplified in the multi-class setting (see BreakHis-2 vs BreakHis-8). Notably, these results are comparable with those in the literature, suggesting the evidence of a data leakage for studies adopting the *Tile-Wise* splitting strategy (Table 20 and Table 21). Results on the *GTEx* dataset do not suggest significant differences using the two protocols; however both *MCC* and *ACC* metrics lie in a very high range. Analogous results (not reported here) were obtained using the DenseNet-201 backbone model, further confirming the generality of the derived conclusions.

³ Original image, not available in the published manuscript.

Table 17: Random Labels (RLab) results using the ResNet-152 as backbone model, and *Tile-Wise* and *Patient-Wise* train-test split protocols. The average MCC_{RL} and ACC_{RL} with 95% CI are reported.

Dataset	MCC_{RL}		ACC_{RL}	
	<i>TW</i>	<i>PW</i>	<i>TW</i>	<i>PW</i>
HF	0.107 (0.078, 0.143)	0.004 (-0.042, 0.048)	0.553 (0.534, 0.570)	0.502 (0.474, 0.530)
BreaKHis-2	0.354 (0.319, 0.392)	-0.065 (-0.131, 0.001)	0.637 (0.613, 0.662)	0.560 (0.506, 0.626)
BreaKHis-8	0.234 (0.173, 0.341)	0.013 (-0.042, 0.065)	0.318 (0.215, 0.506)	0.097 (0.056, 0.143)

Random Labels detects signal in the Tile-Wise split

A data leakage effect is signalled for the *Tile-Wise* partitioning with a MCC consistently positive in the RLab validation schema (Sect. 5.2). For instance, as for BreaKHis-2 coupled with MLP , $MCC_{RL} = 0.354$ (0.319, 0.392) in the *Tile-Wise* setting, to be compared with $MCC_{RL} = -0.065$ (-0.131, 0.001) using the *Patient-Wise* protocol. Full MCC_{RL} results considering 5 trials of the RLab test are reported in Table 17, with corresponding ACC_{RL} values also included for completeness. Notably, all the tests using the *Patient-Wise* split perform as expected, i.e. with median values near 0, whereas results of the *Tile-Wise* case exhibit a high variability, especially for the BreaKHis-2 dataset (Figure 28).

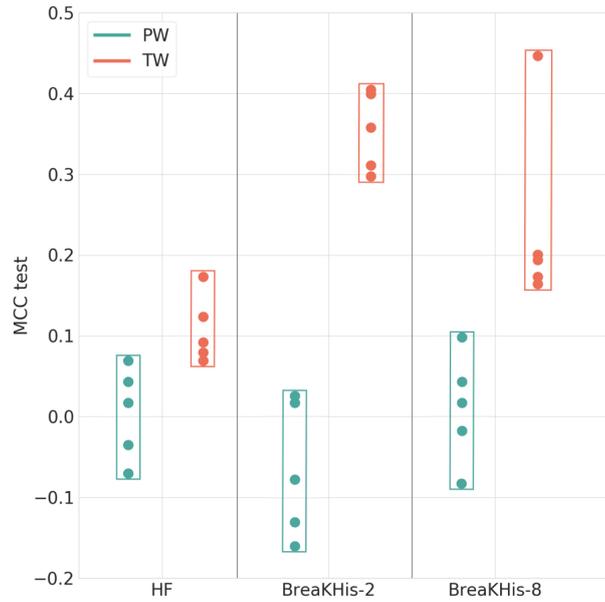


Figure 28: MCC_{RL} results on the test set using the *Tile-wise* (TW) and the *Patient-wise* (PW) protocols.

Benefits of domain-specific transfer learning

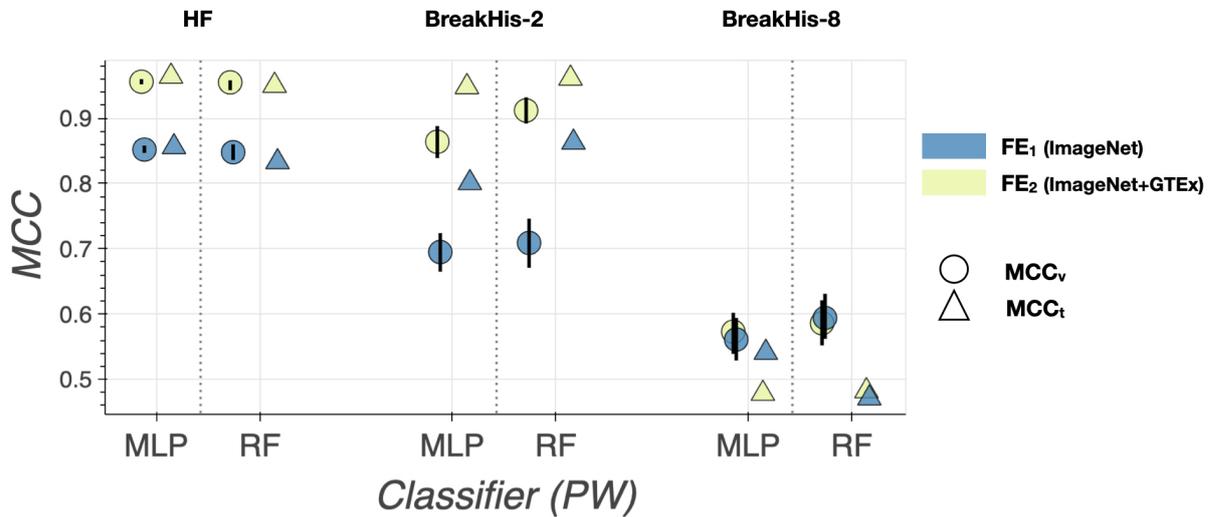


Figure 29: DAP results for each classifier head, using the *Patient-Wise* partitioning protocol, and the FE₁ and the FE₂ feature embedding with ResNet-152 as backbone model. The average MCC_v cross validation MCC_v with 95% CI are reported, along with MCC_t on the test set.⁴

The adoption of the **GTE_x** domain-specific dataset for transfer learning proves to be beneficial over the use of ImageNet only (Figure 29, Table 22, and Table 21). Notably, the *Patient-Wise* partitioning protocol with the FE_2 embedding has comparable performance with FE_1 and the inflated *Tile-Wise* splitting (Table 20). However, minor improvements are achieved on the BreakHis-8 task, with results not reaching state of the art. It must be observed that the BreakHis dataset is highly imbalanced in the multi-class task. As a countermeasure, authors in [11, 562] adopted a balancing strategy during data augmentation, which we did not introduce here for comparability with the other experiments.

To verify how much of previous domain-knowledge can be still re-used for the original task, we devised an additional experiment on the **GTE_x** dataset: on the *Feature Extractor* component (i.e. Convolutional Layers) of the model trained on **GTE_x** and fine-tuned on BreakHis-2, we add back the **MLP** classifier of the model trained on **GTE_x**. Notably, this configuration recover high predictive performance (i.e. $MCC_t=0.983$) on the classification task after only a single epoch of full training on **GTE_x**.

Patient-level Performance Analysis

We report patient-wise performance using the ResNet-152 backbone model with either the FE_1 feature embedding and both *Tile-Wise* and *Patient-Wise* protocols (Table 18), or with the FE_2 strategy and the *Patient-Wise* split (Table 19).

4 Original image, not available in the published manuscript.

Table 18: Patient-level results for each classifier head, using the *Patient-Wise* and *Tile-Wise* partitioning protocols, and the FE_1 feature embedding with the ResNet-152 backbone model. The average CV Patient-level ACC_v with 95% CI, and corresponding ACC_t on the test set are reported. *Others* reports the highest ACC among the compared papers.

Dataset	Metric Partitioning		MLP		RF		Others	
	Protocol		ACC_v	ACC_t	ACC_v	ACC_t	ACC_v	ACC_t
HF	WA	TW	0.984 (0.982, 0.987)	0.995	0.984 (0.981, 0.986)	0.995	-	-
		PW	0.981 (0.975, 0.986)	0.951	0.977 (0.971, 0.983)	0.927	0.940 [343]	
BreaKHis-2	PS	TW	0.995 (0.994, 0.996)	0.997	0.997 (0.996, 0.998)	0.998	0.872 [550]	
		PW	0.864 (0.851, 0.877)	0.885	0.883 (0.869, 0.898)	0.893	0.976 [554]	
BreaKHis-8	PS	TW	0.963 (0.960, 0.967)	0.950	0.957 (0.955, 0.959)	0.962	0.964 [553]	
		PW	0.687 (0.667, 0.709)	0.752	0.705 (0.685, 0.728)	0.725	0.967 [554]	

Table 19: Patient-level results for each classifier head, with the *Patient-Wise* partitioning protocol and the FE_2 feature embedding with the ResNet-152 model. The average cross-validation Patient-level accuracy with 95% CI (ACC_v) and corresponding scores on the test set (ACC_t) are reported. The *Others* column reports the highest accuracy achieved among the compared papers.

Dataset	Patient-level Metric	MLP		RF		Others
		ACC_v	ACC_t	ACC_v	ACC_t	ACC_v
HF	WA	0.992 (0.989, 0.995)	0.976	0.989 (0.984, 0.992)	0.976	0.940 [343]
BreaKHis-2	PS	0.941 (0.930, 0.951)	0.971	0.958 (0.948, 0.968)	0.991	0.976 [554]
BreaKHis-8	PS	0.691 (0.669, 0.716)	0.721	0.699 (0.676, 0.723)	0.724	0.967 [554]

5.4 Discussion

All the related works considered in this study uses different experimental protocols; factors such as preprocessing, data augmentation, and transfer learning methods can explain different accuracy scores obtained for a deep learning model on the same database. We report here, for completeness, a short description of the approach employed in the compared papers; we refer to a *Patient-Wise* partitioning protocol when the authors clearly state the adoption of a train-test split consistent with the patient, or when the code is provided as reference.

Nirschl et al. [343] train a CNN on the HF dataset to distinguish patients with or without heart failure. They systematically apply the *Patient-Wise* rule for the initial train-test split (50-50) and for the training partition into three-folds for cross-validation. Data augmentation strategies are also applied, including random cropping, rotation, mirroring, and staining augmentation. As for the BreaKHis dataset, Alom et al. [554] use a 70-30 *Patient-Wise* partitioning protocol to train a CNN with several (not specified) hidden layers, reporting average results from 5-fold cross-validation. Further, the authors apply augmentation strategies (i.e., rotation, shifting, flipping) to increase the dataset by a factor of $21 \times$ for each magnification level. The work of Han et al. [562] proposes a novel CNN adopting a *Tile-Wise* partition with the

training set accounting for the 50% of the dataset. Data augmentation (i.e. intensity variation, rotation, translation, and flipping) is used to adjust for imbalanced classes. Jiang et al. [561] train two different variants of the ResNet model to address the binary and the multi-class task, for each magnification factor. They adopt a *Tile-Wise* partitioning protocol for the train-test split, using 60% and 70% of the data in the training set for BreakHis-2 and BreakHis-8, respectively. Data augmentation is also exploited in the training process, and experiments are repeated 3 times.

Other authors employed a similar protocol to address the BreakHis-8 task by training a CNN pretrained on ImageNet: Nawaz et al. [553] implemented a DenseNet-inspired model, while Nguyen et al. [556] chose a custom CNN model, instead. Both studies use a *Tile-Wise* partition on the BreakHis dataset (70-30 and 90-10, respectively), and do not apply any data augmentation. Xie et al. [550] adapt a pre-trained ResNet-V2 to the binary and multiclass tasks of BreakHis, at different magnification factors, using a 70-30 *Tile-Wise* partition. Data augmentation has been applied to balance the least represented class in BreakHis-8. Jannesary et al. [552] used a 90-10 *Tile-Wise* train-test split with data augmentation (i.e. resizing, rotations, cropping and flipping) to fine-tune a ResNet-V1 for binary and multi-class prediction. Moreover, experiments in [552] were performed combining images at different magnification factors in a unified dataset. Finally, both [544] and [548] used a *Tile-Wise* train-test split for prediction of malignant vs benign samples using a pre-trained CNN and [548] also employed data augmentation (rotation and flipping).

5.5 Appendix

Table 20: DAP results for each classifier head, using the *Tile-Wise* partitioning protocol, and the FE_1 feature embedding with the ResNet-152 as backbone model. The average cross validation ACC_v with 95% CI are reported for each classification task, along with ACC_t on the test set. The *Others* column reports the highest accuracy achieved among the compared papers.

Dataset	MLP		RF		Others
	ACC_v	ACC_t	ACC_v	ACC_t	ACC_t
GTE _x	0.999 (0.999, 0.999)	0.999	0.999 (0.999, 0.999)	0.998	-
HF	0.980 (0.978, 0.982)	0.978	0.978 (0.977, 0.980)	0.980	-
BreaKHis-2	0.995 (0.994, 0.996)	0.994	0.996 (0.995, 0.997)	0.997	0.993 [561]
BreaKHis-8	0.959 (0.956, 0.962)	0.940	0.946 (0.943, 0.949)	0.940	0.985 [552]

Table 21: DAP results for each classifier head, using the *Patient-Wise* partitioning protocol, and the FE_1 feature embedding with the ResNet-152 as backbone model. The average cross validation ACC_v with 95% CI are reported for each classification task, along with ACC_t on the test set. The *Others* column reports the highest accuracy achieved among the compared papers.

Dataset	MLP		RF		Others
	ACC_v	ACC_t	ACC_v	ACC_t	ACC_t
GTE _x	0.998 (0.998, 0.998)	0.998	0.997 (0.997, 0.998)	0.997	-
HF	0.927 (0.924, 0.929)	0.915	0.924 (0.918, 0.930)	0.915	0.932 [343]
BreaKHis-2	0.870 (0.856, 0.882)	0.924	0.876 (0.859, 0.892)	0.946	0.973 [554]
BreaKHis-8	0.679 (0.655, 0.703)	0.644	0.701 (0.681, 0.732)	0.600	0.973 [554]

Table 22: DAP results for each classifier head, using the *Patient-Wise* partitioning protocol, and the FE_2 feature embedding with ResNet-152 as backbone model. The average cross validation ACC_v with 95% CI are reported, along with ACC_t on the test set. The *Others* column reports the highest accuracy achieved among the compared papers.

Dataset	MLP		RF		Others
	ACC_v	ACC_t	ACC_v	ACC_t	ACC_t
HF	0.978 (0.976, 0.980)	0.982	0.977 (0.975, 0.979)	0.978	0.932 [343]
BreaKHis-2	0.941 (0.930, 0.952)	0.980	0.963 (0.955, 0.971)	0.984	0.973 [554]
BreaKHis-8	0.685 (0.661, 0.712)	0.603	0.699 (0.675, 0.724)	0.606	0.973 [554]

6

THE histolab LIBRARY

histlab

Authors: A. Marcolini, N. Bussola, E. Arbitrio, M. Amgad, G. Jurman, C. Furlanello. *Title:* histolab: a Python library for reproducible Digital Pathology preprocessing with automated testing. *Submitted* (May 2022)

THE HISTOLAB LIBRARY

Highlights

- `histolab` is a novel Python library developed to improve standardization and reproducibility of WSI pre-processing.
- `histolab` is compatible with nine WSI formats and includes seven main modules for image manipulation, and automated tile extraction.
- `histolab` is public, multi-platform, fully documented, and designed to be efficiently embedded into AI pipelines for Digital Pathology.
- The robustness of `histolab` relies on a comprehensive suite of software tests, and it is the first open source library for WSI pre-processing with 100% of code covered by unit and integration tests.

Personal contribution I equally contributed to software development and testing of all `histolab` modules, with A. Marcolini and E. Arbitrio. In addition, I designed and evaluated the nuclei segmentation algorithm (Section 6.4.0.2) and wrote the online `histolab` documentation.

6.1 Abstract

DL is rapidly permeating the field of Digital Pathology with algorithms successfully applied both to ease daily clinical practice and to provide novel information to pathologists. Most DL workflows for Digital Pathology however include custom code for data preprocessing, usually tailored to data and tasks of interest, resulting in software that is hard to understand, peer-review and test. In this work, we introduce `histolab`, a Python package designed to standardize the pre-processing of WSIs in a reproducible environment, supported by automated testing. The package provide functions for building datasets of WSI subwindows (tiles), including augmentation

and morphological operators, a tile scoring framework and stain normalization methods. `histolab` is designed to be modular, extensible, and easily integrable in DL pipelines for Digital Pathology. To guarantee robustness, `histolab` embraces software engineering best practices such as multi-platform automated testing and Continuous Integration.

Summary

Here we introduce `histolab`, a novel open-source Python package for [WSI](#) preprocessing, in order to provide a standard environment for [DL](#) applications in Digital Pathology that is endowed with automated testing. First, `histolab` is designed to be easily integrated in a computational pipeline and contributes to a reproducible setting. Notably, `histolab` includes a comprehensive software test suite for automated early detection of design flaws (see [Section 6.4](#)), thus mitigating the risk of hidden technical debts [\[414\]](#). In particular, by evaluating early and with industry standards if the software can be trusted, we take a critical effort to avoid domino effects that could possibly jeopardize a whole pipeline.

In research, `histolab` has been adopted to standardize [WSI](#) preprocessing on the [GTEx](#) and [TCGA](#) public repositories.

The experimental setting of `histolab` is based on a high-level interface, which can be used for concise and reproducible scripting of tasks such as fine-tuned tile extraction, or sequences of preprocessing steps over intrinsic (size, color variability) and extrinsic (file format) factors. A step-by-step tutorial for experimenting with `histolab` is included in its official documentation ¹.

Related work

Among tools for [WSI](#) analysis and processing, there is a growth of solutions that are publicly available and open-source, suggesting an effort towards promoting repro-

¹ <https://histolab.readthedocs.io/en/latest/quickstart.html>

ducibility and collaborative research, also in clinical tasks [300]. To the best of our knowledge, four other projects have been developed with aims similar to `histolab`, *i.e.*, providing a reproducible, high-level tool for WSI preprocessing that can be easily plugged into a Digital Pathology workflow or in a larger bioimaging analysis solution: `PyHIST` [330], `deep-histopath` [130], `compay-syntax` [64], and `py-wsi` [437]. Further, two general-purpose research toolkits for computational pathology, more oriented to support AI models, have been recently introduced: both are named `PathML` [437]. `PyHIST` [330] is a semi-automatic command-line tool that implements different tile generation methods. `PyHIST` and `histolab` common features include: (i) automatic tissue detection, (ii) assessment of tissue area on the tiles; (iii) grid and random sampling approaches; (iv) multi-platform availability. However, `histolab` offers the following additional features: (v) a pool of filters to manipulate the images; (vi) a score-based extractor with customizable scoring functions; (vii) a data module with several example WSIs, (viii) a comprehensive test suite. Moreover, `PyHIST` currently supports only Aperio's SVS or TIFF formats, while `histolab` supports all OpenSlide formats. The `deep-histopath` project has been developed *ad hoc* to preprocess the TUPAC16² dataset and predict breast cancer tumor proliferation [130]. Its public repository provides many image preprocessing utilities useful to apply deep learning for histopathology image classification. Also, a novel nuclei scoring function is implemented, based on tissue percentage and color characteristics, which also allows the extracted tiles to be sorted and filtered. Although `histolab` and `deep-histopath` share several features, `histolab` has three main advantages: (i) the code is not tailored for a specific dataset; (ii) the package is available on PyPI; (iii) the proposed nuclei scoring function does not depend on empirically derived parameters. Moreover, `deep-histopath` does not include tests. `compay-syntax` is an OpenSlide-based tool for basic random tile sampling of WSIs [64]. `compay-syntax` has an intuitive API and it is published on PyPI, but its documentation is limited and software testing is missing.

Finally, `py-wsi` handles tile extraction on top of OpenSlide following a grid strategy and automatic labeling from Aperio ImageScope XML annotation files [437]. Notably,

² <http://tupac.tue-image.nl/>

it provides functions for saving tiles and their metadata into a LMDB database, as well as HDF5 format, and utilities for batch iteration over tiles, epoch counting and tiles shuffling. `py-wsi` has been published on PyPI, however it lacks a software testing suite.

6.2 Software description

Software Architecture

`histolab` is a Python-based software built on top of existing state-of-the-art libraries, *e.g.*, `OpenSlide` [167] for low-level `WSI` operations, `NumPy` [182] for fast numerical computations and `scikit-image` [495] for image processing algorithms. A preprocessing pipeline for extracting informative tiles from `WSI` datasets can be implemented in `histolab` by composing methods and data structures from six main components: `slide`, `filters`, `masks`, `tile`, `tiler`, `scorer`, and `data`.

Software Functionalities

The `histolab` library is composed of seven main modules for `WSI` processing: the `slide` module, the `filters` package, the `masks` module, the `tile` module, the `tiler` module, the `scorer` module, and the `data` module.

6.2.0.1 *The slide module*

The `slide` module provides a simple high-level interface to handle a `WSI`; it contains the `slide` class, which wraps functions, methods and properties of a virtual slide in a single object. The `slide` class encapsulates `OpenSlide`, possibly the most used open-source scientific software for Digital Pathology [167], and relies on the `openslide-python` library for the low-level operations on digital slides. A `WSI` is usually stored in pyramidal format, where each level corresponds to a specific magnification factor. Therefore, two relevant properties of a `WSI` are: (i) its dimensions

at native magnification; (ii) the number of levels and the dimensions at a specified level. OpenSlide identifies each magnification level of the WSI with a positive integer number, starting from 0.

A slide is initialized by providing the path where the WSI is stored and the path where the extracted tiles will be saved. Further, the slide module implements the SlideSet class, which handles a collection of Slide objects stored in the same directory, possibly filtered by the `valid_extensions` parameter.

The `slides_stats` property of a SlideSet computes statistics for the collection, namely the number of available slides; the slide with the maximum (minimum) width; the slide with the maximum (minimum) height; the slide with the maximum (minimum) size; the average width, height, and size of the slides.

6.2.0.2 *Filters*

Filters is a subpackage including 35 functions for image manipulation, such as contrast enhancement, color deconvolution or background removal and image quality control. Further, the filters component implements 9 morphological filters on binary masks, including basic morphological operations (*i.e.*, dilation, erosion, opening, closing). Many filters wrap scikit-image [495] and SciPy [489] functions, and reimplement functions from [130]. Notably, filters are designed and tested to be applied singularly or combined together as a chain of transformations through the Compose object.

6.2.0.3 *The masks module*

The classes implemented in the masks module define how to calculate a binary mask from a slide, which is necessary during the tiles extraction phase. The TissueMask class segments the tissue areas in the slide leveraging a sequence of histolab filters, including conversion to grayscale, Otsu thresholding, binary dilation, small holes and small objects removal. The BiggestTissueBoxMask class applies the same chain of filters as TissueMask and it returns a binary mask corresponding to the bounding box of the largest connected tissue region. As a diagnostic check, it is possible to draw

the contours of the selected mask on the slide thumbnail via the `slide`'s method `locate_mask`, passing the mask instance as a parameter.

6.2.0.4 *The tile module*

Feeding a neural network with a full resolution `WSI` (possibly exceeding 10GB) is currently impractical. Usually, a `WSI` is first divided into tiles covering the region or regions of interest [103]. Then, the tiles are used as sample inputs in a Deep Learning pipeline; in the simplest training procedure, each tile is treated as an independent object, usually associated to a diagnostic label assigned to the `WSI` of origin by an expert pathologist.

The tile module contains the `tile` class to manage a rectangular region cropped from a `slide`. A `tile` object is described by (i) its extraction coordinates at native magnification (corresponding to level 0 in `OpenSlide`), (ii) the level of extraction, (iii) the actual image, stored as a `PIL Image` [99]. A `tile` object will be created internally during the tile extraction process (Figure 30).

A `tile` object can be evaluated to retrieve information. In particular, the method `has_enough_tissue` checks if the proportion of the detected tissue over the total area of the tile is above a specified threshold (by default 80%). Internally, the method quantifies the amount of tissue by applying a chain of `histolab` filters, including conversion to grayscale, Otsu thresholding, binary dilation and small holes filling.

6.2.0.5 *The tiler module*

Different logics are implemented for tile extraction in the `tiler` module. The constructor of the three extractors `randomtiler`, `gridtiler`, and `ScoreTiler` share a similar interface and common parameters that define the extraction design: (P1) `tile_size`: the tile size; (P2) `level`: the extraction level, from 0 to the number of available levels; negative indexing is also possible, counting backward from the number of available levels to 0 (e.g., `level=-1` means selecting the last available level); (P3) `check_tissue`: True if a minimum percentage of tissue over the total area of the tile is required to save the tiles (see Section 6.2.0.4), False otherwise; (P4) `tissue_percent`: number between 0 and 100 representing the minimum required ratio of tissue over

the total area of the image, considered only if `check_tissue` equals to `True` (default is `80.0`); (P5) `prefix`: a prefix to be added at the beginning of the tiles' filename (optional, default is the empty string); (P6) `suffix`: a suffix to be added to the end of the tiles' filename (optional, default is `.png`).

The general mechanism is to (i) create a `tiler` object, (ii) define a `slide` object, used to identify the input image, and (iii) create a `mask` object to determine the area for tile extraction within the tissue. The extraction process starts when the `tiler`'s `extract` method is called, with the `slide` and the `mask` passed as parameters.

RandomTiler: The `randomtiler` extractor allows for the extraction of tiles picked at random within the regions defined by the binary mask object. Since there is no intrinsic upper bound of the number of the tiles that could be extracted (no overlap check is performed), the number of wanted tiles must be specified.

In addition to (P1)-(P6), the `randomtiler` constructor requires as two additional parameters the number of tiles requested (`n_tiles`), and the random seed (`seed`), to ensure reproducibility between different runs on the same [WSI](#). Note that less than `n_tiles` could be extracted from a slide with not enough tissue pixels and a lot of background, which is checked when the parameter `check_tissue` is set to `True`.

GridTiler: A second basic approach consists of extracting all the tiles in the areas defined by the binary mask. This strategy is implemented in the `gridtiler` class. The additional `pixel_overlap` parameter specifies the number of overlapping pixels between two adjacent tiles. Tiles of width w and height h are cropped by using a sliding window with stride $s = (w - p)(h - p)$.

ScoreTiler: Tiles extracted from the same [WSI](#) may not be equally informative; for example, if the goal is the detection of mitotic activity on [HE](#) slides, tiles with no nuclei are of little interest. The `ScoreTiler` extractor ranks the tiles with respect to a scoring function, described in the `scorer` module. In particular, the `ScoreTiler` class extends the `gridtiler` extractor by sorting the extracted tiles in a decreasing order, based on the computed score. Notably, the `ScoreTiler` is agnostic to the scoring

function adopted, thus a custom function can be implemented provided that it inputs a `Tile` object and outputs a number. The additional parameter `n_tiles` controls the number of highest-ranked tiles to save; if `n_tiles=0` all the tiles are kept.

Each extractor will automatically (i) calculate the binary mask on the `WSI` depending on the mask object provided; (ii) generate the tiles within the mask; (iii) save all generated tiles or only the informative ones if the attribute `check_tissue` is set to `True`.

To visualize the position of the tiles that can be extracted, each `Tiler` object provides the `locate_tiles` method, returning a `PIL Image` corresponding to a scaled version of the slide with the tiles' boundaries outlined.

6.2.0.6 *The scorer module*

The goal of the scorer module is to provide the grading functions for the `ScoreTiler` extractor described in Section 6.2.0.5. The scorer objects input a `Tile` object and return their computed score.

The main aim of the `NucleiScorer` class is to compute a score on tiles extracted from H&E slides based on the amount of nuclei detected. The automatic detection of cell nuclei on histological images is a well-known challenging task due to the high heterogeneity of slide preparation and image acquisition techniques [218]. Moreover, the morphology and texture of nuclei vary according to several factors such as tissue structure or disease type, often resulting in complex, overlapping clusters [189]. Among the automated approaches employed to perform the nuclei segmentation, the threshold-based methods have the advantage of requiring no parameter tuning or training procedure. Threshold-based techniques are usually integrated with other methods, such as color space conversion [4, 295], watershed transformation [378, 410], and morphological operations [360, 506].

The `NucleiScorer` class implements a hybrid algorithm that combines thresholding and morphological operations to segment nuclei on H&E-stained histological images. The proposed method is build upon native `histolab` filters, namely the `HematoxylinChannel` filter, the `YenThreshold` filter, and the `WhiteTopHat` filter. Performance of the proposed nuclei segmentation algorithm have been evaluated on the

UCSB public dataset [162] (Section 6.4.0.2). Notably, the algorithm does not require any parameter tuning.

The NucleiScorer class defines the score of a given tile t as:

$$s_t = N_t \cdot \tanh(T_t), 0 \leq s_t < 1 \quad (4)$$

where N_t is the nuclei ratio on t , computed as number of white pixels on the segmented mask over the tile size, and T_t the fraction of tissue in t . Notice that we introduced the hyperbolic tangent to bound the weight of the tissue ratio over the nuclei ratio.

6.2.0.7 *The data module*

The data module, based on Pooch [476], gives access to a collection of 11 slides dyed with different staining techniques (HE and Immunohistochemistry (IHC)). In particular, WSIs are retrieved from three public archives: TCGA³, OpenSlide⁴, and IDR⁵. The slides are downloaded and stored in the system cache and they can then be imported in a Python script (see Table 25 for a detailed description of the sample data provided with histolab).

³ <https://bit.ly/3G1RpQA>

⁴ <https://bit.ly/3AKrccW>

⁵ <https://bit.ly/34lmQx5>

6.3 Illustrative Example



Figure 30: histolab workflow to retrieve a tile dataset from a WSI with different extraction strategies (Slide of breast cancer from the TCGA repository, see Table 25). Left: code snippets of histolab major functions, from slide downloading and initialization (A, B), to the definition of the extractor for the tiling procedure (C_r, C_g, C_s), and the choice of the tissue mask (D) for the extraction of the tiles (E). Right: Visualization of each step on the slide thumbnail (A-D) and (E) representative examples of extracted tiles by using the C_r extractor on the whole tissue mask.

Figure 30 illustrates the main steps to extract a reproducible dataset of tiles from a tissue slide with `histolab`. In particular, a `WSI` of breast invasive carcinoma is retrieved from the `TCGA` repository through the data module (Figure 30A and Appendix 25) and initialized with the processed path where the tile dataset will be saved (Figure 30B). Different extraction strategies are then defined with custom parameters (Figure 30C_r- 30C_s), such as the tile size, the extraction level, and the minimum required percentage of tissue over the total area of the tile (see Section 6.2.0.5). The result of each extractor can be visualized via the `locate_tiles` method, as illustrated on the right of Figure 30C_r- 30C_s. The `TissueMask` class (see Section 6.2.0.3) is selected in order to consider the whole tissue area detected on the slide (Figure 30D, left) as highlighted by the `locate_tiles` function on the slide thumbnail (Figure 30D, right). Finally, the extraction procedure is performed by the `extract` method on the chosen extractor, passing the slide and the mask as parameters (Figure 30E). The retrieved tiles are automatically saved in the specified processed path.

6.4 Appendix

Automated Testing Methods

6.4.0.1 *Continuous Integration pipeline*

The intent of the `histolab` project is to warrant reliability and quality of scientific software for preprocessing high-throughput digital pathology images. This effort is specifically aimed at improving reproducibility of deep learning frameworks applied to `WSIs`, possibly in combination with other bioinformatics tools. In practice, we embrace industrial software standards that are commonly used in software development and, less frequently, in open source software development communities.

Automated software testing is a specific feature of `histolab`. This Python package includes a comprehensive test suite, composed currently of 320 unit tests and 280 integration tests. The tests are developed using the `pytest` framework (v6.2.5), a full-featured Python testing tool that supports complex functional testing for applications

and libraries. In addition to `pytest`, the Python standard library packages `unittest` and `mock` are used. A synthesis of the Continuous Integration (CI) pipeline adopted for `histolab` is outlined in Figure 31. CI has been previously introduced in major Python scientific open source projects such as `Pandas` and `Numpy` [182], for which a coverage fraction of 93% and 85% are documented. For `DP`, the `histomicsTK` library [13] has adopted CI / automated testing, currently resulting in a 73% coverage. However, the data preprocessing section in known DL pipelines for digital pathology such as `PyHIST` [330], and `compay-syntax` [64] does not appear to explicitly include CI.

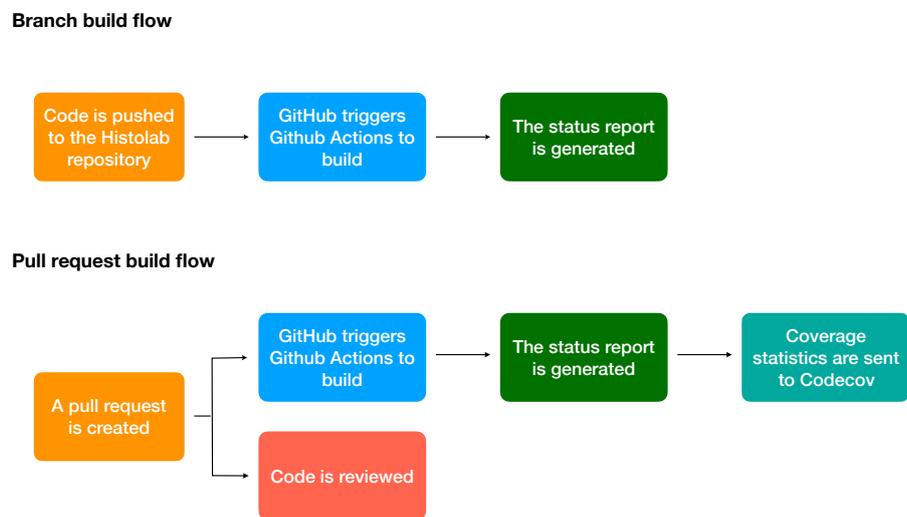


Figure 31: Overview of the `histolab` branch build workflow (top) and the Pull Request build workflow (bottom).

6.4.0.2 Coverage statistics

We adopted the Python package `coverage.py` to compute detailed test coverage statistics for `histolab`. The coverage percentage for each file is computed as:

$$\text{coverage} = \frac{\# \text{ executed lines} + \# \text{ execution paths (branches) hit by at least one test}}{\# \text{ relevant lines} + \# \text{ execution paths}} \cdot 100$$

Note that contributing software is accepted only if the total coverage is not reduced, currently corresponding to 100%. Thus any contribution is expected to be fully covered by tests.

Nuclei segmentation

To assess the validity of our automated nuclei segmentation algorithm (Figure 35), we estimated its accuracy on the public UCSB breast cancer dataset [162]. This public dataset includes 59 HE-stained images of 30 benign and 29 malignant breast cancer samples of size $896\text{px} \times 768\text{px}$ and for each image it provides ground truth binary masks of segmented nuclei on a smaller region of size $200\text{px} \times 200\text{px}$. To compute the number of nuclei on the binary masks, we first applied histolab's Watershed filter and then we counted labeled regions with area greater than 100 pixels (Figure 32-34).

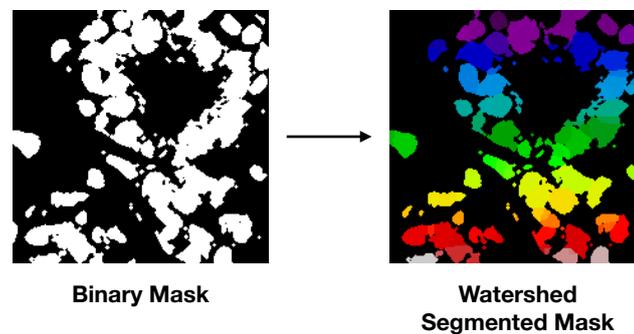


Figure 32: Example of Watershed filter applied on a nuclei binary mask of a tile extracted from a WSI sample from the UCSB breast cancer dataset.

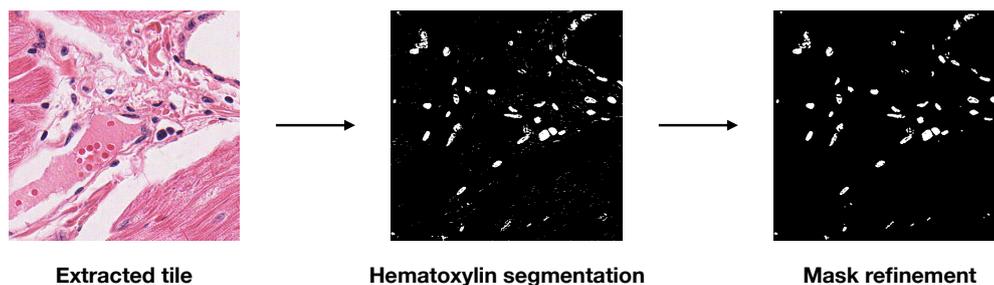


Figure 33: Nuclei segmentation procedure adopted by the `NucleiScorer` class: the hematoxylin channel is filtered with the Yen threshold and the mask is then refined with the white top hat morphological operation. Tile from the OS-HEART WSI (Table 25).

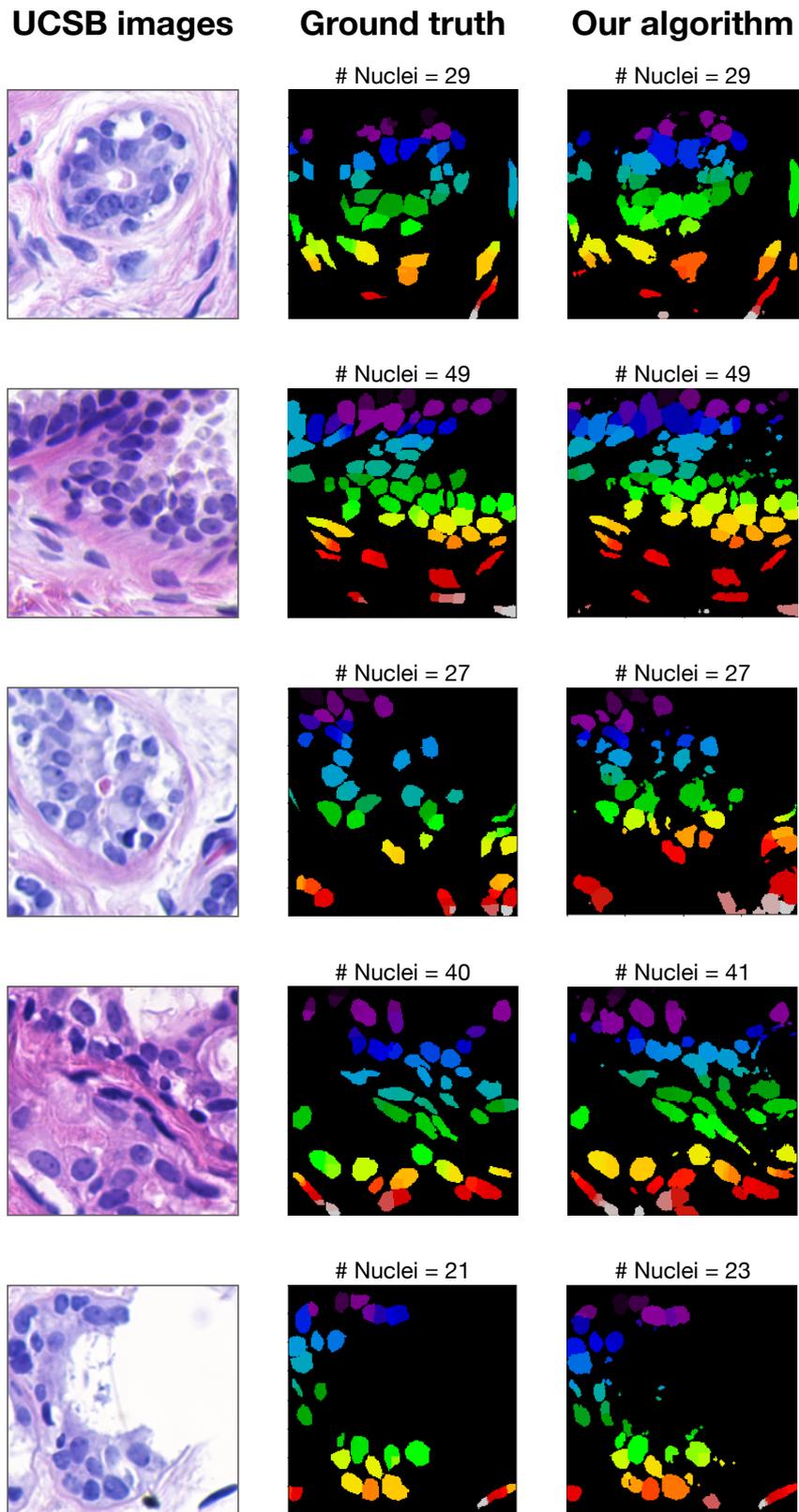


Figure 34: Nuclei segmentation performance of benign samples from the UCSB dataset, compared to the ground truth.

We quantified the segmentation results using pixel-wise metrics as defined in [291] (Table 23). Cell nuclei in malignant samples of the UCSB dataset present the salt-and-pepper chromatin patterns, typically detected in endocrine tumors [395]. The fine granularity of the chromatin results in a dispersion of the hematoxylin staining, explaining the lower pixel-wise performance of the segmentation in malignant samples.

Table 23: Pixel-wise performance of the NucleiScorer (NS) segmentation algorithm on the UCSB dataset, for benign (NS-Benign), malignant (NS-BRC), and all data (NS-All). Results are expressed as a percentage.

Samples	Accuracy (std)	Precision (std)	recall (std)	F1 score (std)	Jaccard index (std)
NS-Benign	90.81 (0.31)	74.27 (0.75)	93.93 (0.50)	82.54 (0.40)	70.47 (0.58)
NS-BRC	87.95 (0.65)	77.19 (1.03)	73.66 (2.64)	70.31 (1.94)	57.04 (1.89)
NS-All	89.48 (0.52)	75.62 (0.91)	84.55 (2.09)	76.88 (1.49)	64.26 (1.52)

Table 24: Pixel-wise performance of nuclei segmentation methods on the UCSB dataset: threshold-based methods (Buggenthin *et al.* [59], Al-Kofahi *et al.* [5]); clustering-based by Tang *et al.* [451]. Results are expressed as a percentage.

Method	accuracy	precision	F1 score	Jaccard index
[59]	84.36	61.37	72.50	–
[5]	82.27	61.03	74.15	60.16
[451]	86.60	64.04	76.34	–

In terms of pixel-wise accuracy (a), precision (p), F1 measure ($f1$), and Jaccard index (j) the NucleiScore algorithm outperforms the threshold-based approaches proposed by Buggenthin *et al.* [59] and Al-Kofahi *et al.* [5], and the precision of the proposed algorithm is significantly higher than the clustering-based technique defined by Tang *et al.* [451], as summarized in Table 24. The lower recall and higher precision with respect to the compared methods implies a reduced number of false negatives detected.

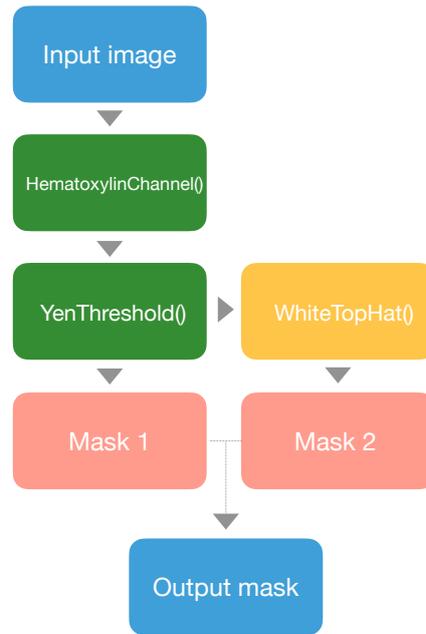
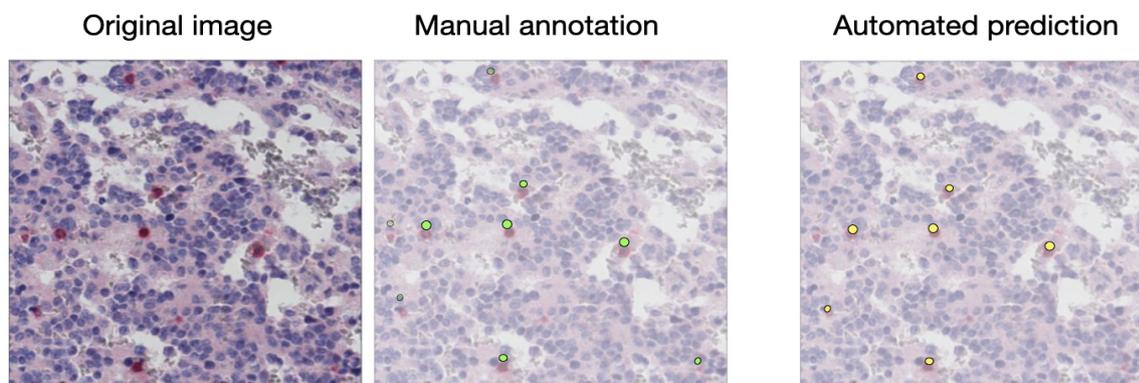
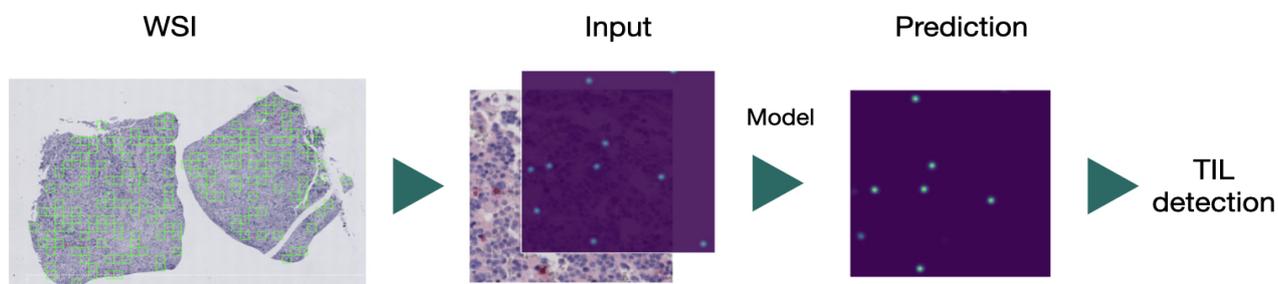


Figure 35: Schematic workflow of the algorithm implemented for the automated nuclei segmentation.

Table 25: Set of WSIs downloadable with the histolab data module. TCGA-BRCA: TCGA Breast Invasive Carcinoma dataset; TCGA-PRAD: TCGA Prostate Adenocarcinoma dataset; TCGA-OV: Ovarian Serous Cystadenocarcinoma dataset. IDR: Image Data Resource. H&E: Hematoxylin and Eosin. IHC: Immunohistochemistry

WSI ID	Tissue	Dimensions ($w \times h$)	Size (MB)	Repository	Staining
OS-AORTA	Aorta	15,374px \times 17,497px	63.8	OpenSlide	H&E
OS-HEART	Heart	32,672px \times 47,076px	289.3	OpenSlide	H&E
TCGA-BREAST	Breast	96,972px \times 30,682px	299.1	TCGA-BRCA	H&E
TCGA-PROSTATE	Prostate	16,000px \times 15,316px	46.1	TCGA-PRAD	H&E
TCGA-OVARY	Ovary	30,001px \times 33,987px	389.1	TCGA-OV	H&E
TCGA-BREAST-RED	Breast	60,928px \times 75,840px	510.9	TCGA-BRCA	H&E
TCGA-BREAST-GREEN	Breast	98,874px \times 64,427px	719.6	TCGA-BRCA	H&E
TCGA-BREAST-BLACK	Breast	121,856px \times 94,697px	1740.8	TCGA-BRCA	H&E
IDR-BREAST	Breast	9960px \times 7121px	218.3	IDR [411]	IHC
IDR-KIDNEY	Kidney	5179px \times 4192px	66.1	IDR [411]	IHC

TILs DETECTION IN NEUROBLASTOMA



Authors: N. Bussola, B. Papa, O. Melaiu, A. Castellano, D. Fruci, G. Jurman. *Original title:* Quantification of the immune content in Neuroblastoma: Deep Learning and Topological Data Analysis in Digital Pathology. *Published in:* International Journal of Molecular Sciences. (Aug. 2021)

TILS DETECTION IN NEUROBLASTOMA

Highlights

- A novel DL pipeline is introduced for automated quantification of TILs on histopathology images in precision oncology.
- The proposed framework is first tested on an original collection of Neuroblastoma WSIs with IHC staining for T-cell detection over the whole tissue area.
- WSI preprocessing is performed with *histolab* to ensure reproducibility, and patient-wise data partitioning protocols are adopted to avoid data-leakage effects.
- TDA descriptors are used to investigate the inner states of DL architectures during training for model interpretability.
- The unsupervised analysis of the deep features reveals clusters of tiles with similar histological patterns, e.g. pseudo-necrotic tissues or stroma poor areas.
- Agreement with human estimates indicates our framework a promising tool in the clinical setting, to support pathologists for the precise quantification of immune content in cancer samples.

Personal contribution I personally scanned the WSI collection with Dr. O. Melaiu. I designed the overall approach, supervised the experiments (performed by B. Papa), and evaluated the results, in particular the deep feature analysis. I contributed to the annotation of histology tiles under the supervision of Dr. O. Melaiu. I also drafted the paper and prepared the figures.

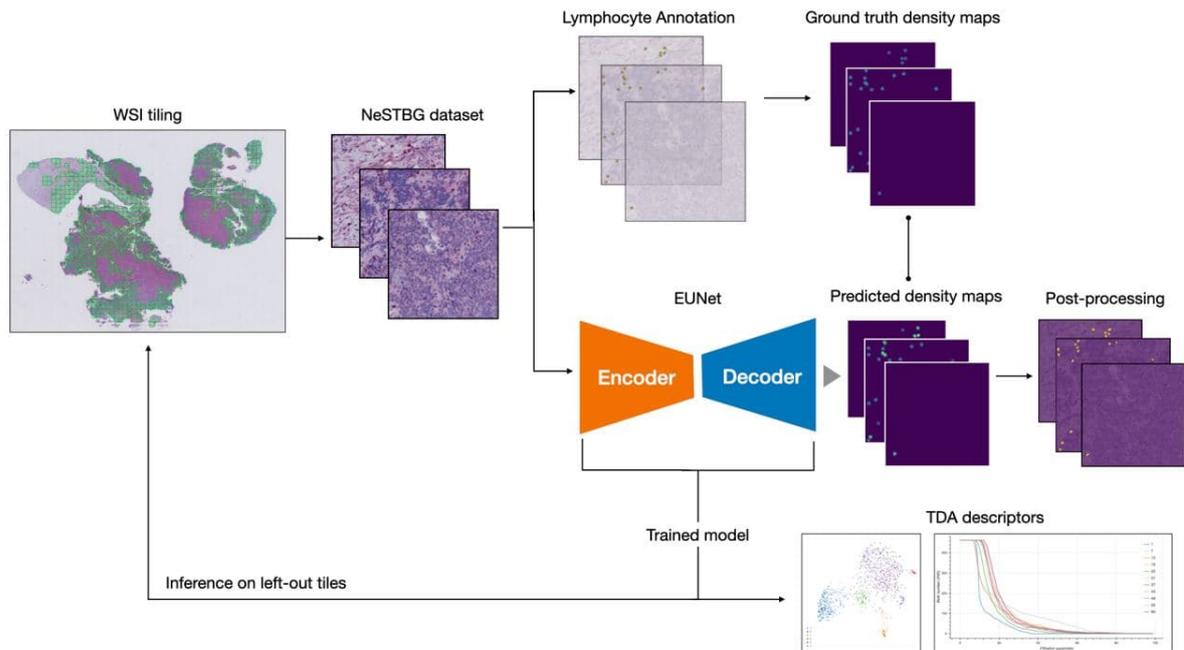
7.1 Abstract

We introduce here a novel Machine Learning (ML) framework to address the issue of the quantitative assessment of the immune content in Neuroblastoma (NB) specimens. First, the EUNet, a U-Net with an EfficientNet encoder, is trained to detect

lymphocytes on tissue digital slides stained with the CD3 T-cell marker. The training set consists of 3782 images extracted from an original collection of 54 [WSIs](#) manually annotated for a total of 73751 lymphocytes. Resampling strategies, data augmentation, and transfer learning approaches are adopted to warrant reproducibility and to reduce the risk of overfitting and selection bias. [TDA](#) is then used to define activation maps from different layers of the neural network at different stages of the training process, described by Persistence Diagrams (PD) and Betti curves. [TDA](#) is further integrated with the [UMAP](#) dimensionality reduction and the [HDBSCAN](#) algorithm for clustering by deep features the relevant subgroups and structures, across different levels of the neural network. Finally, the recent TwoNN approach is leveraged to study the variation of the intrinsic dimensionality of the U-Net model. As the main task, the proposed pipeline is employed to evaluate the density of lymphocytes over the whole tissue area of the [WSIs](#). The model achieves good results with mean absolute error 3.1 on test set, showing significant agreement between densities estimated by our EUNet model and by trained pathologists, thus indicating the potentialities of a promising new strategy in the quantification of the immune content in NB specimens. Moreover, the [UMAP](#) algorithm unveiled interesting patterns compatible with pathological characteristics, also highlighting novel insights on the dynamics of the intrinsic dataset dimensionality at different stages of the training process.

Summary

Figure 36: Graphical representation of the full analysis workflow. From the original WSIs the collection of tiles constituting the NeSTBG dataset is generated and annotated to obtain the ground truth for the model. Tiles are then used as the input for the DL architecture EUNet to predict density maps, that are then post-processed and analysed via TDA descriptors to interpret the detected deep features.



Here we propose a novel AI procedure, sketched in Figure 36, for the quantitative evaluation of the immune content in NB specimens in a IHC / Digital Pathology framework, *i.e.*, using WSIs as the input data: in detail, a Deep Learning predictive model is trained to estimate the density of lymphocytes over the whole tissue area of the WSIs. The approach is demonstrated on the Neuroblastoma Specimens with T-Lymphocytes - Bambino Gesù (NeSTBG), an original dataset of samples from NB patients, provided by Ospedale Pediatrico Bambino Gesù (OPBG) in Rome, achieving a satisfying performance ($MAE \approx 3.1$). To boost reproducibility and interpretability of the DL model, the extracted deep features are analyzed by TDA methods [72, 428] and, in particular, Persistent Homology (PH) [212]. To date, this is the first realization of an Explainable Artificial Intelligence (XAI) reproducible platform encompassing all the analysis steps from WSI preprocessing to clinical feature interpretation inte-

grating topological concepts with deep architectures. Although currently more an effective proof of concepts than a fully fledged infrastructure, the novel link established between DL and TDA in DP can lead to further developments along this research line.

Neuroblastoma

Neuroblastoma (NB) is the most common cancer diagnosed in the first year of life [94] affecting the sympathetic nervous system. NB is a heterogeneous disease with different outcomes ranging from spontaneous regression to aggressive progression, metastasis and death. Two main staging systems have been created to stratify patients based on the wide range of outcome and tumor biology: the International Neuroblastoma Staging System (INSS), introduced in 1988 [54] and revised in 1993 [53], and the International Neuroblastoma Risk Group Staging System (INRGSS), introduced in 2009 by the International Neuroblastoma Risk Group task force [104]. INRGSS enhances INSS by defining a series of Imaging Defined Risk Factors based on radiological data such as CT scans and MRI assessing whether the tumor is circumscribed, if it has metastasized or if it develops near vital parts of the body. However, the effort of establishing an accurate staging system is still ongoing: for instance, additional factors like the MYCN status, the histopathologic classification, and the DNA content have also proven to be significant [114], and they are currently evaluated in clinical practice. Tumor microenvironment (TME) has been shown to play a role in the tumor development. In particular, several pieces of evidence highlighted the importance of the composition, density and distribution of tumor-infiltrating lymphocytes as prognostic markers in several human cancers, including NB. Such observations are stimulating a growing research flow targeting the dynamics of the immune system during the NB evolution [502], following the Precision Medicine paradigm [29, 106, 145, 219].

Related works

Immunohistochemistry

The **IHC** technique is particularly suitable to evaluate the density of tumor infiltrating immune cells on histology specimens [240]; by employing the right binding antibody, it is indeed possible to highlight specific immune cells on the tissue, allowing pathologists to obtain information about their localization within the tumor microenvironment. Given the importance of **TILs** in recognizing and neutralizing cancer cells, several studies have been conducted on different tumor types [100, 148, 433], and the key prognostic significance of these cells has been highlighted. In NB, the study of the immune response can be traced back to more than 50 years ago [70, 140, 262, 303]. However, the adoption of **IHC** to evaluate the role of the immune infiltration for the prognosis of NB patients has its landmark in the work by Mina and colleagues [320]. The authors demonstrated that tumor-infiltrating T cells have a prognostic value greater than, and independent of, the criteria currently used to stage NB. In this thorough study different **IHC** biomarkers were used, including the Cluster of Differentiation 3 (CD3). CD3 is a complex of trans-membrane proteins representing an appropriate target for T Cells, also used as a representative marker in the present work. As a major result of [320], a positive correlation is found between the density of CD3 positive cells (*i.e.*, CD3+ cells) and the overall patient survival.

Lymphocyte detection and density maps

As pointed out in several references [443, 483], detecting and quantifying lymphocytes represent a powerful tool to identify strong prognostic and predictive biomarkers for evaluating cancer progression and targeting novel therapeutic solutions. Nonetheless, it is widely acknowledged that the technical challenges to be solved towards the goal are numerous and difficult, making the aforementioned tasks very hard to tackle and indeed still an open problem. No shared consensus has been currently reached

by the community on an optimal methodology: automatic localization and quantification of lymphocytes have represented a major goal in DP in the last decade, resulting in a constant stream of publications featuring the emerging solutions in both imaging and learning, together with dedicated reviews detailing such evolution [408]. Focusing on the methods adopting DL algorithms, convolution is the natural tool common to many proposals, starting from [221], where CNNs were combined with a probability map to identify lymphocytes' centers. Other approaches employed CNNs as a classifier to discriminate lymphocytes from the image background [274], generating a heatmap representing the probability of each pixel being a lymphocyte. The strategy to move from the heatmap to the lymphocytes identification was later improved in [406] through thresholding, while clinical relevance was made explicit by detecting local spatial features [26]. Further technical improvements on the same directions were achieved in [43] with the development of a non-maxima suppression (NMS) algorithm to locate the center of each lymphocyte. Finally, combination with a more classical morphologically-oriented procedure [488] allowed Li and coauthors [267] to better identify the center of each candidate cell.

The landscape of solutions is quite rich even when restricting to the detection of lymphocytes in IHC-stained WSIs. A first approach combining CNNs with color deconvolution to produce a probabilistic heatmap [91] was later improved in [159] again via a NMS procedure. An important breakthrough came when the general YOLO architecture [382] was adapted to the lymphocyte detection task. After the first attempt [444] where a non-maxima suppression algorithm was used during inference to consider overlapping bounding boxes as detecting the same lymphocyte, in 2019 Van Rijthoven and colleagues proposed YOLLO [567], a modified version of YOLO [382], as a fast method to detect lymphocytes in IHC-stained WSIs. The proposed modifications to YOLO included a guided sampling strategy and a simplified architecture, resulting in both a performance gain and a procedure speed-up. Finally, in [443], YOLLO combined with non-maxima suppression is compared to other approaches based on U-Net, a fully convolutional neural network, and Locality Sensitive Method (LSM).

In particular, two families of algorithms deserve being mentioned for the quite large popularity gained in the last few years, both stemming from the original R-CNN model [166]. The former set of architectures is mainly aimed at quick object detection, with Fast R-CNN [165] as the first implementation, followed by its improved version Faster R-CNN [386]. These models still work as building blocks for recent solutions in DP, as in [280, 282, 523]. The latter family of models stem from the prototypal structure Mask R-CNN [192, 193], obtained by optimizing the Faster R-CNN for pixel-level segmentation tasks. Use of Mask R-CNN and derived models is also currently quite widespread in the DP community, with several examples published in the literature [12, 129, 150, 499].

At the same time, crowd counting has always been a challenging task in computer vision. The idea of tackling the counting problem with density maps begun with [265]. Then Zhang and colleagues [527] started using DL models to predict object density maps, later refined in [22] through a new encoder-decoder CNN for crowd counting in aerial images. Similar strategies have been recently used in computational biology for yeast cells detection [246] and also in DP [224], where density maps are used to count cells in histology images of bone marrow tissues.

7.2 Materials and Methods

Figure 37: Summary of clinical features and age distribution at diagnosis (month) for the 54 patients of the NeSTBG dataset.

GENDER	Male 30 (55%) Female 24 (45%)
INSS	1 18 (33%) 2 10 (19%) 3 5 (9%) 4 16 (30%) 4S 5 (9%)
INRGSS	L1 26 (48%) L2 7 (13%) M 16 (30%) MS 16 (9%) 4S 5 (9%)
COG	Low 24 (44%) Intermediate 19 (35%) High 11 (20%)
PROGNOSIS	Favourable 40 (74%) Unfavourable 14 (26%)
MYTOTIC FREQUENCY	Low 24 (44%) Medium 1 (2%) High 24 (44%) Unknown 5 (9%)
MYCN AMPL.	Absent 31 (57%) Present 13 (24%) Gain 8 (15%) Unknown 2 (4%)
1P36 DELETION	Yes 8 (15%) No 36 (67%) Imbalance 5 (9%) Yes + Imbalance 1 (2%) Unknown 4 (7%)
MORPHOLOGY	Stroma poor, poorly differentiated 42 (78%) Stroma poor, differentiated 4 (7%) Stroma poor, undifferentiated 3 (6%) Other 5 (9%)

The NeSTBG dataset

The NeSTBG dataset is a collection of 3782 tiles with annotations for the centers of lymphocytes for 54 IHC-stained WSIs of as many NB specimens, previously characterized for density of tumor infiltrating immune cells, including T cells [320], dendritic cells and natural killer cells [316], as well as the expression of PD-L1 and PD-1 [315]. CD3 stained slides were digitized by the Menarini D-Sight scanner at native magnification 40x (resolution $0.25 \frac{\mu m}{pixel}$) and employed for digital analysis. The 54 patients are reasonably gender balanced (30 males vs. 24 females), mostly younger than 4 years at diagnosis. INSS, INRGSS and COG values are quite heterogeneous, as well as the tumour location, with suprarenal position as the most frequent (24 patients, 44%); less frequent locations include lymph nodes, aorta, scapula, eye, pharynx, and spleen. The full set of clinical characteristics of the 54 patients are summarised in Fig-

ure 37. Morphologically, the large majority of the tumours in the cohort are stroma poor (91%), and in particular poorly differentiated (42 patients, 78%). The remaining 12 samples include 4 differentiated and 3 undifferentiated stroma poor cases, together with an undifferentiated case and 4 ganglioneuroblastoma (GNBL), with only a single stroma rich case. Furthermore, at a 40x magnification level, all samples have about 560 tumoral cells in each sector, while pseudonecrosis areas are mainly present in Stage 4 samples. Note that the heterogeneity of the stroma in the cohort does not represent a confounding factor in the analysis: our experience suggests that immune cells can infiltrate the tumour tissue regardless of the morphology of the stroma, thus yielding that tissue composition is not directly correlated to the immune content. Further, CD3 staining is extremely clean and specific, and the background noise is reduced by precise stain tuning and by blocking the non specific binding sites, with no need for preprocessing procedures reducing stain variability. Each tile in NeSTBG is a 512 by 512 pixels RGB image stored in *png* format, randomly extracted from a [WSI](#) at 20x magnification.

Annotations refer to the x and y coordinates of the centers of the lymphocytes found in each tile. The tiles are extracted from digitised tumor samples of NB patients, diagnosed between the 2002 and 2013 at OPBG [320]. In particular, NeSTBG is generated from the subset of the samples treated with monoclonal antibodies against CD3, and visualized with the Fast Red chromogen substrate. All the slides are also counterstained with Haematoxylin, the dye used to visualize cell nuclei. Among the ten available stainings, only the subset of slides treated with anti-CD3 antibodies was selected and coloured with Fast-Red. Moreover, the specimens obtained through needle ago-biopsy were excluded, as suggested by an expert pathologist, as they may be less representative of the tumoral status. Level 1 in the OpenSlide standard [167], corresponding to 20x magnification and $0.5 \frac{\mu m}{pixel}$ resolution, was selected for the images as a trade-off between image details and computational load, being sufficiently detailed to detect marked cells and also to describe [WSIs](#) using a limited number of tiles. Segmentation of the tissue region within the slide was also needed: a large portion of a [WSI](#) is background, and restricting computations only on the tissue area saves both time and resources. However, the original slides in the NeSTBG dataset in-

cluded many types of artifacts, for instance different appearances of the background surrounding the tissue; *WSIs* presented a wide range of shades, from pure white to greys with different levels of detail.

To address the above issues, a sequence of filters was applied to mask out low frequency areas, and morphological operations were used to refine the result. The extraction scheme was designed by overlaying a grid on the tissue area detected on each slide, where each cell of the grid represented a tile. A random number of tiles ranging from 20 to 175 were extracted with random uniform probability, in order to have a representative sample of tissue per slide. The pre-processing steps have been performed with the *histolab* library (See Chapter 6). An example of the tile extraction procedure is shown in Figure 38.

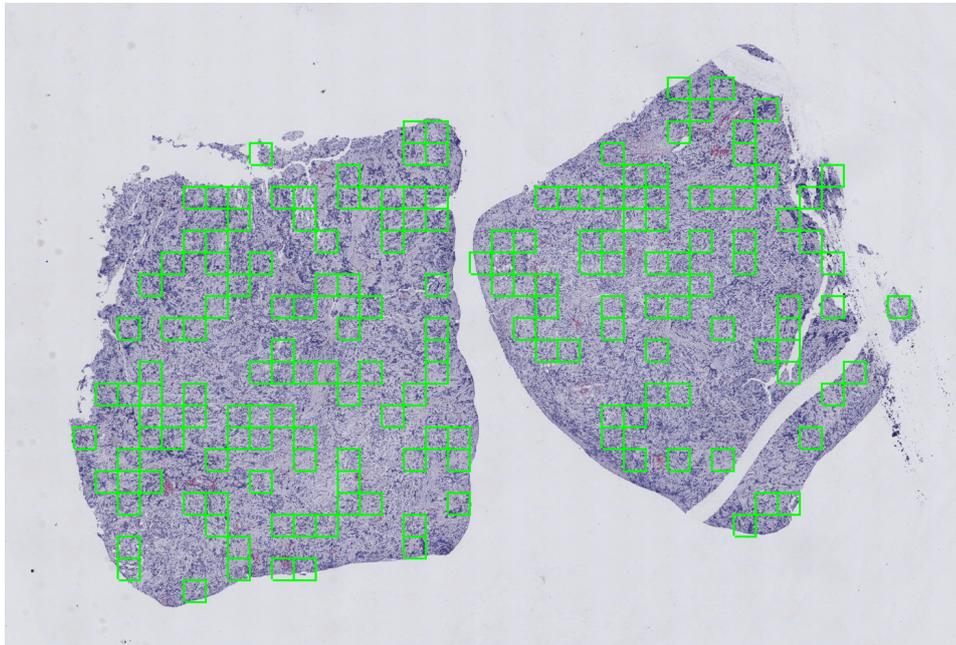


Figure 38: Visualization of the random extraction pattern performed with *histolab*. A CD3+ stained *WSI* used for the NeSTBG dataset is portrayed (at magnification 1.25x). The size of each tile is representative of the real portion of tissue captured with a 512×512 tile at 20x magnification.

The point-wise annotations for the centers of the lymphocytes were manually performed using the *VIA* annotation tool (version 2) [131] by four trained annotators, generating 73571 annotations for 3782 tiles extracted from a total of 54 *WSIs*. Examples of annotations are reported in Figure 39.

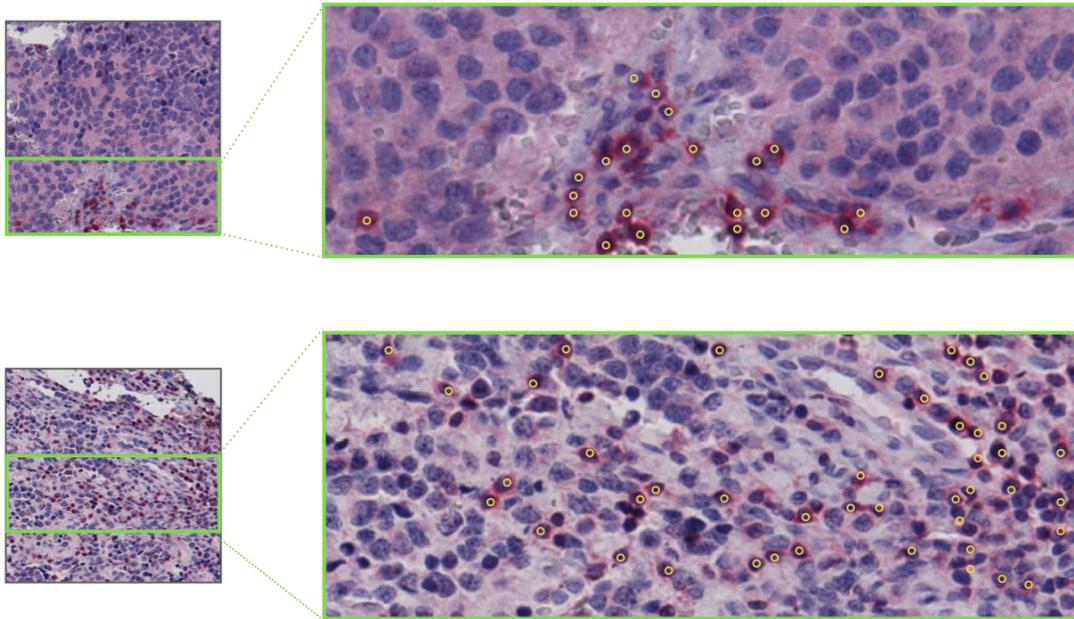


Figure 39: Example tiles from NeSTBG with corresponding manual point-wise annotations for the centers of the lymphocytes by the VIA software.

Given the non negligible irregularity in the shapes of lymphocytes, the staining variability, and the presence of really packed clusters of T-cells, a relaxed constraint for the annotations was chosen, following the strategy introduced in [265] for object counting in crowded scenes; the authors defined a density map of the objects in a crowded scene by centering at each annotated point a Gaussian curve, and normalizing such that the integral over the whole scene would result in the number of objects. When used for lymphocyte detection, the density-map approach associates to the annotated centers the highest confidence of *objectness* [8], a measure that decreases with radial distance from the center.

The point-like annotations were used to build targets to train the deep learning model to reproduce the density maps instead of bounding boxes typically used in an object detection task. This approach allows the model to encode the confidence of the annotation during the training phase, and also to leverage the surrounding context for the prediction. To define a density map, let T be an RGB tile of shape $(N \times N \times 3)$, and A its set of annotations

$$A = \{c_k = (x_k, y_k) | k \in [0, n], x_k \in [0, N - 1], y_k \in [0, N - 1] \text{ for } n \in \mathbb{N}, n \leq \infty\}$$

The density map is then computed as:

1. Assign a value d to each annotated pixel and define \hat{M} as:

$$\hat{M}(i, j) = \begin{cases} d & \text{if } (j, i) = c_k \text{ for } k \in [0, n] \\ 0 & \text{otherwise} \end{cases}$$

2. Define a Gaussian kernel

$$G(x, y; \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

and a squared structuring element GK, with side length $l \ll N$ and values given by G centered on the midpoint of GK;

3. Convolve \hat{M} with GK to obtain the target density map $M = \hat{M} * GK$.

EUNet architecture

EUNet, the chosen architecture for the predictive model, is based on the fully convolutional U-Net [394] in its encoder-decoder version. The aim of the encoder is to extract feature maps at different depth; the corresponding decoder blocks will up-sample feature maps from preceding layers and use feature maps of the encoder to refine the spatial information. Specifically, for each layer of the decoder:

1. The feature map from the preceding layer is up-sampled with standard up-sampling operations, without any trainable parameter.
2. The up-sampled feature map is concatenated with the feature map from the symmetric level of the encoder path on the depth dimension (*i.e.*, adding more feature channels).
3. The concatenated feature map is fed to convolution operations to refine the spatial information and reduce the number of feature channels.

In this work, we leveraged the PyTorch [354] U-Net implementation provided by Yakubovskiy in [512], which includes several encoder architectures and provides pre-trained ImageNet weights [121, 251, 252].

EfficientNet [449] (b3 version) was chosen as encoder; moreover, the spatial and channel Squeeze & Excitation blocks (scSE) [397] were also introduced in the decoder to improve model performance (see Appendix A.3). The proposed framework is illustrated in Figure 40 and includes:

- encoder and decoder each composed of five blocks;
- scSE blocks at the end of each decoder block;
- Decoder blocks with output feature channels of size: 256, 128, 64, 32, 16;
- Identity function as activation map in the output layer.

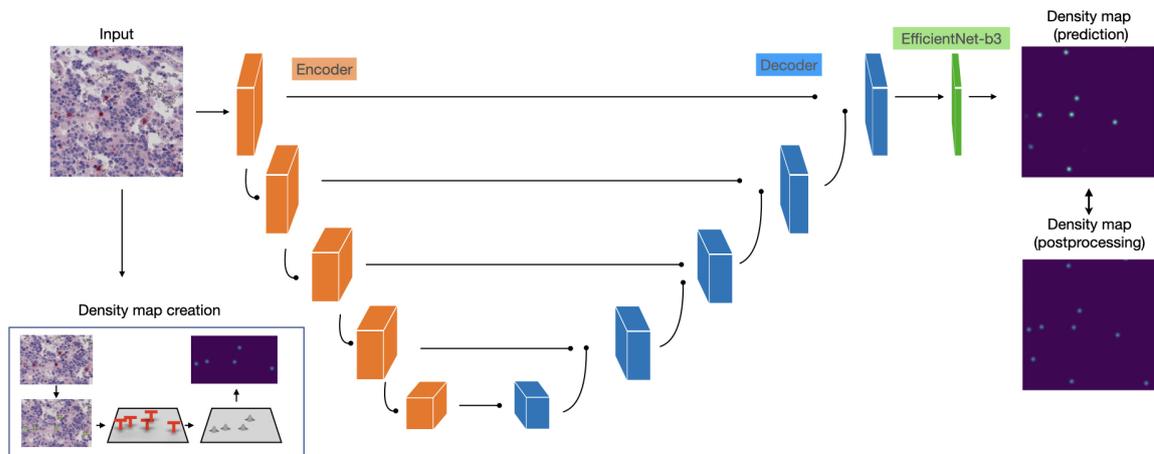


Figure 40: The full EUNet analysis pipeline.

EUNet training and evaluation

The lymphocyte counting task was censored as a classification task, by manually defining classes of lymphocyte density. The density classes used can be represented by the set $C = \{0, 1, 2, 3, 4, 5, 6\}$, as shown in Table 26.

Table 26: Lymphocyte count binning in ordinal classes.

Class	0	1	2	3	4	5	6
No. of Lymphocytes	0	1-5	6-10	11-20	21-50	51-200	> 200

Let \mathcal{D} be a dataset represented by a collection of n tiles: then $\gamma \in \mathbb{N}^n$ is the vector of ground truth class for the target lymphocytes in each tile and $\hat{\gamma} \in \mathbb{R}^n$ is the vector containing class predictions. As model performance metrics we used the Mean Absolute Error (MAE), the Mean Squared Error (MSE), the ACC, the Cohen's Kappa and the MCC. MAE and MSE are the L_1 and L_2 averaged difference between predicted counts and ground truth counts, respectively, while ACC is the averaged matching between the predicted class and the ground truth class.

$$\text{MAE} = \frac{\sum_{i=1}^n |\gamma_i - \hat{\gamma}_i|}{n}$$

$$\text{MSE} = \frac{\sum_{i=1}^n (\gamma_i - \hat{\gamma}_i)^2}{n}$$

$$\text{ACC} = \frac{\sum_{i=1}^n \delta_{\gamma_i \hat{\gamma}_i}}{n}$$

where δ is the Kronecker delta $\delta_{\gamma_i \hat{\gamma}_i} = \begin{cases} 1 & \text{if } \gamma_i = \hat{\gamma}_i \\ 0 & \text{otherwise.} \end{cases}$

The Cohen Kappa K [102] is a statistical measure used to evaluate agreement between two classifier, and it is defined as: $K = \frac{\text{ACC} - p_e}{1 - p_e}$, where p_e is the sum of the probabilities of the two classifiers agreeing on each class by chance. K takes values in $[-1, 1]$ where 1 means perfect agreement between classifiers and 0 or lower values mean that the two classifiers are agreeing by chance. In this work K has quadratic

weights for non agreeing values, thus attributing less importance to errors among nearby classes, in concordance with our classes having ordinal values.

The loss function used for training is the MSE between the ground truth and the predicted density map, computed pixel-wise. Since the L_2 metric penalizes large differences between pixels accordingly to their magnitude, the larger the values of the peaks in the constructed density maps, the higher the relevance: as a result, pixels in proximity of the lymphocyte centers (where the peaks are located) are more easily predicted than pixels of lymphocyte boundaries. Coupling the Gaussian kernel density maps with the MSE loss drives the network to focus on lymphocytes centers using the context in close proximity, but without great penalty for the exact margin reconstruction.

Hyperparameter optimization is done by the Ranger algorithm [508], combining the Lookahead procedure [524], and the Radam stabilization strategy [278]. The rectification strategy of [278] works by tuning the variance parameters of adaptive learning rate optimizers (e.g., Adam [543]) for the first iterations, until variance stabilizes with data from a sufficient number of iterations, thus avoiding the optimizer to remain stuck in local minima. The *Lookahead* strategy [524] improves parameter exploration speed and stability by using two sets of weights for the optimizer. One set of weights is used for fast exploration of the loss landscape, the other set of weights updates with smaller speed and serves as a stabilizing mechanism if the state of the the optimizer get stuck in unwanted local minima of the loss function. Overall, Ranger proved to be more robust and fast with respect to Adam, warranting a stable optimization providing a high optimal learning rate $\eta = 10^{-2}$, resulting in a faster training phase, especially for the ResNet50, whose training could be reduced from more than 300 epochs to about 80 epochs.

Networks were initialized by using pretrained weights from ImageNet [121, 251, 252]; alternative strategies such as using weights from fine-tuned ResNet50 pre-trained on the public DP dataset Lysto¹, did not lead to a significant performance improvement.

¹ <https://lysto.grand-challenge.org/>

To guarantee robustness and reproducibility to the modeling, a preliminary training/test split with ratio $\frac{3}{4} / \frac{1}{4}$ was operated and on the training set a 4×5 —cross validation resampling strategy was employed, following the guidelines recommended by the US-FDS in their MAQC/SEQC initiatives [179, 463]. Metrics are reported indicating average and standard deviation. Moreover, throughout the model training a particular care has been devoted into avoiding overfitting effects such as data (or information) leakage [62]: tiles extracted from the same WSI were not distributed in different training/test data subsets, a careful approach which is now becoming standard in the most recent works being published [146]. Finally, we adopted a plateau learning rate scheduler by monitoring metrics on validation set and reducing the learning rate if no improvements occurred for at least ten epochs: the new learning rate was computed as $\eta_{t+1} = \alpha\eta_t$ with $\alpha = 0.2$.

Lymphocytes spatial identification

The predicted lymphocytes density map is post-processed through a 3-step pipeline in order to refine the coordinates of the lymphocytes' centers:

1. First, the predicted density map values are corrected by setting to zero all pixels with negative values. Indeed, the model learns to predict near-zero values for pixels not belonging to lymphocytes, but the prediction may tend to zero in both positive and negative direction, and for the prediction to be a valid density map the negative values should be removed.
2. Secondly, Otsu thresholding algorithm [348] is used to find an optimal value to discretize the density maps in two levels: lymphocytes and background. The Otsu algorithm is the de-facto standard for discriminating foreground and background pixels within an image. In detail, the optimal threshold is identified by minimizing intra-class intensity variance (equivalent to maximizing inter-class variance). Since the Otsu algorithm is the one-dimensional discrete analog of Fisher's Discriminant Analysis, this procedure coincides with globally optimizing k-means clustering on the intensity histogram. Pixels with values under the

threshold are assigned to the background, while pixels with values over the threshold are assigned to the lymphocyte class.

3. Thirdly, in crowded scenarios, the simple segmentation may still result in connected components including more than one pixel. To split connected components on the Otsu mask, the Watershed segmentation algorithm [392] is used to effectively separate a dense single connected component into multiple sub-components. The result of the Watershed technique is a matrix with n connected components with different labels.

Finally, in order to obtain the coordinates of the center, for each connected component in the mask, the coordinates of the center of mass are computed and used as a proxy for the coordinates of the predicted lymphocytes. The goodness of the detection is evaluated by the three metrics Precision, Recall and F1-score, using as input the two sets of points T and P , defined for each tile as:

$$T = \{t_k = (x_k, y_k) | k \in [0, n_1], x_k \in [0, N - 1], y_k \in [0, N - 1] \text{ for } n_1 \in \mathbb{N}, n_1 \leq \infty\}$$

$$P = \{p_k = (x_k, y_k) | k \in [0, n_2], x_k \in [0, N - 1], y_k \in [0, N - 1] \text{ for } n_2 \in \mathbb{N}, n_2 \leq \infty\},$$

corresponding to the ground truth and the predicted center's coordinates, respectively. The Hungarian algorithm [253] is used to find the best assignment between ground truth points and predicted lymphocyte centers. Since optimal assignment can fail if the matched points are too far away, each possible match is accepted only if the distance between points is lower than a given threshold t , with $\Theta(t) = s_l$ for $s_l \simeq 4\mu\text{m}$, *i.e.*, the average size of a lymphocyte [567], corresponding to 8 pixels. Accepted matches are labeled as True Positive, while unmatched ground truths are considered False Negatives and unmatched predictions False Positives.

Deep Features interpretation

To explore the hidden layers of the model, a subset S_N of 1080 tiles was extracted from the NeSTBG dataset, 20 tiles for each available WSI, and then analysed by three different methods, namely UMAP-HDBSCAN clustering (Chapter 2, Section 2.6.3.1), TDA

descriptors (Appendix A.2) and TwoNN dimensionality estimation (Appendix A.4). First, UMAP is used to project data into a low b -dimensional space with $b \in [2, 12]$, where the upper bound is recommended in [312] for later feeding the projection into the HDBSCAN algorithm without falling into a computationally intractable task. Indeed, feature maps from raw images or from hidden layers of the network can be up to 10^6 dimensions. For instance, in an intermediate step of the trained U-Net the feature map has height and width 128 and 64 feature channels, resulting in a flattened vector of length $128 \times 128 \times 64 = 1,048,576$ elements, for each tile. Estimating densities in 10^6 dimensions with approximately 10^3 data points would not be feasible without the UMAP dimensionality reduction step. Value of b needs to be a trade-off between computational constraints (lower b) and the effort of preserving most of the original structure of the dataset (higher b): for the current tasks, $b = 4$ was chosen. Furthermore, the UMAP minimum distance parameter was set to zero to let the embedding be free of arranging points close together; the number of neighbors parameter was set to 25 so that at each iteration UMAP is forced to compare tiles from more than one patient, since at most 20 tiles are extracted from the same WSI; finally, L2, L1 and cosine norm were used as the distance in the original feature space.

The obtained 4-dim projection was thus used as the input for HDBSCAN to extract the dense regions of the embeddings; the clustering was subsequently visualized using a different 2-dim UMAP projection for a qualitative analysis of its global structure.

Next, Betti curves are used to highlight the topological dynamics of the deep features and finally the estimate of the intrinsic dimensionality of S_N as a point cloud is provided by the TwoNN algorithm.

Ansuini et al [17] experimented standard Convolutional Neural Network architectures for classification tasks (VGG, AlexNet, ResNet) and observed a characteristic pattern of Intrinsic Dimensionality of the deep features along layers in a well trained model. However, EUNet is more complex, with connections across multiple layers and two main branches with inherently different behaviors (encoding and decoding).

7.3 Results

Quantification of the immune content

To quantify the immune content in NB in terms of lymphocyte detection, a suite of DL experiments were run on the NeSTBG dataset employing a U-Net network with an EfficientNet-b3 architecture as encoder (EUNet for short). The whole dataset was first partitioned into Training (TR) and Test (TS) subsets, with ratio $\frac{3}{4} - \frac{1}{4}$; on the TR portion, a 5-fold cross validation was run four times (TR-CV), and the model trained on the whole TR was then evaluated on the left-out TS. The outcome of the prediction on TS was finally postprocessed (TSp) to enhance the lymphocyte detection: for this model, Precision=0.73, Recall=0.85 and F1-score=0.75. The complete set of classification performance is summarized in Table 27.

Table 27: Classification results (in terms of mean and standard deviation *mean (st.dev)* of the performance metrics for repeated experiments) for different EUNet models: in cross-validation on the training set (TR-CV), in training on the whole TR and inference on the test set, before (TS) and after (TSp) postprocessing (see Sec. 7.2). MCC: Matthews Correlation Coefficient; K: Cohen Kappa; MAE: Mean Absolute Error; MSE: Mean Squared Error.

Subset	MCC	K	ACC	MAE	MSE
TR-CV	0.50(0.10)	0.87(0.04)	0.70(0.10)	7.0(5.0)	881(1560)
TS	0.55	0.85	0.69	3.4	47
TSp	0.59	0.84	0.71	3.1	30

The EUNet was later applied to the entire tissue area of the 54 NeSTBG WSIs to obtain a patient-wise estimate of T-cell density. The tiles already included in NeSTBG were discarded during the training phase to avoid data leakage effects. Note that, for each WSI, NeSTBG includes approximately $\frac{1}{100}$ of all possible tiles. In Figure 41 the process of density estimation is graphically summarised on two tiles, while in Figure 42 the effect of postprocessing on the same two tiles is shown.

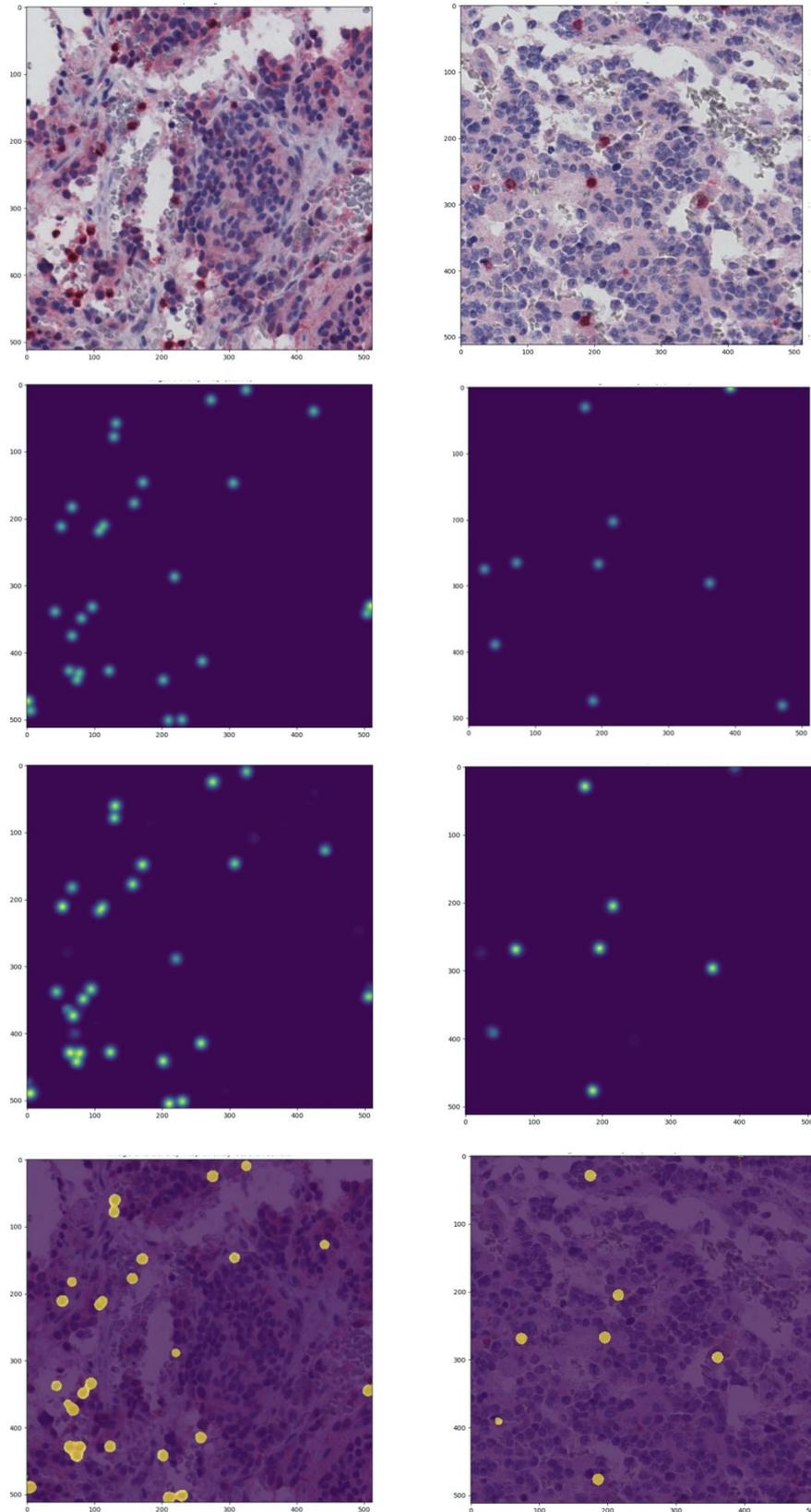


Figure 41: Examples of the density map prediction process on two tiles. First row: original images. Second row: ground truth density maps. Third row: predicted density maps. Fourth row: predicted density maps discretized with Otsu threshold overlaid on the original image.

To compute the density, the area (mm^2) of a single tile of size 512×512 pixels can be approximated as $A_{\text{tile}} = l^2 = 0.655 \text{ mm}^2$ where $l_{\text{tile}} = (512 \frac{\text{pixel}}{\text{mm}} \cdot \rho \cdot 10^{-3}) = 0.256 \text{ mm}$ is the tile side length and $\rho = 0.5 \frac{\mu\text{m}}{\text{pixel}}$ is the resolution ($20\times$) used for the tile extraction.

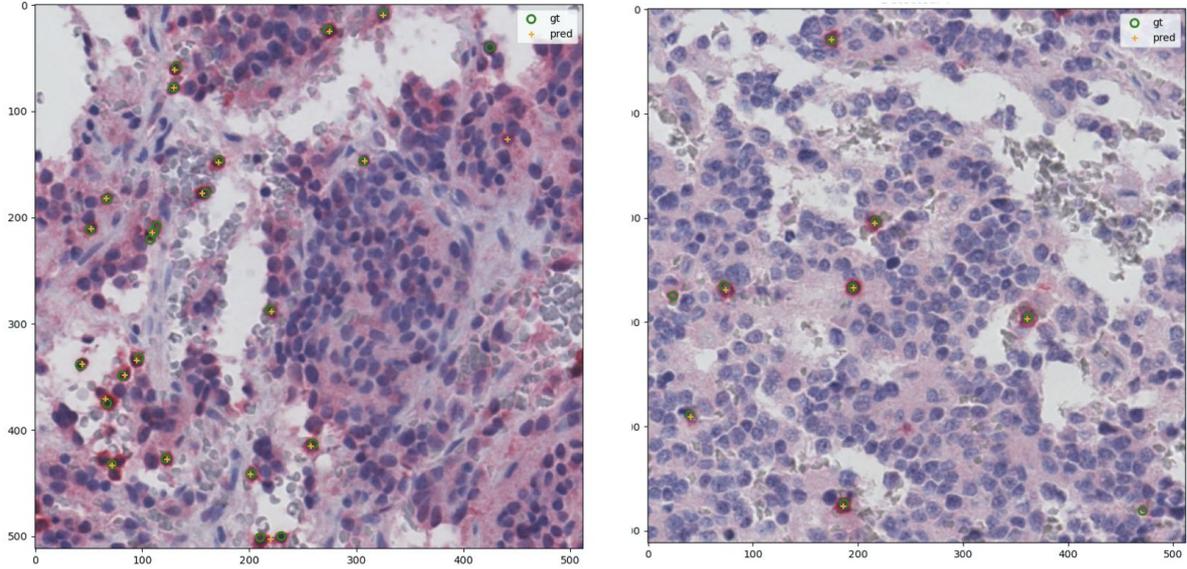


Figure 42: Examples of lymphocytes detection obtained through postprocessing of predicted density maps on the original tiles of Figure 41. gt: ground truth after postprocessing, pred: prediction.

As a benchmark, the DL estimate is compared with the manual estimate of a pathologist through the formula proposed in the reference work [320], expressing the density estimate L for each slide as the natural logarithm of the number of lymphocytes per mm^2 :

$$L = \log \left(\frac{1}{n} \sum_{i=0}^n \frac{c_i}{A_i} \right), \quad (5)$$

where n is the number of regions of interest (5 or 10), \log is the natural logarithm, c_i is the number of lymphocytes in the i -th selected region of interest, and A_i is the area of the i -th region of interest expressed in mm^2 . The two density estimates have a Pearson correlation coefficient of 0.47 with P-value $3 \cdot 10^{-4}$: in details, in Figure 43 the corresponding correlation plot is shown, together with the residual plot display-

ing the difference between DL predicted density value and pathologist estimation, indicating a positive offset.

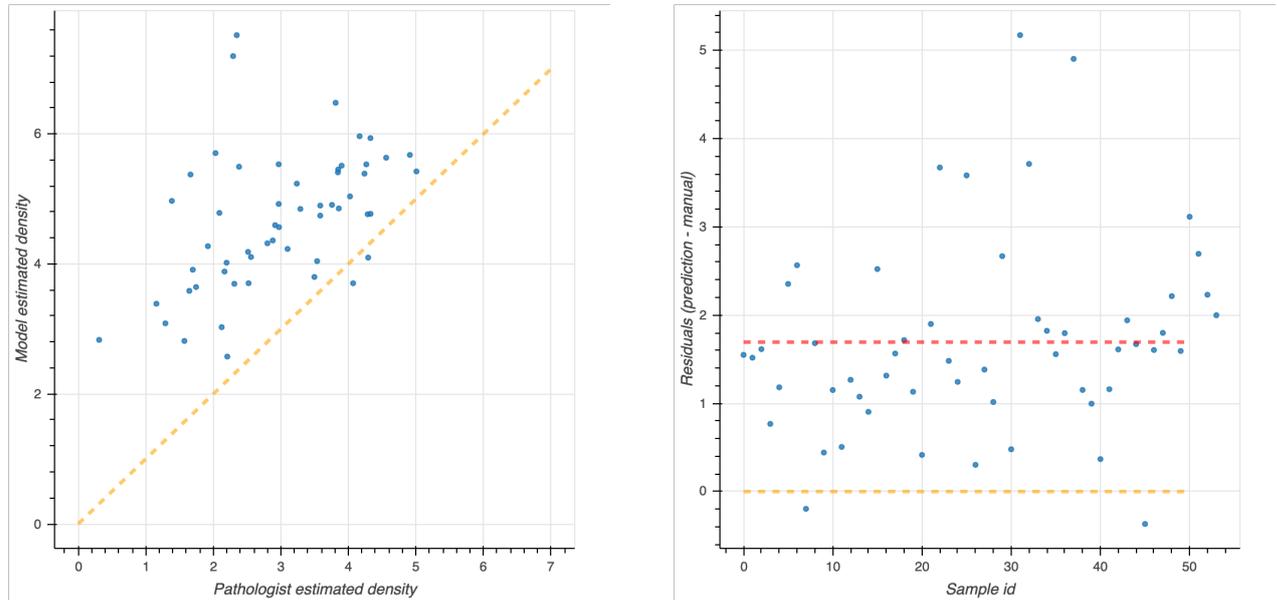


Figure 43: Comparison between DL and human expert density estimations computed for the 54 patients of the NeSTBG dataset. Notice that the pathologist estimation is computed on 10 regions of interest while the DL predicted densities are computed on the whole slide . Left panel: correlation plot for predicted densities and density estimated by pathologists. Right panel: residual plot for difference between DL and pathologist density estimation. Yellow line, both panels: perfect correlation. Red Line: average difference between DL and pathologist prediction.

Clinical assessment of the topological features

Clustering analysis was performed by HDBSCAN on the deep features projected by UMAP from the deepest (central) layers of the EUNet. Notably, these features are represented by vectors $v_i \in \mathbb{R}^D$, with dimension $D = 524,288$, as the output of the feature maps in the deepest layers has spatial dimensions 128×128 and 64 feature channels. The most interesting structure emerged in the second block of the EUNet decoder; Figure 44 shows the cluster assignment using cosine similarity as metric in the higher dimensional space, 15 neighbours and zero minimum distance for UMAP, minimum cluster size 5 and minimum number of samples 16 for HDBSCAN.

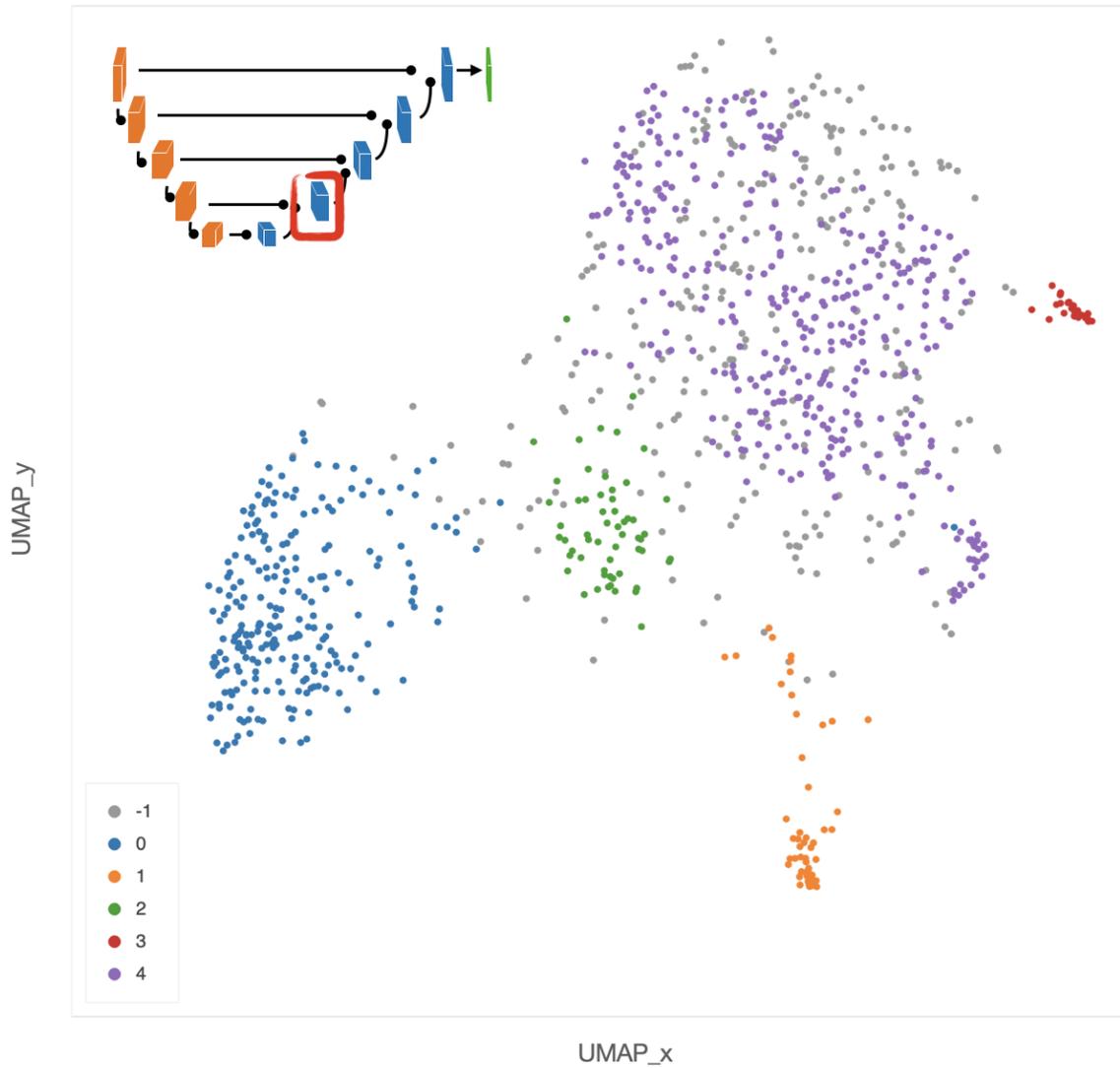


Figure 44: UMAP 2-dimensional embedding of a deep EUNet layer with cosine metric, and HDBSCAN clustering assignments. Gray points (label -1) are classified as noise by the clustering algorithm, while colored points belong to the clusters 0-4 detailed in Figs. 45-49: stroma rich areas with low TILS level (0), tissue with septa infiltration (1), tissue with pseudo-necrotic tissue infiltration (2), intermediate level of lymphocyte infiltration in stroma poor areas (3) and low level of infiltration in stroma poor areas (4). In the upper-left corner, the graphical schema of the corresponding layer in EUNet.

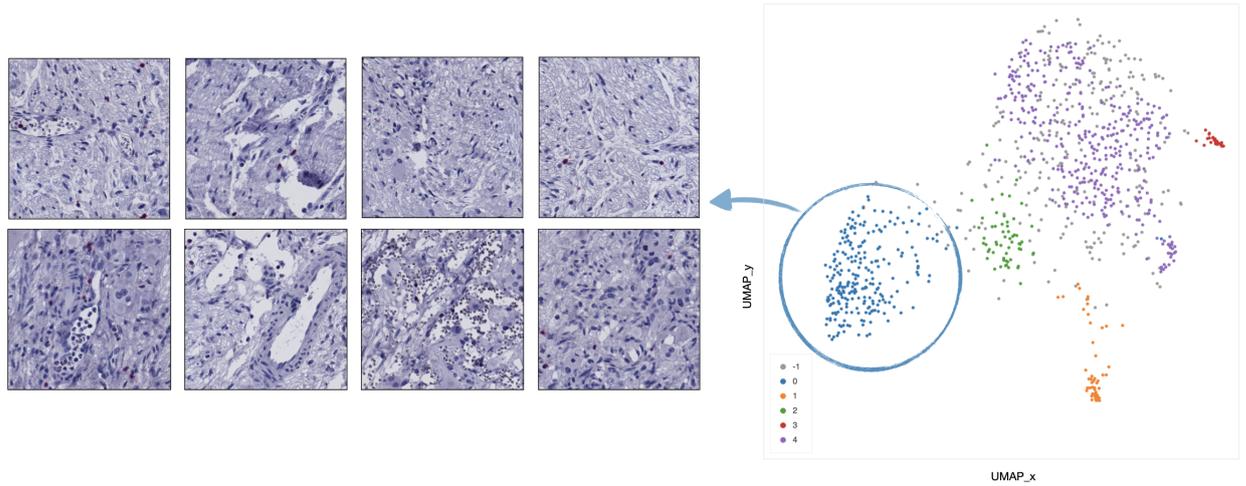


Figure 45: Example of 8 representative tiles with stroma rich areas with low level of TILs, grouped as cluster 0 by UMAP and HDBSCAN.

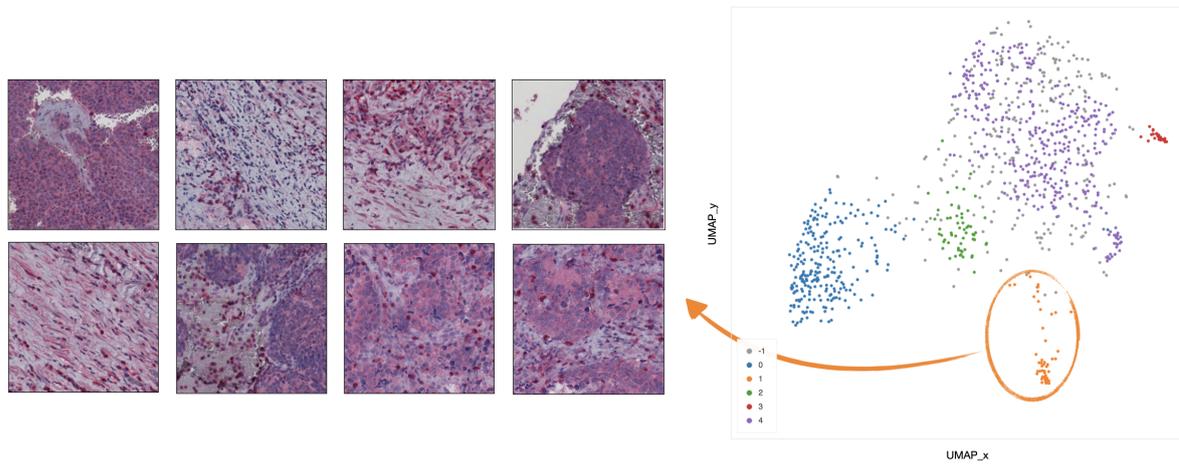


Figure 46: Example of 8 representative tiles with infiltration inside septa, grouped as cluster 1 by UMAP and HDBSCAN.

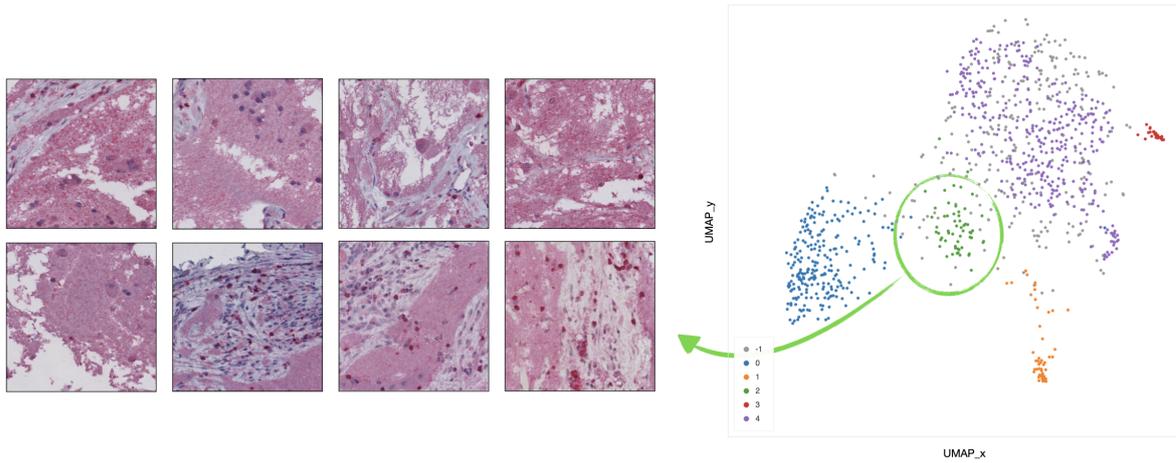


Figure 47: Example of 8 representative tiles with infiltration of lymphocytes in pseudonecrotic tissue, grouped as cluster 2 by UMAP and HDBSCAN.

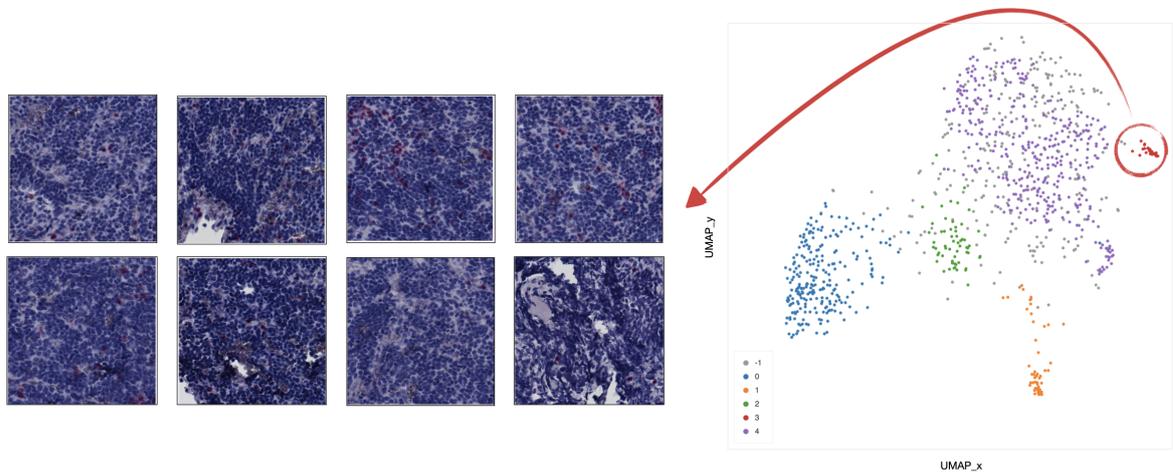


Figure 48: Example of 8 representative tiles with an intermediate level of lymphocyte infiltration in stroma-poor areas, grouped as cluster 3 by UMAP and HDBSCAN.

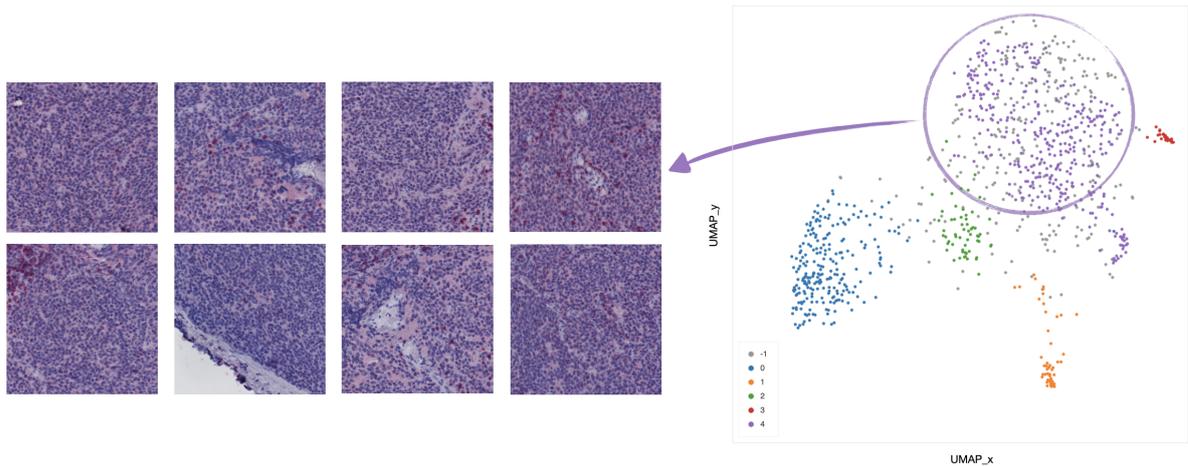


Figure 49: Example of 8 representative tiles with a low level of infiltration in stroma poor areas, grouped as cluster 4 by UMAP and HDBSCAN.

The tiles belonging to the 5 clusters identified by HDBSCAN can be clinically characterized according to their spatial arrangement. In particular,

- In cluster 0 (blue), the majority of tiles represents stroma rich areas with low level of TILs (Figure 45).
- In cluster 1 (orange), the majority of the tiles represents tissue with infiltration inside septa (Figure 46).
- In cluster 2 (green), the corresponding tiles present infiltration of lymphocytes in pseudo-necrotic tissue (Figure 47).
- In cluster 3 (red), the corresponding tiles show an intermediate level of lymphocyte infiltration in stroma poor areas (Figure 48).
- In cluster 4 (purple), the corresponding tiles display a low level of infiltration in stroma poor areas (Figure 49).

The cosine metric seems to be more effective in detecting sub-structures among samples described by DL features than alternative distances such as L_1 or L_2 , as shown in Figure 50.

Here sparsity plays a crucial role: data represent activation maps returned by a rectifier linear unit (ReLU) layer [332] inducing sparsity on the data. Indeed, the



Figure 50: Alternative UMAP 2-dimensional embedding on the features extracted from the second layer of the decoder path of the EUNet network, with L_1 (a) and L_2 (b) metric. Colors correspond to the detected clusters.

extracted feature vectors are quite sparse, with about 60% of the entries being zero, on average. Given the high-dimensionality and the sparsity, cosine similarity is more effective than L_p alternatives [441].

Nonetheless, an interesting pattern emerges also from the UMAP projection of the second layer of the decoder using the Euclidean distance, shown in the two panels of Figure 51. In the left panel, colors represent the INRGS stage of the NB patients, while in the right panel, NB patients are represented according to their MYCN amplification status. In the left scatterplot, high-risk NB patients from stage M are mostly localized on the left portion of the point cloud. Tiles from patients in the L_1 stage can be mainly found along the sides of the triangular shape and, finally, most of the tiles from MS patients (with metastases but with favorable prognosis) lie in the centers spreading to the upper and bottom-right vertex of the triangle. Notably, patients with MYCN amplification are clustered together in the upper-left area of the scatter cloud, similarly to high-risk NB patients.

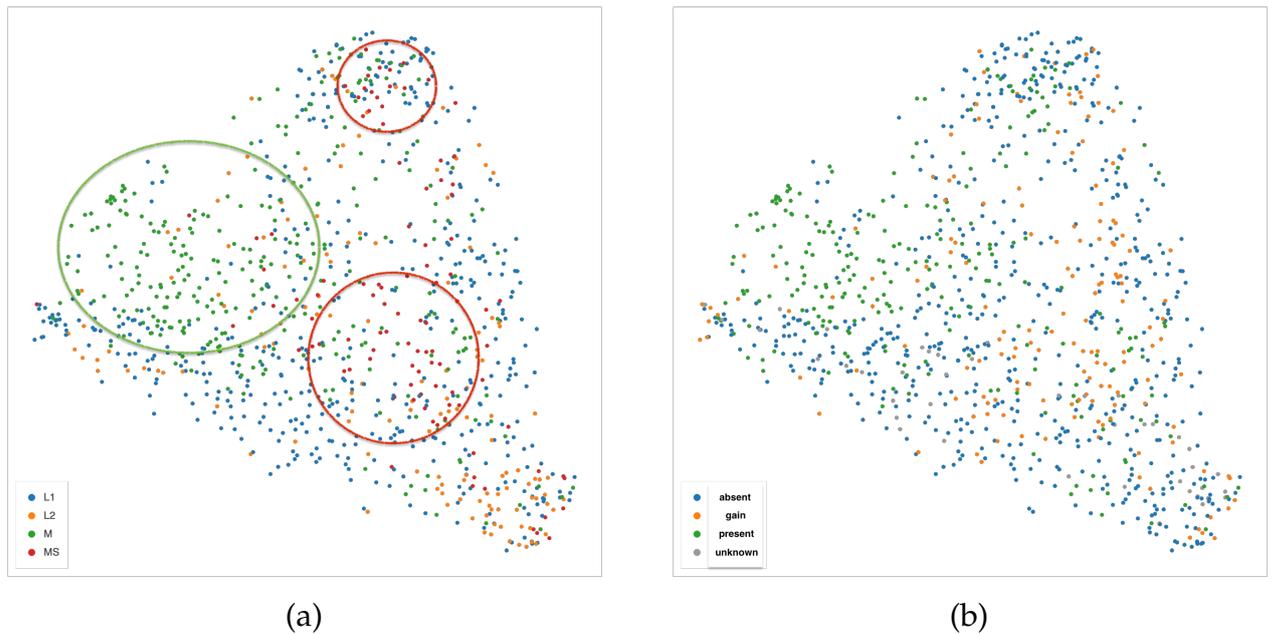


Figure 51: UMAP 2-dimensional embedding on the second layer of the decoder path of the EUNet, with Euclidean metric, minimum distance 0.02 and 15 neighbours. (a) Color indicates INRGSS. The red and green ovals mark the plot areas enclosing the majority of NB patients of stage MS and M, respectively. (b) Color indicates MYCN amplification.

Topological analysis of the deep features

We computed Persistence Diagrams (PD) to extract Betti curves (See Appendix A.2) from six selected EUNet blocks at different stages during model training. In Figure 52 Betti curves are shown for the 0-th homology group H_0 from the third decoder block at different epochs (left panel), with a focus on first three and last three epochs (right panel).

Notice that the Betti curves become smoother as the training proceeds, suggesting that the EUNet is progressively learning a meaningful representation of the data. At earlier training stages, several groups of connected components are merged together at uniformly-spaced thresholds; later in the process the curves decrease slower, implying that, from a set of points lying at uniform distances, there are larger groups at non-uniform mutual distances. Finally, towards the end of the training, points become less and less uniformly distributed as indicated by the smoother profiles of the curves.

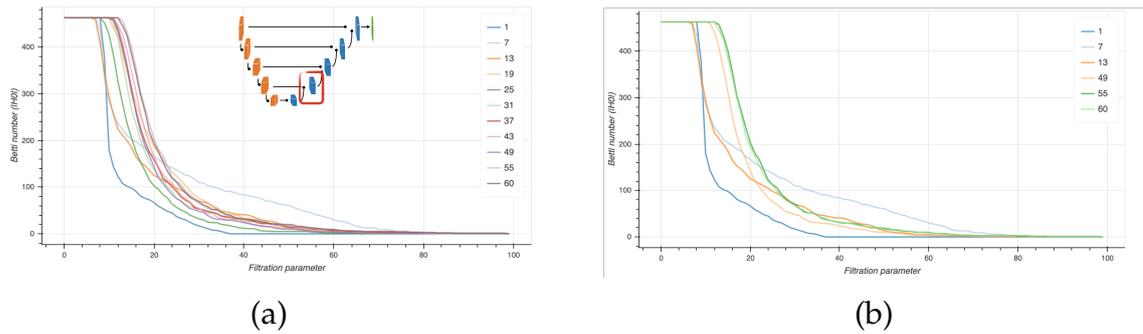


Figure 52: Betti curves for the 0-th homology group H_0 from the third decoder block (inset) at different epochs (a) and in particular for the first and last three epochs (b).

Intrinsic Dimensionality of datasets

The Intrinsic Dimensionality (ID) of NeSTBG is computed by the TwoNN algorithm (See Appendix A.4) in the six inner blocks of the EUNet (Figure 53) at different stages of the training process. Despite the high dimensionality of the deep feature space, the NeSTBG dataset possibly lives on a manifold of much lower dimension, similarly to the findings in [17]. Specifically, we computed the activation map from the EUNet model state every six epochs, and we estimated the dataset ID. Notably, $ID= 125$ for the original dataset (computed on 20 random tiles extracted from each patient), while $ID= 26$ for the predicted density map.

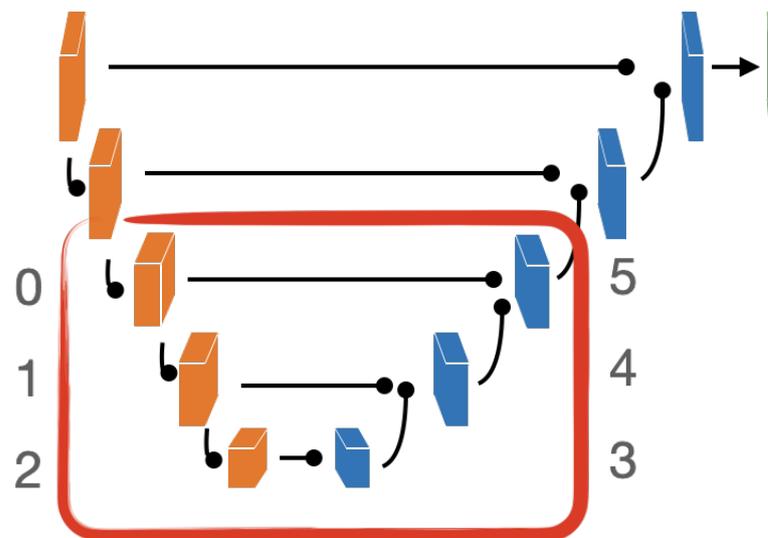


Figure 53: Graphical representation of the EUNet architecture. In the red encircling, the 6 inner blocks computing the Intrinsic Dimensionality at different stages of the training process using the TwoNN algorithm.

Detailed dynamics of the ID estimates are reported in Figure 54. In the top panel, the ID is plotted for each inner block for all the training epochs. In panel b), ID is plotted for the first three epochs (1,7,13) and for the last epoch (60), which corresponds to the highest peak of the encoder. During the central epochs (panel c) ID values of the encoder are stable, while the ID values of the decoder still show some variability; in particular (panel d) a ID peak on the third block. Thus, ID dynamics share a similar trend in both the encoding and the decoding phase, at different magnitudes.

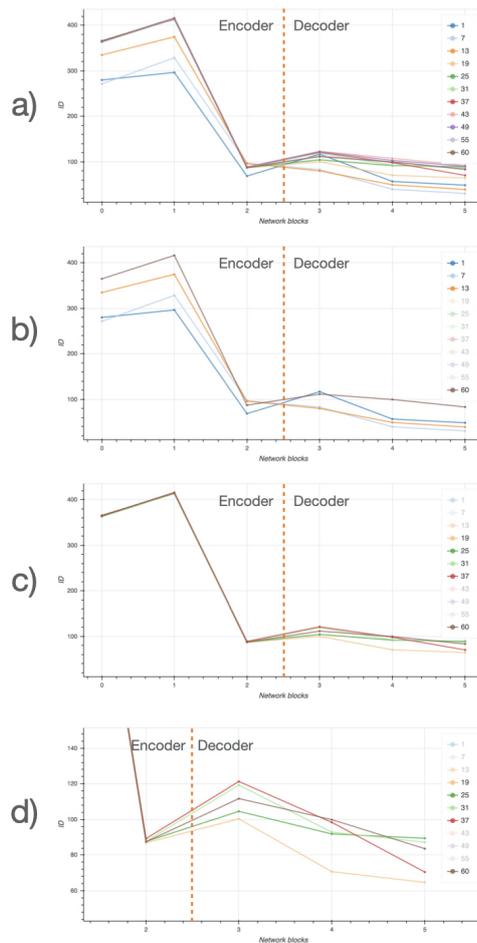


Figure 54: Intrinsic Dimensionality (ID) along different layers of the EUNet network, at different training stages: a) all training checkpoints, b) first three training checkpoints, c) intermediate training checkpoints, d) intermediate checkpoints, zoomed on the decoder. Legend includes all different epochs; ID curves corresponding to transparent elements of the legend are not shown.

7.4 Discussion

WSI data from DP are leveraged here to design a human-in-the-loop ML framework that could aid clinicians in NB risk assessment. As a major novelty in the pipeline, cloud computing is used to train a DL model with state-of-the-art architectures to predict density maps, an approach rarely found in DP for IHC-stained specimens. The predictive model is trained on the task of counting lymphocytes, while a post processing pipeline able to detect nuclei is implemented from the predicted density maps, with results aligned with the pathologist's estimates.

Furthermore, novel TDA approaches are employed to study the hidden representation of data as processed by the network. As future developments, different strategy for data augmentation (such as elastic transformations) or different techniques to construct the predicted target density maps can be explored, as well as possible optimization of the model architecture, and different activation and loss functions. Moreover, the current work focused on the CD3 T-cell marker as a proof of principle that can be extended to other immune cell markers to gain a deeper understanding of the role played by the immune system on NB progression.

Finally, the ML framework would strongly benefit from the ability to simultaneously recognise the tumour regions where lymphocytes are localized, *e.g.*, septa, or tumoral nests, and also to observe tiles within a larger portion of the slides, in order to gain a higher level of information.

Overall, the promising results emerging from the the current study pave the way towards the development of an effective learning tool aimed at timely and precisely quantifying the immune content within tumoral cells. Building on the awareness raising from the experience gained by previously published works [315, 316, 320], such a tool can work as a precious support for the pathologist, with an effective impact on the daily routine in a clinical setting.

As a future development, we plan to complement the current methodological work by deepening the reported analysis through the study of the contribution of additional markers such as PD-1 and PDL-1, investigating their correlation level with

both cell infiltration and patients outcome to strengthen the derived biological insights on NB.

Implementation and code availability

All the experiments were run on the Microsoft Azure cloud platform. The source code is written in Python/PyTorch as a deep learning framework, and it is available at the GitHub repository <https://bit.ly/3Gf4XwX>.

Funding

This work was supported by grants from the Associazione Italiana Ricerca sul Cancro (AIRC) IG 18495 and AIRC IG 24345.

Acknowledgments

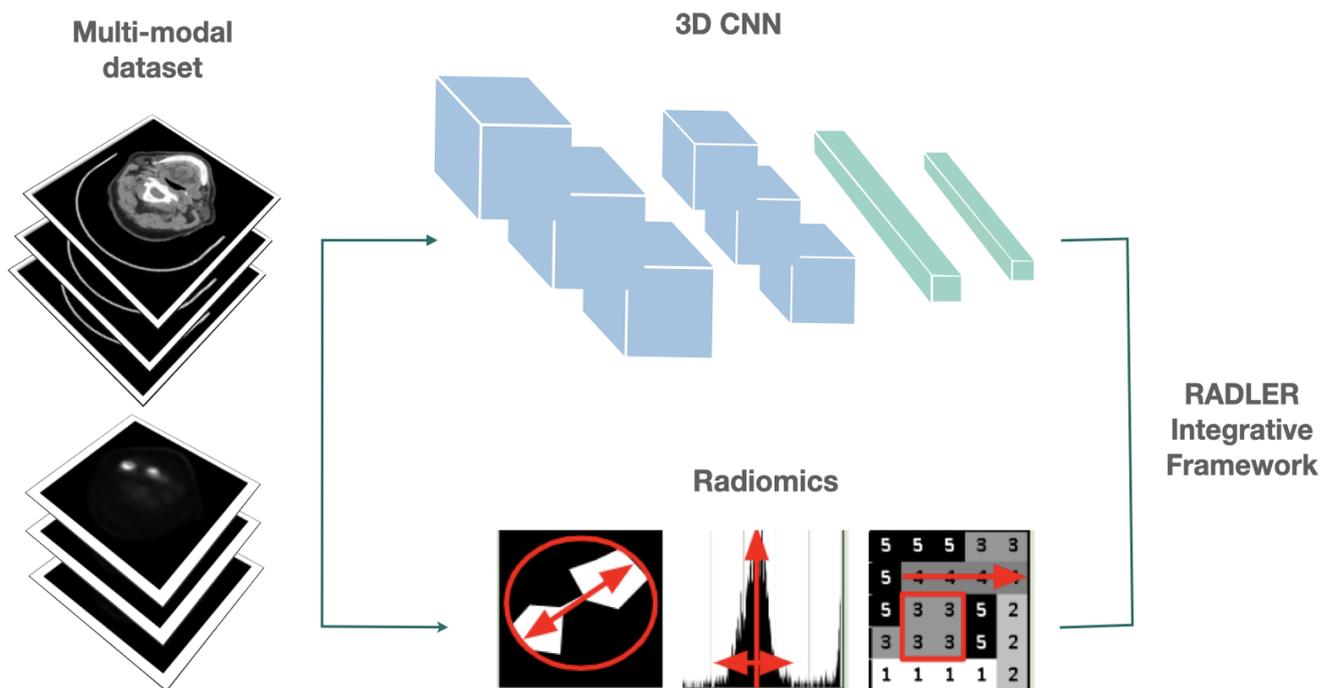
The authors are grateful to Alessia Marcolini for her inspiring passion and helpful advice and to A. Spilimbergo and C. Riccadi for their help with the tile annotation.

Part III

MULTI-MODAL INTEGRATION IN RADIOLOGY

Chapter 8 introduces the RADLER framework to Deep and Machine learning models simultaneously trained on multiple radiology scan collections. RADLER encodes functional and morphological information from PET and CT scans in an integrated reproducible framework, and further evaluates the contribution of deep and handcrafted features to extract predictive biomarkers in precision oncology.

THE RADLER FRAMEWORK



Authors: A. Bizzego, N. Bussola, D. Salvalai, M. Chierici, V. Maggio, G. Jurman, C. Furlanello. *Adapted from:* Integrating Deep and Radiomics features in cancer biomaging. *Published in:* CIBCB 2019 Conference on Computational Intelligence in Bioinformatics and Computational Biology. (Aug. 2019)

THE RADLER FRAMEWORK

Highlights

- The RADLER framework integrates hand-crafted and deep learning features for clinical endpoint prediction on multi-modal datasets in Radiology.
- To boost reproducibility, RADLER relies on a Data Analysis Plan for model evaluation.
- RADLER is evaluated on the prediction of loco-regional recurrence in HNSCC on PET-CT scans from the TCIA collection, improving on published results (MCC=0.926).
- Classification results on integrated and single feature sets indicate a higher accuracy of ML models combining deep and radiomics features in a unique predictive framework.

Personal contribution I contributed to the implementation of the RADLER framework and run part of the experiments on the TCIA data. I co-designed the transfer learning strategies and performed the UMAP analysis. I also contributed to the writing and prepared the figures for the article.

8.1 Abstract

In this study, we introduce a machine learning framework for medical imaging that combines the current pattern recognition approach (“radiomics”) with DL. We apply the framework in cancer bioimaging for prognosis of locoregional recurrence in head and neck squamous cell carcinoma (N=298) from PET and CT imaging. The DL architecture is composed of two parallel cascades of CNNs layers merging in a softmax classification layer. The network is first pretrained on head and neck tumor stage diagnosis, then fine-tuned on the prognostic task by internal transfer learning. In parallel, radiomics features (*e.g.*, shape of the tumor mass, texture and pixels inten-

sity statistics) are derived by predefined feature extractors on the PET-CT pairs. We first compare and then mix deep learning and radiomics features into a unifying classification pipeline (RADLER), where model selection and evaluation are based on a data analysis plan developed in the MAQC initiative for reproducible biomarkers. On the multimodal PET-CT cancer dataset, the mixed deep learning/radiomics approach is more accurate than using only one feature type, or image mode. Further, RADLER significantly improves over published results on the same data.

Summary

Radiomics is grounded on the underlying biological assumption that imaging features can capture distinct phenotype morphology, thus achieving both classification and clinical understanding in the machine learning process. On the other hand, DLR have a remarkably different role in the context of clinical diagnosis. However, DL models typically need a much larger amount of data for training for optimal results than statistical machine learning models; thus these models are often bootstrapped with the transfer learning approach, *i.e.* borrowing weights of models trained on different domains, and possibly retraining only a sector of interest of the network with the data from the novel task [447]. This trick is extensively used in non-medical domains, based on the availability of large-scale data and pretrained architectures [186, 355]. Recently, these resources are becoming available in cancer research. For example, the *DeepLesion* dataset, containing over 32,000 annotated lesions in CT scans [513], and TCIA, which provides medical images of different modalities (MRI, CT, etc.) [485].

The success of transfer learning schemas is clearly contributing to approaching DL models as powerful extractors of useful feature sets (*i.e.* *deep features*). However, linking deep features to meaningful clinical properties interpretable by physicians remains a key challenge [468]. Statistical machine learning approaches are also still widely used in radiomics [105, 537]. This state of the art has naturally led to the idea of an hybrid combination of HCR and DLR in an integrated system [48, 350, 565]. These systems can provide objective characterizations of tumor and a more effective

decision support environment, activating expertise in interpretation by clinicians, biologists and bioinformaticians [487].

In this work, we propose RADLER, an automatic pipeline for the integration of HCR and DLR features for medical images analysis, in a first application on multimodal PET-CT scans. Here we expand a preliminary version of this approach reported [46]. To support reproducibility, models are trained with a DAP that includes repeated cross-validation, model selection and feature ranking techniques. To validate the framework, an application is shown on a dataset of two-modality 3D PET-CT scans for prognosis of locoregional recurrence (LR) in HNSCC (N=298), previously solved with a HCR approach and a logistic regression model [479]. The multimodal network architecture is derived from a multi-stream multi-scale architecture for lung cancer screening [98]. The network is first pretrained on head and neck tumor stage (T-stage) diagnosis, then fine-tuned on the prognostic task (internal transfer learning). The RADLER model integrates in this case up to four feature types (CT-HCR, CT-DLR, PET-HCR, PET-DLR) improving over the published results on the same data [479]. Moreover the mixed deep learning/radiomics approach is more accurate than using only one feature type, or image mode.

Related works

Deep learning based solutions have demonstrated superior quantitative performance for OARs (Organs at Risk) contouring in CT scan images, using both 2D slices [174, 213] and 3D volumes [156, 450, 484, 516]. Approaches for head-and-neck cancer PET-CT images have used either 2D slices or the whole volumetric representation; although 3D networks are computationally more expensive, volumetric images can indeed provide comprehensive information (e.g. spatial context) in any direction rather than just one 2D view [198]. Zhao *et al.* [531] introduced a 3D segmentation method based on a multi-task training module and validated on a PET-CT dataset of 84 patients with lung cancer. Diamant *et al.* [125] developed a convolutional-based framework to predict HNSCC cancer treatment outcomes based on a patient's pre-treatment

CT image. Kann *et al.* [233] implemented a 3D CNN to identify nodal metastasis and extranodal extension on segmented lymphonodes.

Regarding radiotherapy treatment, Daoud *et al.* [113] have implemented a CNN to predict the optimal radiation dose distribution from contoured CT images of patients with nasopharyngeal carcinoma. On the same task, Neph *et al.* [339], adopted a U-Net based architecture on head and neck CT volumes.

Deep learning based reconstruction pipeline have been also implemented for low dose CT [32] to remove the noise and artefacts from the reconstructed images; for example, Sumida *et al.* [439] proposed a 2D convolutional neural network to perform a contrast reduction from contrast-enhanced CT images of head and neck cancer.

Though progress has been made in standardizing PET protocols in oncologic imaging [309], PET processing presents a lot of pitfalls and challenges mainly due to inter-institutional (or even intra-institutional) lack of standardized image acquisition.

Several works have attempted to automate the tumor segmentation process from PET-CT images of head and neck cancer through DL models. In particular, the Gross Tumor Volume (GTV) has been delineated with a CNN on 2D slices by Huang *et al.* [210]. Moreover, U-Net [326] and 3D DenseNet [156] have been leveraged for automatic segmentation of GTV.

The RADLER integrative framework

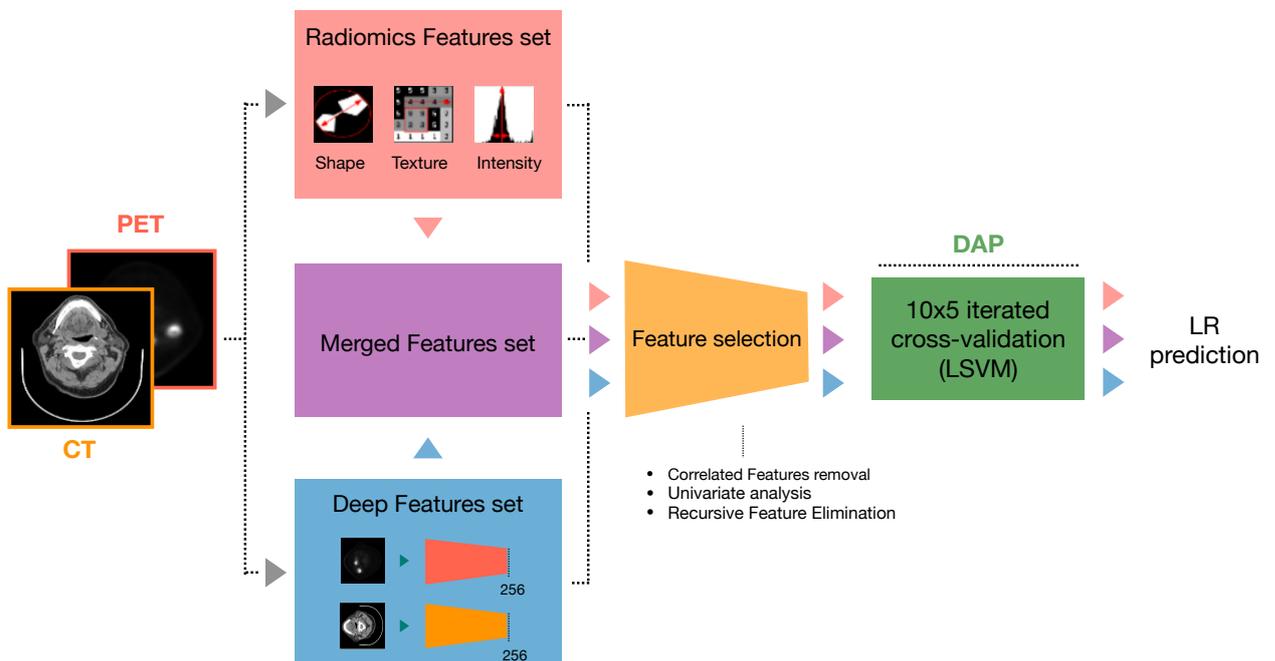
RADLER is a general framework for predictive medicine that can integrate Deep Learning and predefined features from multi-modal radiomics image datasets (see Figure 55). A study was designed to demonstrate the RADLER framework on the clinical task of predicting the recurrence of HNSCC. We can consider RADLER as a pipeline composed by four main stages:

- A) Processing of input images
- B) Feature extraction
- C) Feature selection

D) Classification within a Data Analysis Protocol framework.

Each stage of the RADLER pipeline is described in the corresponding subsection. An additional section (Section 8.2.0.4) describes the pre-trained network which operates on 3D multi-modal images.

Figure 55: The RADLER pipeline on PET-CT data: predictive models from the integration of radiomics and deep features.



8.2 Material and Methods

Medical images of 298 patients with head and neck squamous cell carcinoma have been downloaded from the Head-Neck-PET-CT (HN) dataset [125, 479], publicly available from TCIA¹.

¹ <https://bit.ly/3s1F8LH>

For each patient, the HN dataset provides **CT** and **PET** images and the Gross Tumor Volume (GTV) mask, together with several clinical variables, in particular: Locoregional Recurrence (LR) within the follow-up period (median: 43 months; range: 6-113 months) and the secondary diagnostic label tumour stage (T-stage).

Only 14.4% of the patients presented a LR (Table 28), which is the clinical endpoint of this study, while 3.3% of patients were assigned an undefined T-stage label (T_x).

Images were collected from four different hospitals. Notably, each hospital has its own equipment and image acquisition settings, which cause high heterogeneity in image characteristics, such as the resolution and scale of the **PET** images. In the approach by Vallieres and colleagues [479], images from two hospitals (cohorts: HGJ and CHUS) are used to train the predictive models, while the remaining images (cohorts: HMR and CHUM) are used to test the models. To enable the application of a **DL** framework to a small dataset and to improve model generalization, we adopted a random split 80/20 transversely to the four cohorts, stratified for endpoint classes.

The first stage of the RADLER pipeline aims to homogenize the image characteristics of the input data. The pixel values of **PET** images are converted to Standardized Uptake Values (SUVs), applying the protocol proposed by the Quantitative Imaging Biomarkers Alliance, which also considers vendor-dependent parameters. SUVs are then thresholded in the range between 0 and 50 to avoid outliers due to artifacts or errors in sensor readings. The intensity in **CT** images is associated with tissue density and it is measured with the Hounsfield Unit (HU); the pixel values are thresholded between $HU=-1050$ (air density score) and $HU=3050$ (bone density score). The GTV is converted to a 3D binary mask with the same size of **CT** and **PET** images. An isotropic voxel resampling (1 mm^3) based on cubic interpolation was performed on each of the 3 modalities (**CT**, **PET** and GTV binary mask).

In the last preprocessing step, a fixed-size sub-volume centered on the centroid of the GTV mask was extracted for each patient. This step is necessary due to the limited memory of GPUs, which prevents training with full-size 3D images. The drawback is a loss of information from the regions near the GTV. Thus, the size of the sub-volume was set to 128 mm^3 , which we deemed a reasonable trade-off

Table 28: Clinical data for original HN cohorts (N=298).

Cohort	Age (years)	Sex (% female)	no-LR	LR	total
HGJ	61 ± 10.89	17	79	12	91
CHUS	63 ± 10	27	86	15	101
HMR	67 ± 9.35	25	32	9	41
CHUM	63 ± 9.12	25	58	7	65

between the size of the GTVs in the dataset, the size of the GPU memory and the amount of context included.

Feature Extraction

Three different feature sets are considered in this study and are described in this subsection.

8.2.0.1 Hand-crafted Radiomics Features

A total of 3263 radiomics features are extracted for each patient, replicating [479]. Radiomics features are computed only on the voxels of the GTV and are chosen to describe three main image properties: shape (13 features, based on the GTV contours), intensity (18 features, based on the voxel intensities), and texture (1,600 features, based on four Gray-Level Matrices). Following [479], 40 types of texture features were considered, each one computed on 40 sets of parameters that define voxel spacing, quantization method and number of gray levels. As a result, since the same types of texture features are extracted several times with different sets of parameters, the radiomics feature set might include highly correlated features.

Table 29: HN Data: class distribution of the diagnostic T-stage tasks (4-class and binary) in train and test sets (N=284).

	T ₁	T ₂	T ₃	T ₄	T ₁ /T ₂	T ₃ /T ₄	total
Train	31	85	75	36	116	111	227
Test	8	21	19	9	29	28	57

8.2.0.2 *Deep Learning Radiomics Features*

A total number of 512 deep features are extracted (256 from PET images and 256 from CT images) from a pre-trained multimodal neural network. As in the case of generic 2D images in the recent 5 years, it is now common to pre-train networks on the largest and general image datasets and then apply some type of transfer learning procedure to obtain a performing model into novel medical domains. Often the transfer learning approach can boost the accuracy of Deep Learning models (e.g. see [45]). Unfortunately, although several solutions have been proposed and implemented [90], no research team has released a pre-trained network for multi-modal 3D PET-CT images yet. Thus, we first trained our own network to predict the T-stage (see Section 8.2.0.4). The input images are processed by the convolutional branches of the trained network to obtain the set of the DLR features.

8.2.0.3 *Merged Features*

HCR and DLR features are concatenated into an integrative dataset. A more accurate model is expected from two feature sets as they should capture different and complementary information from the input images.

8.2.0.4 *Feature selection*

The feature selection section in RADLER leverages a combination of three methods and is performed for each feature set. Table 30 summarizes the three feature sets and the results of the feature selection.

After mean imputation of missing values (*inf* and *Nan*), features are standardized to zero mean and unit variance. Overall, the procedure consists of three steps (see also [479]):

- Removal of correlated features: the Pearson's correlation matrix is computed, and in case of highly correlated features pairs ($\rho > 0.95$), one of the two features is removed from the feature set. Notably, no feature pair from the DLR feature sets showed high correlation and was removed at this stage.

- Univariate analysis: an association score (ANOVA F-test) is computed between each feature and the target. Features are ranked based on the association score, and the top 1,000 features are kept;
- Recursive Feature Elimination (RFE): the remaining features are ranked based on their predictive power and ordered by decreasing importance. RFE and alternative methods are discussed in [152, 175].

Table 30: Summary of the three feature sets considered for the LR prediction

Feature Set	# samples	# Initial features	# Features after selection
HCR	287	3263	692
DLR	284	512	512
HCR + DLR	284	1203	1000

Classification within DAP

A linear Support Vector Machine (LSVM) model is trained on the three feature sets within the DAP framework (see Chapter 2, Section 2.6.1.1). A grid search is performed to select the optimal value of the regularization parameter C with $C \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$ and of the number of features $n_f = n_{\text{step}} \times N$ with $n_{\text{step}} \in \{1\%, 5\%, 10\%, 25\%, 50\%, 75\%, 100\%\}$ and N the total number of features.

In our experiments, for each parameter point, the training set is randomly split into 5 folds: 4 folds are cyclically used to train the model which is evaluated on the left out fold. The procedure is repeated 10 times. The parameters that maximize the average performance are selected and used to set the final model which is trained on the whole training set and evaluated on the test set.

Training of the Deep Learning network

The first step of our transfer learning strategy aimed at training a multi-modal network to predict the T-stage, i.e. a descriptor of the size of the primary tumor [123]

that is considered significantly correlated with the Locoregional Recurrence [479]. The multi-modal network is designed following [98] and [46]. Briefly: a convolutional branch, composed of a stack of Convolutional blocks, operates on each modality independently to compute 256 convolutional features, which are then processed by a set of linear layers to output the T-stage classification (see Figure 56).

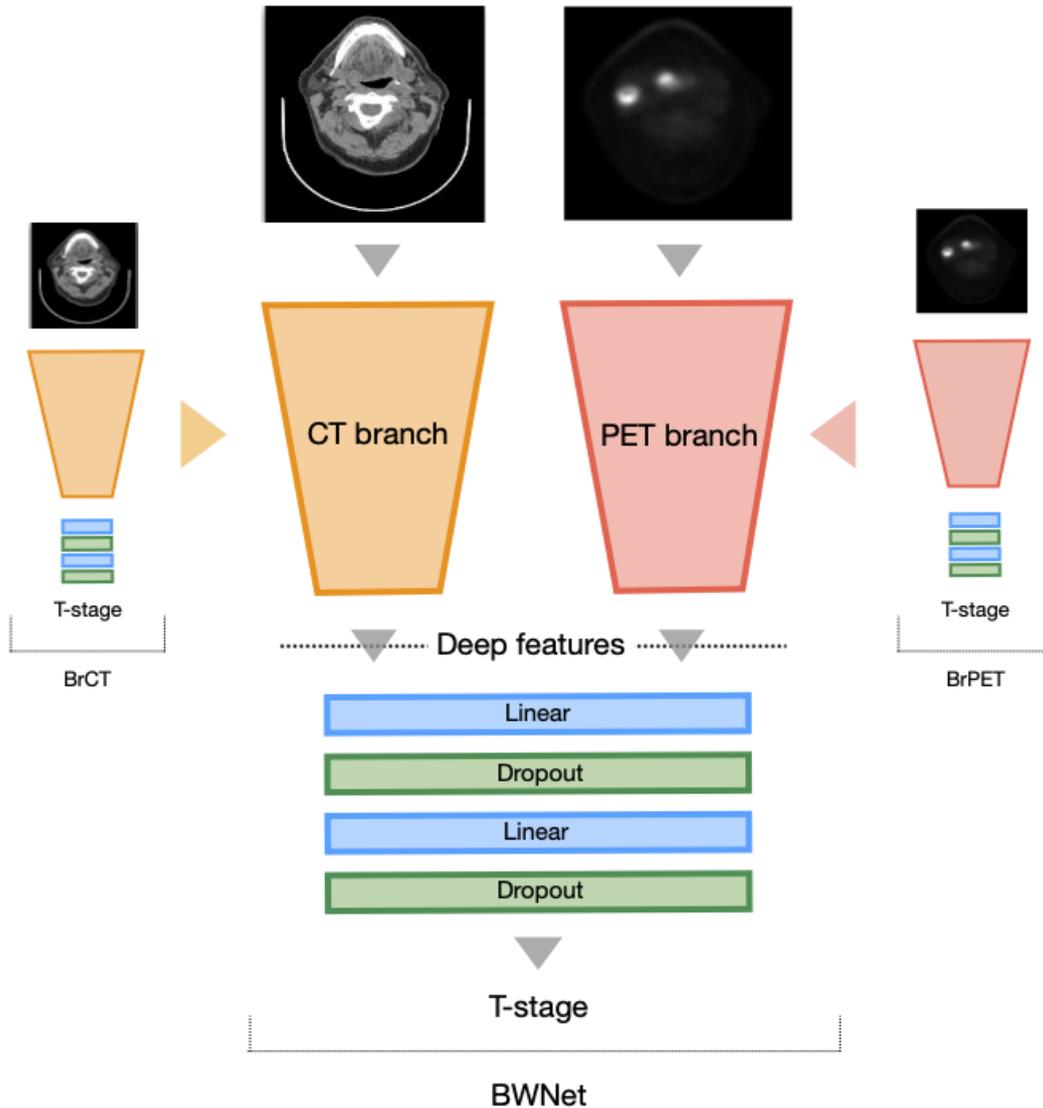


Figure 56: Multimodal network architecture for PET-CT scans. The network inputs are pairs of volumes of size $64 \times 64 \times 64$, one for each channel (CT and PET). The total number of output features from the convolutional branches is 512.

As the network on the 4-class problem (T_1, T_2, T_3, T_4) was prone to overfitting [46], in this study we also considered a binary classification task, by grouping T_1/T_2 and T_3/T_4 stages (Table 29). In addition, to leverage the single modality contribu-

tion, the convolutional branch of each modality (CT and PET) was first trained independently within two single-modality networks (BrCT and BrPET). The network trained on CT-volumes is more accurate than the BrPET network: $MCC_{tr} = 0.308$ and $MCC_{ts} = 0.385$ versus $MCC_{tr} = 0.240$ and $MCC_{ts} = 0.209$ respectively. This result was expected, as the PET modality produces qualitative images and the conversion to SUV is still an open clinical problem [41, 347].

The trained weights were thus transferred to the CT-branch and PET-branch of the multi-modal network (see Figure 56), respectively. This branch-wise network (BWNet) was fine-tuned for further 50 epochs. Notably, the strategy both reduced the overfitting and improved the performance on the T-stage prediction, when considering the 80/20 split. As expected, the original cohort-based split (*Vallières split*) produced an higher overfitting effect (see Table 31).

To improve the performance and reduce overfitting, we applied data augmentation procedures (i.e., minimal rotations and translations and addition of Gaussian noise) and dropout layers. All the training procedures adopted the same hyper-parameters setting: 80/20 train/test split with fixed seed or Vallières-like split, a batch size of 32 samples and the Adam [543] optimizer with a learning rate of 10^{-5} . The input images were resized to cubes of $64 \times 64 \times 64$ pixels to better fit the GPU memory size.

To qualitatively investigate the embedding resulting in the binary T-stage classification, we considered the UMAP dimensionality reduction method (see Chapter 2, Section 2.6.3.1). The UMAP projection of the deep features extracted from the combined PET-CT images with the pre-trained BWNet is displayed in Figure 57. Notably, the T_x -stage data (primary tumor cannot be assessed [123]), which were not considered during training, mostly clustered with the low-stage (T_1/T_2) data.

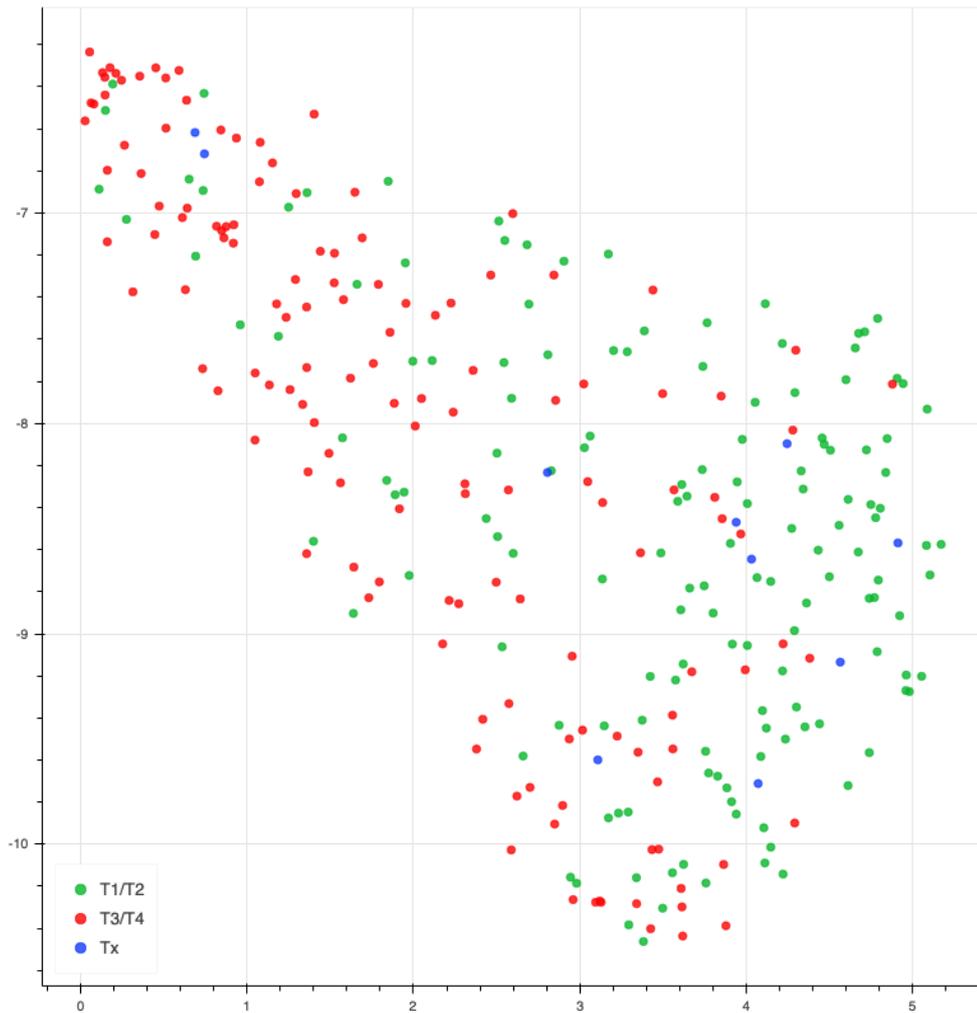


Figure 57: UMAP projection of the deep features extracted from the combined PET-CT images with the pre-trained BWNet.

8.3 Results

We compared the performance of the LSVM model trained within the [DAP](#) on the different feature sets (see [Table 33](#) and [Table 34](#)) and against the reference study [\[479\]](#) (see [Table 32](#)).

Note that the single modality ([CT](#) and [PET](#)) [DLR](#) feature sets are obtained from the corresponding single branches (BrCT and BrPET networks) trained on the T-stage classification task. In validation we found that:

Table 31: Transfer learning strategy: results on binary T-stage prediction. MCC is reported on the training and test set for both train/test split types.

Split	BrCT		BrPET		BWNet	
	train	test	train	test	train	test
80/20	0.308	0.385	0.240	0.209	0.563	0.432
50/50	0.362	0.076	0.424	0.258	0.643	0.233

Table 32: LR prediction scores for the different feature sets on the test set, compared to reference results ("VS PET-CT"). Deep features are extracted from the multi-modal network pretrained on the T-stage binary task.

	Sensitivity	Specificity	Accuracy	MCC
VS PET-CT	0.560	0.670	0.650	0.184
HCR	0.250	0.979	0.873	0.355
DLR	0.250	0.854	0.768	0.099
HCR+DLR	0.875	1.000	0.982	0.926

Table 33: LR prediction task on deep features from network trained on T-stage binary task. MCC: median value with 95% bootstrapped confidence intervals.

MCC	CT	PET	PET-CT
HCR	0.485 (0.433, 0.534)	0 (0, 0)	0.356 (0.300, 0.410)
DLR	0.177 (0.123, 0.229)	0.127 (0.073, 0.178)	0.095 (0.053, 0.139)
HCR+DLR	0.581 (0.532, 0.628)	0.091 (0.046, 0.139)	0.833 (0.807, 0.861)

- In single modality, both **HCR** and **DLR** have better performance on **CT** than on **PET** (as expected);
- In single modality, **DLR** has better performance on **PET** than **HCR**;
- Performance on **PET**-only are poor for all feature sets. Limits of **PET** scans for quantitative interpretation are a known open problem in the clinical practice, particularly for head and neck pathologies [203, 372]. In fact, despite the applied conversion to SUV, technical differences between **PET** scanners and

Table 34: LR prediction task on deep features from network trained on T-stage 4-class task. MCC: median value with 95% bootstrapped confidence intervals.

MCC	CT	PET	PET-CT
HCR	0.485 (0.433, 0.534)	0 (0, 0)	0.356 (0.300, 0.410)
DLR	0.124 (0.081, 0.175)	0.065 (0.017, 0.114)	0 (0, 0)
HCR+DLR	0.628 (0.590, 0.664)	0.112 (0.068, 0.156)	0.982 (0.969, 0.992)

non-linear effects associated to GTV segmentation, as well as other patient-dependent parameters are hard to compensate [41, 185, 347];

- The merged feature set (**HCR+DLR**) gives improves classification. The overall best performance is obtained for the merged feature set and the multi-modal dataset.

On test set, we improve the original results with the merged feature set (**HCR+DLR**).

8.4 Discussion

The RADLER framework can be applied for the integration of deep and radiomics features in medical image analysis and classification tasks. Its application to estimate prognostic locoregional recurrence (LR task) of head and neck cancer has helped design a model more accurate than the original work [479], both in terms of sensitivity (0.875 vs 0.56) and specificity (1.00 vs 0.67). In summary, the **DAP** analysis shows that the feature set that integrates radiomics and deep features is more effective in predicting LR with respect to single feature types.

Notably, the use of the MAQC **DAP** within RADLER helps to evaluate the variability injected by the choice of partitioning and control for possible bias during the model selection phase. As a sanity check for selection bias effects, the **DAP** is repeated stochastically scrambling the training set labels, which should result in

$MCC \sim 0$ (for both train and test). For example, this random labels experiment results in $MCC_{tr} = 0.02$ ($-0.012, 0.075$) for the HCR features set (50 replicates).

We have shown that the RADLER framework can be used with fairly complex DL architectures for multimodal imaging. In our PET-CT application, we adapted a 3D multimodal CNN from a 2D solution originally designed to classify lung nodules from CT imaging [98]. As is often the case in medical imaging, we had to apply an internal transfer learning approach, starting from the diagnostic classification of tumor stage. This domain adaptation approach is useful in dealing with class unbalance and a relatively low number of samples, while achieving good predictive performance, as shown on the HN dataset.

We were motivated by the need of integrating PET and CT images in a clinical context. Further work is needed to confirm the robustness of the approach on different cohorts and hospital systems. However, this design makes the RADLER pipeline and its DL network potentially effective to model other tasks of clinical interest when different image modalities (e.g., MRI) and anatomical regions (e.g., lung, brain) are considered.

This work aimed mainly at investigating the integration of radiomics and deep learning; limited effort was focused on tuning the DL model, and we restricted the types of radiomics features to those proposed in the reference paper [479]. We expect that accuracy can be improved by adopting specific DL architectures or considering more complex methods to extract radiomics features, for instance applying Wavelet filters [566]. A further option to complete the approach should include an automatic segmentation module [156, 531] to avoid the manual annotation of the GTV, moving towards a fully automated pipeline.

Implementation and code availability

The whole pipeline was developed in Python and is available at <https://bit.ly/35wfiIg>. Upstream image processing is based on SimpleITK; the extraction of the HCR features is based on pyRadiomics [565]. The Data Analysis Protocol was devel-

oped based on scikit-learn [357] and pytorch [353]. Deep Learning networks were trained on a NC24 Azure virtual machine featuring a 4xK80 Nvidia GPU.

Acknowledgments

The authors thank the WebValley2018 Students Team for initial development of the radiomics environment. The project has been motivated by a discussion on PET-CT modeling with M. Farsad and A. Fracchetti. Part of this work has been supported by the Microsoft Azure Research Award “Deep Learning for Precision Medicine”, assigned to Cesare Furlanello.

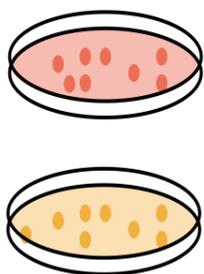
Part IV

PREDICTIVE MODELS ON OMICS FOR TOXICOLOGY

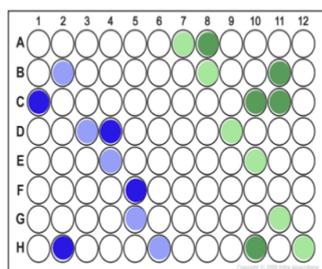
Chapters 9 and 10 explore predictive models on public omics data in pre-clinical research. Chapter 9 investigates Deep and Machine Learning models on gene expression data derived from cancer cell lines to assess hepatotoxicity of chemical compounds. In Chapter 10, shallow and deep learning techniques are implemented on molecular descriptors of endocrine-disrupting compounds within the ML4Tox framework, in order to extract predictive biomarkers for drug response assessment. All models are evaluated within a robust DAP to boost reproducibility.

DILI PREDICTION FROM OMICS DATA

Cell lines



GE data

Modeling (Data
Analysis Plan)

DILI

No DILI

Authors: M. Chierici, N. Bussola, M. Francescato, G. Jurman, C. Furlanello. *Original title:* Deep Learning for drug-induced liver injury prediction. *Published in:* Biology Direct (Feb. 2020)

DILI PREDICTION FROM OMICS DATA

Highlights

- ML and DL strategies are tested to predict DILI status from normalized gene expression (GE) data on two cancer cell lines provided within the CAMDA cMAP Challenge.
- All models are evaluated within a Data Analysis Plan to control for selection bias, using either all genes or a compact signature proposed in literature.
- Independently on cell line, normalization approach or predictive model consistently poor results are obtained.
- Lack of classification power might be caused by poor signal in the data or biological processes causing DILI that are not elicited in the studied cell lines.

Personal contribution I designed, implemented, and trained the DL models on the GE data. I also contributed to the writing and the figures for the article.

9.1 Abstract

DILI is a major concern in drug development, as hepatotoxicity may not be apparent at early stages but can lead to life threatening consequences. The ability to predict DILI from *in vitro* data would be a crucial advantage. In 2018, the Critical Assessment Massive Data Analysis group proposed the CMap Drug Safety challenge focusing on DILI prediction. The challenge data included Affymetrix GeneChip expression profiles for the two cancer cell lines MCF7 and PC3 treated with 276 drug compounds and empty vehicles. Binary DILI labeling and a recommended train/test split for the development of predictive classification approaches were also provided. We devised

three deep learning architectures for [DILI](#) prediction on the challenge data and compared them to random forest and multi-layer perceptron classifiers. On a subset of the data and for some of the models we additionally tested several strategies for balancing the two [DILI](#) classes and to identify alternative informative train/test splits. All the models were trained with the MAQC [DAP](#), *i.e.*, 10x5 cross-validation over the training set. In all the experiments, the classification performance in both cross-validation and external validation gave Matthews correlation coefficient ([MCC](#)) values below 0.2. We observed minimal differences between the two cell lines. Notably, deep learning approaches did not give an advantage on the classification performance.

We extensively tested multiple machine learning approaches for the [DILI](#) classification task obtaining poor to mediocre performance. The results suggest that the CMap expression data on the two cell lines MCF7 and PC3 are not sufficient for accurate [DILI](#) label prediction.

Summary

We developed three [DL](#) models to predict [DILI](#) on the challenge data and compared their accuracy with shallow machine learning models (SL), namely a [RF](#) classifier and a baseline [MLP](#). Models combining response to both drug and corresponding vehicles were investigated, as well as strategies for class balancing and identification of alternative informative TR/TS splits. [MCC](#) was used to assess the performance of our models, as it effectively conveys in a single number the confusion matrix of a classification task, thus making it possible to evaluate classifier performance even in presence of unbalanced classes.

9.2 Material and Methods

Data

The data provided by the CAMDA2018 organizers included microarray expression derived from two cell lines (MCF7 and PC3), either treated with one of 276 chemical compounds or dimethyl sulfoxide (DMSO) vehicle alone, part of the larger Connectivity Map build 02 resource [259]. A spreadsheet containing information to link compound filename identifiers to the corresponding vehicles, the DILI labels for the 276 compounds and the split into TR and test TS sets was also provided. To complement these information, we downloaded from the CMap project website a sample annotation file (Additional file 2) including information such as chip platform used for the assay, processing batch identifiers, compound CMap names, treatment duration and compound concentration during treatment. Experiments were performed in 96-well plates and a graphical representation of the experimental design is provided in Figure 58 along with the data pre-processing overview.

The original dataset provided by the organizers globally included 1095 CEL files (Table 35).

Table 35: CEL files available in the original CAMDA2018 Drug Safety challenge dataset.

Affymetrix chip	MCF7	PC3
HT_HG-U133A	588	475
HG-U133A	7	25

Two distinct Affymetrix chips were used for the expression data assays: HG-U133A and HT_HG-U133A. To avoid potential confounding effects in the analysis, since HG-U133A was used only for a handful of samples, these were removed from the list of input CEL files prior to normalization. Consequently, the starting dataset consisted of a total of 1057 samples, distributed across cell lines as shown in Table 36.

Table 36: Number of samples available after removing CEL files profiled with the HG-U133A chip. Sample numbers are reported according to three categories: samples treated with a compound assigned to the TR test, samples treated with a compound assigned to the TS set and samples treated with DMSO vehicle only.

category	MCF7	PC3
compound train	180	180
compound test	86	86
vehicle	316	209

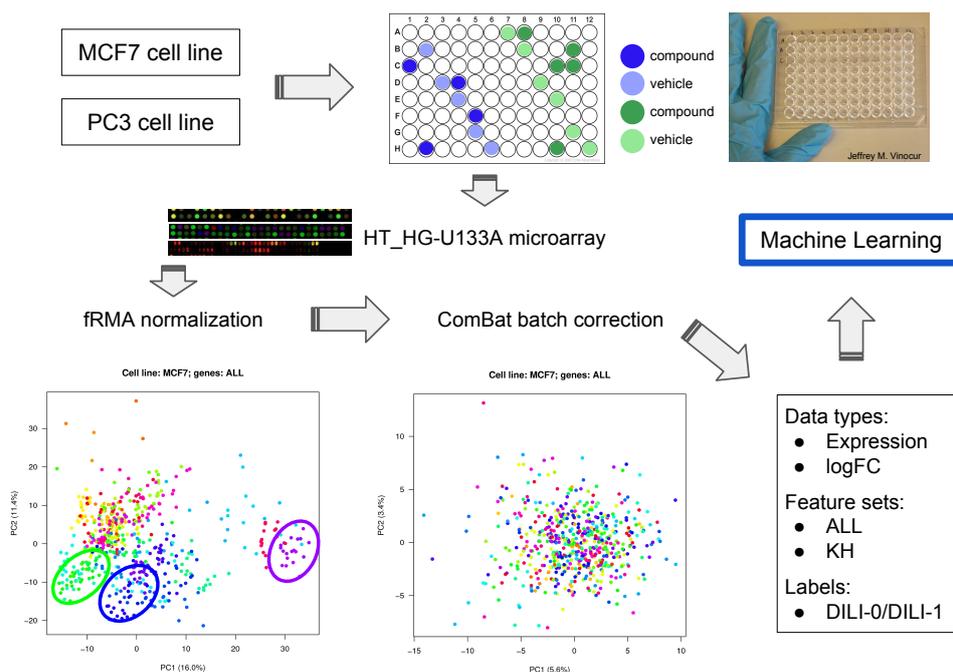


Figure 58: Experimental design scheme and batch correction. The figure represents schematically the data processing approach adopted in the article.

9.2.0.1 Microarray data preprocessing

The microarray data was normalized using the fRMA function of the Bioconductor package fRMA [308] with default parameters. Briefly, the function performs background correction according to the robust multi-array average algorithm, quantile normalization and robust weighted average summarization over probesets. Using the Bioconductor annotation package hgu133a.db [71], the expression data was further summarized considering the mean expression value for each gene and gene symbols were used as reference. Since a batch effect related to the actual microarray processing batches was observed, a batch correction adjustment was applied to the

normalized expression data using the ComBat function of the Bioconductor package *sva* [263]. The resulting normalized and batch adjusted data was used as input for the subsequent analyses, either directly in the form of compound expression or as the \log_2 -transformed fold change (logFC) between compound and vehicle treated samples. If a given compound was associated to multiple vehicles, their median expression value was considered in the calculation. All data were simultaneously normalized, neglecting the TR/TS partition due to their mutual heterogeneity. We note that part of the vehicles were shared between the TR and the TS set. We considered two feature sets. A first dataset included all the 12437 genes resulting from the processing of the microarray data (named ALL feature set). A second, more compact, consisted of 1234 genes (KH feature set) representing the intersection between ALL and the 1331 genes most associated to the predictive toxicogenomics space defined by Kohonen and colleagues in [244].

9.2.0.2 *Random splits*

We randomly split either the whole dataset or the original TR set into new TR/TS pairs, containing 75% and 25% of the data respectively with balanced classes, 100 times. Since previous experiments showed fundamentally homogeneous results across classifiers and feature sets, the “random split” experiments were performed using the RF classifier and the ALL feature set for both cell lines.

9.2.0.3 *Class balancing*

Since the TR and TS classes were unbalanced (including about two thirds vs. one third of the initial data respectively, Table 37) three oversampling strategies were considered for balancing, as follows:

- naïve random over-sampling, i.e. resampling either both classes (*all*) or the minority class only (*minority*);
- synthetic minority oversampling technique (SMOTE, [85]) and variants *borderline1*, *borderline2*, *svm* [180, 341];
- adaptive synthetic sampling approach for imbalanced learning (ADASYN, [190]).

Table 37: Number of samples belonging to DILI-0 and DILI-1 classes for TR and TS sets.

	DILI-1	DILI-0
TR	120	60
TS	67	19

Oversampling was performed using `imbalanced-learn` v0.3.3 Python package [264]. The experiments were performed on the cell line MCF7, on the feature set KH, using expression as input and either `RF` or `NMBDeep` as classifier.

Deep learning architectures

The DL models were trained following two distinct strategies dealing with vehicle expression differently, as sketched in Figure 59A. In the first strategy (“single”) each cell line was treated independently and either the logFC values or the expression of each compound were considered as input for the models, creating samples of size $(1 \times N)$, with $N = 12437$ (ALL) or $N = 1234$ (KH). In the second strategy (“end-to-end”), we considered the expression of each compound along with the median of the corresponding vehicles, creating homogeneous samples of size $(2 \times N)$ for each cell line, with $N = 12437$ (ALL) or $N = 1234$ (KH).

We designed three neural network architectures with increasing depths: `NBM1`, `NBM2`, and `NMBDeep` (Figure 59B). The `NBM1` architecture includes a first layer taking as input the whole set of 12437 (ALL) or 1234 (KH) features, concatenated according to the two strategies. This is followed by two fully connected layers with 1000K and 100K nodes (with $K = 2$ for ALL and $K = 1$ for KH) and by the output layer. `NBM2` was created doubling the 1000K and 100K inner layers of `NBM1`. `NMBDeep` is the deepest network, created further expanding the inner layers of `NBM2` as detailed in Figure 59B, obtaining a total of 12 hidden layers.

For each architecture the weights and biases of the fully connected layers were initialized before training with values drawn from the uniform distribution. The rectified linear unit (ReLU) functions [332] were used as activations for all the inner layers while SoftMax was used for the output layer. For the ReLU layers a batch

normalization with $\text{eps } 10^{-5}$ and momentum 0.1 was applied. The categorical cross-entropy was chosen as loss function, with weights proportional to the class sizes. To avoid overfitting, dropout layers were added with rate 0.5 after each of the inner layers. The networks were trained over 1000 (NBM₁, NBM₂) or 5000 (NBMD_{deep}) epochs, using minibatches of 60 samples.

9.2.0.4 *Parameter tuning*

The optimizer type and the learning rate (LR) of the networks were selected among the alternatives described below by training NBM₁ over 1000 epochs on 70% of the training set (randomly chosen) and evaluating the performance on the left-out 30% portion. With the stochastic gradient descent (SGD) optimizer, the net was trained with $\text{LR} \in [10^{-2}, 5 \times 10^{-3}, 2 \times 10^{-3}, 10^{-3}]$. Using Adam optimizer, the net was trained with $\text{LR} \in [10^{-7}, 10^{-6}, 5 \times 10^{-6}, 7 \times 10^{-6}, 8 \times 10^{-6}, 9 \times 10^{-6}, 10^{-5}, 10^{-4}, 5 \times 10^{-4}, 10^{-3}]$, as Adam requires smaller LR with respect to SGD [241]. We compared the training and validation performance and losses of the network using the two optimizers. As detailed in the Results sections, the performances were generally poor without strong dependence on the parameters. We decided to use Adam as optimizer with $\text{LR} = 1 \times 10^{-5}$ as it was giving slightly better performance (not shown).

Shallow machine learning

We considered a basic MLP and a RF as baseline machine learning strategies to compare our DL models to. MLP consisted of three fully connected hidden layers with 30 nodes each, and an input layer with 12437 or 1234 nodes for ALL and KH feature sets, respectively. All activations were ReLU functions [332], with neither dropout nor batch normalization. As optimizer we used Adam [241] with the number of iterations bounded at 200. RF was initialized with 500 trees and the Gini impurity as criterion to evaluate the quality of a split.

Predictive modeling strategy

All shallow and DL models (including class balancing experiments) were trained within the DAP (See Chapter 2, Section 2.6.1.1). Data were rescaled in the interval $[-1, 1]$ (for shallow learning) or centered and scaled to unit variance (for DL) before undergoing classification: rescaling parameters from TR were used for rescaling both TR and TS subsets, so to avoid information leakage. The DL models were run in the DAP without feature selection, which was enabled for MLP and RF.

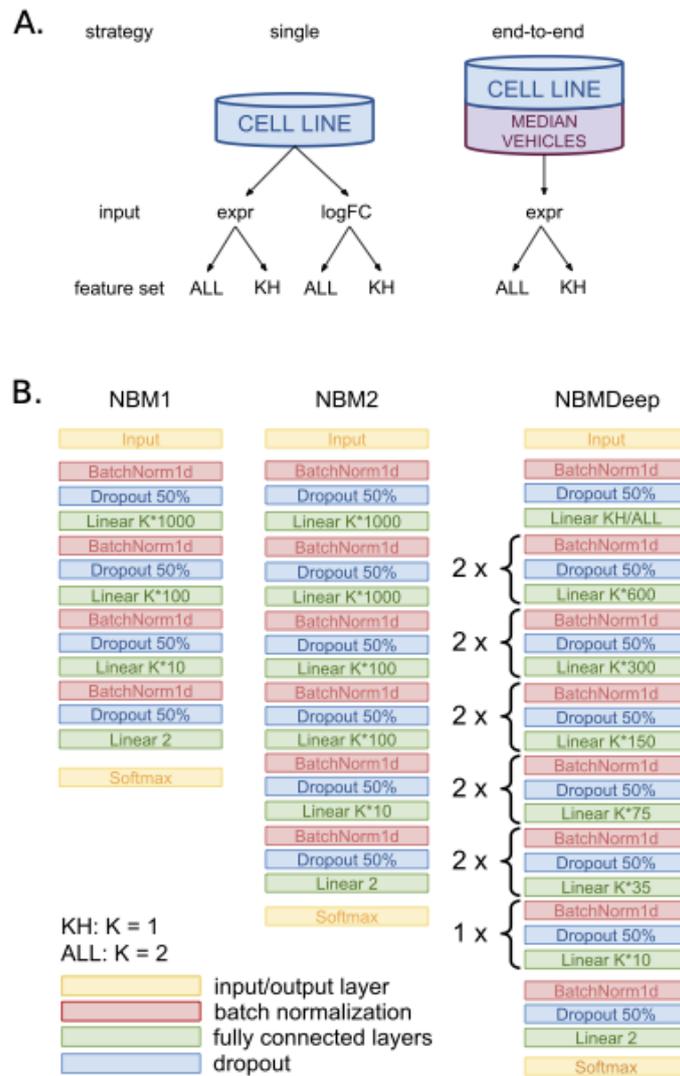
9.3 Results

Data production and processing layout are outlined in Figure 58. Briefly, the microarray data for compounds and vehicles was pre-processed, normalized and batch corrected following a standard procedure. Two distinct feature sets were extracted: ALL (including all 12437 genes with detectable expression) compared with KH (the 1234-gene PTGS signature proposed in [244]). All the models were trained on 187 drugs within a standard data analysis protocol (DAP) and validated on 79 different drugs, using as input either the compound expression values or the log-fold change (logFC) of compounds vs. vehicles. All processing steps are detailed in the Methods section. Considering our results globally, the general classification performance for the DILI status was poor. MCC values in CV ranged from -0.04 to 0.21 , while MCC in validation ranged from -0.16 to 0.11 (details below). These results are comparable with a random labels experiment on the same data. We did not identify a model that performs systematically better than the others, nor important differences in classification performance when considering separately the two cell lines, the different feature sets or the different input types. The results of all experiments performed are collected in the online manuscript.

Deep Learning

We devised three DL architectures of increasing depth, namely NBM₁, NBM₂, NBMDeep (Figure 59; see Methods for details), with 4, 6, and 13 hidden layers, respectively. All DL models operated in two modes: “single”, with the logFC values or the expression of each compound as inputs, or “end-to-end”, with the expression values of each compound concatenated with its corresponding vehicles as inputs.

Figure 59: DL strategies and architectures. A. Strategies used for the analysis. “single” indicates that the logFC values or the expression of each compound were considered as input for the models; “end-to-end” indicates that the expression values of each compound are considered along with its corresponding vehicles. B. Schematic representation of the DL architectures used.



Overall, the classification performance was poor independently of the architecture, the DL strategy, and the cell line (Figure 60 and Table 38). In particular, all DL models performed poorly on the two cell lines (median $MCC_{cv,MCF7} = MCC_{cv,PC3} = 0.02$; $MCC_{val,MCF7} = 0$, $MCC_{val,PC3} = -0.02$), using the two feature sets or input types. The MCC values of the DL “end-to-end” experiments were higher in CV than the “single” experiments (median $MCC_{cv,end-to-end} = 0.09$, $MCC_{cv,single} = 0.01$; Wilcoxon $p = 0.003$), but close to 0 in validation for both strategies. Notably, the NBMDeep architecture performed worse than NBM1 and NBM2, achieving median $MCC = 0$ both in cross-validation and validation for each experiment. Qualitatively, NBM1 performed slightly better than NBM2 in CV (median $MCC_{cv,NBM1} = 0.07$, $MCC_{cv,NBM2} = 0.03$; $p = 0.31$), showing opposite behavior in validation (median $MCC_{val,NBM1} = -0.06$, $MCC_{val,NBM2} = -0.02$; $p = 0.25$).

Figure 60: Classification results. A. DL results. B. SL results. C. Random TR/TS splits results. D. Results obtained testing various strategies to balance classes. MCC CV = MCC in CV; MCC val = MCC in validation.

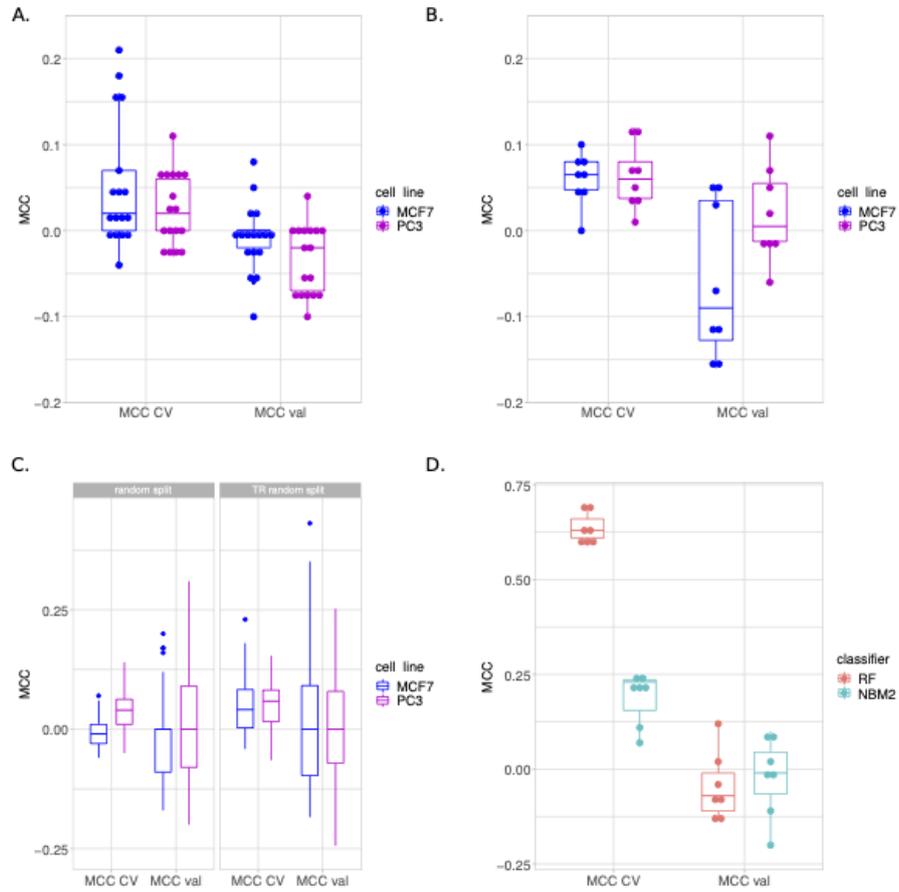


Table 38: Results obtained for RF and NBM2 classifiers using different class balancing strategies.

balancing strategy	classifier	MCC_{cv}	MCC_{val}
adasyn	RF	0.63 (0.60, 0.66)	0.12
oversampled_all	RF	0.69 (0.65, 0.71)	-0.13
oversampled_minority	RF	0.69 (0.65, 0.71)	-0.13
smote	RF	0.63 (0.60, 0.66)	0.02
smote_svm	RF	0.61 (0.59, 0.65)	-0.09
smote_borderline1	RF	0.61 (0.58, 0.64)	-0.04
smote_borderline2	RF	0.59 (0.55, 0.63)	-0.07
adasyn	NBM2	0.07 (0.03, 0.10)	0.02
oversampled_all	NBM2	0.24 (0.19, 0.29)	-0.02
oversampled_minority	NBM2	0.23 (0.19, 0.28)	0.07
smote	NBM2	0.20 (0.15, 0.25)	-0.2
smote_svm	NBM2	0.24 (0.20, 0.29)	0.1
smote_borderline1	NBM2	0.23 (0.19, 0.29)	-0.11
smote_borderline2	NBM2	0.11 (0.06, 0.16)	-0.01

9.4 Discussion

In the context of the CAMDA2018 CMap Drug Safety Challenge we performed an array of machine learning experiments to assess the capability of classifying **DILI** status from expression data derived from the two cancer cell lines MCF7 and PC3. We built three **DL** architectures to solve the assigned **DILI** classification task and compared their performance to two shallow machine learning algorithms (**RF** and **MLP**). Overall, we observed very poor classification performance both in **CV** and in validation, independently on cell line, feature set and classifier. Notably, the NBMDeep architecture performed significantly worse than the two shallower **DL** architectures, possibly due to a much larger number of parameters to train with limited data. A reduced number of samples is notoriously a limit for the applicability of **DL**. We investigated the existence of a better TR/TS split by randomly splitting the 266 samples into 100 artificial TR/TS splits containing 75 and 25% of the data. The results on these simulated TR/TS splits did not highlight the presence of a more informative partition

of the data. We additionally questioned whether the low MCC values obtained in validation indicate that the TR and TS samples are extracted from two distinct data distributions regardless of normalization. To indirectly test this hypothesis we randomly split the 180 samples of the TR set into 100 artificial TR/TS splits. The results obtained were in line with the random splits on the full dataset. As the two **DILI** classes were fairly imbalanced we tested two of our classifiers on a subset of the data (MCF7 expression data restricted to the KH feature set) with classes artificially balanced following multiple strategies. The results show a sharp improvement for **MCC** in **CV** (9.7 and 7.7 times for the **RF** and **DL** classifiers, respectively) with essentially no improvement in external validation, suggesting that the balancing strategies give rise to overfitting. An objective comparison with previous efforts aiming at **DILI** prediction is challenging, as most studies relied on compound chemical structures and molecular descriptors to assess **DILI** risk [88, 134, 205, 538]. The closest study we can consider for comparison is Kohonen et al. [244] as they also used CMap transcriptomics data for the creation of a **DILI** prediction score. However, the authors used the full CMap dataset, including ca. 1300 compounds and three cell lines, combined with the NCI-60 cytotoxicity data [422]. As the input is fundamentally much larger and therefore more suitable for training a model, a direct comparison with the classification strategies presented here is difficult to interpret.

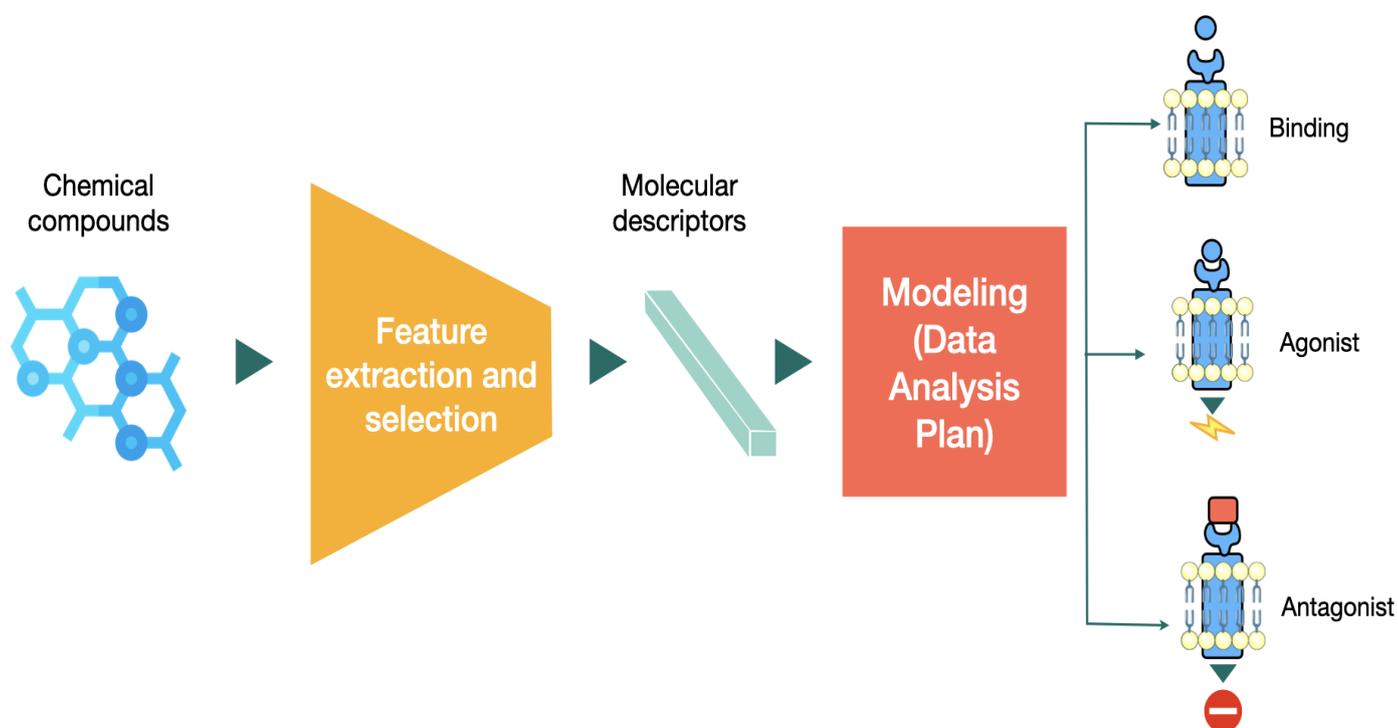
All our experiments point to the major conclusion that the data provided in the context of the CAMDA2018 CMap Drug Safety Challenge do not grant the capability of classifying the **DILI** status.

Implementation and code availability

The NBM₁, NBM₂ and NBMDDeep architectures were implemented in PyTorch v0.4.0 [353]. The **MLP** network and the **RF** models were implemented in scikit-learn v0.19.1 [357]. The whole **DAP** was written in Python. All **DL** computations were run on either a Microsoft Azure platform with 4x NVIDIA Tesla K80 GPU cards or on a Linux workstation with 2x NVIDIA GeForce GTX 1080 cards. Shallow learning models were run on the FBK KORE high-performance computing Linux cluster. All plots

were produced using the `ggplot2` R package [501]. Comparisons between conditions of interest were assessed by Wilcoxon test using the `wilcox.test` R function. The datasets supporting the conclusions of this article are available in the CAMDA2018-cmap-DILI repository, the code is written in Python 3.6 and available at <https://bit.ly/3ITfZcS>.

ML4TOX FOR DRUG BINDING ACTIVITY



Authors: M. Chierici, M. Giulini, N. Bussola, G. Jurman, C. Furlanello. *Original title:* Machine learning models for predicting endocrine disruption potential of environmental chemicals. *Published in:* Journal of Environmental Science and Health Part C. (Jan. 2019)

ML₄TOX FOR DRUG BINDING ACTIVITY

Highlights

- The novel AI framework ML₄Tox is designed to determine compound activities for the ligand-binding domain in predictive toxicology.
- On the CERAPP dataset, ML₄Tox is applied to classify agonist, antagonist, and binding properties of endocrine-disrupting chemical compounds.
- ML₄Tox, evaluated by DAP, significantly improves sensitivity (Sens) over previous studies, with Sens > 0.69 on all tasks.

Personal contribution I designed and implemented the deep learning models of the ML₄Tox framework. I also contributed to the writing and prepared the figures for the published manuscript.

10.1 Abstract

We introduce here ML₄Tox, a framework offering Deep Learning and Support Vector Machine models to predict agonist, antagonist, and binding activities of chemical compounds, in this case for the estrogen receptor ligand-binding domain. The ML₄Tox models have been developed with a 10x5-fold cross-validation schema on the training portion of the CERAPP ToxCast dataset, formed by 1677 chemicals, each described by 777 molecular features. On the CERAPP “All Literature” evaluation set (agonist: 6319 compounds; antagonist 6539; binding 7283), ML₄Tox significantly improved sensitivity over published results on all three tasks, with agonist: 0.78 vs 0.56; antagonist: 0.69 vs 0.11; binding: 0.66 vs 0.26.

Summary

With the aim to target faster, reproducible, and more cost-effective *in silico* safety assessment, we propose here the machine learning framework ML4Tox, including both deep and shallow learning in a classification pipeline. The framework can employ a deep multilayer network or a SVM (linear or gaussian) as predictive models. The central element in the approach is the DAP that takes care of massive replication of experiments on data partitions (See Chapter 2, Section 2.6.1.1). While in biomarker studies based on high-throughput -omics features the MCC is adopted as reference error function [228], in this study we also use sensitivity as a target performance metric, in order to give priority to compounds detected as active for potential toxicity.

We applied ML4Tox in three toxicology tasks defined in the Collaborative Estrogen Receptor Activity Prediction Project (CERAPP) [297]. Environmental exposure to endocrine-disrupting chemical compounds (EDC) poses high risks for human health, with potential impact on the endocrine system causing adverse immune, neurological and developmental effects. The interest of the toxicological community in datasets describing effects of estrogen receptor (ER) related compounds has steadily grown in the last few years, due to the key role of the ER molecular complex in the reproductive function.

Supported by accuracy improvements reported for autoencoder architectures trained on the ToxCast invitrodb dataset [61], we aim in particular at extending the application of the QSAR data-driven approach to deep learning architectures to improve sensitivity of CERAPP tasks. Specifically, the CERAPP assessed the application of predictive modeling to evaluate the binding interactions of environmental chemicals to the ligand-binding domain of human ER from *in vitro* high-throughput screening (HTS) assay data. These interactions are differentiated into three classes: agonist, antagonist and binding. The CERAPP has defined a training set of 5031 compounds (1677 per class) and the “Literature Evaluation set” (6319, 6539 and 7283 respectively for agonist, antagonist and binding) labelled as positive or negative. The data defines three distinct learning tasks, which are tackled by ML4Tox with a deep multilayer network, a linear SVM and a Gaussian SVM respectively, yielding superior performances

over those published. We present here the general architecture of the ML4Tox framework, its main methods and experimental application to the CERAPP tasks, finally discussing the potential for the future development of deep learning architectures in predictive toxicology.

10.2 Materials and Methods

Datasets

In this study, for model development we used the CERAPP training set (TR in the following), derived from ToxCastTM and Tox21 programs and consisting of 1677 chemicals [127, 211, 297] (Table 39). Each chemical was assigned a binary label representing their agonist, antagonist, and binding activities for the ligand-binding domain of ER. Models were tested on the CERAPP “Literature” evaluation set (EV), consisting of 6319 chemicals for agonist, 6539 for antagonist, and 7283 for binding (see Table 40). EV was derived from the 7547 compounds in the full CERAPP evaluation set after exclusion of chemicals with relatively high (> 20%) disagreement amongst literature sources [297].

Table 39: CERAPP training set for binary classification tasks.

Class	Active	Inactive	Total
Agonist	219	1458	1677
Antagonist	41	1636	1677
Binding	237	1440	1677

Table 40: CERAPP “Literature” evaluation set for binary classification tasks.

Class	Active	Inactive	Total
Agonist	350	5969	6319
Antagonist	284	6255	6539
Binding	1982	5301	7283

Feature extraction and filtering.

All chemicals were described by 777 molecular descriptors extracted from their bi-dimensional chemical structure by the software Mold2 [204, 206], following previous analyses [297]. Additionally, we explored the use of Extended-Connectivity Fingerprints [390] (ECF), a class of circular topological fingerprints, as an alternative set of descriptors. The ECF features were generated from the canonical SMILES of each compound using the Morgan algorithm [329] with four iterations, as implemented in the DeepChem [455] and RDKit^{rdkit} Python 2.7 modules. As a filtering step, features with constant values across the training samples were removed from the TR and EV sets before further analysis.

Machine learning

We applied [DL](#) and [SVM](#) models to predict compound agonist, antagonist and binding activities. The general ML4Tox architecture is sketched in [Figure 61](#).

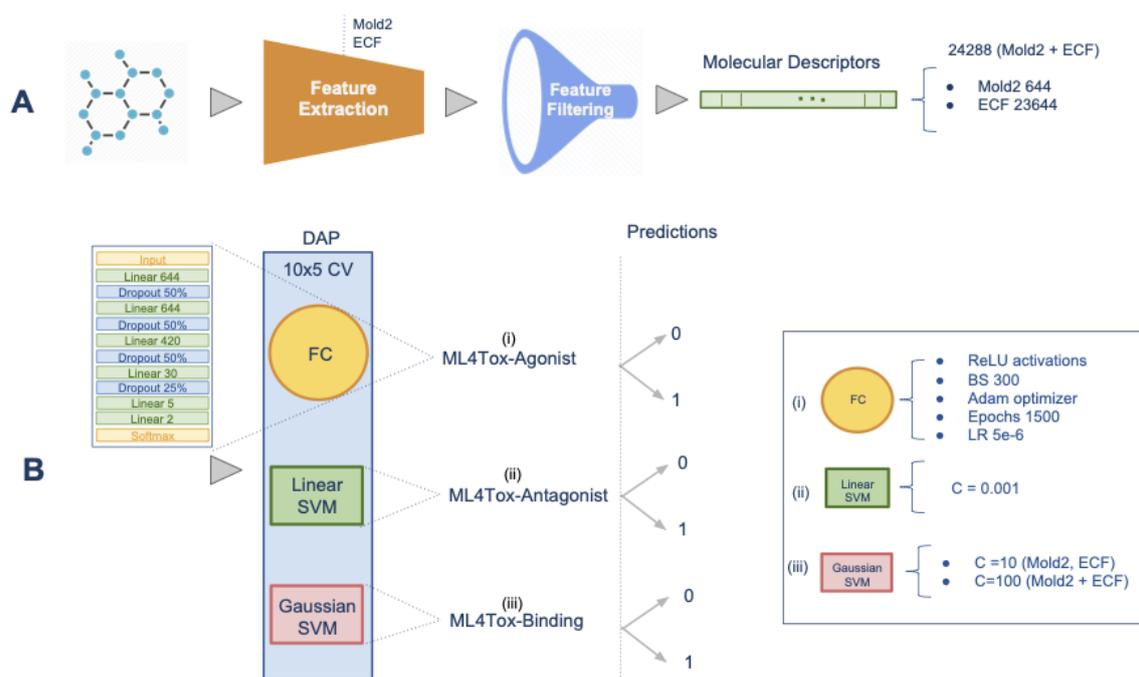


Figure 61: Overview of ML4Tox framework. A: Preprocessing of molecular descriptors from chemical compounds. Features are extracted by Mold2 or the Extended-Connectivity Fingerprint (ECF) algorithms. Descriptors constant across all training data are filtered out. B: Model training. Descriptors and binary labels (agonist, antagonist, binding) for each compound are used to train three *in-silico* models: ML4Tox-Agonist, ML4Tox-Antagonist, and ML4Tox-Binding. DAP: Data analysis protocol; FC: fully-connected layers; 10x5 CV: 10 rounds of 5-fold cross-validation.

We first developed ML4Tox-Agonist, a deep multilayer neural network to predict Agonist activity with five fully-connected hidden layers of 644, 644, 420, 30, and 5 nodes. As activation functions, we considered the Rectified Linear Unit [332] for the inner layers and a SoftMax one for the output layer. The optimizer was Adam [adam_optimizer] with a learning rate of $LR = 5 \times 10^{-6}$ and Cross Entropy as the loss function, with weights proportional to the class sizes. The batch size was 300 and the number of epochs 1500. To avoid overfitting, dropout layers were added with rate 0.5 after each of the first three inner layers and rate 0.25 after the fourth layer, with no dropout applied to the last hidden layer. Optimizer type, LR, number of epochs, and the dropout strategy were selected amongst alternatives by training over 4000 epochs on 70% of TR (randomly chosen) and evaluating the performance on the left-out 30%. Using a grid of LR values, we compared the training and val-

idation performance and losses of the net using stochastic gradient descent (SGD) (Figure 63) and Adam (Figure 62) as optimizers, with no dropout at first.

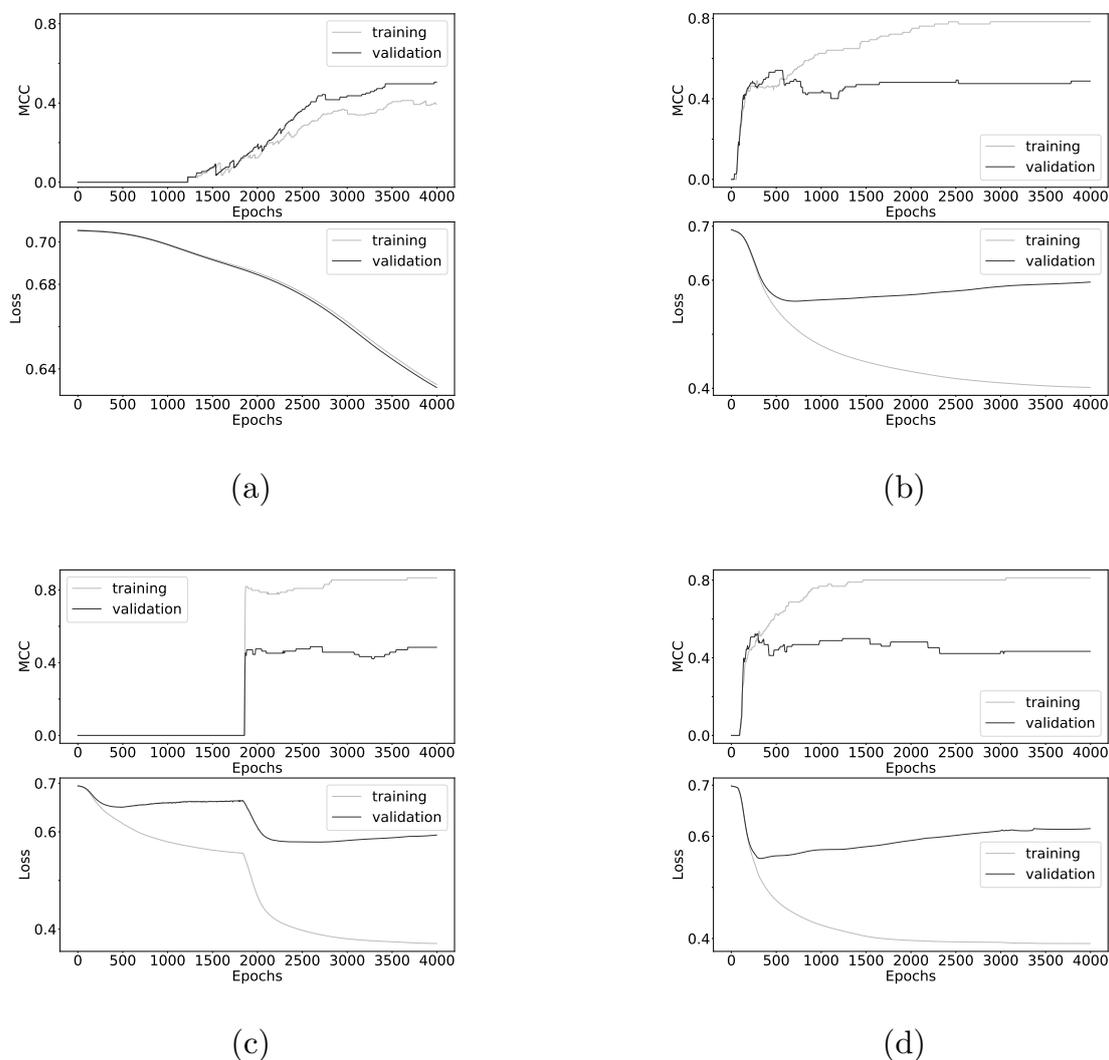


Figure 62: ML4Tox-Agonist: training with the Adam optimizer at different learning rates (LRs). (a) $LR=10^{-6}$; (b) $LR=5 \times 10^{-6}$; (c) $LR=7.5 \times 10^{-6}$; (d) $LR=10^{-5}$. Upper panels: training (grey) and validation (black) MCC; lower panels: training and validation losses (cross-entropy).

For SGD, the net was trained with $LR \in [10^{-4}, 5 \times 10^{-4}, 10^{-3}, 2 \times 10^{-3}]$: with the two lowest LR values (Figure 63, panels A-B) we observed a slow decay of training/validation losses combined with lower MCC (defined below) on training; with the two highest LR values (Figure 63, panels C-D), the net is slowly learning from data (MCC ~ 0.8), while also quickly overfitting.

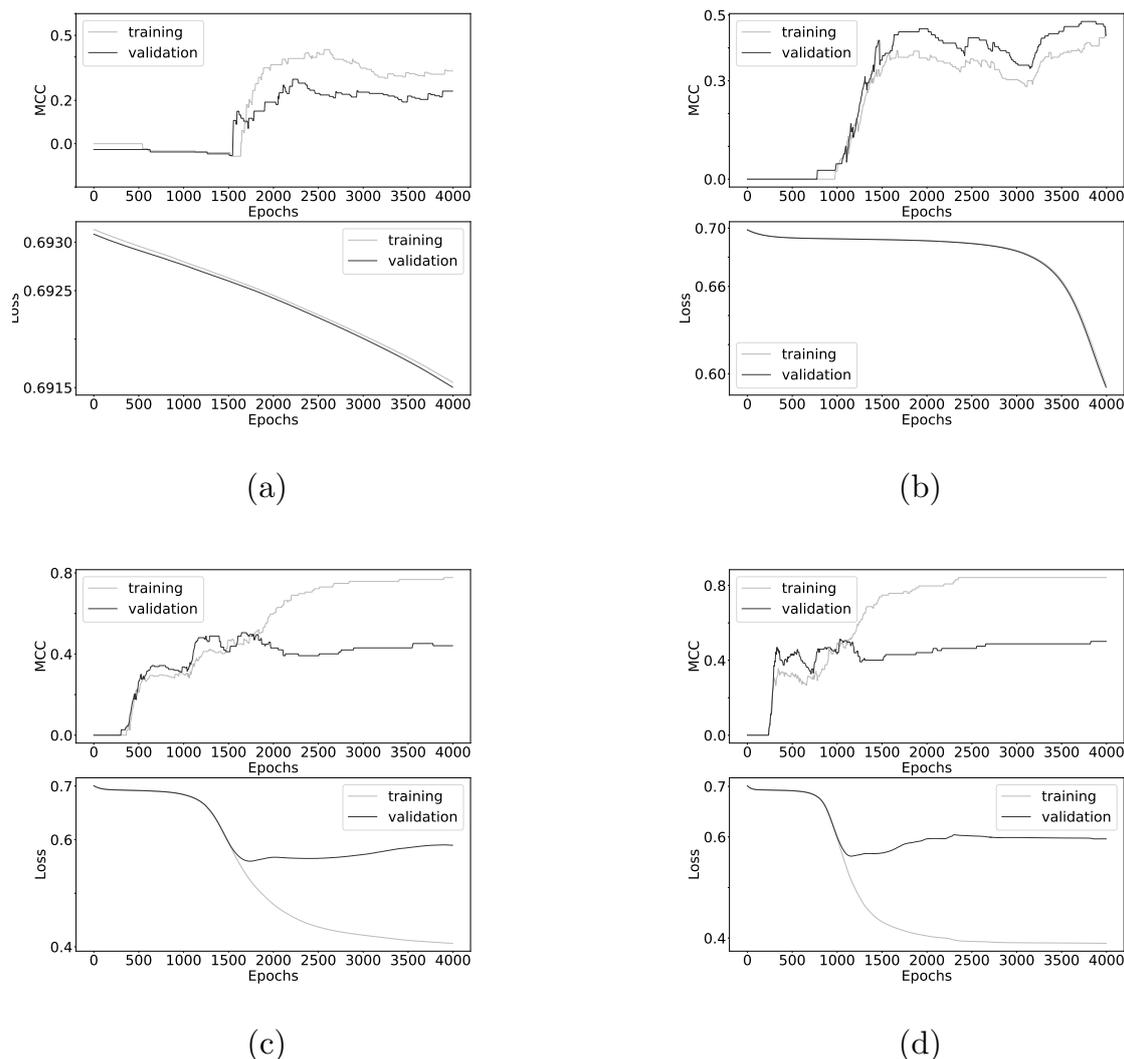


Figure 63: ML4Tox-Agonist: training with the SGD optimizer at different learning rates (LRs). (a) LR=0.0001; (b) LR=0.0005; (c) LR=0.001; (d) LR=0.002. Upper panels: training (grey) and validation (black) MCC; lower panels: training and validation losses (cross-entropy).

For Adam, we trained the net with $LR \in [10^{-6}, 5 \times 10^{-6}, 7.5 \times 10^{-6}, 10^{-5}]$, as Adam requires smaller LR with respect to SGD `[adam_optimizer]`. We chose $LR = 5 \times 10^{-6}$ as the best compromise for the LR; further, to delay overfitting and to narrow the gap between training and validation MCC, we added the dropout layers, with different dropout rates. The overall training performance with Adam and the optimal dropout rates are displayed in Figure 64. We fixed the number of epochs to 1500, to optimize for loss minimization on training and internal validation, *i.e.* ensuring the development of an accurate model while limiting overfitting.

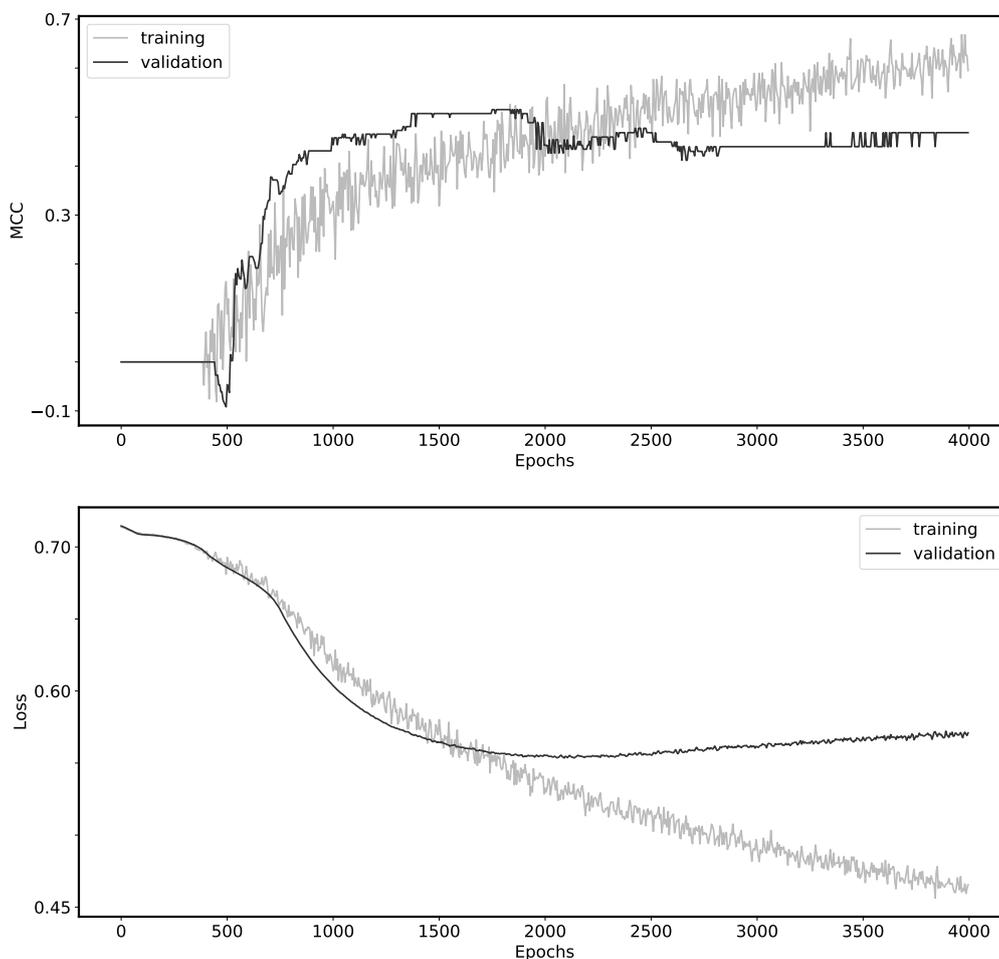


Figure 64: ML₄Tox-Agonist model: training convergence curves. Optimizer=Adam; LR= 5×10^{-6} and dropout rate=0.5 (inner layers 1–4) and 0.25 (layer 5).

For the Antagonist task, which is highly unbalanced with less than 3% of positive labels in the training set (Table 39), we developed ML₄Tox-Antagonist, a linear SVM with a regularization parameter $C = 0.001$ and balanced class weights. Linear, polynomial and Gaussian kernels were tested as possible kernels with a cross-validation strategy only on the training set. Details on the SVM metaparameter selection are not reported here for brevity. For the Binding Task, we designed ML₄Tox-Binding, a Gaussian kernel SVM trained on Mold2 and ECF features used alone or in combination (“Mold2+ECF”). The SVM regularization parameter was $C = 10$ for Mold2 and

Mold2+ECF, and $C = 100$ for ECF features; the kernel coefficient $\gamma = \frac{1}{n_f}$, for n_f the number of features. The optimal SVM regularization parameter was chosen from the grid [0.001, 0.01, 0.1, 1, 10, 100, 1000].

Predictive modeling strategy.

All machine learning models were trained and evaluated within the DAP framework and evaluated for a set of metrics, including balanced accuracy (BA) [52], sensitivity (SN), specificity (SP), and the MCC. In the binary case, BA and MCC are defined as $BA = \frac{1}{2} \left(\frac{TP}{AP} + \frac{TN}{AN} \right)$ and $MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$, respectively, for TN, TP, FN, FP the entries of the binary confusion matrix and $AP = TP + FN$, $AN = TN + FP$ (TN: true negatives; TP: true positives; FN: false negatives; FP: false positive). Features were rescaled in the interval [0, 1] before undergoing classification: to avoid information leakage, rescaling parameters from TR were used for rescaling both TR and EV sets. Random label experiments were also run to test against selection bias. After obtaining CV performance estimates, models were retrained on the whole TR set and evaluated on the EV set.

10.3 Results

After filtering out descriptors with constant values across training samples, a total of 24288 features were kept (644 Mold2 features, 23644 ECF features) for model selection, training and evaluation as described above. Classification results by ML4Tox on the CERAPP tasks are reported in Table 41, for Training (CV estimates) and Literature data. We compared ML4Tox with the FDA_NCTR_DBB model (DBB), which was among the best performing models on the same CERAPP tasks and datasets [297] (see Figs. 65–67).

On the Agonist task, the deep learning ML4Tox model scored a balanced accuracy $BA = 0.71$ (CI: 0.69, 0.73; 95% studentized bootstrap confidence interval) on Training and $BA_{EV} = 0.73$ on Evaluation. Notably, the model had fair sensitivity both in

Table 41: Classification performance of ML4Tox models in cross-validation on the training set, and in evaluation. CI: 95% studentized bootstrap confidence interval; MCC: Matthews Correlation Coefficient; BA: balanced accuracy; SN: sensitivity; SP: specificity. EV subscript indicates performance on the independent evaluation set. M2: Mold2 feature set.

Task	Feature set	MCC (CI)	BA (CI)	SN (CI)	SP (CI)	MCC _{EV}	BA _{EV}	SN _{EV}	SP _{EV}
Agonist	M2	0.32 (0.29, 0.35)	0.71 (0.69, 0.73)	0.60 (0.54, 0.65)	0.82 (0.80, 0.83)	0.22	0.73	0.78	0.68
Antagonist	M2	0.07 (0.05, 0.08)	0.60 (0.58, 0.62)	0.49 (0.44, 0.54)	0.70 (0.69, 0.72)	0.19	0.71	0.69	0.73
Binding	M2	0.37 (0.36, 0.39)	0.74 (0.73, 0.75)	0.70 (0.68, 0.72)	0.79 (0.78, 0.80)	0.22	0.60	0.37	0.84
Binding	ECF	0.42 (0.40, 0.44)	0.67 (0.66, 0.68)	0.38 (0.36, 0.40)	0.96 (0.96, 0.96)	0.25	0.59	0.23	0.94
Binding	M2+ECF	0.24 (0.23, 0.26)	0.67 (0.66, 0.68)	0.76 (0.74, 0.79)	0.59 (0.57, 0.60)	0.24	0.64	0.66	0.61

training $SN = 0.60$ (CI: 0.54, 0.65) and evaluation $SN_{EV} = 0.78$, improving on the original DBB model ($SN_{EV}^{DBB} = 0.56$), as shown in Figure 65.

On the Antagonist task, with $BA_{EV} = 0.71$ ($SN_{EV} = 0.69$, $SP_{EV} = 0.73$) in evaluation, the ML4Tox significantly improved over DBB both for balanced accuracy and sensitivity: $BA_{EV}^{DBB} = 0.55$ ($SN_{EV}^{DBB} = 0.11$, $SP_{EV}^{DBB} = 0.98$), as shown in Figure 66.

On the binding task, the most accurate ML4Tox model was Gaussian SVM over combined Mold2 and ECF features (M2+ECF), with $BA = 0.67$ (CI: 0.66, 0.68) and $SN = 0.76$ (CI: 0.74, 0.79) in cross-validation, and $BA_{EV} = 0.64$ ($SN_{EV} = 0.66$, $SP_{EV} = 0.61$), also significantly improved the tradeoff between sensitivity and specificity of the DBB model ($BA_{EV}^{DBB} = 0.60$, $SN_{EV}^{DBB} = 0.26$, $SP_{EV}^{DBB} = 0.94$), see Figure 67. Notably, the improvement in sensitivity of the ML4Tox-Binding model was obtained with the M2+ECF feature combination.

The consistent gain in sensitivity with ML4Tox in the antagonist and binding tasks is important in the context of predictive toxicology, where the prioritization of chemicals for *in vivo* risk assessment is one of the main goals.

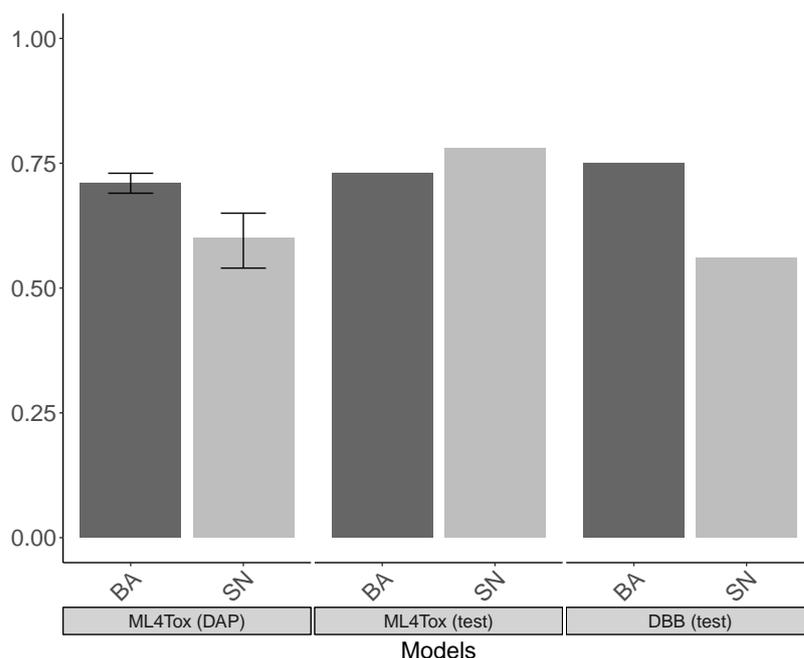


Figure 65: Agonist prediction task. ML4Tox-Agonist classification metrics in cross-validation on training (DAP: leftmost panel) and on independent evaluation set (test: middle panel), compared with evaluation metrics of the FDA-NCTR model (DBB: rightmost panel). Vertical bars represent 95% studentized confidence intervals. BA: balanced accuracy; SN: sensitivity.

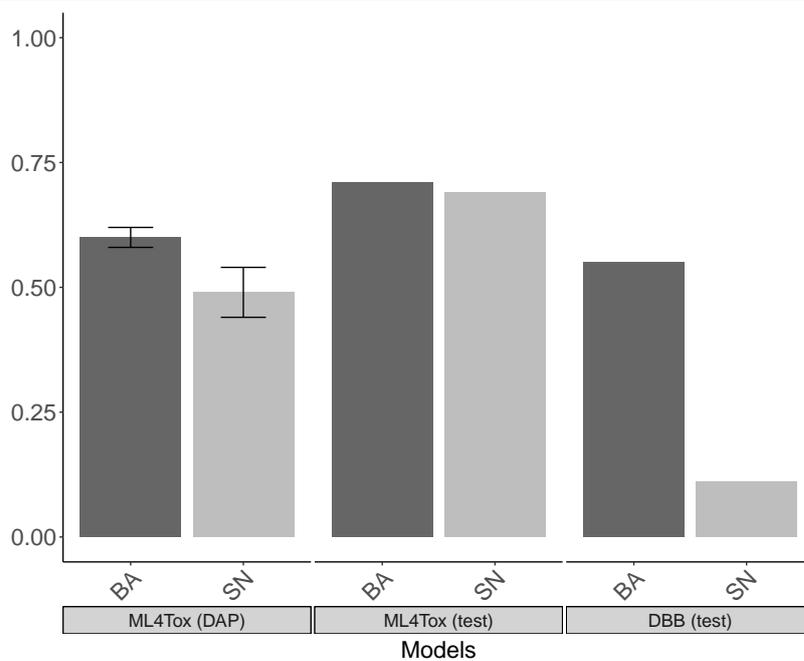


Figure 66: Antagonist prediction task. ML₄Tox-Antagonist classification metrics in cross-validation (DAP: leftmost panel) and on independent evaluation set (test: middle panel), compared with evaluation set metrics by NCTR model (DBB: rightmost panel). Vertical bars represent confidence intervals. BA: balanced accuracy; SN: sensitivity.

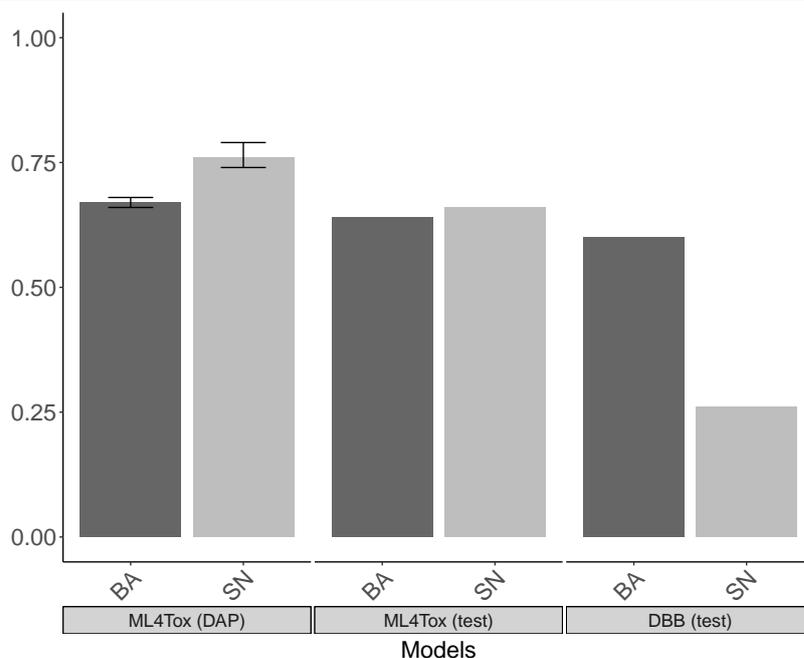


Figure 67: Binding prediction task. ML4Tox-Binding (Mold2+ECF features) classification metrics in cross-validation (DAP: leftmost panel) and on independent evaluation set (test: middle panel), compared with evaluation set metrics by NCTR model (DBB: rightmost panel). Vertical bars represent confidence intervals. BA: balanced accuracy; SN: sensitivity.

10.4 Discussion

In this study, we introduced ML4Tox, a predictive toxicology computational framework for modeling the potential endocrine disruption of environmental chemicals, based on machine learning. Our approach is motivated by the recent availability of QSAR-ready datasets from the literature.

We demonstrated ML4Tox by developing Deep Learning and Support Vector Machine models to predict chemical compound agonist, antagonist and binding activity for the human estrogen receptor ligand-binding domain, using data sourced from the CERAPP initiative.

We have explored the use of Deep Learning models for the first task (Agonist), and of shallow methods on the other two (Antagonist and Binding). We also test the potential of a different class of molecular descriptors (circular topological fingerprints) in the Binding task. While the sensitivity of these methods remains fair at

best, all three ML4Tox models improve over published results (Agonist: 0.78 vs 0.56; Antagonist: 0.69 vs 0.11; Binding: 0.66 vs 0.26).

We did not test the use of the richer set of features for predictive toxicology with deep architectures: combining the two improvements is expected to provide an advantage. In general, the recent availability of curated datasets has been used until now only with shallow statistical machine learning models [287] and thus it opens new potential applications for deep learning. In particular, as already proposed in the combination of diagnostic-prognostic tasks [290], we aim to exploit multi-task learning architectures to simultaneously solve the agonist, antagonist and binding tasks. The hypothesis that supports multi-task architectures in predictive toxicology is that when a shared core structure is trained to target several tasks simultaneously, each covering different aspects of toxicity, the model may be driven to better describe pathway disruption, thus improving its potential for prioritizing chemicals.

Implementation and code availability

The ML4Tox-Agonist models were implemented in PyTorch v0.4.1 [353], with code available at <https://bit.ly/3oeymkw>; ML4Tox-Antagonist and ML4Tox-Binding were built on top of the scikit-learn v0.19.1 [357] Python library. The whole DAP was written in Python based on scikit-learn functions. Computations were run on the FBK KORE high-performance computing cluster for shallow learning models, and on a Linux workstation with 2x NVIDIA GeForce GTX 1080 cards and on a cloud instance with 4x NVIDIA Tesla K80 GPU cards, funded by the Azure Research grant “Deep Learning for Precision Medicine”.

Funding

The authors gratefully acknowledge financial support of the FBK institutional program for Data Science (Big Data Analytics 2018).

Acknowledgments

The authors would like to thank Luca Coviello (FBK and Université Nice Sophia Antipolis) for useful technical comments on deep learning models. They particularly thank Drs. Huixiao Hong and Bohu Pan (FDA/NCTR) for kindly providing the Mold2-preprocessed CERAPP training and evaluation data sets. The Microsoft Azure resources used in training models was funded by the Azure Research grant “Deep Learning for Precision Medicine”, endowed to CF.

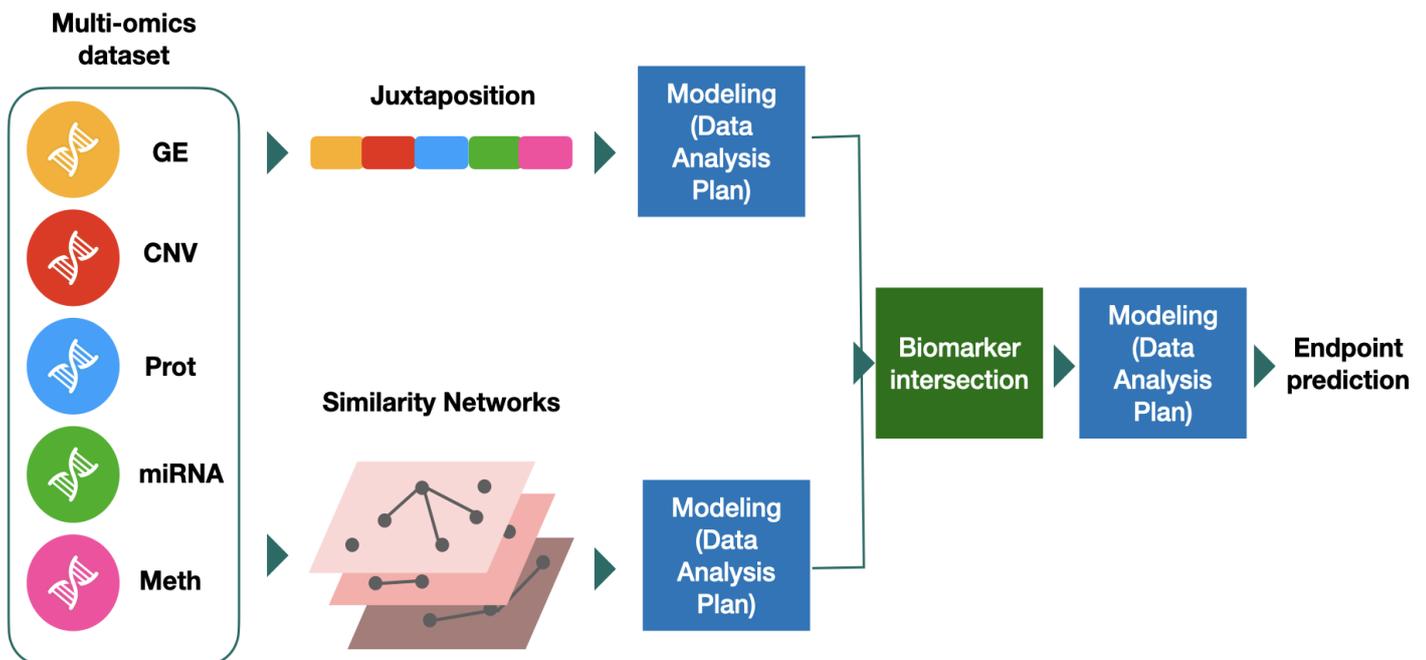
Part V

MULTI-OMICS INTEGRATION

Chapter 11 presents the network-based INF framework for biomarker development on multi-omics datasets. INF leverages ML models to extract compact signatures of predictive biomarkers in a reproducible environment. The diagnostic/prognostic potential of INF in precision oncology is demonstrated on public oncogenomics datasets for prediction of (i) estrogen receptor (ER) status in Breast Cancer, (ii) overall survival in acute myeloid leukemia and (iii) kidney renal clear cell carcinoma, and (iv) subtypes of breast invasive carcinoma. Following the approach in MAQC II [463], the ER status is used as the first target outcome in which models have relatively higher accuracy, as compared to outcomes such as overall survival. For example, MCC ranged from $MCC > 0.7$ on ER status to $MCC = 0.4$ on overall survival on the microarray Neuroblastoma data [463], and similarly for RNA-Seq data [525] (from $MCC = 0.3$ to $MCC = 0.6$ for overall survival of Neuroblastoma patients).

While INF is not proposed as a clinical alternative for the ER status prediction in breast cancer, we show that it supports the integration of multiple layers deriving from omics and possibly non-omics data. Of interest, the combination of diverse omics data for ER-status prediction can overcome the limits of IHC-stained analysis, due to the variability in slide preparation and pathologists' subjectivity [56]. For example, Alakwaa et al. [6] have analyzed metabolomics data to predict ER status and extract compact signatures for targeted therapies.

THE INF FRAMEWORK



Authors: M. Chierici, N. Bussola, A. Marcolini, M. Francescato, A. Zandonà, L. Trastulla, C. Agostinelli, G. Jurman, C. Furlanello. *Original title:* Integrative Network Fusion: a multi-omics approach in molecular profiling. *Published in:* Frontiers in Oncology, section Cancer Genetics. (June 2020)

THE INF FRAMEWORK

Highlights

- INF is a reproducible network-based framework for multi-omics integration that leverages ML models within a robust DAP, to extract predictive biomarkers from an arbitrary number of omics layers.
- INF was applied to predict estrogen receptor status (BRCA-ER, N=381) and breast invasive carcinoma subtypes (BRCA-subtypes, N=305) from gene expression, protein abundances and copy number variants. Further INF was applied to methylation, gene expression, and microRNA expression to predict overall survival in acute myeloid leukemia (AML, N=157) and kidney renal clear cell carcinoma (KIRC, N=181).
- INF improves over the performance of single layers and naive juxtaposition on (three) oncogenomics tasks, extracting a biologically meaningful compact set of predictive biomarkers.
- Notably, gene expression provides the largest number of features to the biomarkers' lists identified by the INF workflow.
- Feature analysis with UMAP on the INF signature identifies refined clusters of cancer subtypes consistent with published findings.

Personal contribution I contributed to the implementation of the overall INF framework and the design of the experimental pipeline. I run most of the experiments on the oncogenomics datasets, and I performed the UMAP analysis on the INF signature. I also contributed to the writing and prepared the figures for the article.

11.1 Abstract

Recent technological advances and international efforts, such as [TCGA](#), have made available several pan-cancer datasets encompassing multiple omics layers with de-

tailed clinical information in large collection of samples. The need has thus arisen for the development of computational methods aimed at improving cancer subtyping and biomarker identification from multi-modal data. Here we apply the Integrative Network Fusion (INF) pipeline, which combines multiple omics layers exploiting Similarity Network Fusion (SNF) within a machine learning predictive framework. INF includes a feature ranking scheme (rSNF) on SNF-integrated features, used by a classifier over juxtaposed multi-omics features (juXT). In particular, we show instances of INF implementing RF and linear Support Vector Machine (LSVM) as the classifier, and two baseline RF and LSVM models are also trained on juXT. A compact RF model, called rSNFi, trained on the intersection of top-ranked biomarkers from the two approaches juXT and rSNF is finally derived. All the classifiers are run in a 10x5-fold cross-validation schema to warrant reproducibility, following the guidelines for an unbiased Data Analysis Plan by the US FDA-led initiatives MAQC/SEQC. INF is demonstrated on four classification tasks on three multi-modal TCGA oncogenomics datasets. Gene expression, protein abundances and copy number variants are used to predict estrogen receptor status (BRCA-ER, N=381) and breast invasive carcinoma subtypes (BRCA-subtypes, N=305), while gene expression, miRNA expression and methylation data is used as predictor layers for acute myeloid leukemia and renal clear cell carcinoma survival (AML-OS, N=157; KIRC-OS, N=181). In test, INF achieved similar Matthews Correlation Coefficient (MCC) values and 97% to 83% smaller feature sizes (FS), compared with juXT for BRCA-ER (MCC: 0.83 vs 0.80; FS: 56 vs 1801) and BRCA-subtypes (0.84 vs 0.80; 302 vs 1801), improving KIRC-OS performance (0.38 vs 0.31; 111 vs 2319). INF predictions are generally more accurate in test than one-dimensional omics models, with smaller signatures too, where transcriptomics consistently play the leading role. Overall, the INF framework effectively integrates multiple data levels in oncogenomics classification tasks, improving over the performance of single layers alone and naive juxtaposition, and provides compact signature sizes.

Summary

Our study introduces Integrative Network Fusion (INF), a reproducible network-based framework for high-throughput omics data integration that leverages machine learning models to extract multi-omics predictive biomarkers. Originally conceptualized and tested on multi-omics metagenomics data in an early preliminary version [472, 519], INF combines the signatures retrieved from both the early-integration approach of variable juxtaposition (juXT) and an intermediate-integration approach (SNF [496]), to find the optimal set of predictive features. In particular, first a set of top-ranked features is extracted by juXT by a classifier, here Random Forest (RF) and linear Support Vector Machine (LSVM). Then, a feature ranking scheme (rSNF) is computed on SNF-integrated features and finally a RF model (rSNFi) is trained on the intersection of two set of top-ranked features from juXT and rSNF, obtaining an approach that effectively integrates multiple omics layers and provides compact predictive signatures. Selection bias and data-leakage effects are controlled by performing the experiments within a rigorous Data Analysis Plan (DAP) to warrant reproducibility, following the guidelines of the US FDA-led initiatives MAQC/SEQC [419, 464, 465]. In particular, to alleviate the computational burden of the full DAP pipeline, an approximating DAP is designed to lighten computing without significantly affecting the results. Further, experiments are run on samples with randomly shuffled labels as a sanity check versus overfitting effects and, finally, INF robustness is verified by testing on different train/test splits.

We test INF on three datasets retrieved from the TCGA repository, to predict either the estrogen receptor status (ER) or the cancer subtype on the breast invasive carcinoma (BRCA) dataset, and to predict the overall survival (OS) on the kidney renal clear cell carcinoma (KIRC) and acute myeloid leukemia (AML) datasets. Overall, INF improves over the performance of single layers and naive juxtaposition on all four oncogenomics tasks, extracting a biologically meaningful compact set of predictive biomarkers. Notably, the transcriptomics layer is prevalent inside the inferred INF signatures, consistently with published findings [80].

The INF framework is currently designed to integrate an arbitrary number of one-dimensional omics layers. We plan to further extend the framework by enabling the integration of histopathological features extracted from whole slide images [45] or deep features from radiological images [47] extracted by deep neural network architectures, carefully addressing all potential caveats [283].

11.2 Materials and Methods

Data

Three multi-modal cancer datasets generated by TCGA and four classification tasks are considered in this study. Protein abundance (*prot*), gene expression (*gene*) and copy number variants (*cnv*) are used to predict breast invasive carcinoma (BRCA) estrogen receptor status (0: negative; 1: positive) and subtypes (luminal A, luminal B, basal-like, HER2-enriched). Methylation (*meth*), gene expression (*gene*) and microRNA expression (*mirna*) are used to predict acute myeloid leukemia (AML) and kidney renal clear cell carcinoma (KIRC) overall survival (0: alive; 1: deceased). The number of samples and features for each omic layer and classification task are detailed in Table 43; class balance, split by dataset, is reported in Table 42.

Dataset-task	labels (#samples)
BRCA-ER	Negative (95), Positive (286)
BRCA-subtypes	LuminalA (170), LuminalB (102), Basal-like (81), HER2-enriched (48)
AML-OS	Dead (101), Alive (56)
KIRC-OS	Dead (133), Alive (48)

Table 42: Class balance. BRCA: breast invasive carcinoma; AML: acute myeloid leukemia; KIRC: kidney renal clear cell carcinoma; ER: estrogen receptor; subtypes: breast cancer subtypes; OS: overall survival.

For AML [461] and KIRC [460], gene expression is profiled using the Illumina HiSeq2000 and quantified as log₂-transformed RSEM normalized counts; miRNA mature strand expression is profiled using the Illumina Genome Analyzer and quan-

tified as reads per million miRNA mapped; and methylation is assessed by Illumina Human Methylation 450K and expressed as beta values. For BRCA [459], gene expression is profiled with Agilent 244K custom gene expression microarrays; protein abundance is assessed by reverse phase protein arrays; copy number profiles are measured using Affymetrix Genome-Wide Human SNP Array 6.0 platform, copy number variants are segmented by the TCGA Firehose pipeline using GISTIC2 method, and then mapped to genes.

The original data is publicly accessible on the National Cancer Institute GDC Data Portal¹ and the Broad GDAC Firehose², where further details on data generation can be found. The data was retrieved in December, 2019 and January, 2020 using the RTCGA R library [249].

Furthermore, the INF pipeline has been tested on a synthetic dataset with 380 observations in two classes (70% class 1 and 30% class 2, defining the synthetic target ST), 3 pseudo-omics layers, and 400 features (layer 1: 100; layer 2: 50; layer 3: 250). The dataset is generated in-house using *scikit-learn*'s `make_classification` function with the arguments `shuffle=False` and `flip_y=0`. The number of informative features and the difficulty of the task were set on a per-layer basis, as summarized in Table 45.

Dataset-task	#samples	layers (#features)
BRCA-ER	381	<i>gene</i> (17814), <i>cnv</i> (18050), <i>prot</i> (142)
BRCA-subtypes	305	
AML-OS	157	<i>gene</i> (10265), <i>meth</i> (2500), <i>mirna</i> (352)
KIRC-OS	181	<i>gene</i> (10265), <i>meth</i> (2500), <i>mirna</i> (484)
Synthetic-ST	380	layer1 (100), layer2 (50), layer3 (250)

Table 43: Data summary. BRCA: breast invasive carcinoma; AML: acute myeloid leukemia; KIRC: kidney renal clear cell carcinoma; *gene*: gene expression; *cnv*: copy number variants; *prot*: protein abundance; *meth*: methylation; *mirna*: microRNA expression; ER: estrogen receptor; subtypes: breast cancer subtypes; OS: overall survival; ST: synthetic target.

¹ <https://portal.gdc.cancer.gov/>

² <https://gdac.broadinstitute.org/>

In silico workflow

The INF pipeline integrates two or more omics layers, e.g. gene expression, protein abundance, or methylation, in a machine learning framework for improved patient classification and biomarker identification in cancer. The core consists of three main components, structured as in Figure 68, managing the integration of the omics layers and their predictive modeling. A baseline integration method (juXT) is first considered by training a RF [51] or a linear Support Vector Machine (LSVM) [107] classifier on juxtaposed multi-omics data, ranking features by ANOVA F-value. Secondly, the multi-omics features are integrated by Similarity Network Fusion (SNF) [496], a method that computes a sample similarity network for each data type and fuses them into one network. INF introduces a novel feature ranking scheme (rSNF) that sorts multi-omics features according to their contribution to the SNF-fused network structure. A RF or LSVM classifier is trained on the juxtaposed multi-omics data, ranking features by rSNF. A compact RF model (rSNFi) is finally trained on the juxtaposed dataset restricted to the intersection of top-ranked biomarkers from juXT and rSNF.

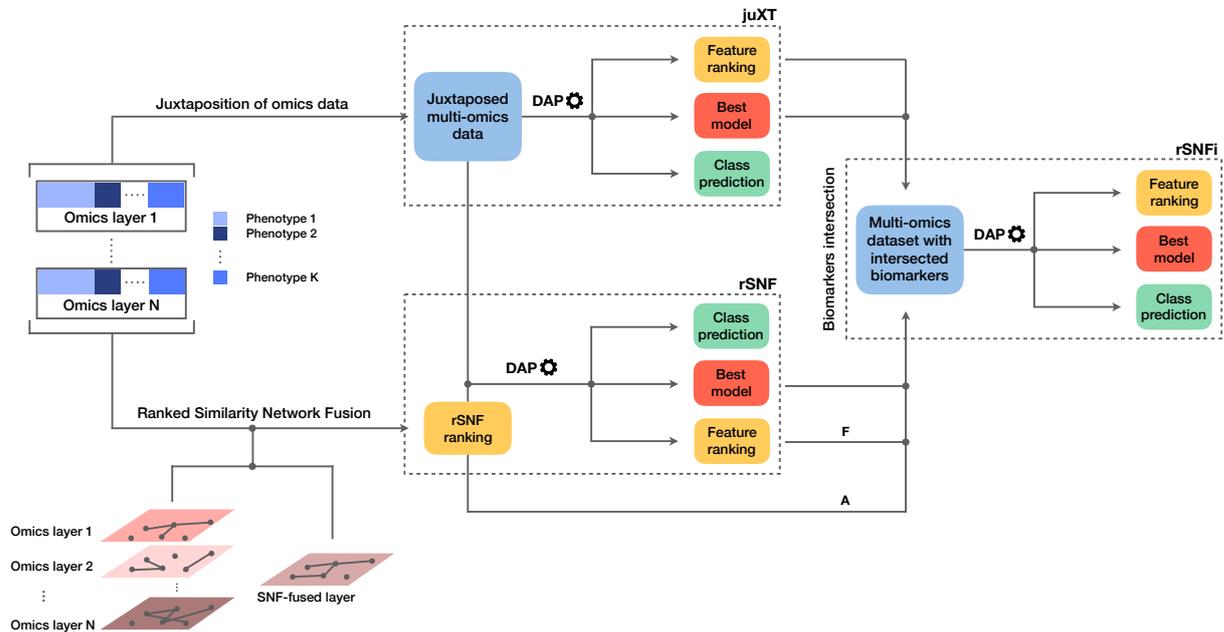


Figure 68: Graphical representation of the INF workflow for N omics datasets with K phenotypes. A first RF or LSVM classifier is trained on the juxtaposed data, ranking features by ANOVA F-value ($juXT$). The data sets are then integrated by Similarity Network Fusion, the features are ranked by rSNF and a RF or LSVM model is developed on the juxtaposed dataset with the rSNF feature ranking ($rSNF$). Finally, a RF or LSVM classifier is trained on the juxtaposed dataset restricted to the intersection of $juXT$ and $rSNF$ top discriminant feature lists ($rSNFi$). The classifier is either RF or LSVM throughout the INF workflow. All the predictive models are developed within the DAP described in the methods and graphically represented in Figure 6. The alternative and mutually exclusive paths A and F are followed by the “accelerated DAP” and the “full DAP” procedures, respectively (see Methods)

Omics integration

In a comparative review of scientific literature, SNF [496] emerged as one of the most reliable alternatives to simple juxtaposition-based integration. SNF is a non-Bayesian network-based method that can be divided into two main steps: the first step builds

a sample-similarity network for each omics dataset, where nodes represent samples and edges encode a scaled exponential Euclidean distance kernel computed on each pair of samples; the second step implements a nonlinear combination of these networks into a single similarity network through an iterative procedure. The multi-omics datasets are first converted into graphs, and for each graph two matrices are computed: a patient pairwise similarity matrix (“status matrix”), and a matrix with similarity of each patient to the K most similar patients, through K -nearest neighbors (“local affinity matrix”). At each iteration, the status matrix is updated through the local affinity matrix, generating two parallel interchanging processes. The status matrices are finally fused together into a single network. Spectral clustering is performed on the fused network, in order to identify sub-communities of samples, potentially reflecting phenotypes. The clustering performance is evaluated with respect to a ground truth, *i.e.*, the real phenotype each sample belongs to, by the Normalized Mutual Information (NMI) score. SNF integrates multiple omics datasets into a single comprehensive network in the space of samples rather than measurements (*e.g.*, gene expression values).

This work proposes multi-omics integration as an approach to identify robust biomarkers of sample phenotypes or cancer subtypes (*e.g.*, survival status vs breast cancer subtyping); consequently, it is necessary to extract measurement information from the SNF-fused network of samples. To this aim, we extended SNF by implementing *rSNF* (ranked SNF), a feature-ranking scheme based on SNF-fused network clustering. In detail, a patient network W_i is built for each feature f_i , based on f_i alone, and spectral clustering is performed on it. Then, NMI score is computed comparing the clusters found inside W_i with those in the fused network; the higher the score, the more similar the clustering between the fused network and W_i . Thus, each feature f_i is associated to a consistency score, ranking all multi-omics features with respect to their relative contribution to the whole network structure.

The entire procedure of similarity networks inference and fusion relies on two hyperparameters: α , the scaling variance in the scaled exponential similarity kernel used for similarity networks construction, and K , the number of nearest neighbors in sparse kernel and scaled exponential similarity kernel construction. While the

original method [496] assigned fixed values to α and K , in this study the optimal hyperparameters are chosen among the grids $\alpha_{\text{grid}} = \{0.3, 0.35, 0.4, 0.45, \dots, 0.8\}$ and $K_{\text{grid}} = \{i \in \mathbb{N}, 10 \leq i \leq 30\}$ in a 10×5 -fold cross-validation schema.

Predictive profiling

To ensure the reproducibility of results and limit overfitting, the development of classification models is performed inside the DAP (see Chapter 2, Section 2.6.1.1). Data is split in a training set (TR) and two non-overlapping test sets (TS, TS₂), preserving the original proportion of patient phenotypes (classes). The TR/TS/TS₂ partitions are 50%/30%/20% of the entire data set, respectively. The data splitting procedure is repeated 10 times to obtain 10 different TR/TS/TS₂ splits. Predictive models are trained and developed on TR and TS for juXT and rSNF; in the case of rSNFi, the models are trained and developed on TS and TS₂ to avoid information leakage due to using the same data both for feature selection and model training (see Figure 69).

For each split, RF or linear kernel Support Vector Machine (LSVM) classifiers are trained on the training partition within a stratified 10×5 -fold cross-validation (10×5 -MCVCC!). The model performance is assessed in terms of average precision, recall, and MCC. At each MCVCC! round, features are ranked either by ANOVA F-value (for juXT, rSNFi) or by the rSNF ranking (see Section 11.2) and different classification models are trained for increasing numbers of ranked features, namely 5%, 10%, 25%, 50%, 75%, and 100% of the total features. A unified list of top-ranked features is then obtained by Borda aggregation of all the ranked MCVCC! lists [227, 230]. The best model is later retrained on the whole training set restricted to the features yielding the maximum MCC in MCVCC!, and validated on the test partition. A global list of top-ranked features is derived for juXT, rSNF, and rSNFi by Borda aggregation of the Borda lists of each TR/TS split (Borda of Bordas, “BoB”). The signatures for juXT, rSNF, and rSNFi are defined by the top N features of the corresponding BoB lists, with N being the median size of top features across all experiments.

In the “full” version of the DAP (*fDAP*), described above, the rSNF ranking is performed at each MCVCC! round on the training portion of the data. Since this

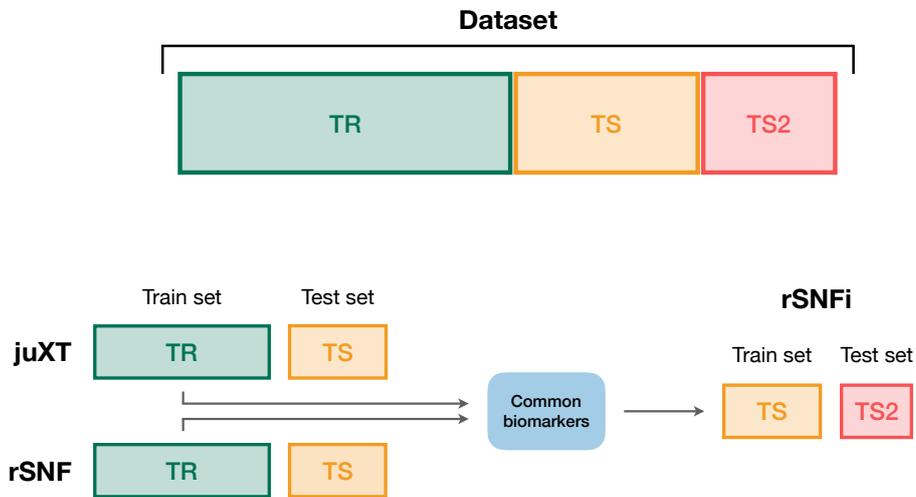


Figure 69: Data splitting procedure. To avoid information leakage due to the use of the same data both for feature selection and model training, we considered different train and test sets according to the integration scheme. In particular, each data set is split into three non-overlapping partitions (TR/TS/TS₂), corresponding to the 50%/30%/20% of the entire data set, respectively. The TR/TS/TS₂ partitions preserve the original proportion of patient phenotypes. Predictive models for juXT and rSNF are trained on TR and validated on TS, while for rSNFi the train set is TS (with features restricted to the intersected biomarkers of juXT and rSNF) and TS₂ the test set.

procedure is quite demanding in terms of computational time, even if parallelized (≈ 9 feature/min), we devised an “accelerated” version of the DAP (*aDAP*), where the rSNF ranking is precomputed on the whole TR data and used as is at each MCVCC! round. We assessed the *fDAP* vs *aDAP* performance on the synthetic dataset as well as BRCA-ER and BRCA-subtypes by comparing the overall metrics and measuring the dissimilarity of the rSNF BoB of the two DAPs by the Canberra distance [227].

RF models are trained using 500 trees, measuring the quality of a split as mean decrease in the Gini impurity index [51]; the regularization parameter C of LSVM models is tuned over the grid $C_{\text{grid}} = \{10^i, i \in \mathbb{N}, -2 \leq i \leq 3\}$ within a $10 \times$ stratified

Monte Carlo cross-validation (50% training/validation proportion). Results for RF models are summarized in Table 44, while LSVM models performance is detailed in the Supplementary Tables online.

To ensure that the predictive profiling procedure is not affected by selection bias, the whole INF workflow, including the rSNF procedure, is also repeated after randomly scrambling the training set labels (“random labels” mode): in the absence of systematic bias, MCC is expected to be close to the random guess value of zero.

Implementation

The complete INF pipeline is implemented through the workflow management tool Snakemake [250], which allows automatic handling of all dependencies required to generate the INF output. The pipeline operates on N omics input files, one for each layer that should be integrated, and a single file describing the patient labels. The omics files are tab-separated text matrices with patients on the rows and features on the columns, with row and column identifiers. The label file is a single column file with patient phenotypes, with no header. This input structure, with one file per omic layer and a label file, simplifies the downstream analysis and reduces to a minimum the preprocessing burden for the end user.

The predictive profiling module, including the DAP, is written in Python 3.6 on top of NumPy [346] and scikit-learn methods [357]. The ranked SNF (rSNF) procedure is implemented in R [375] leveraging the original R scripts provided by SNF authors [496], extended by a dedicated script for SNF tuning and a main script for SNF analysis and the post-SNF feature selection procedure, which is parallelized over the features for efficiency using the foreach R library.

Table 44: Summarized best predictive performances for each classification task using RF model and three omics layers. CI: 95% bootstrap confidence interval; {MCC,PREC,REC}_cv: best average MCC, precision, recall in cross-validation on training set splits; {MCC,PREC,REC}_ts: average MCC, precision, recall on validation set splits; Nf: median number of features leading to MCC_cv. Bold indicates best performance (highest MCC and smallest signature size). Precision and recall were computed for binary classification tasks only.

Task	Method	MCC_cv (CI)	MCC_ts (CI)	PREC_cv (CI)	PREC_ts (CI)	REC_cv (CI)	REC_ts (CI)	Nf
BRCA-ER	juXT	0.785 (0.776, 0.795)	0.797 (0.778, 0.819)	0.935 (0.932, 0.938)	0.946 (0.935, 0.957)	0.962 (0.959, 0.965)	0.955 (0.949, 0.962)	1801
	rSNF	0.792 (0.782, 0.801)	0.804 (0.779, 0.830)	0.938 (0.935, 0.941)	0.947 (0.934, 0.961)	0.961 (0.958, 0.965)	0.958 (0.949, 0.966)	1801
	rSNFi	0.820 (0.808, 0.831)	0.830 (0.803, 0.857)	0.955 (0.951, 0.959)	0.951 (0.939, 0.962)	0.956 (0.952, 0.960)	0.967 (0.956, 0.977)	55.5
BRCA-subtypes	juXT	0.778 (0.771, 0.785)	0.795 (0.771, 0.817)	-	-	-	-	1801
	rSNF	0.769 (0.762, 0.777)	0.811 (0.787, 0.835)	-	-	-	-	1801
	rSNFi	0.788 (0.778, 0.798)	0.838 (0.794, 0.879)	-	-	-	-	301.5
KIRC-OS	juXT	0.266 (0.243, 0.289)	0.305 (0.229, 0.382)	0.540 (0.509, 0.570)	0.579 (0.494, 0.664)	0.299 (0.280, 0.317)	0.343 (0.300, 0.393)	2319
	rSNF	0.253 (0.230, 0.276)	0.274 (0.189, 0.348)	0.539 (0.505, 0.571)	0.628 (0.507, 0.739)	0.253 (0.235, 0.270)	0.257 (0.200, 0.314)	3313
	rSNFi	0.268 (0.239, 0.298)	0.378 (0.288, 0.464)	0.485 (0.449, 0.521)	0.594 (0.512, 0.668)	0.321 (0.296, 0.347)	0.490 (0.380, 0.600)	111
AML-OS	juXT	0.141 (0.120, 0.163)	0.223 (0.146, 0.307)	0.675 (0.669, 0.681)	0.704 (0.682, 0.725)	0.860 (0.849, 0.870)	0.880 (0.850, 0.907)	6559
	rSNF	0.180 (0.157, 0.202)	0.263 (0.175, 0.366)	0.685 (0.679, 0.691)	0.717 (0.692, 0.743)	0.876 (0.867, 0.886)	0.873 (0.847, 0.903)	656
	rSNFi	0.274 (0.245, 0.301)	0.176 (0.068, 0.278)	0.726 (0.718, 0.735)	0.673 (0.639, 0.706)	0.870 (0.858, 0.882)	0.835 (0.785, 0.880)	91.5

Computational details

The INF computations were run on the FBK Linux high-performance computing facility KORE, on a 8-core i7 3.4 GHz Linux workstation, and on a 72-vCPU 2.7 GHz Platinum Intel Xeon 8168 Microsoft Azure cloud machine (F72s v2 series).

11.3 Results

The INF workflow was run on all tasks considering 3-layer integration and all 2-layer combinations; the [DAP](#) was also run separately on all single-layer datasets in order to obtain a baseline. All results presented here refer to experiments performed with [RF](#) classifier. Experiments using LSVM were performed on BRCA-ER and KIRC-OS obtaining similar classification performances, top features and layer contributions (see Supplementary Material tables online). The classifier performance for 3-layer integration is summarized in [Table 44](#), in terms of average cross-validation [MCC](#) on the 10 training set splits (MCC_{cv}) with 95% Studentized bootstrap confidence intervals (CI) as (MCC_{min} , MCC_{max}), average [MCC](#) on the 10 test set splits (MCC_{ts}) with CI, and median number of features (N_f) yielding MCC_{cv} . Similarly, precision (PREC) and recall (REC) are reported in [Table 44](#) as average cross-validation and test set values with CI. As expected, whenever there is a non negligible unbalance towards the positive class, the number of false positives tends to increase, with more false positives yielding a comparatively low precision with higher recall, and vice versa. In both cases, the [MCC](#) efficiently works in balancing the two effects. The classifier performance on single-layer and 2-layer data is summarized in [Figure 70](#).

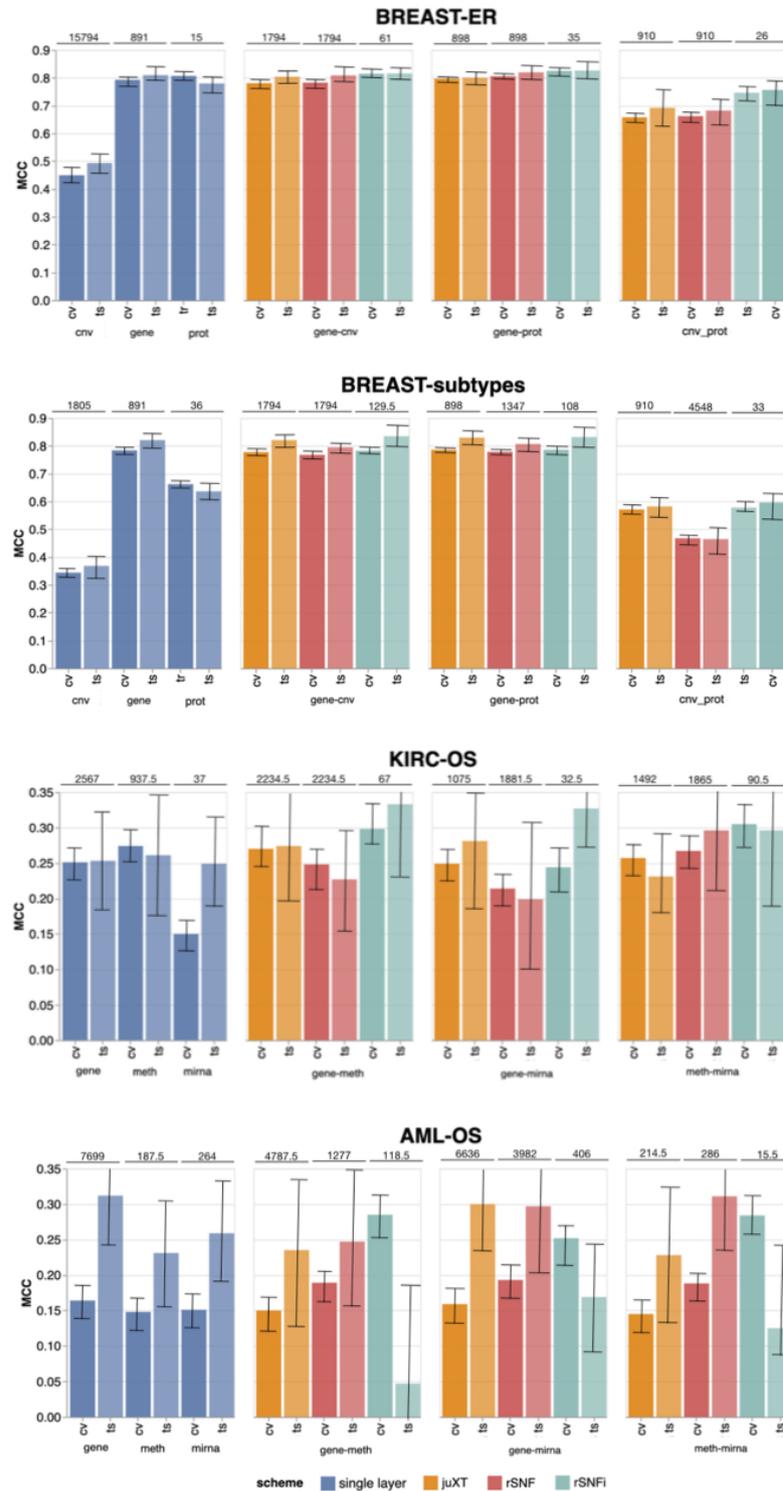


Figure 70: Overview of Random Forest classification performance (MCC, Matthews Correlation Coefficient) on the four tasks in cross validation (“CV”) and test (“ts”), on single-layer (blue shades) and on all two-layer combinations for juXT (orange), rSNF (red) and rSNFi (green). Bars indicate 95% confidence intervals. On top of each MCVCCI-ts pair is the median number of features leading to best MCVCCI! performance.

A comparison between the “accelerated” flavour of the DAP (*aDAP*) and the full DAP (*fDAP*) was run on synthetic data, BRCA-ER and BRCA-subtypes data, with *aDAP* yielding similar performance metrics and top-ranked biomarker lists as *fDAP* (Results reported Supplementary Material online), while being $\approx 30\times$ faster (for BRCA-ER, approx. 2h vs 64h, or 300 features/min vs 9 features/min). All the results presented here were thus obtained using *aDAP*. Moreover, the INF workflow running in “random labels” mode achieved an average cross-validation MCC ≈ 0 , as expected by a procedure unaffected by systematic bias.

Overall, integrating multiple omics layers with INF yields better or comparable classification performance than using only features from a single layer or naïve omics juxtaposition, at the same time with much more compact signature sizes. On 3-layer BRCA-subtypes and 2- or 3-layer KIRC-OS, INF outperforms the single layers, as well as juXT and rSNF (Figure 70, Table 44). On 2-layer BRCA-subtypes, INF performance on *gene-cnv* and *gene-prot* is comparable to the best-performing single-layer data (*gene*) and superior to *cnv* and *prot* single layers, while INF on *cnv-prot* only improves over the *cnv* single layer. On the BRCA-ER task, the performance with INF integration of 2 or 3 layers is still better than using single layers, nevertheless to a smaller extent, except for *cnv-prot* integration which performs better than *cnv* alone but slightly worse than *gene* and *prot* single layers. The good performances achieved at the *gene* and *prot* single layers do not come unexpected, since the biological nature of the target ER-status is defined at transcriptomics level. On the more difficult AML-OS task, INF has better performance over both rSNF and juXT on *gene-mirna* and *meth-mirna* integration, still improving over single-layer performance both in terms of MCC and reduced signature sizes.

One or multi-omics layers vs juXT/rSNF/rSNFi.

For BRCA-ER, three-layer INF (rSNFi) integration performs better than either rSNF or juXT (MCC test 0.830 vs 0.804, 0.797 for rSNF and juXT, respectively). All two-layer INF integrations perform similarly to, or better than, the corresponding rSNF and

juXT integrations, in particular for *cnv-prot* integration (MCC test 0.746 vs 0.682, 0.692 resp. for rSNF and juXT).

On BRCA-subtypes, the 3-layer INF integration performs better than either rSNF or juXT (MCC test 0.838 vs 0.811, 0.795 resp. for rSNF and juXT), nevertheless without improving over the *gene* single-layer performance (MCC test 0.821). However, the INF median signature size is only 301.5, compared to 1801 for rSNF and juXT, and 891 for the *gene* layer alone. All two-layer INF integrations yield better performance than their corresponding juXT or rSNF integrations.

Omics integration is particularly effective for KIRC-OS, as all 2- and 3-layer INF integrations outperform juXT, rSNF, and each of the single-layer classifiers. In fact, 3-layer rSNFi achieves MCC test 0.378 vs 0.274, 0.305 (resp. for juXT, rSNF), 0.296, 0.327, 0.333 (resp. rSNFi *meth-mirna*, *gene-mirna*, *gene-meth*), and 0.253, 0.261, 0.249 (resp. *gene*, *meth*, *mirna*).

For AML-OS, INF feature sets are always more compact than either juXT or rSNF, with three-layer integration giving better MCC than any of the INF two-layer integrations (MCC test 0.176 vs 0.125, 0.169, 0.047, respectively three-layer vs *meth-mirna*, *gene-mirna*, *gene-meth*). Moreover, cross-validation MCCs corresponding to INF integration are better than any single layer MCC as well as rSNF and juXT.

Characterization of the signatures identified by INF.

For all tasks, INF signatures are markedly more compact with respect to both juXT and rSNF. With 91.5 vs 6559 (1.4%) median features (rSNFi vs juXT), the largest reduction in size occurs for AML-OS 3-layer integration, while the least reduction is observed for BRCA-subtypes task, with 301.5 vs 1801 (16.7%) median features (rSNFi vs juXT).

In terms of contributions from the omics datasets being integrated, the *gene* layer generally provides the largest number of features to the signatures identified by the INF workflow. In particular for the BRCA dataset, in both ER and subtypes tasks, the *gene* layer contributes over 95% of the top features for juXT and rSNFi, with rSNF signatures being slightly more balanced (*prot* contribution remains marginal, while

cnv provides 28.3% and 17.7% of the top features in ER and subtypes tasks respectively). This is expected as the class label is defined mainly at transcriptomics level. In AML-OS experiments, the layer contributing the most is still *gene*, accounting for ca. 78%, 73% and 81% of the top feature sets for RF juXT, rSNF and rSNFi experiments, respectively. In KIRC-OS experiments, *gene* is the layer contributing the most to the top juXT and rSNF feature sets, while *meth* is the major contributor for rSNFi. The percentage of features from each omic layer contributing to the top signatures for juXT, rSNF and rSNFi 3-layer integrations are reported in the Supplementary Material online. The RF rSNFi signatures for all tasks are available in Supplementary Material online.

Even though a systematic biological interpretation of the signatures identified is beyond the scope of this work, to ascertain the reliability of our results we compared them with published data. The top features in the BRCA-ER rSNFi signature include multiple genes known to be associated with breast carcinoma progression and outcome such as *AGR3*, *B3GNT* and *MLPH* [160, 367, 458]. In addition we find the estrogen receptor gene (*ESR1* from the *gene* and ER-alpha from the *prot* layer) and the transcription factor *GATA3* (from both *gene* and *prot* layers) [173]. Both the BRCA-ER and BRCA-subtypes signatures include genes previously identified as novel biomarkers for intrinsic breast carcinoma subtype prediction [319]. Interestingly there is only partial overlap between the top features identified in BRCA ER vs subtypes tasks. Considering AML-OS task, it is noteworthy to mention that the top feature identified has been recently reported as a potential biomarker predicting overall survival in a subset of AML patients [23].

Within the *mirna* features of the AML-OS signature, *MIR-203* expression was recently found to be associated with AML patient survival [172]; *MIR-100* is highly expressed in AML and was found to regulate cell differentiation and survival [533]; high expression of *miR-504-3p* was reported to be associated with favorable AML prognosis [269]. Given that the rSNFi signature identified in the KIRC-OS task contains a large percentage of methylation data (86.5%), its direct interpretation is more difficult. It is however interesting to observe that all the 15 *gene* features in the sig-

nature are identified as prognostic markers for renal carcinoma according to the Human Protein Atlas [475].

Unsupervised analysis

The features selected by juXT, rSNF and rSNFi are projected on a bi-dimensional space using the UMAP unsupervised multidimensional projection method (see Chapter 2, Section 2.6.3.1). Here we show an example on the BRCA-subtypes 3-layer dataset, with a UMAP projection of the features selected by juXT (Figure 72) compared to the UMAP projection of the INF signature (Figure 71) for one of the 10 data splits (the UMAP plots for the remaining 9 splits are reported in Figure 73 and Figure 74). Colors represent cancer subtypes and shapes represent training/test partitions. Using the 1801 juXT features, cancer subtypes are roughly clustered, with HER2-enriched and Luminal B being more dispersed (Figure 72). The clusters appear to be more sharply defined in the projection of the 302-feature INF signature: in particular, Basal-like patients form a distinct cluster, while Luminal A, Luminal B and HER2-enriched patient clusters are close to each other, slightly overlapping yet hinting to a trajectory pattern (Figure 71). The HER2/luminal cluster contains two patients classified as basal-like subtype, consistently with the findings of Koh and colleagues [243].

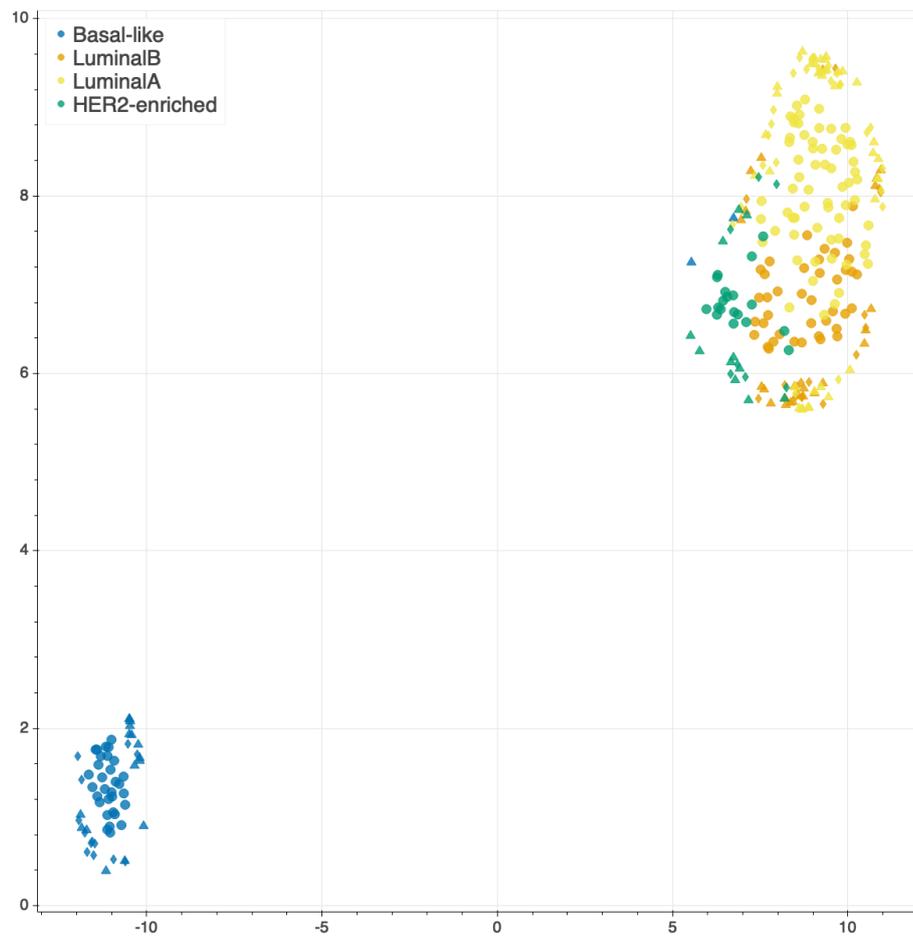


Figure 71: UMAP projection on the BRCA-subtypes task with 3-layer juxtaposed data restricted to the rSNFi signature. Circle: TR set; triangle: TS set; diamond: TS2 set.

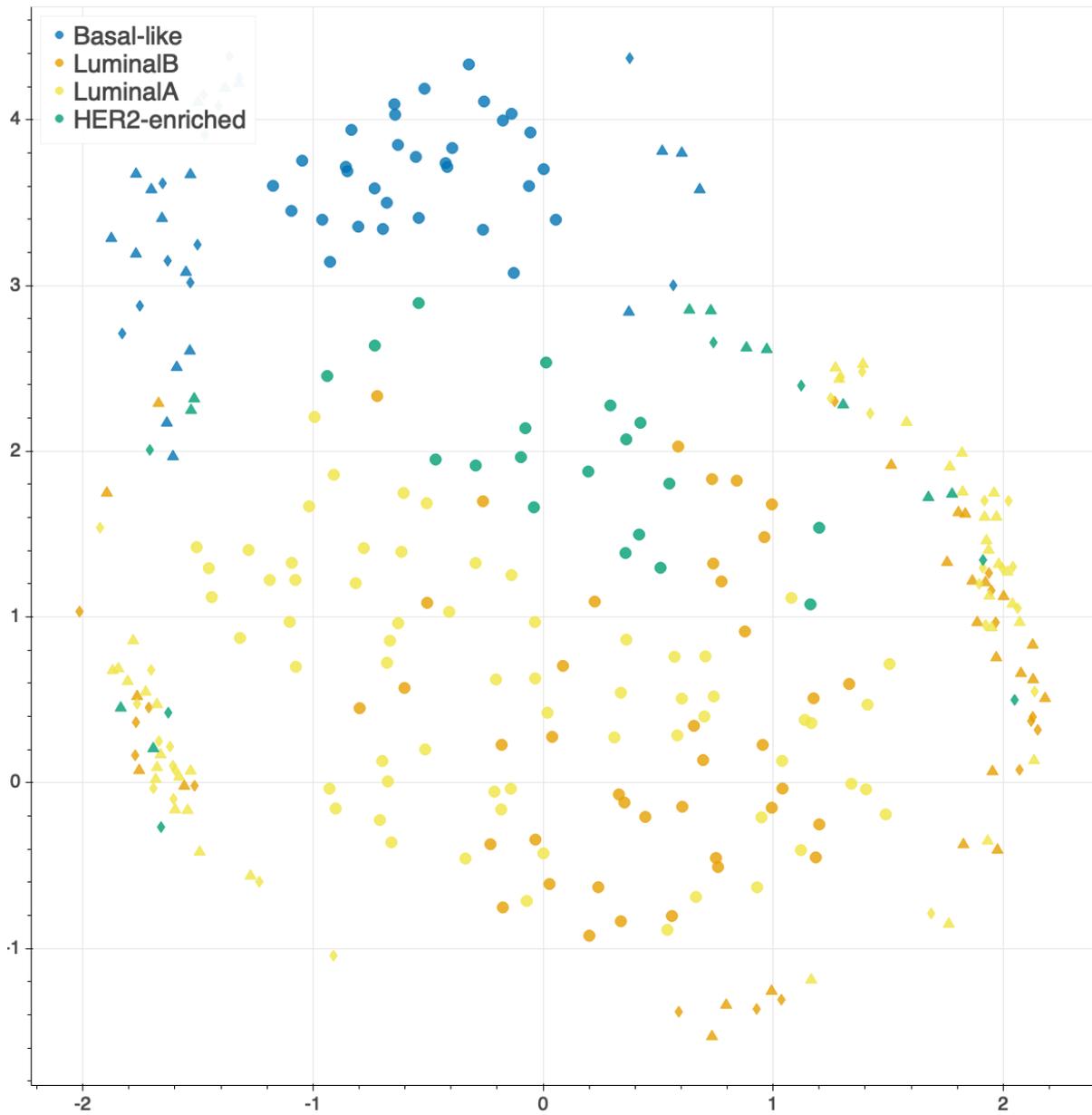


Figure 72: UMAP projection on the BRCA-subtypes task with 3-layer juxtaposed data. Circle: TR set; triangle: TS set; diamond: TS2 set.

11.4 Discussion

We present the INF framework for the characterization of cancer patient phenotypes by integrated multi-omics signatures, combining an improved version of a state-of-the-art integration technique [496] with predictive models developed inside a Data Analysis Plan [464] for machine learning. The framework is applied to TCGA data to predict clinically relevant patient phenotypes such as the overall survival or cancer subtypes.

The simplest approach for multi-omics data integration consists in juxtaposition of normalized measurements into one joint matrix, followed by the development of a predictive model. Juxtaposition-based integration is considered as a baseline technique, since it is the most naïve approach to combine two datasets; moreover, it enables to identify multi-omics signatures by borrowing discriminatory strength from information derived by all datasets. Juxtaposition further dilutes the already possible low signal-to-noise ratio in each data type, affecting the understanding of the biological interactions at the different omics levels.

Conversely, theour INF method for omics data integration is an improvement of the popular Similarity Network Fusion (SNF) approach [496], which has inspired several studies in the scientific literature, specifically in cancer genomics [79, 96, 225, 288, 365, 432, 486]. SNF maximizes the shared or correlated information between multiple datasets by combining data through inference of a joint network-based model, accounting for how informative each data type is to the observed similarity between samples.

Two innovative solutions have been implemented in this study: (i) we devised a SNF-based procedure to rank variables according to their importance in clustering samples with similar phenotypes; and (ii) predictive models were developed exploiting the SNF-ranked variables, inside a rigorous Data Analysis Plan which ensures reproducibility [464, 526].

The performance of INF was assessed both in terms of statistical properties as well as biological interest. Concerning the statistical aspect, INF was compared with pre-

dictive models developed on the juxtaposed datasets (juXT technique), as well as on the single-layer datasets. With INF, smaller signature sizes were systematically derived to achieve comparable or even better performance both in cross-validation and in test. This is an added value for INF, as biological validation of biomarkers can definitely benefit from signatures of small size in terms of both costs and required time. This main achievement is mainly due to the novel rSNF ranking, which increases the signal-to-noise ratio from the combined layers by prioritizing the most discriminant biomarkers in terms of network mutual information. rSNF exploits two main SNF advantages: integration of heterogeneous data and clustering of sample networks. The main peculiarity of the SNF integrative procedure is its robustness to noise [496], because weak similarities among samples (low-weight edges) disappear, except for low-weight edges supported by all networks, which are conserved depending on how tightly connected their neighborhoods are across networks. Moreover, the rSNFi step further increases the signal-to-noise ratio by training a predictive classifier on multi-omics juxtaposed data restricted to the top-ranked biomarkers shared by juXT and rSNF models. The resulting signatures are compact in size (up to 99% reduction w.r.t. juXT) while allowing predictive models to achieve equal or better performance compared to naïve juxtaposition or the single layers alone. While a comprehensive evaluation of the biological meaning of the signatures identified through the INF framework is beyond the scope of this work, we assessed their general validity with a thorough literature search. Our investigation shows that the signatures identified through the INF framework include biological markers that are relevant in the tasks under analysis and are consistent with previously published data. Further, as in [80], the largest contribution in the biomarkers' lists is provided by gene expression, while epigenomics, proteomics and miRNA transcriptomics play a minor role.

It should be noted that, especially in computational biology, multicollinearity between pairs of predictors and/or layers is intrinsic in the problem. Nevertheless, most machine learning models are indeed designed to identify the relevant predictors even in the presence of strong linear or nonlinear correlations, provided that an appropriate DAP, feature ranking method, and diagnostic tools (e.g. random labels)

are adopted against selection bias. To this aim, the application of a [DAP](#) derived from the MAQC-II initiative for model selection is a core attribute of the INF framework.

A fair comparison of INF results with other integration methods is currently unfeasible due to the number and variety of computational pipelines with dissimilar datasets, preprocessing methods, data analysis plans, and performance metrics.

This work is based on the original R implementation of the SNF algorithm [496]. However, we are aware that Open Source implementations exist in other programming languages, in particular *snfpy* for Python [301]. In a future release of the INF workflow, we plan to migrate the SNF-related parts to *snfpy* or a similar Python-based implementation, in order to drop the dependency on R and to potentially improve the overall performance.

In its current version, the INF framework supports the integration of two or more one-dimensional omics layers. As part of our future effort we will add support for the integration of medical imaging layers, for example leveraging the extraction of histopathological features from whole slide images by deep learning [45] or using radiomics or deep features from radiological images [47]. In both cases, further issues will emerge from the interactions between the omics and the non-omics data, needing particular care in the integration [283].

Conflict of Interest Statement

Author AZ was employed by the company NIDEK Technologies Srl. Author CF was employed by the company HK3 Lab. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Implementation and code availability

The full code of this benchmark is publicly shared on the GitLab repository <https://gitlab.fbk.eu/MPBA/INF>. Additional information is included in the Supplementary

Material available on the publisher's website, while the full set of experimental data can be accessed at <https://bit.ly/3ucB7a4>.

11.5 Appendix

Layer	# features	# informative features	Multiplicative factor	Class separation	Random state
Layer 1	100	10	default	1.0	1
Layer 2	50	5	default	1.2	2
Layer 3	250	25	10	0.8	3

Table 45: Synthetic data summary for each simulated layer. Multiplicative factor, class separation, and random state refer to the parameters `scale`, `class_sep`, and `random_state` of the `make_classification` function in *scikit-learn*.

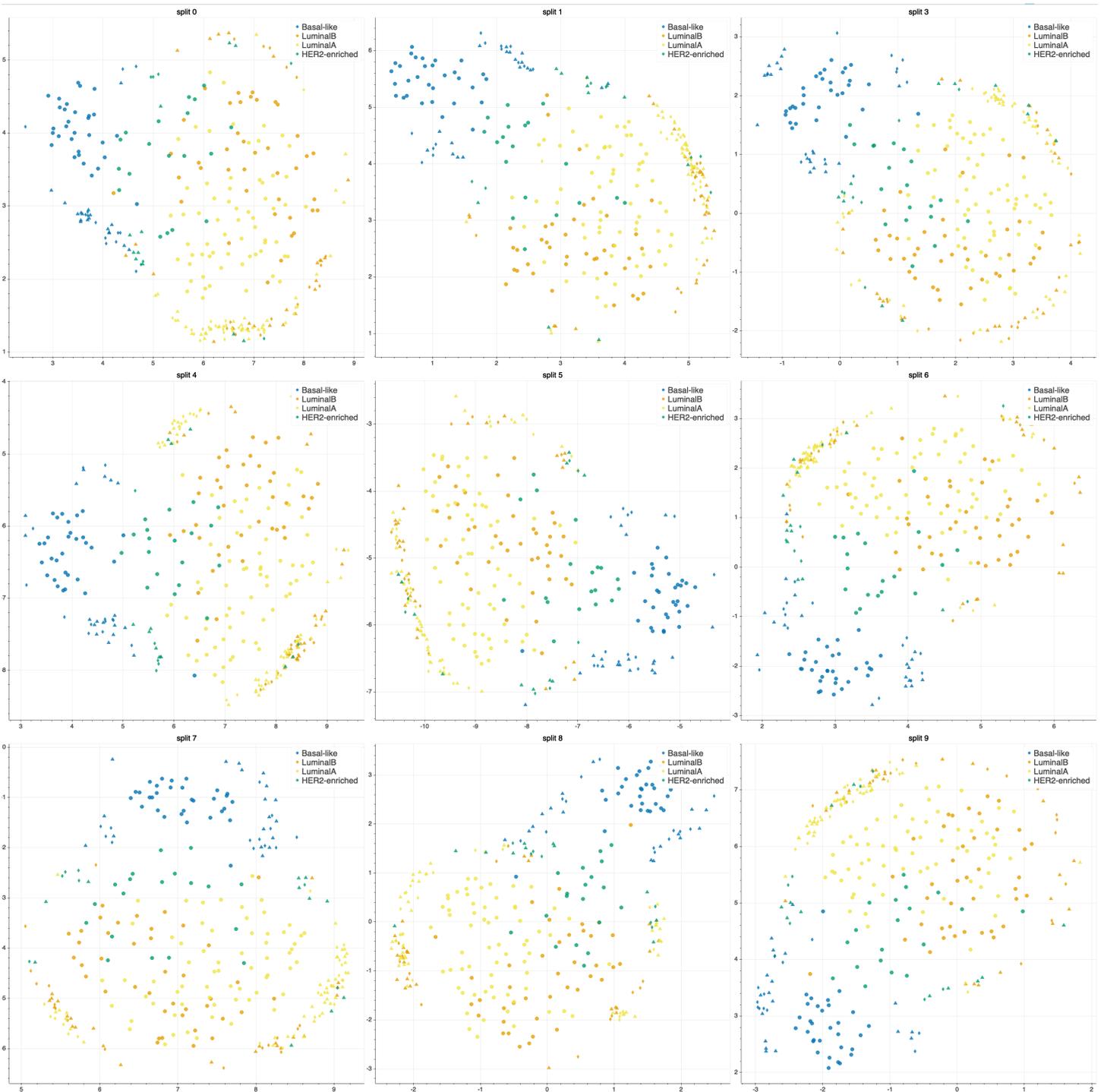


Figure 73: UMAP projections on the BRCA-subtypes task with 3-layer juxtaposed data. Each subplot represents the projection of the TR/TS/TS₂ partition for the remaining 9 splits not reported in the main text. Circle: TR set; triangle: TS set; diamond: TS₂ set.

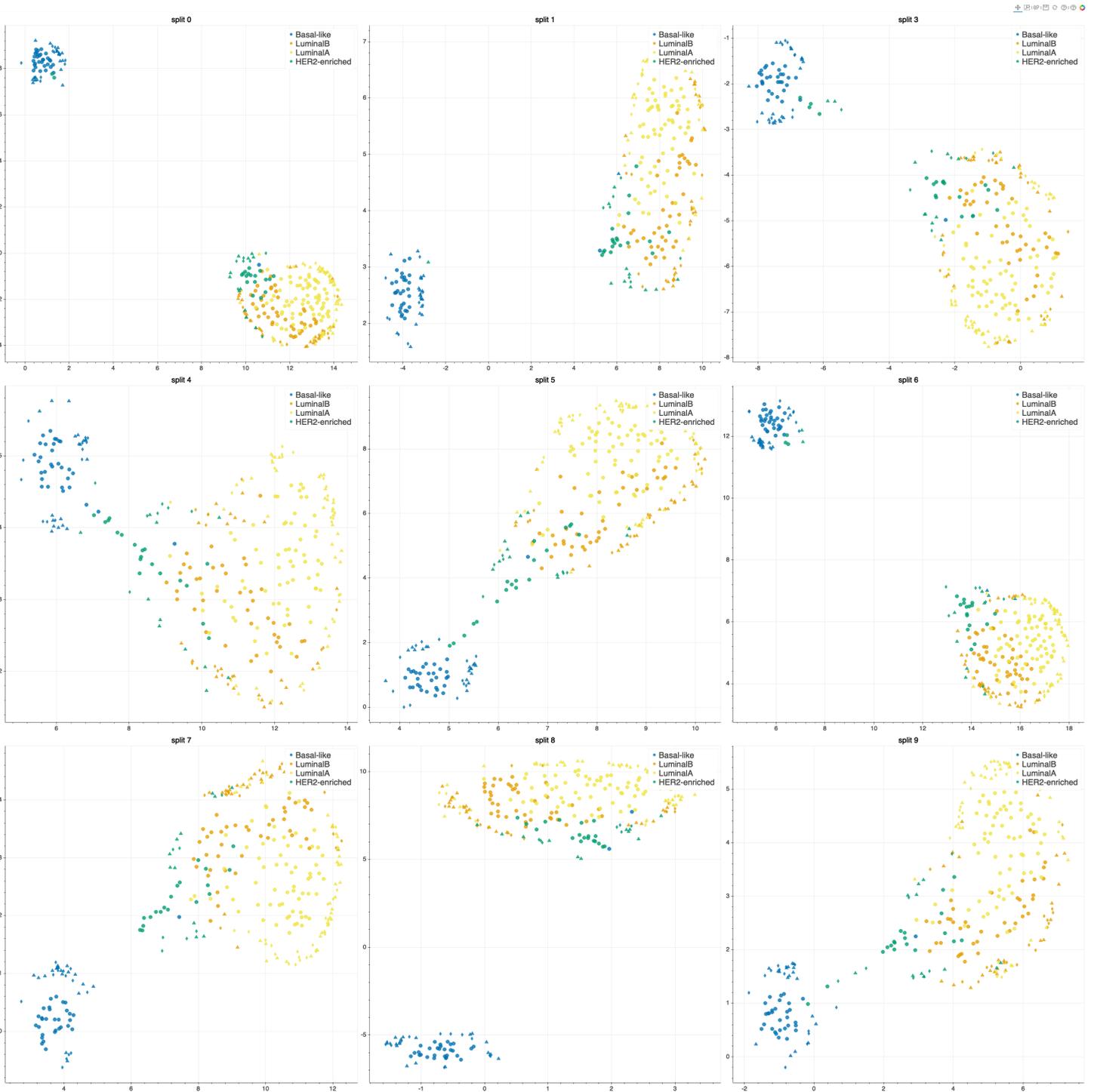


Figure 74: UMAP projections on the BRCA-subtypes task with 3-layer juxtaposed data restricted to the INF signature. Each subplot represents the projection of the TR/T-S/TS₂ partition for the remaining 9 splits not reported in the main text. Circle: TR set; triangle: TS set; diamond: TS₂ set.

Part VI

AI ON OPEN RESEARCH QUESTIONS IN TOXICOLOGY

Chapters 12 and 13 address open questions concerning AI applications in Toxicology. The PathologAI deep learning framework is introduced in Chapter 12 to identify lesion type and location on WSIs by weak label learning. The PathologAI framework is further applied in Chapter 13 to assess the advantages of combining molecular data with histopathology for improved biomarkers. The work presented in Chapter 12 and Chapter 13 is the result of a one-year long collaboration with Prof. Weida Tong's team (FDA / NCTR, Arkansas, U.S.), and HK3Lab (Rovereto, Italy). In particular, the PathologAI project is one of the initiatives promoted by the AI Research Force (AIRForce) in the Division of Bioinformatics & Biostatistics of FDA's National Center for Toxicological Research (NCTR).

CAN AI IMPROVE ON TOXICOLOGICAL PATHOLOGY REPORTS?

Highlights

- PathologAI is one of the first Deep Learning frameworks in preclinical pathology that operates in a weak-label setting.
- PathologAI exploits an unsupervised encoding network (BiGAN) to compress WSIs and preserve spatial information during model training.
- PathologAI is applied for DILI classification using $n=816$ WSIs from the TG-GATEs repository and generalized necrosis as pathologic endpoint.
- The HistoMAP approach is designed to pinpoint on original slides injury locations predicted by the model, further calibrated with pathologists' annotations.
- Although trained on extreme pathologic endpoints, PathologAI can accurately discriminate mild findings and dose level on external validation data. Notably, spontaneous lesions are differentiated from article test related injuries.

Personal contribution I am one of the main contributors of the PathologAI framework. In particular, I implemented the encoding network and run most of the classification experiments; performed deep feature analysis; contributed to the design of HistoMAP, and collaborated with the pathologists for review of original annotations.

12.1 The PathologAI weak-label framework

PathologAI is a multistage deep learning framework for digital pathology analysis of [WSIs](#) that can be effectively trained to identify lesion type and location by weak label learning, i.e. based on a label at slide level, without a previous manual annotations of [ROIs](#). The PathologAI framework is applied to predict Necrosis and Single Cell

Necrosis (Generalized Necrosis) on Liver data in [TG-GATEs](#) slides, on a total of $n=816$ [WSIs](#) (377 controls).

To boost reproducibility, PathologAI integrates methods for reproducible [WSI](#) pre-processing with deep-learning architectures evaluated within the Data Analysis Plan (See Chapter 2, Section 2.6.1.1). Based on `histolab` functions (see Chapter 6), datasets of non-overlapping square tiles are extracted from all the detected tissue. The PathologAI architecture for weak label training is a generalization of the work of Tellez *et al.* [457]. It is based on a combination of three main components (Figure 75):

An embedding map based on the deep learning Bidirectional Generative Adversarial Network ([BiGAN](#)) is trained to learn a low-dimensional representation of histological images (embedding vectors z) at tile level. Each tile is then compressed individually with the trained encoder, resulting in a dataset of N vectors. The embedding map is applied, obtaining a new image (actually a tensor) preserving the original locality structure of the tissue. The positions corresponding to the background or empty tiles are filled with zero vectors. Each compressed [WSI](#) is further packed by removing empty rows and columns and adding a padding of fixed size. The dataset with the transformed images is used to train a [CNN](#) classifier with the [WSI](#) label as target value. The HistoMAP mapping approach, based on the ScoreCAM (Score-Weighted Class Activation Mapping) [497] method is applied to localize lesions based on the activation of the fourth convolutional layer of the [CNN](#). Deep features from the last fully connected layer ($d = 128$) are analyzed with [UMAP](#) projection and [HDBSCAN](#) unsupervised clustering (see Chapter 2, Section 2.6.3.1).

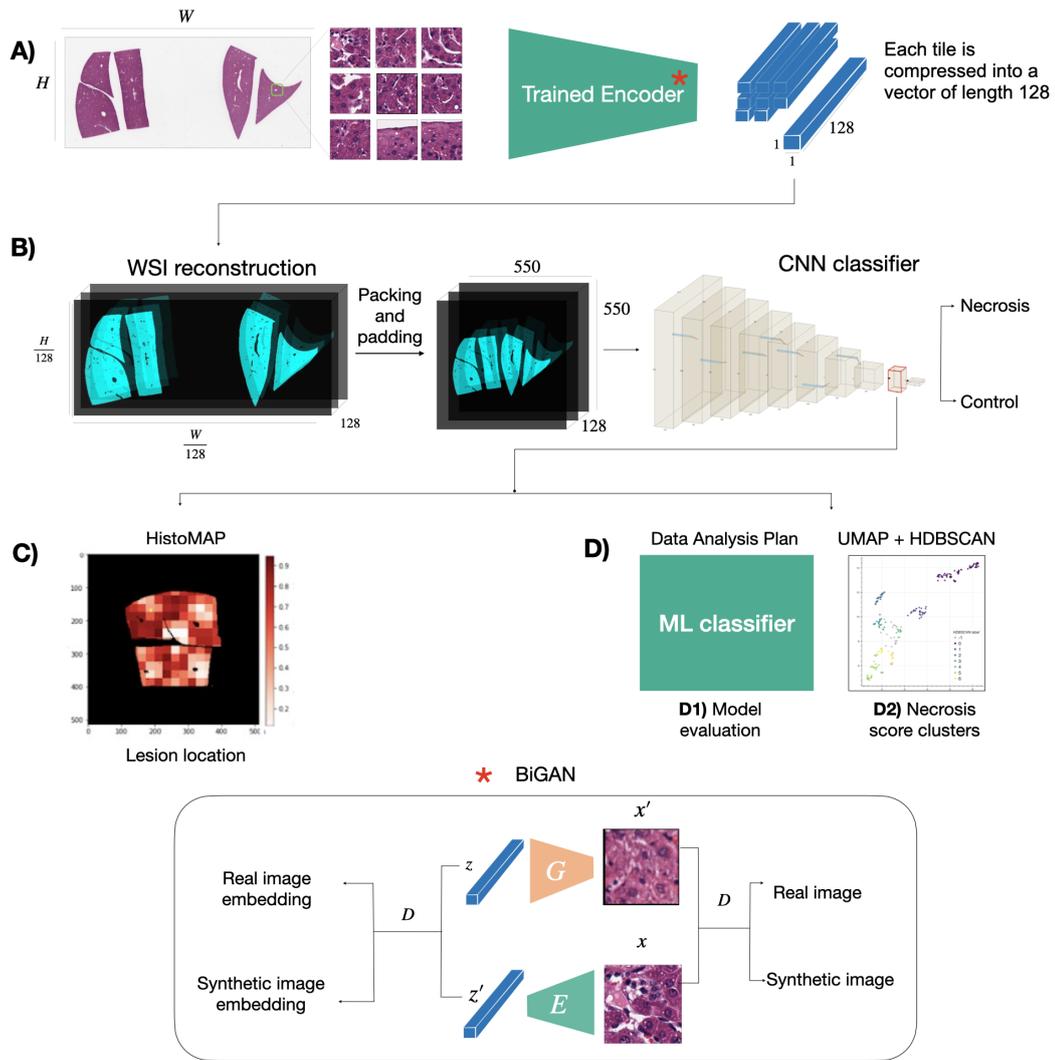


Figure 75: Structure of the PathologAI architecture experimental workflow. A) WSIs are pre-processed into tiles; each tile is compressed with the BiGAN encoder into a vector of length $L=128$. B). The embedding maps the WSI into a compressed WSI, preserving the original locality structure of the tissue. The transformed images are used to train a CNN classifier with the WSI label as target value. C). Output: (1) Locate the lesions on the slide (2) Provide severity score.

The proposed approach enables a re-analysis of the [TG-GATEs](#) toxicological data collection. PathologAI is applied to predict Necrosis and Single Cell Necrosis (Generalized Necrosis) on Liver data in [TG-GATEs](#) slides, on a total of $n=816$ WSIs (377 controls). The PathologAI system includes HistoMAP, a saliency mapping method for automated lesion location, where ROIs are prioritized as a preprocessor for human expert analysis. A validation applied to the [TG-GATEs](#) images analyzed with PathologAI has proven that injury identification is facilitated (pathologists can easily

pinpoint necrosis locations). Correct predictions have been validated, showing non inferiority to pathologists. Further, the PathologAI has been used to identify and revise misclassified samples. The method is general and it can be equally applied to other endpoints, which is critical to preclinical research.

12.2 Related works

12.2.1 *Weakly-supervised approaches for clinical tasks*

The most straightforward approach in a weakly-supervised framework in Digital Pathology is to independently evaluate tiles using global [WSI](#) labels, and then aggregate the predictions at slide level. This technique has been successfully employed in clinical tasks such as lung cancer subtyping on TCGA slides [109]. More sophisticated Multi-instance Learning (MIL) frameworks have been evaluated to avoid the need for effortful pixel-wise annotations on large [WSI](#) collections [65, 119]; in this setting, the slide is a single labeled bag comprising multiple instances (the tiles), which can be used to learn predictive features for the bag. MIL approaches have been tested on public datasets for oncological tasks, such as prostate cancer classification, Breast Cancer classification, metastasis detection, or molecular subtypes classification. Procedures based on tiling preprocessing require an optimal strategy for slide-level aggregation that might depend on the dataset or task addressed. Moreover, tile aggregation is based on the assumption that the [WSI](#) global label is dominant on the slide. To overcome these limitations, compression techniques to train [CNNs](#) on the whole [WSI](#) have been introduced. The StreamingCNNs [362, 363] are designed as memory efficient neural networks for megapixel images; StreamingCNNs retain operations computed on all tiles from a single [WSI](#) during [CNN](#) training thus leveraging information at slide-level while reducing memory usage. In the clinical setting, StreamingCNN has been evaluated on CAMELYON17 for metastases detection in lymph nodes, and on TUPAC16 for predicting a proliferation score based on gene expression. However, technical constraints in StreamingCNNs disrupt tech-

niques that can improve CNN training, such as batch normalization layers. Chen *et al.* [87] employed GPU memory optimization strategies to accelerate the processing of very large WSIs (> 20kx20k pixels) by downsizing by a factor of 4 without affecting model architectures or training procedures. Although this approach has been proven on TCGA datasets for lung cancer classification, it is best suited for easier tasks on lower magnification images.

PathologAI generalizes the compression NIC approach designed by Tellez *et al.* [457] that employs a DL-based feature encoding of tiles. The embedding vectors are then reassembled in a compressed WSI to recover the spatial relationships in the original slide; compressed images are finally used as input to a CNN classifier for whole slide classification. NIC has been evaluated in clinical histology for metastasis detection on CAMELYON16, TUPAC16, and Colorectal liver, and tissue type classification on a dataset of rectal carcinoma samples [456, 457]. In the original work, several embedding networks have been explored, including Variational Autoencoder (VAE), a contrastive training based discriminative model, and a BiGAN architecture.

12.2.2 AI in Preclinical Toxicologic Pathology

Several works have implemented weakly supervised AI techniques for toxicologic evaluation on preclinical pathology WSIs. Kuklyte *et al.* [254] explored popular CNN architectures in a multi-magnification approach for lesion segmentation on different organs. Notably, the authors performed an extensive curation of the TG-GATEs original data (650 hours effort, > 3600 slides reviewed, 1300 re-annotated). On 4319 Liver WSIs from TG-GATEs, improved necrosis segmentation results were obtained with a single-magnification model. However, a Slide-Wise train/test split protocol on tile collections was not applied, possibly incurring in the risk of data leakage (See Chapter 5). Pischon *et al.* [364] exploited a U-Net architecture to segment hepatocellular hypertrophy on liver samples from rats treated with Phenobarbital compound. A total of 34 HE-stained original WSIs from 8 different studies were used, and manually selected ROIs were annotated by pathologists. Selected ROIs were into 80-20 for model training, although the tile partitioning protocol is not discussed for data leak-

age. Mudry *et al.* [116] trained a VGG network to quantify rodent retinal toxicity by determination of nuclear layer thickness. In particular, manually annotated tiles of different severity grades were extracted from 112 *WSIs* for model training. Quantification of retinal thickness was accurate on the control group but partial results were obtained on diseased eyes. Tokarz *et al.* [466] designed a *DL* algorithm to segment and score severity of Progressive cardiomyopathy (PCM)-related abnormalities, classifying necrosis, fibrosis, Mononuclear Cell Infiltration, and mineralization. A total of 300 *HE*-stained *WSIs* were used for model development, manually annotated from veterinary pathologists. On the necrosis endpoint, a Spearman rank-order correlation $\rho = 0.82$ of model and median grade of five pathologists was obtained on the test set. However, results are possibly affected by data leakage as the tiling splitting procedure adopted for *WSI* preprocessing did not consider the slide's origin.

12.3 Imaging data

Digital images of the *TG-GATEs* collection were reviewed by expert pathologists and for each *WSI* pathological annotations are provided, including (*DOSE_LEVEL*): the dose level of the test article (Control, Low, Middle, High); (*SACRIFICE_PERIOD*): the sampling time of the tissue section (3hr, 6hr, 9hr, 24hr, 4 day, 8 day, 15 day, 29 day); (*SINGLE_REPEAT_TYPE*): the type of treatment exposure (single dose or repeat dose), (*FINDING_TYPE*): the type of histological lesion, (*TOPOGRAPHY_TYPE*): the region of the organ where the corresponding lesion was found; (*GRADE_TYPE*): the severity of pathological changes based on a 4-point scoring system [296] (0: within normal limits, 1: minimal, 2: slight, 3: moderate, 4: severe), and (*SP_FLG*): the spontaneity of corresponding lesion (True/False). Necrosis/Single cell necrosis (Generalized necrosis) were selected as target *DILI* and downloaded 559 slides from treated samples with the generalized necrosis among pathological findings, and 257 slides from control samples with no pathological findings identified. The *WSI* collection was partitioned into four main sub-datasets, according to the original *TG-GATEs'* annotations (Table 46):

- **Control-NF** (n=257): **WSIs** of liver tissue from non-treated rats, no pathological changes annotated.
- **Control-F** (n=120): **WSIs** of liver tissue from non-treated rats, generalized necrosis is found among pathological findings with slight/minimal **GRADE_TYPE**. Notice that all **acsWSIs** in this group have **SP_FLG=True** for the target lesion.
- **Mild-F** (n=324): **WSIs** of liver tissue from treated rats, generalized necrosis is found among pathological findings with slight/minimal **GRADE_TYPE**.
- **Positive-F** (n=115): **WSIs** of liver tissue from treated rats, generalized necrosis is found among pathological findings with moderate/severe **GRADE_TYPE**.

Notice that when more than one target finding was detected on a single slide (e.g. for different topography types), we considered only the annotation with the highest severity score. **WSIs** were selected from the four datasets to address the main tasks of our experimental pipeline, namely the **BiGAN** training (n = 102 **WSIs**), the Necrosis classifier development (n = 233 **WSIs**) and testing (n = 481 **WSIs**) (Figure 76).

Dataset name	Treated	Pathology finding	Severity score	Lesion spontaneity	Total WSIs (# TG-GATEs slides)
Control-NF	No	None	None	None	257 (5856)
Control-F	No	Necrosis/Single cell necrosis	1-2	True	120 (120)
Mild-F	Yes	Necrosis/Single cell necrosis	1-2	True	157 (292)
				False	167 (472)
Positive-F	Yes	Necrosis/Single cell necrosis	3-4	True	3 (3)
				False	112 (112)

Table 46: Summary of the datasets used in this work.

12.3.1 *BiGAN training data*

We selected a set of 102 **WSIs** including 50 **WSIs** from the Control-NF group, and 52 **WSIs** from the treated group with the target finding of generalized Necrosis (Necrosis/single cell necrosis). In particular, the control group included 25 **WSIs** from the single dose time point and 25 **WSIs** from the repeat dose time point; the treated group consisted of 50 Mild-F slides evenly distributed on the topography type of

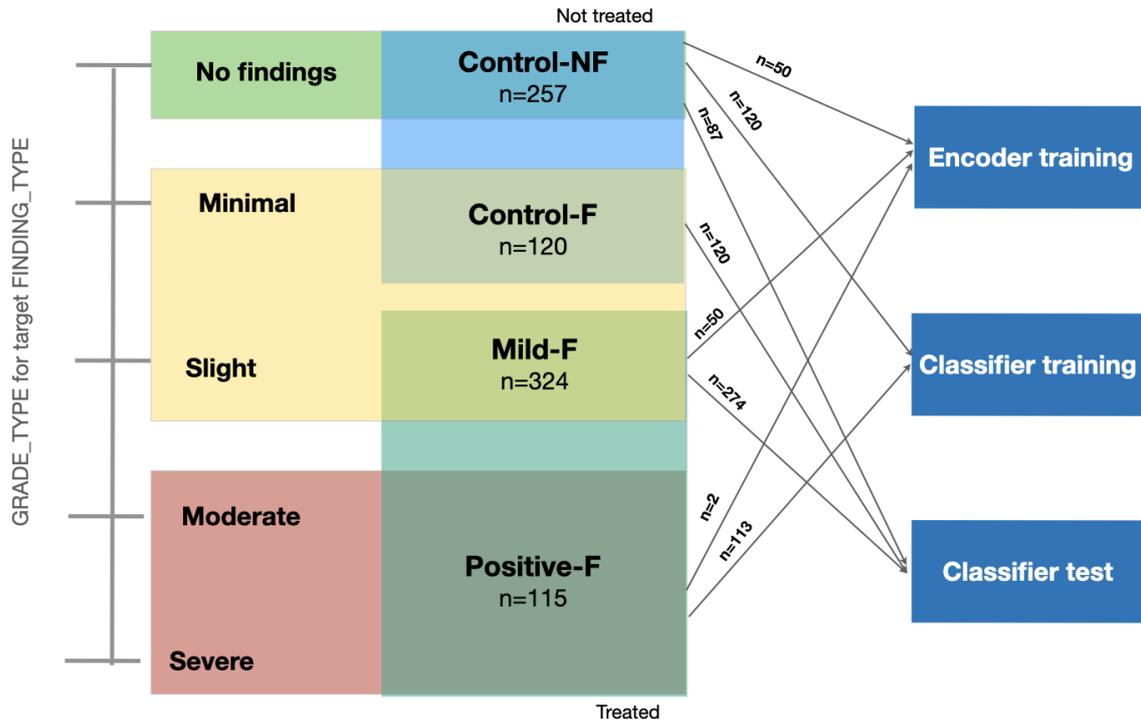


Figure 76: Main sub-datasets used in this work and corresponding task. Control-NF: control (DOSE=0) with no finding (NF). Control-F: control (DOSE=0), with target finding (F). Mild-F: treated (DOSE≠ 0), with target finding (F) and GRADE_TYPE slight or minimal. Positive-F: treated (DOSE≠ 0), with target finding (F) and GRADE_TYPE moderate or severe. For each sub-dataset, the number n of WSIs selected for each task is indicated.

the Positive-F data (Figure 77), and 2 Positive-F slides including one WSI from the single dose group with (SP_FLG, DOSE_LEVEL) = (True, Middle), and one WSI from the repeat dose group with (SP_FLG, DOSE_LEVEL) = (False, High). From the WSI collection, we extracted a total of 1.08M tiles of size 128x128 pixels used to train the BiGAN model (see Section 12.4).



Figure 77: Sunburst diagrams for FINDING_TYPE, GRADE_TYPE, TOPOGRAPHY_TYPE, and EXP_ID for (A) the BiGAN training data with findings (n=52), and (B) the Positive-F group (n=115). For each cell, (n) indicates the number of WSIs. Note: slides can be labelled for both SC Necrosis and Necrosis.

12.3.2 Classification data

The training dataset for the Necrosis classifier included 113 WSIs in the Positive group (the two WSIs used for the embedding network were excluded from the dataset), and 120 WSIs from the Control-NF group, for a total of 233 WSIs. The slides in the control group were selected to match (i) the 23 EXP_IDS of the positive group; (ii) the dose-time point (SACRIFICE_PERIOD) of the positive group, when available. Each slide was compressed with the trained BiGAN encoder (see Section 12.4) and the compressed slides were split into five training and validation sets corresponding to 80% and 20% of the total dataset, respectively. To avoid data leakage effects, train-

test splits were stratified by label, balanced for topography type (Figure 78), and segregated by experiment (no EXP_ID is simultaneously in the training and test sets). Moreover, all validation sets have empty intersections. The remaining 481 WSIs included the Control-F group ($n = 120$), The Mild-F group excluded the slides used for the BiGAN training ($n = 274$), and 87 WSIs from the Control-NF group with no EXP_ID in common with the training data. After compression and packing, these slides were used for model inference as external validation data.

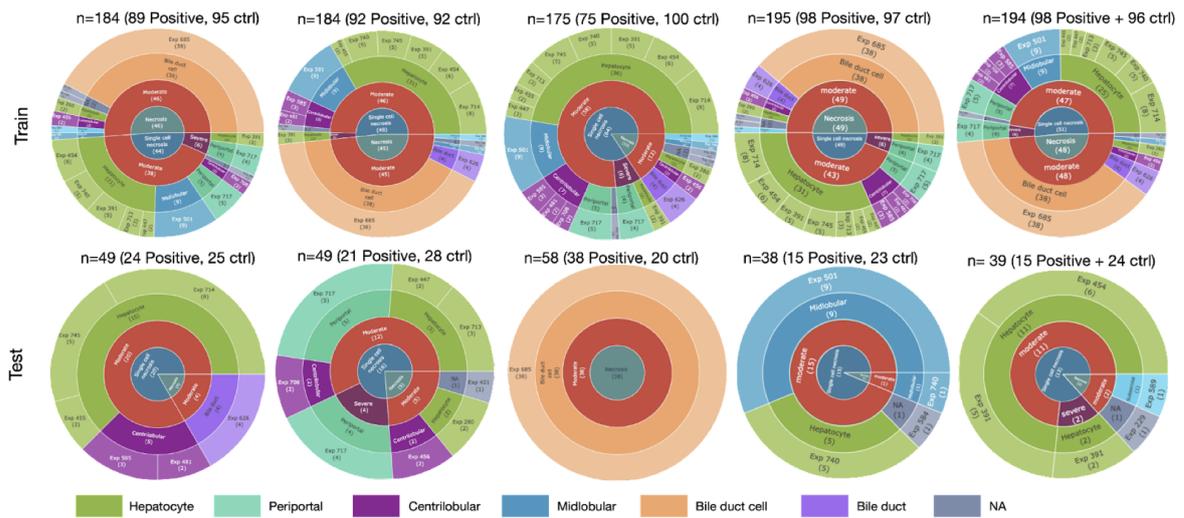


Figure 78: Sunburst diagrams for slides with findings in the train/validation splits of the 5 classifiers. For each diagram, FINDING_TYPE, GRADE_TYPE, TOPOGRAPHY_TYPE, EXP_ID and the number of slides (n) is indicated.

12.3.3 Tile extraction and preprocessing

BIGAN TRAINING. We extracted $\sim 10K$ non-overlapping tiles from each WSI of the Control-NF ($n=50$) and the Mild-F group ($n=50$) using the `RandomTiler` function of `histolab`, and 85000 tiles from the WSIs in the Positive group ($n=2$) using the `GridTiler` function of `histolab` (see Chapter 6). Tiles were extracted at 20x magnification (level 0) considering all the tissue detected on the slide (`TissueMask` function), and only the tiles with at least 80% of tissue were selected.

CNN TRAINING AND EXTERNAL VALIDATION. From each slide, N non-overlapping tiles covering all the tissue detected were extracted at level o ($\min(N) = 28656$, $\max(N) = 85104$, $\text{mean}(N) = 52299$), using the `GridTiler` function with `check_tissue` `True` to keep only the tiles with a minimum of 80% of tissue. All the extracted tiles were normalized with the Reinhard stain normalization method [383] using TCGA-A2-A3XS-DX1¹ as target image [14].

12.4 Unsupervised encoding network

We adopted the `BiGAN` architecture [128] as the unsupervised encoding strategy to learn the mapping between the images and the embedding space (see Appendix A.5). The `BiGAN` encoder had the best performance when used as feature extractor for classifying necrotic tissue among the six encoders (VAE, `BiGAN`, Siamese Network) compared in the work of Tellez *et al.* [457]. In our work, we fixed the dimension of the latent space to $C = 128$. To avoid instability and mode collapse during training, we implemented several strategies, including noise addition to both the real and synthetic data, one-sided label smoothing [405], spectral normalization [324], and two time-scale update rules [199]. Moreover, we exploited data augmentation techniques, *i.e.*, random horizontal flipping, random rotation with degree $\theta \in [-\pi, -\frac{\pi}{2}, 0, \frac{\pi}{2}]$, and we finally resized the input image to half its size before feeding the model. We minimized the binary cross entropy loss with the Adam optimizer with different learning rate (`lr`) for the discriminator (`lr` = 10^{-4}) and the generator (`lr` = 4×10^{-4}) and with fixed coefficients $\beta_1, \beta_2 = (0.5, 0.999)$. We trained the network with batch size 144 for a maximum of 2000 epochs, and selected the encoder corresponding to the epoch with the lowest Fréchet Inception Distance (FID) value [199], which measures the quality of generated images by computing the distance between the Inception-v3 activation distributions for real and generated samples. Examples of tiles generated with the `BiGAN` architecture are reported in Figure 79.

¹ <https://bit.ly/35yZDbf>

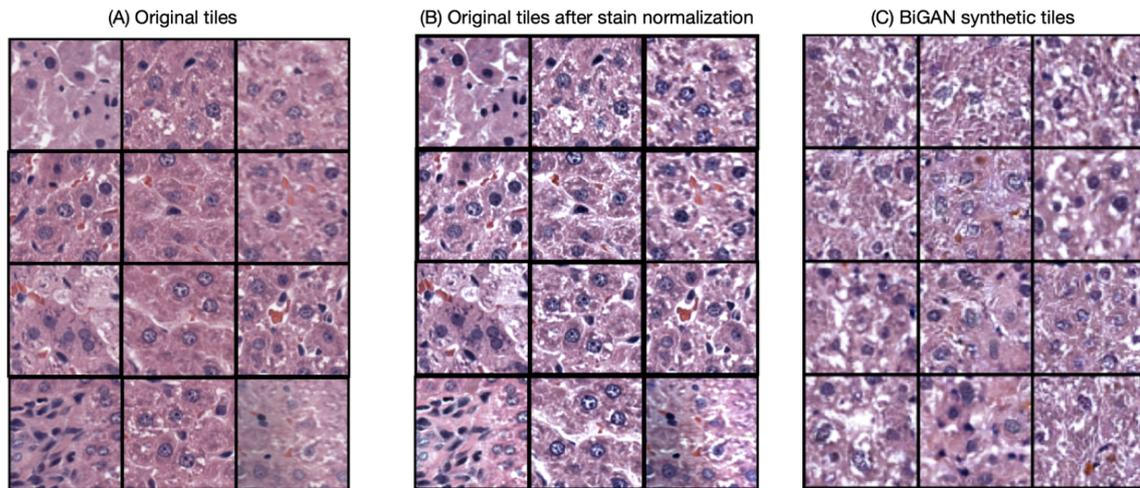


Figure 79: (A) Examples of tiles extracted from the BiGAN training dataset before and (B) after the stain normalization, and (C) examples of synthetic tiles generated by the trained BiGAN network.

12.4.1 Compression and packing

As a variation to the NIC approach [457], we included a "packing and padding" step during WSI compression to cope with the general case of multiple, distinct slices of tissue on a single slide. First, each tile was independently compressed by the trained BiGAN model, resulting in a set of N (N =number of extracted tiles) embedding vectors for each slide. To preserve the original spatial structure of the tissue in a slide S , we filled a matrix of size $\frac{H_s}{h} \times \frac{W_s}{w} \times c$, where $(H_s, W_s, 3)$ is the original size of S , $(h, w) = (128, 128)$, and $c = 128$ is the dimension of the latent space, with the N vectors so that the embedding of a tile at position (i, j) on S occupies the (i, j) -th position of the matrix. The remaining positions of the matrix, corresponding to the background of the slide or to tiles that were not extracted, were filled with empty vectors $0 \in \mathbb{R}^c$. As the slides in the TG-GATEs dataset include two or more slices of tissue, often separated by empty background, we further removed the empty rows and columns of the matrix and we added a padding of fixed size 532, obtaining a dataset of highly-compact WSIs ($\sim \frac{1}{10}$ of the original size).

12.5 CNN architecture

Similarly to [457], we implemented a CNN model comprising seven convolutional layers with 128 2-strided filters, batch normalization, leaky ReLU activation, 20% dropout, and stride 2, and a dense layer with sigmoid activation for the final classification. We applied data augmentation strategies on both training and validation set, including random crop to size 512, discrete random rotation with degree $\theta \in [-\pi, -\frac{\pi}{2}, 0, \frac{\pi}{2}]$, and random horizontal flip with $p = 0.5$. Moreover, we normalized the input channel-wise with mean and standard deviation (std) computed on the training set. Specifically, let $W = \{W_0, \dots, W_L\}$ be the set of compressed WSIs in the training set, where $W_i \in \mathbb{R}^{H \times W \times C}$ for $i \in [1, \dots, L]$, and let $W_{ic} \in \mathbb{R}^{H \times W}$ be the c -th channel of W_i with mean value μ_{ic} and std σ_{ic} . We computed the mean and std per channel as $\mu_c = \text{mean}(\{\mu_{ic}, \forall i \in [1 \dots L]\})$, and $\sigma_c = \text{mean}(\{\sigma_{ic}, \forall i \in [1 \dots L]\})$. We then map each $x_c \in W_{ic}$ to

$$\bar{x}_c = \frac{x_c - \mu_c}{\sigma_c} \text{ for each } c \in [1 \dots C] \tag{6}$$

We used Adam [543] as optimizer with learning rate 10-2 and we adopted the ReduceLR0nPlateau scheduler, thus lowering the initial learning rate by a factor of 10 if the validation loss had not decreased for 30 epochs (*patience*), up to the lower bound of 10^{-7} . We maximized the binary cross-entropy loss and trained the classifier with batch size 25 for a maximum of 500 epochs, saving the first three models with the highest MCC values on the training set.

12.6 Feature projection and clustering

We adopted the UMAP (See Chapter 2, Section 2.6.3.1 and Appendix A.1) for the analysis of the 128-dimensional deep features extracted from the last fully connected layer of the trained CNN classifier. To avoid the projection being affected by outliers,

features were first scaled according to the inter-quartile range computed on the training set (Robust scaler). We fit the [UMAP](#) algorithm on the training data using 40 neighbors and the Euclidean metric to construct the graph in the high-dimensional input space, and transform the test data ($n = 481$) on the learnt 2D manifold.

For the unsupervised clustering analysis of the external validation data, we adopted the [HDBSCAN](#) algorithm (See Chapter 2, Section 2.6.3.1). We first reduce the dimensionality of the scaled features down to 50 dimensions via the PCA algorithm fitted on training data (99.99% explained variance). To enhance clustering [10], we further used [UMAP](#) on the obtained 50-dimensional features to learn a 20-dimensional feature space where data were clustered via [HDBSCAN](#).

12.7 Lesion mapping

Inspired by the work of Wang *et al.* [497] we designed the HistoMAP approach to locate the decisive patterns for [CNN](#) classifiers on a compressed [WSI](#). First, the tissue portion in a compressed [WSI](#) is partitioned in a grid of 25×25 patches, corresponding to mid-size patches (3200×3200 px) in the original space. Each mid-size patch is then used as a region mask on the compressed [WSI](#). For each [CNN](#) classifier, the Necrosis Score is computed on the masked inputs, and the value assigned to the corresponding mid-size patch, resulting in a global mapping on the whole tissue region of the compressed slide. The HistoMAPs are then thresholded to scores between T_1 and T_2 quantiles, for each model, where T_1 and T_2 are uniquely selected during calibration (see Section 12.7.1). Notice that T_2 is needed to exclude maximal values due to artifacts (*e.g.*, pen marking). As the HistoMAP approach is applied on a single model, we designed a strategy to combine the maps of the base estimators for the ensemble model (Figure 80). In particular, the ensemble HistoMAP is obtained by combining the thresholded HistoMap of each model with the mean operator at mid-size patch level.

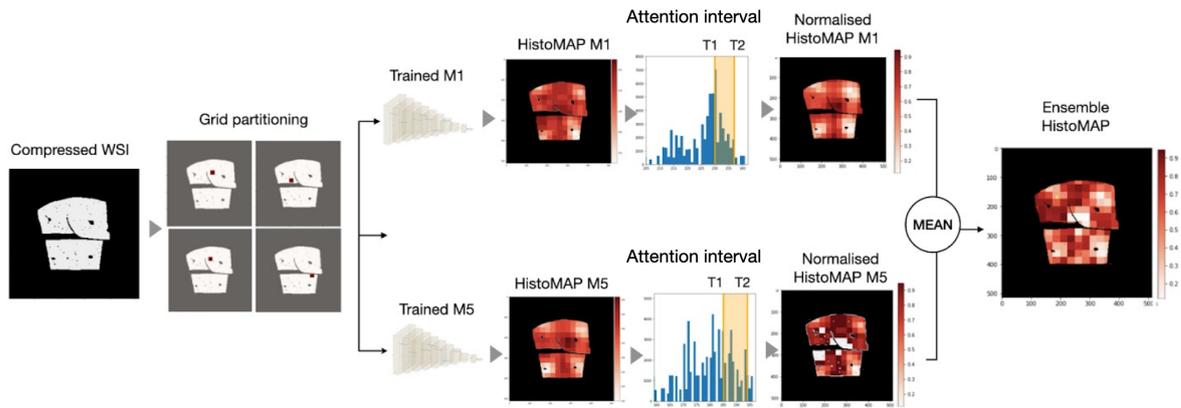


Figure 80: Workflow for automated lesion location with the HistoMAP mapping.

12.7.1 HistoMAP calibration with pathologists

The optimal T_1, T_2 thresholds for the HistoCAM mapping were determined on a grid of values based on pathologists' reports from TG-GATEs. Specifically, selected ROIs are collected from WSIs tested for 92 different drugs and accessible from the online repository, along with pathologists' revisions (Finding type and topography type). For the target endpoint of generalized necrosis, 23 WSIs (corresponding to 19 EXP_ID, and 17 compounds), and 33 annotated regions are available. We manually located and annotated 26/33 regions on the corresponding slides with the QuPath software [27]; six regions were excluded due to the lack of identifiable landmarks. Two expert pathologists (Dr. M. Cadei, and Dr. E. Villanacci) were asked to review the original annotations with respect to necrosis findings: 33 regions have been confirmed with consensus, and 14 were discarded (Figure 81). The confirmed regions were thus mapped back on the compressed slides and the corresponding WSIs were input to the histoMAP mapping. T_1, T_2 were finally selected to maximize the Necrosis Score on the confirmed annotations, namely $T_1=65$, and $T_2=96$.

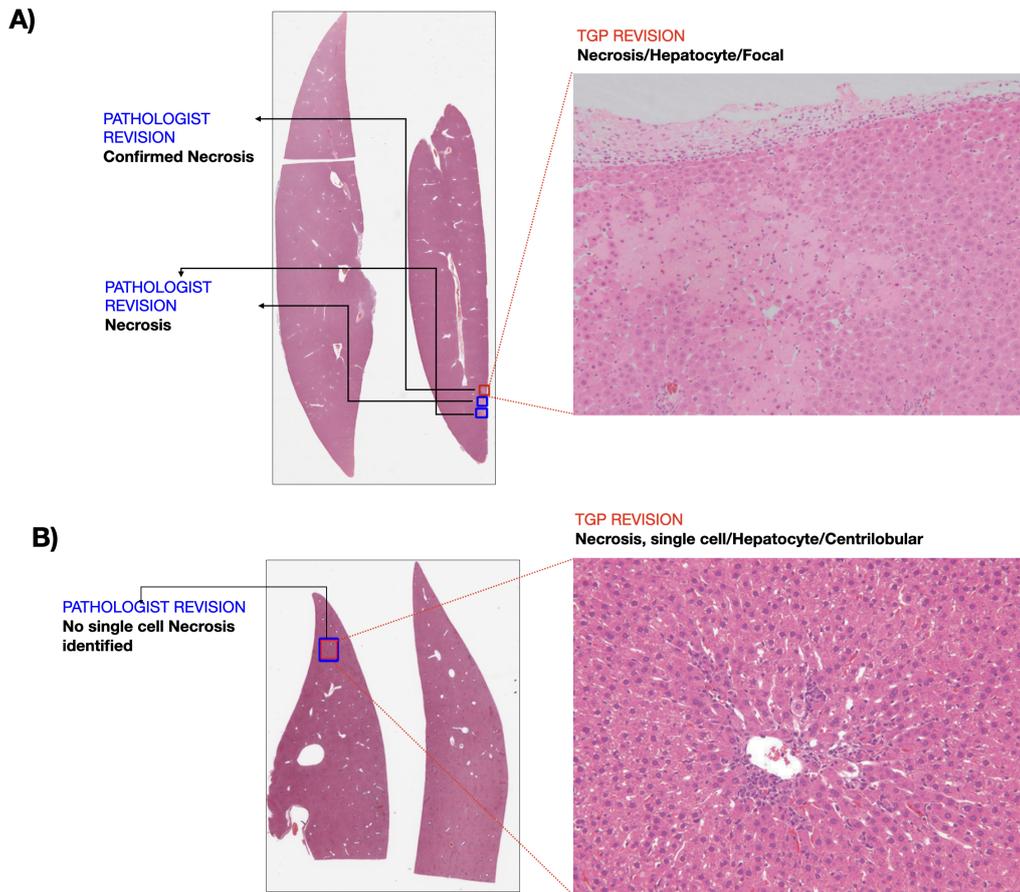


Figure 81: Examples of original TG-GATEs annotations and pathologist revisions. A) Original necrosis finding confirmed, B) Original necrosis finding not confirmed.

12.8 Experiments and Results

12.8.1 Necrosis Score

We defined the Necrosis score (NS) as the probability of the positive class output by the CNN classifier, $NS \in (0, 1)$. The retrieved NS provides a quantitative evaluation of lesion severity, possibly more robust than a semiquantitative grading system that may affect the quality of toxicological assessments [296]. We analyzed the NS distribution of the five classification models, *i.e.* trained on different splits, on the training set; results were visualized on the 2D UMAP projection of the 128-dimensional deep features extracted from the last fully connected layer of the network. Notably, the NS

gradually increases from Control-NF slides (grade 0) to Positive-F samples (grade 3-4), for all five models (Figure 12.8.1).

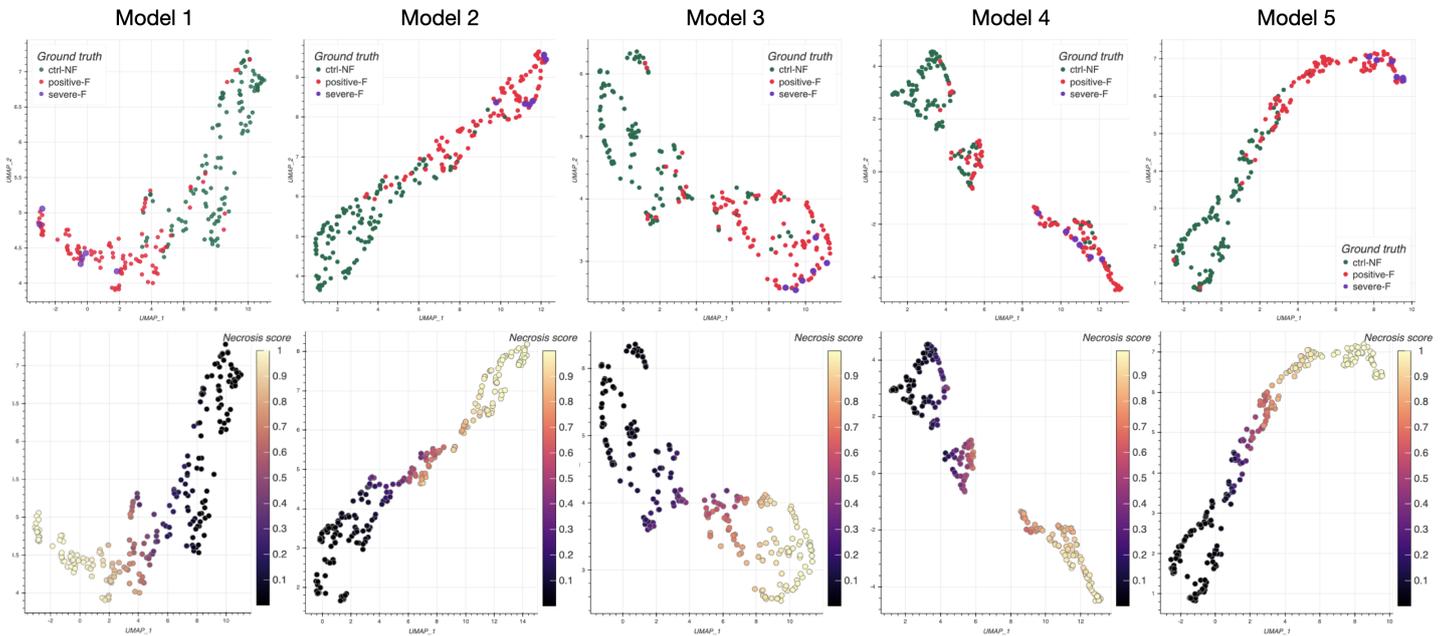


Figure 82: 2D UMAP projection of the deep features ($d = 128$) from the last fully connected layer of Model 1 on the training set ($n = 233$). Each point represents a compressed WSI colored by (first row) ground truth labels and (second row) by Necrosis score.

12.8.2 Data Analysis Plan and Ensemble model

Classification results were evaluated with a DAP for Machine Learning models (See Chapter 2, Section 2.6.1.1). In the DAP setting, features extracted from the last fully connected layer of the CNN model were input to a RF classifier (100 trees), after Robust Scaler normalization (see Section 12.6).

On the 5 internal validation sets, classification results varied from $\min(\text{MCC}) = 0.389$ on Model 3 to $\max(\text{MCC}) = 0.707$ on Model 1 (Table 47). Notably, 176 WSIs were correctly classified on all models, including the six positive slides with severe grade for generalized necrosis, and $n = 15$ WSIs were misclassified in at least 4 models (Figure 83). Lower performance in validation are obtained when the target lesion is found in different regions (TOPOGRAPHY_TYPE) than the training slides. In particular, Model 3 is strongly unbalanced due to a large EXP_ID = 685 (corresponding

MCC	Split 1	Split 2	Split 3	Split 4	Split 5	Mean MCC
Train (CI)	0.803 (0.782, 0.824)	0.801 (0.782, 0.824)	0.749 (0.714, 0.781)	0.699 (0.674, 0.723)	0.825 (0.807, 0.844)	0.775
Internal validation	0.658	0.707	0.389	0.587	0.679	0.604

Table 47: Training and internal validation results for the 5 base estimators of the ensemble model. CI: 95% bootstrapped confidence intervals.

to compound *methylene dianiline*), thus slides with annotated lesions in the bile duct cell are not learnt during model training (Figure 78). To overcome the variability of a single estimator, we adopted an ensemble model combining the predictions of the five classifiers; specifically, the NS per sample is defined as the probability of the Necrosis class output by the best RF model selected by DAP, for all the base estimators. Predicted labels are then computed by mapping the mean NS (mNS) of the base estimators to a binary value with a cut-off $t = 0.5$. As expected, MCC and sensitivity of the training set are improved for the ensemble model (Table 48).

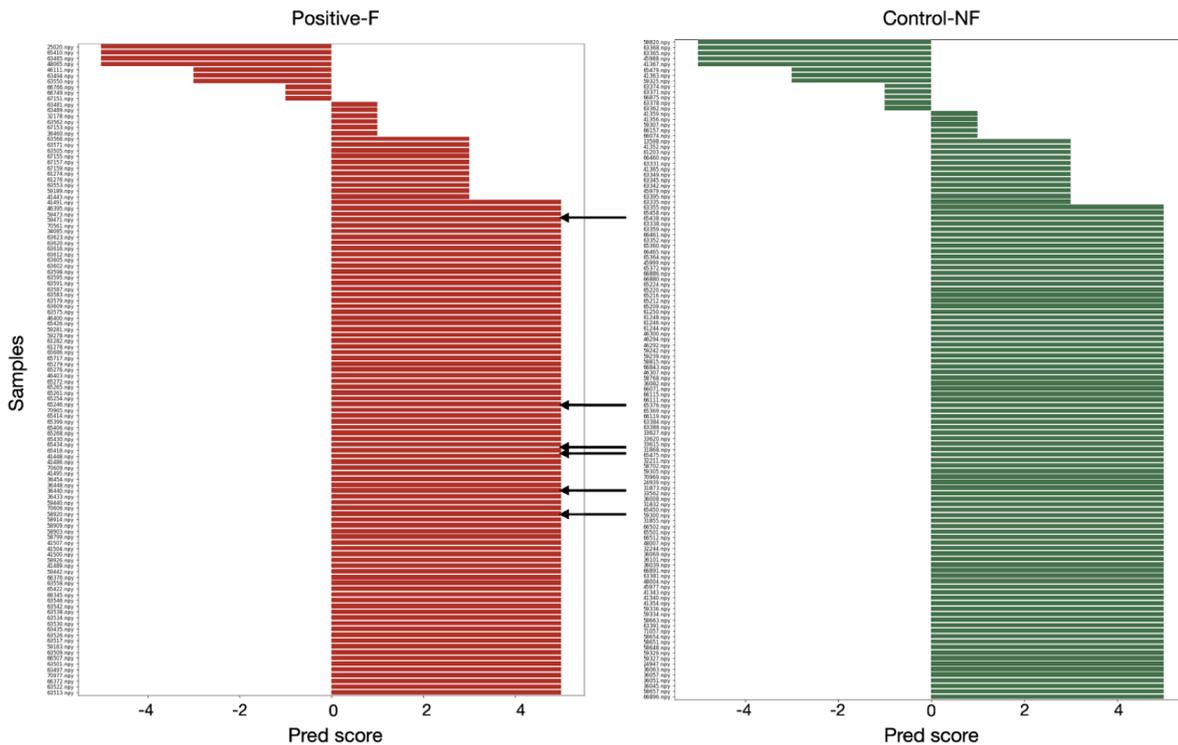


Figure 83: Predictions by sample for the 5 classification models. The Pred score is defined, per sample, as the difference between correct and wrong models. As splits have disjoint validations, exactly one prediction is on the internal validation set. Black arrows indicate samples with GRADE_TYPE=severe for generalized necrosis.

Model	MCC	Sensitivity	Specificity
Single	0.838	0.930	0.906
Ensemble	0.889	0.958	0.930

Table 48: Single and ensemble model results for the training dataset (n=233). Single model metrics are computed as the average of the metrics for the single estimators.

12.8.3 External validation

To evaluate the ability of the ensemble model to distinguish between spontaneous and treatment-related lesions, we further partitioned the Mild-F group into Mild-F slides with label SP_FLG=True (Mild-F-SP) and Mild-F slides with label SP_FLG=False (Mild-F-NSP). Remarkably, the mNS for test slides in Control-NF, Control-F, and Mild-F-SP groups distributes similarly to the Control-NF samples in the training set, while the mNS is higher for the Mild-F-NSP group (Figure 84). We quantitatively assessed the performance on the external validation data by considering the positive class (Necrosis) as ground truth label for the Mild-F-NSP data, and the negative class (Control) for the Control-NF, the Control-F, and the Mild-F-SP groups (Table 49).

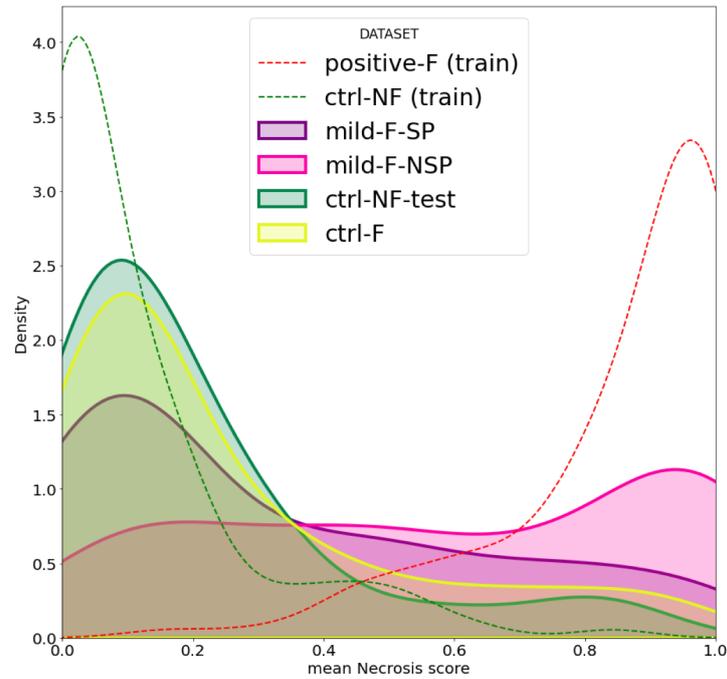


Figure 84: Univariate distribution of mNS on training data (dashed: Positive-F and Control-NF) and external validation data (solid: Ctrl-F, mild-F, and ctrl-NF-test), with Mild-F data separated into spontaneous and nonspontaneous lesions (SP: spontaneous; NSP: non spontaneous).

Dataset	# WSIs	Label	ACC (%)
Control-NF (test)	87	Control	89.7
Control-F	120	Control	83.3
Mild-F-SP	141	Control	73.1
Mild-F-NSP	133	Necrosis	57.9

Table 49: Ensemble model accuracy on external validation datasets.

Moreover, the ensemble model classified Control-NF vs MILD-F-NSP slides with $MCC=0.476$, and mild spontaneous (Mild-F-SP) vs mild treatment-related (Mild-F-NSP) injuries with $MCC=0.313$. The analysis of the deep features extracted from the last fully connected layer of the base estimators indicates that external validation data lie on the manifolds learnt by the classifiers with Control-NF and Control-F test samples embedded close to the Control-NF training data (Figure 85).

Notably, the mNS distribution on the external validation data demonstrates the PathologAI framework, although developed only on extreme endpoints, *i.e.* high-

severity necrosis and control samples with no pathology findings, to be effective in (i) predicting the severity level of the pathology finding, even when minimal (ii) differentiating sample treated with diverse dosage (iii) discerning between spontaneous and treatment-related injury (Figure 86).

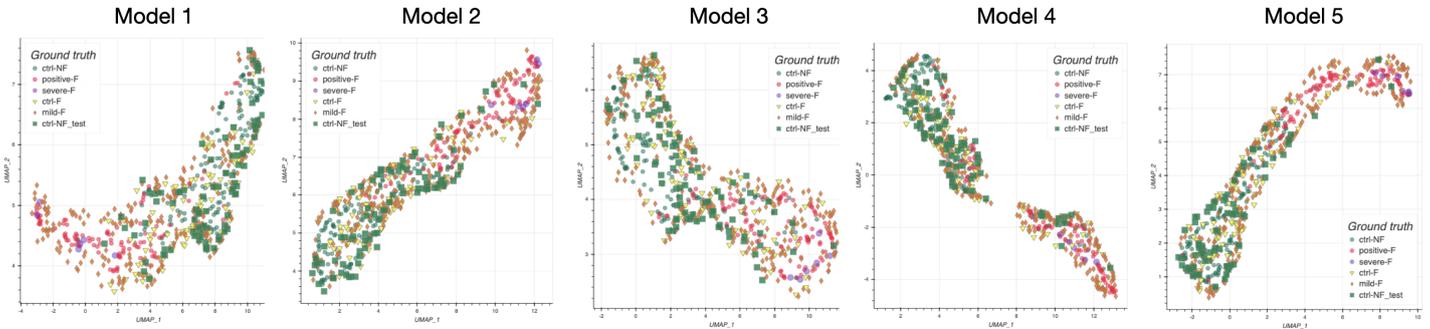


Figure 85: 2D UMAP projections of the 128-dimensional deep features of the five base estimators on all data (n=714) colored by ground truth labels.)

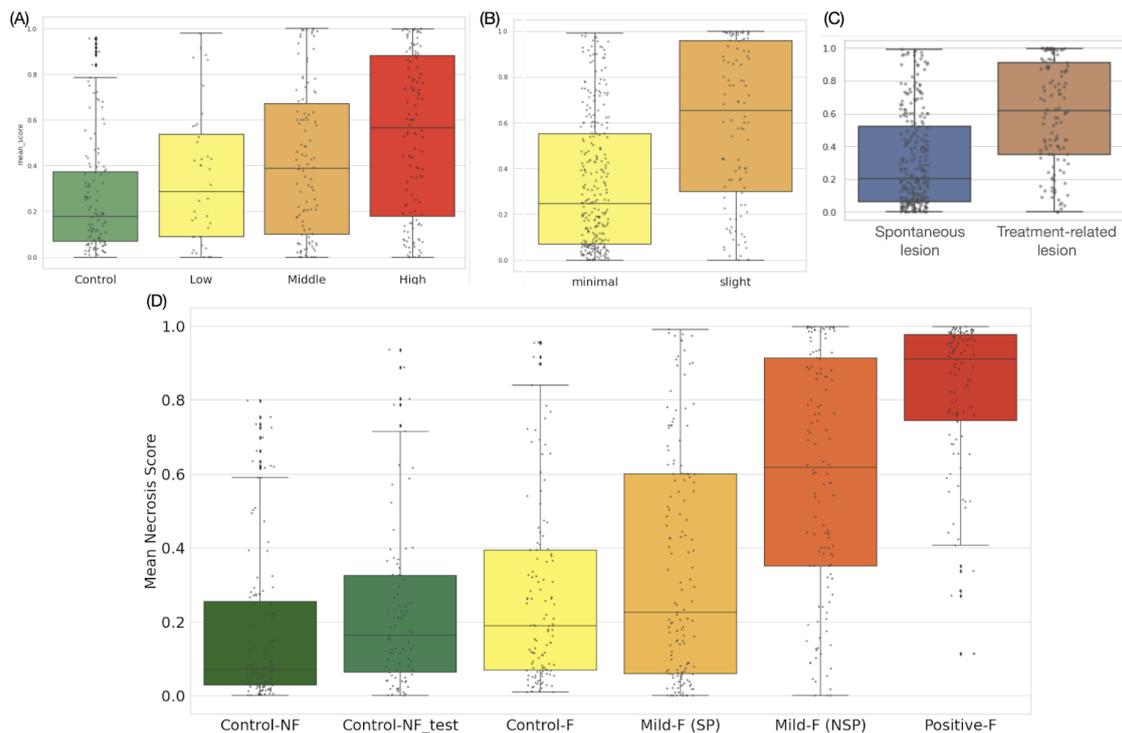


Figure 86: (A,B,C) Boxplots of mean Necrosis score distribution on the external validation data grouped by (A) dose level; (B) severity grade of target finding, and (C) SP_FLG for target lesion. (D) Boxplots of mean Necrosis score on classifier development and external validation data grouped by datasets. SP: spontaneous; NSP: non spontaneous.

IMAGING AND OMICS COMBINATION

Highlights

- A Machine Learning pipeline for DILI prediction is developed on gene expression data in the intersection with PathologAI imaging datasets (Chapter 12).
- Omics features are highly accurate for liver necrosis prediction; mean MCC=0.871 is obtained on internal validation for the five base estimators.
- Overall, predictive models based on Imaging features and selected genes combination did not improve classification results over single modalities.
- Imaging and MAS5-normalized omics features show a higher consistency with respect to original pathologists' annotations.

Personal contribution I designed the Machine Learning pipeline for GE data and performed all the experiments on omics and combined (imaging+omics) data.

The PathologAI framework described in Chapter 12 aims at extracting final phenotypic readings from pathology *WSIs* in preclinical applications, possibly reaching human level performance. Despite the methodological and ethical limitations of *in vivo* testing, the histopathology assessment is still considered the gold standard to assess potential adverse effects in drug safety studies [226]. However, several works have explored alternate *in silico* models to complement, or substitute, animal based assays for the early detection of drug-related hepatotoxicity (see for example Chapter 9). In particular, gene expression analysis has been evaluated to detect underlying biological mechanisms that can precede histological manifestations [226, 276]. For example, Smith *et al.* [429] tested different *ML* classifiers to predict *DILI* on gene expression data from the *TG-GATEs* collection; they identified a compact signature of ten genes predictive of Liver necrosis. Recently, the FDA/NCTR team proposed the ToxGAN

architecture to completely simulate *in silico* the GE profiles for different compounds, dosage and time points from TG-GATEs [92].

To investigate whether the integration of Imaging and Omics data can provide better safety biomarkers in toxicology, we first developed a computational pipeline on gene expression data using the same endpoint used for the PathologAI framework, *i.e.* Liver generalized necrosis. The predictive power of imaging and omics combination is then evaluated with an early integration strategy (see Chapter 2, Section 2.6.4.1).

13.1 Gene expression data

Gene expression (GE) data were retrieved by the FDA/NCTR team from the TG-GATEs repository, and processed either with Robust Multiarray Averaging (RMA) or Microarray Analysis Suite 5 (MAS5) normalization. Although techniques like FARMS [202] or CrossNorm [93] have been proven more efficient/accurate than classical methods [281], MAS5 and RMA were used as baseline preprocessing for a direct comparison with results recently published on the use of GAN methods from the same TG-GATEs gene expression dataset [92]. Specifically, The FDA/NCTR team proposed an AI approach alternative to animal experiments (Tox-GAN) and a case study on TG-GATEs.

For comparability with the imaging model, only samples in the intersection with the datasets described in Chapter 12, Section 12.3 were considered to develop the Machine Learning pipeline (Table 50). Notice that GE data were available for ~ 58% of the imaging samples. Moreover, the five splits stratified for experiment and topography detailed in Chapter 12, Section 12.3.2 were also used to develop corresponding predictive models on the GE data.

Dataset	GE+WSI	WSI only
Total	418	713
Control-NF (train)	74	120
Positive-F (train)	76	113
Control-NF_test	41	86
Control-F	67	120
Mild-F-SP	81	147
Mild-F-NSP	80	127

Table 50: Summary of data for the necrosis classifier: GE+WSI: samples with both omics and imaging data; WSI: samples with only imaging data available.

After normalization, dimensionality reduction was performed on each GE feature vector ($d = 13788$) to obtain 128-dimensional vectors; the output dimension was selected to match the dimension of the deep features extracted from the imaging model (see Chapter 12, Section 12.4).

In particular, two sequential steps were considered for dimensionality reduction: (i) Removal of correlated features, which discards one of the two features with a high Pearson’s correlation coefficient ($\rho > 0.95$), and (ii) Univariate analysis, which computes an association score (ANOVA F-test) between each feature and the target label. Features are then ranked based on the association score, and only the top $K = 128$ features are kept. To avoid selection bias, feature (gene) selection was implemented, for each split, on the training data, and the corresponding genes were selected on internal and external validation data.

13.2 Machine Learning pipeline

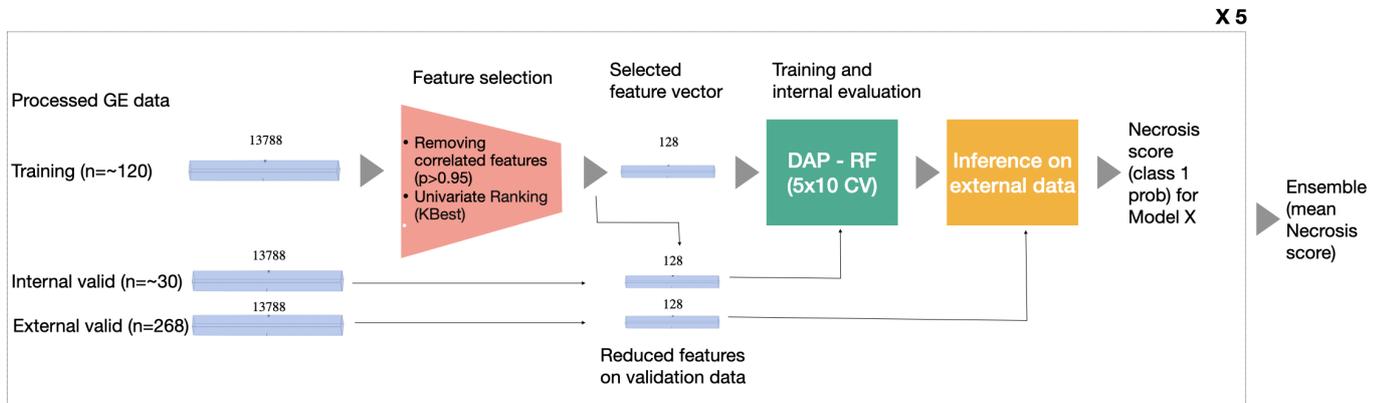


Figure 87: Workflow of the Machine Learning pipeline to predict Liver necrosis from GE data. After data partitioning and dimensionality reduction, a RF classifier is trained on the five splits within the DAP. Results on the external validation datasets are evaluated by an ensemble model.

The Machine Learning pipeline implemented for Necrosis prediction on GE data is illustrated in Figure 87. After reduction from 13788 to 128 on all datasets by KBoost univariate ranking, the five splits adopted in the histopathology study are considered for the intersection data. For each split, a RF model is trained within a DAP for predictive models (see Chapter 2, Section 2.6.1.1). Results of the best model selected in the DAP were evaluated in terms of MCC on the five internal validation sets and on the external validation data. An ensemble model for omics data was also adopted to evaluate predictions on the external validation data, as for the imaging framework (Chapter 12, Section 12.8.2); in particular, the probability of the class Necrosis (Necrosis Score) for the best RF classifier is computed for each of the five models, and the mean value (mean Necrosis Score) is thresholded ($t = 0.5$) to obtain the predicted label.

13.3 Experiments and Results

13.3.1 Omics data

The Machine Learning pipeline was applied to predict liver Necrosis on the five train/validation splits, by using both RMA and MAS5 normalized GE data. Similarly to the imaging only framework (Chapter 12, Section 12.8.2), results on the unbalanced split 3 are significantly worse with respect to the other models, regardless of the normalization method (Table 51, and Table 52). Overall, performance of RMA and MAS5 normalization are comparable, with narrow CIs and high predictive ability (MCC > 0.663 on internal validation data).

Training	Split 1	Split 2	Split 3	Split 4	Split 5	Mean MCC
Train (CI)	0.921 (0.903, 0.939)	0.950 (0.934, 0.967)	0.966 (0.954, 0.976)	0.931 (0.912, 0.949)	0.951 (0.935, 0.968)	0.944
Internal validation	0.935	0.889	0.663	1.0	0.921	0.882

Table 51: Training and Internal validation results on the five splits for the machine learning pipeline on omics data normalized with the RMA method. CI: 95% studentized bootstrap confidence interval

Training	Split 1	Split 2	Split 3	Split 4	Split 5	Mean MCC
Train (CI)	0.942 (0.924, 0.959)	0.949 (0.933, 0.965)	0.947 (0.934, 0.961)	0.928 (0.910, 0.947)	0.931 (0.909, 0.951)	0.939
Internal validation	0.935	0.837	0.663	1.0	0.921	0.871

Table 52: Training and Internal validation results on the five split for the machine learning pipeline on omics data normalized with the MAS5 method. CI: 95% studentized bootstrap confidence interval

To identify a compact gene signature for liver necrosis, we computed the Borda aggregation of the Borda lists [227] on the 10x5 replicates in DAP for the 5 splits, resulting in a ranked list of genes estimated on 250 experiments. The top 10 genes for each normalization method are reported in Table 56. Notably, the RMA and MAS5 normalization approaches produce different ranked lists that intersect on only one gene on the top 10 positions. In particular, the top-1 ranked feature for the RMA normalization is *Bmf* (Bcl2 Modifying Factor): this gene belongs to the BCL2 family,

and acts as an apoptotic activator¹. The MAS5 normalization identifies CXCL12 (C-X-C Motif Chemokine Ligand 12) as top-1 gene for necrosis prediction: CXCL12 encodes for a protein that is involved in immune surveillance, inflammation response, tissue homeostasis, and tumor growth and metastasis². Notice that the top-1 feature for both lists has a mean position close to 0, indicating that the corresponding gene is selected first for most experiments.

Table 53: RMA normalization

	Gene	mean_pos
1	Bmf	0.284
2	Adamts7	5.480
3	Ctsl	6.316
4	Pprc1	7.908
5	Tnfrsf12a	8.305
6	Rbm3	8.968
7	Mcl1	9.068
8	Klf	11.920
9	Enc1	13.404
10	Btg3	13.460

Table 54: MAS5 normalization

	Gene	mean_pos
1	Cxcl12	0.700
2	Nrep	6.188
3	Ndrp2	6.520
4	Bmf	6.764
5	Slc19a2	8.112
6	Ptprd	8.764
7	Pign	9.216
8	Nfib	9.808
9	Kmo	12.784
10	Slc22a8	16.8

Table 55: Total Borda lists for (A) RMA normalization, and (B) MAS5 normalization. Mean_pos: average position of feature on 250 experiments (50 DAP replicates for 5 models)

On the external validation data, results of the ensemble model are slightly improved for the RMA normalization (Table 56). Notably, the Machine Learning framework predicts all non-treated sample with spontaneous lesion (Control-F) as control, regardless of the normalization technique.

¹ provided by RefSeq, Jul 2008

² provided by RefSeq, Sep 2014

Accuracy on External Data	Label	GE (RMA)	GE (MAS5)
Control-NF test	0	0.878	0.853
Control-F	0	1.000	1.000
Mild-F SP	0	0.788	0.700
Mild-F NSP	1	0.725	0.712

Table 56: Predictive results on the external validation data for the ensemble model applied on GE data with RMA or MAS5 normalization.

13.3.2 Omics and Imaging combination

The PathologAI framework aims to extract the final phenotypic readings from pathology [WSIs](#), possibly reaching human level performance. Human pathologists can be incorrect, or inconsistent but histopathology evaluation based on human experts is still the gold standard to assess phenotypic endpoints. The gene expression data come from a different category; a GE based model tries to find any relationship between genotypes and phenotypes. The goal of a GE model is to discover any molecular explanation for the phenotypic prediction. Such signatures can also be used to predict the phenotypic endpoint when phenotypic measurements are not available (Joshua Xu, personal communication). Both PathologAI and the GE model were trained with extreme labels, Positive-F (treated with non-spontaneous and high-risk injuries) and Control-NF (controls with no pathology findings) samples. Moreover, imaging and GE were able to predict samples with non drug related lesions, despite no spontaneous necrosis was reported in the [TG-GATEs](#) findings for the training set.

To evaluate the benefits of an integrative model for necrosis prediction, we implemented an early-integration strategy that combines selected omics features ($d_O = 128$), and imaging deep features extracted by the PathologAI model ($d_W = 128$), in a unique 256-dimensional vector per sample. The retrieved feature vectors are then used as input to [RF](#) classifiers in the [DAP](#) environment. Following the protocols implemented for the imaging-only and GE-only models, we adopted the ensemble approach to evaluate predictive performance on the external validation datasets (Table 57).

Accuracy on External Data	Label	WSI	GE (RMA)	GE (MAS5)	GE (RMA) + WSI	GE (MAS5) + WSI
Control-NF test	0	0.902	0.878	0.853	0.951	0.951
Control-F	0	0.866	1.000	1.000	0.985	0.955
Mild-F-SP	0	0.700	0.788	0.700	0.800	0.775
Mild-F-NSP	1	0.660	0.725	0.712	0.663	0.688

Table 57: Comparison on external validation datasets of predictive models trained on imaging only (WSI), normalized gene expression only (GE), and integration of imaging and gene expression data (GE+WSI). RMA: Robust Multiarray Averaging normalization; MAS5: Microarray Analysis Suite 5 normalization

Overall, the integration of imaging features and GE signatures into a single ML model did not significantly improve predictions of the single modalities. We further investigated if GE and imaging models exhibit a higher consistency than human pathologists in discerning spontaneous and drug-related necrosis in difficult cases. In particular, we compared the predictive performance on the classification task of Mild-F-SP vs Mild-F-NSP of models trained on GE using as target labels predicted by PathologAI, and *viceversa*. Notably, the imaging-only model on labels predicted by GE improves the performance with respect to the original labels in the [TG-GATEs](#) pathology reports (Table 58).

Features	Target labels	MCC (GE=RMA)	MCC (GE=MAS5)
GE	Original	0.514	0.413
GE	WSI model predictions	0.512	0.613
WSI	GE model predictions	0.513	0.613
GE + WSI	Original	0.467	0.464

Table 58: MCC results on Mild-F external data for spontaneous (Mild-F-SP) vs test article related necrosis (Mild-F-NSP) prediction, using imaging (WSI), normalized omics (GE) or their integration (GE+WSI). Four sets of target labels are used as ground truth for Mild-F-SP and Mild-F-NSP samples. Original: labels reported in TG-GATEs findings.

Part VII

DISCUSSION

DISCUSSION

In the last 24 months, the worldwide healthcare system focused on improving patient stratification and accelerating drug development in the fight against COVID-19. To this aim, the whole arsenal of cutting-edge technologies has been employed, including AI-based models. Patient stratification is of course a key need for all diseases, in particular for complex non communicable diseases. The urgent need to extract precise biomarkers from biomedical data requires efficient computational tools to disentangle the complexity of the generated information. AI techniques, particularly Deep Learning, have improved the efficiency and accuracy of statistical methods to analyze high-throughput imaging and omics data. AI can provide a fully data-driven approach that learns to extract high-level features, avoiding the need for predefined instructions that might introduce cognitive biases. Once trained, AI models can deliver fast and accurate predictions on unseen input, possibly revealing unknown relationships within the data. As a result, AI in Precision Medicine promises to aid decision-making, reduce diagnostic errors, and avoid unnecessary tests or ineffective treatments. By improving the economic cost and time management of the medical practice, AI will impact the overall healthcare system and promote patient-centered care.

This thesis investigates the opportunities of Deep Learning and Machine Learning techniques on heterogeneous data in clinical and preclinical tasks, exploring multi-modal and multi-omics integration strategies for a composite and broader view of biological mechanisms underlying complex diseases. The application of AI pipelines to the biomedical domain presented unique technical issues that are known to exacerbate the current discontinuity between research and deployment of medical AI.

Tools to overcome technical challenges in medical AI

Reproducibility of AI models is the most urgent concern in the medical setting; poorly designed experiments can result in overconfident estimates, undermining the reliability of AI models. As a countermeasure, the components presented in this thesis have systematically adopted a robust [DAP](#) that controls model selection and dataset partitioning biases, providing diagnostic tests to evaluate the ability of predictive models to generalize on original data.

Developing AI pipelines on biomedical data requires considerable effort in data curation and preprocessing, especially [DL](#) applications in Digital Pathology. The lack of standardization and robust tools for [WSI](#) data preprocessing is a major issue for the reproducibility of AI in Digital Pathology. In particular, flawed protocols for data partitioning on tile collections extracted from [WSI](#) datasets can introduce selection bias during the development of [DL](#) models, resulting in highly inflated performance (Chapter 5). The `histolab` library for reproducible [WSI](#) preprocessing introduced in Chapter 6 is the first open-source project that embraces best practices in software engineering including automated testing, and Continuous Integration. `histolab` allowed for maximum reliability of the preprocessing step in AI pipelines that support particularly complex designs, such as density-map guided training of [DL](#) models for small object detection on crowded images (Chapter 7) or sequential employment of diverse [DL](#) components ([BiGAN](#) and [CNN](#)) for classification within a weak-label setting (Chapter 12).

Opportunities in Precision Oncology

Deep Learning, particularly [CNNs](#), excels at pattern recognition on imaging datasets; its application to the biomedical domain can accelerate routine clinical tasks and discover sub-visual features that cannot be detected with traditional methods. For example, integrating Deep Learning and Radiomics has proven effective in estimating prognostic endpoints on complex multi-modal radiological datasets. In particular, the `RADLER` framework exploits 3D [CNNs](#) simultaneously trained on [PET](#) and [CT](#) scans to summarize functional and morphological information into compact rep-

resentations. Results of RADLER indicate that deep and hand-crafted features can provide complementary perspectives for improved biomarkers in precision oncology. Notably, the RADLER system significantly improved the original published solution.

In clinical histopathology, examining tissue slides remains a laborious and challenging task, even when performed on digital **WSIs**. For example, **TILs** are increasingly studied as promising predictive biomarkers in solid tumors; the quantification of **TILs**, however, requires pathologists to manually count nuclei (~ 100 pixels) on gigapixel images ($> 10^9$ pixels), and thus it is usually performed on selected **ROIs**. Moreover, when staining techniques other than **HE** are applied, as **IHC**, the identification of marked lymphocytes can be highly subjective. An AI approach that rapidly and objectively detects **TILs** on the whole tissue in **WSIs** (Chapter 7) is thus a promising tool for the automated quantification of the immune content in precision oncology.

The rationale for integrating different omics sources into a unique computational framework relies on the central dogma of molecular biology; multi-omics AI models can efficiently model causal pathways between different omics (e.g., genomics, transcriptomics, proteomics) to develop a better understanding of predictive biomarkers. For example, the reproducible network-based INF framework (Chapter 11) integrates various oncogenomics layers to extract biologically meaningful compact sets of sensitive biomarkers. Notably, INF improves the performance of single omics and naive juxtaposition.

Opportunities in Predictive Toxicology

In toxicological research, traditional animal testing to evaluate the potential of new drugs is slow, expensive, and raises highly controversial ethical issues. *In silico* models would have a crucial advantage in assessing organ toxicity in preclinical studies. We tested AI approaches on expression data from cancer cell lines to predict hepatotoxicity of chemical compounds (Chapter 9). Unfortunately, results on the CAMDA CMAP Challenge data were not sufficiently accurate for **DILI** label prediction, as confirmed by similar studies. Technical and biological reasons might cause a lack of classification power, including poor signals in the available data, lack of biological replicates for compound microarrays, or **DILI**-causing compounds that are harmful

to the liver but not to the two cell lines assayed. On the other hand, the ML4Tox framework (Chapter 10) to classify properties of endocrine-disrupting chemical compounds from molecular features significantly improved sensitivity over published results, suggesting the ability of AI models to develop faster, reproducible, and more cost-effective *in silico* drug safety assessment.

Digital Pathology has emerged among imaging modalities as one of the most exciting and challenging fields of application for medical AI. The histopathology analysis of tissue samples is the benchmark for diagnostic evaluation and treatment decisions in clinical practice and drug safety assessment in toxicology. Several AI models have been successfully applied to address relevant tasks in clinical pathology, mostly adopting tiles extracted from WSIs to train DL architectures in a weak-label setting. However, healthy tissue in histopathology slides for preclinical studies usually exceeds diseased tissue, or multiple lesion types of progressively increasing severity may be present in the same slide. Deriving the tile lab from the WSI before applying a standard supervised method is therefore a suboptimal procedure.

PathologAI is the first AI-based framework in preclinical pathology for DILI classification in a weak-label setting. PathologAI was developed to classify samples from extreme endpoints (treated samples with moderate/severe injury vs. controls with no injuries); remarkably, the trained model additionally differentiate spontaneous injury from the test article related injury. Moreover, PathologAI can stratify samples according to lesion severity level, even between mild findings, and dose-dependent experimental design. The availability of PathologAI contributes to experiment on the overarching question of whether AI can perform as humans or better when it comes to pre-clinical digital pathology. Although several teams experienced the need to re-annotate the TG-GATEs preclinical findings, in our review lesion categorization changes were suggested by expert pathologists for up to 38% of ROIs. Notably, the PathologAI mapping function prioritizes the ROIs with reviewed labels in treated cases. Second, the PathologAI framework enables the integration between molecular level and pathology reading. The analysis at a molecular level presents several advantages concerning histopathology assessment, including (i) the development of

alternative assays that are cheap and non-animal based. (ii) an improved translation from non-clinical to human by understanding the underlying mechanisms. (iii) the detection of targets for therapeutic options. First, the availability of an automated tool for quantitative lesion detection opens the possibility of scaling up the use of the molecular level as a measuring and modeling tool.

Limitations

It is important to acknowledge a set of due limitations to the experimental approach described in this thesis. In practice, for each of the different components discussed in the previous chapters, specific and general technical limitations can be found at the end of the chapter. As an overall caution, the reduced availability of large scale datasets limits capability of training AI models with guaranteed stability along time and robustness against variability or changes of new bioimaging and omics platforms. Of notice, the [FDA](#) has recently shared the GMLP – Good Machine Learning Practice for Medical device development guidelines ([139], Point #10), where it is required that Deployed Models should be monitored in “real world” use. It is expected that both in the clinical and preclinical domains of application, AI developers will need to focus on maintained or improved safety and performance. Most of the main themes discussed but not solved in this thesis are expected to be addressed and extended to include periodic or continual trained after deployment, with appropriate controls in place to manage risks of overfitting, bias, or lack of accuracy along time (degradation) of the model due to data drifts. All such aspects may impact the safety and performance of the model as it is used by the so-called Human-AI team. Hopefully, this thesis has shed some light on issues related to data leakage and selection bias of interest for automated management of performance and re-training risk.

Future perspectives

Although medical AI solutions for the automation of repetitive and subjective tasks are going to revolutionize the current *modus operandi* in clinical and preclinical settings, precise patient stratification remains a very ambitious goal of AI applications in medicine. The extraction of novel reliable, reproducible, and interpretable biomark-

ers from biomedical data will require the synergy of experts and heterogeneous computational tools to overcome the limitations for AI deployment in healthcare. We expect that unsupervised machine learning and computational geometry methods will be used in integration with the tools described in this thesis. For example, [TDA](#) descriptors can provide a robust and reproducible characterization of unsupervised clustering performed on features extracted from ML pipelines, as well as a deeper understanding on the geometrical structure of the inner states of deep learning models. Such descriptors can be used to optimize complex architectures like GAN networks, and to identify data drifts of relevance. Moreover, GAN architectures can create realistic synthetic data providing a world of simulations to overcome the lack of curated annotated datasets; for example, [Chen *et al.*](#) designed the ToxGAN approach to generate transcriptomic profiles for toxicogenomics studies without animal experiments [92]. We plan to integrate ToxGAN with PathologAI in a GAN-based framework for the spatial location of gene expression signatures on histopathology images. To this end, the framework described in Chapter 13 adopts the RMA preprocessing both for reference to the results in ToxGAN and for further integration. The impact of different preprocessing and normalization methods in combination with the GANs will be explored in the future.

Additionally, PathologAI will be trained on the more general task of normal vs abnormal histology to include multiple lesions in a single class, given the inconsistent annotations found in the [TG-GATEs](#) metadata for [WSI](#) samples. Despite the adoption of fair evaluation frameworks (*e.g.* the [DAP](#) variants described in this thesis), more sophisticated methods are needed to assess reproducibility of AI pipelines on biomedical data. In particular, data and concept drifts must be evaluated before deployment, considering the adoption of approaches such as online training, or focused re-training of models.

The path to the clinic for AI applications is at its early stages, but the rapid growth of applications and tools makes the promise of medical AI a real opportunity in the near future. Overall, AI will not replace medical specialists, but physicians who do not employ AI will likely be replaced by those who do.

14.1 Personal contributions

- Chapter 4. I contributed to the design of the DAPPER framework and performed several main steps of the experimental pipeline, including data preprocessing and DL/ML model training (50% of the experiments). I collaborated personally with the expert pathologist (L. Cima) for model accuracy validation. I also significantly contributed to the writing and figure preparation for the article.
- Chapter 5. I co-designed the overall pipeline and run 50% of the experiments on the considered datasets. I implemented 100% of the public protocols for data preprocessing on GTEx and TCGA data. I also drafted the article and prepared 100% of the figures.
- Chapter 6. I equally contributed to the software development and testing of all `histolab` modules, with A. Marcolini and E. Arbitrio. In addition, I independently designed and evaluated the nuclei segmentation algorithm (Section 6.4.0.2), and wrote 100% of the online `histolab` documentation.
- Chapter 7. I personally scanned the WSI collection with Dr. O. Melaiu. I designed the overall approach, supervised the experiments (performed by B. Papa), and evaluated the results, in particular the deep feature analysis. I contributed to the annotation of the histology tiles (25%) under the supervision of Dr. O. Melaiu. I also drafted the paper and prepared the figures.
- Chapter 8. I contributed to the implementation of the RADLER framework and run 50% of the experiments on the TCIA data. I co-designed the transfer learning strategies and performed the whole UMAP analysis. I also substantially contributed to the writing and prepared 100% of the figures in the published article.
- Chapter 9. I designed, implemented, and trained 100% of the DL models on the GE data. I also contributed to the writing and prepared the figure for the deep learning strategies.

- Chapter 10. I designed and implemented 100% of the deep learning models within the ML4Tox framework. I also contributed to the writing and prepared 75% of the figures for the published manuscript.
- Chapter 11. I equally contributed to the implementation of the overall INF framework in Python and the design of the experimental pipeline. I run 50% of the experiments on the oncogenomics datasets, and I independently performed the UMAP analysis on the INF signatures. I also substantially contributed to 50% of the writing, and prepared 100% of the figures for the published article.
- Chapter 12. I am one of the main contributors to the PathologAI framework. In particular, I implemented the encoding network and run 75% of the classification experiments; I implemented the UMAP analysis of the deep features; I equally contributed (50%) to the design of HistoMAP, and independently collaborated with the pathologists for the review of original annotations.
- Chapter 13. I independently designed the whole Machine Learning pipeline for GE data, implemented the Borda aggregation, and analyzed the resulting gene signatures. I performed 100% of the experiments on omics and combined imaging+omics data.

APPENDIX A

A.1 UMAP Background

UMAP creates a topological representation of a high-dimensional manifold by exploiting fuzzy topology. First, the **UMAP** algorithm constructs the Čech complex $\check{C}_\epsilon(X)$ of a point cloud X . Formally, the Čech complex is the nerve of the set of balls centered on each points and having radius ϵ (*open cover*). By the Nerve theorem [412], from $\check{C}_\epsilon(X)$ we can recover the key topological structures of the original data. Notice that in practice **UMAP** constructs the Vietoris Rips Complex (see Section A.2), given that $VR_\epsilon(X) \supset \check{C}_\epsilon(X) \subset VR_\epsilon(X) \subset \check{C}_{2\epsilon}(X)$, as it is computationally easier. $VR_\epsilon(X)$ implicitly defines a graph where an edge between two points is created if the balls centered at each point have non-empty intersection. To ensure an optimal construction of the high-dimensional graph **UMAP** requires two assumptions. First, data points are assumed to be uniformly distributed across the manifold to avoid gaps and clumps in the cover. This condition however is rarely met in a real world scenario; **UMAP** thus adapt the notion of distance on the manifold by stretching each ball to the k -nearest neighbor of its center. Formally, this step corresponds to replacing classical balls with a *fuzzy* cover where the radius ϵ is defined as a function in $[0, 1]$ decreasing further away from the center of the ball. To avoid completely isolated points that might affect the local structure of the manifold, **UMAP** introduces the local connectivity as a second assumption; this translates into requiring the fuzziness to decay only beyond the nearest neighbor of each ball's center. The local connectivity assumption also helps to avoid the curse of dimensionality [10]. The use of fuzzy topology results into a compatibility issue between the local metric defined for each ball; the theoretical solution relies on the definition of a fuzzy union as the probability that at least one edge exists

between two vertices. This process results into a weighted graph containing all the topological information of the high-dimensional manifold. The higher dimensional representation of the manifold is then adapted to the target lower dimensional space via optimization techniques. In particular, the cross-entropy function between edges' weights in high and low dimensional spaces is minimized.

Attractive force

$$\sum_{e \in E} w_h(e) \log \left(\frac{w_h(e)}{w_l(e)} \right) + (1 - w_h(e)) \log \left(\frac{1 - w_h(e)}{1 - w_l(e)} \right)$$

Repulsive force

Figure 88: Cross Entropy function minimized in the UMAP construction. The attractive force aims at optimizing clumps, the repulsive force aims at optimizing gaps. E : set of graph's edges, w_h : edges' weights of the high-dimensional graph, w_l : edges' weights of the low-dimensional graph¹.

In this adaptation, the exact points coordinates lose their spatial meaning but points that are close together are more similar than points far apart. The steps of the UMAP algorithm for a point cloud X in a high-dimensional space can be summarized as following (Figure 89):

- For all points $x_i \in X$ determine the optimal radius ϵ_i of each ball $B_{\epsilon_i}(x_i)$ centred at x_i , based on the distance to its K -nearest neighbors (fuzzy cover).
- Within each $B_{\epsilon_i}(x_i)$ attribute a probability to each connection, with the constraint that each x_i is connected, at least, to its closest neighbor. This results in a high-dimensional weighted graph G .
- Construct a lower-dimensional graph H by minimizing the cross-entropy function between G and H .

¹ Original image created with <https://bit.ly/3s2v84U>

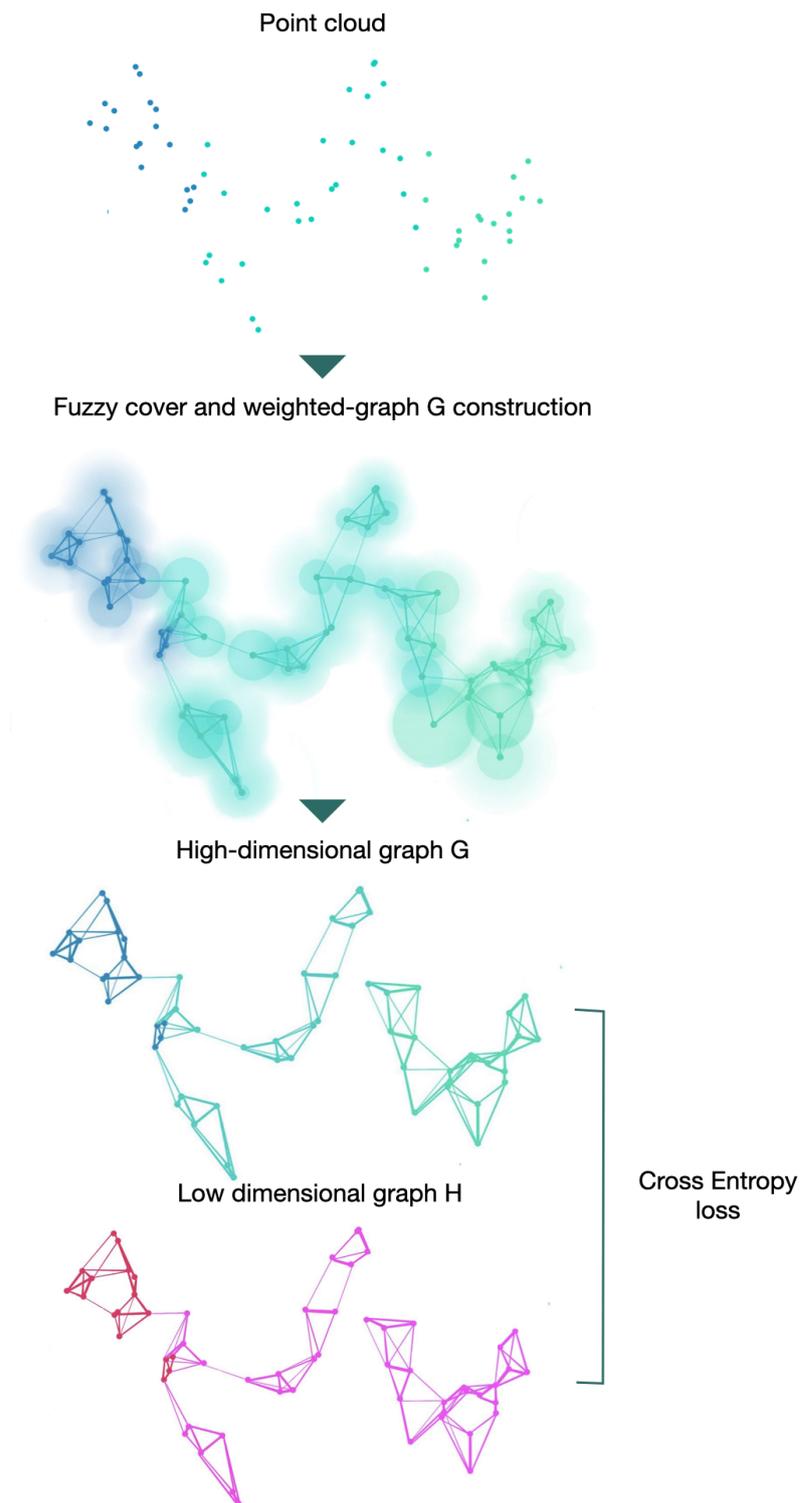


Figure 89: UMAP steps for dimensionality reduction of a high-dimensional point cloud.²

² Original image created with Keynote software. The visualization of the fuzzy cover has been created with [356].

A.2 Topological Descriptors

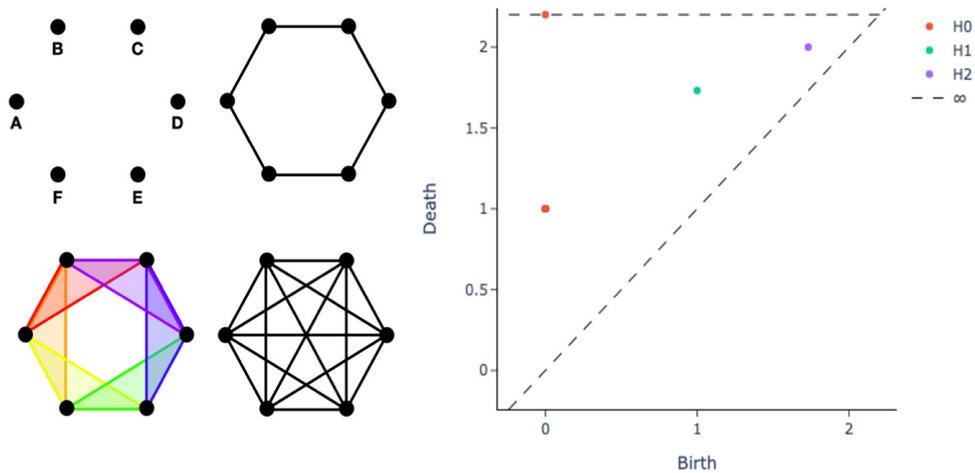


Figure 90: Persistence Diagram (right panel) for different Vietoris-Rips Complexes on equilateral hexagons with side length of 1 (left panel). In the left panel we display the four different categories of Vietoris-Rips complexes generated by 6 points forming the vertices of a regular hexagon of side length 1 in the Euclidean plane: 0-simplices (top left), 0- and 1-simplices (top right), 0-, 1- and 2-simplices (bottom left) and complexes including simplices of degree higher than 2 (bottom right). In the right panel, each point in the scatterplot represents a specific topological feature of the dataset, where the axes denote the values of the distance for which topological features appear (“birth” on the x axis) and vanish (“death” on the y axis)

A fundamental building block of TDA is Persistent Homology (PH), the geometric technique for studying a system at different length scales and discerning noise from actual topological features, the notion of how *persistent* a feature is throughout all the possible length scales. Hereafter we briefly outline the construction of a PD, whose starting point is the geometric concept of a simplex. Consider a finite set of points $S = \{x_0, \dots, x_n\}$ that are in general position with respect to the universe \mathbb{R}^d , *i.e.*, S should not be contained in an affine subspace of \mathbb{R}^d . If this condition is satisfied, S can be associated to a simplex $\sigma(S)$, the convex hull of S . Define the diameter of a simplex as the maximum distance between any two points on the simplex itself, or equivalently, between any of the two vertices, since the set is convex. Given a

set of points S with diameter r we can define the Vietoris Rips Complex as the set of simplices³ with diameter $d \leq r$. Moreover, given a Vietoris Rips Complex, it is possible to compute its Betti numbers, where the k -th Betti number is denoted as $\beta_k(X)$, for a simplicial complex X ; in layman's terms, $\beta_k(X)$ represents the number of k -dimensional holes on S . For example, consider the Vietoris Rips complexes shown in the left panel of Fig. 90 (adapted from Fig.5.2 in [337]) for different values of r .

The four different complexes can be described by Betti numbers as follows:

- The first complex ($r = 0$) is composed of 0-simplices, *i.e.*, the single points. Therefore, $\beta_0 = 6$ and $\beta_k = 0, \forall k > 0$. Note that β_0 indicates the number of connected components.
- The second complex ($r = 1$) includes 6 0-simplices and 6 1-simplices, denoted by dots and lines, respectively. Here $\beta_0 = 1$ and $\beta_1 = 1$ as there is one connected component and one 1-dimensional hole, namely the circle originated by the connection of the points.
- In the third step we have six 0-dimensional simplices, six 1-dimensional simplices, and six 2-dimensional simplices. The 2-dimensional simplices are the triangles, that is, the connection of 3 points. Thus $\beta_0 = 1$ and $\beta_1 = 1$.
- The last complex ($r = 2$) has simplices of degree greater than 2. Here $\beta_0 = 1$ but $\beta_1 = 0$: for this choice of r the 1-dimensional hole is filled.

The example in Fig. 90 illustrates that features of points arranged, for instance, in a circular shape can be recovered from their topological descriptors. In particular, $\beta_1 = 1$ for a large range of possible distance values: this is thus defined as a persistent feature of the dataset $\{A,B,C,D,E,F\}$. PD provides a compact representation of the topological insights provided by Betti numbers, as shown in the right panel of Fig. 90.

Betti numbers can be encoded into a two-dimensional scatterplot, each point representing a specific topological feature of the dataset. The x and y coordinates denote the values of the distance for which topological features appear (*birth* x) and vanish (*death* y), respectively. Considering (x, y) as coordinates of the scatterplot, only half

³ A simplex is the generalization of a triangle in an arbitrary dimension.

of the plot is relevant and, the closer a point is to the diagonal, the shorter its lifetime, and thus the point may represent topological noise. The k -th Betti number β_k is the rank of the k -th homology group H_k and thus each feature counted by β_k belongs to H_k . Considering now the plot in Fig. 90:

- The point at coordinates $(0, 1)$ represents 6 overlapping points. The 6 connected components (points) appear at $r = 0$ and vanish at $r = 1$, the side length of the equilateral hexagon, when each point is connected to its neighbors by a line.
- There is an H_1 point (a 1-dimensional hole) with the same birth value of the death of the 6 connected components ($r = 1$), as this topological feature arises from the union of the 6 features.
- A H_0 point (a 0-dimensional hole) lies at ∞ ; indeed, the connected components represented by the union of the 6 points persist for every value of r : for every value of $r > 1$ there exists only one connected component.

A compact representation to efficiently encode PH information is offered by the Persistent Diagram (PD) and Barcodes, which represent a different visualization of PD but encode the same information. If a PD is a scatterplot with coordinates given by the length scale for which topological features arise or vanish, a barcode can be regarded as a dumbbell plot where each bar represents a different topological feature, and the start and end values of the bar represent its life span.

Since both PDs and barcodes cannot be handle in a ML framework, several approaches have been recently introduced to translate PDs into discrete vector spaces [58]. Finally, Betti curves represent the magnitude of an homology group at different length scales of the filtration. Betti curves are an intuitive way to visualize the evolution of topological features within the dataset. Take as an example the equilateral hexagon and its persistence in Fig. 90. Recall that in the persistence diagram there is only one point at $(0, 1)$, which is the collapse of original connected components. By using Betti curve, it is possible to visualize the number of elements belonging to an homology group at every length scale. In this way we could have easily observed the Betti curve starting at 6 and decreasing to 1 for $r = 1$. A less trivial example

is reported in Fig. 91. On the top row, a point cloud with the shape of a lemniscate is created without noise, *i.e.*, the points are equally spaced. In the top row are also illustrated the persistence diagram and the Betti curves for homology groups H_0 and H_1 . Similarly, the bottom row contains a lemniscate-shaped point cloud with corresponding Persistence Diagram and Betti curves for H_0 and H_1 , but the point cloud construction involves some noise. The bottom row shows that it is still possible to appreciate the same topological structure, but the persistence diagram is more crowded with points near the diagonal, representing noisy features and thus not persistent features of the input point cloud. The different spatial organization of the two point clouds is also reflected by the the H_0 Betti curve; for the noisy dataset it has indeed slower decay rate.

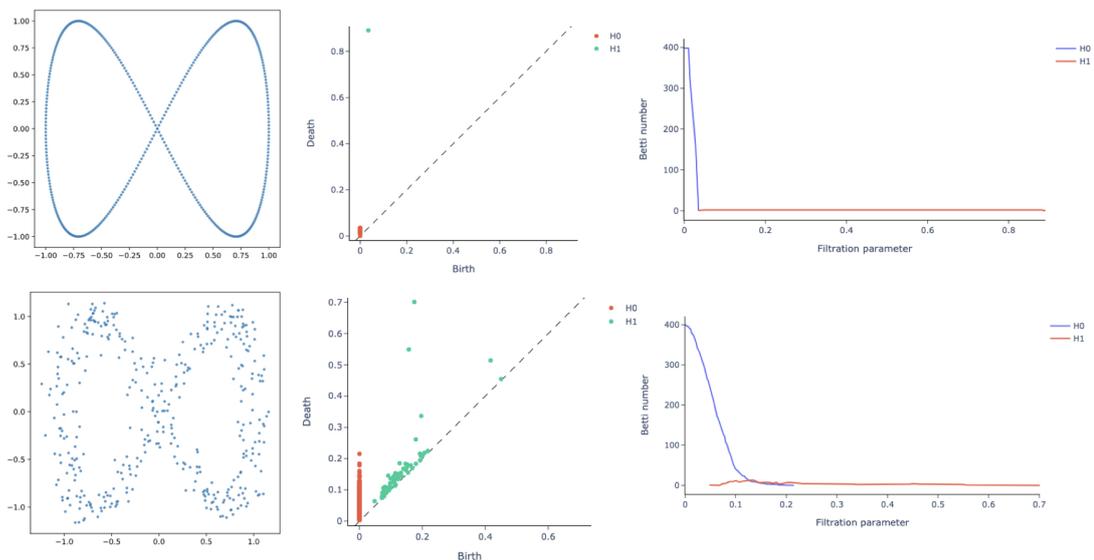


Figure 91: Topological descriptors on a lemniscate-shaped synthetic dataset with (top) and without (bottom) noise: dataset scatterplot (left column), Persistence Diagram (central column) and Betti curves for homology groups H_0 and H_1 (right column). Although it is still possible to appreciate the same topological structure, the Persistence Diagram for the noisy lemniscate has a cluster of points near the diagonal, representing not persistent features of the input point cloud. The different spatial organization of the two structures is also reflected by the Betti curve for H_0 , displaying a slower decay rate

A.3 EfficientNets

The Squeeze and Excitation blocks (SE), originally introduced in [209], implement a self-attention mechanism to make a network focus on the most relevant feature channels; SE work by first squeezing the spatial dimensions, and then using global information on feature channels to learn a vector of coefficients used as weights for each channel in the input feature map. See [209, Fig. 2] for a graphical scheme of the SE block. In particular, SEs exploit the global average pooling to resize the input feature map $M_{C \times H \times W}$ to a vector $z_{C \times 1 \times 1}$, where C is the number of feature channels, H and W the height and the width, respectively. The vector $z_{C \times 1 \times 1}$ is processed through a pipeline including a linear layer that halves its size, a Rectified Linear Unit (ReLU) activation layer, a second linear layer that recovers the original number of channels C , and finally a sigmoid activation feeding the vector of weights to the channels of the input feature map. Two different versions of SE were later introduced in [396, 397], aimed at improving segmentation models by introducing spatial attention components. The former, named sSqueeze and Excitation, works by first learning a mapping that reduces the number of channels in the input feature map from C to 1, hence summarizing information from the C input channels to a single number for each pixel, resulting in a two dimensional feature map. A sigmoid activation function is applied to each pixel of the two-dimensional feature map, providing weights in the range $[0, 1]$ for each pixel of the original feature map. The latter, called scSE block and shown in [397, Fig. 1], had the goal of combining the benefits of learning weights for spatial locations and feature channels. The two approaches work in parallel on the input feature map: a 1×1 convolution kernel is applied to obtain a two-dimensional one-channel matrix CM , while preserving the spatial dimensions. A pixel-wise sigmoid activation function is then applied to CM , finally obtaining the weight matrix, then multiplied by the input feature matrix on each channel. Two coefficients are obtained for each entry in the input feature map, and choosing their maximum value leads to best results in terms of performance and complexity added to the model [397].

EfficientNets have been introduced in [449], where the authors exploited the *network scaling practice*, namely, developed a novel baseline network, which can then be scaled up to obtain a more powerful model. Typically there are 3 main dimensions along which is possible to scale a network: depth, width and image resolution. ResNet is a good example of the first two cases: depth ranges from basic ResNet-18 with 18 layers to architectures with 1000+ layers, while width scaling allows reaching the same accuracy as very deep ResNets with reduced training time [518]. Scaling the third dimension, image resolution, is based on the idea that better resolution of input images implies learning patterns that were not discernible at low-res; however, there is a fundamental technical limit in the memory available on the machine used for training. EfficientNets, based on MnasNet [448], implement a novel strategy – called compound scaling – for scaling base neural network architectures by depth, width and resolution together using a set of coefficients for each dimension. Compound scaling has been validated also on common ResNet architectures and MobileNets, and can improve network performances, provided the existence of a strong baseline model. The available architectures range from the EfficientNet-b0 to the biggest EfficientNet-b7, achieving top performance on ImageNet with many fewer parameters, thus improving in efficiency. In particular, EfficientNet-b3 has 12×10^6 parameters [449] and, tested on ImageNet for a 1000-class classification task, EfficientNet-b3 scores 81.6% in top-1 accuracy, computed as the comparison between the ground truth and the most confident prediction of the model. Further, because of compound scaling, EfficientNets models support interpretability, since they focus on relevant regions when making predictions, as verified by the Class Activation Map [535]. Therefore, using EfficientNets as the encoder in a U-net architecture, allows the decoder to take advantage of the improved spatial attention mechanism of the encoder, and ultimately to improve the reconstruction of high-resolution density maps.

A.4 Intrinsic dimensionality

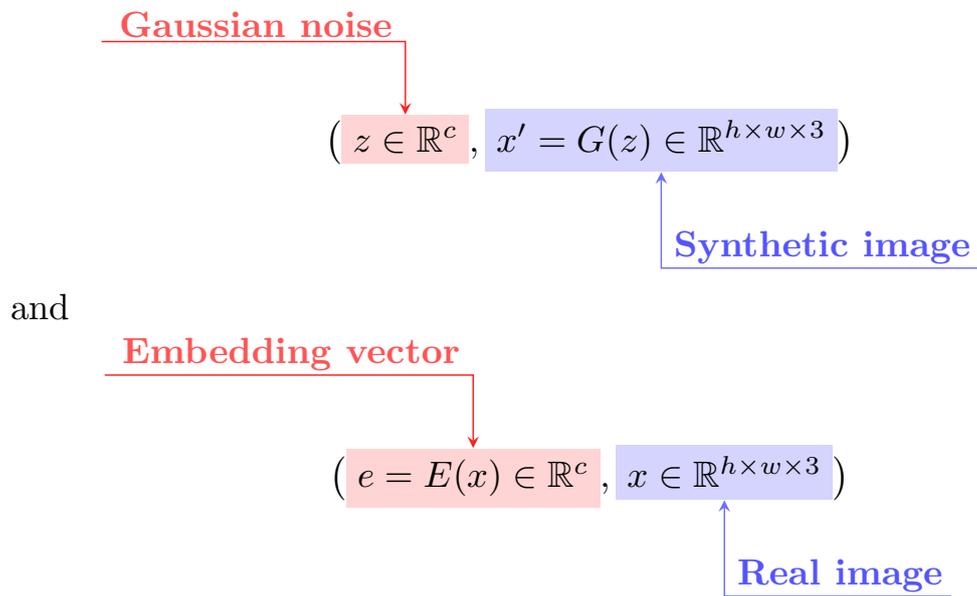
TwoNN

In a first attempt to understanding deep features, Odena and coworkers [345] used deconvolution layers to explore the filters learned by a CNN, while few years later Carter and colleagues [76] used [UMAP](#) to explore activation maps coming from different layers of an Inception network. More recently, the Mapper algorithm has been used in [153] to study the structure of CNN filters, while Ansuini and colleagues [17] employed TwoNN [141] to estimate the intrinsic dimensionality of a dataset and how such dimension changes when the dataset is transformed by the different CNN layers. TwoNN is a recent method that can be used for the estimation of Intrinsic Dimensionality of high dimensional and sparse data [141]. TwoNN assumes that the density of points is approximately constant on the length scale of the distance between a point and its two neighbors. With the former hypothesis, TwoNN uses information only from a restricted neighborhood of the point to measure properties of the data manifold [17]. The quantity $\rho_i = \frac{d_{i,2}}{d_{i,1}}$ is assumed to be a random variable following a Pareto distribution; if points are uniformly sampled and the hypothesized manifold has intrinsic dimensionality $d \in [0, +\infty]$, then $p(\rho_i; d) = d\rho_i^{-(d+1)}$. From this formula, the parameter d can be estimated by fitting the likelihood function to the data $P(d; \rho_i) = d\rho_i^{-(d+1)}$, where ρ_i is known.

A.5 BiGAN architecture

The [BiGAN](#) model extends the classic GAN structure, composed of a generator G and a discriminator D , by adding an encoder E as a third component. The encoder maps an image $x \in \mathbb{R}^{h \times w \times 3}$ of size $h \times w$ into an embedding vector $e = E(x) \in \mathbb{R}^c$, where $c \ll h \times w \times 3$. The objective of the generator is to fool the discriminator by creating realistic synthetic images x' from Gaussian noise, $G : z \rightarrow x'$, where $z \in \mathbb{R}^c$, and $x' \in \mathbb{R}^{h \times w \times 3}$. The discriminator D in the [BiGAN](#) framework is altered to distin-

guish the combinations of images and corresponding embeddings. In particular, the discriminator learns to separate:



As proven in [128], the BiGAN objective forces the encoder E to learn to invert the generator, such that $E(G(z)) = z$ and z defines a reliable representation of the image in a lower dimensional space.

RESEARCH OUTPUTS

- [1] Andrea Bizzego, Nicole Bussola, Marco Chierici, Valerio Maggio, Margherita Francescato, Luca Cima, Marco Cristoforetti, Giuseppe Jurman, and Cesare Furlanello. "Evaluating reproducibility of AI algorithms in digital pathology with DAPPER." In: *PLoS computational biology* 15.3 (2019), e1006269.
- [2] Andrea Bizzego, Nicole Bussola, D Salvalai, Marco Chierici, Valerio Maggio, Giuseppe Jurman, and Cesare Furlanello. "Integrating deep and radiomics features in cancer bioimaging." In: *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE. 2019, pp. 1–8.
- [3] Nicole Bussola, Alessia Marcolini, Valerio Maggio, Giuseppe Jurman, and Cesare Furlanello. "AI Slipping on Tiles: Data Leakage in Digital Pathology." In: *Pattern Recognition. ICPR International Workshops and Challenges, 2021, Proceedings*. 2021, pp. 167–182.
- [4] Nicole Bussola, Bruno Papa, Ombretta Melaiu, Aurora Castellano, Doriana Fruci, and Giuseppe Jurman. "Quantification of the Immune Content in Neuroblastoma: Deep Learning and Topological Data Analysis in Digital Pathology." In: *International Journal of Molecular Sciences* 22.16 (2021), p. 8804.
- [5] Marco Chierici, Nicole Bussola, Alessia Marcolini, Margherita Francescato, Alessandro Zandonà, Lucia Trastulla, Claudio Agostinelli, Giuseppe Jurman, and Cesare Furlanello. "Integrative Network Fusion: a multi-omics approach in molecular profiling." In: *Frontiers in oncology* 10 (2020), p. 1065.
- [6] Marco Chierici, Margherita Francescato, Nicole Bussola, Giuseppe Jurman, and Cesare Furlanello. "Predictability of drug-induced liver injury by machine learning." In: *Biology direct* 15.1 (2020), pp. 1–10.

- [7] Marco Chierici, Marco Giulini, Nicole Bussola, Giuseppe Jurman, and Cesare Furlanello. "Machine learning models for predicting endocrine disruption potential of environmental chemicals." In: *Journal of Environmental Science and Health, Part C* 36.4 (2018), pp. 237–251.
- [8] Cesare Furlanello, Manlio De Domenico, Giuseppe Jurman, and Nicole Bussola. "Towards a scientific blockchain framework for reproducible data analysis." In: *arXiv preprint arXiv:1707.06552* (2017).
- [9] Alessia Marcolini, Nicole Bussola, Ernesto Arbitrio, Mohamed Amgad, Giuseppe Jurman, and Cesare Furlanello. "histolab: a Python library for reproducible Digital Pathology preprocessing with automated testing." *Submitted*.
- [10] Simone Monaco, Nicole Bussola, Sara Butto, Diego Sona, Daniele Apiletti, Giuseppe Jurman, Elisa Viola, Marco Chierici, Christodoulos Xinaris, and Vincenzo Viola. "Cyst segmentation on kidney tubules by means of U-Net deep-learning models." In: *2021 IEEE International Conference on Big Data (Big Data)*. IEEE. 2021, pp. 3923–3926.

ACKNOWLEDGEMENTS

This section will include specific acknowledgments and credits to hosting laboratories and other scientific collaborators.

BIBLIOGRAPHY

- [1] Esther Abels, Liron Pantanowitz, Famke Aeffner, Mark D Zarella, Jeroen Van Der Laak, Marilyn M Bui, Venkata NP Vemuri, Anil V Parwani, Jeff Gibbs, Emmanuel Agosto-Arroyo, et al. "Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association." In: *The Journal of pathology* 249.3 (2019), pp. 286–294.
- [2] Balazs Acs, Mattiaas Rantalainen, and Johan Hartman. "Artificial intelligence as the next step towards precision pathology." In: *Journal of Internal Medicine* 288 (2020), pp. 62–81.
- [3] P. Afshar, A. Mohammadi, K. N. Plataniotis, A. Oikonomou, and H. Benali. "From Hand-Crafted to Deep Learning-based Cancer Radiomics: Challenges and Opportunities." In: *arXiv* 1808.07954 (2018), pp. 1–31.
- [4] Rakibull Ahasan, Ahasan Ulla Ratul, and ASM Bakibillah. "White blood cells nucleus segmentation from microscopic images of strained peripheral blood film during leukemia and normal condition." In: *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*. IEEE. 2016, pp. 361–366.
- [5] Y Al-Kofahi and *et al.* "Improved automatic detection and segmentation of cell nuclei in histopathology images." In: *IEEE Trans Biomed Eng* 57.4 (2009), pp. 841–852.
- [6] Fadhl M Alakwaa, Kumardeep Chaudhary, and Lana X Garmire. "Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data." In: *Journal of proteome research* 17.1 (2018), pp. 337–347.
- [7] Benjamin Alexander-Dann, Lavinia Lorena Pruteanu, Erin Oerton, Nitin Sharma, Ioana Berindan-Neagoe, Dezső Módos, and Andreas Bender. "Developments in toxicogenomics: understanding and predicting compound-induced toxicity from gene expression data." In: *Molecular omics* 14.4 (2018), pp. 218–236.
- [8] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. "Measuring the Objectness of Image Windows." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.11 (2012), pp. 2189–2202.
- [9] T. J. Alhindi, S. Kalra, K. H. Ng, A. Afrin, and H. R. Tizhoosh. "Comparing LBP, HOG and Deep Features for Classification of Histopathology Images." In: *Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–7.

- [10] Mebarka Allaoui, Mohammed Lamine Kherfi, and Abdelhakim Cheriet. "Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study." In: *International Conference on Image and Signal Processing*. Springer. 2020, pp. 317–325.
- [11] Md Zahangir Alom, Theus Aspiras, Tarek M Taha, Vijayan K Asari, TJ Bowen, Dave Billiter, and Simon Arkeel. "Advanced deep convolutional neural network approaches for digital pathology image analysis: A comprehensive evaluation with different use cases." In: *arXiv preprint arXiv:1904.09075* (2019).
- [12] Nicola Altini, Giacomo Donato Cascarano, Antonio Brunetti, Irio De Feudis, Domenico Buongiorno, Michele Rossini, Francesco Pesce, Loreto Gesualdo, and Vitoantonio Bevilacqua. "A Deep Learning Instance Segmentation Approach for Global Glomerulosclerosis Assessment in Donor Kidney Biopsies." In: *Electronics* 9.11 (2020), p. 1768.
- [13] Mohamed Amgad, Beezeley Beezeley, Deepak Roy Chittajallu, Lee Cooper, David Gutman, Brian Gutman, Sanghoon Lee, David Manthey, and Zach Mullen. *HistomicsTK*. <https://bit.ly/3rmVy2a>. 2020.
- [14] Mohamed Amgad, Habiba Elfandy, Hagar Hussein, Lamees A Atteya, Mai AT Elsebaie, Lamia S Abo Elnasr, Rokia A Sakr, Hazem SE Salem, Ahmed F Ismail, Anas M Saad, et al. "Structured crowdsourcing enables convolutional segmentation of histology images." In: *Bioinformatics* 35.18 (2019), pp. 3461–3467.
- [15] Erik J Amézquita, Michelle Y Quigley, Tim Ophelders, Elizabeth Munch, and Daniel H. Chitwood. "The shape of things to come: Topological data analysis and biology, from molecules to organisms." In: *Developmental Dynamics* 249.7 (2020), pp. 816–833.
- [16] Nicole M Anderson and M Celeste Simon. "The tumor microenvironment." In: *Current Biology* 30.16 (2020), R921–R925.
- [17] Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. "Intrinsic dimension of data representations in deep neural networks." In: *Proc. Advances in Neural Information Processing Systems 2019 (NeurIPS)*. Vol. 32. Curran Associates, Inc., 2019, pp. 6111–6122.
- [18] R Argelaguet, B Velten, D Arnol, S Dietrich, T Zenz, J C Marioni, F Buettner, W Huber, and O Stegle. "Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets." In: *Mol Syst Biol* 14.6 (2018), e8124.
- [19] H. Arimura, M. Soufi, H. Kamezawa, K. Ninomiya, and M. Yamada. "Radiomics with artificial intelligence for precision medicine in radiation therapy." In: *Journal of Radiation Research* 60.1 (2019), pp. 150–157.

- [20] M. Babaie, S. Kalra, A. Sriram, C. Mitcheltree, S. Zhu, A. Khatami, S. Rahnamayan, and H. R. Tizhoosh. "Classification and Retrieval of Digital Pathology Scans: A New Dataset." In: *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2017, pp. 8–16.
- [21] Andrea Baccarelli and Valentina Bollati. "Epigenetics and environmental chemicals." In: *Current opinion in pediatrics* 21.2 (2009), p. 243.
- [22] Reza Bahmanyar, Elenora Vig, and Peter Reinartz. "MRCNet: Crowd Counting and Density Map Estimation in Aerial and Ground Imagery." arXiv:1909.12743. 2019.
- [23] H Bai, M Zhou, M Zeng, and L Han. "PLA2G4A Is a Potential Biomarker Predicting Shorter Overall Survival in Patients with Non-M3/NPM1 Wildtype Acute Myeloid Leukemia." In: *DNA and Cell Biology Online* (2020), 20 Feb 2020.
- [24] M. Baker. "1,500 scientists lift the lid on reproducibility." In: *Nature News* 533.7604 (2016), p. 452.
- [25] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen. "Assessing the accuracy of prediction algorithms for classification: an overview." In: *Bioinformatics* 16.5 (2000), pp. 412–424.
- [26] Jeffrey D. Banfield and Adrian E. Raftery. "Model-Based Gaussian and Non-Gaussian Clustering." In: *Biometrics* 49.3 (1993), p. 803.
- [27] Peter Bankhead, Maurice B Loughrey, José A Fernández, Yvonne Dombrowski, Darragh G McArt, Philip D Dunne, Stephen McQuaid, Ronan T Gray, Liam J Murray, Helen G Coleman, et al. "QuPath: Open source software for digital pathology image analysis." In: *Scientific reports* 7.1 (2017), pp. 1–7.
- [28] Laura Barisoni, Kyle J Lafata, Stephen M Hewitt, Anant Madabhushi, and Ulysses G J Balis. "Digital pathology and computational image analysis in nephropathology." In: *Nature Reviews Nephrology* 16.11 (2020), pp. 669–685.
- [29] Ivraym Barsoum, Eriny Tawedrous, Hala Faragalla, and George M Yousef. "Histo-genomics: digital pathology at the forefront of precision medicine." In: *Diagnosis* 6.3 (2019), pp. 203–212.
- [30] A. Basavanhally, S. Ganesan, M. Feldman, N. Shih, C. Mies, J. Tomaszewski, and A. Madabhushi. "Multi-field-of-view framework for distinguishing tumor grade in ER+ breast cancer from entire histopathology slides." In: *IEEE Transactions on Biomedical Engineering* 60 (2013), pp. 2089–2099.
- [31] Ulrich Bauer. "Ripser: efficient computation of Vietoris-Rips persistence barcodes." arXiv:1908.02518. 2019.
- [32] Shabab Bazrafkan, Vincent Van Nieuwenhove, Joris Soons, Jan De Beenhouwer, and Jan Sijbers. "Deep Learning Based Computed Tomography Whys and Wherefores." In: *arXiv preprint arXiv:1904.03908* (2019).

- [33] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. "Dimensionality reduction for visualizing single-cell data using UMAP." In: *Nature Biotechnology Online* (2018), pp. 2018/12/03.
- [34] Stan Benjamens, Pranavsingh Dhunoo, and Bertalan Meskó. "The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database." In: *NPJ digital medicine* 3.1 (2020), pp. 1–8.
- [35] Eyal Bercovich and Marcia C Javitt. "Medical imaging: from roentgen to the digital revolution, and beyond." In: *Rambam Maimonides medical journal* 9.4 (2018).
- [36] Jesse A. Berlin, Susan C. Glasser, and Susan S. Ellenberg. "Adverse Event Detection in Drug Development: Recommendations and Obligations Beyond Phase 3." In: *American Journal of Public Health* 98.8 (2008), pp. 1366–1371.
- [37] M Bersanelli, E Mosca, D Remondini, E Giampieri, C Sala, G Castellani, and L Milanese. "Methods for the integration of multi-omics data: mathematical aspects." In: *BMC Bioinformatics* 17 (2016), S15.
- [38] Donald M Berwick. "Elusive waste: the fermi paradox in US health care." In: *JAMA* 322.15 (2019), pp. 1458–1459.
- [39] Donald M Berwick and Andrew D Hackbarth. "Eliminating waste in US health care." In: *Jama* 307.14 (2012), pp. 1513–1516.
- [40] Sébastien Besson, Roger Leigh, Melissa Linkert, Chris Allan, Jean-Marie Burel, Mark Carroll, David Gault, Riad Gozim, Simon Li, Dominik Lindner, et al. "Bringing open data to whole slide imaging." In: *European Congress on Digital Pathology*. Springer. 2019, pp. 3–10.
- [41] T. Beyer, J. Czernin, and L. S. Freudenberg. "Variations in clinical PET/CT operations: results of an international survey of active PET/CT users." In: *Journal of Nuclear Medicine* 52.2 (2011), pp. 303–310.
- [42] J.-E. Bibault et al. "Deep Learning and Radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer." In: *Scientific Report* 8.1 (2018), p. 12611.
- [43] Rene Bidart, Mehrdad J. Gangeh, Mohammad Peikari, Anne L. Martel, Ali Ghodsi, Sherine Salama, and Sharon Nofech-Mozes. "Localization and classification of cell nuclei in post-neoadjuvant breast cancer surgical specimen using fully convolutional networks." In: *Proc. Medical Imaging 2018: Digital Pathology*. Vol. 10581. SPIE, 2018, p. 1058100.
- [44] S Bilal, F Serrano, EJ Blutinger, C Vargas-Torres, C Counts, M Straight, and MP Lin. "84 Racial/Ethnic Disparities in Hospitalization And Clinical Outcomes Among COVID-19 Patients in an Integrated Health Care System In New York City." In: *Annals of Emergency Medicine* 78.2 (2021), S39.

- [45] A. Bizzego, N. Bussola, M. Chierici, M. Cristoforetti, M. Francescato, V. Maggio, G. Jurman, and C. Furlanello. "Evaluating reproducibility of AI algorithms in digital pathology with DAPPER." In: *PLOS Computational Biology* (2019).
- [46] A. Bizzego, N. Bussola, D. Salvalai, M. Chierici, V. Maggio, G. Jurman, and C. Furlanello. "Integrating deep and radiomics features in cancer bioimaging." In: *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. 2019, pp. 1–8.
- [47] A. Bizzego, N. Bussola, D. Salvalai, M. Chierici, V. Maggio, G. Jurman, and C. Furlanello. "Integrating deep and radiomics features in cancer bioimaging." In: *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. 2019, pp. 1–8.
- [48] Z. Bodalal, S. Trebeschi, and R. Beets-Tan. "Radiomics: a critical step towards integrated healthcare." In: *Insights into Imaging* 9.6 (2018), p. 911.
- [49] Zuhir Bodalal, Stefano Trebeschi, Thi Dan Linh Nguyen-Kim, Winnie Schats, and Regina Beets-Tan. "Radiogenomics: bridging imaging and genomics." In: *Abdominal radiology* 44.6 (2019), pp. 1960–1984.
- [50] Laura Bravo-Merodio, Animesh Acharjee, Dominic Russ, Vartika Bisht, John A Williams, Loukia G Tsaprouni, and Georgios V Gkoutos. "Translational biomarkers in the era of precision medicine." In: *Advances in Clinical Chemistry*. Vol. 102. Elsevier, 2021, pp. 191–232.
- [51] L Breiman. "Random Forests." In: *Mach Learn* 45.1 (2001), pp. 5–32.
- [52] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. "The balanced accuracy and its posterior distribution." In: *Proc. of the 20th International Conference on Pattern Recognition (ICPR-10)*. IEEE, 2010, pp. 3121–3124.
- [53] G M Brodeur, J Pritchard, F Berthold, N L Carlsen, V Castel, R P Castleberry, B De Bernardi, A E Evans, M Favrot, and F Hedborg. "Revisions of the international criteria for neuroblastoma diagnosis, staging, and response to treatment." In: *Journal of Clinical Oncology* 11.8 (1993), pp. 1466–1477.
- [54] G M Brodeur, R C Seeger, A Barrett, F Berthold, R P Castleberry, G D'Angio, B De Bernardi, A E Evans, M Favrot, and A I Freeman. "International criteria for diagnosis, staging, and response to treatment in patients with neuroblastoma." In: *Journal of Clinical Oncology* 6.12 (1988), pp. 1874–1881.
- [55] Shannon Brownlee, Kalipso Chalkidou, Jenny Doust, Adam G Elshaug, Paul Glasziou, Iona Heath, Somil Nagpal, Vikas Saini, Divya Srivastava, Kelsey Chalmers, et al. "Evidence for overuse of medical services around the world." In: *The Lancet* 390.10090 (2017), pp. 156–168.

- [56] Christian Brueffer, Johan Vallon-Christersson, Dorthe Grabau, Anna Ehinger, Jari Häkkinen, Cecilia Hegardt, Janne Malina, Yilun Chen, Pär-Ola Bendahl, Jonas Manjer, et al. "Clinical value of RNA sequencing-based classifiers for prediction of the five conventional breast cancer biomarkers: a report from the population-based multicenter Sweden Cancerome Analysis Network—Breast Initiative." In: *JCO Precision Oncology* 2 (2018), pp. 1–18.
- [57] Gema Bruixola, Elena Remacha, Ana Jiménez-Pastor, Delfina Dualde, Alba Viala, Jose Vicente Montón, Maider Ibarrola-Villava, Ángel Alberich-Bayarri, and Andrés Cervantes. "Radiomics and radiogenomics in head and neck squamous cell carcinoma: Potential contribution to patient management and challenges." In: *Cancer Treatment Reviews* 99 (2021), p. 102263.
- [58] Peter Bubenik. "Statistical topological data analysis using persistence landscapes." In: *The Journal of Machine Learning Research* 16.1 (2015), pp. 77–102.
- [59] F Buggenthin and et al. "An automatic method for robust and fast cell detection in bright field images from high-throughput microscopy." In: *BMC Bioinformatics* 14.1 (2013), pp. 1–12.
- [60] Marilyn M Bui, Sylvia L Asa, Liron Pantanowitz, Anil Parwani, Jeroen van der Laak, Christopher Ung, Ulysses Balis, Mike Isaacs, Eric Glassy, and Lisa Manning. "Digital and computational pathology: Bring the future into focus." In: *Journal of Pathology Informatics* 10 (2019).
- [61] Lyle D Burgoon. "Autoencoder Predicting Estrogenic Chemical Substances (APECS): An improved approach for screening potentially estrogenic chemicals using in vitro assays and deep learning." In: *Computational Toxicology* 2 (2017), pp. 45–49.
- [62] Nicole Bussola, Alessia Marcolini, Valerio Maggio, Giuseppe Jurman, and Cesare Furlanello. "AI slipping on tiles: Data leakage in digital pathology." In: *International Conference on Pattern Recognition*. Springer. 2021, pp. 167–182.
- [63] D. Bychkov, N. Linder, R. Turkki, S. Nordling, P. E. Kovanen, C. Verrill, M. Walliander, M. Lundin, C. Haglund, and J. Lundin. "Deep learning based tissue analysis predicts outcome in colorectal cancer." In: *Scientific Reports* 8.1 (2018), p. 3395.
- [64] P Byfield and J Gamper. *compay-syntax*. <https://bit.ly/3ggk2Un>. 2020.
- [65] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Mirafior, Victor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images." In: *Nature medicine* 25.8 (2019), pp. 1301–1309.

- [66] Ricardo J G B Campello, Davoud Moulavi, and Joerg Sander. "Density-based clustering based on hierarchical density estimates." In: *Proc. Advances in Knowledge Discovery and Data Mining 2013 – Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. Vol. 7819. Lecture Notes in Computer Science. Springer International Publishing, 2013, pp. 160–172.
- [67] Ian G Cannell, Yi Wen Kong, and Martin Bushell. "How do microRNAs regulate gene expression?" In: *Biochemical Society Transactions* 36.6 (2008), pp. 1224–1231.
- [68] L Cantini, P Zakeri, C Hernandez, A Naldi, D Thieffry, E Remy, and A Baudot. "Benchmarking joint multi-omics dimensionality reduction approaches for cancer study." *bioRxiv* 905760. 2020.
- [69] A. Canziani, A. Paszke, and E. Culurciello. "An analysis of deep neural network models for practical applications." In: *arXiv* 1605.076784 (2017), pp. 1–7.
- [70] Lena-Maria Carlson, Anna De Geer, Baldur Sveinbjörnsson, Abiel Orrego, Tommy Martinsson, Per Kogner, and Jelena Levitskaya. "The microenvironment of human neuroblastoma supports the activation of tumor-associated T lymphocytes." In: *OncImmunity* 2.3 (2013), e23618.
- [71] Marc Carlson. *hgu133a.db: Affymetrix Human Genome U133 Set annotation data (chip hgu133a)*. R package version 3.2.3. 2016.
- [72] Gunnar Carlsson. "Topological methods for data modelling." In: *Nature Reviews Physics* 2.12 (2020), pp. 697–708.
- [73] Gunnar Carlsson, Tigran Ishkhanov, Vin de Silva, and Afra Zomorodian. "On the Local Behavior of Spaces of Natural Images." In: *International Journal of Computer Vision* 76.1 (2007), pp. 1–12.
- [74] A. E. Carpenter et al. "CellProfiler: image analysis software for identifying and quantifying cell phenotypes." In: *Genome Biology* 7.10 (2006), R100.
- [75] Rickey E Carter, Zachi I Attia, Francisco Lopez-Jimenez, and Paul A Friedman. "Pragmatic considerations for fostering reproducible research in artificial intelligence." In: *NPJ digital medicine* 2.1 (2019), pp. 1–3.
- [76] Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. *Exploring Neural Networks with Activation Atlases*. <https://bit.ly/3Gzg6ZX>. Distill. 2019.
- [77] David Casarett. "The Science of Choosing Wisely—Overcoming the Therapeutic Illusion." In: *The New England journal of medicine* 374.13 (2016), pp. 1203–1205.
- [78] Quassim Cassam. "Diagnostic error, overconfidence and self-knowledge." In: *Palgrave Communications* 3.1 (2017), pp. 1–8.

- [79] Florence M.G. Cavalli et al. "Intertumoral Heterogeneity within Medulloblastoma Subgroups." In: *Cancer Cell* 31.6 (2017), 737–754.e6.
- [80] H Chai, X Zhou, Z Cui, J Rao, Z Hu, and Y Yang. "Integrating multi-omics data with deep learning for predicting cancer prognosis." bioRxiv 807214. 2019.
- [81] S Chakraborty, M I Hosen, M Ahmed, and H U Shekhar. "Onco-Multi-OMICS Approach: A New Frontier in Cancer Research." In: *BioMed Res Int* 2018 (2018), p. 9836256.
- [82] Naga P Chalasani, Haripriya Maddur, Mark W Russo, Robert J Wong, and K Rajender Reddy. "ACG Clinical Guideline: Diagnosis and Management of Idiosyncratic Drug-Induced Liver Injury." In: *Official journal of the American College of Gastroenterology | ACG* 116.5 (2021), pp. 878–898.
- [83] Mark R Chassin, Robert W Galvin, et al. "The urgent need to improve health care quality: Institute of Medicine National Roundtable on Health Care Quality." In: *Jama* 280.11 (1998), pp. 1000–1005.
- [84] Farhan Chaudhry, Jenna Isherwood, Tejeshwar Bawa, Dhruvil Patel, Katherine Gurdziel, David E Lanfear, Douglas M Ruden, and Phillip D Levy. "Single-cell RNA sequencing of the cardiovascular system: new looks for old diseases." In: *Frontiers in cardiovascular medicine* 6 (2019), p. 173.
- [85] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.
- [86] Frédéric Chazal and Bertrand Michel. "An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists." arXiv:1710.04019. 2017.
- [87] Chi-Long Chen, Chi-Chung Chen, Wei-Hsiang Yu, Szu-Hua Chen, Yu-Chan Chang, Tai-I Hsu, Michael Hsiao, Chao-Yuan Yeh, and Cheng-Yu Chen. "An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning." In: *Nature communications* 12.1 (2021), pp. 1–13.
- [88] Minjun Chen, Huixiao Hong, Hong Fang, Reagan Kelly, Guangxu Zhou, Jürgen Borlak, and Weida Tong. "Quantitative Structure-Activity Relationship Models for Predicting Drug-Induced Liver Injury Based on FDA-Approved Drug Labeling Annotation and Using a Large Collection of Drugs." In: *Toxicological Sciences* 136.1 (2013), pp. 242–249.
- [89] S. Chen, J. Qin, X. Ji, B. Lei, T. Wang, D. Ni, and J.-Z. Cheng. "Automatic Scoring of Multiple Semantic Attributes With Multi-Task Feature Leverage: A Study on Pulmonary Nodules in CT Images." In: *IEEE Transactions on Medical Imaging* 36.3 (2017), pp. 802–814.

- [90] Sihong Chen, Kai Ma, and Yefeng Zheng. "Med3D: Transfer Learning for 3D Medical Image Analysis." In: *arXiv preprint arXiv:1904.00625* (2019).
- [91] Ting Chen and Christophe Chefd'hotel. "Deep Learning Based Automatic Immune Cell Detection for Immunohistochemistry Images." In: *Proc. Machine Learning in Medical Imaging 2014 (MLMI)*. Springer International Publishing, 2014, pp. 17–24.
- [92] Xi Chen, Ruth Roberts, Weida Tong, and Zhichao Liu. "Tox-GAN: An AI Approach Alternative to Animal Studies—a Case Study with Toxicogenomics." In: *Toxicological Sciences* (2021).
- [93] Lixin Cheng, Leung-Yau Lo, Nelson LS Tang, Dong Wang, and Kwong-Sak Leung. "CrossNorm: a novel normalization strategy for microarray data in cancers." In: *Scientific reports* 6.1 (2016), pp. 1–11.
- [94] N-K V Cheung and M A Dyer. "Neuroblastoma: developmental biology, cancer genomics and immunotherapy." In: *Nature Reviews Cancer* 13.6 (2013), pp. 397–411.
- [95] Davide Chicco and Giuseppe Jurman. "The advantages of the Matthews Correlation Coefficient (MCC) over F1 score and accuracy in binary classification evaluation." In: *BMC Genomics* 21.1 (2020), p. 6.
- [96] A M Chiu, M Mitra, L Boymoushakian, and H A Coller. "Integrative analysis of the inter-tumoral heterogeneity of triple-negative breast cancer." In: *Scientific Reports* 8.1 (2018), p. 11807.
- [97] Michał Chromiak and Krzysztof Stencel. "A data model for heterogeneous data integration architecture." In: *Beyond Databases, Architectures and Structures. BDAS 2014*. Ed. by S Kozielski, D Mrozek, P Kasprowski, B Małysiak-Mrozek, and D Kostrzewa. Vol. 424. Springer, Cham, 2014, pp. 547–556.
- [98] F. Ciompi et al. "Towards automatic pulmonary nodule management in lung cancer screening with deep learning." In: *Scientific Reports* 7 (2017), p. 46479.
- [99] Alex Clark. *Pillow (PIL Fork) Documentation*. <https://bit.ly/3ojaFrB>. 2015.
- [100] Claudio G Clemente, Martin C Mihm, Rosaria Bufalino, Stefano Zurrada, Paola Collini, and Natale Cascinelli. "Prognostic value of tumor infiltrating lymphocytes in the vertical growth phase of primary cutaneous melanoma." In: *Cancer: Interdisciplinary International Journal of the American Cancer Society* 77.7 (1996), pp. 1303–1310.
- [101] Personalized Medicine Coalition. "The personalized medicine report: opportunity, challenges, and the future." In: (2020).
- [102] Jacob Cohen. "A coefficient of agreement for nominal scales." In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46.
- [103] S. Cohen. *Artificial Intelligence and Deep Learning in Pathology*. Elsevier, 2020.

- [104] Susan L Cohn et al. "The International Neuroblastoma Risk Group (INRG) Classification System: An INRG Task Force Report." In: *Journal of Clinical Oncology* 27.2 (2009), pp. 289–297.
- [105] G. S. Colafati et al. "MRI features as a helpful tool to predict the molecular subgroups of medulloblastoma: state of the art." In: *Therapeutic Advances in Neurological Disorders* 11 (2018), pp. 1–14.
- [106] Francis S Collins and Harold Varmus. "A new initiative on precision medicine." In: *New England Journal of Medicine* 372.9 (2015), pp. 793–795.
- [107] C Cortes and V N Vapnik. "Support-vector networks." In: *Mach Learn* 20.3 (1995), pp. 273–297.
- [108] C. Cortes and V. Vapnik. "Support-vector networks." In: *Machine Learning* 20.3 (1995), pp. 273–297.
- [109] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, and A. Tsirigos. "Classification And Mutation Prediction From Non-Small Cell Lung Cancer Histopathology Images Using Deep Learning." In: *Nature Medicine* 24.10 (2018), pp. 1559–1567.
- [110] David Crosby, Nicole Lyons, Emma Greenwood, Samantha Harrison, Sara Hiom, Jodie Moffat, Talisia Quallo, Emlyn Samuel, and Ian Walker. "A roadmap for the early detection and diagnosis of cancer." In: *The Lancet Oncology* 21.11 (2020), pp. 1397–1399.
- [111] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. "AutoAugment: Learning Augmentation Policies from Data." In: *arXiv* 1805.09501 (2018).
- [112] M C Dao et al. "A Data Integration Multi-Omics Approach to Study Calorie Restriction-Induced Changes in Insulin Sensitivity." In: *Front Physiol* 9 (2019), p. 1958.
- [113] Bilel Daoud, Ken'ichi Morooka, Shoko Miyauchi, Ryo Kurazume, Wafa Mnejja, Leila Farhat, and Jamel Daoud. "Dose Distribution Prediction for Optimal Treatment of Modern External Beam Radiation Therapy for Nasopharyngeal Carcinoma." In: *Workshop on Artificial Intelligence in Radiation Therapy*. 2019, pp. 128–136.
- [114] Andrew M Davidoff. "Neuroblastoma." In: *Seminars in Pediatric Surgery* 21.1 (2012), pp. 2–14.
- [115] J. De Fauw et al. "Clinically applicable deep learning for diagnosis and referral in retinal disease." In: *Nature Medicine* 24.9 (2018), pp. 1342–1350.
- [116] Maria Cristina De Vera Mudry, Jim Martin, Vanessa Schumacher, and Raghavan Venugopal. "Deep Learning in Toxicologic Pathology: A New Approach to Evaluate Rodent Retinal Atrophy." In: *Toxicologic Pathology* 49.4 (2021), pp. 851–861.

- [117] Kimberley DeMerle, Derek C Angus, and Christopher W Seymour. "Precision Medicine for COVID-19: Phenotype Anarchy or Promise Realized?" In: *JAMA* 325.20 (2021), pp. 2041–2042.
- [118] DeepAI Team. *What are Deep Features?* <https://bit.ly/3gAfE2M>, note = Accessed: 2022-08-02.
- [119] Olivier Dehaene, Axel Camara, Olivier Moindrot, Axel de Lavergne, and Pierre Courtiol. "Self-Supervision Closes the Gap Between Weak and Strong Supervision in Histology." In: *arXiv preprint arXiv:2012.03583* (2020).
- [120] F Del Chierico et al. "Gut microbiota profiling of pediatric nonalcoholic fatty liver disease and obese patients unveiled by an integrated meta-omics-based approach." In: *Hepatology* 65.2 (2017), pp. 451–464.
- [121] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "ImageNet: A large-scale hierarchical image database." In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition 2009 (CVPR)*. IEEE, 2009, pp. 248–255.
- [122] C. Denkert et al. "Standardized evaluation of tumor-infiltrating lymphocytes in breast cancer: results of the ring studies of the international immunoncology biomarker working group." In: *Modern Pathology* 29.10 (2016), pp. 1155–1164.
- [123] Daniel G. Deschler and Terry Day. "TNM staging of head and neck cancer and neck dissection classification." In: *American Academy of Otolaryngology–Head and Neck Surgery Foundation* (2008).
- [124] T. J. Di Ciccio and B. Efron. "Bootstrap confidence intervals (with Discussion)." In: *Statistical Science* 11 (1996), pp. 189–228.
- [125] A. Diamant, A. Chatterjee, M. Vallières, G. Shenouda, and J. Seuntjens. "Deep learning in head & neck cancer outcome prediction." In: *Scientific Report* 9.1 (2019), p. 2764.
- [126] C Dimitrakopoulos, S Kumar Hindupur, L Häfliger, J Behr, H Montazeri, M N Hall, and N Beerenwinkel. "Network-based integration of multi-omics data for prioritizing cancer genes." In: *Bioinformatics* 34.14 (2018), pp. 2441–2448.
- [127] David J Dix, Keith A Houck, Matthew T Martin, Ann M Richard, R Woodrow Setzer, and Robert J Kavlock. "The ToxCast program for prioritizing toxicity testing of environmental chemicals." In: *Toxicological Sciences* 95.1 (2007), pp. 5–12.
- [128] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. "Adversarial feature learning." In: *arXiv preprint arXiv:1605.09782* (2016).

- [129] Madeleine S. Durkee, Rebecca Abraham, Junting Ai, Jordan D. Fuhrman, Marcus R. Clark, and Maryellen L. Giger. “Comparing Mask R-CNN and U-Net architectures for robust automatic segmentation of immune cells in immunofluorescence images of Lupus Nephritis biopsies.” In: *Proc. Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues 2021 XIX*. Ed. by James F. Leary, Attila Tarnok, and Irene Georgakoudi. Vol. SPIE 11647. SPIE, 2021, p. 116470X.
- [130] Mike Dusenberry, Fei Hu, Nakul Jindal, and Deron Eriksson. *deep-histopath*. <https://github.com/CODAIT/deep-histopath>. 2019.
- [131] Abhishek Dutta and Andrew Zisserman. “The VIA Annotation Software for Images, Audio and Video.” In: *Proc. ACM International Conference on Multimedia 2019 (MM)*. ACM, 2019, pp. 2276–2279.
- [132] Amelie Echle, Niklas T Rindtorff, Titus Josef Brinker, Tom Luedde, Alexander T Pearson, and Jakob N Kather. “Deep learning in cancer pathology: a new generation of clinical biomarkers.” In: *British Journal of Cancer Review articles* (2020), 18 November 2020.
- [133] Editorial. “Towards trustable machine learning.” In: *Nature Biomedical Engineering* 2 (2018), pp. 709–710.
- [134] S. Ekins, A. J. Williams, and J. J. Xu. “A Predictive Ligand-Based Bayesian Model for Human Drug-Induced Liver Injury.” In: *Drug Metabolism and Disposition* 38.12 (2010), pp. 2302–2308.
- [135] Olivier Elemento. “The future of precision medicine: Towards a more predictive personalized medicine.” In: *Emerging Topics in Life Sciences* 4.2 (2020), pp. 175–177.
- [136] Alison M Elliott. “Genetic counseling and genome sequencing in pediatric rare disease.” In: *Cold Spring Harbor perspectives in medicine* (2019), a036632.
- [137] Martin Ester, Hans-Peter Kriegel, Jiirg Sander, and Xiaowei Xu. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.” In: *Proc. Knowledge Discovery and Data Mining 1996 (KDD)*. Vol. 96. 1996, pp. 226–231.
- [138] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. “Dermatologist-level classification of skin cancer with deep neural networks.” In: *Nature* 542.7639 (2017), pp. 115–118.
- [139] FDA. *Good Machine Learning Practice for Medical Device Development: Guiding Principles*. <https://bit.ly/3hrytptu>, note = Accessed: February 15, 2022.

- [140] Paola Facchetti, I Prigione, Fabio Ghiotto, Paola Tasso, Alberto Garaventa, and Vito Pistoia. "Functional and molecular characterization of tumour-infiltrating lymphocytes and clones thereof from a major-histocompatibility-complex-negative human tumour: neuroblastoma." In: *Cancer Immunology, Immunotherapy* 42.3 (1996), pp. 170–178.
- [141] Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. "Estimating the intrinsic dimension of datasets by a minimal neighborhood information." In: *Scientific Reports* 7.1 (2017), pp. 1–8.
- [142] Michael D Farwell, Daniel A Pryma, and David A Mankoff. "PET/CT imaging in cancer: current applications and future directions." In: *Cancer* 120.22 (2014), pp. 3433–3445.
- [143] Fera Science Limited, London, UK. *In Silico Predictive Toxicology (INSPECT)*. <https://bit.ly/3rjVpMU>. 2018.
- [144] Jose F Figueroa, Rishi K Wadhwa, and Ashish K Jha. "Eliminating wasteful health care spending—is the united states simply spinning its wheels?" In: *JAMA cardiology* 5.1 (2020), pp. 9–10.
- [145] Fabian V Filipp. "Opportunities for Artificial Intelligence in Advancing Precision Medicine." In: *Current Genetic Medicine Reports* 7.4 (2019), pp. 208–213.
- [146] Steven J Frank. "Resource-frugal classification and analysis of pathology slides using image entropy." In: *Biomedical Signal Processing and Control* 66 (2021), p. 102388.
- [147] L. Fu, J. Ma, Y. Ren, Y. S. Han, and J. Zhao. "Automatic Detection of Lung Nodules: False Positive Reduction Using Convolutional Neural Networks and Handcrafted Features." In: *Proc. SPIE 10134, Medical Imaging 2017: Computer-Aided Diagnosis*. SPIE, 2017, 101340A.
- [148] Qiaofen Fu et al. "Prognostic value of tumor-infiltrating lymphocytes in melanoma: a systematic review and meta-analysis." In: *Oncology* 8.7 (2019), e1593806.
- [149] Thomas J Fuchs and Joachim M Buhmann. "Computational pathology: challenges and promises for tissue analysis." In: *Computerized Medical Imaging and Graphics* 35.7-8 (2011), pp. 515–530.
- [150] Seiya Fujita and Xian-Hua Han. "Cell Detection and Segmentation in Microscopy Images with Improved Mask R-CNN." In: *Proc. Computer Vision 2020 Workshops (ACCV)*. Springer, 2021, pp. 58–70.
- [151] C. Furlanello, M. Serafini, S. Merler, and G. Jurman. "Entropy-based gene ranking without selection bias for the predictive classification of microarray data." In: *BMC Bioinformatics* 4.1 (2003), p. 54.
- [152] Cesare Furlanello, Maria Serafini, Stefano Merler, and Giuseppe Jurman. "An accelerated procedure for recursive feature ranking on microarray data." In: *Neural Networks* 16.5-6 (2003), pp. 641–648.

- [153] Rickard B Gabrielsson and Gunnar Carlsson. "Exposition and Interpretation of the Topology of Neural Networks." In: *Proc. IEEE International Conference On Machine Learning And Applications 2019 (ICMLA)*. IEEE, 2019, pp. 1069–1076.
- [154] Vijay Gadepally, Timothy Mattson, Michael Stonebraker, Fusheng Wang, Gang Luo, Yanhui Laing, and Alevtina Dubovitskaya. *Heterogeneous Data Management, Polystores, and Analytics for Healthcare: VLDB 2019 Workshops, Poly and DMAH, Los Angeles, CA, USA, August 30, 2019, Revised Selected Papers*. Vol. 11721. Springer Nature, 2019.
- [155] M E Gallo Cantafio, K Grillone, D Caracciolo, F Scionti, M Arbitrio, V Barbieri, L Pensabene, P Hiram Guzzi, and M T Di Martino. "From Single Level Analysis to Multi-Omics Integrative Approaches: A Powerful Strategy towards the Precision Oncology." In: *High-Throughput* 7 (2018), p. 33.
- [156] Yunhe Gao, Rui Huang, Ming Chen, Zhe Wang, Jincheng Deng, Yuanyuan Chen, Yiwei Yang, Jie Zhang, Chanjuan Tao, and Hongsheng Li. "FocusNet: Imbalanced Large and Small Organ Segmentation with an End-to-End Deep Neural Network for Head and Neck CT Images." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 829–838.
- [157] Miren García-Cortés, Aida Ortega-Alonso, M. Isabel Lucena, and Raúl J. Andrade. "Drug-induced liver injury: a safety review." In: *Expert Opinion on Drug Safety* 17.8 (2018), pp. 795–804.
- [158] Miren Garcia-Cortes, Mercedes Robles-Diaz, Camilla Stephens, Aida Ortega-Alonso, M Isabel Lucena, and Raúl J Andrade. "Drug induced liver injury: An update." In: *Archives of Toxicology* (2020), pp. 1–27.
- [159] Emilio Garcia, Renato Hermoza, Cesar Beltran Castanon, Luis Cano, Miluska Castillo, and Carlos Castañeda. "Automatic Lymphocyte Detection on Gastric Cancer IHC Images Using Deep Learning." In: *Proc. IEEE 30th International Symposium on Computer-Based Medical Systems 2017 (CBMS)*. IEEE, 2017, pp. 200–204.
- [160] S. Garczyk et al. "AGR3 in breast cancer: prognostic impact and suitable serum-based biomarker for early cancer detection." In: *PLoS ONE* 10.4 (2015), e0122106.
- [161] Laura-Jayne Gardiner, Anna Paola Carrieri, Jenny Wilshaw, Stephen Checkley, Edward O Pyzer-Knapp, and Ritesh Krishna. "Using human in vitro transcriptome analysis to build trustworthy machine learning models for prediction of animal drug toxicity." In: *Scientific reports* 10.1 (2020), pp. 1–8.
- [162] Elisa Drelie Gelasca, Jiyun Byun, Boguslaw Obara, and BS Manjunath. "Evaluation and benchmark for biological image segmentation." In: *2008 15th IEEE International Conference on Image Processing*. IEEE. 2008, pp. 1816–1819.

- [163] Angela Giardino, Supriya Gupta, Emmi Olson, Karla Sepulveda, Leon Lenchik, Jana Ivanidze, Rebecca Rakow-Penner, Midhir J Patel, Rathan M Subramaniam, and Dhakshinamoorthy Ganeshan. "Role of imaging in the era of precision medicine." In: *Academic radiology* 24.5 (2017), pp. 639–649.
- [164] Robert J Gillies, Paul E Kinahan, and Hedvig Hricak. "Radiomics: images are more than pictures, they are data." In: *Radiology* 278.2 (2016), pp. 563–577.
- [165] Ross Girshick. "Fast R-CNN." In: *Proc. IEEE International Conference on Computer Vision 2015 (ICCV)*. IEEE, 2015, pp. 1440–1448.
- [166] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation." In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition 2014 (CVPR)*. IEEE, 2014, pp. 580–587.
- [167] Adam Goode, Benjamin Gilbert, Jan Harkes, Drazen Jukic, and Mahadev Satyanarayanan. "OpenSlide: A vendor-neutral software foundation for digital pathology." In: *Journal of pathology informatics* 4 (2013).
- [168] Mark Graber. "Diagnostic errors in medicine: a case of neglect." In: *The joint commission journal on quality and patient safety* 31.2 (2005), pp. 106–113.
- [169] Peter Greaves. *Histopathology of preclinical toxicity studies: interpretation and relevance in drug safety evaluation*. Academic Press, 2011.
- [170] Jon Griffin and Darren Treanor. "Digital pathology in clinical use: where are we now and what is holding us back?" In: *Histopathology* 70.1 (2017), pp. 134–145.
- [171] FDA-NIH Biomarker Working Group et al. "BEST (Biomarkers, EndpointS, and other Tools) Resource [Internet]." In: (2016).
- [172] Y. Guo. "Clinical significance of serum MicroRNA-203 in patients with acute myeloid leukemia." In: *Bioengineered* 10.1 (Dec. 2019), pp. 345–352.
- [173] Y. Guo, P. Yu, Z. Liu, Y. Maimaiti, C. Chen, Y. Zhang, X. Yin, S. Wang, C. Liu, and T. Huang. "Prognostic and clinicopathological value of GATA binding protein 3 in breast cancer: A systematic review and meta-analysis." In: *PLoS ONE* 12.4 (2017), e0174843.
- [174] Pooja Gupta and Avleen Kaur Malhi. "Using deep learning to enhance head and neck cancer diagnosis and classification." In: *2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCA)*. 2018, pp. 1–6.
- [175] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. "Gene selection for cancer classification using support vector machines." In: *Machine learning* 46.1-3 (2002), pp. 389–422.

- [176] Freedman David H. "The Gene Bubble: Why We Still Aren't Disease-Free." In: *Fast Company* (2009), pp. 116–22.
- [177] Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. "A Survey of Deep Learning Techniques for Medical Diagnosis." In: *Information and Communication Technology for Sustainable Development*. 2020, pp. 161–170.
- [178] Miriam Hägele, Philipp Seegerer, Sebastian Lapuschkin, Michael Bockmayr, Wojciech Samek, Frederick Klauschen, Klaus-Robert Müller, and Alexander Binder. "Resolving challenges in deep learning-based analyses of histopathological images using explanation methods." In: *Scientific Reports* 10.1 (2020), p. 6423.
- [179] Benjamin Haibe-Kains et al. "Transparency and reproducibility in artificial intelligence." In: *Nature* 586.7829 (2020), E14–E16.
- [180] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning." In: *International Conference on Intelligent Computing*. Springer. 2005, pp. 878–887.
- [181] Xiaoping Han, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, Ziming Zhou, Haide Chen, Fang Ye, et al. "Mapping the mouse cell atlas by microwell-seq." In: *Cell* 172.5 (2018), pp. 1091–1107.
- [182] C. R Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N J Smith, et al. "Array Programming with NumPy." In: *Nature* 585 (2020), pp. 357–362.
- [183] Wanda M Haschek, Colin G Rousseaux, Matthew A Wallig, and Brad Bolon. "Toxicologic pathology: An introduction." In: *Haschek and Rousseaux's Handbook of Toxicologic Pathology*. Elsevier, 2022, pp. 1–12.
- [184] M Hatt, C Parmar, J Qi, I El Naqa IEEE Transactions on, and 2019. "Machine (deep) learning methods for image processing and radiomics." In: *ieeexplore.ieee.org* ().
- [185] M. Hatt, F. Tixier, L. Pierce, P. E. Kinahan, C. C. Le Rest, and D. Visvikis. "Characterization of PET/CT images using texture analysis: the past, the present... any future?" In: *European Journal of Nuclear Medicine and Molecular Imaging* 44.1 (2017), pp. 151–165.
- [186] M. Hatt, F. Tixier, D. Visvikis, and C. Cheze Le Rest. "Radiomics in PET/CT: More Than Meets the Eye?" In: *Journal of Nuclear Medicine* 58.3 (2017), pp. 365–366.
- [187] Mathieu Hatt, Catherine Cheze Le Rest, Florent Tixier, Bogdan Badic, Ulrike Schick, and Dimitris Visvikis. "Radiomics: Data Are Also Images." In: *Journal of nuclear medicine : official publication, Society of Nuclear Medicine* 60.Suppl 2 (Sept. 2019), 38S–44S.

- [188] Mathieu Hatt et al. "The first MICCAI challenge on PET tumor segmentation." In: *Medical Image Analysis* 44 (2018), pp. 177–195.
- [189] T. Hayakawa, VB S. Prasath, H. Kawanaka, B. J. Aronow, and S. T. "Computational Nuclei Segmentation Methods in Digital Pathology: A Survey." In: *Arch Comput Methods Eng* 2019 (2019), pp. 1–13.
- [190] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning." In: *Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on.* IEEE. 2008, pp. 1322–1328.
- [191] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition." In: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 770–778.
- [192] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask R-CNN." In: *Proc. IEEE International Conference on Computer Vision 2017 (ICCV)*. 2017, pp. 2980–2988.
- [193] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask R-CNN." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.2 (2020), pp. 386–397.
- [194] Felix Hensel, Michael Moor, and Bastian Rieck. "A Survey of Topological Machine Learning Methods." In: *Frontiers in Artificial Intelligence* 4 (2021), p. 52.
- [195] Jean-Karim Hériché, Stephanie Alexander, and Jan Ellenberg. "Integrating imaging and omics: Computational methods and challenges." In: *Annual Review of Biomedical Data Science* 2 (2019), pp. 175–197.
- [196] Christian J Herold, Jonathan S Lewin, Andreas G Wibmer, James H Thrall, Gabriel P Krestin, Adrian K Dixon, Stefan O Schoenberg, Rena J Geckle, Ada Muellner, and Hedvig Hricak. "Imaging in the age of precision medicine: summary of the proceedings of the 10th Biannual Symposium of the International Society for Strategic Studies in Radiology." In: *Radiology* 279.1 (2016), pp. 226–238.
- [197] Fedorov A Doyle SW Pieper S Klepeis V Herrmann MD Clunie DA. "Implementing the DICOM standard for digital pathology." In: *J Pathol Inform* 9.37 (2020).
- [198] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. "Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges." In: *Journal of digital imaging* (2019), pp. 1–15.
- [199] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. "Gans trained by a two time-scale update rule converge to a local nash equilibrium." In: *Advances in neural information processing systems* 30 (2017).

- [200] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. "Improving neural networks by preventing co-adaptation of feature detectors." In: *arXiv 1207.0580* (2012), pp. 1–18.
- [201] T. K. Ho. "Random decision forests." In: *Proceedings of 3rd International Conference on Document Analysis and Recognition (ICDAR 1995)*. IEEE, 1995, 278–282).
- [202] Sepp Hochreiter, Djork-Arne Clevert, and Klaus Obermayer. "A new summarization method for Affymetrix probe level data." In: *Bioinformatics 22.8* (2006), pp. 943–949.
- [203] M. S. Hofman and R. J. Hicks. "How We Read Oncologic FDG PET/CT." In: *Cancer Imaging 16.1* (2016), p. 35.
- [204] Huixiao Hong, Svetoslav Slavov, Weigong Ge, Feng Qian, Zhenqiang Su, Hong Fang, Yiyu Cheng, Roger Perkins, Leming Shi, and Weida Tong. "Mold² Molecular Descriptors for QSAR." In: *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*. Wiley-Blackwell, 2012. Chap. 3, pp. 65–109.
- [205] Huixiao Hong, Shraddha Thakkar, Minjun Chen, and Weida Tong. "Development of Decision Forest Models for Prediction of Drug-Induced Liver Injury in Humans Using A Large Set of FDA-approved Drugs." In: *Scientific Reports 7.1* (2017).
- [206] Huixiao Hong, Qian Xie, Weigong Ge, Feng Qian, Hong Fang, Leming Shi, Zhenqiang Su, Roger Perkins, and Weida Tong. "Mold², Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics." In: *Journal of Chemical Information and Modeling 48.7* (2008), pp. 1337–1344.
- [207] Mingye Hong, Shuang Tao, Ling Zhang, Li-Ting Diao, Xuanmei Huang, Shaohui Huang, Shu-Juan Xie, Zhen-Dong Xiao, and Hua Zhang. "RNA sequencing: new technologies and applications in cancer research." In: *Journal of hematology & oncology 13.1* (2020), pp. 1–16.
- [208] Yasushi Horai, Mao Mizukawa, Hironobu Nishina, Satomi Nishikawa, Yuko Ono, Kana Takemoto, and Nobuyuki Baba. "Quantification of histopathological findings using a novel image analysis platform." In: *Journal of Toxicologic Pathology 32.4* (2019), pp. 319–327.
- [209] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-Excitation Networks." In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition 2018 (CVPR)*. IEEE, 2018, pp. 7132–7141.
- [210] Bin Huang, Zhewei Chen, Po-Man Wu, Yufeng Ye, Shi-Ting Feng, Ching-Yee Oliver Wong, Liyun Zheng, Yong Liu, Tianfu Wang, Qiaoliang Li, et al. "Fully Automated Delineation of Gross Tumor Volume for Head and Neck Cancer on PET-CT Using Deep Learning: A Dual-Center Study." In: *Contrast media & molecular imaging 2018* (2018).

- [211] R. Huang et al. "Profiling of the Tox21 10K compound library for agonists and antagonists of the estrogen receptor alpha signaling pathway." In: *Scientific Reports* 4 (2014), p. 5664.
- [212] Stefan Huber. "Persistent Homology in Data Science." In: *Data Science – Analytics and Applications – Proc. 3rd International Data Science Conference 2020 (iDSC)*. Springer Fachmedien Wiesbaden, 2021, pp. 81–88.
- [213] Bulat Ibragimov and Lei Xing. "Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks." In: *Medical physics* 44.2 (2017), pp. 547–557.
- [214] Yoshinobu Igarashi, Noriyuki Nakatsu, Tomoya Yamashita, Atsushi Ono, Yasuo Ohno, Tetsuro Urushidani, and Hiroshi Yamada. "Open TG-GATES: a large-scale toxicogenomics database." In: *Nucleic acids research* 43.D1 (2015), pp. D921–D927.
- [215] Osamu Iizuka, Fahdi Kanavati, Kei Kato, Michael Rambeau, Koji Arihiro, and Masayuki Tsuneki. "Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours." In: *Scientific Reports* 10.1 (2020), p. 1504.
- [216] Jesse C Ikeme, James W Salazar, and Richard W Grant. "Reappraising Medical Syntax—Does Race Belong in the First Line of the Patient History?" In: *JAMA Internal Medicine* (2021).
- [217] J. P. A. Ioannidis et al. "Repeatability of published microarray gene expression analyses." In: *Nature Genetics* 41.2 (2009), pp. 149–155.
- [218] Humayun Irshad, Antoine Veillard, Ludovic Roux, and Daniel Racoceanu. "Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential." In: *IEEE reviews in biomedical engineering* 7 (2013), pp. 97–114.
- [219] Meredith S Irwin and Julie R Park. "Neuroblastoma: paradigm for precision medicine." In: *Pediatric Clinics of North America* 62.1 (2015), pp. 225–256.
- [220] Helen H. Wang Janina A. Longtine Ann Dvorak James L. Connolly Stuart J. Schnitt and Harold F. Dvorak. *Holland-Frei Cancer Medicine*. 6th. BC Decker, 2003.
- [221] Andrew Janowczyk and Anant Madabhushi. "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases." In: *Journal of Pathology Informatics* 7.1 (2016), p. 29.
- [222] Daniel Jarrett, Eleanor Stride, Katherine Vallis, and Mark J Gooding. "Applications and limitations of machine learning in radiation oncology." In: *The British journal of radiology* 92.1100 (2019).

- [223] Frédéric Jehl, Fabien Degalez, Maria Bernard, Frédéric Lecerf, Laetitia Lagoutte, Colette Désert, Manon Coulée, Olivier Bouchez, Sophie Leroux, Behnam Abasht, et al. “RNA-Seq data for reliable SNP detection and genotype calling: Interest for coding variant characterization and cis-regulation analysis by allele-specific expression in livestock species.” In: *Frontiers in genetics* 12 (2021), p. 1104.
- [224] Ni Jiang and Feihong Yu. “Multi-column network for cell counting.” In: *OSA Continuum* 3.7 (2020), pp. 1834–1846.
- [225] Yi-Zhou Jiang et al. “Genomic and Transcriptomic Landscape of Triple-Negative Breast Cancers: Subtypes and Treatment Strategies.” In: *Cancer Cell* 35.3 (2019), 428–440.e5.
- [226] Pius Joseph. “Transcriptomics in toxicology.” In: *Food and Chemical Toxicology* 109 (2017), pp. 650–662.
- [227] G. Jurman, S. Merler, A. Barla, S. Paoli, A. Galea, and C. Furlanello. “Algebraic stability indicators for ranked lists in molecular profiling.” In: *Bioinformatics* 24.2 (2008), pp. 258–264.
- [228] G. Jurman, S. Riccadonna, and C. Furlanello. “A Comparison of MCC and CEN Error Measures in Multi-Class Prediction.” In: *PLoS ONE* 7.8 (2012), e41882.
- [229] G. Jurman, S. Riccadonna, and C. Furlanello. “A comparison of MCC and CEN error measures in multi-class prediction.” In: *PLOS ONE* 7.8 (2012), e41882.
- [230] G Jurman, S Riccadonna, R Visintainer, and C Furlanello. “Algebraic comparison of partial lists in bioinformatics.” In: *PLoS One* 7.5 (2012), e36540.
- [231] Wetterstrand KA. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. <https://bit.ly/34tPuMz>. Accessed: January 9, 2022.
- [232] K Kalecky, R Modisette, S Pena, Y-R Cho, and J Taube. “Integrative analysis of breast cancer profiles in TCGA by TNBC subgrouping reveals novel microRNA-specific clusters, including miR-17-92a, distinguishing basal-like 1 and basal-like 2 TNBC subtypes.” In: *BMC Cancer* 20.1 (2020).
- [233] Benjamin H Kann, Sanjay Aneja, Gokoulakrichenane V Loganadane, Jacqueline R Kelly, Stephen M Smith, Roy H Decker, B Yu James, Henry S Park, Wendell G Yarbrough, Ajay Malhotra, et al. “Pretreatment identification of head and neck cancer nodal metastasis and extranodal extension using deep learning neural networks.” In: *Scientific reports* 8.1 (2018), p. 14036.
- [234] K Karczewski and M Snyder. “Integrative omics for health and disease.” In: *Nat Rev Genet* 19 (2018), pp. 299–310.
- [235] P. A. Keane and E. J. Topol. “With an eye to AI and autonomous diagnosis.” In: *NPJ Digital Medicine* 1.1 (2018), p. 40.

- [236] B D Kechavarzi, H Wu, and T N Doman. "Bottom-up, integrated -omics analysis identifies broadly dosage-sensitive genes in breast cancer samples from TCGA." In: *PLoS ONE* 14.1 (2019), e0210910.
- [237] S Khan, G Ince-Dunn, A Suomalainen, and L L Elo. "Integrative omics approaches provide biological and clinical insights: examples from mitochondrial diseases." In: *J Clin Invest* 130.1 (2020), pp. 20–28.
- [238] B. Kieffer, M. Babaie, S. Kalra, and H. R. Tizhoosh. "Convolutional Neural Networks for Histopathology Image Classification: Training vs. Using Pre-Trained Networks." In: *Proceedings of the Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA 2017)*. IEEE, 2017, pp. 1–6.
- [239] B C Kim, Y S Sung, HI Suk 2016 4th International Winter, and 2016. "Deep feature learning for pulmonary nodule classification in a lung CT." In: *ieeexplore.ieee.org* ().
- [240] So-Woon Kim, Jin Roh, and Chan-Sik Park. "Immunohistochemistry for pathologists: protocols, pitfalls, and tips." In: *Journal of Pathology and Translational Medicine* 50.6 (2016), pp. 411–418.
- [241] DP Kingma and J Ba. "Adam: A method for stochastic optimization." In: *International Conference on Learning Representations (ICLR)*. 2015.
- [242] Dmitry Kobak and George C. Linderman. "Initialization is critical for preserving global data structure in both t-SNE and UMAP." In: *Nature Biotechnology* 39 (2021), pp. 156–157.
- [243] H W L Koh, D Fermin, C Vogel, K Pui Choi, R M Ewing, and H Choi. "iOmicSPASS: network-based integration of multiomics data for predictive subnetwork discovery." In: *NPJ Syst Biol Appl* 5 (2019), p. 22.
- [244] Pekka Kohonen, Juuso A Parkkinen, Egon L Willighagen, Rebecca Ceder, Krister Wennerberg, Samuel Kaski, and Roland C Grafström. "A transcriptomics data-driven gene space accurately predicts liver cytopathology and drug-induced liver injury." In: *Nature Communications* 8 (2017), p. 15932.
- [245] Yoriaki Komeda et al. "Computer-Aided Diagnosis Based on Convolutional Neural Network System for Colorectal Polyp Classification: Preliminary Experience." In: *Oncology* 93.1 (2017), pp. 30–34.
- [246] Yan Kong, Hui Li, Yongyong Ren, Georgi Z Genchev, Xiaolei Wang, Hongyu Zhao, Zhiping Xie, and Hui Lu. "Automated yeast cells segmentation and counting using a parallel U-Net based two-stage framework." In: *OSA Continuum* 3.4 (2020), pp. 982–992.
- [247] D. Kontos and R. M. Summers. "Radiomics and Deep Learning." In: *Journal of Medical Imaging* 4.4 (2018), p. 041301.

- [248] B. Korbar, A. M. Olofson, A. P. Miraflor, C. M. Nicka, M. A. Suriawinata, L. Torresani, A. A Suriawinata, and S. Hassanpour. "Deep Learning for Classification of Colorectal Polyps on Whole-Slide Images." In: *Journal of Pathology Informatics* 8 (2017), p. 30.
- [249] M Kosinski and P Biecek. *RTCGA: The Cancer Genome Atlas Data Integration*. R package version 1.16.0, <https://rtcga.github.io/RTCGA>. 2019.
- [250] J Köster and S Rahmann. "Snakemake—a scalable bioinformatics workflow engine." In: *Bioinformatics* 28.19 (2012), pp. 2520–2522.
- [251] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." In: *Proc. Advances in Neural Information Processing Systems 2012 (NIPS)*. Curran Associates, Inc., 2012, pp. 1106–1114.
- [252] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." In: *Communications of the ACM* 60.6 (2017), pp. 84–90.
- [253] Harold W Kuhn. "The Hungarian Method for the assignment problem." In: *Naval Research Logistics* 2 (1955), pp. 83–97.
- [254] Jogile Kuklyte, Jenny Fitzgerald, Sophie Nelissen, Haolin Wei, Aoife Whelan, Adam Power, Ajaz Ahmad, Martyna Miarka, Mark Gregson, Michael Maxwell, et al. "Evaluation of the Use of Single-and Multi-Magnification Convolutional Neural Networks for the Determination and Quantitation of Lesions in Nonclinical Pathology Studies." In: *Toxicologic Pathology* 49.4 (2021), pp. 815–842.
- [255] M. D. Kumar, M. Babaie, S. Zhu, S. Kalra, and H. R. Tizhoosh. "A Comparative Study of CNN, BoVW and LBP for Classification of Histopathological Images." In: *Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2017, pp. 1–7.
- [256] Jeroen van der Laak, Geert Litjens, and Francesco Ciompi. "Deep learning in histopathology: the path to the clinic." In: *Nature medicine* 27.5 (2021), pp. 775–784.
- [257] Sarah Labib, Carole Yauk, Andrew Williams, Volker M Arlt, David H Phillips, Paul A White, and Sabina Halappanavar. "Subchronic oral exposure to benzo (a) pyrene leads to distinct transcriptomic changes in the lungs that are related to carcinogenesis." In: *Toxicological Sciences* 129.1 (2012), pp. 213–224.
- [258] Jeffrey A Lam, MD Edward Feller, et al. "Are We Right When We're Certain? Overconfidence in Medicine." In: *Rhode Island Medical Journal* 103.2 (2020), pp. 11–12.

- [259] Justin Lamb, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N Ross, et al. "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease." In: *science* 313.5795 (2006), pp. 1929–1935.
- [260] P. Lambin et al. "Radiomics: extracting more information from medical images using advanced feature analysis." In: *European Journal of Cancer* 48.4 (2012), pp. 441–446.
- [261] Tze Kin Lau, Xiaofan Zhu, Yvonne Ka Yin Kwok, Tak Yeung Leung, and Kwong Wai Choy. "Recent Advances in the Noninvasive Prenatal Testing for Chromosomal Abnormalities Using Maternal Plasma DNA." In: *Journal of Fetal Medicine* 7.1 (2020), pp. 17–23.
- [262] I Lauder and W Aherne. "The Significance of Lymphocytic Infiltration in Neuroblastoma." In: *British Journal of Cancer* 26.4 (1972), pp. 321–330.
- [263] Jeffrey T. Leek, W. Evan Johnson, Hilary S. Parker, Andrew E. Jaffe, and John D. Storey. "The sva package for removing batch effects and other unwanted variation in high-throughput experiments." In: *Bioinformatics* 28.6 (2012), pp. 882–883.
- [264] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning." In: *Journal of Machine Learning Research* 18.17 (2017), pp. 1–5.
- [265] Victor Lempitsky and Andrew Zisserman. "Learning To Count Objects in Images." In: *Proc. Advances in Neural Information Processing Systems 2012 (NIPS)*. Curran Associates Inc., 2010, pp. 1324–1332.
- [266] P Leon-Mimila, J Wang, and A Huertas-Vazquez. "Relevance of Multi-Omics Studies in Cardiovascular Diseases." In: *Front Cardiovasc Med* 6 (2019), p. 91.
- [267] Chao-Ting Li, Pau-Choo Chung, Hung-Wen Tsai, Nan-Haw Chow, and Kuo-Sheng Cheng. "Inflammatory Cells Detection in H&E Staining Histology Images Using Deep Convolutional Neural Network with Distance Transformation." In: *Communications in Computer and Information Science*. Springer Singapore, 2019, pp. 665–672.
- [268] Jiawei Li, Xin Guan, Zhimin Fan, Lai-Ming Ching, Yan Li, Xiaojia Wang, Wen-Ming Cao, and Dong-Xu Liu. "Non-invasive biomarkers for early detection of breast Cancer." In: *Cancers* 12.10 (2020), p. 2767.
- [269] S. M. Li, Y. Q. Zhao, Y. L. Hao, and Y. Y. Liang. "Upregulation of miR-504-3p is associated with favorable prognosis of acute myeloid leukemia and may serve as a tumor suppressor by targeting MTHFD2." In: *Eur Rev Med Pharmacol Sci* 23.3 (2019), pp. 1203–1213.

- [270] Y Li, F-X Wu, and A Ngom. "A review on machine learning principles for multi-view biological data integration." In: *Brief Bioinform* 19.2 (2018), pp. 325–340.
- [271] Yu Li, Chao Huang, Lihong Ding, Zhongxiao Li, Yijie Pan, and Xin Gao. "Deep learning in bioinformatics: Introduction, application, and perspective in the big data era." In: *Methods* 166 (2019), pp. 4–21.
- [272] Z. Li, Y. Wang, J. Yu, Y. Guo, and W. Cao. "Deep learning based radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma." In: *Scientific Reports* 7.1 (2017), p. 5467.
- [273] Anna Licata. "Adverse drug reactions and organ damage: The liver." In: *European Journal of Internal Medicine* 28 (2016), pp. 9–16.
- [274] Nina Linder, Jenny C Taylor, Richard Colling, Robert Pell, Edward Alveyn, Johnson Joseph, Andrew Protheroe, Mikael Lundin, Johan Lundin, and Clare Verrill. "Deep learning for detecting tumour-infiltrating lymphocytes in testicular germ cell tumours." In: *Journal of Clinical Pathology* 72.2 (2018), pp. 157–164.
- [275] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghahfoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez. "A survey on deep learning in medical image analysis." In: *Medical Image Analysis* 42 (2017), pp. 60–88.
- [276] Anika Liu, Namshik Han, Jordi Munoz-Muriedas, and Andreas Bender. "Deriving time-concordant event cascades from gene expression data: A case study for Drug-Induced Liver Injury (DILI)." In: *bioRxiv* (2021).
- [277] Chao Liu, Cuiyun Gao, Xin Xia, David Lo, John Grundy, and Xiaohu Yang. "On the replicability and reproducibility of deep learning in software engineering." In: *arXiv preprint arXiv:2006.14244* (2020).
- [278] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. "On the Variance of the Adaptive Learning Rate and Beyond." In: *Proc. International Conference on Learning Representation 2020 (ICLR)*. OpenReview.net, 2019, pp. 1–14.
- [279] S-H Liu et al. "DriverDBv3: a multi-omics database for cancer driver gene research." In: *Nucleic Acids Res* 48.D1 (2019), pp. D863–D870.
- [280] Shanglong Liu, Yuejuan Zhang, Yiheng Ju, Ying Li, Xiaoning Kang, Xiaojuan Yang, Tianye Niu, Xiaoming Xing, and Yun Lu. "Establishment and Clinical Application of an Artificial Intelligence Diagnostic Platform for Identifying Rectal Cancer Tumor Budding." In: *Frontiers in Oncology* 11 (2021), p. 320.

- [281] Xueyan Liu, Nan Li, Sheng Liu, Jun Wang, Ning Zhang, Xubin Zheng, Kwong-Sak Leung, and Lixin Cheng. "Normalization methods for the analysis of unbalanced transcriptome data: a review." In: *Frontiers in bioengineering and biotechnology* (2019), p. 358.
- [282] Ying-Chih Lo, I-Fang Chung, Shin-Ning Guo, Mei-Chin Wen, and Chia-Feng Juang. "Cycle-consistent GAN-based stain translation of renal pathology images with glomerulus detection application." In: *Applied Soft Computing* 98 (2021), p. 106822.
- [283] E López de Maturana, L Alonso, P Alarcón, I A Martín-Antoniano, S Pineda, L Piorno, M L Calle, and N Malats. "Challenges in the Integration of Omics and Non-Omics Data." In: *Genes* 10.3 (2019), p. 238.
- [284] David N Louis, Michael Feldman, Alexis B Carter, Anand S Dighe, John D Pfeifer, Lynn Bry, Jonas S Almeida, Joel Saltz, Jonathan Braun, John E Tomaszewski, et al. "Computational pathology: a path ahead." In: *Archives of pathology & laboratory medicine* 140.1 (2016), pp. 41–50.
- [285] Eve J Lowenstein. "Patient safety and the mother of all biases: Overconfidence." In: *International journal of women's dermatology* 6.2 (2020), p. 127.
- [286] Thomas Luechtefeld, Dan Marsh, Craig Rowlands, and Thomas Hartung. "Machine Learning of Toxicological Big Data Enables Read-Across Structure Activity Relationships (RASAR) Outperforming Animal Test Reproducibility." In: *Toxicological Sciences* 165 (1 2018), pp. 198–212.
- [287] Thomas Luechtefeld, Craig Rowlands, and Thomas Hartung. "Big-data and machine learning to revamp computational toxicology and its use in risk assessment." In: *Toxicology Research* 7 (5 2018), pp. 732–744.
- [288] T Ma and A Zhang. "Affinity network fusion and semi-supervised learning for cancer patient clustering." In: *Methods* 145 (2018), pp. 16–24.
- [289] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets." In: *Cell* 161.5 (2015), pp. 1202–1214.
- [290] Valerio Maggio, Marco Chierici, Giuseppe Jurman, and Cesare Furlanello. "A multiobjective deep learning approach for predictive classification in Neuroblastoma." In: *arXiv preprint arXiv:1711.08198* (2017).
- [291] A Mahmoud and et al. "Segmentation of Heavily Clustered Nuclei from Histopathological Images." In: *Sci Rep* 9.1 (2019), p. 4551.

- [292] Amal I Malik, Andrea Rowan-Carroll, Andrew Williams, Christine L Lemieux, Alexandra S Long, Volker M Arlt, David H Phillips, Paul A White, and Carole L Yauk. "Hepatic genotoxicity and toxicogenomic responses in Muta™ Mouse males treated with dibenz [a, h] anthracene." In: *Mutagenesis* 28.5 (2013), pp. 543–554.
- [293] Timothy Malloy and Elizabeth Beryt. "Leveraging the new predictive toxicology paradigm: alternative testing strategies in regulatory decision-making." In: *Environmental Science: Nano* 3 (6 2016), pp. 1380–1395.
- [294] Sayan Mandal, Aldo Guzmán-Sáenz, Niina Haiminen, Saugata Basu, and Laxmi Parida. "A Topological Data Analysis Approach on Predicting Phenotypes from Gene Expression Data." In: *Proc. Algorithms for Computational Biology 2020 (AlCoB)*. Vol. 12099. Lecture Notes in Computer Science. Springer International Publishing, 2020, pp. 178–187.
- [295] Shubham Manik, Lalit Mohan Saini, and Nikhil Vadera. "Counting and classification of white blood cell using artificial neural network (ANN)." In: *2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)*. IEEE. 2016, pp. 1–5.
- [296] Peter C Mann, John Vahle, Charlotte M Keenan, Julia F Baker, Alys E Bradley, Dawn G Goodman, Takanori Harada, Ronald Herbert, Wolfgang Kaufmann, Rupert Kellner, et al. "International harmonization of toxicologic pathology nomenclature: an overview and review of basic principles." In: *Toxicologic pathology* 40.4_suppl (2012), 7S–13S.
- [297] K. Mansouri et al. "CERAPP: Collaborative estrogen receptor activity prediction project." In: *Environmental Health Perspectives* 124.7 (2016), p. 1023.
- [298] Tuomo Mantere, Simone Kersten, and Alexander Hoischen. "Long-read sequencing emerging in medical genetics." In: *Frontiers in genetics* 10 (2019), p. 426.
- [299] R. Marée. "The Need for Careful Data Collection for Pattern Recognition in Digital Pathology." In: *Journal of Pathology Informatics* 8.1 (2017), p. 19.
- [300] R Marée. "Open practices and resources for collaborative digital pathology." In: *Front Med* 6 (2019), p. 255.
- [301] Ross Markello. *snfpy: Similarity Network Fusion in Python*. <https://bit.ly/3Hn9NcR>. 2019.
- [302] Marsh. *REPORT MEDMAL - Risk trend study from medical malpractice in health-care Italian public and private*. 12th ed. 2021.
- [303] Richard F Martin and J Bruce Beckwith. "Lymphoid infiltrates in neuroblastomas: Their occurrence and prognostic significance." In: *Journal of Pediatric Surgery* 3.1 (1968), pp. 161–164.

- [304] B. W. Matthews. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme." In: *Biochimica et Biophysica Acta* 405.2 (1975), pp. 442–451.
- [305] Marius E Mayerhoefer, Andrzej Materka, Georg Langs, Ida Häggström, Piotr Szczypiński, Peter Gibbs, and Gary Cook. "Introduction to radiomics." In: *Journal of Nuclear Medicine* 61.4 (2020), pp. 488–495.
- [306] Morgan P McBee, Omer A Awan, Andrew T Colucci, Comeron W Ghobadi, Nadja Kadom, Akash P Kansagra, Srinu Tridandapani, and William F Aufmann. "Deep learning in radiology." In: *Academic radiology* 25.11 (2018), pp. 1472–1480.
- [307] S D McCabe, D-Y Lin, and M I Love. "Consistency and overfitting of multi-omics methods on experimental data." In: *Brief Bioinform Advance Access* (2019), bbz070.
- [308] Matthew N McCall and Rafael A Irizarry. "Thawing Frozen Robust Multi-array Analysis (fRMA)." In: *BMC Bioinformatics* 12.1 (2011), p. 369.
- [309] Wendy McDougald, Christian Vanhove, Adrienne Lehnert, Barbara Lewellen, John Wright, Marco Mingarelli, Carlos Corral, Jurgen Schneider, Sven Plein, David Newby, et al. "Standardization of preclinical PET/CT imaging to improve quantitative accuracy, precision and reproducibility: a multi-center study." In: *Journal of Nuclear Medicine* (2019), jnumed–119.
- [310] L McInnes, J Healy, and J Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." arXiv:1802.03426. 2018.
- [311] L McInnes, J Healy, N Saul, and L Großberger. "UMAP: Uniform Manifold Approximation and Projection." In: *Journal of Open Source Software* 3.29 (2018), p. 861.
- [312] Leland McInnes, John Healy, and Steve Astels. "hdbscan: Hierarchical density based clustering." In: *Journal of Open Source Software* 2.11 (2017), p. 205.
- [313] Shima Mehrvar, Lauren E Himmel, Pradeep Babburi, Andrew L Goldberg, Magali Guffroy, Kyathanahalli Janardhan, Amanda L Krempley, and Bhupinder Bawa. "Deep learning approaches and applications in toxicologic histopathology: Current status and future perspectives." In: *Journal of Pathology Informatics* 12 (2021).
- [314] J Mehtonen et al. "Data-driven characterization of molecular phenotypes across heterogeneous sample collections." In: *Nucleic Acids Res* 47.13 (2019), e76.
- [315] Ombretta Melaiu et al. "PD-L1 Is a Therapeutic Target of the Bromodomain Inhibitor JQ1 and, Combined with HLA Class I, a Promising Prognostic Biomarker in Neuroblastoma." In: *Clinical Cancer Research* 23.15 (2017), pp. 4462–4472.

- [316] Ombretta Melaiu et al. "Cellular and gene signatures of tumor-infiltrating dendritic cells and natural-killer cells predict prognosis of neuroblastoma." In: *Nature Communications* 11.1 (2020), p. 5992.
- [317] C Meng, A Basunia, B Peters, A M Gholami, B Kuster, and A C Culhane. "MOGSA: Integrative Single Sample Gene-set Analysis of Multiple Omics Data." In: *Mol Cell Proteomics* 18.8 suppl 1 (2019), S153–S168.
- [318] "Method of the Year 2013." In: *Nature methods* 1.11 (2014).
- [319] H. H. Milioli, R. Vimieiro, C. Riveros, I. Tishchenko, R. Berretta, and P. Moscato. "The Discovery of Novel Biomarkers Improves Breast Cancer Intrinsic Subtype Prediction and Reconciles the Labels in the METABRIC Data Set." In: *PLoS ONE* 10.7 (2015), e0129711.
- [320] Marco Mina, Renata Boldrini, Arianna Citti, Paolo Romania, Valerio D'Alicandro, Maretta De Ioris, Aurora Castellano, Cesare Furlanello, Franco Locatelli, and Doriana Fruci. "Tumor-infiltrating T lymphocytes improve clinical outcome of therapy-resistant neuroblastoma." In: *OncoImmunology* 4.9 (2015), e1019981.
- [321] Martina Mirlacher, Marlis Kasper, Martina Storz, Yvonne Knecht, Ursula Dürmüller, Ronald Simon, Michael J Mihatsch, and Guido Sauter. "Influence of slide aging on results of translational research studies using immunohistochemistry." In: *Modern pathology* 17.11 (2004), pp. 1414–1420.
- [322] B B Misra, C Langefeld, M Olivier, and L A Cox. "Integrated omics: tools, advances and future approaches." In: *J Mol Endocrinol* 62 (2019), R21–R45.
- [323] Takashi Misu, Cindy M Kortepeter, Monica A Muñoz, Eileen Wu, and Gerald J Dal Pan. "An evaluation of "drug ineffective" postmarketing reports in drug safety surveillance." In: *Drugs-real world outcomes* 5.2 (2018), pp. 91–99.
- [324] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. "Spectral normalization for generative adversarial networks." In: *arXiv preprint arXiv:1802.05957* (2018).
- [325] P. Mobadersany, S. Yousefi, M. Amgad, D. A. Gutman, J. S. Barnholtz-Sloan, J. E. Velázquez Vega, D. J. Brat, and L. A. D. Cooper. "Predicting cancer outcomes from histology and genomics using convolutional networks." In: *Proceedings of the National Academy of Sciences* 115.13 (2018), E2970–E2979.
- [326] Yngve Mardal Moe, Aurora Rosvoll Groendahl, Martine Mulstad, Oliver Tomic, Ulf Indahl, Einar Dale, Eirik Malinen, and Cecilia Marie Futsaether. "Deep learning for automatic tumour segmentation in PET/CT images of patients with head and neck cancers." In: (2019).
- [327] Attayeb Mohsen, Lokesh P Tripathi, and Kenji Mizuguchi. "Deep Learning Prediction of Adverse Drug Reactions in Drug Discovery Using Open TG-GATEs and FAERS Databases." In: *Frontiers in Drug Discovery* (2021), p. 3.

- [328] Thomas J Moore, Hanzhe Zhang, Gerard Anderson, and G Caleb Alexander. "Estimated costs of pivotal trials for novel therapeutic agents approved by the US Food and Drug Administration, 2015-2016." In: *JAMA internal medicine* 178.11 (2018), pp. 1451–1457.
- [329] HL Morgan. "The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service." In: *Journal of Chemical Documentation* 5.2 (1965), pp. 107–113.
- [330] Manuel Muñoz-Aguirre, Vasilis F Ntasis, Santiago Rojas, and Roderic Guigó. "PyHIST: A Histological Image Segmentation Tool." In: *PLoS computational biology* 16.10 (2020), e1008349.
- [331] C. Nader Vasconcelos and B. Nader Vasconcelos. "Increasing deep learning melanoma classification by classical and expert knowledge based image transforms." In: *arXiv* 1702.07025 (2017).
- [332] Vinod Nair and Geoffrey E. Hinton. "Rectified Linear Units improve Restricted Boltzmann Machines." In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. 2010, pp. 807–814.
- [333] Raouf E Nakhleh. "Lost, mislabeled, and unsuitable surgical pathology specimens." In: *AJSP: Reviews & Reports* 8.3 (2003), pp. 98–102.
- [334] National Academies of Sciences, Engineering, and Medicine, Policy and Global Affairs. *Reproducibility and Replicability in Science*. National Academies Press, 2019.
- [335] Nicholas E Navin. "The first five years of single-cell cancer genomics and beyond." In: *Genome research* 25.10 (2015), pp. 1499–1507.
- [336] Sidra Nawaz and Yinyin Yuan. "Computational pathology: Exploring the spatial dimension of tumor ecology." In: *Cancer letters* 380.1 (2016), pp. 296–303.
- [337] Tom Needham. "Introduction to Applied Algebraic Topology." Course notes <https://bit.ly/32Ucqng>. 2019.
- [338] Mark D. Nelms, Claire L. Mellor, Steven J. Enoch, Richard S. Judson, Grace Patlewicz, Ann M. Richard, Judith M. Madden, Mark T. D. Cronin, and Stephen W. Edwards. "A Mechanistic Framework for Integrating Chemical Structure and High-Throughput Screening Results to Improve Toxicity Predictions." In: *Computational Toxicology* 8 (2018), pp. 1–12.
- [339] Ryan Neph, Yangsibo Huang, Youming Yang, and Ke Sheng. "DeepMCDose: A Deep Learning Method for Efficient Monte Carlo Beamlet Dose Calculation by Predictive Denoising in MR-Guided Radiotherapy." In: *Workshop on Artificial Intelligence in Radiation Therapy*. 2019, pp. 137–145.
- [340] Hannah T. Neprash, Alexander Everhart, Donna McAlpine, Laura Barrie Smith, Bethany Sheridan, and Dori A. Cross. "Measuring Primary Care Exam Length Using Electronic Health Record Data." In: *Medical Care* 59 (2021), pp. 62–66.

- [341] Hien M Nguyen, Eric W Cooper, and Katsuari Kamei. "Borderline over-sampling for imbalanced data classification." In: *Proceedings: Fifth International Workshop on Computational Intelligence & Applications*. Vol. 2009. 1. IEEE SMC Hiroshima Chapter. 2009, pp. 24–29.
- [342] Z. Ning, J. Luo, Y. Li, S. Han, Q. Feng, Y. Xu, W. Chen, C. Tao, and Y. Zhang. "Pattern Classification for Gastrointestinal Stromal Tumors by Integration of Radiomics and Deep Convolutional Features." In: *IEEE Journal of Biomedical and Health Informatics* May 29 (2018), [Epub ahead of print].
- [343] J. J. Nirschl, A. Janowczyk, E. G. Peyster, R. Frank, K. B. Margulies, M. D. Feldman, and A. Madabhushi. "A deep-learning classifier identifies patients with clinical heart failure using whole-slide images of H&E tissue." In: *PLOS ONE* 13.4 (2018), e0192726.
- [344] OECD. *Tackling Wasteful Spending on Health*. 2017, p. 304.
- [345] Augustus Odena, Vincent Dumoulin, and Chris Olah. *Deconvolution and Checkerboard Artifacts*. <https://bit.ly/34cw9j1>. Distill. 2016.
- [346] Travis E Oliphant. *A guide to NumPy*. Vol. 1. Trelgol Publishing USA, 2006.
- [347] F. Orhac, S. Boughdad, C. Philippe, H. Stalla-Bourdillon, C. Nioche, L. Champion, M. Soussan, F. Frouin, V. Frouin, and I. Buvat. "A postreconstruction harmonization method for multicenter radiomic studies in PET." In: *Journal of Nuclear Medicine* 59.8 (2018), pp. 1321–1328.
- [348] Nobuyuki Otsu. "A Threshold Selection Method from Gray-Level Histograms." In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.1 (1979), pp. 62–66.
- [349] Liron Pantanowitz, Ashish Sharma, Alexis B Carter, Tahsin Kurc, Alan Sussman, and Joel Saltz. "Twenty years of digital pathology: an overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives." In: *Journal of pathology informatics* 9 (2018).
- [350] R. B. Parikh, Z. Obermeyer, and A. S. Navathe. "Regulation of predictive analytics in medicine." In: *Science* 363.6429 (2019), pp. 810–812.
- [351] Prithvirajsinh Parmar, Himan Patel, Ashvin Mishra, Miteshkumar Malaviya, and Keyur Parmar. "Personalized medicine: Time for One person, one medicine." In: *International Journal of Pharmaceutics and Drug Analysis* (2021), pp. 86–92.
- [352] A. Parthasarathi and A. Dhawan. "In Silico Approaches for Predictive Toxicology." In: *In Vitro Toxicology*. Ed. by A. Dhawan and S. Kwon. Academic Press, 2018. Chap. 5, pp. 91–109.
- [353] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. "Automatic differentiation in PyTorch." In: (2017).

- [354] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In: *Proc. Advances in Neural Information Processing Systems 2019 (NeurIPS)*. Curran Associates, Inc., 2019, pp. 8024–8035.
- [355] R. Paul, S. H. Hawkins, Y. Balagurunathan, M. B. Schabath, R. J. Gillies, L. O. Hall, and D. B. Goldgof. "Deep Feature Transfer Learning in Combination with Traditional Features Predicts Survival Among Patients with Lung Adenocarcinoma." In: *Tomography* 2.4 (2016), pp. 388–395.
- [356] Adam Pearce. *Understanding UMAP*. <https://bit.ly/3Gk6sKn>. 2019.
- [357] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python." In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [358] C Peng, Y Zheng, and D-S Huang. "Capsule Network based Modeling of Multi-omics Data for Discovery of Breast Cancer-related Genes." In: *IEEE/ACM Trans Comput Biol Bioinform Early Access* (2019), pp. 1–8.
- [359] Jose A Perea and John Harer. "Sliding windows and persistence: an application of topological methods to signal analysis." In: *Foundations of Computational Mathematics* 15.3 (2015), pp. 799–838.
- [360] Sokol Petushi, Fernando U Garcia, Marian M Haber, Constantine Katsinis, and Aydin Tozeren. "Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer." In: *BMC medical imaging* 6.1 (2006), pp. 1–11.
- [361] Kathryn A Phillips, Patricia A Deverka, Gillian W Hooker, and Michael P Douglas. "Genetic test availability and spending: where are we now? Where are we going?" In: *Health affairs* 37.5 (2018), pp. 710–716.
- [362] Hans Pinckaers, Wouter Bulten, Jeroen Van der Laak, and Geert Litjens. "Detection of prostate cancer in whole-slide images through end-to-end training with image-level labels." In: *arXiv preprint arXiv:2006.03394* (2020).
- [363] Hans Pinckaers, Bram van Ginneken, and Geert Litjens. "Streaming convolutional neural networks for end-to-end learning with multi-megapixel images." In: *arXiv preprint arXiv:1911.04432* (2019).
- [364] Hannah Pischon, David Mason, Bettina Lawrenz, Olivier Blanck, Anna-Lena Frisk, Frederic Schorsch, and Valeria Bertani. "Artificial Intelligence in Toxicologic Pathology: Quantitative Evaluation of Compound-Induced Hepatocellular Hypertrophy in Rats." In: *Toxicologic Pathology* 49.4 (2021), pp. 928–937.
- [365] Sean P Pitroda and Ralph R Weichselbaum. "Integrated molecular and clinical staging defines the spectrum of metastatic cancer." In: *Nature Reviews Clinical Oncology* 16.9 (2019), pp. 581–588.
- [366] O Poirion, K Chaudhary, S Huang, and L X Garmire. "Multi-omics-based pancreatic cancer prognosis prediction using an ensemble of deep-learning and machine-learning models." medRxiv 19010082. 2019.

- [367] Ivan O Potapenko, Torben Lüders, Hege G Russnes, Åslaug Helland, Therese Sørli, Vessela N Kristensen, Silje Nord, Ole C Lingjærde, Anne-Lise Børresen-Dale, and Vilde D Haakensen. "Glycan-related gene expression signatures in breast cancer subtypes; relation to survival." In: *Molecular oncology* 9.4 (2015), pp. 861–876.
- [368] L Prélot et al. "Machine Learning in Multi-Omics Data to Assess Longitudinal Predictors of Glycaemic Trait Levels." bioRxiv 358390. 2018.
- [369] Ross L Prentice. "Surrogate endpoints in clinical trials: definition and operational criteria." In: *Statistics in medicine* 8.4 (1989), pp. 431–440.
- [370] B M Pucher, O A Zeleznik, and G G Thallinger. "Comparison and evaluation of integrative methods for the analysis of multilevel omics data: a study based on simulated and experimental cancer data." In: *Brief Bioinform* 28 (2018), pp. 1–11.
- [371] Munish Puri. "Automated machine learning diagnostic support system as a computational biomarker for detecting drug-induced liver injury patterns in whole slide liver pathology images." In: *Assay and drug development technologies* 18.1 (2020), pp. 1–10.
- [372] B. S. Purohit, A. Ailianou, N. Dulguerov, C. D. Becker, O. Ratib, and M. Becker. "FDG-PET/CT pitfalls in oncological head and neck imaging." In: *Insights into Imaging* 5.5 (2014), pp. 585–602.
- [373] Dahui Qin. "Next-generation sequencing and its clinical application." In: *Cancer biology & medicine* 16.1 (2019), p. 4.
- [374] C Qiu et al. "Multi-omics Data Integration for Identifying Osteoporosis Biomarkers and Their Biological Interaction and Causal Mechanisms." In: *ISCIENCE Journal pre-Proof* (2020), p. 100847.
- [375] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017.
- [376] Arwa B. Raies and Vladimir B. Bajic. "In silico toxicology: computational methods for the prediction of chemical toxicity." In: *WIREs Computational Molecular Science* 6.2 (2016), pp. 147–172.
- [377] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. "AI in health and medicine." In: *Nature Medicine* (2022), pp. 1–8.
- [378] Uppada Rajyalakshmi, S Koteswara Rao, and K Satya Prasad. "Supervised classification of breast cancer malignancy using integrated modified marker controlled watershed approach." In: *2017 IEEE 7th International Advance Computing Conference (IACC)*. IEEE. 2017, pp. 584–589.
- [379] N Rappoport and R Shamir. "NEMO: cancer subtyping by integration of partial multi-omic data." In: *Bioinformatics* 35.18 (2019), pp. 3348–3356.

- [380] Nimrod Rappoport and Ron Shamir. "Multi-omic and multi-view clustering algorithms: review and cancer benchmark." In: *Nucleic acids research* 46.20 (2018), pp. 10546–10562.
- [381] S. Raschka. "Model evaluation, model selection, and algorithm selection in machine learning." arXiv:1811.12808v3. 2020.
- [382] Joseph Redmon and Ali Farhadi. "YOLO9000: Better, faster, stronger." In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition 2017 (CVPR)*. IEEE, 2016, pp. 6517–6525.
- [383] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. "Color transfer between images." In: *IEEE Computer graphics and applications* 21.5 (2001), pp. 34–41.
- [384] Reisfeld, Brad and Mayeno, Arthur N. "What is computational toxicology?" In: *Computational Toxicology*. Springer, 2012, pp. 3–7.
- [385] Steven Reisman, Thomas Hatzopoulos, Konstantin Läufer, George K Thiruvathukal, and Catherine Putonti. "A Polyglot Approach to Bioinformatics Data Integration: A Phylogenetic Analysis of HIV-1." In: *Evol Bioinform Online* 12 (2016), pp. 23–7.
- [386] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (2017), pp. 1137–1149.
- [387] Bastian Rieck, Tristan Yates, Christian Bock, Karsten Borgwardt, Guy Wolf, Nicholas Turk-Browne, and Smita Krishnaswamy. "Uncovering the topology of time-varying fMRI data using cubical persistence." In: *Proc. Advances in Neural Information Processing Systems 2020 (NeurIPS)*. Vol. 33. Curran Associates Inc., 2020, pp. 6900–6912.
- [388] Henri Riihimäki, Wojciech Chachólski, Jakob Theorell, Jan Hillert, and Ryan Ramanujam. "A topological data analysis based classification method for multiple measurements." In: *BMC Bioinformatics* 21.1 (2020), p. 336.
- [389] Daniela Rodrigues, Terezinha Souza, Danyel GJ Jennen, Lieve Lemmens, Jos CS Kleinjans, and Theo M de Kok. "Drug-induced gene expression profile changes in relation to intestinal toxicity: state-of-the-art and new approaches." In: *Cancer treatment reviews* 77 (2019), pp. 57–66.
- [390] David Rogers and Mathew Hahn. "Extended-connectivity fingerprints." In: *Journal of Chemical Information and Modeling* 50.5 (2010), pp. 742–754.
- [391] F Rohart, B Gautier, A Singh, and K-A Lê Cao. "mixOmics: An R package for 'omics feature selection and multiple data integration." In: *PLoS Comput Biol* 13.11 (2017), e1005752.

- [392] Rocio Romero-Zaliz and Juan Francisco Reinoso-Gordo. "An Updated Review on Watershed Algorithms." In: *Soft Computing for Sustainability Science*. Springer, 2017, pp. 235–258.
- [393] Guoguang Rong, Arnaldo Mendez, Elie Bou Assi, Bo Zhao, and Mohamad Sawan. "Artificial intelligence in healthcare: review and prediction case studies." In: *Engineering* 6.3 (2020), pp. 291–301.
- [394] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." In: *Proc. Medical Image Computing and Computer-Assisted Intervention 2015 (MICCAI)*. Vol. 9351. Lecture Notes in Computer Science. Springer International Publishing, 2015, pp. 234–241.
- [395] L E Rosen and P Gattuso. "Neuroendocrine tumors of the breast." In: *Arch Pathol Lab Med* 141.11 (2017), pp. 1577–1581.
- [396] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger. "Concurrent Spatial and Channel 'Squeeze & Excitation' in Fully Convolutional Networks." In: *Proc. Medical Image Computing and Computer-Assisted Intervention 2018 (MICCAI)*. Vol. 11070. Lecture Notes in Computer Science. Springer International Publishing, 2018, pp. 421–429.
- [397] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger. "Recalibrating Fully Convolutional Networks With Spatial and Channel "Squeeze and Excitation" Blocks." In: *IEEE Transactions on Medical Imaging* 38.2 (2018), pp. 540–549.
- [398] Royal Chemical Society. *Royal Chemical Society view on predictive toxicology*. <https://rsc.li/3ukfvJ5>. Accessed: 2018-08-29. 2016.
- [399] Daniel Rudmann, Jay Albretsen, Colin Doolan, Mark Gregson, Beth Dray, Aaron Sargeant, Donal O'Shea D, Jogile Kuklyte, Adam Power, and Jenny Fitzgerald. "Using deep learning artificial intelligence algorithms to verify N-nitroso-N-methylurea and urethane positive control proliferative changes in Tg-RasH2 mouse carcinogenicity studies." In: *Toxicologic Pathology* 49.4 (2021), pp. 938–949.
- [400] Héctor A. Rueda-Zárate, Iván Imaz-Rosshandler, Roberto A. Cárdenas-Ovando, Juan E. Castillo-Fernández, Julieta Noguez-Monroy, and Claudia Rangel-Escareño. "A computational toxicogenomics approach identifies a list of highly hepatotoxic compounds from a large microarray database." In: *PLOS ONE* 12.4 (2017). Ed. by Alok Deoraj, e0176284.
- [401] Kane SP. *The Top 300 of 2014, ClinCalc DrugStats Database*. ClinCalc: <https://bit.ly/3rj3iSJ>. Updated September 15, 2021. Accessed: January 6, 2022.

- [402] Tim Sainburg, Leland McInnes, and Timothy Q Gentner. "Parametric UMAP: learning embeddings with deep neural networks for representation and semi-supervised learning." arXiv:2009.12981. 2020.
- [403] Neha Saini, Shikha Bakshi, and Sadhna Sharma. "In-silico approach for drug induced liver injury prediction: Recent advances." In: *Toxicology Letters* 295 (2018), pp. 288–295.
- [404] R. Salgado and L. Sherene. "Tumour infiltrating lymphocytes in breast cancer: increasing clinical relevance." In: *The Lancet Oncology* 19.1 (2018), pp. 3–5.
- [405] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. "Improved techniques for training gans." In: *Advances in neural information processing systems* 29 (2016), pp. 2234–2242.
- [406] Joel Saltz et al. "Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images." In: *Cell Reports* 23.1 (2018), 181–193.e7.
- [407] Massimo Salvi, U Rajendra Acharya, Filippo Molinari, and Kristen M Meiburger. "The impact of pre-and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis." In: *Computers in Biology and Medicine* (2020), p. 104129.
- [408] Massimo Salvi, U. Rajendra Acharya, Filippo Molinari, and Kristen M. Meiburger. "The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis." In: *Computers in Biology and Medicine* 128 (2021), p. 104129.
- [409] A Sathyanarayanan, R Gupta, E W Thompson, D R Nyholt, D C Bauer, and S H Nagaraj. "A comparative study of multi-omics integration tools for cancer driver gene identification and tumour subtyping." In: *Brief Bioinform Advance Access* (2019), bbz121.
- [410] TM Shahriar Sazzad, LJ Armstrong, and Amiya K Tripathy. "An automated ovarian tissue detection approach using type P63 non-counter stained images to minimize pathology experts observation variability." In: *2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*. IEEE. 2016, pp. 155–159.
- [411] Nadine S Schaadt and *et al.* "Graph-based description of tertiary lymphoid organs at single-cell level." In: *PLoS computational biology* 16.2 (2020), e1007385.
- [412] Hal Schenck. *Computational Algebraic Geometry*. Vol. 58. Cambridge University Press, 2003.
- [413] M V Schneider and R C Jimenez. "Teaching the Fundamentals of Biological Data Integration Using Classroom Games." In: *PLoS Comput Biol* 8.12 (2012), e1002789.

- [414] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. "Hidden technical debt in machine learning systems." In: *Proc. 28th International Conference on Neural Information Processing Systems (NIPS)*. ACM, 2015, pp. 2503–2511.
- [415] Consolato M Sergi. "Digital pathology: the time is now to bridge the gap between medicine and technological singularity." In: *Interactive Multimedia-Multimedia Production and Digital Storytelling*. IntechOpen, 2019.
- [416] H Sharifi-Noghabi, O Zolotareva, C C Collins, and M Ester. "MOLI: multiomics late integration with deep neural networks for drug response prediction." In: *Bioinformatics* 35.14 (2019), pp. i501–i509.
- [417] Chandra K Sharma and Monika Sharma. "Biomarkers are the Need of the Present Era." In: *The Open Biomarkers Journal* 9.1 (2019).
- [418] Chen Shen and Vic Patrangenaru. "Topological Object Data Analysis Methods with an Application to Medical Imaging." In: *Proc. Functional and High-Dimensional Statistics and Related Fields 2020 (IWFOS)*. Contribution to Statistics. Springer International Publishing, 2020, pp. 237–244.
- [419] L. Shi et al. "The international MAQC Society launches to enhance reproducibility of high-throughput technologies." In: *Nature Biotechnology* 35.12 (2017), pp. 1127–1128.
- [420] M. Shibusaki, K. Maeda, H. Nagahara, T. Fukuoka, Y. Iseki, S. Matsutani, S. Kashiwagi, H. Tanaka, K. Hirakawa, and M. Ohira. "Tumor-infiltrating Lymphocytes Predict the Chemotherapeutic Outcomes in Patients with Stage IV Colorectal Cancer." In: *In Vivo* 32.1 (2018), pp. 151–158.
- [421] Daichi Shigemizu, Taiki Mori, Shintaro Akiyama, Sayuri Higaki, Hiroshi Watanabe, Takashi Sakurai, Shumpei Niida, and Kouichi Ozaki. "Identification of potential blood biomarkers for early diagnosis of Alzheimer's disease through RNA sequencing analysis." In: *Alzheimer's research & therapy* 12.1 (2020), pp. 1–12.
- [422] Robert H. Shoemaker. "The NCI60 human tumour cell line anticancer drug screen." In: *Nature Reviews Cancer* 6.10 (2006), pp. 813–823.
- [423] William H Shrank, Teresa L Rogstad, and Natasha Parekh. "Waste in the US health care system: estimated costs and potential for savings." In: *Jama* 322.15 (2019), pp. 1501–1509.
- [424] Lin Shui, Haoyu Ren, Xi Yang, Jian Li, Ziwei Chen, Cheng Yi, Hong Zhu, and Pixian Shui. "The era of radiogenomics in precision medicine: an emerging approach to support diagnosis, treatment decisions, and prognostication in oncology." In: *Frontiers in Oncology* (2021), p. 3195.

- [425] Joshua D Shur, Simon J Doran, Santosh Kumar, Derfel Ap Dafydd, Kate Downey, James PB O'Connor, Nikolaos Papanikolaou, Christina Messiou, Dow-Mu Koh, and Matthew R Orton. "Radiomics in oncology: A practical guide." In: *RadioGraphics* 41.6 (2021), pp. 1717–1732.
- [426] K Simonyan and A Zisserman. "Very deep convolutional networks for large-scale image recognition." In: *arXiv preprint arXiv:1409.1556* (2014).
- [427] Dalmeet Singh Chawla. "Critiqued coronavirus simulation gets thumbs up from code-checking efforts." In: *Nature* (2020), pp. 323–324.
- [428] Gurjeet Singh, Facundo Memoli, and Gunnar Carlsson. "Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition." In: *Proc. IEEE/Eurographics Symposium on Point-Based Graphics 2007 (PBG)*. The IEEE/Eurographics Association, 2007, pp. 91–100.
- [429] Brandi Patrice Smith, Loretta Sue Auvil, Michael Welge, Colleen Bannon Bushell, Rohit Bhargava, Navin Elango, Kamin Johnson, and Zeynep Madak-Erdogan. "Identification of early liver toxicity gene biomarkers using comparative supervised machine learning." In: *Scientific reports* 10.1 (2020), pp. 1–27.
- [430] Zhigang Song et al. "Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning." In: *Nature Communications* 11.1 (2020), p. 4294.
- [431] Brian B Spear, Margo Heath-Chiozzi, and Jeffrey Huff. "Clinical application of pharmacogenetics." In: *Trends in molecular medicine* 7.5 (2001), pp. 201–204.
- [432] N K Speicher and N Pfeifer. "Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery." In: *Bioinformatics* 31.12 (2015), pp. i268–i275.
- [433] Sasha E Stanton and Mary L Disis. "Clinical significance of tumor-infiltrating lymphocytes in breast cancer." In: *Journal for Immunotherapy of Cancer* 4.1 (2016), p. 59.
- [434] G L Stein-O'Brien et al. "Enter the Matrix: Factorization Uncovers Knowledge from Omics." In: *Trends Genet* 34.10 (2018), pp. 790–805.
- [435] Sarah L Stenton and Holger Prokisch. "The clinical application of RNA sequencing in genetic diagnosis of Mendelian disorders." In: *Clinics in Laboratory Medicine* 40.2 (2020), pp. 121–133.
- [436] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. "A deep learning approach to antibiotic discovery." In: *Cell* 180.4 (2020), pp. 688–702.
- [437] R Stone. *py-wsi*. <https://github.com/ysbecca/py-wsi>. 2020.

- [438] E. S. Stovgaard, D. Nielsen, E. Hogdall, and E. Balslev. "Triple negative breast cancer—prognostic role of immune-related factors: a systematic review." In: *Acta Oncologica* 57.1 (2018), pp. 74–82.
- [439] Iori Sumida, Taiki Magome, Hideki Kitamori, Indra J Das, Hajime Yamaguchi, Hisao Kizaki, Keiko Aboshi, Kyohei Yamashita, Yuji Yamada, Yuji Seo, et al. "Deep convolutional neural network for reduction of contrast-enhanced region on CT images." In: *Journal of radiation research* (2019).
- [440] Wenqing Sun, Bin Zheng, and Wei Qian. "Automatic feature learning using multichannel ROI based on deep structured algorithms for computerized lung cancer diagnosis." In: *Computers in Biology and Medicine* 89 (2017), pp. 530–539.
- [441] Yujia Sun and Jan Platoš. "High-Dimensional Text Clustering by Dimensionality Reduction and Improved Density Peak." In: *Wireless Communications and Mobile Computing* 2020 (2020), Article ID 8881112.
- [442] Karen Boaz Supriya Nikita Kapila and Srikant Natarajan. "The post-analytical phase of histopathology practice: Storage, retention and use of human tissue specimens." In: *International journal of applied & basic medical research* 6.1 (2016), pp. 3–7.
- [443] Zaneta Swiderska-Chadaj, Hans Pinckaers, Mart van Rijthoven, Maschenka Balkenhol, Margarita Melnikova, Oscar Geessink, Quirine Manson, Mark Sherman, Antonio Polonia, Jeremy Parry, et al. "Learning to detect lymphocytes in immunohistochemistry with deep learning." In: *Medical image analysis* 58 (2019), p. 101547.
- [444] Zaneta Swiderska-Chadaj, Hans Pinckaers, Mart van Rijthoven, Maschenka Balkenhol, Margarita Melnikova, Oscar Geessink, Quirine Manson, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. *Convolutional Neural Networks for Lymphocyte detection in Immunohistochemically Stained Whole-Slide Images*. Poster at Medical Imaging with Deep Learning 2018 (MIDL). 2018.
- [445] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. "Rethinking the Inception Architecture for Computer Vision." In: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 2818–2826.
- [446] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition 2015 (CVPR)*. IEEE, 2015, pp. 1–9.
- [447] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. "Convolutional neural networks for medical image analysis: Full training or fine tuning?" In: *IEEE Transactions on Medical Imaging* 35.5 (2016), pp. 1299–1312.

- [448] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. "MnasNet: Platform-Aware Neural Architecture Search for Mobile." In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition 2019 (CVPR)*. 2019, pp. 2820–2828.
- [449] Mingxing Tan and Quoc Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." In: *Proc. International Conference on Machine Learning 2019 (ICML)*. Vol. 97. Proc. Machine Learning Research. PMLR, 2019, pp. 6105–6114.
- [450] Hao Tang, Xuming Chen, Yang Liu, Zhipeng Lu, Junhua You, Mingzhou Yang, Shengyu Yao, Guoqi Zhao, Yi Xu, Tingfeng Chen, et al. "Clinically applicable deep learning framework for organs at risk delineation in CT images." In: *Nature Machine Intelligence* (2019), pp. 1–12.
- [451] Jing Rui Tang, Nor Ashidi Mat Isa, and Ewe Seng Ch'ng. "A fuzzy-c-means-clustering approach: Quantifying chromatin pattern of non-neoplastic cervical squamous cells." In: *PloS one* 10.11 (2015), e0142830.
- [452] Xiaoning Tang, Yongmei Huang, Jinli Lei, Hui Luo, and Xiao Zhu. "The single-cell sequencing: new developments and medical applications." In: *Cell & bio-science* 9.1 (2019), pp. 1–9.
- [453] S Tarazona, L Balzano-Nogueira, and A Conesa. "Multiomics Data Integration in Time Series Experiments." In: *Data Analysis for Omic Sciences: Methods and Applications*. Vol. 82. Comprehensive Analytical Chemistry. Elsevier, 2018. Chap. 18, pp. 505–532.
- [454] Guillaume Tautin, Umberto Lupo, Lewis Tunstall, Julian Burella Pérez, Matteo Caorsi, Anibal Medina-Mardones, Alberto Dassatti, and Kathryn Hess. "giotto-tda: A Topological Data Analysis Toolkit for Machine Learning and Data Exploration." arXiv:2004.02551, NeurIPS 2020 TDA and Beyond Workshop. 2020.
- [455] DeepChem Team. *DeepChem: Democratizing Deep-Learning for Drug Discovery, Quantum Chemistry, Materials Science and Biology*. <https://bit.ly/3L13XQP>. Accessed: 2018-08-29. 2016.
- [456] David Tellez, Diederik Höppener, Cornelis Verhoef, Dirk Grünhagen, Pieter Nierop, Michal Drozdal, Jeroen Laak, and Francesco Ciompi. "Extending unsupervised neural image compression with supervised multitask learning." In: *Medical Imaging with Deep Learning*. PMLR. 2020, pp. 770–783.
- [457] David Tellez, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. "Neural Image Compression for Gigapixel Histopathology Image Analysis." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.2 (2021), pp. 567–578.

- [458] Arvind Thakkar, Hemanth Raj, Ravishankar, Bhaskaran Muthuvelan, Arun Balakrishnan, and Muralidhara Padigaru. "High Expression of Three-Gene Signature Improves Prediction of Relapse-Free Survival in Estrogen Receptor-Positive and Node-Positive Breast Tumors." In: *Biomarker Insights* 10 (2015).
- [459] The Cancer Genome Atlas Research Network. "Comprehensive molecular portraits of human breast tumours." In: *Nature* 490.7418 (2012), p. 61.
- [460] The Cancer Genome Atlas Research Network. "Comprehensive molecular characterization of clear cell renal cell carcinoma." In: *Nature* 499.7456 (2013), p. 43.
- [461] The Cancer Genome Atlas Research Network. "Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia." In: *New England Journal of Medicine* 368.22 (2013), pp. 2059–2074.
- [462] The GTEx Consortium. "The genotype-tissue expression (GTEx) project." In: *Nature Genetics* 45.6 (2013), pp. 580–585.
- [463] The MAQC Consortium. "The MAQC-II Project: A comprehensive study of common practices for the development and validation of microarray-based predictive models." In: *Nature Biotechnology* 28.8 (2010), pp. 827–838.
- [464] The MAQC Consortium. "The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models." In: *Nat Biotechnol* 28.8 (2010), pp. 827–838.
- [465] The SEQC/MAQC-III Consortium. "A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequence Quality Control consortium." In: *Nature Biotechnology* 32 (2014), 903–914.
- [466] Debra A Tokarz, Thomas J Steinbach, Avinash Lokhande, Gargi Srivastava, Rajesh Ugalmugle, Carroll A Co, Keith R Shockley, Emily Singletary, Mark F Cesta, Heath C Thomas, et al. "Using Artificial Intelligence to Detect, Classify, and Objectively Score Severity of Rodent Cardiomyopathy." In: *Toxicologic Pathology* 49.4 (2021), pp. 888–896.
- [467] K. Tomczak, P. Czerwińska, and M. Wiznerowicz. "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge." In: *Contemporary oncology* 19.1A (2015), A68.
- [468] E. J. Topol. "High-performance medicine: the convergence of human and artificial intelligence." In: *Nature Medicine* 25.1 (2019), p. 44.
- [469] Eric J Topol, Sarah S Murray, and Kelly A Frazer. "The genomics gold rush." In: *Jama* 298.2 (2007), pp. 218–221.
- [470] Eric Topol. *The creative destruction of medicine: How the digital revolution will create better health care*. Basic books, 2012.

- [471] V A Traag, L Waltman, and N J van Eck. "From Louvain to Leiden: guaranteeing well-connected communities." In: *Sci Rep* 9 (2019), p. 5233.
- [472] L Trastulla. "Techniques of integration for high-throughput omics data." MA thesis. Trento, Italy: Department of Mathematics, University of Trento, 2016.
- [473] Jan G Van den Tweel and Clive R Taylor. "A brief history of pathology." In: *Virchows Archiv* 457.1 (2010), pp. 3–10.
- [474] U.S. Food and Drug Administration. *FDA's predictive toxicology roadmap*. <https://bit.ly/35IrKEZ>. Accessed: 2018-08-29. 2017.
- [475] M Uhlen et al. "A pathology atlas of the human cancer transcriptome." In: *Science* 357.6352 (2017), eaan2507.
- [476] Leonardo Uieda, Santiago Rubén Soler, Rémi Rampin, Hugo van Kemenade, Matthew Turk, Daniel Shapero, Anderson Banihirwe, and John Leeman. "Pooch: A friend to fetch your data files." In: *Journal of Open Source Software* 5.45 (2020), p. 1943.
- [477] Rakesh Vaja and Meenal Rana. "Drugs and the liver." In: *Anaesthesia & Intensive Care Medicine* (2020).
- [478] Andreu Vall, Yogesh Sabnis, Jiye Shi, Reiner Class, Sepp Hochreiter, and Günter Klambauer. "The promise of AI for DILI prediction." In: *Frontiers in Artificial Intelligence* 4 (2021), p. 15.
- [479] M Vallieres, E Kay-Rivest, and et al. "Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer." In: *Scientific reports* 7.1 (2017), pp. 1–14.
- [480] L. Van der Maaten and G. Hinton. "Visualizing data using t-SNE." In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605.
- [481] V Vantaku et al. "Multi-omics Integration Analysis Robustly Predicts High-Grade Patient Survival and Identifies CPT1B Effect on Fatty Acid Metabolism in Bladder Cancer." In: *Clin Cancer Res* 25.12 (2019), pp. 3689–3701.
- [482] Ashley J Vargas and Curtis C Harris. "Biomarker development in the precision medicine era: lung cancer as a case study." In: *Nature Reviews Cancer* 16.8 (2016), pp. 525–537.
- [483] Frederick S. Varn, Yue Wang, David W. Mullins, Steven Fiering, and Chao Cheng. "Systematic Pan-Cancer Analysis Reveals Immune Cell Interactions in the Tumor Microenvironment." In: *Cancer Research* 77.6 (2017), pp. 1271–1282.
- [484] J Van der Veen, S Willems, S Deschuymer, D Robben, W Crijs, F Maes, and S Nuyts. "Benefits of deep learning for delineation of organs at risk in head and neck cancer." In: *Radiotherapy and Oncology* 138 (2019), pp. 68–74.

- [485] C. Vendt. "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository." In: *Journal of Digital Imaging* 26.6 (2013), pp. 1045–57.
- [486] L P C Verbeke, J Van den Eynden, A C Fierro, P Demeester, J Fostier, and K Marchal. "Pathway Relevance Ranking for Tumor Samples through Network-Based Data Integration." In: *PLoS One* 10.7 (2015), e0133503.
- [487] A. Vial, D. Stirling, M. Field, M. Ros, C. Ritz, M. Carolan, L. Holloway, and A.A. Miller. "The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review." In: *Translational Cancer Research* 7.3 (2018), pp. 803–816.
- [488] L. Vincent. "Morphological grayscale reconstruction in image analysis: applications and efficient algorithms." In: *IEEE Transactions on Image Processing* 2.2 (1993), pp. 176–201.
- [489] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. "SciPy 1.0: fundamental algorithms for scientific computing in Python." In: *Nature methods* 17.3 (2020), pp. 261–272.
- [490] Sophie Visvikis-Siest, Danai Theodoridou, Maria-Spyridoula Kontoe, Satish Kumar, and Michael Marschler. "Milestones in Personalized Medicine: From the Ancient Time to Nowadays—the Provocation of COVID-19." In: *Frontiers in Genetics* 11 (2020), p. 1442.
- [491] Dimitris Visvikis, Catherine Cheze Le Rest, Vincent Jaouen, and Mathieu Hatt. "Artificial intelligence, machine (deep) learning and radio(geno)mics: definitions and nuclear medicine imaging applications." In: *European journal of nuclear medicine and molecular imaging* 41.6 (July 2019), p. 1369.
- [492] Antonia Vlahou, Fulvio Magni, Harald Mischak, and Jerome Zoidakis. *Integration of Omics Approaches and Systems Biology for Clinical Applications*. John Wiley & Sons, 2018.
- [493] Aleksandar Vodovnik. "Diagnostic time in digital pathology: A comparative study on 400 cases." In: *Journal of pathology informatics* 7 (2016).
- [494] Kieran Walsh, Mircea A Voineagu, Fatemeh Vafae, and Irina Voineagu. "TDAview: an online visualization tool for topological data analysis." In: *Bioinformatics* 36.18 (2020), pp. 4805–4809.
- [495] S Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu. "scikit-image: image processing in Python." In: *PeerJ* 2 (2014), e453.
- [496] Bo Wang, Aziz M Mezlini, Feyyaz Demir, et al. "Similarity network fusion for aggregating data types on a genomic scale." In: *Nature methods* 11.3 (2014), pp. 333–337.

- [497] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. "Score-CAM: Score-weighted visual explanations for convolutional neural networks." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2020, pp. 24–25.
- [498] J. Wang and L. Perez. "The effectiveness of data augmentation in image classification using deep learning." In: *arXiv 1712.04621* (2017).
- [499] Shidan Wang et al. "Computational Staining of Pathology Images to Study the Tumor Microenvironment in Lung Cancer." In: *Cancer Research* 80.10 (2020), pp. 2056–2066.
- [500] P.C.H. Weaver. *Understanding Your Biopsy Results and Pathology Report*. <https://bit.ly/3Gk5ZYD>. Updated October 2021. Accessed: January 6, 2022.
- [501] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer, 2016. URL: <http://ggplot2.org>.
- [502] Judith Wienke, Miranda P Dierselhuis, Godelieve A M Tytgat, Annette Künkele, Stefan Nierkens, and Jan J Molenaar. "The immune landscape of neuroblastoma: Challenges and opportunities for novel therapeutic strategies in pediatric oncology." In: *European Journal of Cancer* 144" (2021), pp. 123–150.
- [503] M. D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship." In: *Scientific Data* 3 (2016), p. 160018.
- [504] Martin J Willeminck, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R Folio, Ronald M Summers, Daniel L Rubin, and Matthew P Lungren. "Preparing medical imaging data for machine learning." In: *Radiology* 295.1 (2020), pp. 4–15.
- [505] E. Williams et al. "Image Data Resource: a bioimage data integration and publication platform." In: *Nature Methods* 14.8 (2017), p. 775.
- [506] K Y Win and S Choomchuay. "Automated segmentation of cell nuclei in cytology pleural fluid images using OTSU thresholding." In: *International Conference on Digital Arts, Media and Technology (ICDAMT 2017)*. IEEE, 2017, pp. 14–18.
- [507] David S Wishart, Brendan Bartok, Eponine Oler, Kevin YH Liang, Zachary Budinski, Mark Berjanskii, AnChi Guo, Xuan Cao, and Michael Wilson. "MarkerDB: an online database of molecular biomarkers." In: *Nucleic Acids Research* 49.D1 (2021), pp. D1259–D1267.
- [508] Less Wright. *Ranger-Deep-Learning-Optimizer*. <https://bit.ly/3IUDFhc>. 2020.
- [509] C Wu, F Zhou, J Ren, X Li, Y Jiang, and S Ma. "A Selective Review of Multi-Level Omics Data Integration Using Variable Selection." In: *High-Throughput* 8 (2019), p. 4.

- [510] Fengying Wu, Jue Fan, Yayi He, Anwen Xiong, Jia Yu, Yixin Li, Yan Zhang, Wencheng Zhao, Fei Zhou, Wei Li, et al. "Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer." In: *Nature communications* 12.1 (2021), pp. 1–11.
- [511] A Xu, J Chen, H Peng, G Q Han, and H Cai. "Simultaneous Interrogation of Cancer Omics to Identify Subtypes With Significant Clinical Differences." In: *Front Genet* 10 (2019), p. 236.
- [512] Pavel Yakubovskiy. *Segmentation Models Pytorch*. <https://bit.ly/3L1DlZ9>. GitHub repository. 2020.
- [513] K. Yan, X. Wang, L. Lu, and R. M. Summers. "DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning." In: *Journal of Medical Imaging (Bellingham)* 5.3 (2018), p. 036501.
- [514] B Yang, Y Zhang, S Pang, X Shang, X Zhao, and M Han. "Integrating Multi-Omic Data with Deep Subspace Fusion Clustering for Cancer Subtype Prediction." In: *IEEE/ACM Trans Comput Biol Bioinform Early Access* (2019), pp. 1–12.
- [515] Hongbin Yang, Lixia Sun, Weihua Li, Guixia Liu, and Yun Tang. "In Silico Prediction of Chemical Toxicity for Drug Design Using Machine Learning Methods and Structural Alerts." In: *Frontiers in Chemistry* 6 (2018), p. 30.
- [516] Qiming Yang, Hongyang Chao, Dan Nguyen, and Steve Jiang. "A Novel Deep Learning Framework for Standardizing the Label of OARs in CT." In: *Workshop on Artificial Intelligence in Radiation Therapy*. 2019, pp. 52–60.
- [517] X-T Yu and T Zeng. "Integrative Analysis of Omics Big Data." In: *Computational Systems Biology*. Vol. 1754. Methods in Molecular Biology. Springer, 2018. Chap. 7, pp. 109–135.
- [518] Sergey Zagoruyko and Nikos Komodakis. "Wide Residual Networks." In: *Proc. British Machine Vision Conference 2016 (BMVC)*. BMVA Press, 2016, pp. 87.1–87.12.
- [519] A Zandonà. "Predictive networks for multi meta-omics data integration." <http://eprints-phd.biblio.unitn.it/2547/>. PhD thesis. Trento, Italy: Doctoral Programme in Biomolecular Sciences, University of Trento, 2017.
- [520] Mark D Zarella, Douglas Bowman, Famke Aeffner, Navid Farahani, Albert Xthona, Syeda Fatima Absar, Anil Parwani, Marilyn Bui, and Douglas J Hartman. "A Practical Guide to Whole Slide Imaging: A White Paper From the Digital Pathology Association." In: *Archives of Pathology & Laboratory Medicine* 143.2 (2018), pp. 222–234.
- [521] I S L Zeng and T Lumley. "Review of Statistical Learning Methods in Integrated Omics Studies (An Integrated Information Science)." In: *Bioinform Biol Insights* 12 (2018), pp. 1–16.

- [522] Haowen Zhang, Chirag Jain, and Srinivas Aluru. "A comprehensive evaluation of long read error correction methods." In: *BMC genomics* 21.6 (2020), pp. 1–15.
- [523] Jing Zhang, Xiangzhou Wang, Guangming Ni, Juanxiu Liu, Ruqian Hao, Lin Liu, Yong Liu, Xiaohui Du, and Fan Xu. "Fast and accurate automated recognition of the dominant cells from fecal images based on Faster R-CNN." In: *Scientific Reports* 11.1 (2021), p. 10361.
- [524] Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. "Lookahead Optimizer: k steps forward, 1 step back." In: *Proc. Advances in Neural Information Processing Systems 2019 (NeurIPS)*. Curran Associates Inc., 2019, pp. 9597–9608.
- [525] W. Zhang, Y. Yu, F. Hertwig, et al. "Comparison of RNA-seq and microarray-based models for clinical endpoint prediction." In: *Genome Biology* 16.1 (2015), p. 133.
- [526] W Zhang et al. "Comparison of RNA-seq and microarray-based models for clinical endpoint prediction." In: *Genome Biol* 16 (2015), p. 133.
- [527] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. "Single-Image Crowd Counting via Multi-Column Convolutional Neural Network." In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition 2016 (CVPR)*. 2016, pp. 589–597.
- [528] Yuanyuan Zhang, Hongyan Chen, Hongnan Mo, Xueda Hu, Ranran Gao, Yahui Zhao, Baolin Liu, Lijuan Niu, Xiaoying Sun, Xiao Yu, et al. "Single-cell analyses reveal key immune cell subsets associated with response to PD-L1 blockade in triple-negative breast cancer." In: *Cancer Cell* 39.12 (2021), pp. 1578–1593.
- [529] L Zhao and H Yan. "MCNF: A novel method for cancer subtyping by integrating multi-omics and clinical data." In: *IEEE/ACM Trans Comput Biol Bioinform Early Access* (2019), pp. 1–11.
- [530] Shanrong Zhao. "Alternative splicing, RNA-seq and drug discovery." In: *Drug discovery today* 24.6 (2019), pp. 1258–1267.
- [531] Xiangming Zhao, Laquan Li, Wei Lu, and Shan Tan. "Tumor co-segmentation in PET/CT using multi-modality fully convolutional neural network." In: *Physics in Medicine and Biology* 64.1 (Jan. 2019), pp. 015011–.
- [532] Xiuran Zheng, Dan Zhang, Mengying Xu, Wanqin Zeng, Ran Zhou, Yiming Zhang, Chao Tang, Li Chen, Lu Chen, and Jing-wen Lin. "Short-read and long-read RNA sequencing of mouse hematopoietic stem cells at bulk and single-cell levels." In: *Scientific Data* 8.1 (2021), pp. 1–10.

- [533] Y S Zheng et al. "MiR-100 regulates cell differentiation and survival by targeting RBSP3, a phosphatase-like tumor suppressor in acute myeloid leukemia." In: *Oncogene* 31.1 (2012), pp. 80–92.
- [534] Amy Zhou, Maya Sabatello, Gil Eyal, Sandra Soo-Jin Lee, John W. Rowe, Deborah F. Stiles, Ashley Swanson, and Paul S. Appelbaum. "Is precision medicine relevant in the age of COVID-19?" In: *Genetics in Medicine* 23.6 (2021), pp. 999–1000. ISSN: 1098-3600.
- [535] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Learning Deep Features for Discriminative Localization." In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition 2016 (CVPR)*. IEEE, 2016, pp. 2921–2929.
- [536] G Zhou, S Li, and J Xia. "Network-Based Approaches for Multi-omics Integration." In: *Computational Methods and Data Analysis for Metabolomics*. Vol. 2104. Methods in Molecular Biology. Springer, 2020. Chap. 23, pp. 469–487.
- [537] M. Zhou et al. "Radiomics in Brain Tumor: Image Assessment, Quantitative Feature Descriptors, and Machine-Learning Approaches." In: *American Journal of Neuroradiology* 39.2 (2018), pp. 208–216.
- [538] Xiang-Wei Zhu and Shao-Jing Li. "In Silico Prediction of Drug-Induced Liver Injury Based on Adverse Drug Reaction Reports." In: *Toxicological Sciences* 158.2 (2017), pp. 391–400.
- [539] A. Barla et al. "Machine learning methods for predictive proteomics." In: *Briefings in Bioinformatics* 9.2 (2008), pp. 119–128.
- [540] A. Bizzego et al. "Evaluating reproducibility of AI algorithms in digital pathology with DAPPER." In: *PLOS Computational Biology* 15.3 (2019), pp. 1–24.
- [541] B. O. Turner et al. "Small sample sizes reduce the replicability of task-based fMRI studies." In: *Communications Biology* 1.1 (2018), pp. 1–10.
- [542] D. Komura et al. "Machine Learning Methods for Histopathological Image Analysis." In: *Computational and Structural Biotechnology Journal* 16 (2018), pp. 34–42.
- [543] D. P. Kingma et al. "Adam: A Method for Stochastic Optimization." arXiv:1412.6980; Published as a conference paper at ICLR 2015. 2014.
- [544] E. Deniz et al. "Transfer learning based histopathologic image classification for breast cancer detection." In: *Health Information Science and Systems* 6.1 (2018), p. 18.
- [545] F. A. Spanhol et al. "A Dataset for Breast Cancer Histopathological Image Classification." In: *IEEE Transaction in Biomedical Engineering* 63.7 (2016), pp. 1455–1462.

- [546] F. Shahidi et al. "Breast Cancer Classification Using Deep Learning Approaches and Histopathology Image: A Comparison Study." In: *IEEE Access* 8 (2020), pp. 187531–187552.
- [547] G. Huang et al. "Densely Connected Convolutional Networks." In: *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018, pp. 2261–2269.
- [548] J. L. Myung et al. "Deep Convolution Neural Networks for Medical Image Analysis." In: *International Journal of Engineering & Technology* 7.3 (2018), pp. 115–119.
- [549] J. T. Leek et al. "Tackling the widespread and critical impact of batch effects in high-throughput data." In: *Nature Reviews Genetics* 11.10 (2010), p. 733.
- [550] J. Xie et al. "Deep Learning Based Analysis of Histopathological Images of Breast Cancer." In: *Frontiers in Genetics* 10 (2019), p. 80.
- [551] L. Peixoto et al. "How data analysis affects power, reproducibility and biological insight of RNA-seq studies in complex datasets." In: *Nucleic Acids Research* 43.16 (2015), pp. 7664–7674.
- [552] M. Jannesari et al. "Breast Cancer Histopathological Image Classification: A Deep Learning Approach." In: *Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2018, pp. 2405–2412.
- [553] M. Nawaz et al. "Multi-class breast cancer classification using deep learning convolutional neural network." In: *International Journal of Advanced Computer Science and Applications* 9.6 (2018), pp. 316–332.
- [554] M. Z. Alom et al. "Breast Cancer Classification from Histopathological Images with Inception Recurrent Residual Convolutional Neural Network." In: *Journal of Digital Imaging* 32.4 (2019), pp. 605–617.
- [555] N. Saravanan et al. "Data wrangling and data leakage in machine learning for healthcare." In: *International Journal of Emerging Technologies and Innovative Research* 5.8 (2018), pp. 553–557.
- [556] P. T. Nguyen et al. "Multiclass Breast Cancer Classification Using Convolutional Neural Network." In: *Proceedings of the 2019 International Symposium on Electrical and Electronics Engineering (ISEE)*. IEEE, 2019, pp. 130–134.
- [557] R. Mormont et al. "Comparison of Deep Transfer Learning Strategies for Digital Pathology." In: *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2018, pp. 2343–234309.
- [558] S. A. Iqbal et al. "Reproducible research practices and transparency across the biomedical literature." In: *PLoS Biology* 14.1 (2016), e1002333.
- [559] S. Moossavi et al. "Repeatability and reproducibility assessment in a large-scale population-based microbiota study: case study on human milk microbiota." [bioRxiv:2020.04.20.052035](https://doi.org/10.1101/2020.04.20.052035). 2020.

- [560] T. Ching et al. "Opportunities and obstacles for deep learning in biology and medicine." In: *Journal of The Royal Society Interface* 15.141 (2018), p. 20170387.
- [561] Y. Jiang et al. "Breast cancer histopathological image classification using convolutional neural networks with small SE-ResNet module." In: *PLOS ONE* 14.3 (2019), e0214587.
- [562] Z. Han et al. "Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model." In: *Scientific Reports* 7.1 (2017), p. 4172.
- [563] I M de Mas. "Multiomic Data Integration and Analysis via Model-Driven Approaches." In: *Data Analysis for Omic Sciences: Methods and Applications*. Vol. 82. Comprehensive Analytical Chemistry. Elsevier, 2018. Chap. 16, pp. 447–476.
- [564] Í F do Valle, G Menichetti, G. Simonetti, S Bruno, I Zironi, D Fernandes Durso, J C M Mombach, G Martinelli, G Castellani, and D Remondini. "Network integration of multi-tumour omics data suggests novel targeting strategies." In: *Nat Commun* 9 (2018), p. 4514.
- [565] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts. "Computational Radiomics Ssystem to Decode the Radiographic Phenotype." In: *Cancer Research* 77.21 (2017), e104–e107.
- [566] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J. C. Fillon-Robin, S. Pieper, and H. J. W. L. Aerts. "Computational Radiomics System to Decode the Radiographic Phenotype." In: *Cancer Research* 77.21 (2017), e104—e107.
- [567] Mart van Rijthoven, Zaneta Swiderska-Chadaj, Katja Seeliger, Jeroen van der Laak, and Francesco Ciompi. "You only look on lymphocytes once." In: *Proc. Conference on Medical Imaging with Deep Learning 2018 (MIDL)*. 2018, pp. 1–3.