



UNIVERSITY OF LEEDS

**Artificial Intelligence for Subgrouping
of Non-Hodgkin Lymphoma**

Volodymyr Chapman

**Submitted in accordance with the requirements for the
degree of Doctor of Philosophy**

The University of Leeds

**Faculty of Biological Sciences & Leeds Institute for
Data Analytics**

School of Cellular and Molecular Biology

January 2025

Intellectual Property

I confirm that the work submitted is my own, except where work which has formed part of jointly authored publications has been included. My contribution and the other authors to this work has been explicitly indicated. I confirm that appropriate credit has been given within the thesis where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

The right of Volodymyr Chapman to be identified as Author of this work has been asserted by Volodymyr Chapman in accordance with the Copyright, Designs and Patents Act 1988.

© 2025 The University of Leeds, Volodymyr Chapman

Signed:

Volodymyr Chapman



Acknowledgements

All work presented, unless otherwise stated, has been conducted by me - **Volodymyr Chapman** (University of Leeds & Leeds Institute for Data Analytics). The PhD project was funded by an **Engineering & Physical Sciences Research Council** Doctoral Training Partnership. I am grateful to the British taxpayer for ultimately funding this studentship. I would also like to thank the Faculty of Biological Sciences, University of Leeds for financial assistance through their hardship fund.

I am grateful to all those who have maintained their humanity; who have felt the unthinkable and found the kindness to laugh; who, burdened by the unimaginable, reach out a helping hand. I owe these people everything. Chief among them are the people that raised me - **Oksana, Lidiya, Oleksiy, Nick & Zhanna** - my wonderful partner, **Hannah**, siblings **Ollie & Fifi** and siblings in all but blood - **Tom, Polly, Uzma, Pandy, Will, Oleksandr L. & Vanya**.

Thank you to my **family and friends defending Ukraine** against the russian invasion. From the front to small acts of resistance, preserving our language, culture, freedom and peaceful existence - ***Glory to the Heroes***.

I have been beyond fortunate with work colleagues, who without exception have been kind, cooperative and immensely talented. Work-related thanks begin with **Prof. David Westhead** (University of Leeds & Leeds Institute for Data Analytics) for his continuous feedback and support. His was a fantastic example of humility, understanding and perseverance in the face of challenges much larger than ourselves.

Prof. Reuben Tooze (UoL / Leeds Teaching Hospitals NHS Trust) provided feedback and support in pathology aspects of the project. Through Reuben's example, I learned that when choosing what to be good at in life, 'all the above' is an acceptable choice. Regular, disease-specific feedback was provided by clinicians of the Haematological Malignancy Diagnostic Service, Leeds: **Dr. Alireza Behzadnia, Mr. Phillip Thompson,**

Dr. Sharon Barrans and **Dr. Catherine Burton**.

Prof. Andrew Janowczyk (Emory University, USA and Geneva University Hospitals, Switzerland) provided invaluable feedback and mentorship, with signature enthusiasm and vivacity, on use of AI in digital pathology.

Continuous feedback was provided by members of the Leeds Institute for Data Analytics (LIDA), namely **Drs. Mark Iles, Marlous Hall, Elton Rosas de Vasconcelos, Keiran Suchak, Amber Emmett, Euan McDonnell, Nazia Ahmed, Michal Zulcinski, Anna-Grace Linton, Zoe Hancox, Lucy Godson & Jonathan Batty**. Administrative support was provided by **Kimberley Wright, Renata Ferreira Lopes, Hannah Jackson & Karen Fletcher** of LIDA and **Emma Trowsdale** of the Faculty of Biological Sciences. Thank you also to **Joan Boyes** for pastoral support.

Mr. Dan Painter and **Prof. Alex Smith** of the University of York provided support with the Hematological Malignancy Research Network dataset (Research Ethics Committee reference 04/Q1205/69). Access to the GALLIUM clinical trial dataset (National Clinical Trials identifier: 01332968) was provided by **F. Hoffmann–La Roche AG** as part of a research partnership. Thank you to **Lin Pang** and **Dr. Alessia Bottos of Roche** for facilitating data sharing.

Every one of these people generously enriched my educational journey, understanding of the World and future journey - thank you. Finally, thank you to the **patients whose tissue and data were used in this work**. I hope that I have done justice to your donation.

Contents

List of Figures	xi
List of Tables	xii
Abstract	xiii
Abbreviations and Glossary	xv
Introduction	1
Problem definition	1
Nuances of non-Hodgkin lymphoma	4
Opportunities	8
Computational Pathology	10
Aims and output	17
2 Cytological Grading	21
1 Abstract	21
2 Introduction	23
2.1 Challenges in Follicular Lymphoma	23
2.2 Cytological Grading	23
2.3 Limitations of grading	24
2.4 Previous Work	24
2.5 Work summary	28
3 Methods	30
3.1 Experiment Workflow	30
3.2 Patient Data	31
3.3 Ethics statement	31

3.4	Whole slide images	31
3.5	Data partitions	32
3.6	Nuclear Segmentation	34
3.7	Naive B-cell Region Classification	34
3.8	Centroblast Classification	36
3.9	Automated Cytological Grading	40
3.10	Inference on validation data	41
3.11	Survival analysis	42
4	Results and Discussion	44
4.1	Dataset characteristics	44
4.2	Nuclear segmentation methods	46
4.3	Naive B-cell region classifier	50
4.4	Centroblast Classification	56
4.5	Centroblast counts evaluation	61
4.6	Cytological Grading	65
4.7	Survival Analysis	69
4.8	Overall discussion	73
5	Conclusions and Future Work	76
5.1	Cytological grade	76
5.2	Limitations	78
5.3	Key takeaways	78
5.4	Future directions	79
3	Patient nuclear subgroups	81
1	Abstract	81
2	Introduction	83
2.1	Successes in non-Hodgkin Lymphoma	83
2.2	Unsupervised classification	83
2.3	Aims	84

3	Methods	85
3.1	Experiment Workflow	85
3.2	Patient data	86
3.3	Whole slide images	86
3.4	Data partitions	86
3.5	Pan Non-Hodgkin lymphoma nuclear dataset	87
3.6	Unsupervised Non-Hodgkin cell classifier	87
3.7	Nuclear Phenotype Clusters	88
3.8	Nuclear Distribution Clusters	90
3.9	External validation	92
3.10	Survival Analysis	94
3.11	Software	95
3.12	Hardware	95
4	Results and discussion	96
4.1	Unsupervised Non-Hodgkin cell classifier	96
4.2	Nuclear Phenotypes Clusters	98
4.3	Nuclear Distribution Clusters	100
4.4	Comparison of clustering strategies	102
4.5	Survival Analysis	110
4.6	Overall discussion	116
4.7	Limitations	122
5	Conclusions and future work	125
5.1	Patient groups	125
5.2	Clustering methods	126
5.3	Summary of Findings	127
5.4	Limitations and future work	127
4	Immunohistochemistry clusters	129
1	Abstract	129

2	Introduction	130
2.1	Value of IHC in Lymphoma	130
2.2	Limitations of IHC	131
2.3	Recent advances in IHC	132
2.4	Aims	132
2.5	Novel Contributions	133
3	Methods	133
3.1	Experimental workflow	133
3.2	Data	134
3.3	Slide selection	135
3.4	Data Partitions	136
3.5	IHC quantification	137
3.6	IHC clustering	139
3.7	Survival Analysis	140
3.8	Validation	140
3.9	Computational specifications	145
4	Results and Discussion	146
4.1	IHC quantification	146
4.2	IHC clustering	150
4.3	Survival Analysis	165
4.4	Validation	174
5	Conclusions	189
5.1	Limitations	192
5.2	Summary of Findings	193
5.3	Future Work	194
	Conclusions and Future Work	195
	Summary of Main Findings	195
	Findings in context	199

Real-world merit of this thesis	200
Bibliography	204
Appendix	245
1 Chapter 1: Introduction	245
2 Chapter 2: Cytological Grading	246
3 Chapter 3: Nuclear clusters	255
4 Chapter 4: Immunohistochemistry clusters	260
4.1 Evaluation metrics	261

List of Figures

Introduction	1
1.1 Histopathology of non-Hodgkin Lymphoma	7
1.2 Example CNN and GCN task in non-Hodgkin Lymphoma	16
 Chapter 2: Cytological Grading	 21
2.1 Comparison of naive B-cell region staining in histopathology . .	26
2.2 HoVernet inference on reactive germinal centre	28
2.3 Automated Cytological Grading workflow	30
2.4 Visual quality control and partitioning of the datasets	33
2.5 Naive B-cell region classifier	36
2.6 Nuclear images in Centroblast classification	40
2.7 Nuclear segmentation algorithm evaluation	48
2.8 Nuclear Segmentation Algorithm Visual Inspection	49
2.9 Follicle Segmentation Algorithm performance	53
2.10 Follicle Segmentation Algorithm visual inspection	54
2.11 Centroblast classifier performance	59
2.12 Bland-Altman plots of pathologist and ML model centroblast counts	63
2.13 Confusion matrices of machine performance against pathologist grade	68
2.14 Hazard ratios for cytological grade	71
2.15 Kaplan-Meier plots of overall survival, with respect to high/low cytological	72
2.16 Summary of key results	77

Chapter 3: Patient nuclear subgroups	81
3.1 Experiment workflow for patient clusters	85
3.2 Cluster validation on external data	93
3.3 Nuclear clustering visual inspection	97
3.4 Nuclear phenotype cluster	99
3.5 Nuclear distribution clusters composition	101
3.6 Slide clustering contingency matrices	102
3.7 Contingency tables of molecular features	105
3.8 Supervised metrics on nuclear clusters	106
3.9 Relationships between cell clusters	109
3.10 Hazard ratios of patient clusters	113
3.11 Kaplan-Meier plots of overall survival for patient clusters	114
3.13 PFS Kaplan-Meier plots for patient clusters	116
3.14 Proposed relationships between nuclear and molecular features	126
 Chapter 4: Immunohistochemistry clusters	 129
4.1 Experimental workflow for IHC profiles	134
4.2 Evaluation of IHC quantification against pathologist	147
4.3 Comparison of BCL6 quantification between DeepLIIF and pathol- ogist	148
4.4 Optimal clustering of IHC	153
4.5 Comparison of higher DLBCL cluster number against expression	154
4.6 Interpretation of IHC clusters	158
4.7 Clustermap with 3 Macenko IHC clusters	161
4.8 Overall survival in IHC clusters	166
4.9 Overall survival in more DLBCL clusters	169
4.10 FL IHC cluster hazard against established indices	171
4.11 DLBCL IHC cluster hazard against established indices	173

4.12 Evaluation of UNI and nuclear features in prediction of IHC clusters	175
4.13 Evaluation of IHC cluster prediction by quantification method	176
4.14 Characteristics of IHC clusters predicted on validation	179
4.15 UNI feature importance	183
4.16 Interpretation of H&E prediction of IHC clusters	188
4.17 Summary of Key clusters	190
Appendix	245
A.1 Follicle Segmentation against IHC ground truth	248
A.2 Visual inspection of H&E Follicle Segmentation against IHC ground truth	251
A.3 Nuclear feature selection	252
A.4 Correlations between patient and nuclear features	254
A.5 PFS on HMRN FL patients	254
A.6 Evaluation of optimal nuclear cluster number	256
A.7 Nuclear phenotype patient clusters	257
A.8 Cellular composition of external dataset phenotypes	258
A.9 Molecular composition of groups	259
A.10 Divergence of imputed IHC data	260

List of Tables

2.1	Nuclear segmentation algorithm evaluation	47
2.2	Naive B-cell region segmentation algorithm evaluation	52
2.3	Evaluation of nuclear classifier performance	57
2.4	Published centroblast classifier performance	60
2.5	Evaluation of cell counts	62
2.6	Cytological grade regression metric results	66
2.7	Classification of high cytological grade	67
A.1	Patient characteristics	248

Abstract

Introduction

Non-Hodgkin Lymphoma (nHL) is a common type of cancer, accounting for over 10,000 annual diagnoses in England. Diagnostic procedures for nHL require analysis of tissue biopsies by highly skilled histopathologists. Procedural demand has increasingly outpaced supply, with potential for negative impact on patient care. In this thesis, methods of automating high burden, subjective procedures and of extracting additional value from routine samples are proposed, with a focus on the most common forms of nHL: Follicular Lymphoma (FL) and DLBCL.

Methods

A pipeline of processes are proposed for automation of FL cytological grading, involving quantification of cancerous B-cells (centroblasts). Two novel methods of clustering FL and DLBCL patients on cell populations are presented. Finally, FL and DLBCL are grouped on presence of informative proteins (immunohistochemistry).

Results

High cytological grade was found to associate with increased survival risk but, the automated procedure classified a group of FL patients with good outcome. Cell-based clusters associated with mutation in FL and survival in DLBCL. Automated immunohistochemistry estimates clustered patients on gene expression profiles, informative of survival. Prediction of immunohistochemistry clusters using routine (H&E) biopsy slides was successful with state-of-the-art pathology image representations.

Conclusions and future work

Classification of a low risk patient group through automated cytological grading evidenced potential utility of centroblast subtypes as indicators of good outcome in FL. Performance of automated immunohistochemistry quantification in clustering gene expression groups evidenced utility of automated procedures for reduction of histopathologist burden. Successful classification of immunohistochemistry with H&E evidenced feasibility, by extension, of gene expression prediction.

Abbreviations and Glossary

Disease terms

- **FL: Follicular Lymphoma** (a common, often indolent Non-Hodgkin Lymphoma characterised by arrangement of tumour cells in follicles and BCL2 gene translocation)
- **DLBCL: Diffuse Large B-cell Lymphoma** (a common, aggressive group of Non-Hodgkin Lymphomas characterised by sheets of large neoplastic cells)
- **HL: Hodgkin Lymphoma** (an aggressive Lymphoma, characterised by sparse, multinucleated, neoplastic Reed-Sternberg cells)
- **BL: Burkitt's Lymphoma** (an aggressive Non-Hodgkin Lymphoma characterised by MYC gene translocation and classic 'starry sky' appearance, due to sparse tingible body macrophages within uniform sheets of medium-sized neoplastic cells)
- **Ig: Immunoglobulin** (Scientific term for an antibody protein, produced by B-cells. Though often associated with secreted forms, the antigen-binding portion is derived from the membrane-bound B-cell receptor)
- **BCR: B-cell receptor** (Surface protein of B-cells responsible for antigen binding, consisting of a heavy chain, IgH, and light chain, IgL)
- **GC: Germinal Centre** (The sites within lymph tissue where naive B-cell somatic mutation events occur, with controlled selection of cells exhibiting affinity for desired antigens)
- **SHM: Somatic Hypermutation** (Somatic mutation performed by naive B-cells within GCs. Error-prone DNA repair mechanisms introduce mutations within DNA regions encoding antigen binding sites within the B-cell receptor. This mechanism leads to diversification of antibody binding characteristics and thus, target antigens)

in clonal B-cell populations)

- **CSR: Class Switch Recombination** (A controlled somatic mutation event in activated B-cells wherein DNA regions encoding effector cell binding sites (not antigen binding sites) of the B-cell receptor are substituted for alternative isotypes. This leads to changes in downstream responses of antibody binding and kinetics of antibodies, optimising performance for a range of environments and pathogens)
- **fDC: Follicular Dendritic cell** (Antigen-presenting cells responsible for coordination of GC events. These include GC initiation upon immune challenge, selection of naive B-cells with preferable antigen-binding characteristics and recruitment of macrophages for clearance of apoptotic cells)
- **TBM: Tingible-body Macrophage** (Specialised macrophages found within GCs responsible for clearance of apoptotic naive B-cells)

Clinical terms

- **NHS: National Health Service** (The public health authority in England)
- **WHO: World Health Organisation** (A key international organisation in recommendations for classification and treatment of disease)
- **ICC: International Consensus Classification of myeloid neoplasms** (An international group of experts in cancers affecting blood and lymphoid tissues that provide specialist recommendations for those conditions)
- **HMRN: Haematological Malignancy Research Network** (A consortium of experts in blood-related cancers in the North of England, involved in data collection and reporting on those conditions)
- **ICD-10: 10th version of the International Statistical Classification of Diseases** (Standardised numeric classifications of disease, intended to improve consistency in reporting between sites and states)
- **LDH: Lactate dehydrogenase** (An enzyme responsible for catalysing reactions

of glycolysis and anaerobic respiration within cells. Serum concentration is used to assess extent of cellular damage and patient risk)

- **B2M: β_2 microglobulin** (A subunit of the antigen-presenting HLA complex responsible for cytotoxic immune responses, expressed on nucleated cells. Serum concentration in the context of lymphomas is used to assess patient risk, though the mechanism behind this is not known)
- **ECOG: Eastern Cooperative Oncology Group Performance Status** (An index for assessing impact of disease on patient livelihood/independence, ranging from no impact to complete loss of motility and death)
- **IPI: International Prognostic Index** (a prognostic index for DLBCL, accounting for patient age, stage, serum LDH concentration, performance status (ECOG) and involvement of non-lymphoid tissue, termed extranodal sites)
- **FLIPI: Follicular Lymphoma International Prognostic Index** (A prognostic index for FL, comparable to IPI in DLBCL, taking into account age, disease stage, serum LDH, serum haemoglobin and extranodal involvement)
- **PRIMA-PI: PRIMA Prognostic Index** (A prognostic index for FL taking into account serum B2M concentration and bone marrow involvement)
- **POD24: Progression of disease within 24 months** (A binary measure of patient outcome accounting for poor response to initial measures of disease management. Evidenced to be informative of longer term patient outcomes in at least some lymphomas.)
- **R-CHOP: Rituximab, cyclophosphamide, doxorubicin, vincristine and prednisolone** (A common immuno-chemotherapy regimen for the treatment of aggressive B-cell lymphomas consisting of Rituximab, an antibody specific to the B-cell marker CD20 (immuno- portion of therapy), cytotoxic molecules (CHO; - chemo portion of therapy) and a steroid to modulate inflammation (prednisolone; P))

- **W&W: Watch and Wait** (Also referred to as ‘watchful waiting’, it is a common management strategy in asymptomatic and indolent cancers like FL, where immuno- and/or chemotherapy are not considered immediately beneficial. Instead, patients undergo a period of regular observation and re-assessment of optimal management strategy)

Pathology terms

- **FFPE: Formalin Fixed Paraffin Embedded** (A procedure for tissue preservation for long-term storage. This involves dehydration and fixation through polymerisation with formalin - a mixture of methanol and formaldehyde - and ‘sealing’ of tissue within paraffin wax)
- **H&E: Haematoxylin and Eosin** (A routine pathology staining panel, leading to purple colouration of negatively-charged, nucleic acids (chromatin) and pink to red staining of positively-charged amino groups within proteins (cytoplasmic proteins and vesicular nuclei, such as in the case of immunoblasts))
- **IHC: Immunohistochemistry** (Staining of specific proteins within pathology samples for the purposes of differential diagnosis, subgrouping and prognosis. Commonly, this involves oxidation of 3,3’-diaminobenzidine, DAB, using peroxidase-conjugated antibodies specific to proteins of interest, leading to deposition of brown precipitate in regions of protein presence)

Image terms

- **WSI: Whole Slide Image** (A scanned biopsy slide. These are large images, often measuring over 10,000 pixels in height and width)
- **ROI: Region of Interest** (A portion of a WSI, often representing areas with prognostic value. These are much smaller than the WSI and thus, are often more feasible to parse and analyse in computational efforts)

Computational terms

- **ML: Machine Learning** (Iterative processes (algorithms) for determination of a process, given that starting data and at least some conditions of the outcome of the process are defined)
- **DL: Deep Learning** (A subset of ML characterised by training of processes involving many, often hundreds, of ‘learned’ steps and subsequently, use of complex, abstract relationships within data)
- **Epoch: An iteration in fitting ML models** (A single cycle of modifying internal parameter values with the aim of better achieving required output conditions using input data)
- **Weight: Linear co-efficient** (A co-efficient value which is optimised during ML model training, with the objective of meeting the stated output conditions)
- **ANN: Artificial Neural Network** (An ML model type involving sequential matrix multiplications with optimised weights and potentially non-linear transformations of data)
- **Perceptron: A single transformation unit of an ANN** (Comprising matrix multiplication with optimised weight values and transformation using an activation function)
- **Activation function: A decision function** (A transformation, such as Sigmoid or Rectified Linear (ReLu), applied in perceptrons to values resulting from matrix multiplication, often to bound values within a desired range)
- **Loss function: Transformation to determine distance (loss) between data and desired output conditions** (Used in training of ML/DL methods to inform which parameters require tuning)
- **Backpropagation: The process of optimising ANN parameters to better reach desired output conditions** (Starting from the perceptron layer closest to

the output, loss (between values and desired output) and extent of changes required to individual weights (called gradients) are calculated. An optimisation function is used to change weight values with the aim of better predicting the required output in the next iteration)

- **Optimiser: Function informing weight value changes in backpropagation**
(A function, such as stochastic gradient descent, that uses gradients to adjust weight values with the aim of better predicting required output data)

Introduction

Problem definition

Healthcare in crisis

Histopathology is the study of disease through microscope analysis of tissue samples[1]. Pathologists specialised in this work (histopathologists) are involved in a wide range of clinical investigations including disease diagnosis, subclassification and determination of cause of death. Supply of pathology specialists does not meet demand[2]. This deficit likely contributes to general increases in diagnostic wait times[2], with $\frac{1}{4}$ of patients waiting over 6 weeks for test results[3]. This has been evidenced to directly impact patient outcome in aggressive cancers[4, 5], with up to 8.7% increase in overall survival risk for each additional day before treatment in lung cancer.

The number of microscopy slide analyses required annually in England may have grown to over 37 million by 2025, following a 4.5% annual increase[6, 7]. With limited scope for drastically increasing workforce size to meet demand, the NHS has prescribed adoption of digital technologies to increase productivity of existing specialists[8].

Non-Hodgkin Lymphoma

Non-Hodgkin Lymphomas are common cancers of lymphoid tissues (such as lymph nodes, bone marrow and thyroid) with over 10,000 annual diagnoses in England[9]. The vast majority involve B-lymphocytes, responsible for antibody responses. Between 2,100-2,500 cases of Follicular Lymphoma (FL; ICD-10 code: C82) and 5,800-6,400 cases of Diffuse Large B-cell Lymphoma (DLBCL; ICD-10 code: C83) are diagnosed in England, annually [9, 10].

These diseases have specific needs of histopathology. With respect to aforementioned delays in NHS diagnosis time, median time for non-Hodgkin Lymphoma diagnosis following

referral to specialist care was 28 days in 2014. Local wait times in Kent (34 days), the West Midlands (32 days) and Lancashire (31 days) were substantially higher[11]. Given additional waits for referral to specialist from primary care (GP practices) and evidence of almost doubled waiting list size since 2014[3], it is reasonable to estimate median wait times of over 35 days for diagnosis from an initial GP visit. For low risk cases, with low prognostic index values (International Prognostic Index, IPI, score below 3), these wait times may not significantly impact outcome[12]. In fact, watchful waiting (observation, without immuno-/chemotherapy till evidence of benefit) remains a recommended course of action in asymptomatic FL, especially in elderly patients[9]. Yet, higher risk patients with disease spread throughout the body (Ann Arbor stage III-IV), potentially including non-lymphoid tissue (extranodal site involvement) and experiencing greater impact on their independence (ECOG performance status > 1) have substantially worse outcomes with increased wait times[12]. This highlights the first histopathology challenge in non-Hodgkin Lymphoma: faster diagnosis for high risk patients (with IPI score above 2). Secondly, survival of FL and DLBCL patients has seemingly benefitted less from technological advancements and improving standards of care than other non-Hodgkin Lymphomas. Mortality has decreased by 31% in FL and 29% in DLBCL (between 2004 and 2019), while Burkitt Lymphoma (53% decrease), Marginal Zone Lymphoma (47% decrease) and Mantle Cell Lymphoma (38% decrease) have seen greater improvements in long-term care[13].

In the commonly indolent lymphoma FL, outcomes are often good with over 90% 5-year survival in patients under the age of 60[14]. The indolent nature of the disease means that it is not curable, so, clinical challenges lie in determining an optimal strategy for long term treatment. Existing prognostic indices stratify risk well in patients undergoing immuno-/chemotherapy[15], though they do not make use of a plethora of information provided by histopathology.

Progression of disease within 24 months (POD24) is a key indicator of outcome, with 86% of patients that respond well to initial disease management surviving for over 10 years compared to 66% of those experiencing early progression[16]. The presence of this effect

across treatment regimens evidence a subgroup of FL patients for whom current gold standards have limited benefit[17]. Early classification of these patients, before initial treatment failure and disease progression, may have utility in optimal treatment choice and patient wellbeing.

In the aggressive lymphoma, DLBCL, 65% of patients have been cured, in some settings, with current standards of care[18]. Like in FL, challenges lie in determining which patients would respond poorly to standard of care. A range of mutational and gene expression profiles have been identified, with significant survival differences in some cases[19–22]. The tests for determining these features remain costly and so, are not routine[23, 24]. In the case of cell-of-origin expression in DLBCL[19], expression must commonly be predicted using microscopy slides, further increasing histopathologist workloads[22, 23, 25]. There are considerable limitations to these analyses compared to gene expression, with greater subjectivity, imperfect agreement with expression (around 85%) and limited utility, since only one expression profile is classified against all others[26].

Even with information on expression profile, clinical trials have evidenced limited benefits of specific therapies for expression groups[27]. Despite substantial increases in risk of chronic heart failure ($>3\times$, compared to the population[28]), R-CHOP immunochemotherapy has remained the standard of care for almost two decades. Access to effective means of determining mutational and expression status of patients in ways that minimise additional histopathologist burden, and limited utility of results in treatment choice limit improvements in DLBCL patient care.

Summary of requirements

In summary, major challenges in Non-Hodgkin lymphoma lie in:

- Prediction of FL patients likely to respond poorly to the standard of care, often experiencing disease progression within 2 years
- Prediction of a range of mutational and expression profiles (rather than one against all others), with minimal increases to pathologist burden

- Prediction of patient groups informative in treatment choice

Nuances of non-Hodgkin lymphoma

Utility of lymphoid tissue

B-cells are vital for humoral responses to immune challenge, those involving secreted molecules over direct cell-cell contact with pathogens. B-cell effectors, termed plasma cells, secrete high volumes of immunoglobulin (Ig; antibodies) which lead to recruitment of effector cells and sequestration of pathogens upon binding to targets[29–31]. The importance of humoral defence is highlighted in individuals deficient in key B-cell components, who often experience severe and recurring bacterial infections[32].

B-cell responses must be, paradoxically, specific yet versatile to maintain delicate homeostasis between host cells and the microscopic ecosystem on the body's border. They must overcome intracellular threats (cancer, viruses and microbacteria); eukaryotic parasites (helminth and Plasmodium) and extracellular pathogens (bacteria and fungi) while maintaining beneficial interactions with microbes and modulating responses to prevent autoimmunity[33–36].

Somatic mutations

This contradictory set of requirements are fulfilled in B-cells through three controlled, somatic mutation events targeting the B-cell receptor (BCR). This complex ultimately forms the antigen-binding portion of antibodies. Somatic mutations create a pool of B-cells with immense immunoglobulin diversity from identical starting genetics[30, 37–39]. These processes are:

V(D)J recombination In the bone marrow, progenitor B-cells undergo ‘assembly’ of the genes encoding the BCR. Unique to each cell, these are the result of random choice of Variable (V), Diversity (D) and Joining (J) regions. Double strand breaks and ligation

are catalysed in genes encoding antigen-binding sites (Ig heavy chain, *IgH*, and light chain, *IgL*, genes). Joining combinations of V,D and J regions (VDJ in *IgH* and VJ in *IgL*) forms the completed BCR[37–39].

Somatic hypermutation (SHM) Upon immune challenge, B-cells are recruited to nearby lymph nodes. B-cells without prior activation through simultaneous binding to an appropriate antigen and activation (inflammatory) signals, termed naive B-cells, migrate to germinal centres (GCs). In these temporary structures, antigens are presented to naive B-cells, their intrinsic apoptosis escape mechanisms (mediated by the BCL-2 protein) are suppressed and point mutation mechanisms are activated (SHM). Naive B-cells with adequate binding to the required antigen receive survival signals and undergo successive rounds of mutation and error-prone repair (SHM) in V regions of immunoglobulin genes (*IgH* and *IgL*)[30, 37, 38].

Class Switch Recombination (CSR) While V(D)J recombination and SHM introduce variation in antigen-binding kinetics, diversity is required in the mechanics of Ig and downstream immune responses caused by antigen binding. Adaptations for humoral defence in mucosa differ from those needed in blood; neutralisation of cancerous cells require distinct downstream processes compared to parasitic fungi. Partially random exchange of the *IgH* constant region, responsible for downstream signalling and Ig transport creates diversity in immunoglobulin responses. Alternative constant region isotypes are adapted for specific transport behaviours (such as crossing the placenta) and downstream immune effects (such as complement fixation), facilitating appropriate responses for a wide range of pathogens. Ultimately, this process produces a pool of differentiated B-cells (plasmablasts), producing antibodies with similar antigen specificity but a range of environmental and situational adaptations[30, 38, 40, 41].

Non-Hodgkin Lymphoma pathogenesis

Despite being crucial in defence, errors in these somatic mutation processes lead to development of non-Hodgkin lymphomas:

- CSR and SHM can lead to translocation of the highly-expressed *IgH* gene to the *c-Myc* oncogene, leading to aggressive Burkitt and diffuse large B-cell lymphomas[22, 42–44]
- V(D)J recombination leading to translocation of *IgH* to the anti-apoptotic *BCL-2* gene, facilitating continued *BCL-2* expression in the GC. The mutation, allowing apoptosis escape in GCs, is a hallmark of classical Follicular Lymphoma[45, 46]
- SHM leading to successive cycles of mutation in DLBCL and higher risk FL[47, 48]

An additional risk in B-cell lymphomas uncommon in many other cancers is the high initial motility of affected cells. Disease is often delocalised at diagnosis, limiting treatment options. This effect is evidenced by the 82% of DLBCL patients and 84% of FL patients presenting with more than one affected site at diagnosis (Ann Arbor Stage > 1)[46, 47], and involvement of vital organs such as lung, central nervous system, liver or bone marrow in almost 40% of DLBCL patients[49].

Ultimately, a ‘perfect storm’ of features required for effective defence against a myriad of potential pathogens lead to development of B-cell cancers, primed with innate mutation activities and high motility.

A comparison of non-cancerous (reactive) lymph node features, FL and DLBCL are presented in Figure 1.1, below. In reactive tissue (left column), the GC architecture is clear: visible ‘light’ zones composed of tingible-body macrophages (TBMs) resolving apoptotic debris and dark mantle zones surrounding GCs. GCs are largely concentrated on the lymph node periphery, closest to where molecules from surrounding tissues drain into the lymph node and contact antigen presenting cells. In FL (middle column), GC-like structures appear to form but their architecture is unlike GCs (therefore named follicles, not GCs): no observable ‘light’ zones and unclear mantle zones. These follicles are distributed throughout the lymph nodes. Closer inspection reveals absence of TBMs, with large numbers of centrocytes and some centroblasts. In DLBCL (right column), no GC-like structures are observed. Neoplastic B-cells (centroblasts) form sheets, pervading the tissue.

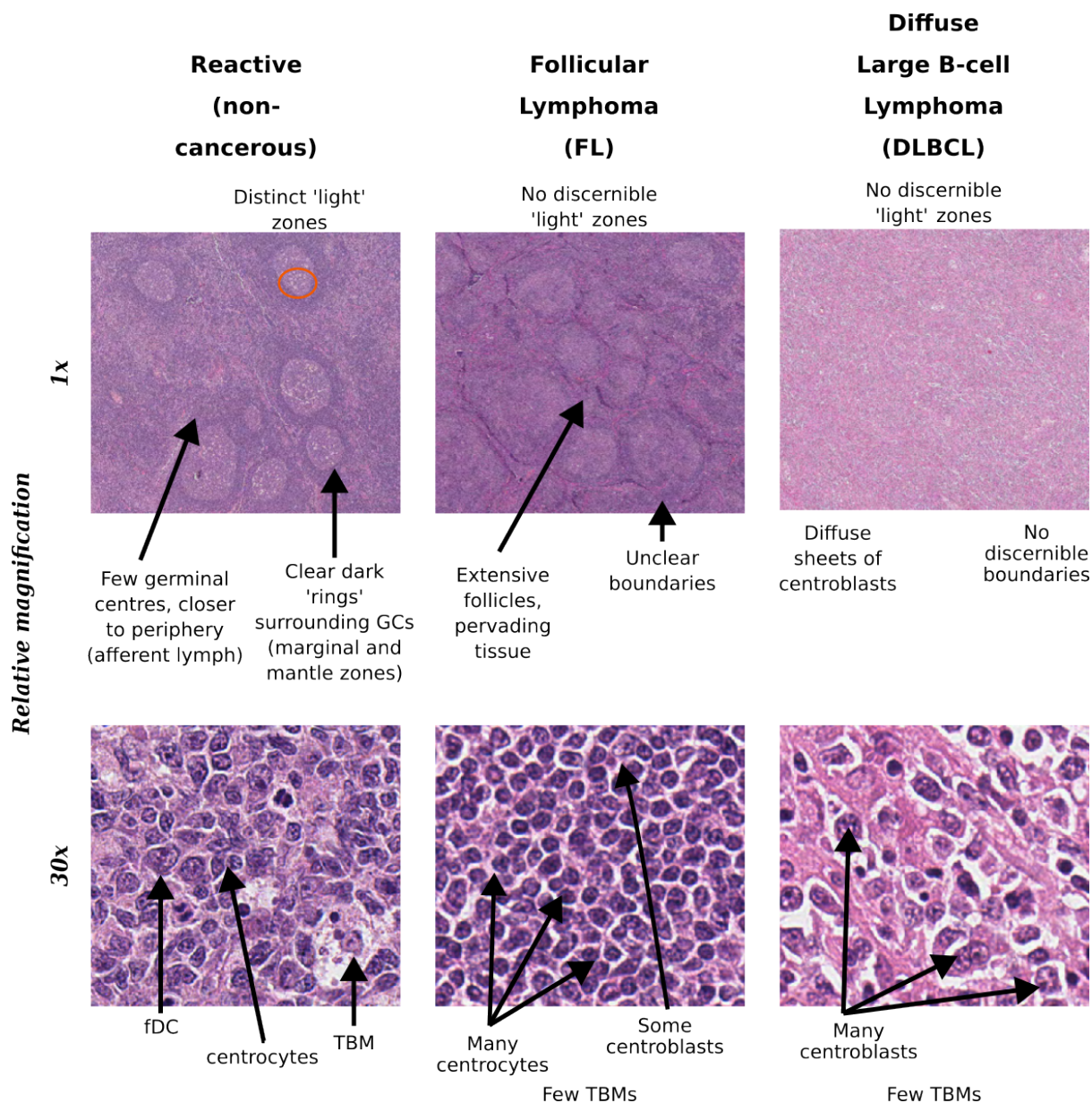


Figure 1.1: Overview of classic histopathology features of FL and DLBCL in H&E slides, with respect to reactive (non-cancerous inflammation) lymph nodes. Key features are annotated with arrows. Abbreviations: fDC = Follicular Dendritic Cell; TBM = Tingible-Body Macrophage; GC = Germinal Centre. Relative magnification refers to viewer magnification (QuPath[50]) used to extract images. Further scaling was performed to fit images to an A4 page size. Cell annotations were made by VC.

Opportunities

Previous efforts

Subgrouping of Non-Hodgkin Lymphoma

Recent work has highlighted value of mutational signatures in FL and DLBCL. In FL, patients with mutations of *TP53* and related genes appear at significantly increased overall survival (OS) and progression-free survival (PFS) risk[51, 52]. Classification of DLBCL and FL mutation identified a subgroup of FL harbouring similarities to DLBCL and significantly higher mutation in genes associated with errors in SHM (such as *BCL7A*)[48]. This group was found to be at higher risk of transformation to aggressive disease (DLBCL), often carrying worse prognosis.

Likewise, *TP53* mutation appears to lead to significantly worse PFS in DLBCL[53]. Besides *TP53*, simultaneous mutations or over-expression of two or more lymphoma-associated genes (anti-apoptotic *BCL2*, the oncogene *c-MYC* and/or the germinal transcription regulator *BCL6*) are associated with poor 5-year OS of 22-43%[54].

A recent, novel mutational signature incorporating markers of B-cell differentiation and BCR signalling appears to predict lower PFS risk in a range of datasets[55].

Gene expression profiles have long been used to stratify risk in DLBCL, with evidence of higher risk in late stage (plasmablast-like) germinal centre and Burkitt Lymphoma-like expression[19, 56]. Work aiming to resolve expression to the level of specific cells demonstrated significantly increased OS and PFS risk with abnormally low (‘depleted’) and high (‘inflammatory’) levels of immune cell expression[57].

These studies present clear utility of mutation and expression assays in stratifying patient risk, though, the procedures remain prohibitively expensive for routine use.

Cytological grading of FL

A series of efforts by the laboratory of Metin Gurcan aimed to improve inter-observer agreement in H&E cytological grading of FL[58–63], a burdensome and highly subjective

task. Despite their efforts, utility of the FL automated grading system (FLAGS[63]) was limited due to reliance on additional staining (CD20 immunohistochemistry), incurring additional costs, and the lack of survival analyses to investigate potential clinical benefit. Automation of cytological grading was not accepted as standard. The manual process is no longer recommended by the WHO[23] due to poor reproducibility and limited evidence of clinical utility. Yet, independent groups of experts continue to recommend cytological grading for potential benefits in stratifying at-risk patients[24].

Recent work has evidenced relationships between uncommon profiles of mutation (involving the *STAT6* gene) and diffuse arrangements of lymphoma follicles[64], indicating potential of mutation profile prediction in histopathology. Additionally, other work in FL has evidenced clinical utility in cell heterogeneity and presence of specific cell types[65]. Together, these results evidence potential utility in subgrouping and quantifying cells, as well as their arrangements in FL.

Cell types in DLBCL

Recent work evidenced clinical utility of nuclear morphology in DLBCL[66]. This echoed previous findings in which heterogeneity of cells and nuclear morphology were either informative of genetics or of survival[67, 68]. As yet, diagnostic procedures in classification of morphology and protein markers, as recommended by international organisations[23, 24], are conducted manually. As with FL, classification of these features likely suffers from low reproducibility due to high reliance on subjective features and variation in equipment[61, 69–71]. The work of Vrabac et al.[66] evidence utility of automated approaches, with potential for higher consistency in analyses. Use of such systems for ‘second opinions’ may improve histopathologist consistency, as demonstrated with the FLAGS system[63], while providing information on features not feasibly quantified by human observers (such as cell heterogeneity).

Key questions

Given findings above, the following questions remain in FL and DLBCL:

- Is there clinical utility to cytological grading in FL?
- Can cellular subtypes be resolved in routine histopathology slides?
- Can molecular subtypes, originating from mutation or gene expression profiles, be resolved in histopathology slides?
- What is the clinical utility, with respect to OS of histopathology clusters in FL and DLBCL?

Computational Pathology

Considerations in Histopathology

Tissue biopsies must be preserved, sliced and stained for routine histopathology assessments in non-Hodgkin Lymphoma. A common procedure for this formalin fixation and paraffin embedding (FFPE).

First, tissue is dehydrated and ‘fixed’, a term to describe formation of covalent bonds to preserve microscopic structures, using an ethanol/formaldehyde mix (formalin). Next, biopsies are sealed in paraffin wax for long term storage, forming structures commonly referred to as ‘tissue blocks’. Tissue blocks are sliced finely (up to $5\mu\text{m}$ in thickness) using a microtome to allow light transmission through the tissue. Next, tissue is stained. Commonly, this consists of haematoxylin and eosin (H&E), which stain nucleic acids purple (DNA and RNA) and proteins pink (cytoplasm and ribosomes), respectively. In addition, specific proteins can be targeted using enzyme-conjugated antibodies in a process termed immunohistochemistry (IHC). In DLBCL, this process is standard for detecting cells producing markers of naive and plasma B-cells, as part of estimations of their expression profile (Hans algorithm[25]). Deposition of coloured precipitate by the antibodies, commonly brown precipitate from 3,3'-diaminobenzidine (DAB), stains regions enriched for the target protein.

The above steps can each lead to noticeable image features not representative of the underlying pathology (artefacts), highlighting need for adequate quality control measures

in histopathology[72]. For example, delays in fixation and use of blunt blades in tissue slicing can lead to fragmentation of tissue layers. Loose blades can lead to tissue sections of variable thicknesses and hence, hues. Exposure time with each of the stains mentioned can lead to variation in staining between slides, with variable thresholds for regions containing the target biological entities.

These phenomena must be distinguished from ‘real’ features of disease, those resulting from the underlying pathology.

Digital slides

Histopathology slides can be digitised through scanning at high resolution. The advantages of using digital images over glass slides for histopathology analyses include greater efficiency in indexing; scope for consultations from geographically-distant specialists (tele-pathology) and scope for computer-aided analysis[73]. Individual case studies have evidenced long-term financial benefits to digitisation, albeit high initial costs[74].

This process presents unique challenges, though. To effectively resolve informative features of tissue, slides must be scanned at high magnifications of $20\times$ and $40\times$ (further information on magnification in digital images provided in Appendix Section 1). This results in images on the gigapixel scale ($> 10^9$ pixels), displaying only several centimetres squared of tissue, while using dozens of gigabytes of computer memory to store (without image compression). Storage and use of these images for routine histopathology present technical challenges. These become more acute where computational analyses are desired, as the image files are too large to parse whole for consumer grade computers.

For this reason, digitised slides must often be split into many smaller image ‘patches’ or parsed at a lower magnification (termed ‘downsampling’ in this thesis) for use with computational analyses.

Machine Learning

Machine learning (ML) refers to a group of mathematical procedures aimed at determining the steps behind a desired transformation: given an input X and conditions of a

desired output $f(X)$, ML methods aim to estimate the nature of $f()$ [75]. For instance, lifestyle characteristics, X , may be known for a set of individuals, such as weekly alcohol consumption, social media use, compliance with traffic safety protocols etc. Whether they have ever been at fault in a motor vehicle incident, $f(X)$, may also be known. A machine learning task may be developed to find a series of steps, the function $f()$, which determines likelihood of ever being at fault in a vehicle incident, given lifestyle habits.

This differs from traditional statistics such as correlation, which can identify relationships between data but makes no assumptions of processes involved in interconversion of data. ML is especially useful in automating or making inferences on relationships between data where the relationships are complex, involving interactions between multiple variables. Common ML tasks include grouping of data points (clustering), classifying data into types (classification) or predicting numerical values from a set of starting conditions (regression).

ML methods are ‘fitted’ to perform a process through iterative optimisation of internal parameters often called ‘training’. Through comparisons of input data to desired output conditions, parameters are modified according to defined criteria in an algorithmic process. When a defined set of criteria are met, training ends and the modified parameters are collectively commonly referred to as the trained model.

The underlying assumptions, characteristics and limitations of ML methods differ between distinct model types and the choice of model parameters. The number of optimisable parameters roughly indicate the complexity of the process that an ML method will fit: more parameters often mean training of a more complex sequence of processing steps. ML methods also differ in the nature of processing steps, meaning that some ‘train’ more complex processes, inherently. For example, logistic regression models train a series of linear relationships (‘weights’) between data to make predictions. On the other hand, Classification and Regression Tree (CaRT or decision tree) models train a series of binary conditions to make predictions. Relationships between input data are more complex in CaRT, where not only increase or decrease in values influence predictions (as in logistic regression), but whether values meet thresholds.

The advantages of more complex models lie in identification of novel relationships between data, too abstract to otherwise be formulated. If the complexity more accurately represents the relationship between features in input data, complex models can lead to better performance in tasks (where less complex models ‘underfit’ the true complexity of relationships).

Disadvantages in complex models lie in increased computational demands, and potentially worse performance, if the model trained over-estimates the complexity of relationships in data (termed ‘over-fitting’)[75, 76]. Fitting ML methods for a given task depends on identifying an acceptable balance of complexity, to accurately model the general relationships between features in the input data and desired output criteria.

Deep Learning

Deep Learning (DL) methods are a subgroup of ML methods, with hundreds to many billions of optimisable parameters, united by the use of artificial neural networks (ANNs) somewhere in their design[75, 77]. The larger scale of parameters compared to ML increases the complexity of processes fitted, meaning that typically, more complex relationships can be identified between input features. For example, prediction of exact plant height from genetic sequences may potentially be more accurately performed with DL than ML, since height is influenced by complex relationships between thousands of so-called quantitative trait loci (qTLs), as well as thousands of other sites influencing interactions with the environment, metabolism and resource management.

ANN models generally involve sequential layers of nodes, commonly termed perceptrons. Input data is either passed directly to the first layer of perceptrons or preprocessed. Between layers of perceptrons, matrix multiplication is performed with data values from previous layers and trainable parameters. The results are transformed by a decision function (commonly called an activation function) to obtain new data values, passed to subsequent layers. Trainable parameters mentioned above are optimised in model training (often termed ‘weights’). This has the effect of modulating influence of each perceptron to each one in the next layer, facilitating an immense range of possible processes that can be

fitted to the data. Considerations such as number of perceptrons in each layer, number of perceptron layers, activation function chosen, data preprocessing steps and interactions between distant perceptrons can lead a vast range of different trained processes and thus, qualities of model[75–77]. A reference on ML/DL specific terminology is provided in the Glossary.

Histopathology methods

ML and DL paradigms common in computational histopathology[78] and relevant to the work presented include the following:

- **Convolutional Neural Networks (CNNs):** ANNs which use image ‘convolution’ transformations similar to differentiation in pre-processing of data. Progressively more complex properties of image regions, from boundaries between objects to textures, are captured by successive layers of convolution[77]. CNN variants which use values from a range of convolution layers, and so, encode features over a range of morphological complexities are called UNet models[79]. These have demonstrated success in encoding distinctive image features, as evidenced through classification of nuclei from images and image reconstruction[80–83].
- **Graph Convolutional Networks (GCNs):** ANNs which use a topographical representation of input data (graphs), describing not only feature values but relationships between data points. While the ANNs described above have grid-like structures, with data passed sequentially through layers of perceptrons, GCN architecture is not bound to a grid structure. Perceptrons can be connected to any other perceptrons, meaning that investigators have a wide range of choice in initial architectures. Data flow is also not restricted to a unidirectional progression through the network. It is typically passed and transformed between adjacent nodes (weighted by respective distance), using trained parameters to either model degree of influence from each feature or update distances between perceptrons[84]. Graph convolutions relate to progressive ‘rounds’ of data transfer and transformation between perceptrons, with gradual increase in influence from distant perceptrons. Such representations

have outperformed CNNs in some contexts[85, 86], highlighting context-dependent importance of interactions (near, far and bi-directional) in biological systems.

Example use of CNN or GCN networks in non-Hodgkin lymphoma are presented in Figure 1.2, below. In the presented case, the objective of models is to predict whether input images show cancerous B-cell masses. The major differences between typical CNN and GCN embeddings of histopathology data can be seen, with topographical maps in GCN and extraction of abstract morphological features with CNN. Also exemplified are the key differences in data transfer direction and network architecture constraints between typical ANNs and GCNs.

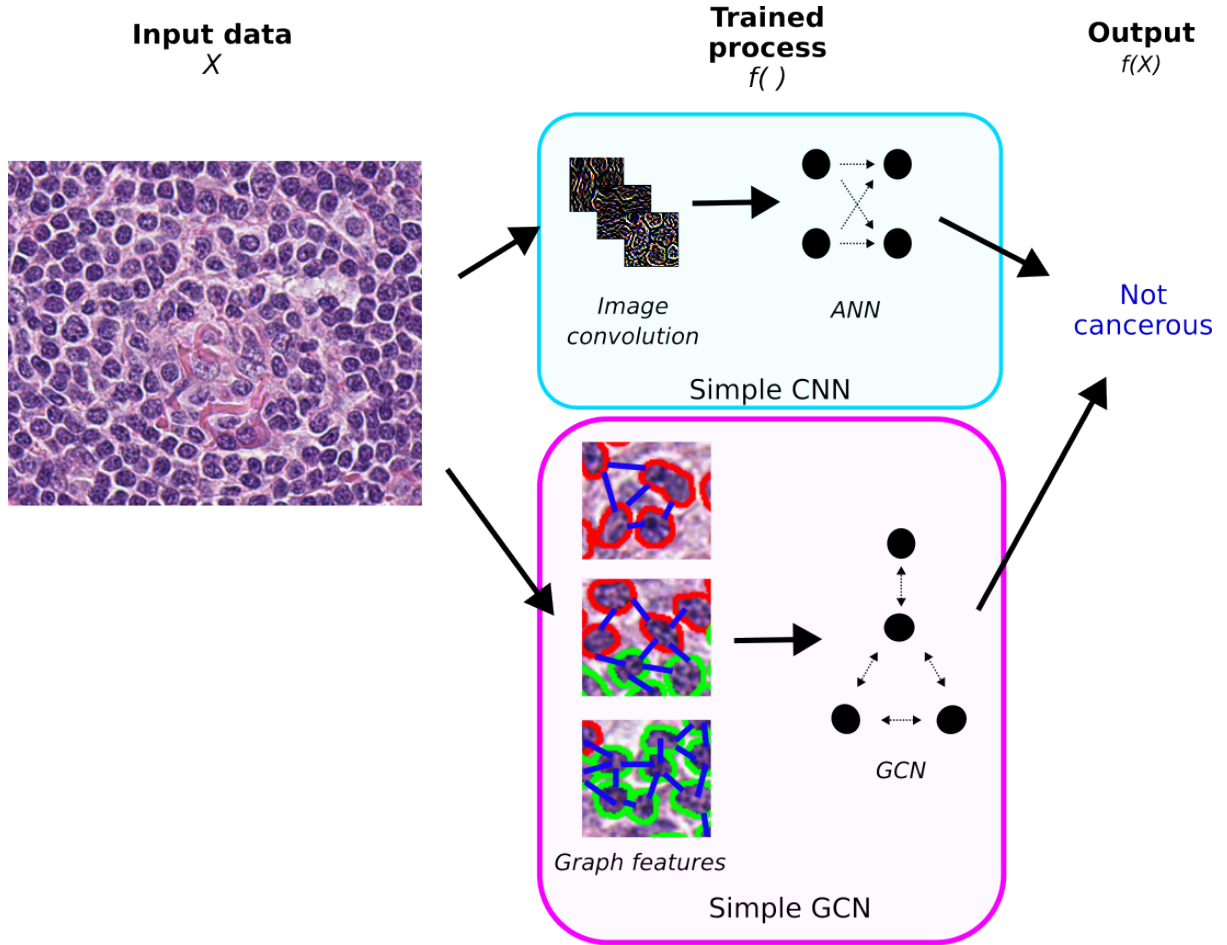


Figure 1.2: Simple examples of a convolutional neural network (CNN) and graph convolutional network (GCN) in a non-Hodgkin Lymphoma context: classification of cancer given a naïve B-cell region image.

In the CNN (top), convolution transformations are performed on parts of the input image, typically increasing contrast of morphological features. Convolution images are linearised and passed through an ANN. Finally, data in the final ANN layer are transformed to make a prediction of whether the input image showed a cancerous sample, or not.

In the GCN example (bottom), the image is converted to a graph representation. In this case, nuclear subgraphs were constructed to represent cells likely to interact. Subgraph data is passed to the GCN, which transforms features of nuclei according to connected nuclei. Aggregation of data across subgraphs can provide a prediction of whether the input image contained a cancerous mass. Line colour key: black arrow = process step; Dashed black arrow = data transfer with learned parameters; blue line = distance between adjacent objects in graph representation; nuclear border in graph representation = a token set of nuclear features, potentially cell type or expression profile

Aims and output

The aims of this thesis are broadly to use developments in computational analysis of histopathology data to subgroup non-Hodgkin lymphoma patients. These efforts are focused on subgroups informative on patient overall survival; molecular features (such as mutation, expression and protein presence) and treatment-specific effects. Each chapter includes use of nuclear morphology features, which inform histopathologist classifications in lymphoma. Motivation for this focus over more abstract features was that they were generally interpretable and, empirically, a bridge between the language of those confined to the microscope (histopathologists) and the keyboard (me).

Thesis Structure

Chapter 1: Cytological Grading

In Chapter 1, we present an effort to automate cytological grading of Follicular Lymphoma. Composed of follicle (naive B-cell region) and naive B-cell (centroblast) classifiers, it aimed to follow a similar procedure to histopathologists in cytological grading. Novel aspects of this model are operation solely on H&E, routine in FL histopathology, and (to the best of my knowledge) the first comparison of automated grades to mutation profiles and overall survival. The classifier was validated against pathologist classifications and survival in internal and external (GALLIUM clinical trial) data.

Work was presented at ECDP 2023 (details below).

Chapter 2: Nuclear Signatures in Non-Hodgkin Lymphoma

In Chapter 2, we present unsupervised approaches of clustering non-Hodgkin Lymphoma patients on nuclei detected within B-cell regions of histopathology samples. Two methods are proposed with one operating on median cell concentrations (nuclear phenotype) and the other on the distribution of all cells detected (nuclear distributions). Patient clusters were compared on features deemed to have clinical utility: overall survival and mutation in FL/cell-of-origin expression in DLBCL. Finally, validation was performed on external

datasets (GALLIUM trial for FL and DLBCL-MORPH for DLBCL). Nuclear distribution clusters were presented at ECDP 2024 and MIUA 2024, with subsequent publication of proceedings (details below).

Chapter 3: Immunohistochemistry prediction in Non-Hodgkin Lymphoma

In Chapter 3, we present unsupervised FL and DLBCL clusters, grouped on protein presence. Using popular methods of immunohistochemistry (IHC) data quantification, protein presence was estimated across panels of 11 and 13 lymphoma-associated proteins in FL and DLBCL, respectively. Estimates were compared with pathologist estimates for CD10, BCL6 and MUM1. Patients were clustered on estimates of IHC staining, with comparison against mutation (FL) and expression profiles (DLBCL). Finally, Graph Neural Networks were trained to predict protein (IHC) clusters using H&E data, with comparisons of overall survival and mutation/expression profiles between groups. DLBCL cluster characteristics were validated externally using the DLBCL-MORPH dataset. A research paper manuscript is being written to report this work, with inclusion of nuclear clusters from chapter 2 (details below).

Research output

Papers

- **(Manuscript in progress) Volodymyr Chapman**, Alireza Behzadnia, Cathy Burton, Dan Painter, Alex Smith, Reuben Tooze, Andrew Janowczyk & David Westhead
Cellular phenotypes of disease in Follicular and Diffuse Large B-cell Lymphomas.
Related to nuclear phenotype clusters in Chapter 3 and IHC clusters in Chapter 4.

Contributions:

- VC: Conception, analyses and writing.
- AB: Annotation of nuclei.

– CB, DP, AS, RT, AJ & DW: Feedback on methods and manuscript.

- **Volodymyr Chapman**, Alireza Behzadnia, Cathy Burton, Dan Painter, Alex Smith, Reuben Tooze, Andrew Janowczyk & David Westhead (2024)

The FLY-HEi index for overall survival prediction in Follicular Lymphoma patients undergoing watchful waiting. *In Thomas, C., Kendrick, C., Cootes, T., Reeves, N., Yap, M. H., Zwigelaar, R. ed. Medical Image Understanding and Analysis, 24th-26th July 2024, Manchester UK.* Frontiers in Medical Technology, pp. 136-141. ISBN:9782832512449. DOI:10.3389/978-2-8325-1244-9.

Related to nuclear distribution clusters in Chapter 2: Patient nuclear subgroups.

Contributions:

- VC: Conception, analyses and writing.
- AB: Annotation of nuclei.
- CB, DP, AS, RT, AJ & DW: Feedback on methods and manuscript.

Conference Presentations

- **Volodymyr Chapman**, Alireza Behzadnia, Reuben Tooze, Cathy Burton, Dan Painter, Alex Smith, Andrew Janowczyk & David Westhead (2024) Nuclear distribution clusters associate with mutation profile in Follicular Lymphoma. **20th European Congress on Digital Pathology (ECDP)**, Vilnius, Lithuania.
- **Volodymyr Chapman**, Alireza Behzadnia, Reuben Tooze, Cathy Burton, Dan Painter, Alex Smith, Andrew Janowczyk & David Westhead (2023) Automation of cytological grading in Follicular Lymphoma evidences polarising impact of immunochemotherapy in high grade cases. **19th European Congress on Digital Pathology (ECDP)**, Budapest, Hungary.

Conference Posters

- **Volodymyr Chapman**, Alireza Behzadnia, Reuben Tooze, Cathy Burton, Dan Painter, Alex Smith, Andrew Janowczyk & David Westhead (2024) FLY-HEi nuclear distribution clusters associate with clinical features in Follicular Lymphoma. **28th UK Conference on Medical Image Understanding and Analysis (MIUA)**, Manchester, UK. **Winner of Best Poster Award - Abstract paper category.**
- **Volodymyr Chapman**, Alireza Behzadnia, Reuben Tooze, Cathy Burton, Dan Painter, Alex Smith, Andrew Janowczyk & David Westhead (2024) Histopathology classifier for high risk Diffuse Large B-Cell Lymphoma. **20th European Congress on Digital Pathology (ECDP)**, Vilnius, Lithuania.

Chapter 2

Cytological Grading

1 Abstract

Introduction

Cytological grading in Follicular Lymphoma (FL) involves quantification of neoplastic cells, centroblasts, within naive B-cell regions of lymph nodes (follicles). Manual grading is laborious, suffers from poor reproducibility, and is susceptible to sampling bias, limiting evaluation of its clinical utility. Previous efforts have successfully improved inter-observer agreement with computer assistance but did not investigate relationships between grade and survival. We present an automated pipeline with associated survival analysis.

Methods

335 H&E whole slide images (WSIs) of the HMRN dataset, scanned at $40\times$ magnification were used for development ($n = 27$) and internal validation ($n = 308$) of an automated cytological grading pipeline. External validation was performed on 526 WSIs of the GALLIUM trial dataset.

First, naive B-cell regions (follicles) were classified using an ensemble of 3 UNet models, trained on 9 exhaustively-annotated WSIs and evaluated on 6 further WSIs.

Then, HoVerNet was used for segmentation of nuclei within follicle regions. 67,000 pathologist-labelled nuclei were used to train a centroblast classifier, utilising the 5 most informative features from 66 nuclear features. Grade was assigned on number of centroblasts per unit area, calibrated against pathologist annotations. The automated pipeline was evaluated against pathologist grade using a range of regression and classification met-

rics. Clinical utility of cytological grading was assessed using Cox Proportional Hazards models.

Results

The follicle region classifier performed similarly to published classifiers with naive B-cell region DICE scores of 0.70 and 0.41 against H&E and CD10 IHC ground truth, respectively. Centroblast counts showed high agreement with a pathologist assisted by a quadrat, achieving R2 score of 0.68 on validation data. Grading of regions of interest, with fixed magnification and field of view, resulted in agreement comparable to published inter-pathologist agreement, with a Cohen κ of 0.55. The model performed similarly to the published FLAGS system, achieving AUC-ROC of 0.88. Classification of high grade at the slide level resulted in Cohen κ close to 0, evidencing poor agreement with pathologists. Pathologist high grade was associated with significantly increased hazard against low grade (HR high grade = 2.62 (95% CI: 1.29-5.32) while machine high grade was associated with reduced hazard on the same data (HR high grade = 0.22; 95% CI: 0.03-1.63).

Conclusions

Results evidenced significant clinical utility of cytological grading. The proposed pipeline achieved centroblast classification and quantification with high pathologist agreement, but substantial differences were observed in aggregation of results for a whole slide image classification.

2 Introduction

2.1 Challenges in Follicular Lymphoma

Follicular Lymphoma (FL) is a success story of treatment innovation in indolent cancers, with a 5-year relative survival increase from 66% to 92% in 30 years [87]. Prognostication still relies heavily on clinical features such as age, Ann Arbor stage and presence of B-symptoms[88, 89], despite burdensome pathologist analyses. Ideal outcomes in terms of analysis time, efficiency of pathology experts and patient care would be for procedures with little clinical utility to be limited or delegated to software and/or for greater value to be extracted from routine pathology procedures. These needs become ever more acute with global pathologist shortages[1]. Cytological grading of Follicular Lymphoma is one such time-intensive pathology procedure, with conflicting recommendations for use[23, 24] and an algorithmic process seemingly well-suited for computer automation. In this study, we aim to:

1. Automate cytological grading
2. Assess its clinical utility
3. Ascertain whether
 - A) It has value in the modern age of immunochemotherapy
 - B) Automation provides value for clinicians and ultimately, patients

2.2 Cytological Grading

In the highly dynamic follicular lymphoma lymph node environment, tumour burden from motile, migratory B-cells is assessed indirectly using ‘cytological grade’. In this procedure, concentration of neoplastic B-cells at the early, centroblast stage is quantified within abnormal B-cell regions, follicles, with loss of cell polarisation present in otherwise non-cancerous inflammation (reactive germinal centres; GCs). Using arbitrary thresholds, patients are assigned cytological grade from one to three on aggregation of ten sample

follicles. Grade 3 is further subdivided into 3A, where more developed centrocyte cells are present and 3B, characterised by homogeneous centroblast regions.

Some evidence exists that classification of high grade has clinical utility in treatment choice, especially in patients with symptomatic disease [90–93].

2.3 Limitations of grading

The grading procedure suffers critical limitations which lead to poor consistency, reproducibility and inter-observer agreement:

- Sampling bias through evaluation of a small fraction of naive B-cell regions[63]
- Subjectivity in nuclear classification[59]
- Error introduced through inconsistencies in grading procedure[61, 94]

Agreement between experts can vary wildly with reported Cohen’s kappa values between 0.18 and 0.71[59, 61]. Such inconsistencies have led to conflicting recommendations for the procedure, with re-assignment as an optional assay by the WHO[23] but guidance on continued usage by an international committee of haematopathology experts[24].

2.4 Previous Work

Immunohistochemistry Dependence

Previous efforts have successfully improved inter-observer agreement in centroblast classification and grading[61, 63]. These strategies relied on immunohistochemistry (IHC) for selection of naive B-cell regions to be graded[63, 95, 96]. Though clearly effective in those efforts, IHC presents a considerable additional expense[97], has limited utility in differential diagnosis of FL and so, is not routinely or consistently performed. Where it is available, error can be introduced through:

- Serial sections of biopsy slides, cut 2-5 μ m apart potentially resulting in large distances between IHC slide regions and the corresponding H&E, for which they inform region selection

- Variable presence of the naive B-cell markers CD10 and BCL2. These markers are more frequently absent in the higher grade cases for which grading may hold clinical utility[98, 99](Figure 2.1 second and third rows, respectively)
- Potential for poor contrast with the general B-cell marker CD20, previously used[63, 95, 96](Figure 2.1 right). High concentrations of B-cells within all areas of the lymph node, even in high grade disease[99], may potentially limit scope for selection of specific regions

Region selection with H&E staining also presents technical challenges, namely in variable contrast between lighter, lower cell density naive B-cell regions and classically darker, high cell density interfollicular regions (Figure 2.1 top row). Despite this, an automated approach relying solely on H&E would have wider application within the FL diagnosis process.

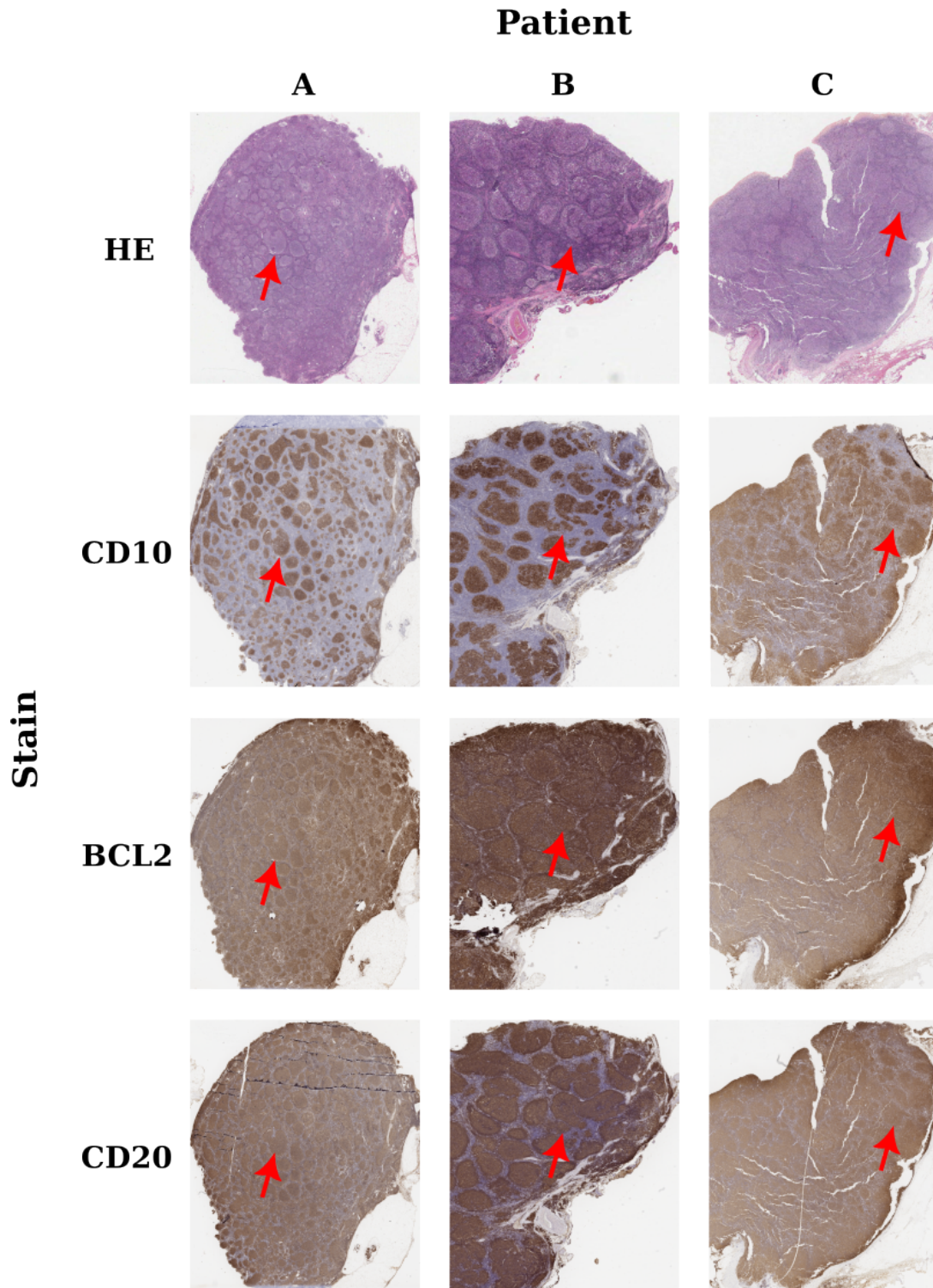


Figure 2.1: Comparison of naive B-cell region staining methods. Red arrows pointing to naive B-cell region (follicle) landmarks between comparable regions of stained serial sections of FFPE tissue from the HMRN dataset[100]. H&E = Haematoxylin and Eosin staining. CD10, BCL2 and CD20 refer to immunohistochemistry staining of the corresponding B-cell protein markers.

Centroblast Classification

A range of methods have been used for centroblast classification, with earlier methods operating on texture, intensity and geometric features[58, 60] while later models utilised nonlinear embeddings of intensity features[62, 101]. Validation of all models identified has involved comparatively small datasets of 30-220 centroblasts.

As in many cases of FL, centroblasts are recognised as the neoplastic cell in Diffuse Large B-cell Lymphoma[102]. In this related disease, interpretable geometric features of nuclei predicted to be neoplastic, such as area and Feret diameter, were used to predict survival, achieving a concordance index of 0.64[66]. This work evidenced clinical utility of geometric features, but interpretation of results is complicated by the choice of model used to classify neoplastic cells (HoVerNet[103]). In empirical analyses, we found that HoVerNet classified almost all cells within naive B-cell regions as neoplastic, even in non-cancerous samples (Figure 2.2). This observation evidenced that cell concentrations were a confounding variable in this analysis, with the method used to aggregate feature values for all nuclei within a sample (mean) sensitive to the balance between small, jagged centrocytes and large, more circular centroblasts. Overall, results from Vrabac et al.[66] evidenced utility of geometric features in lymphoma survival prediction and indirectly, utility in classification of centroblasts with modern nuclear segmentation methods.

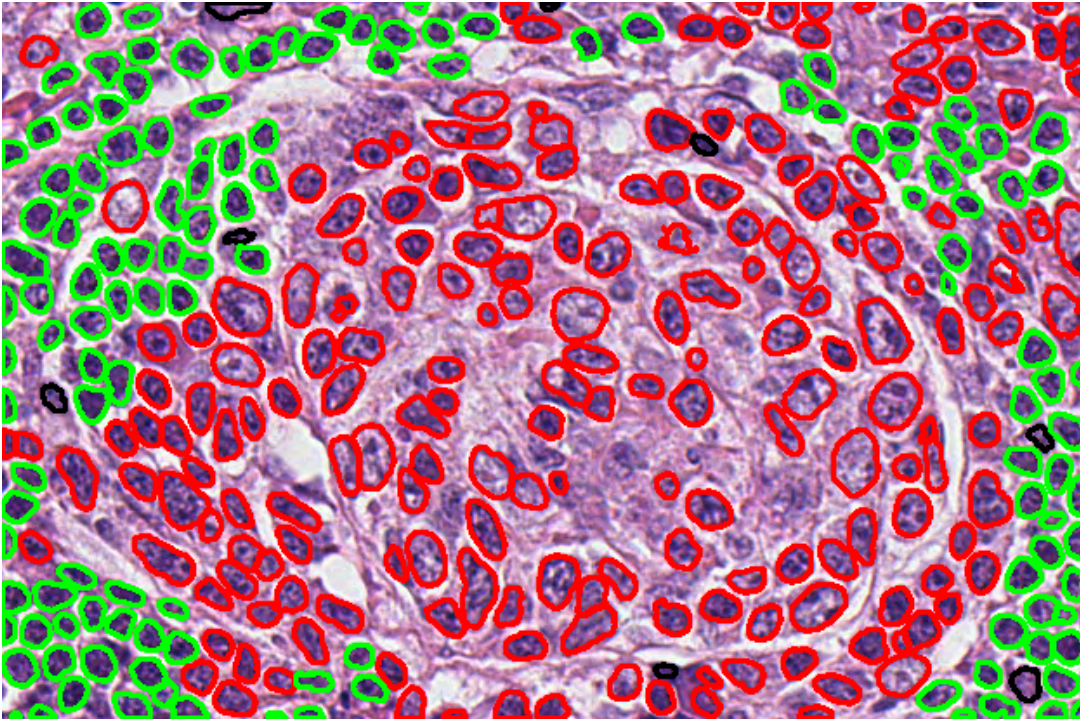


Figure 2.2: HoVerNet[103] inference on non-cancerous, reactive lymph node naive B-cell region (germinal centre), demonstrating erroneous widespread classification of cells in these regions as neoplastic. Border colour prediction key: Red = neoplastic; Green = inflammatory; Black = other. Image from H&E slide of HMRN dataset[100].

In summary, previous efforts have shown promise in centroblast classification using interpretable geometric, texture and intensity features as well as abstract embeddings of colour intensity. Most efforts have suffered from comparatively low centroblast numbers in evaluation, leading to uncertainty of true performance.

2.5 Work summary

In this study, we evaluate methods of centroblast classification from those operating on geometric, texture and intensity features (random forest model on nuclear features) to more recent methods operating on abstract embeddings of intensity (DenseNet169 and EfficientNetV2s on nuclear images). Next, we create a pipeline of supervised naive B-cell region (follicle) and centroblast classification to perform automated cytological grading, operating on entirely on H&E slides. Finally, for the first time (as far as the authors are aware) we perform survival analyses with predicted grades.

The particular items of novelty in this study are:

1. Development of a supervised H&E follicle classifier, with performance evaluated against H&E and IHC
2. Benchmarks of a supervised centroblast classifiers utilising objective nuclear features against deep learning approaches
3. Development of a cytological grading tool operating solely on H&E WSIs, allowing widest usage
4. Evaluation of the clinical utility of cytological grading, with respect to overall survival

3 Methods

3.1 Experiment Workflow

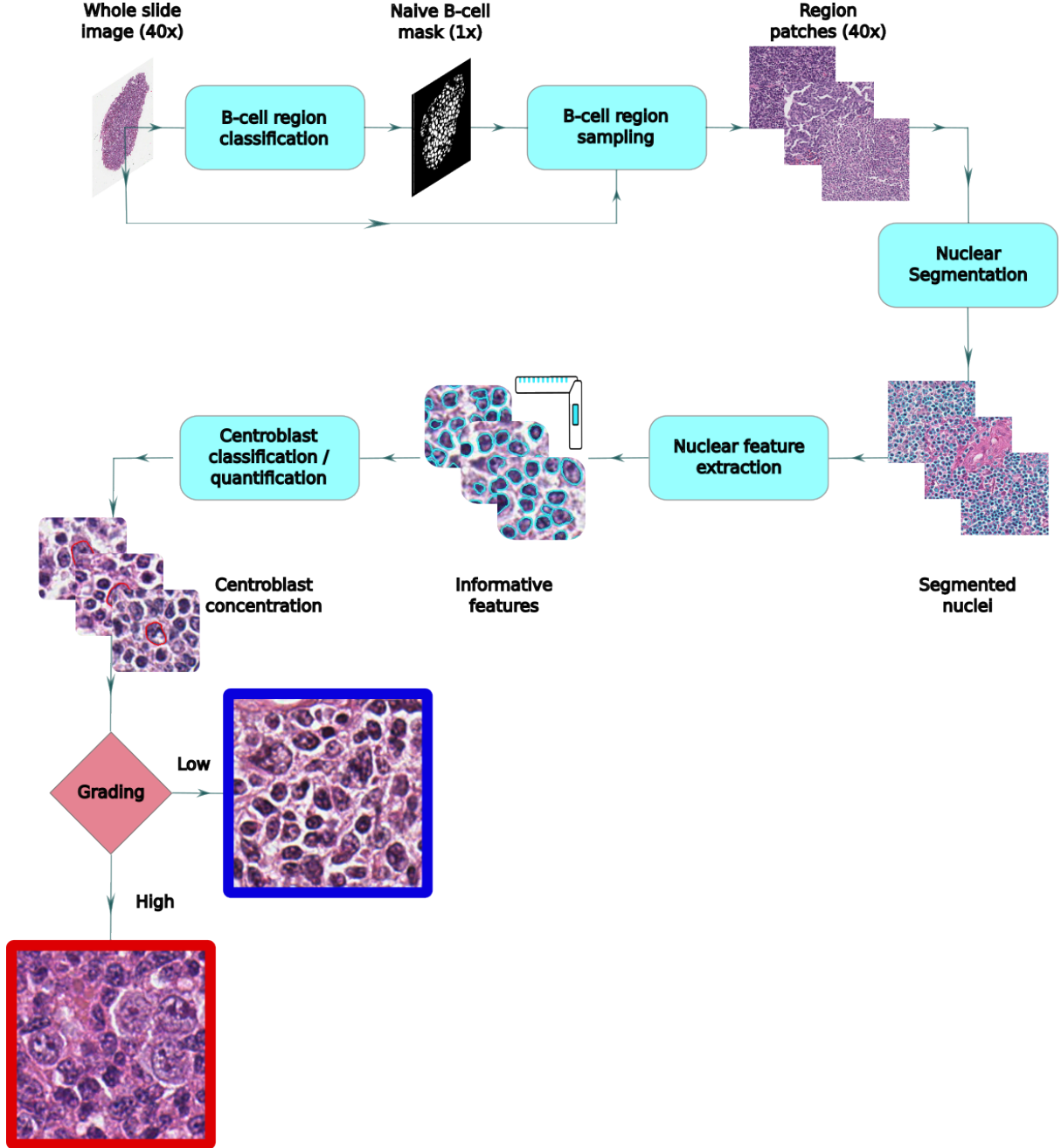


Figure 2.3: **Automated Cytological Grading workflow**

The proposed method aimed to follow the clinical guidelines of cytological grading closely [104], with narrowing of cell analysis to naive B-cell regions (termed follicles) and subsequent quantification of cells predicted to be centroblasts within those regions. Finally, grade was assigned through comparison of centroblast counts against grading thresholds.

3.2 Patient Data

Data used are from the UK population-based Haematological Malignancy Research Network (HMRN) dataset[100], comprising data on all novel hematological malignancies from 14 hospitals in the Yorkshire & Humberside regions from the 1st September 2004 to the 31st August 2012. Clinical characterisation, including diagnosis, treatment choice and tracking of outcome until the 31st December 2018 was conducted by the Hematological Malignancy Diagnostic Service (HMDS; www.HMDS.info).

Slide scanning was initiated on 548 patient samples, diagnosed with Follicular Lymphoma (FL) according to International Classification of Diseases for Oncology, Third Edition [105], with previous targeted sequencing [46].

Whole slide, H&E-stained images of the GALLIUM trial dataset [106, 107], were used for external validation of the classifier.

3.3 Ethics statement

Fully anonymised data was collected from HMDS patient records following ethical approval (Research Ethics Committee reference 04/Q1205/69).

3.4 Whole slide images

Formalin fixed paraffin embedded (FFPE) tissue was prepared and stained with H&E by the HMDS as part of a routine diagnostic process. Prepared slides were scanned by the Slide Scanning Service, St. James’s University Hospital, Leeds using a Leica Aperio AT2 scanner at 40 $\times$ magnification. All downstream analyses were performed using Python 3.8.18. Whole slide images (WSIs) were parsed with OpenSlide 3.4.1, using OpenSlide-Python 1.3.1 for Python bindings. WSIs were visually inspected before use, with exclusion of slides containing features which would likely lead to excess noise in classification tasks, including the following:

- needle core biopsies

- low tissue content, defined as less than 1000×1000 pixels of tissue
- excess fibrotic tissue (evidencing necrosis or sclerosis)
- poor fixation
- anomalous staining
- artefacts, such as excessive dust

Comparable visual quality control was performed on external validation data.

For the majority of patients in both the HMDS and GALLIUM cohorts, a single H&E slide was available for analysis. For 4 patients in the HMDS cohort, multiple slides were available. These were from biopsy samples stored in separate FFPE blocks, likely due to tissue separation or sampling from multiple sites during the surgical procedure. Since there was a chance that samples were taken from different biological sites, results of multiple slides for a patient could not be reliably merged through averaging. Each slide was therefore treated as a separate data point.

3.5 Data partitions

Data acquisition was performed throughout the duration of the reported work. Initial data available was randomly partitioned into training and internal validation partitions. Subsequent WSIs were added to the internal validation partition. All slides underwent visual quality control (QC), as described above. Following QC, the GALLIUM trial dataset was assigned into three (partially overlapping) groups on quality:

1. Sufficient tissue, defined as slides with at least 10 patches of 1000×1000 pixels at the highest magnification level of the image. This included images scanned at $20 \times$ magnification and needle core biopsies.
2. Passes, with further filtering for scanning at $40 \times$ magnification; exclusion of needle core biopsies and other slide level artefacts.
3. Strict passes, with further exclusion of slides on artefacts caused by cover slips,

staining etc.

Partitioning schema and exclusion criteria are outlined in CONSORT diagrams (Figure 2.4) below. For the HMRN dataset, cases were first excluded on slide QC features (79/414 excluded) then split into a model training ($n = 27$) and validation partition ($n = 308$). With GALLIUM trial data, slides containing tissue from outside of lymph nodes or without associated clinical data were excluded (490/1070 excluded). Next, cases were filtered on sufficient tissue (>10 patches; 54/580 excluded) and finally, slides were assigned to groups with increasingly stringent QC selection criteria.

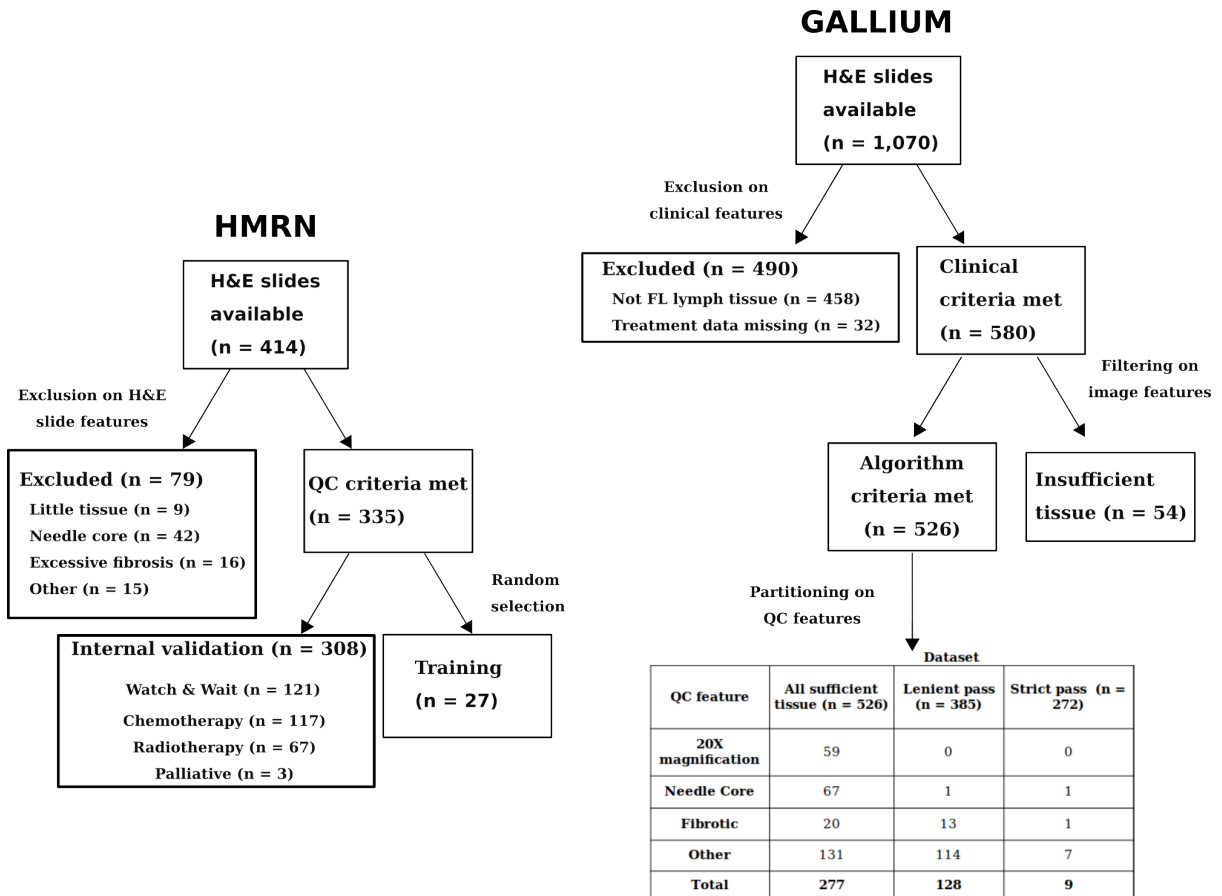


Figure 2.4: Visual quality control and partitioning of the datasets

Partition clinical characteristics are described in Appendix Table A.1 Patient characteristics.

3.6 Nuclear Segmentation

Algorithms

To identify an optimal nuclear segmentation algorithm for the proposed use case, segmentation performance was compared against manually segmented nuclei. In total, 831 ground truth nuclei were segmented in 50 128×128 pixel tiles extracted from naive B-cell regions of 4 FL WSIs.

Three published segmentation algorithms were evaluated: StarDist[108], CellPose (version 1)[109] and HoVerNet[103]. For Cellpose and HoVerNet, multiple publicly available versions of the models were evaluated. These versions, referred to as ‘checkpoints’, were created by the algorithm developers through training of the same underlying model architectures with different datasets. The ‘cyto’ and ‘nuclei’ checkpoints of Cellpose were evaluated, trained on fluorescent cytoplasmic markers and nuclear segmentations, respectively, as reported[109]. The CPM-17[110], CoNSeP[110], PanNuke[111] and Kumar[112] checkpoints of HoVerNet were evaluated (trained as previously reported[103] on the relevant datasets). The MoNuSAC checkpoint was also used[113]. Finally, both Cellpose and StarDist required selection of hyperparameters in inference. To evaluate general model performance, results were aggregated over a wide range of parameters for each method ($n = 216$ sets for Cellpose and $n = 300$ sets for StarDist). Therefore, median performance values presented (Table 2.1) were calculated using 1 prediction for 50 ground truth images in the case of HoVerNet checkpoints; 300 predictions for each ground truth image in StarDist ($n = 1,500$) and 216 predictions per ground truth image for Cellpose checkpoints ($n = 1,080$).

3.7 Naive B-cell Region Classification

Naive B-cell regions (follicles) of tissue were classified to direct nuclear sampling. To begin, preprocessing was conducted through tissue region classification and removal of artefacts based on methods in HistoQC [114]. First, tissue regions were segmented using filtering of grayscale image intensity at $1 \times$ magnification. An empirically determined

minimum intensity threshold of 0.2 removed dust and other dark artefacts in digital slides and maximum threshold of 0.8 removed light background. Next, binary erosion was performed to remove objects with area lower than 5,000 square pixels, such as small pieces of tissue and fat. This was conducted using the Scikit-Image library [115].

Following preprocessing, 3 UNet models[79] were trained on 9 H&E WSIs, exhaustively annotated using Quick Annotator[116] at $1\times$ magnification to classify follicle pixels (process displayed in Figure 2.5, below). These adapted convolutional neural network (CNN) models accepted image patches as input, which underwent local aggregation procedures termed ‘convolution’. During this, filters were used to effectively ‘emphasise’ image features, including horizontal and vertical boundaries. Each sequential convolution layer resulted in emphasis of more complex image properties from lines, to compound shapes and finally, texture features. Following four layers of convolution, aggregated, complex image features were reconstructed into naive B-cell region masks of the same shape as the input image, termed ‘deconvolution’. The UNet architecture used was distinguished from the CNN described through ‘skip connections’ (lines between convolution and deconvolution layers in Figure 2.5), in which image features extracted from each convolution layer were made available during image reconstruction (deconvolution), as opposed to just the most complex features extracted.

A voting ensemble was formed of the three top-performing models, with respect to AUC-ROC on training data, in which classification of each pixel (B-cell region or not) was determined through a majority model vote.

Models were evaluated against exhaustively annotated H&E ($n = 6$) and CD10 IHC ($n = 5$; registered to H&E[117]) ground truth using a range of segmentation metrics. Image patches used in downstream analyses were filtered for a minimum of 90% tissue coverage.

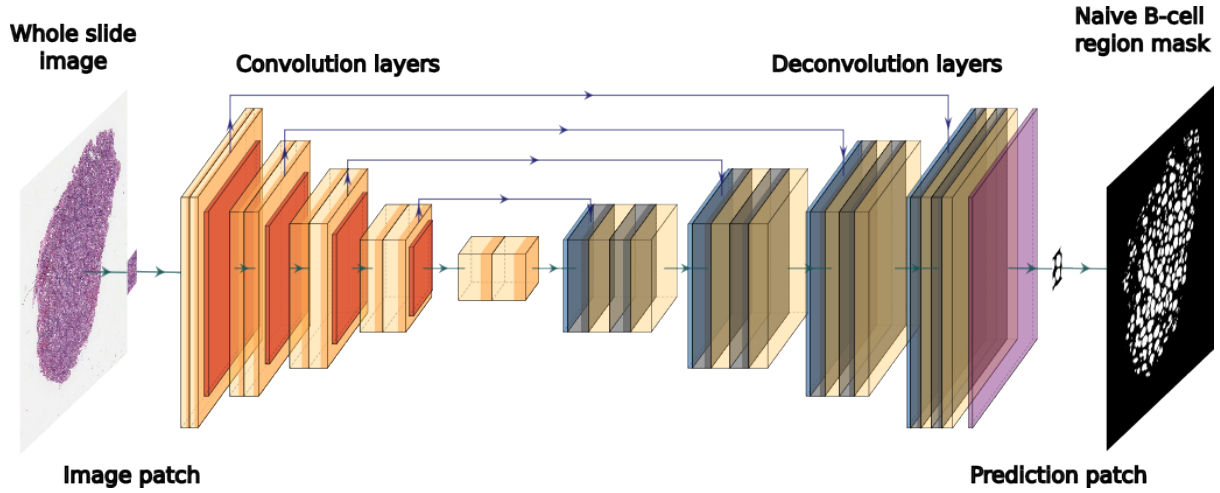


Figure 2.5: The architecture of a single UNet model, as used for prediction of naive B-cell regions (follicles). Final follicle classification was performed using voting with 3 UNet models.

3.8 Centroblast Classification

Two broad methods of centroblast classification were evaluated. The first was a machine learning approach, using random forest classifiers [118] trained on nuclear features describing morphology, colour intensity and texture. The second was a deep learning approach using pretrained computer vision models to classify nuclei, in a process called transfer learning.

A dataset of pathologist labelled nuclei was created for training of algorithms, selection of an optimal one and evaluation of the final model. To enable classification of centroblasts across a range of morphologies, tissue samples were used from three types of non-Hodgkin Lymphoma with naive B-cell (blastoid cell) involvement, as reported by the WHO[23]: Burkitt Lymphoma (BL), Diffuse Large B-cell Lymphoma (DLBCL) and Follicular Lymphoma (FL). 15 RGB region patches were extracted from 3 WSIs each of BL, FL and DLBCL (totalling 9 WSIs and 45 region patches). 67,786 objects were segmented in these patches using the MoNuSAC checkpoint of HoVerNet [103]. Segmented objects were labelled by a pathologist into three classes - Centroblast, non-centroblast nucleus and non-nucleus - using Patch Sorter[119].

Nuclei were randomly split into training ($n = 40,671$; 60%), cross validation (for selection of an optimal model; $n = 13,557$; 20%) and test (for evaluation of the best-performing

model; $n = 13,558$, 20%) partitions.

All approaches to centroblast classification used identical data partitions. Algorithms were trained to predict the multiclass pathologist label of objects (non-nuclear, centroblast and non-centroblast). Once an optimal general approach had been identified, that model was trained on the binary task of predicting centroblasts vs. non-centroblasts, with exclusion of non-nuclear objects from training data.

Machine Learning approach

66 Nuclear features describing texture, morphology and intensity were measured for each nucleus (Scikit-image library features used for morphology and intensity: region features; texture features described in Supplementary files, Chapter_1 folder, Texture_features.xlsx). Feature values were clipped at the 1st and 99th percentiles, to reduce the impact of outliers. Features were then linearly scaled to the range of 0 and 1 to promote equal contribution in model fitting.

To maximise the signal-to-noise ratio, redundant features and those not informative for classification of cell types were removed. Features with Spearman's rank correlation values above an empirically-determined value of 0.7 were considered to be redundant and were filtered off (except the first feature). Next, informative features were filtered on utility in classification and feature importance using permutation importance and recursive feature elimination (RFE), respectively. Both of these methods used a Random Forest classifier of 100 decision tree models (classification and regression tree; CART) with max tree depth of 5, trained on the training data partition. Permutation importance, defined as predictive performance loss due to random permutation of a feature (described in Section 10 of [118]), was used to filter informative features. Features for which permutation led to net model performance loss were deemed to be informative. Permutation importance was calculated over 50 permutation cycles (repeats). Performance was calculated using accuracy loss, defined as the reduction in model accuracy upon random shuffling (permutation) of the feature. Features with 95% confidence intervals above zero were

deemed to be informative. Finally, to filter for the most informative features, RFE was used for step-wise elimination of less informative features. A step size of 1 was used with Gini impurity as a measure of feature importance, averaged over 5 cross validation partitions.

A final random forest classifier was trained using a grid search over 150 parameter sets. 5-fold cross validation was used to select for models with capacity to generalise to unseen data. To reduce the impact of class imbalance in model fitting, SMOTE[120] was used to synthetically oversample under-represented classes of objects (centroblasts and non-nuclei) in training data.

Deep Learning approach

Two general image classification deep learning (DL) algorithms were evaluated in nuclear classification: DenseNet169 [121] and EfficientNet v2-s [122]. Models were initiated with pretrained parameters (also called weights and checkpoint; ImageNet V1). Final layers, responsible for classification, were substituted for ones with 3 classes that corresponded with the pathologist labels. In training, only parameters associated with the final layers were trained.

Images of nuclei were preprocessed to increase signal-noise ratio, ensure pixel intensities were within required thresholds and input image sizes were as required for each DL model - 224×224 pixels for DenseNet169 and 384×384 pixels for EfficientNet v2-s. All masking and padding steps described involved setting pixel values to zero.

Bounding box images of objects segmented using HoVerNet (above, section 3.8) were extracted from WSIs at 40x magnification. Background pixels were masked (with values of zero) to reduce potential noise from nearby nuclei and cytoplasm artefacts (Figure 2.6, left images of each pair). Images of nuclei were transformed as necessary for DL models, consisting of:

1. Padding of masked images on all sides, with $\frac{\text{final image side length}}{8}$ pixels

2. Crop of the centre of images, with crop side lengths equal to $\frac{\text{final image side length}}{4}$ pixels
3. Scaling of the central crop to the required image size
4. Normalisation of pixel intensities for each colour channel, with respect to parameters used in model training:
 - Red: mean = 0.485, SD = 0.229
 - Green: mean = 0.456, SD = 0.224
 - Blue: mean = 0.406, SD = 0.225

Examples of images of nuclei following processing are displayed in Figure 2.6, right images of each pair.

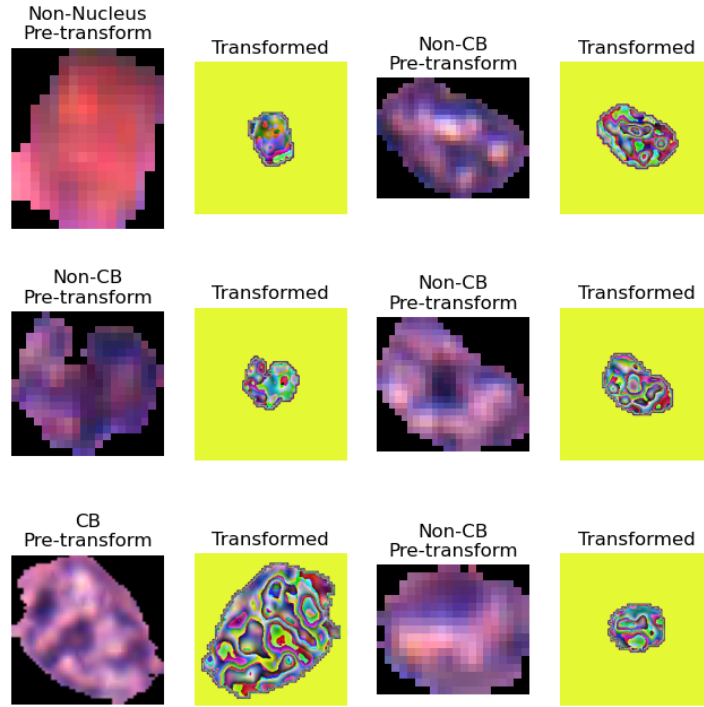


Figure 2.6: Nuclear image pairs displaying nuclear images after masking of background pixels (left of each pair) and after subsequent pre-processing for use in deep learning models (right image in each pair, denoted by ‘Transformed’ title). Pathologist ground truth labels are displayed above pre-transformation objects (left images in each pair). Non-CB = non-centroblast; CB = centroblast; Non-nucleus = labelled as non-nuclear material, such as an erythrocyte or apoptotic body. **Note:** Transformed images displayed were transformed with training augmentations, including rotation. Therefore, transformed images displayed did not necessarily have identical rotation to masked images displayed.

3.9 Automated Cytological Grading

A minimum of 10 and maximum of 200 1000×1000 pixel regions of interest (ROIs) were extracted from naive B-cell regions of slides at 40x magnification, depending on tissue availability. Nuclear segmentation and centroblast classification was performed on each ROI, as described in sections 3.6 and 3.8, respectively. Median number of centroblasts (per ROI) was used for slide-level cytological grading. Grade thresholds were determined with comparison against pathologist grade for 15 ROIs (named the calibration set). 5 ROIs were used for validation.

3.10 Inference on validation data

Cytological grading was performed on GALLIUM trial data[106, 107] using models and parameters trained on HMRN data.

Region selection was conducted using the Naive B-cell region classifier, trained on HMRN FL slides (Section 3.7). Nuclear segmentation was conducted using the MoNuSaC checkpoint of HoVerNet[103]. Nuclear features were measured and linearly scaled using parameters determined on a set of over 177,000 unlabelled nuclei across Non-Hodgkin Lymphomas (Chapter 3 Section 3.5 Pan Non-Hodgkin lymphoma nuclear dataset). Features were also linearly scaled with respect to all nuclei in the GALLIUM dataset. Therefore, 2 sets of scaled nuclear features were created - one scaled to HMRN parameters and one to the GALLIUM trial. This was conducted to allow evaluation of performance across data acquisition workflows.

Centroblasts were classified on scaled nuclear features using a Random Forest model, trained on labelled HMRN FL nuclei (Section 3.8). Finally, classification of slides into low and high grade (relating to cytological grades 1-2 and 3, respectively) was achieved through comparison of median number of centroblasts per ROI against a high grade threshold determined with pathologist annotations (Section 3.9). These grades were compared against pathologist grades assigned during the GALLIUM trial[107].

Predicted centroblast cell counts were validated against pathologist cell counts of:

- centroblasts on the 15 calibration set patches (HMRN dataset)
- large cell nuclei that could conceivably be centroblasts on 64 ROIs of the 'Strict pass' partition (GALLIUM dataset)
- centroblasts, immunoblasts, endothelial cells, macrophages and follicular dendritic cells (fDCs) on 30 ROIs of the 'Strict pass' partition (GALLIUM dataset)

Though not requested, the pathologist reported use of a quadrat for cell counts in GALLIUM data but not HMRN data. For GALLIUM data, only centroblasts predicted in the HMRN-scaled feature set were evaluated.

3.11 Survival analysis

Kaplan-Meier estimators were used for display of survival data.

Association with patient features

The Pearson correlation coefficient was calculated between binary classification of high cytological grade and patient clinical features. Given differences in clinical features measured and potential differences in procedures, correlation was calculated separately for HMRN and GALLIUM trial data.

Cox Proportional Hazards Models

The contribution of cytological grade in survival risk was assessed using Cox Proportional Hazards (Cox) models (Equation 2.1, below). To test the core assumption of Cox models, that hazards were proportional over time, the ‘check_assumption’ function of the lifelines library[123] was used for all Cox models fitted. In this, extent of time-scaled Schoenfeld residual divergence from a Chi-squared distribution was used to assess a null hypothesis (proportional hazards) for each variable. A significance threshold of 0.01 was used. The null hypothesis was not rejected for any variable or model.

For risk features: $\{X_1, X_2, \dots, X_n\}$ and associated risk co-efficients: $\{\beta_1, \beta_2, \dots, \beta_n\}$:

$$\lambda(t|X_1, X_2, \dots, X_n) = \lambda_0(t) e^{\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n} \quad (2.1)$$

Where λ is the hazard function at time point t and λ_0 is the baseline risk to the patient.

Overall survival time and patient vital status at this time, with an event defined as death by any cause, were used for fitting. This definition of an event was used over death attributed solely to lymphoma firstly, to maximise number of observed events in fitting and secondly, to reduce noise in attribution of death to lymphoma. Using death by any cause to define events could conceivably increase noise from deaths with no contribution

from FL but, the systemic burden of FL was assumed to be substantial enough to have contributed to most deaths.

Contribution of cytological grade in overall survival was measured using the exponential coefficient value e^{β} from Cox models, the hazard ratio (HR). The coefficient values β assigned by Cox models were dependent on features used in fitting. In effect, HRs can be inflated if models are fitted with features that predict events poorly. Common practice to prevent such over-estimation of HRs includes fitting of models with the inclusion of features assumed to be important in predicting events - often age, Ann Arbor Stage and performance status (ECOG) - in a process termed ‘correction’.

Statistical significance

P-values were calculated for clinical feature associations (Section 3.11, above) using Welch’s t-test on clinical features of patients assigned to high grade against those assigned to low grade.

P-values were calculated for Cox model coefficients using the Wald test of a single parameter. This calculated the probability of the observed or more extreme data, with respect to a null hypothesis in which individual coefficient values were zero ($HR = 1$).

For all p-values, a significance threshold of 0.05 was used, with Bonferroni multiple testing correction (where applicable).

Evaluation metrics

Definitions and equations for all evaluation metrics used are provided in Appendix section 4.1 Evaluation metrics.

4 Results and Discussion

4.1 Dataset characteristics

The HMRN and GALLIUM datasets displayed considerable differences in patient characteristics, as expected when comparing population and clinical trial datasets. As displayed in Appendix Table A.1 Patient characteristics, key differences between GALLIUM and HMRN datasets were:

- lower male populations than HMRN data (40.1-44.7% vs. 51.3%)
- 2.0-7.9% higher proportions of patients with B-symptoms (nights sweats, fever and weight loss)
- Higher median performance status (ECOG) and Ann Arbor stage in GALLIUM (ECOG 1 vs ECOG 0 & stage 4 vs. 3)
- Substantially greater proportion of patients with disease involvement beyond lymph nodes, with over $2\times$ higher proportion of extranodal involvement and 19.5-22% higher bone marrow involvement
- Over 11% higher proportions of elevated LDH (with respect to centre-specific thresholds) in GALLIUM
- Substantially greater use of immuno- and chemotherapy in GALLIUM, with all patients receiving each. As opposed to 35.1% administration of immuno- and 38% chemotherapy in HMRN. 61.1% of patients in HMRN either underwent radiotherapy or watchful waiting (watch & wait).
- Use of Obinutuzumab and Rituximab immunotherapy in GALLIUM, as opposed to only Rituximab in HMRN
- Preference in GALLIUM for administration of Bendamustine ($\simeq 65\%$ vs. 0.3% HMRN) and CHOP (24.3-28.6% vs. 9.2% in HMRN)
- Preference in HMRN for administration of CVP (24.6% vs. 6.3-10.6%)

- Over $3\times$ higher administration of radiotherapy in HMRN (21.8% vs. 5.6-6.8%)

Current selection criteria for immunochemotherapy are the presence of systemic (Ann Arbor stage 2+) and symptomatic disease [88, 89]. This can explain most differences between datasets including higher B-symptom prevalence; involvement of non-lymphoid tissue; stage; performance status; elevated LDH and administration of both immuno- and chemotherapies.

Conversely, radiotherapy is currently advised for localised (Ann Arbor stage 1) disease, which explains higher administration of this treatment in HMRN.

Finally, collection of the HMRN cohort of patients began in 2004 [100], before publication of results from the STiL and BRIGHT trials evidencing improved complete and partial response to Bendamustine, compared to CHOP and CVP [124, 125]. Since enrollment for the GALLIUM trial commenced several years after publication of initial results from the STiL trial, stark differences in administration of Bendamustine in HMRN and GALLIUM datasets could potentially be explained by changing standard of care after 2009.

Recommendations for management of FL distinguished patients on extent of disease spread (Ann Arbor stage and involvement of organs outside of lymph nodes), symptomatic disease (presence of B-symptoms) and risk of transformation to DLBCL. Considerable confounding variables existed when comparing disease managed through immunochemotherapy, radiotherapy and watchful waiting so, these treatment partitions were evaluated separately. With respect to external validation, the GALLIUM trial data was deemed suitable for comparison with immunochemotherapy patients of the HMRN dataset only. Despite evidence of improved progression-free survival with Obinutuzumab[107] and Bendamustine[125], both administered in far greater proportions in GALLIUM data, no significant differences in overall survival have yet been demonstrated (to the best of my knowledge). Therefore, comparison of overall survival events was chosen over progression-free survival, to allow for direct comparison between GALLIUM and HMRN despite considerable differences in immunochemotherapy regimens.

Finally, there was almost no overlap in scanning hardware used in the HMRN and GALLIUM datasets, with use of Leica Aperio and Hamamatsu scanners, respectively. Scanner

variability has previously been shown to negatively impact performance of deep learning tools[126]. Evidence has led to gold standard practices of stain normalisation[127] with computer vision models, but the utility of these with a random forest classifier operating on tabular data is unknown. To try to account for potential performance differences of nuclear segmentation models in GALLIUM data, nuclear features were linearly scaled to maximum and minimum values of nuclei in HMRN and GALLIUM datasets, separately. Finally, over 90% of patients that received chemotherapy in the HMRN dataset also received immunotherapy (107/117). For accuracy, the term ‘chemotherapy’ was used throughout this thesis to refer to those patients, but it should be noted that almost all technically received immunochemotherapy.

4.2 Nuclear segmentation methods

Published nuclear segmentation algorithm performance was compared against manually-segmented nuclei from the internal HMRN dataset, with results displayed below.

Model	Checkpoint	Median Metric value (IQR)						
		DICE	DICE 2	AJI	AJI+	DQ	SQ	PQ
Cellpose	Cyto	0.52	0.64	0.34	0.34	0.41	0.62	0.26
		(0.43	(0.59	(0.26	(0.26	(0.30	(0.60	(0.18
		-	-	-	-	-	-	-
		0.59)	0.70)	0.39)	0.40)	0.52)	0.65)	0.33)
	Nuclei	0.37	0.47	0.19	0.20	0.14	0.58	0.08
		(0.25	(0.40	(0.14	(0.14	(0.07	(0.53	(0.04
		-	-	-	-	-	-	-
		0.44)	0.57)	0.24)	0.25)	0.24)	0.62)	0.14)
HoVerNet	CPM-17	0.61	0.72	0.41	0.42	0.60	0.69	0.41
		(0.54	(0.64	(0.34	(0.35	(0.49	(0.67	(0.34
		-	-	-	-	-	-	-
		0.67)	0.76)	0.48)	0.48)	0.69)	0.72)	0.49)

	CoNSeP	0.28 (0.09 - 0.63)	0.69 (0.63 - 0.83)	0.14 (0.05 - 0.41)	0.14 (0.05 - 0.41)	0.24 (0.03 - 0.52)	0.74 (0.15 - 0.77)	0.17 (0.03 - 0.39)
	Kumar	0.76 (0.65 - 0.81)	0.84 (0.74 - 0.87)	0.58 (0.48 - 0.64)	0.59 (0.48 - 0.65)	0.75 (0.64 - 0.82)	0.80 (0.78 - 0.82)	0.60 (0.51 - 0.68)
	MoNuSAC	0.77 (0.72 - 0.83)	0.85 (0.78 - 0.89)	0.61 (0.52 - 0.68)	0.61 (0.52 - 0.68)	0.76 (0.67 - 0.83)	0.82 (0.79 - 0.84)	0.62 (0.54 - 0.68)
	PanNuke	0.72 (0.63 - 0.78)	0.81 (0.74 - 0.84)	0.55 (0.44 - 0.62)	0.56 (0.44 - 0.63)	0.72 (0.62 - 0.79)	0.78 (0.74 - 0.80)	0.56 (0.47 - 0.61)
StarDist	2D H&E	0.72 (0.64 - 0.78)	0.83 (0.75 - 0.87)	0.54 (0.45 - 0.62)	0.55 (0.45 - 0.63)	0.71 (0.62 - 0.81)	0.79 (0.76 - 0.81)	0.56 (0.47 - 0.64)

Table 2.1: Results of segmentation algorithm evaluation on 50 manually annotated follicular lymphoma patches. Higher median values indicate better segmentation performance. IQR = Interquartile range of performance across all images. Top performing model for each metric in **bold**.

The MoNuSaC checkpoint of HoVerNet was found to have best performance in all evaluation metrics, with respect to the annotated nuclei, evidencing optimal segmentation performance for the required use case (Table 2.1). Evaluation metric values evidenced comparable performance between Kumar, MoNuSaC and PanNuke checkpoints of HoV-

erNet as well as StarDist, with overlapping interquartile ranges (IQRs; Figure 2.7).

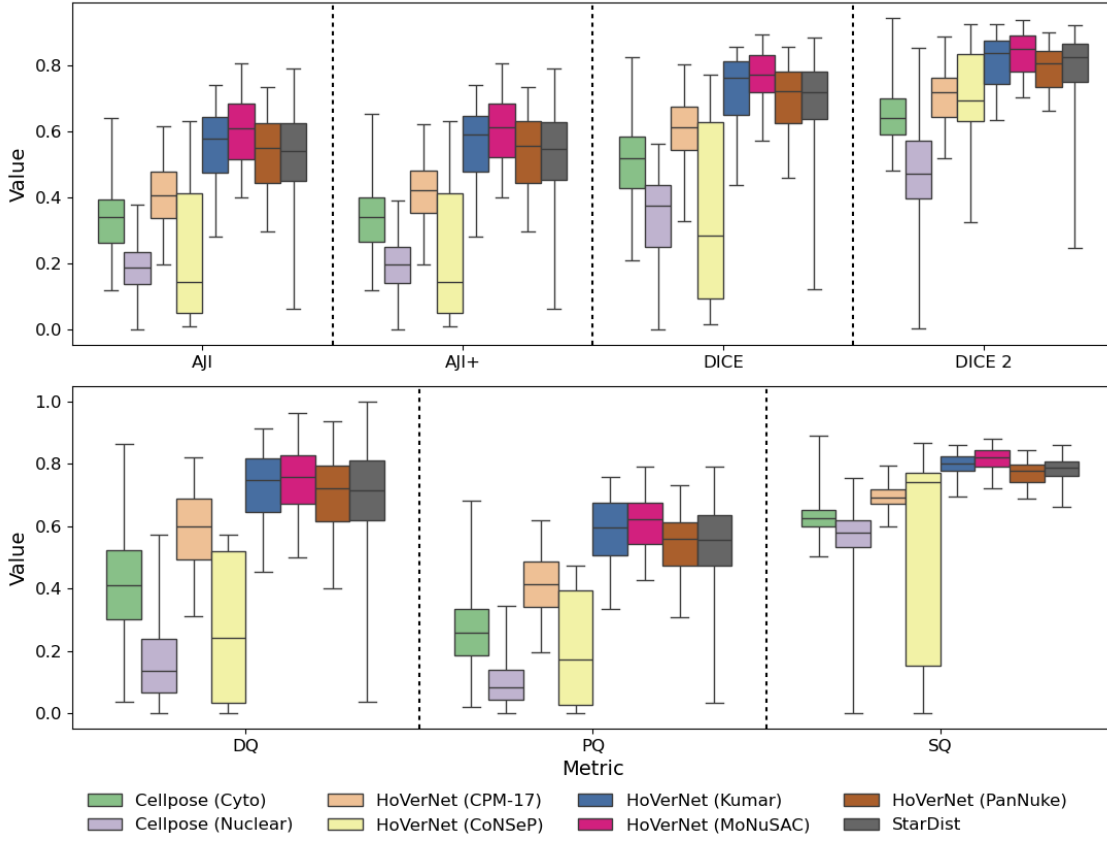


Figure 2.7: Nuclear segmentation algorithm evaluation. Higher values indicate better segmentation performance. Where multiple pretrained parameter sets were tested for a model architecture, model variants are presented in brackets.

Visual inspection evidenced poor nuclear classification as the likely cause of poor performance with the CoNSeP checkpoint of HoVerNet and CellPose (Figure 2.8). Inspection also revealed potential errors in ground truth annotations, especially evident in missed nuclei in images 1, 4 and 5. This likely led to erroneously low performance values for the CPM-17 checkpoint of HoVerNet, which appeared to segment missed nuclei well.

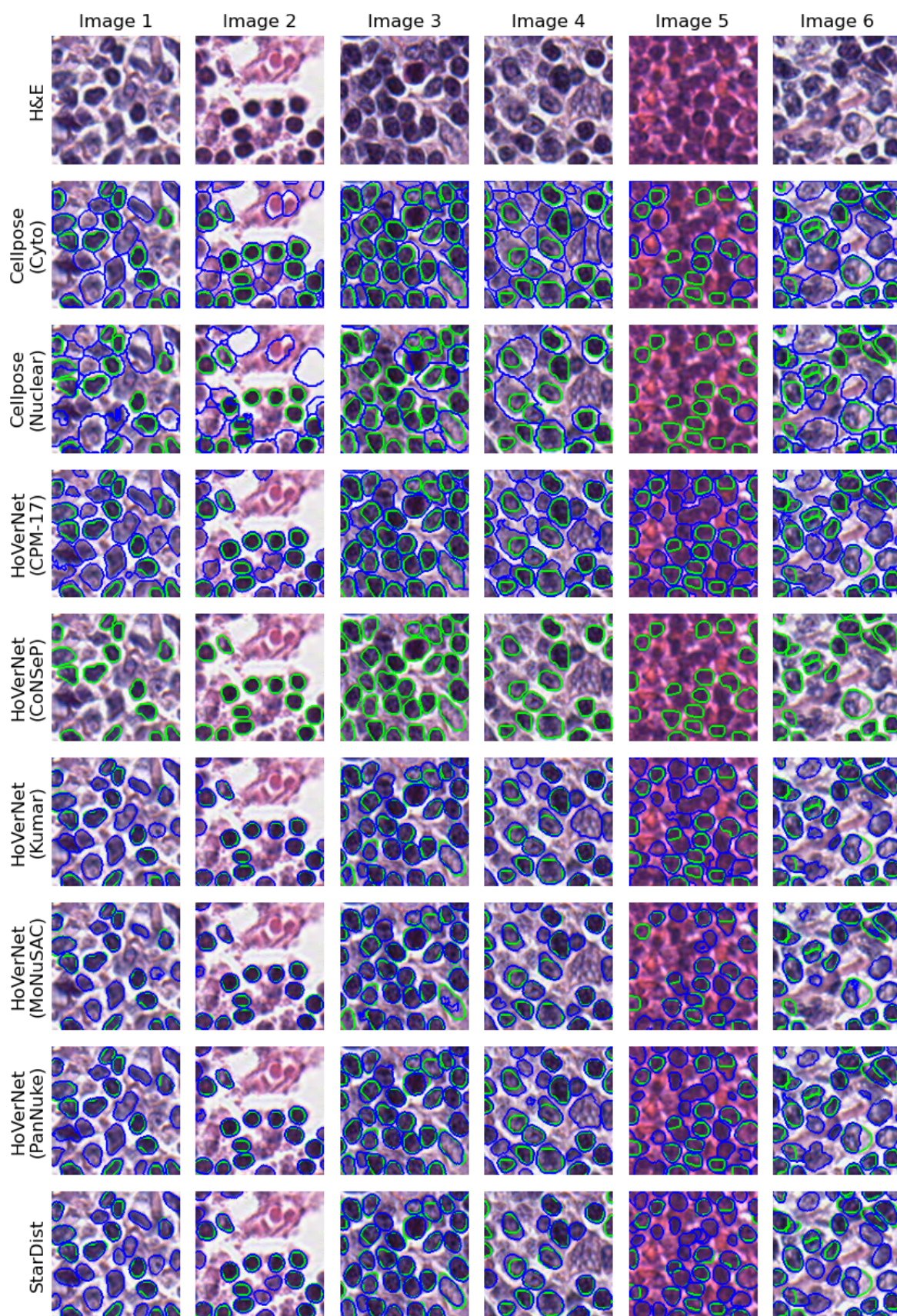


Figure 2.8: Nuclear Segmentation Algorithm Visual Inspection. Green borders indicate ground truth segmentation. Blue borders indicate predicted segmentations.

Discussion

Quantitative evaluation evidenced that the optimal algorithm for nuclear segmentation on the HMRN dataset was the MoNuSaC checkpoint of HoVerNet. In most metrics, the interquartile range of performance overlapped with the Kumar and PanNuke checkpoints of HoVerNet, and the 2D H&E checkpoint of StarDist. This evidenced suitability of those algorithms in the proposed lymphoma use case also.

Qualitative assessment of algorithms through visual inspection firstly evidenced potential mistakes in manual segmentation. This was especially visible to the centre right of image 4, where CPM-17, Kumar and PanNuke checkpoints of HoVerNet segmented an apparently large, jagged nucleus with numerous nucleoli. This object was not segmented by the human observer or the MoNuSaC checkpoint.

Secondly, visual inspection evidenced similar segmentation and nucleus detection performance of the top four models, as well as the CPM-17 checkpoint of HoVerNet.

Together, high similarity in both quantitative and qualitative performance, as well as uncertainty in annotation of a ground truth indicate that an optimal model among the 5 checkpoints referenced was challenging to distinguish.

With respect to segmentation and detection most closely resembling performance of the human observer, both qualitative and quantitative evaluation evidence that this was achieved by the MoNuSaC checkpoint.

This optimal model achieved performance values above 0.75 in DICE and segmentation quality (SQ), evidencing good performance in segmentation of pixels belonging to nuclei. Similar performance in distinction quality (DQ) evidenced good performance in segmentation of individual nuclei. Performance values were lower for AJI than DICE in all cases, evidencing a stricter metric. Similar performance in AJI and the object-wise AJI+ metric evidenced

4.3 Naive B-cell region classifier

Follicle segmentation algorithms, developed on manually annotated H&E slides at 1x magnification were evaluated on 6 exhaustively annotated H&E slides, with results dis-

played below.

Evaluation data	Model	Median Metric value (IQR)						
		DICE	DICE 2	AJI	AJI+	DQ	SQ	PQ
H&E	A	0.69	0.26	0.17	0.37	0.39	0.69	0.27
		(0.63	(0.14	(0.08	(0.27	(0.23	(0.65	(0.15
		-	-	-	-	-	-	-
		0.73)	0.52)	0.40)	0.48)	0.49)	0.71)	0.34)
	B	0.70	0.41	0.28	0.40	0.43	0.70	0.30
		(0.63	(0.19	(0.11	(0.27	(0.25	(0.65	(0.16
		-	-	-	-	-	-	-
		0.75)	0.56)	0.45)	0.51)	0.49)	0.71)	0.35)
	C	0.68	0.49	0.39	0.44	0.42	0.69	0.30
		(0.63	(0.26	(0.17	(0.30	(0.24	(0.67	(0.16
		-	-	-	-	-	-	-
		0.75)	0.58)	0.48)	0.51)	0.49)	0.71)	0.34)
	Ensemble	0.70	0.43	0.30	0.42	0.43	0.70	0.30
		(0.63	(0.20	(0.12	(0.28	(0.25	(0.68	(0.17
		-	-	-	-	-	-	-
		0.75)	0.57)	0.46)	0.51)	0.49)	0.71)	0.34)
CD10 IHC	A	0.41	0.01	0.00	0.04	0.00	0.00	0.00
		(0.38	(0.00	(0.00	(0.03	(0.00	(0.00	(0.00
		-	-	-	-	-	-	-
		0.46)	0.01)	0.01)	0.06)	0.00)	0.57)	0.00)
	B	0.41	0.01	0.00	0.04	0.00	0.56	0.00
		(0.36	(0.00	(0.00	(0.03	(0.00	(0.00	(0.00
		-	-	-	-	-	-	-
		0.46)	0.02)	0.01)	0.10)	0.00)	0.63)	0.00)

	C	0.40 (0.34 - 0.46)	0.01 (0.00 - 0.02)	0.00 (0.00 - 0.01)	0.05 (0.05 - 0.10)	0.01 (0.0 - 0.02)	0.55 (0.53 - 0.57)	0.00 (0.00 - 0.01)
	Ensemble	0.41 (0.36 - 0.46)	0.01 (0.00 - 0.02)	0.00 (0.00 - 0.01)	0.04 (0.04 - 0.10)	0.00 (0.00 - 0.01)	0.56 (0.00 - 0.59)	0.00 (0.00 - 0.00)

Table 2.2: Results of naive B-cell region (follicle) segmentation algorithm evaluation on exhaustively-annotated follicular lymphoma whole slide images. Higher median values indicate better segmentation performance. H&E = Inference on Hematoxylin & Eosin (H&E)-stained slides with evaluation against ground truth segmentations made with the same H&E slides. CD10 IHC = Inference on H&E slides with evaluation against ground truth segmentations made with registered serial CD10 immunohistochemistry (IHC) slides. IQR = Interquartile range of performance across all images. Top performing model(s, in case of ties in median and IQR values) for each metric in **bold**.

Of individual models, B and C performed best, with model B outperforming C in all metrics but DICE 2, AJI and AJI+ (Table 2.2; rows 1 - 4). Median Model A performance was substantially worse than models B and C but, due to low sample number, was within the respective IQRs of the other models (Figure 2.9).

The ensemble performed similarly to model B alone in all metric values, achieving joint best performance in the same metrics.

Segmentation was further evaluated with respect to an immunohistochemistry ground truth, the naive B-cell marker CD10 (Appendix Figure A.1). Again, model B and the ensemble achieved similar results, with joint highest DICE and Segmentation Quality (SQ), evidencing optimal performance in pixel level classification. Performance on the IHC ground truth was close to 0 for all models in all metrics except, DICE and SQ (Table 2.2; rows 5 - 8) evidencing that object-wise segmentation performance (per naive B-cell region) was poor.

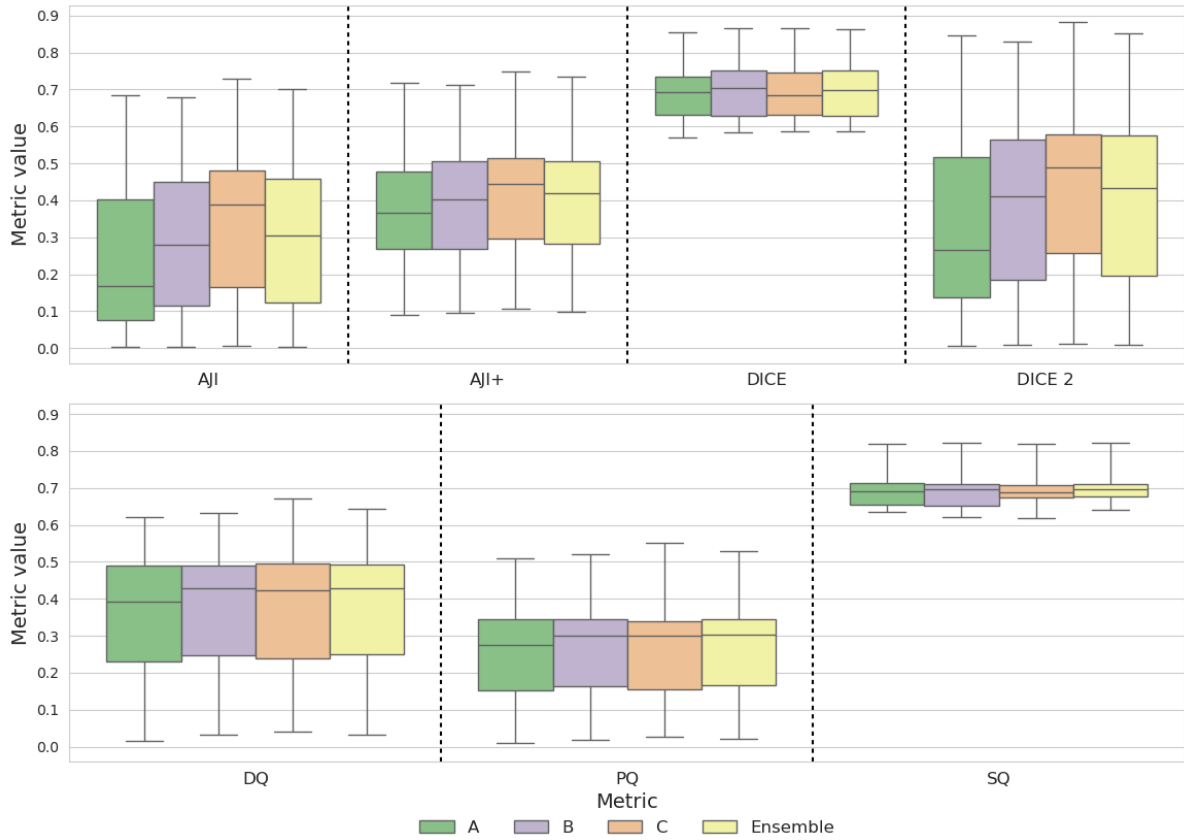


Figure 2.9: Follicle Segmentation Algorithm performance on test data against H&E ground truth. A, B, C = individual UNet models used in follicle segmentation. Ensemble = segmentation using pixel-level voting with models A, B and C, where final classification is the modal classification.

Visual inspection of follicle segmentation (blue lines indicating boundaries; Figure 2.10, below) showed segmentation similar to ground truth (green lines indicating boundaries) where follicles were clearly visible (top row; 2.10). The model appeared to segment far greater areas than in ground truth where follicles were less clearly visible (bottom row; Figure 2.10). This was also the case with IHC ground truth (bottom row; Appendix Figure A.1).

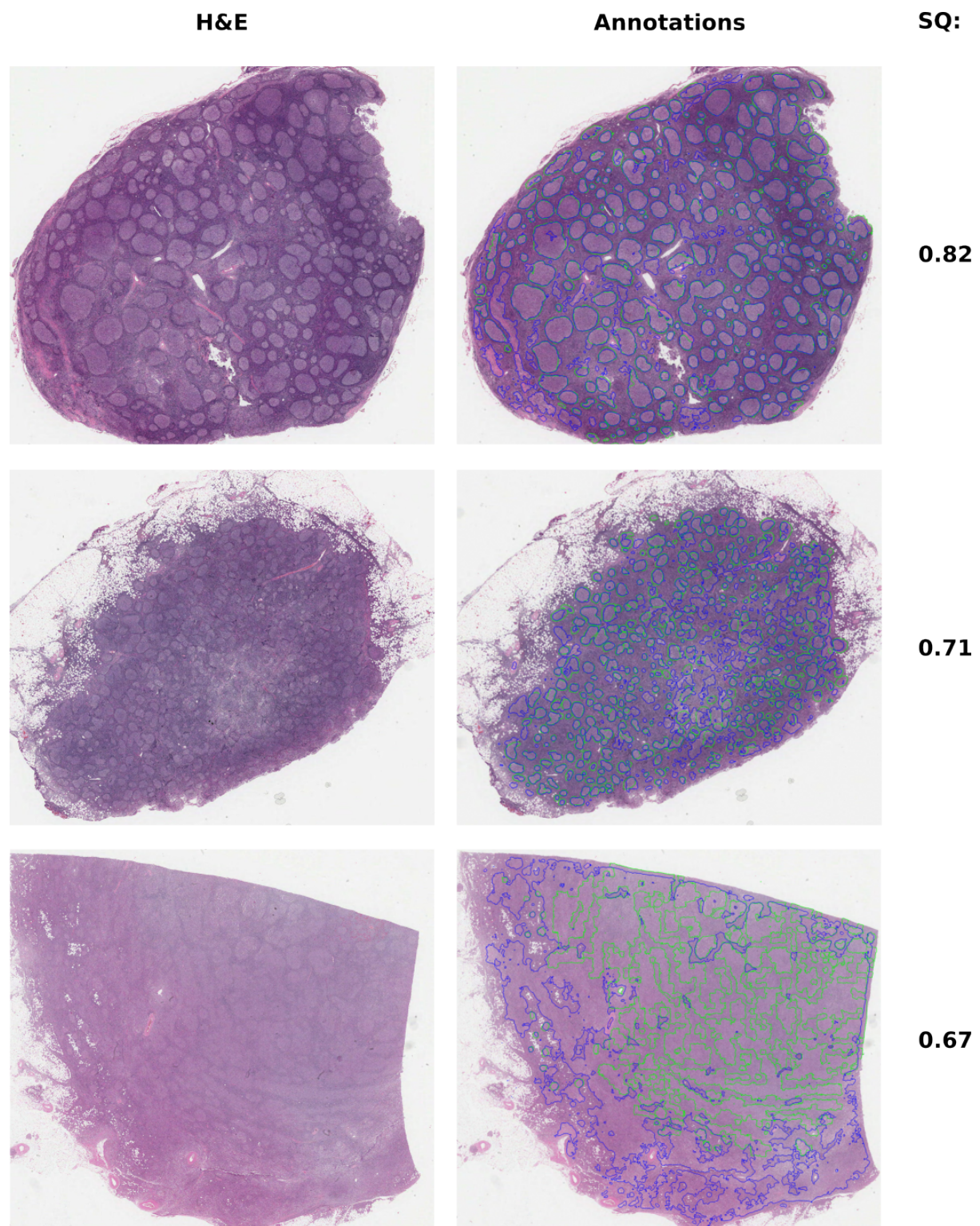


Figure 2.10: Visual inspection of ensemble Follicle Segmentation Algorithm performance on test data. SQ = Segmentation Quality metric score, per slide. Green border = Ground Truth. Blue border = Ensemble prediction.

Discussion

Quantitative evaluation of H&E results evidenced inferior performance of model A, with no apparent redeeming value in the ensemble. Models B and C appeared to perform with complementary strengths, with each achieving top performance in approximately half of the performance metrics used. These results evidence a negative impact of model A in the ensemble, while models B and C each appeared to add value to the final ensemble.

Qualitative evaluation through visual inspection evidenced performance close to a human annotator in images where follicles were clear (top images Figure 2.10). Where follicles were less clear, ensemble and human segmentations diverged, as was seen in Figure 2.10, from top to bottom. In such a low confidence segmentation task, it was not possible to determine whether the cause of lower performance was an increase in model misclassification or greater segmentation errors in the ground truth.

Efforts to determine the causes of poorer performance using an IHC ground truth (the naive B-cell marker CD10) evidenced poor distinction of individual follicles, through low performance (up to 0.10) in AJI+ and distinction quality (DQ) (Table A.1). Ensemble DICE and segmentation quality (SQ) were higher at 0.41 and 0.56, respectively, indicating moderate performance in classification of follicle pixels. Again, model A under-performed with respect to models B and C.

Qualitative evaluation through visual inspection (Figure A.2a bottom) evidenced overly high classification of follicle regions, with respect to ground truth. Comparison of follicles in H&E and CD10 ground truth staining (Figure A.2b middle) showed low intersection. Though H&E and CD10 ground truths were not compared in this study, follicles in H&E appeared to cover far larger areas than regions stained for CD10. This qualitatively evidenced poor agreement between CD10 IHC and H&E in classification of follicle regions, with ground truth annotations in H&E including large regions of false positive pixels, with respect to IHC. An improved H&E follicle classifier would likely require training of models with ground truth generated using CD10 IHC, eliminating some error introduced during annotation of follicles in IHC. Low availability of IHC limited such an effort here. With respect to existing classifiers of naive B-cell regions, performance of the H&E clas-

sifier was close to existing classifiers of germinal centres when compared with the CD10 (germinal centre marker) IHC ground truth. DICE scores of 0.41 compared well with 0.48 for HookNet, 0.30 for other UNet models and 0.38 for MRN[128]. Low DICE values across models further evidence the complexity and challenge of naive B-cell region classification. In classification of tertiary lymphoid structures, which potentially more closely resemble the H&E follicles that our model was evaluated on, our model (DICE 0.70) performed worse than HookNet, MRN and other UNet models (DICE scores: 0.85, 0.83 and 0.81, respectively).

Overall, despite low performance against CD10 IHC ground truth, the proposed naive B-cell region classifier appeared to perform favourably compared to some previous germinal centre classification efforts but less well than the multi-resolution HookNet architecture.

4.4 Centroblast Classification

As described in Methods section 3.8, centroblast classification was performed using complementary methods: deep learning methods (Methods section 3.8) and a random forest-based machine learning method aimed at simpler interpretability (Methods section 3.8).

ML approach

In the ML approach for centroblast classification, elimination of correlated features resulted in filtering of 66 features to 18. Further filtering on permutation accuracy resulted in reduction of features to 13 deemed as informative (Appendix Figure A.3a). Finally, 5-fold CV recursive feature elimination was used to determine the 5 most informative features for nuclear classification on maximal F1-score (Appendix Figure A.3b). These features were:

- Area
- Maximum intensity (red channel)
- Minimum intensity (red channel)
- Edge ratio (proportion of pixels in the first and last quartile of local binary patterns)

- Number of dark spots

Classifier evaluation

Classifiers were evaluated using multiclass (non-nucleus vs. non-centroblast vs. centroblast) and binary classification (centroblast vs. non-centroblast, with exclusion of non-nuclear objects) tasks on a cross validation partition of pathologist labelled nuclei. The top-performing classifier was further evaluated on a test partition. Evaluation metric values are presented in Table 2.3.

Task	Model	partition	AUC-ROC	Acc	Prec	Sens	F1	Spec
MC	DN169	CV	0.95	0.82	0.65	0.86	0.70	0.91
	ENV2s		0.94	0.85	0.69	0.81	0.72	0.89
	RF		0.93	0.79	0.65	0.83	0.70	0.89
BC	DN169		0.90	0.90	0.65	0.90	0.70	0.90
	ENV2s		0.86	0.92	0.66	0.86	0.71	0.86
	RF		0.90	0.94	0.70	0.90	0.76	0.90
MC	RF	Test	0.93	0.80	0.66	0.83	0.71	0.89
BC	RF		0.90	0.94	0.71	0.90	0.77	0.90

Table 2.3: Evaluation of nuclear classifier performance.

CV = Cross validation partition. MC = multiclass classification of centroblasts, non-centroblast nuclei and non-nuclei. Binary = binary classification of centroblasts vs. all other object classes. Acc. = Accuracy. Prec. = Precision. Sens. = Sensitivity. Spec. = Specificity. F1 = F1 score (macro averaged in multiclass mode). AUC-ROC = Area under the receiver-operating curve (macro averaged results used in multiclass mode, using one vs. rest classification). Final classifier performance is presented on the test partition, in terms of multiclass and binary classification of centroblasts.

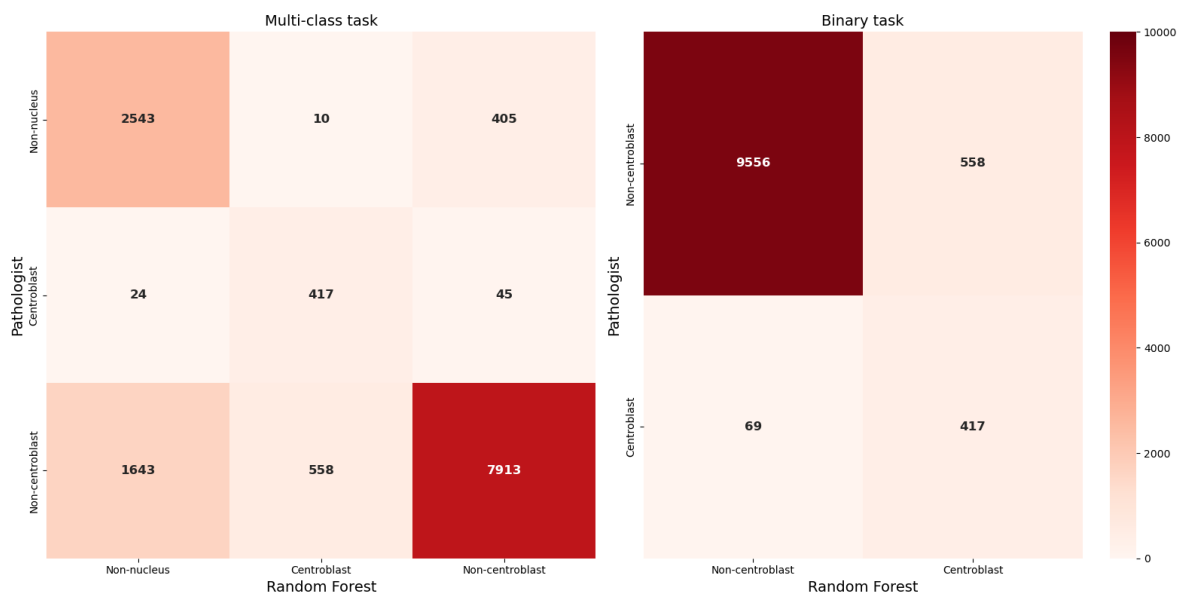
For each metric, higher values indicate better performance. Top performance for each metric in **bold**.

Deep learning (DL) classifiers marginally outperformed the random forest classifier (RF) in the multiclass task, with the largest performance disparity observed in accuracy (0.79 RF vs. 0.85 EfficientNetV2s). In binary classification, the RF achieved highest (or joint

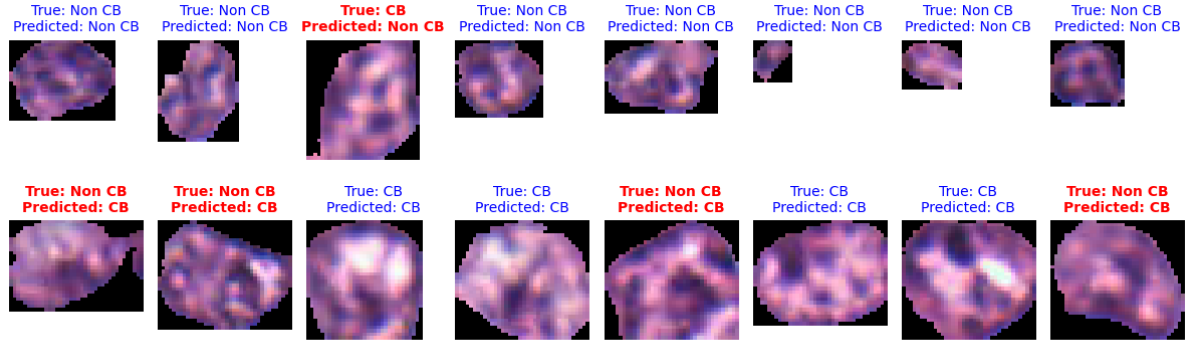
highest) performance in all metrics, with disparities of up to 0.05 with respect to next best performing approaches.

Greater interpretability of the RF approach was considered an advantage over marginally better performance offered by DL approaches in multiclass tasks so, the RF approach was selected for downstream classification tasks.

Metric values between CV and test partitions were comparable for the RF classifier, with maximum disparity of 0.01 in values. Confusion matrices for performance on the test partition (Figure 2.11a) evidenced that the vast majority of classification errors were false positive non-nuclei and centroblasts in the multiclass and binary tasks, respectively. Visual inspection (Figure 2.11b) confirmed this, with classification disparities evident in distinction of large, jagged centrocytes and larger centroblasts, with greater numbers of large nucleoli (bottom row, far left and far right columns).



(a) Confusion matrices of RF classifier performance. Left: In a multiclass task. Right: In a binary classification task.



(b) Visual inspection of binary nuclear classifications. Top row: predicted non-centroblasts. Bottom row: predicted centroblasts. Blue text: Agreement. Red text: Disagreement.

Figure 2.11: Evaluation of the random forest classifier on test partition data using confusion matrices (A) and visual inspection (B).

Discussion

Firstly, stable metric values between CV and test partitions evidenced acceptable generalisability of the RF classifier to unseen data.

Secondly, comparison of the RF classifier against published centroblast classifiers (Table 2.4, below) evidenced performance closest to Saxena et al.[129], which achieved a published F1 score of 0.82. The RF displayed noticeably worse performance than CLEM and COB[62] for which published F1 scores were between 0.95 and 0.99. Despite this, the proposed RF was evaluated on the largest dataset of centroblasts, yet, with over 13,000 involved as compared to the nearest previous efforts of 447 (CLEM and COB). The performance of the RF classifier was therefore possibly more reliable than previous efforts, using far lower numbers of cells in model evaluation.

Method	Precision (95% CI)	Sensitivity (95% CI)	F1 (95% CI)	n (CB)
Proposed RF	0.71	0.90	0.77	13,558 (486)
Saxena et al., 2023 [129]	0.78 (0.62-0.94)	0.90 (0.80-1.00)	0.82 (0.71-0.93)	58 (34)

Method	Precision (95% CI)	Sensitivity (95% CI)	F1 (95% CI)	n (CB)
Michail et al., 2014 ^[101]	0.14 (0.11-0.17)	0.83 (0.77-0.89)	0.24 (0.19-0.29)	180 (70)
COB[62]	0.99 (0.98-1.00)	0.92 (0.91-0.93)	0.95 (0.93-0.97)	447 (213)
CLEM[62]	0.99 (0.97-1.00)	0.99 (0.97-1.00)	0.99 (0.93-1.00)	447 (213)
Sertel et al., 2010[60] Initial CB detection	0.03 (0.02-0.04)	0.92 (0.87-0.97)	0.06 (0.04-0.08)	? (78)
Sertel et al., 2010[60] Final CB detection	0.17 (0.13-0.21)	0.79 (0.73-0.75)	0.27 (0.22-0.32)	? (78)
Belkacem- Boussaid et al., 2010[58]	0.81	0.87	0.84	436 (218)

Table 2.4: Published centroblast classifier performance. Classifier performance was calculated from published results in the relevant publications. All classifier results were reported with respect to binary classification tasks (centroblasts vs. non-centroblasts). Where image level evaluation data was available, 95% confidence intervals (CI) were estimated using a t-distribution samples. For COB and CLEM[62], mean metric values were calculated for sensitivity and precision; F1 was calculated from these means and F1 confidence intervals were calculated through combination of uncertainties in sensitivity and specificity. Confidence intervals for F1 were estimated Where confidence intervals are not provided, metrics were calculated over all evaluation data. n (CB) = number of nuclei that models were evaluated on, with number of centroblasts in brackets. ? = data was not found in the listed publication.

Lastly, confusion matrices (Figure 2.11a) highlighted that most binary classification errors were false positive centroblasts.

Qualitative evaluation through visual inspection showed little distinction between misclassified and correctly classified nuclei in some cases (Figure 2.11b). For example, the

misclassified nucleus on the top row has a similar size and distribution of small nucleoli as the other larger, correctly classified nuclei on the row. Likewise, there are similarities in size, nuclear folds and presence of a single, large nucleolus between the misclassified left-most nucleus on the bottom row and the correctly classified nucleus fourth from the left on the same row. These results further highlight the subjective nature of centroblast classification.

An important factor in evaluation was that the labelled centroblast dataset, used in training, cross validation and evaluation, was created using annotations from one pathologist. Performance reported was likely an overestimation of performance with respect to other pathologists. Though not feasible here, a pathologist consensus as performed by other investigators[58, 62] would be required to achieve gold standard ground truth. Evaluation against such a dataset would be required to determine performance against a general population of pathologists.

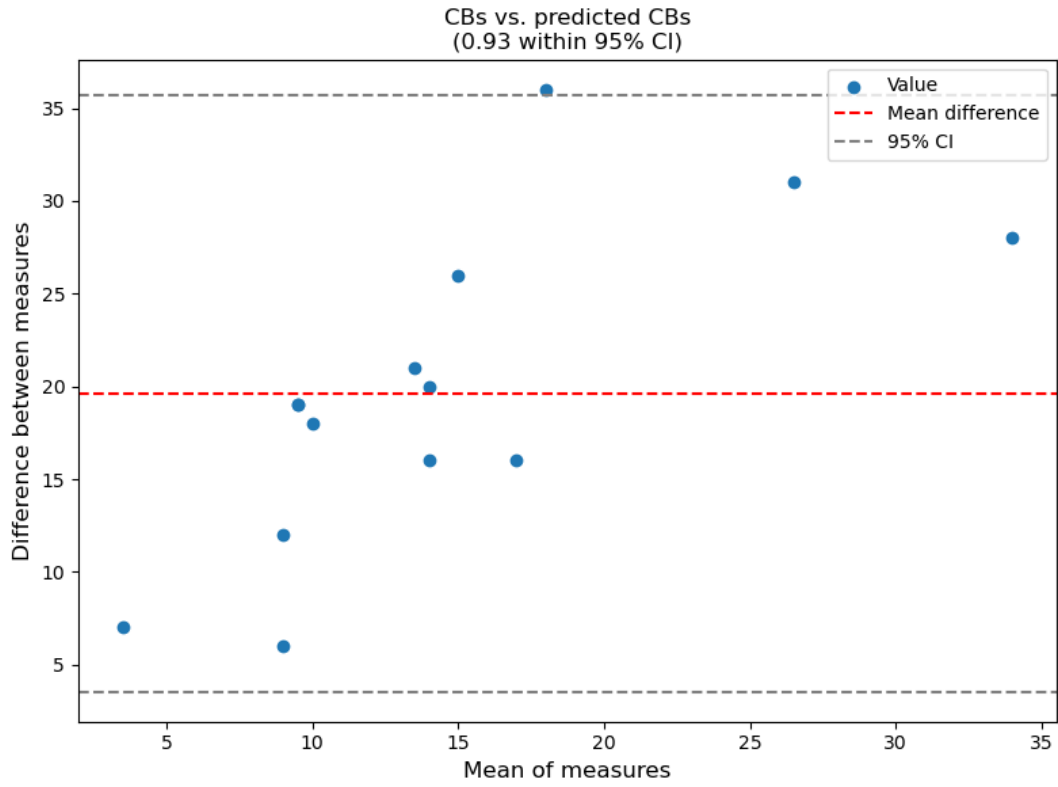
4.5 Centroblast counts evaluation

To validate performance of the classifier in quantifying centroblasts, pathologist counts of cells potentially resembling centroblasts were compared to predicted counts of centroblasts made by the model (Methods section 3.10), reported in Table 2.5. An additional classification of ‘large cells’ was also used, comprising all cells that could plausibly be classified as centroblasts in nuclear morphology alone, not taking into account surrounding cells or cytoplasm. This was used to further evaluate model performance specifically on centroblasts, against centroblast-like nuclei. Maximum metric values were used to evaluate model classifications, with respect to potential misclassification of visually-similar cell types. For evaluation of correlation in cell counts, Pearson and Spearman correlations were used. For evaluation of agreement between the RF and pathologist, the R^2 metric and Bland-Altman plots were used (Table 2.5 and Figure 2.12, respectively; as described in Section 4.1). One region of interest (ROI) in the HMRN dataset ‘calibration’ partition was excluded due to anomalously high number of centroblasts, which a pathologist annotator deemed too high to reliably count.

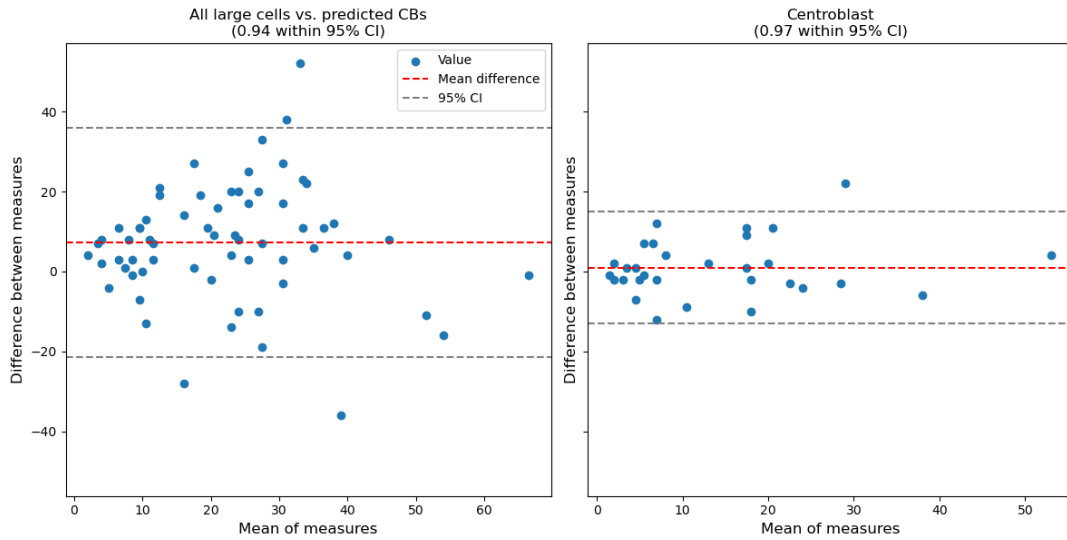
Count task	Data	Partition	n	Pearson r	Spearman r	R2
Centroblast	HMRN	Cal.	14	0.68	0.46	-2.84
	GALLIUM	Pass	30	0.83	0.72	0.68
Large cell	GALLIUM		64	0.53	0.55	-0.17
Immunoblast	GALLIUM		30	0.16	0.18	-21.17
Endothelial	GALLIUM		30	-0.41	-0.39	-25.62
Macrophage	GALLIUM		30	-0.30	-0.28	-42.27
fDC	GALLIUM		30	-0.09	0.05	-148.47

Table 2.5: Evaluation of pathologist cell counts against predicted centroblast number. The count task column relates to the cell types counted by a pathologist. Correlation (r) and agreement (R2) measures indicate agreement between cell counts for each pathologist category and number of centroblasts counted by the machine.

GALLIUM trial partitions: Pass = Strict QC pass slides. HMRN partitions: Cal. = Calibration of grading thresholds. fDC = Follicular Dendritic Cells. For each metric, higher values indicate better performance. Top performance for each metric in GALLIUM external validation data in **bold**. n = number of patches (1000 x 1000px) used in evaluation.



(a) HMRN centroblast count agreement



(b) GALLIUM centroblast count agreement

Figure 2.12: Evaluation of agreement in centroblast counts against pathologist counts of:

- a) centroblasts in the ‘calibration’ partition of the HMRN dataset
- b) large nuclei which could be classified as centroblasts (left) and high certainty centroblasts (right) in the ‘strict pass’ partition of the GALLIUM dataset.

Correlation between predicted centroblast and pathologist cell counts was highest between centroblasts in the GALLIUM dataset, with r of 0.83 and 0.72 for Pearson and Spearman correlation, respectively (Table 2.5). Correlation between centroblast counts in the smaller HMRN set of images was considerably lower (r values of 0.68 and 0.46 for Pearson and Spearman correlation, respectively). Moderate correlation was observed with large cell counts (r values of 0.68 and 0.46 for Pearson and Spearman correlation, respectively) while correlation with all other cell types was either closer to zero or negative (Table 2.5).

Agreement, as measured with the R^2 metric, was negative in all cases except centroblasts in the GALLIUM dataset, where it was 0.68 (Table 2.5).

Bland-Altman plots displayed the difference between cell counts on the Y-axis, calculated as number of predicted centroblasts subtracted from pathologist counts. In the HMRN dataset (Figure 2.12a), the mean difference between cell counts was 19.6. This indicated that the RF consistently predicted around 20 centroblasts more per region of interest than the pathologist. The plot also appeared to show a positive gradient in datapoints, indicating a greater scale of discrepancy with increased number of centroblasts present. 93% of datapoints were within a 95% confidence interval (95% CI) of the mean difference.

These effects were not observed in the GALLIUM dataset (Figure 2.12b right), where mean difference between centroblast counts was 1.0, and no visible gradient existed. 97% of datapoints were within a 95% CI of the mean difference.

With respect to large cell counts, there was a mean difference of 7.2 cells, with 94% of datapoints within a 95% CI and no visible gradient. All other cell types measured had negative mean differences between -9.7 and -11.8, clear negative gradients and 93% of datapoints within a 95% CI (data not displayed).

Discussion

Moderate to high correlation was observed between predicted and pathologist centroblast counts in the HMRN and GALLIUM datasets, respectively. For the GALLIUM data especially, these correlation values were at least 0.30 and 0.18 higher than other cell types

(general large cells) for Pearson and Spearman r , respectively. In addition, moderately high R^2 value and mean difference between pathologist and machine centroblast counts close to zero in the Bland-Altman plot for GALLIUM data indicate moderately strong agreement.

Correlation and agreement values were visibly lower in HMRN data than in GALLIUM. Lozanski et al[61] reported systematic differences in methodology as key sources of error in centroblast count tasks, leading to failure in reaching grading consensus in 6-59% of cases. The methodology presented here attempted to follow the more reliable approach reported by Lozanski et al. but failed to reach acceptable agreement with a pathologist in HMRN data. It was subsequently found that a pathologist annotator had used a quadrat in GALLIUM annotation but not in HMRN data, potentially accounting for the stark differences in agreement. Though long used in ecology and haemocytometry, use of a quadrat is not standard in the developing field of digital pathology. Nonetheless, far higher agreement between computational tools and pathologist cell counts evidence it as a simple but effective aid in increasing reliability.

4.6 Cytological Grading

A ‘calibration’ set of 15 pathologist-graded ROIs was used to set centroblast thresholds for cytological grades:

1. < 1.5 centroblasts
2. $1.5 \leq \text{centroblasts} < 7$
3. ≥ 7 centroblasts

Median centroblast number per ROI was used to aggregate cell counts over all ROIs of a WSI. The above thresholds were used to assign a grade to each WSI. These thresholds were evaluated in regression tasks (Table 2.6), predicting numeric grade and a classification task, predicting high grade (grade 3; Table 2.7).

Two sets of nuclear feature scaling parameters were used: those determined on labelled nuclei in the HMRN dataset (termed HMRN scaling) and those determined on all nuclei extracted in the GALLIUM dataset. This was conducted to assess generalisability of nuclear feature scaling to external centre conditions (such as slide acquisition and scanning procedures).

Data	Scaling	Partition	n	Pearson r	Spearman r	R2
H	H	Cal.	15	0.78	0.77	0.52
		Val.	5	0.07	0.03	-1.14
G	H	A	523	0.10	0.04	-32.61
		L	383	0.13	0.08	-41.93
		P	271	0.13	0.12	-52.37
	G	A	523	0.04	0.00	-3212.32
		L	383	0.06	0.03	-4195.64
		P	271	0.11	0.09	-4779.78

Table 2.6: Evaluation of cytological grade regression, with respect to pathologist estimates of cytological grade (1-3). In HMRN data, estimates of cytological grade made on 1000×1000 pixel regions of interest were compared. In GALLIUM trial data, estimates made on whole slide images were compared.

Datasets: H = HMRN data; G = GALLIUM data. Scaling referred to parameters used to linearly scale nuclear feature values, either limits from HMRN (H) or GALLIUM (G) datasets. HMRN partitions: Cal. = Calibration of grading thresholds; Val. = validation data. GALLIUM partitions: A = All slides with sufficient tissue; L = Slides in ‘A’ additionally filtered for 40× magnification and passing initial quality control, including those with potential concerns flagged; P = Slides in ‘L’ passing visual inspection (with no concerns flagged). For each metric, higher values indicated better performance. Top performance for each metric in **bold**.

Task	Data-set	Scaling	Part-ition	n	Acc.	Prec.	Sens.	F1	Spec.	AUC-ROC	κ
Slide grading	G	H	A	523	0.78	0.51	0.50	0.49	0.50	0.50	0.01
			L	383	0.77	0.51	0.51	0.50	0.51	0.51	0.02
			P	271	0.75	0.51	0.50	0.50	0.50	0.50	0.01
		G	A	523	0.47	0.50	0.50	0.43	0.50	0.50	0.00
			L	383	0.38	0.50	0.51	0.37	0.51	0.51	0.01
			P	271	0.35	0.53	0.53	0.35	0.53	0.53	0.03
ROI grading	H	H	Cal.	15	0.93	0.88	0.96	0.91	0.96	0.96	0.81
			Val.	5	0.80	0.75	0.88	0.76	0.88	0.88	0.55

Table 2.7: Evaluation of high cytological grade classification.

Metric values were calculated in a binary high cytological grade (3A/3B) vs. low cytological grade (1/2) configuration.

‘Slide grading’ referred to comparison of cytological grades (1-3) assigned to whole slide images. ‘ROI grading’ referred to comparison of high grade classifications of 1000×1000 pixel regions of interest.

Dataset abbreviations: G = GALLIUM; H = HMRN. GALLIUM trial partitions: All = All slides with sufficient tissue; Len. = Lenient QC pass slides; Pass = Strict QC pass slides. HMRN partitions: Cal. = Calibration of grading thresholds; Val. = validation data. Metric abbreviations: Acc. = Accuracy; Prec. = Precision; Sens. = Sensitivity; Spec. = Specificity; F1 = F1 score; AUC-ROC = Area under the receiver-operating curve; κ = Cohen’s kappa.

For each metric, higher values indicate better performance. Top performance for each metric in **bold**.

Performance was either similar between GALLIUM- and HMRN-scaled data (precision, sensitivity, specificity, AUC-ROC and κ in classification tasks) or noticeably worse (accuracy and F1 in classification, as well as all regression metrics; (Table 2.6); rows 3-6 compared with rows 7-9). Scaling to the GALLIUM data was ineffective in improving performance of models developed on a different scanner manufacturer.

Cytological grading regression did not evidence association or agreement with pathologist ground truth in any validation partitions (disregarding HMRN calibration partition), with values of Pearson and Spearman r below 0.13 and highly negative R^2 in all cases.

In classification, grading of patches (regions of interest; ROIs) resulted in good performance, with AUC-ROC score of 0.88 (bottom row; Table 2.7). Performance on whole slide images (slide grading) was far worse, with AUC-ROC values close to 0.5, evidencing

prediction no better than chance regardless of nuclear feature scaling strategy used.

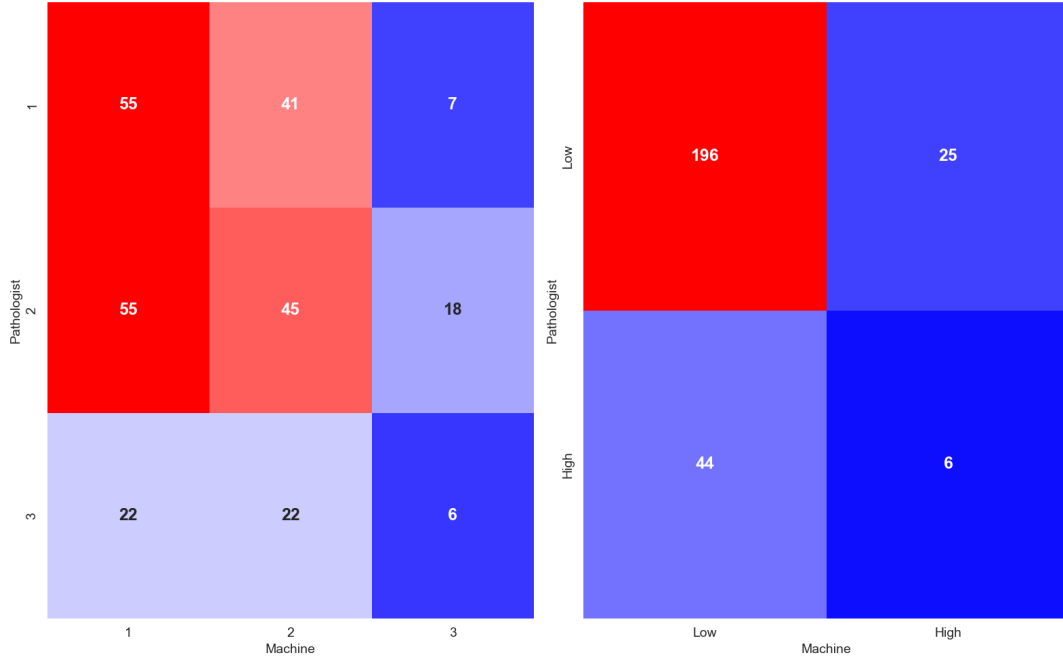


Figure 2.13: Evaluation of automated grading on the ‘strict pass’ QC partition of GALLIUM trial data in:
left) Classification of individual grade
right) Classification of high grade (3)

Confusion matrices (Figure 2.13) identified that major causes of disagreement were in classification of grade 3, with only 6 of 50 pathologist-assigned cases correctly classified.

Discussion

Classification of high grade (grade 3) WSIs evidenced performance close to random choice, with values of precision, sensitivity, F1 and AUC-ROC close to 0.50. Cohen’s κ values were close to 0, indicating no agreement. Despite the low agreement, inter-pathologist agreement was previously reported at similar levels (median κ of 0) in digital pathology assessment of WSIs[61].

Finally, ROI grading on HMRN validation data yielded the highest values for all evaluation metrics, with AUC-ROC close to that reported for the FLAGS system[63] (0.88 ours vs. 0.87 FLAGS), and κ agreement marginally lower than inter-pathologist agreement in ideal technical conditions (0.55 ours vs. 0.65 inter-pathologist)[61]. Notably, sample

number was very low in this task at just 5 ROIs. Additional evidence of good model performance is significantly higher prevalence of patients diagnosed with large cell Follicular Lymphoma (archaic nomenclature for high grade FL[89]) among those predicted to be high grade in HMRN data (Chapter 5.3 Figure A.4a; Pearson $r = 0.40$).

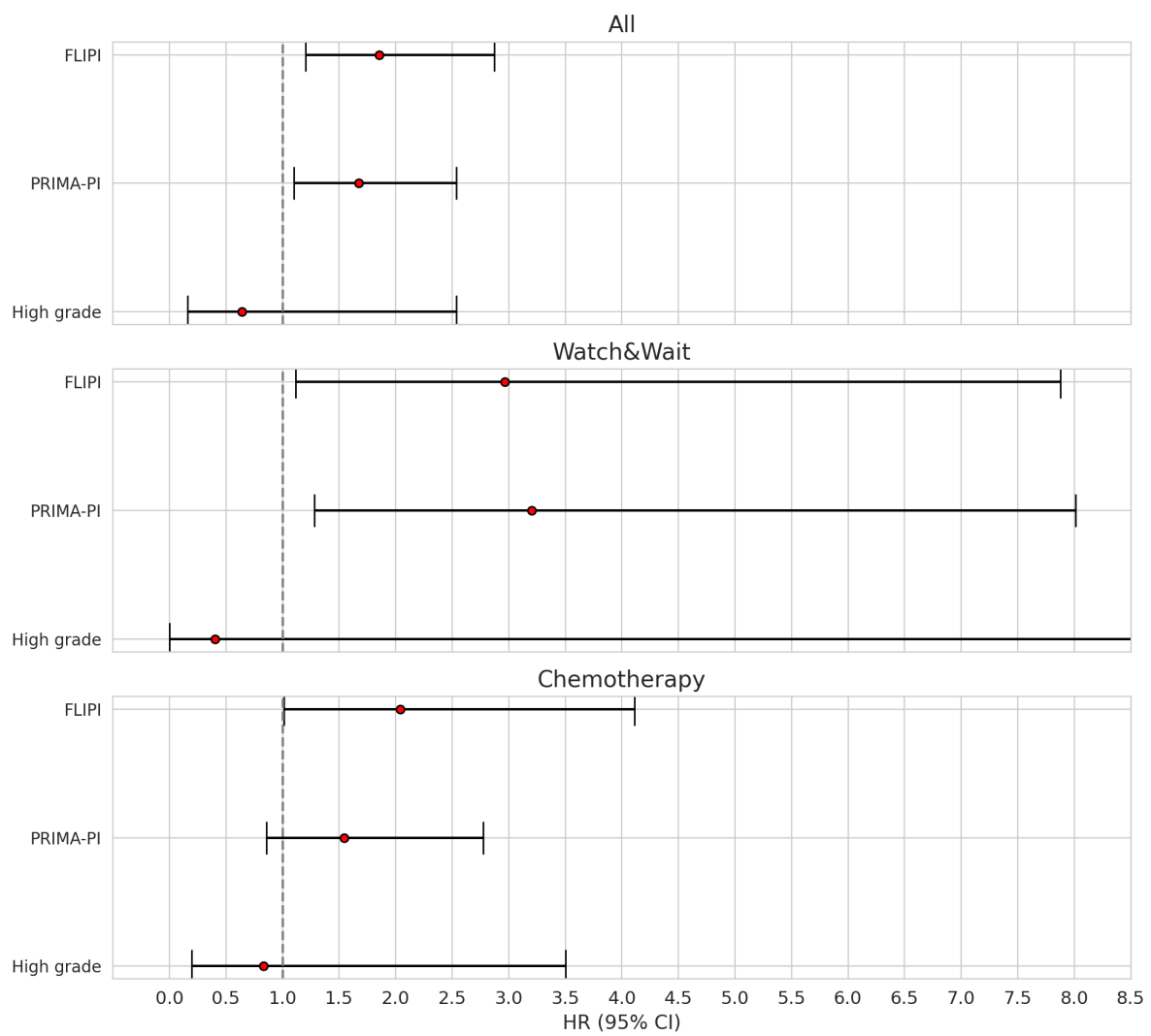
In summary, the classifier performed poorly in regression grading and classification of high grade. Despite this, levels of agreement between the machine and pathologist were close to reported inter-pathologist agreement levels, under less stringent methodologies. Where the classifier demonstrated success in grading - ROI level grading, with fixed field of view and magnification - sample number was too low for comparisons with suitable statistical power. Significant enrichment of large cell FL disease in patients predicted to be high grade provides further evidence of acceptable performance in classification of high grade disease. Further validation efforts under stringent pathologist grading conditions would be required for conclusive evaluation.

4.7 Survival Analysis

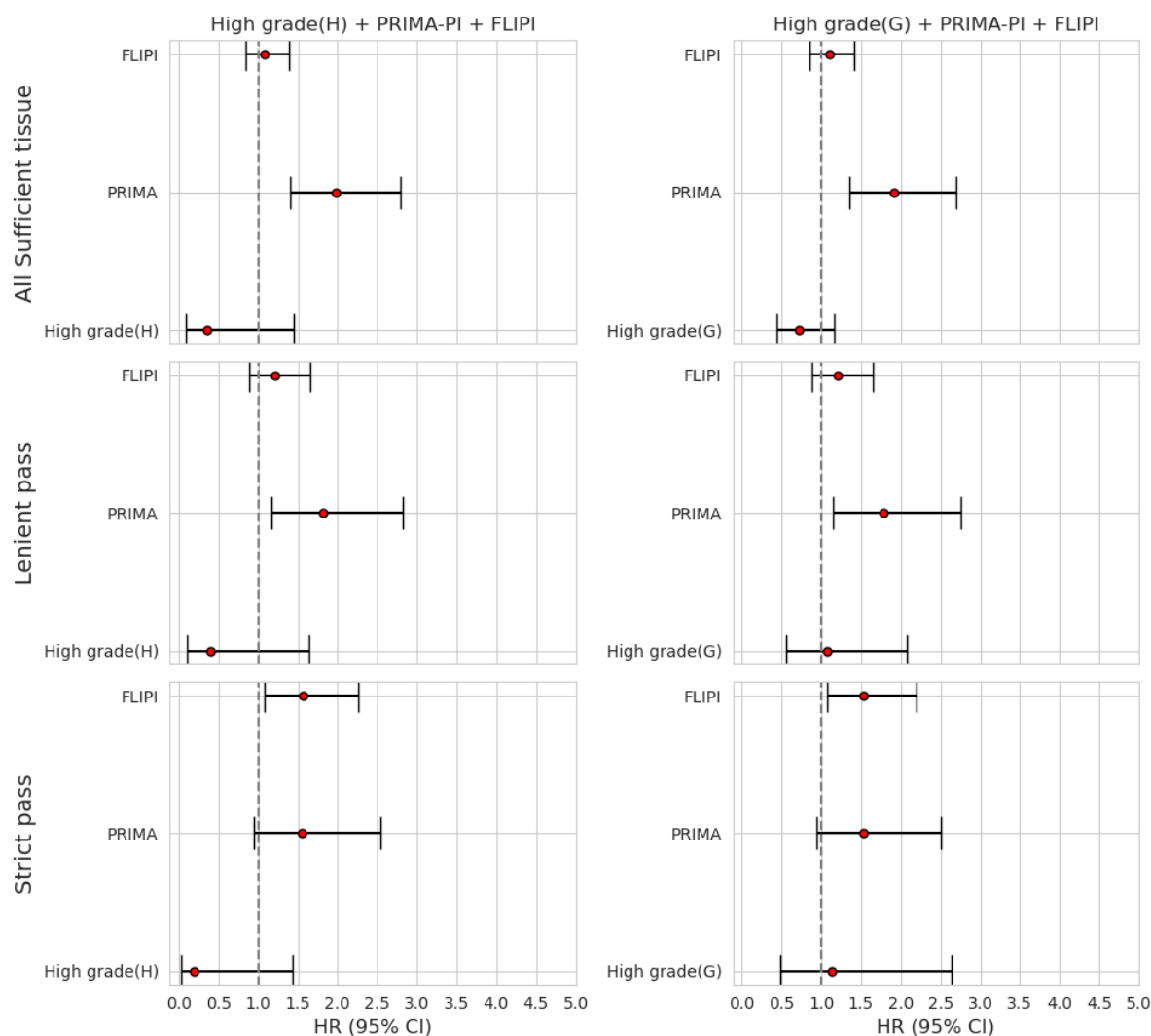
Hazard ratio findings are summarised below in Figure 2.14. Complete Cox model metrics for HMRN and GALLIUM data are provided in Supplementary documents, Cox_model_results.xlsx. In HMRN radiotherapy and watch & wait partitions, few events in high grade partitions led to extremely large 95% confidence intervals.

In HMRN partitions, High grade (predicted) was found to associate with non-significant lower hazard (between 0.4 and 0.9; Figure 2.14a) compared to low grade, adjusted using FLIPI and PRIMA-PI. In GALLIUM partitions using nuclear features scaled to HMRN parameters, this effect was also observed, with hazard between 0.2 and 0.4 (left column; Figure 2.14b). This effect was not observed when features were scaled to GALLIUM dataset parameters (right column). This effect appeared consistent between datasets though hazards were not significant for cytological grade in any partition.

FLIPI quantiles and PRIMA-PI associated with significant increased hazard in HMRN (all treatments) data and in some groups of GALLIUM cases.



(a) HMRN dataset



(b) GALLIUM dataset

Figure 2.14: Cox proportional hazards model hazard ratios for cytological grade, with respect to FLIPI and PRIMA-PI indices. Multivariate models were fitted to overall survival events, defined as death by any cause and the associated time to event. Where 95% confidence interval (95% CI) bars are not shown, they were calculated to be between 0 and infinity.

a) Fitted to treatment partitions of the HMRN dataset.

b) Fitted to quality control partitions of the GALLIUM dataset.

Kaplan-Meier estimators were fitted to overall survival (OS) events in HMRN treatment partitions and the 'Strict Pass' partition of GALLIUM data for pathologist-labelled and predicted high grade, displayed in Figure 2.15. Kaplan-Meier estimators were also fitted to progression-free survival events in HMRN data (Appendix Figure A.5).

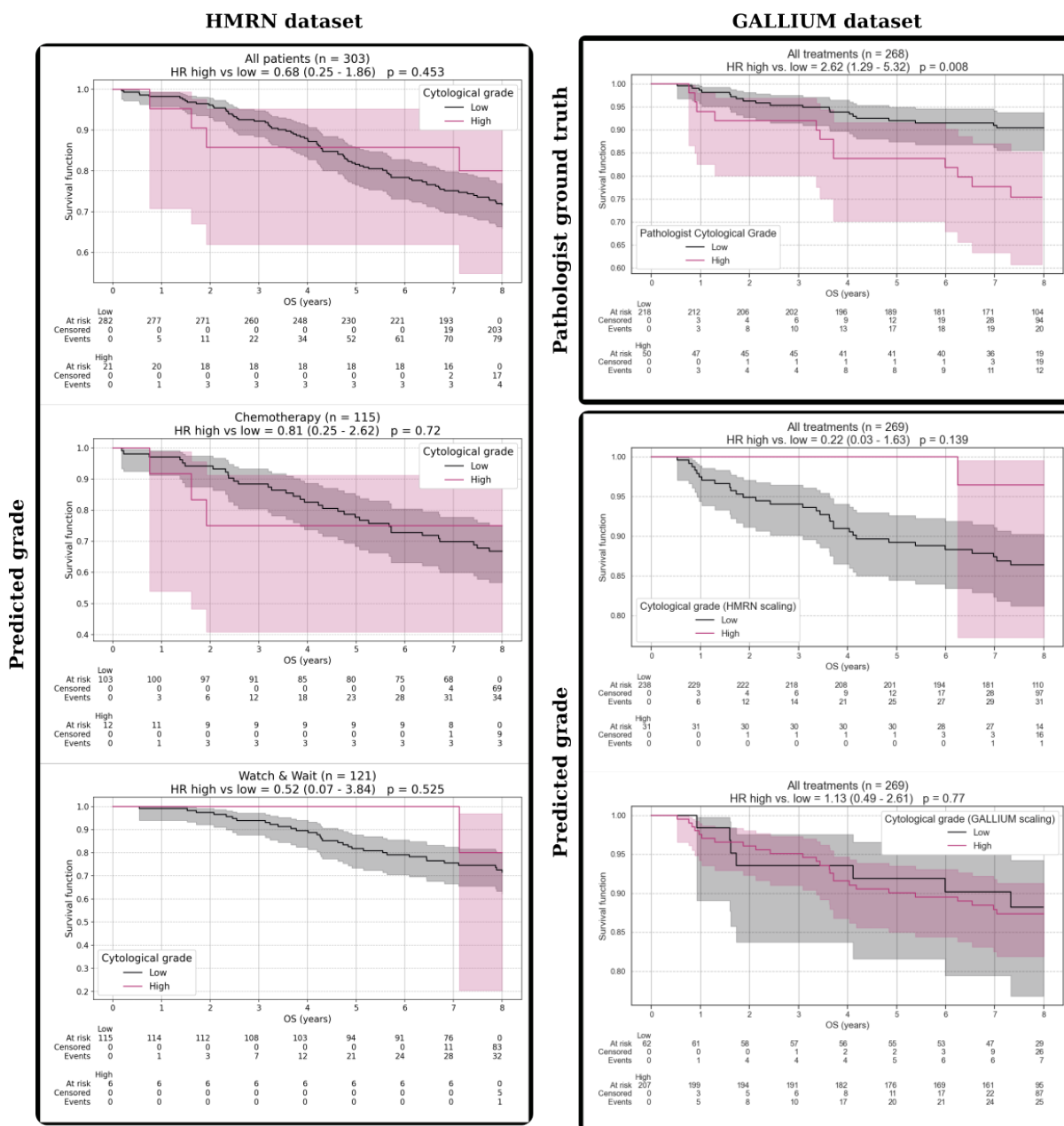


Figure 2.15: Overall survival events defined as death by any cause. High cytological grade defined as grade 3A/3B. Number of events and censoring events indicated below each Kaplan-Meier (KM) plot. Method used for calculation of cytological grade indicated in text to the left of boxes. Dataset indicated above columns of plots.

Pathologist-assigned high cytological grade (top right; Figure 2.15) associated with high hazard (HR vs. low grade: 2.62). No significant differences were observed between groups for predicted cytological grade (predicted grade cells; Figure 2.15).

4.8 Overall discussion

Comparison of correlations between predicted high grade and clinical features did not identify any effects present in both HMRN and GALLIUM datasets (correlation matrices presented in Appendix, Figure A.4). Both significant relationships (following Bonferroni correction) in high grade patients - moderately higher proportion of patients reported to have ‘large cell FL’ and slightly lower cell counts - could be explained by confounding variables. Namely, that enrichment of ‘large cell FL’, meeting requirements of ‘Follicular Large B-cell Lymphoma’ under more recent WHO recommendations and including cytological grade 3B cases[23] would be expected when considering all cases predicted to be grade 3. Secondly, the relationship between high grade disease and lower cell densities could likely be explained by a lower number of large cells, centroblasts, packing to fill the defined tissue space (ROI).

Pathologist-annotated high grade patients in the GALLIUM dataset experienced significantly greater overall survival (OS) hazard than low grade, with hazard ratio of 2.62 (95% CI: 1.29-5.32; $p = 0.008$; Figure 2.15 top right). The observed effect bore similarities to that observed in previous immunochemotherapy trials[92]. These results evidenced increased risk associated with high cytological grade in immunochemotherapy patients. By extension, a performant classifier of high grade, performing at least as well as a pathologist, would be expected to stratify a higher risk group of patients in the GALLIUM dataset.

Analyses of OS with predicted high grade evidenced slightly reduced risk, albeit, not significantly. This was observed in all treatment partitions of HMRN data (Figure 2.14A) and GALLIUM data linearly scaled according to HMRN parameters (Figure 2.14B, left). Kaplan-Meier plots showed that all but 1 event occurred within two years (3 for 21 patients, total). No significant differences were observed for PFS but event number was low leading to underpowered analyses (Appendix Figure A.5).

In GALLIUM data linearly scaled according to feature ranges of GALLIUM nuclear data, HRs for high grade were close to 1 in data filtered on slide quality (1.08 in ‘Lenient Pass’

and 1.13 in ‘Strict pass’; Figure 2.14B, right). In data filtered on availability of tissue alone (‘All sufficient tissue’), HR was slightly below 1 (0.71). Kaplan-Meier showed that, where HMRN scaling parameters were used, only one event was observed in the study period (out of 31 patients).

These results evidenced considerable differences in patient subgroups classified by the automated route and pathologists, where high grade was associated with substantial hazard. The small number of patients classified as high grade by the classifier appeared to consistently experience lower long-term hazard, evidencing a lower risk subgroup.

Potential explanations of lower risk with higher concentrations of cells predicted to be centroblasts may be in enrichment of ‘classical FL’ cases, lacking uncommon morphological features[23, 24]. Comparisons of classifier and pathologist centroblast counts (Figure 2.12) showed good agreement, evidencing that misclassification of dendritic cells, macrophages or other immune cells potentially involved in effective immune responses against neoplastic cells was unlikely. Another potential explanation would be classification of patients with higher proportions of non-neoplastic centroblasts, potentially responding to neoplastic cells. This could also explain disparities in machine/pathologist cytological grading at a slide level, where pathologists may ignore centroblasts not displaying aberrant features or ‘reactive’ germinal centres, mounting responses to lymphoma cells. This further highlighted challenges in region selection for grading, given that canonical B-cell responses to cancer involve formation of germinal centres. Simultaneous presence of neoplastic and reactive germinal centres may have led to predictions that differed to pathologists, since the naive B-cell region classifier did not distinguish these entities.

Training a classifier to distinguish reactive germinal centres (GCs) and neoplastic follicles may be challenging using H&E alone, with loss of so-called ‘light’ and ‘dark’ zone polarisation characterising the ‘follicle’[23]. Entropy measures of cell type disorder or graph representations of ROIs may effectively capture the loss of heterogeneity expected in transition of GCs to neoplastic follicles. Otherwise, presence of the anti-apoptotic BCL2 protein is a classic feature of neoplastic follicles, while absent in GCs[130]. CD10

is a classical GC marker commonly produced by neoplastic (except grade 3B) and reactive centroblasts. It may therefore be feasible to train a classifier of naive B-cell regions using registered CD10 immunohistochemistry (IHC) and further distinguish regions into reactive GCs and follicles using registered BCL2 IHC.

Differences observed between predicted cytological grade groups were not significant, but this is a common problem in analyses of FL. The indolent nature of most FL cases and older age of incidence present substantial challenges in survival analyses. Firstly, low event number (deaths, in the case of OS) mean that longer periods of observation are required to gather sufficient data for estimations of risk, though such lengths of observation are rarely feasible. Progression-free survival (PFS), time between diagnosis and progression of disease (firstline treatment failure) is often used to achieve greater statistical power as treatment failure events are more common than death. Major clinical trials have frequently observed significant differences in PFS hazard but not in OS[107, 131, 132]. Secondly, older patient age, evidenced by median age at diagnosis close to 65 in the UK[46] complicates analyses through increased risk of co-morbidity and death by other causes. For significant OS differences to be observed between groups, effect size must be substantially higher than in younger age groups due to the increased ‘noise’ of death by other causes. Again, significant differences in PFS between clinical trial groups (without corresponding differences in OS) evidence lower impact of confounding factors associated with age on disease progression.

In summary, pathologist classification of high cytological grade was associated with substantial, significant increased hazard against low grade ($HR = 2.62$). On the other hand, classification of high grade by the machine classifier was associated with low hazard (HRs of 0.22 - 0.81 between datasets and partitions), albeit not significantly. Differences between patients classified as high grade by pathologists and the machine may lie in distinction of neoplastic and reactive (inflammatory) naive B-cell regions by pathologists. As such, the machine may preferentially classify a subgroup of patients mounting effective B-cell responses to lymphoma cells as high grade, potentially accounting for survival differences.

5 Conclusions and Future Work

5.1 Cytological grade

Cytological grading involves quantification of naive B cells, centroblasts, in lymphoma-affected lymph nodes and classification of patients on concentration thresholds. The process presents a considerable resource burden for pathologists, with high subjectivity in cell classification limiting reproducibility. With the aim of reducing pathologist burden and subjectivity in this task, we developed an automated cytological grading tool operating on objective nuclear features in H&E slides.

Summary of key results

The cell classifier performed favourably with respect to published classifiers of centroblasts. Centroblast counts agreed well with a pathologist aided by a quadrat (Figure 2.16A), with an average difference in counts of 1. Performance was worse when a quadrat was not used by the pathologist, further evidencing the high levels of subjectivity in the task.

A deep learning model trained to classify naive B-cell regions (follicles), the sites assessed in cytological grading, performed well with respect to ground truth H&E but poorly against immunohistochemistry ground truth (the naive B-cell marker CD10), with preferential classifications of false positive compared to IHC. Overall grades did not agree with pathologists but were comparable to inter-pathologist agreement under certain conditions.

Finally, while pathologist estimates of high cytological in GALLIUM trial data were associated with significantly increased risk (Figure 2.16B), prediction of high grade by the classifier were associated with small but reduced hazard (Figure 2.16C & D for HMRN and GALLIUM datasets, respectively), consistent across datasets. Few OS events were observed among patients classified as high grade, evidencing unintentional classification of a lower risk subgroup of FL patients.

Taken together, these results demonstrated that the automated route performed well in centroblast quantification, with respect to pathologists, but evidenced major differences in how individual readings were aggregated for a slide, with potential exclusion of reactive germinal centres by pathologists but not the machine.

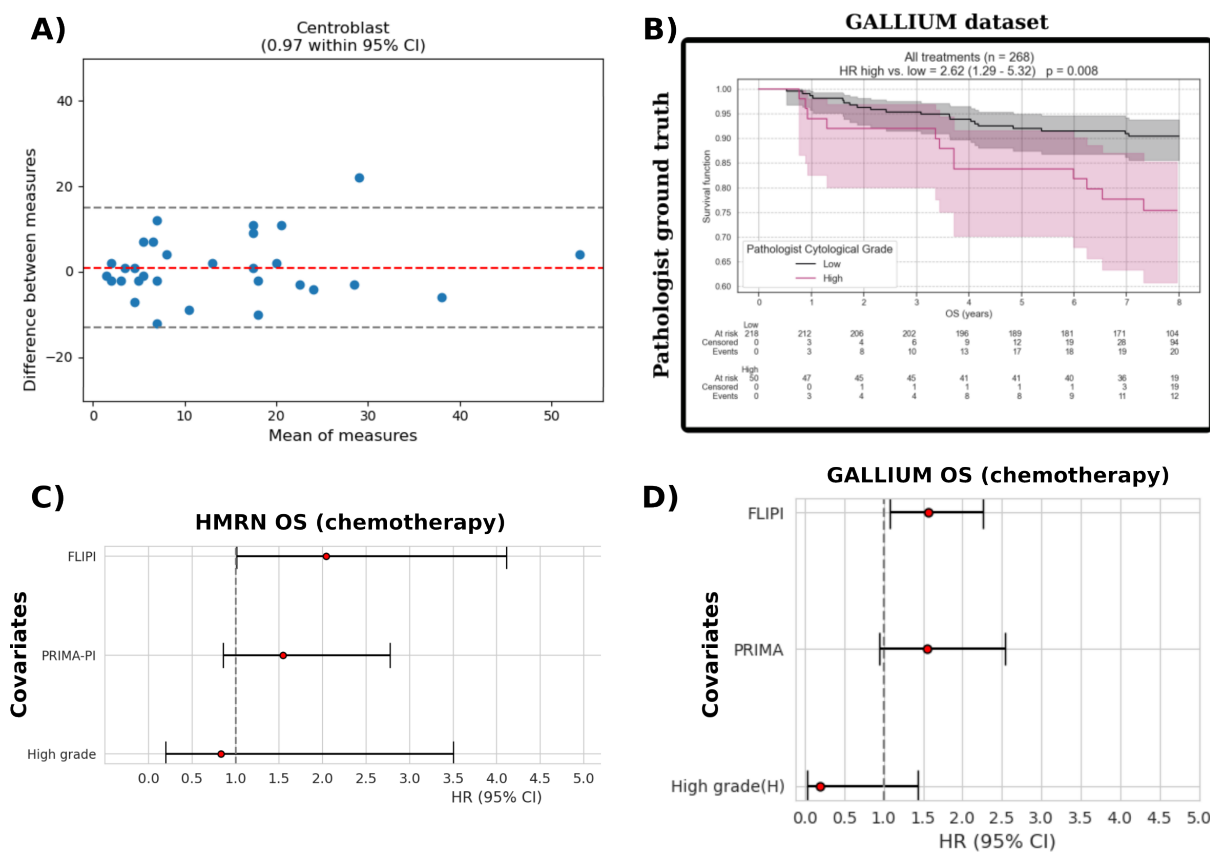


Figure 2.16: Key results of Chapter 2: A) Bland-Altman plot of centroblast counts made by a pathologist (using a quadrat) and the machine centroblast classifier. Blue dots = comparisons for individual image patches; Red dotted line = mean difference in centroblast counts between pathologist and machine; Grey dotted lines = 95% confidence intervals.

B) Kaplan-Meier plot of overall survival events for the strict pass GALLIUM dataset (all treatments), stratified by pathologist assigned high cytological grade. Hazard ratios reported relate to univariate Cox proportional hazard estimators; significance testing performed using the Wald test.

C) Multivariate Cox proportional hazard ratios for overall survival events in the chemotherapy partition of HMRN data. Red dots = HR; whiskers = 95% confidence intervals; High grade = grade predicted using machine learning model. D) Multivariate Cox proportional hazard ratios for overall survival events in the strict pass GALLIUM group. Red dots = HR; whiskers = 95% confidence intervals; High grade (H) = grade predicted using machine learning model on nuclear features scaled with HMRN dataset parameters.

5.2 Limitations

The first major source of error in the automated classifier was low reliability in ground truth annotation of follicles. Low contrast of follicles in some H&E slides complicated annotation, potentially leading to disparities in performance, as compared to pathologist. This was especially visible when evaluating H&E model performance with the more reliable CD10 (follicle B-cell marker) IHC ground truth. The use of IHC was specifically avoided in inference to allow for widest usage, yet, use of IHC to guide annotation of H&E ground truth may have led to more reliable ground truth and improved performance of the region classifier.

Secondly, evaluation on validation data did not evidence any benefit of Model A in the region classification UNet. Removal or replacement of this model may have led to better overall region classifier performance.

Thirdly, region classification was only trialled using one model type (UNet) on a single magnification ($1\times$), despite evidence of higher B-cell region classification performance with the multi-resolution HookNet architecture[128]. Future efforts may benefit from integration of multiple resolutions in classification of naive B-cell regions.

Finally, severe performance disparities between region of interest (ROI) and slide level grading indicated systematic differences between machine and pathologist aggregation of information for a whole slide. Further investigation of pathologist decision-making may aid in increasing grading agreement.

5.3 Key takeaways

- Classification of centroblasts on objective, interpretable nuclear features led to performance comparable to other top-performing techniques
- Classification of naive B-cell regions (follicles) was complicated in H&E by poor contrast. These regions were far clearer in CD10 IHC, allowing for more reliable ground truth annotations.

- A classifier of high cytological grade performed well at the region of interest level but poorly at the slide level, indicating disparities between machine and pathologist region selection or slide-level aggregation
- Pathologist high cytological grade was associated with over $2.5\times$ increased hazard, compared to low grade, while machine high grade on the same data was associated with over $4.5\times$ reduced hazard

5.4 Future directions

Firstly, survival analyses of pathologist-labelled high grade in the GALLIUM trial evidenced a significantly increased risk for high grade patients. This demonstrated a clear clinical utility for cytological grading in prognostication and hence, automated efforts such as presented here.

The weaknesses of the proposed cytological grading tool appeared to lie in region selection and subsequently, agreement with pathologists on grade. The key to improved grading performance may therefore lie in development of an improved region selection strategy. A core aim of this work was for inference on H&E slides, to allow for widest usage. Annotation of a reliable ground truth for training of a region classifier was challenging in H&E due to low contrast with background. On the other hand, there was a clear contrast in CD10 IHC, allowing for a more reliable ground truth. An improved classifier and therefore automated cytological grading may require training on H&E slides, using ground truth annotated on matched, registered IHC slides.

Secondly, machine classifications of high grade were associated with reduced OS hazard. Further explanation of these classifications may aid in identification of determinants of good prognosis in FL, potentially involving presence of reactive B-cell regions. Since such assessments are not conducted in FL, it may represent a novel biomarker in FL. Future work in distinction of reactive and neoplastic B-cell regions, possibly using CD10 and BCL2 immunohistochemistry data in training of H&E classifiers, may improve the range of histopathology-based prognosis methods in FL.

Contributions

Labelling of the pathologist annotated dataset of nuclei was performed by Dr. Alireza Behzadnia of the Leeds Teaching Hospitals NHS Trust (LTHT). Pathologist cell counts and patch grading were performed by Professor Reuben Tooze of LTHT and the University of Leeds (UoL). Cytological grading of GALLIUM data was performed as previously reported[107]. Professors David Westhead (UoL) and Andrew Janowczyk (Geneva University Hospitals, Switzerland and Emory University, USA) provided feedback on methodology throughout. All other work was conducted by Volodymyr Chapman (UoL).

Chapter 3

Patient nuclear subgroups

1 Abstract

Introduction:

Efforts in non-Hodgkin Lymphoma have not bridged relationships between molecular profiles (expression and mutation), survival and cellular composition within routine haematoxylin and eosin (H&E) staining. We address this through clustering of patients on cell populations, using nuclear features extracted from H&E slides. Finally, we assess clinical utility through comparison with molecular features and overall survival.

Methods:

337 follicular lymphoma (FL) and 296 DLBCL FFPE H&E slides of the population-based HMRN dataset were used to train clustering methods. Nuclear features (e.g. area and dark blob number) were measured in unlabelled nuclei sampled from naive B-cell regions. Nuclear data was aggregated for whole slides using two approaches: median concentration of cell clusters (nuclear phenotype) and a distribution of all cells (nuclear distribution). Patients were clustered separately on these aggregations, with optimal cluster number determined using maximum Calinski-Harabasz score. Slide-level clusters were compared to molecular features and overall survival (OS) of patients.

Results:

2 cellular phenotypes were identified in FL, with significantly higher proportions of uncommon mutation profiles (STAT6 and aSHM) in the centrocytic group ($p=0.03$). In DLBCL, patients clustered into centrocytic groups had substantially worse OS, espe-

cially with cell distribution clustering (HR distribution 2 vs. rest = 1.95; 95% CI: 1.11 - 3.43; $p = 0.02$).

Conclusions:

These results evidenced distinct cellular subgroups of FL and DLBCL, with characteristic nuclear morphologies, mutation and survival. Future work may integrate spatial features to further resolve cases.

2 Introduction

2.1 Successes in non-Hodgkin Lymphoma

Recent work has provided an unprecedented insight into the pathogenesis of non-Hodgkin Lymphoma. Major findings have included increased disease progression risk in mutational subgroups of Follicular Lymphoma (FL)[52] and biomarkers associated with treatment success in large B-cell lymphomas[133].

Within cellular studies, the following have been identified:

- Utility of predicted neoplastic cell populations in DLBCL overall survival (OS)[66]
- Follicular dendritic cell marker presence (CD21+) and worse OS in FL cases with transformation to DLBCL[134], otherwise not observed in untransformed cases[65]
- Lower density of CD68+ macrophages, density of immune infiltrate and diversity in immune cells associated with increased risk of early disease progression[65]
- Lower density of specific T-cell populations associated with increased risk of early disease progression[135]
- Increased risk with immunoblastic cell phenotypes in DLBCL[67]

These studies have seldom focused on routine haematoxylin and eosin (H&E) pathology slides. With evidence of substantial clinical value in cellular classifications, there remains an unmet need in classification of such biomarkers in routine pathology.

2.2 Unsupervised classification

Many of the cell types involved in non-Hodgkin Lymphoma, including tingible body macrophages, follicular dendritic cells and centrocytes are unique to lymph nodes, playing crucial roles in controlled somatic cell mutation activities. As such, annotated samples of these cells do not feature in openly-available nuclear datasets, even those that include lymph node biopsies[136, 137].

Distinction of these cells in routine H&E is not trivial, with reliance on subjective fea-

tures and high inter-observer variability[59, 61] complicating labelling of a reliable ground truth dataset for training of nuclear classifiers. Absence of quality public datasets featuring these cell types and high innate uncertainty in expert annotation complicate tool development for routine H&E.

Necessary steps for development of such tools are for extraction of objective nuclear features, to improve consistency and for development of systems less dependent on unreliable, subjective annotations. Unsupervised clustering of cells is one approach for meeting these requirements, where patterns within objective nuclear features may guide classification, greatly reducing reliance on human annotation. Objective morphological features of nuclei have been used successfully for classification of molecular states and risk in H&E[138, 139]. An advantage of these features over more complex features derived from deep learning models, such as image convolutions[81], is improved scope for interpretation, a key requirement for integration of software tools into routine practice[140, 141].

2.3 Aims

In this study, we propose an unsupervised classifier for quantification of a range of cell types in non-Hodgkin Lymphoma. Patients are clustered on cellular features with the aim of identifying distinct patterns of cellular pathology in FL and DLBCL. Finally, patient-level clusters are compared to molecular features (expression and mutation profiles) and overall survival to evaluate clinical utility.

The particular items of novelty in this study are:

1. Development of an unsupervised classifier for a range of nuclei in non-Hodgkin Lymphoma using objective nuclear features
2. Calculation of H&E nuclear signatures in FL & DLBCL
3. Evaluation of the clinical utility of clusters with respect to overall survival
4. The first quantitative evaluation of molecular features and cell morphology in FL and DLBCL

3 Methods

3.1 Experiment Workflow

The ultimate goal of this work was to cluster FL and DLBCL patients on similarities in cell composition. Two complementary methods were used to achieve this. Both methods followed the same region selection, nuclear segmentation and nuclear feature extraction procedures as outlined in Chapter 1 sections 3.7, 3.6 & 3.8, respectively.

Approach 1, termed ‘nuclear phenotype’ aimed to classify cells in biopsy samples and cluster patients on median concentration of cells (Figure 3.1 top).

Approach 2, termed ‘nuclear distribution’ aimed to cluster patients on distributions of cell features, circumventing classification of cells (Figure 3.1 bottom).

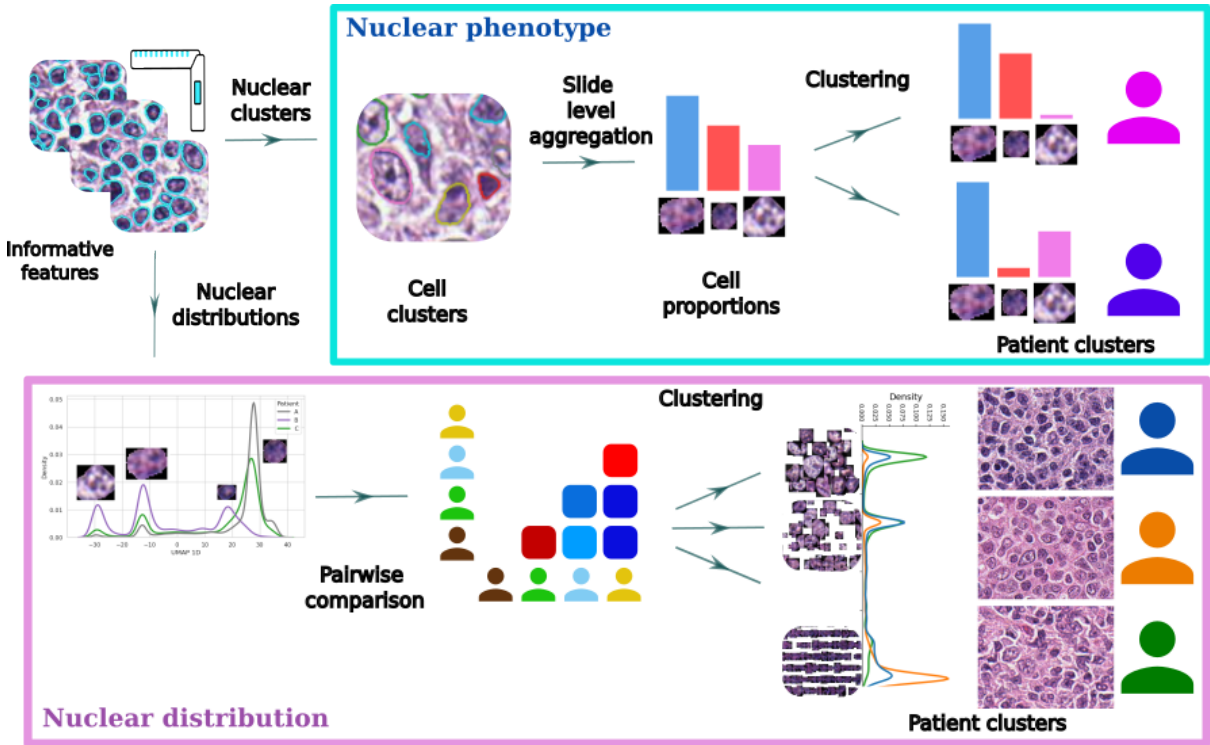


Figure 3.1: Overview of nuclear phenotype and distribution approaches for clustering patients on cell populations.

3.2 Patient data

Models were trained on data from the UK population-based Haematological Malignancy Research Network (HMRN; <https://www.hmrn.org>[100]) dataset. External validation was performed on data from the GALLIUM[106, 107] and DLBCL-MORPH [66, 142, 143] datasets for FL and DLBCL, respectively. Mutation profiles were calculated by the HMRN as previously described for FL[46]. Cell-of-origin expression profiles were calculated by HMRN as previously reported [56, 144–146].

3.3 Whole slide images

Whole slide images retrieval and visual quality control was performed as described in Chapter 2 Methods section 3.4 and associated publications. HMRN and DLBCL-MORPH datasets were composed of images scanned at $40\times$ magnification. GALLIUM trial data underwent quality control including filtration on magnification ($40\times$) and visual inspection (criteria outlined in Chapter 1). The partition used in the presented work was ‘Lenient pass’, otherwise referred to as ‘Pass and follow up’.

1000×1000 pixel patches, termed regions of interest (ROIs), were extracted from WSIs. ROIs were filtered for at least 90% tissue and naive B-cell region coverage (as outlined in Chapter 2 Section 3.7 Naive B-cell Region Classification). Between a minimum of 20 and maximum of 200 ROIs were extracted, depending on tissue availability.

3.4 Data partitions

As outlined in Chapter 2 Methods section 3.4, data collection was ongoing during the duration of the work presented. Due to limitations in data availability, the ‘nuclear phenotype clustering’ work was trained on 144 of 337 FL slides. Inference on HMRN FL and GALLIUM trial data was performed using this model.

Otherwise, results presented relate to models trained on all data available in the HMRN dataset for FL ($n = 337$) or DLBCL ($n = 296$), respectively. Sample numbers displayed may be lower than these figures owing to missing data.

3.5 Pan Non-Hodgkin lymphoma nuclear dataset

Measurements of over 170,000 unlabelled nuclei were used to create a ‘Pan Non-Hodgkin Lymphoma nuclear dataset’. 42 regions of interest were extracted at a magnification of $40\times$ from 15 WSIs (5 Burkitt Lymphoma, 5 FL and 5 DLBCL slides) by an observer. This effort was aimed at capturing diversity in regions and cells and so patches were taken across diverse regions including germinal centres, inter-follicular regions and vascular regions.

177,280 nuclei were segmented using the MoNuSaC checkpoint of HoVerNet. The five most informative features for nuclear classification, identified in Chapter 2 Section 3.8 Machine Learning approach were measured:

- Area
- Maximum intensity (Red channel)
- Minimum intensity (Red channel)
- Number of dark blobs
- Local binary pattern edge ratio

Features were clipped at the highest and lowest percentiles to limit impact of measurement outliers. Features were linearly scaled in the range 0-1 to enforce more equal contribution in clustering.

3.6 Unsupervised Non-Hodgkin cell classifier

Unlabelled nuclei of the pan Non-Hodgkin dataset described above (Section 3.5) were clustered into cell types using K-Means clustering. This distance-based clustering algorithm was chosen specifically to divide the feature space into roughly equal parts, representing cell types across the feature space. Density-based methods (such as DBSCAN) may have biased clusters against infrequent but potentially informative cell types, such as tingible body macrophages in FL, so were not used, despite potentially leading to more meaning-

ful clustering.

Two clustering approaches were evaluated. In the first approach, nuclei were clustered directly on scaled nuclear features using K-means clustering. In the second approach, as previously found to be a valuable preprocessing step for clustering, UMAP was used to embed nuclear features to two dimensions [147–149]. K-means clustering was then performed on embedded features.

Optimal clustering method and cluster number were determined using maximum Calinski-Harabasz score. These were also evaluated with respect to pathologist classifications using homogeneity, completeness and V-measure [150] in labelled data. Details of this dataset are provided in Chapter 2 Section 3.8 Centroblast Classification.

Equations for all clustering metrics are provided in Appendix Section 4.1 Clustering. The output of cell clustering was assignment of nuclei to cluster numbers from 0 to $k-1$ (k being chosen number of clusters). Visual inspection, with feedback from a pathologist (R.T.) was used to suggest likely cell types contained within each cluster.

Finally, the effects of linear scaling and feature reduction (through PCA) on clustering quality were assessed on nuclei from the training dataset (HMRN), combining those from FL and DLBCL ($n = 4.9$ million nuclei from 666 WSIs) and using a greater range of features (all 66 nuclear features in Chapter 2 Section 3.8). This was used to assess whether inclusion of other nuclear features and transformations may have resulted in preferable clustering of cells.

3.7 Nuclear Phenotype Clusters

As outlined in Section 3.1, patients were clustered on median number of cells belonging to each cell cluster, calculated over all ROIs sampled in the whole slide image. This was termed the ‘Nuclear Phenotype’.

Cell type prediction

Naive B-cell regions were classified in whole slide images (WSIs) exactly as described previously (Chapter 2 Methods Section 3.7). Predictions were used to sample 20-200 naive

B-cell region patches, termed ‘regions of interest’ (ROIs), from WSIs at $40\times$ magnification. Number of ROIs extracted depended on tissue availability. Nuclear segmentation and measurement were performed exactly as described for unlabelled nuclei (Section 3.5 Pan Non-Hodgkin lymphoma nuclear dataset). Measurements were clipped and linearly scaled according to parameters determined in the unlabelled nuclear dataset (3.5). Cell types were predicted using the pan Non-Hodgkin cell classifier (Section 3.6, above).

Cell phenotypes

Number of each cell type was aggregated to the whole slide level using the following approach: First, median cell number was calculated over all ROIs per slide. Next, to promote more equal contribution of each cell type in clustering, median concentrations of cells were linearly scaled over each disease partition in HMRN data ($n = 144$ for FL and 334 for DLBCL; scaled between values of 0 and 1). Finally, slides were clustered on median cell concentrations into general patient phenotypes, with use of maximum Calinski-Harabasz score to determine the optimal cluster number. Hierarchical clustering was chosen for this step as performance was found to be close to the top-performing K-Means algorithm (with respect to CH score and arbitrary clustering of slides; 5.3 Figure A.7), with the added value of hierarchical tree construction.

Prediction on unseen data

Since hierarchical clustering methods used on training data did not allow for prediction of patient clusters on unseen data, K-nearest neighbour (KNN) models were trained to perform this task. Non-negative matrix factorisation was tested as a preprocessing step to clustering (as previously used for extraction of gene expression signatures in cancer[151]). Downstream performance of KNN models in prediction of patient cluster were used to assess utility of this step. In FL, classification performance was high without NMF preprocessing of scaled cell concentrations (F1 score of 0.97 on test partition of 25% of slides). In DLBCL, NMF was found to greatly improve performance of nuclear phenotype cluster prediction (F1 scores of 0.77 without vs. 1.0 with, respectively, on test partition of

25% of slides; not shown). Slide clustering was therefore performed on cell concentrations in FL and on NMF-embedded concentrations in DLBCL.

3.8 Nuclear Distribution Clusters

Using median concentrations of cell type in nuclear phenotype clusters (Section 3.7) may have led to loss of information, with respect to subpopulations of cells, not resolved by the cell classifier. The nuclear distribution clustering method proposed here aimed to address potential loss of ‘signal’ through aggregation of nuclear data into summary statistics. Instead, all nuclei of a representative sample of patient tissue were considered, embedded into a distribution. In patient clustering, the entire complement of cells was compared between patient pairs.

Nuclear feature embedding

Nuclear features were extracted from naive B-cell regions exactly as described for Nuclear Phenotype clusters (Section 3.7, above), with sampling from naive B-cell regions, nuclear segmentation and measurement of nuclear features. Methods of comparing nuclear distributions between patients (Anderson-Darling score) required nuclei to be described in a one-dimensional feature space. UMAP was used to embed nuclear features to a single dimension. The UMAP model was trained on unlabelled FL nuclear data (described above in Section 3.5). To improve UMAP convergence, the model was trained on data evenly sampled across the feature space. This was achieved by randomly sampling n nuclei from each cell cluster (Section 3.6), where n was equal to the number of cells in the smallest cluster (process also known as ‘under-sampling’). Following UMAP embedding of nuclei, each WSI was represented as a 1-dimensional vector, with length equal to number of nuclei segmented within the tissue. In pairwise comparison of patients, these vectors were converted to empirical distribution functions of a comparable scale (bounded by the minimum and maximum values of the 1-dimensional UMAP embedding).

Patient clusters

The Anderson-Darling (AD) statistic was used for comparing nuclear distributions between patients in a pairwise manner. The statistic was used as a distance metric between patient samples. AD was chosen over the Kolmogorov-Smirnov statistic due to more equal weighting provided across the entire length of distributions in comparison. The SciPy implementation of the K-sample Anderson-Darling test was used for calculations (`anderson_ksamp` function)[152].

The resultant distance matrix was used for hierarchical clustering of patients into nuclear distribution groups (using Scikit-learn `AgglomerativeClustering` function[153]). Optimal cluster number was determined using maximum Calinski-Harabasz score on complete linkage.

Prediction on unseen data

As for nuclear phenotype clusters, it was not possible to directly predict nuclear distribution clusters on unseen data due to dependence on pairwise relationships. Again, KNN models were trained to predict patient clusters for unseen data. Following naive B-cell region sampling, nuclear segmentation and measurement of nuclear features, features were embedded using the UMAP model trained on unlabelled pan non-Hodgkin lymphoma nuclei (Section 3.8). Next, for each slide, relative frequency of nuclei at equally-spaced intervals across the 1D UMAP feature space were calculated. This was conducted to create feature vectors of the same length for each patient (allowing comparison between patients with different numbers of nuclei detected). Finally, KNN models were trained to predict nuclear distribution cluster using relative frequency feature vectors. The optimal number of intervals used in calculation of relative frequency was determined using grid searches, using 5 fold cross validation. In this, 75% of slides in the respective HMRN datasets (FL and DLBCL; randomly sampled, stratified by distribution cluster membership) were used as training data. The remaining 25% of slides were used for evaluation of the final KNN classifier, following determination of optimal interval number.

3.9 External validation

External datasets were used to validate characteristics of patients assigned to clusters. For each patient, nuclear phenotype and distribution cluster membership was predicted using models trained on the internal HMRN training dataset. The following features were compared between patient clusters:

- survival: hazard ratios of Cox proportional hazards models between datasets
- patient characteristics: correlation between groups and patient features such as prognostic indices, B-symptoms and treatment

Additionally, independent nuclear phenotype clusters were fitted to external data, to compare optimal cluster number and survival of patients. These models were trained on external data exactly as models were trained on internal HMRN data (Section 3.7). This was used to assess whether clusters fitted on training data represented general patient groupings. A general summary of validation is provided in Figure 3.2, outlining how nuclear phenotypes and distribution clusters were predicted on validation slides using models trained on HMRN data and additionally, separate nuclear phenotype models were trained on validation data, also.

Due to computational limitations, nuclear distribution clustering models were not trained on external datasets. More detailed methods for individual steps are provided in subsections below.

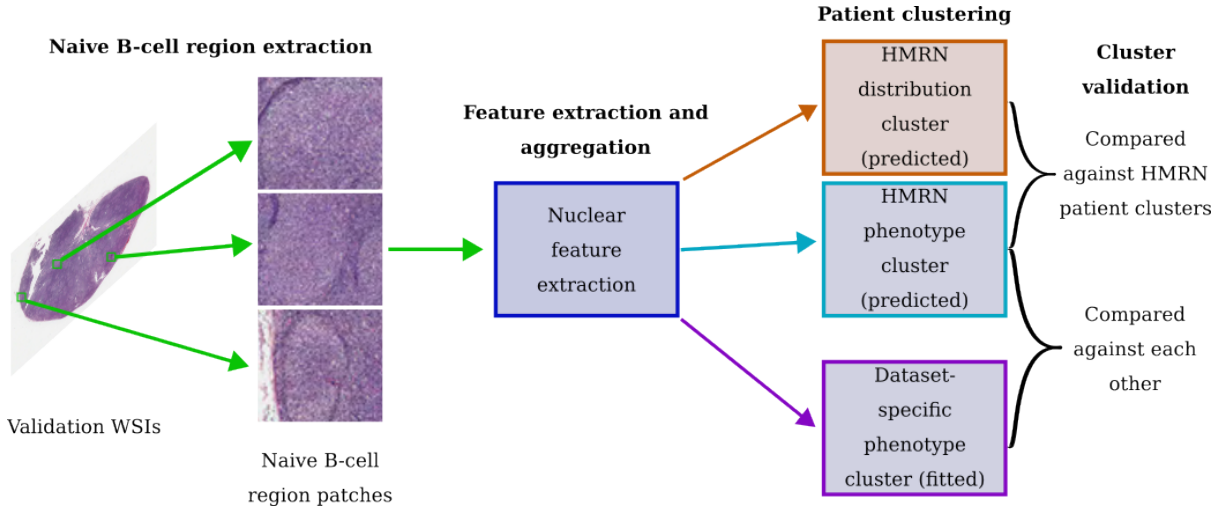


Figure 3.2: Outline of procedure used for external validation of patient clusters. Process comprised of naive B-cell region classification in external slides (left; as in internal HMRN training data), followed by nuclear segmentation (using HoVerNet[103]), nuclear feature extraction and linear scaling of features (according to HMRN feature value range). Nuclear phenotype and distribution clusters were predicted on validation data using models trained on internal data (right, top and middle boxes). Additionally, nuclear phenotype cluster models were fitted to individual external data to compare the optimal number of patient clusters between datasets.

Nuclear feature extraction

Prediction of patient nuclear phenotype and distribution cluster membership required measurement of nuclear features in naive B-cell regions, as in training data. This was performed exactly as described for training data (Section 3.7, above). Following measurement of nuclear features, these were linearly scaled (between 0 and 1) using parameters of the pan non-Hodgkin lymphoma nuclear dataset (as in training data).

Nuclear Phenotype Clusters

For prediction of nuclear phenotype, cell types were first predicted for each nucleus using the pan Non-Hodgkin cell model described above (Section 3.6). Median number of cells belonging to each cluster was calculated for WSIs (per ROI). These values were linearly scaled (between 0 and 1) using parameters from the respective HMRN training data. Negative values (where concentration value was lower than the minimum observed in

training data) was clipped at 0. Nuclear phenotype clusters were predicted using KNN models, trained on HMRN data (as described in Section 3.7, above).

Nuclear Distribution Clusters

Scaled nuclear features were embedded into a single dimension using the UMAP model trained on unlabelled nuclear data (Section 3.5). Nuclear distribution cluster was predicted for each WSI using a KNN model (described in Section 3.8, above), operating on relative frequency of nuclei at equally-spaced points in the UMAP embedding range.

3.10 Survival Analysis

Survival analyses were performed for nuclear phenotype and distribution clusters comparably to cytological grade (Chapter 2 Section 3.11) using Cox Proportional Hazards models (Cox models; Equation 2.1) and Kaplan-Meier estimators for visualisation. Survival models were trained with overall survival data clipped at 8 years. This was done to reduce potential impact of confounding conditions under the assumption that multimorbidity risk increased with age and time.

As with cytological grade:

- Models were fitted to overall survival, with events defined as death by any cause
- Hazard ratios were used to evaluate contribution of cellular features in survival risk
- Cox models were corrected with FLIPI and PRIMA-PI [154, 155] using patients for which this data was available
- Concordance index (C-index) was used to evaluate quality of Cox model fit and hence, extent to which cytological features were appropriately corrected for in calculation
- Significance testing was performed using the Wald test for a single parameter on Cox model co-efficients
- Multiple testing correction was performed using the Bonferroni correction, where

applicable

Further justification and explanation of the above procedures are provided in Chapter 2 Section 3.11: Survival analysis.

Validation of best-performing methods

As outlined in Section 3.4 above, patient cluster inference was performed on external validation data using models trained on HMRN data. This was used for evaluation of survival prediction for those models.

To validate optimal patient groupings observed, the top-performing patient clustering method was additionally fitted on external datasets. Optimal cluster number, constituent cell types and patient survival were compared between HMRN and external validation data.

Evaluation metrics

Definitions and equations for all evaluation metrics used are provided in Appendix section 4.1 Evaluation metrics.

3.11 Software

Whole slide images (WSIs) were parsed with OpenSlide 3.4.1, using OpenSlide-Python 1.3.1 for Python bindings. All downstream analyses were performed using Python 3.8.18. Machine learning was performed using scikit-learn 1.2.2. Full software environment specifications can be provided on request.

3.12 Hardware

Hardware used varied for each dataset given individual requirements and terms outlined by the data owners. Inference was performed using NVIDIA Tesla T4 GPUs and Intel Xeon CPUs (of unknown models) on cloud providers for external data (Vivli platform for GALLIUM and Google Cloud for DLBCL-MORPH). For internal HMRN data, training

and inference were performed on an NVIDIA RTX 3060Ti GPU with an AMD R5 3600 CPU. Full environment specifications, including NVIDIA driver and CUDA versions used can be provided on request.

4 Results and discussion

4.1 Unsupervised Non-Hodgkin cell classifier

177,000 unlabelled nuclei were clustered on nuclear features with and without prior UMAP embedding into 2 dimensions. Evaluation was performed using clustering metrics to identify optimal cluster number (Appendix Figure A.6) **Evaluation of optimal nuclear cluster number**. No clear optimum, as defined by maximum metric values, was identified, with monotonic changes in Calinski-Harabasz (CH) score for increasing cluster number on raw (decreases in CH) and UMAP-embedded features (increases in CH). A similar effect was observed on clustering of training set nuclei using a greater range of nuclear features and preprocessing steps (Appendix Figure A.6C). Clustering of 10 cell types on raw nuclear features was empirically chosen for downstream tasks. The justification for this decision was a compromise between greater homogeneity than the V-measure maximum at 8 clusters (evaluated on labelled cells) and greater CH score than the homogeneity score maximum at 19 clusters.

Qualitative evaluation of clustering (through visual inspection) displayed in Figure 3.3, below, was used to assign likely cell type to nuclei. Centroblasts were observed to be assigned to nuclear cluster 6 (Figure 3.3A; top right of plots) with centrocytes (clusters 1, 2, and 8) neighbouring in the UMAP feature space.

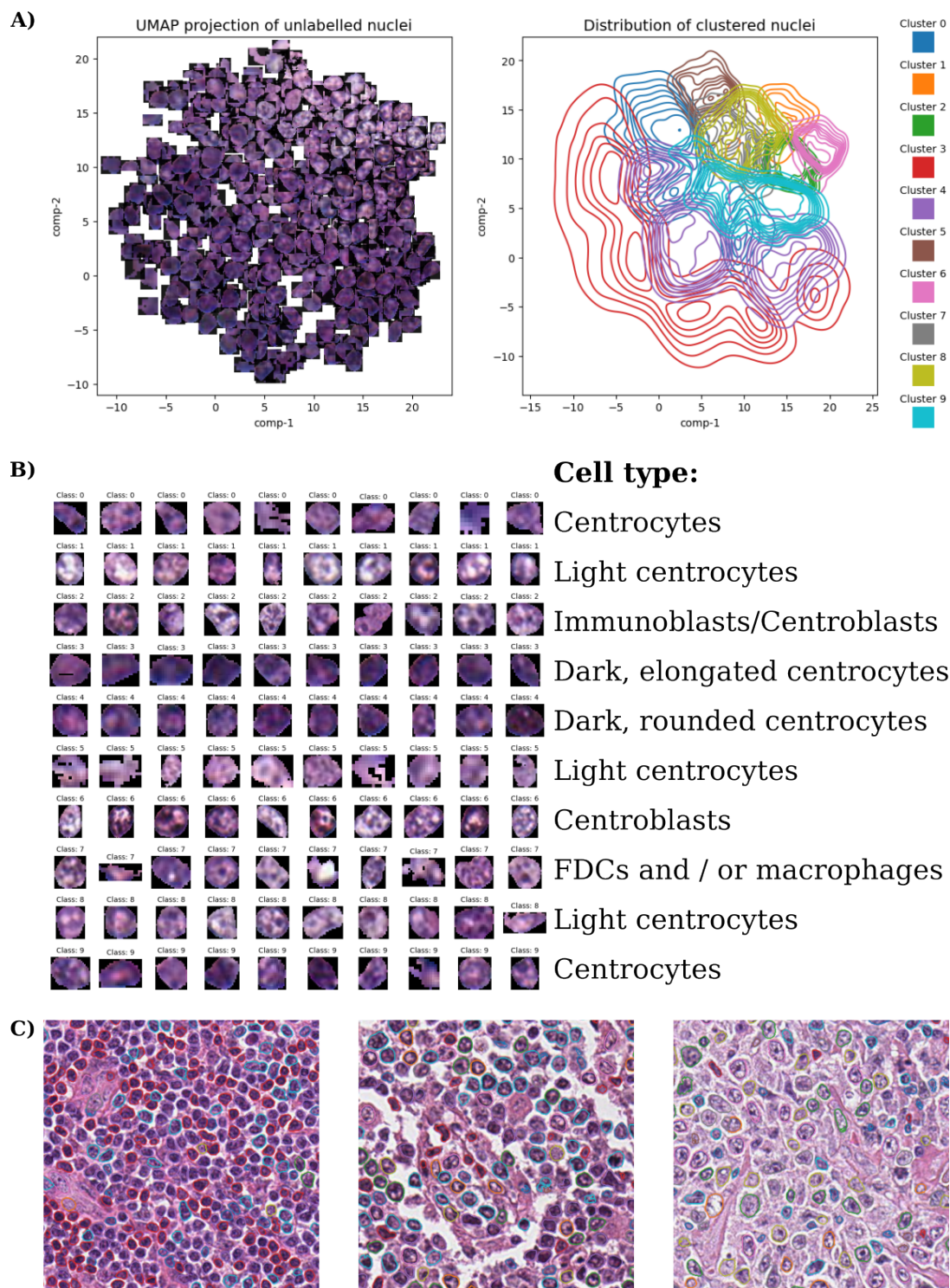


Figure 3.3: Visual inspection of clustering on unlabelled FL, DLBCL and BL nuclei.

A) Projection of nuclei and clustering boundaries in 2-dimensional UMAP space

B) Examples of nuclei and possible cell types (assigned by me)

C) Contextual images, with nuclear boundary colour indicating cluster

Discussion

Quantitative evaluation did not identify a clear optimum cluster number or input feature type (nuclear features or UMAP-embedded features), as previously displayed for other data modalities[147–149]. Efforts to expand number of nuclear features to all those previously assessed for centroblast classification (Chapter 2 Section 3.8) and feature pre-processing methods did not identify clear optima either (Appendix Figure A.6c). For this reason, the features used were those informed by pathologist annotations of key cells in FL and DLBCL (Chapter 2 Section 3.8).

Overall, these observations evidenced that either:

- The 66 nuclear features extracted were suboptimal for distinction of cells
- Feature transformation methods tested were ineffective for distinction of cells

The effect of suboptimal clustering, as evidenced by qualitative evaluation was overlap in cell types and characteristics between clusters. This was especially visible in clusters 3 and 4 (Figure 3.3a and b), both consisting of dark centrocytes.

The purposes of cell clustering were to measure cell number across the nuclear feature space. It was a delicate compromise between sufficiently high cluster number to measure subtle differences between patients while keeping cluster number low enough to prevent excess noise through trivial cluster differences.

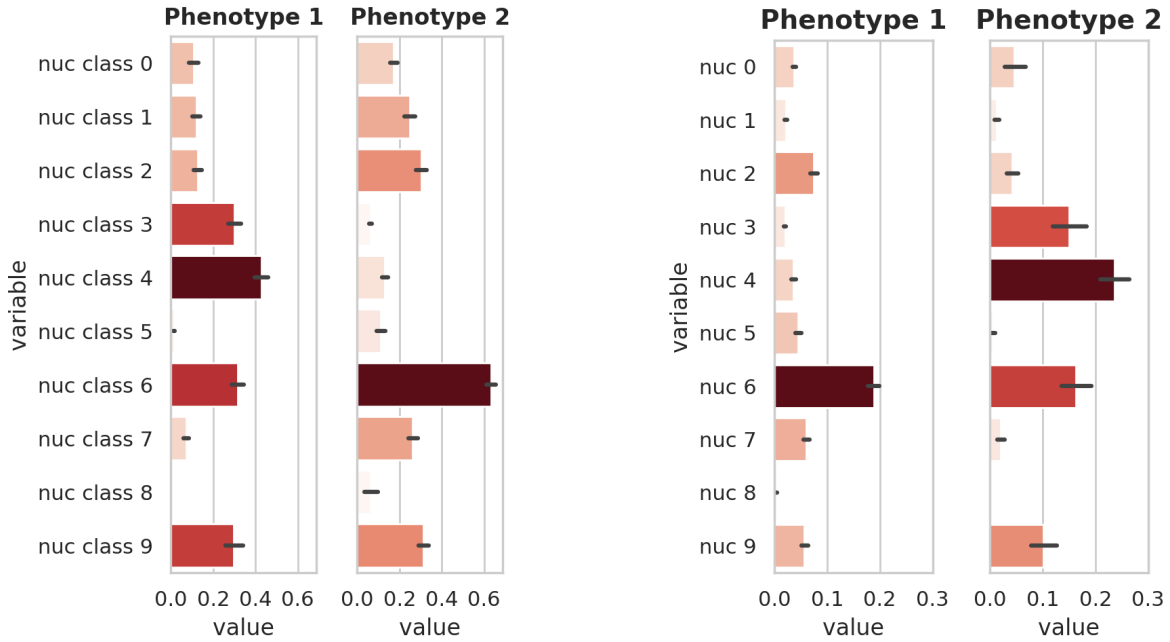
Quantitative and qualitative evaluation evidenced that the clustering schema selected was potentially not optimal for the intended purpose.

4.2 Nuclear Phenotypes Clusters

Patients were clustered on cell populations using discrete cell types (nuclear phenotypes) and continuous cell distributions (nuclear distribution clusters).

Following clustering of cells, the number of each cell type was calculated for all ROIs. These counts were aggregated on a WSI level using median number of each cell type per ROI.

2 patient clusters were identified in FL and DLBCL (Appendix Figures A.7). Though greater numbers of clusters were evidenced in DLBCL, this resulted in cluster membership of less than 10 cases.



(a) Follicular Lymphoma

(b) Diffuse Large B-cell Lymphoma

Figure 3.4: Proportion of cell clusters in nuclear phenotype groups. Calculated using median number of each cell type per ROI, aggregated over WSIs and linearly scaled between 0 and 1. Error bars represent 95% confidence intervals. Nuc class = Nuclear cluster

In FL (Figure 3.4a), phenotype 2 was characterised by an especially high proportion of centroblasts (nuclear class 6) so was termed a ‘centroblastic’ phenotype. Phenotype 1 was characterised by moderate proportions of centroblasts and high proportions of dark centrocytes (nuclear classes 3 and 4) and was therefore termed ‘centrocytic’. Slides were divided between these groups in the following proportions:

- 1 (Centrocytic): 39% (n=129)
- 2 (Centroblastic): 61% (n=206)

Like in FL, one DLBCL phenotype (1) was characterised by high proportions of centroblasts (nuclear cluster 6 and 2; Figure 3.4b) with few other cells, while the other phenotype (2) had high proportions of centroblasts with dark centrocytes (classes 6 along with 3

and 4, respectively).

Slides were split between these groups:

- 1 (Centroblastic): 93% (n = 275)
- 2 (Centroblasts and centrocytes): 7% (n = 21)

4.3 Nuclear Distribution Clusters

The optimal number of nuclear distribution clusters were deemed to be 3 for both FL and DLBCL. Inspection of mean cell densities for each patient cluster (Figure 3.5, below) revealed that groups noticeably differed at four points in the UMAP embedding, with values:

- -30,: large nuclei with multiple nucleoli (primarily centroblasts[156])
- -13: large nuclei with a few nucleoli (immunoblasts, FDCs, macrophages)
- 18: mixed size, dark and generally jagged nuclei (centrocytes)
- 27: small, mixed shape, dark nuclei (centrocytes and mature lymphocytes)

Clusters in FL and DLBCL bore resemblances:

- Cluster 1 of FL and DLBCL cluster 2 both had higher proportions of centrocyte/mature lymphocyte cells than other disease clusters. In each, proportion of these cells was approximately double proportion of immunoblasts and several times greater than centroblast densities (10× in FL and 4× in DLBCL)
- DLBCL cluster 1 and FL cluster 3 displayed high cell densities of centroblasts, with lymphocyte populations of less than 3%. Centroblasts comprised over 20% of all cells in DLBCL, (4× higher than other large nuclei). In FL, this disparity was smaller at 12% centroblast density (35% higher than the large nuclei peak at UMAP value of -13).
- Each disease had a ‘balanced’ cluster, where two majority cell types were present in almost equal proportions. In FL, this was observed in cluster 2, with similar

proportions of immunoblast and centrocyte cells, and $2\times$ lower centroblast density than centrocytes (CC/IB cluster).

In DLBCL cluster 3, near equal proportions of centroblast and immunoblast cells were coupled with $2\times$ lower density of centrocytes (CB/IB cluster).

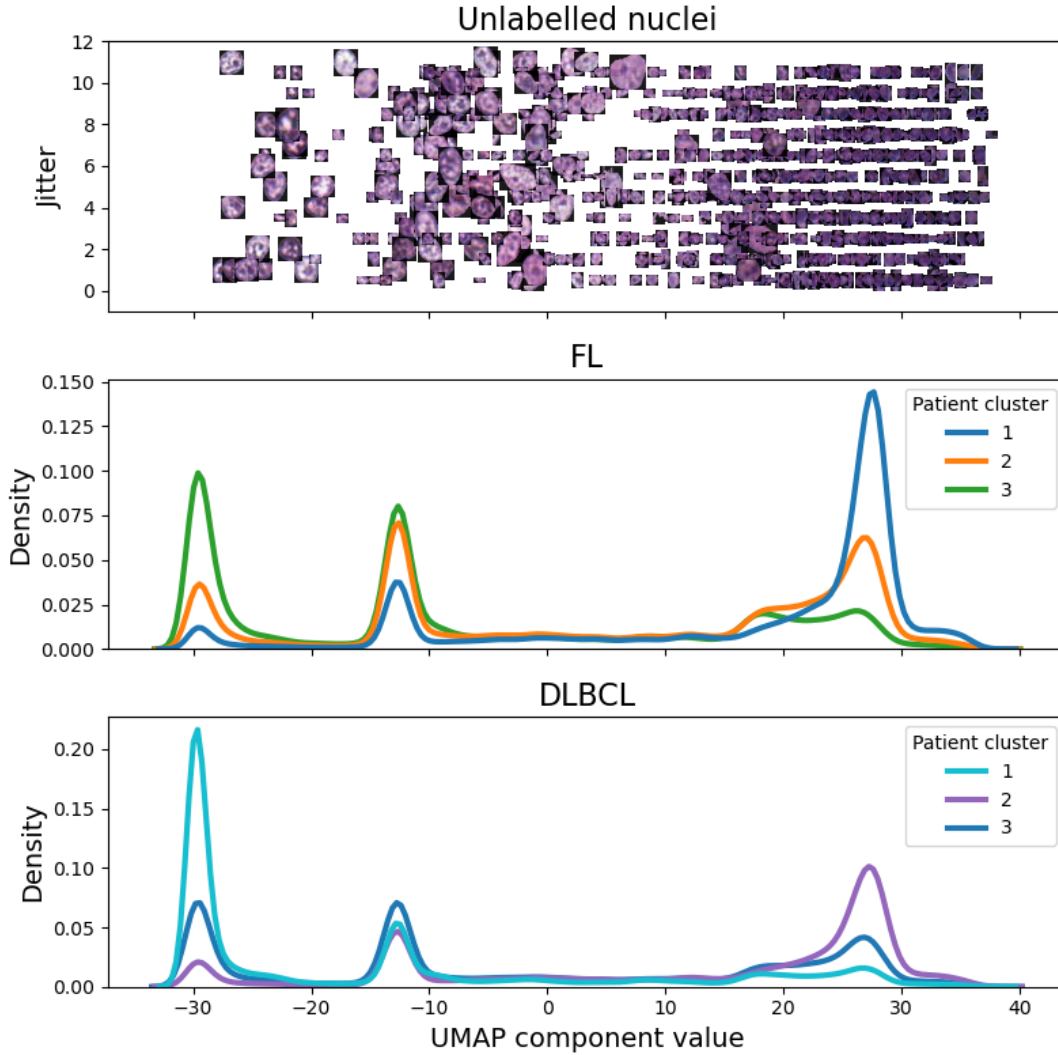


Figure 3.5: Cell types present in nuclear distribution clusters of FL (middle) and DLBCL (bottom), calculated across all nuclei from slides assigned to each cluster and aligned with images of nuclei across the 1D UMAP embedding.

FL slides were divided between groups as follows:

- 1 (Centrocytic): 41% (n=137)
- 2 (Centroblasts and immunoblasts): 20% (n=66)

- 3 (Centroblastic): 39% (n=132)

While DLBCL slides were divided between groups:

- 1 (Centroblastic): 66% (n=220)
- 2 (Centrocytic): 10% (n=33)
- 3 (Centroblasts and immunoblasts): 24% (n=81)

4.4 Comparison of clustering strategies

Contingency matrices were used to compare clustering of slides, in order to assess redundancy and unique value in each technique.

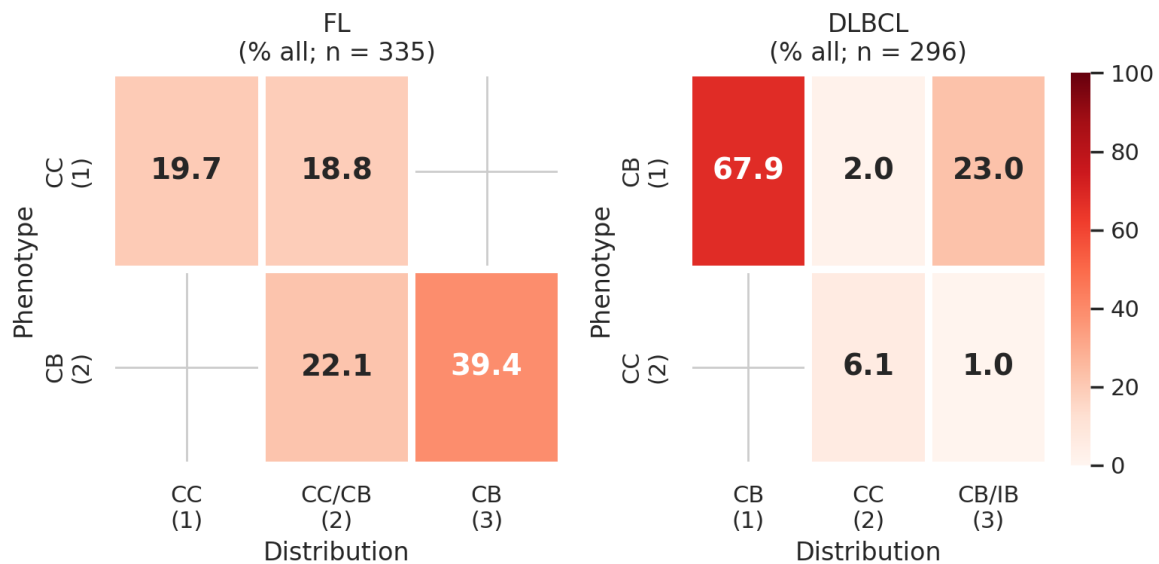


Figure 3.6: Contingency matrices visualising membership of FL and DLBCL slides between nuclear phenotype and distribution groups. Nuclear group numbers denoted in brackets. Values represent percentage of all FL or DLBCL slides. These may not sum to 100% due to rounding. Dominant cell types present: CB = centroblastic; CC = Centrocytic; IB = Immunoblastic; Large = large nuclei (CB, IB and dendritic cells). Values of zero masked.

In FL, centroblast-dominant clusters comprising almost 40% of cases (nuclear phenotype 2 and nuclear distribution 3) perfectly overlapped. Centrocytic clusters (distribution 1 and phenotype 1) also overlapped, accounting for almost 20% of patients. Finally, cases

belonging to distribution cluster 2 (CC/IB) were almost evenly split between nuclear phenotype 1 and 2. These results evidenced a high level of redundancy between clustering methods in FL, with nuclear distribution cluster 2 representing a midpoint between nuclear phenotypes 1 and 2.

In DLBCL, all cases assigned to the centroblastic distribution cluster 1 were assigned to the centroblastic nuclear phenotype (1). Likewise, over 95% of cases assigned to nuclear distribution cluster 3 (with high proportions of centroblasts and immunoblasts) were also assigned to nuclear phenotype 1. This was expected as nuclear phenotype 1 was found to have noticeable populations of nuclear class 2, identified as a mix of blastoid cells. 75% of nuclear distribution cluster 2 (characterised by centrocytes) was assigned to the centrocytic phenotype (2). Again, these results evidenced a high level of redundancy between clustering methods. Both methods agreed on a majority of cases in centroblastic and centrocytic clusters (1 and 2 for both methods). Nuclear distributions additionally distinguished cases containing other blastoid cells in cluster 3, a potential advantage over phenotypes.

Molecular characteristics

Nuclear clusters were characterised with respect to published molecular subgroups. In FL, comparisons were made against the following published mutation groups[46]:

- ‘STAT6’, with high rates of *STAT6* and commonly co-mutated genes, characteristic of diffuse pattern FL[157–159]
- ‘aSHM’, with high mutation at loci associated with aberrant somatic hypermutation (aSHM, described in the Introduction Section Somatic mutations)
- ‘COM’, a majority group consisting of all remaining patients

In DLBCL, clusters were compared against cell-of-origin expression groups (COO)[56, 144–146]:

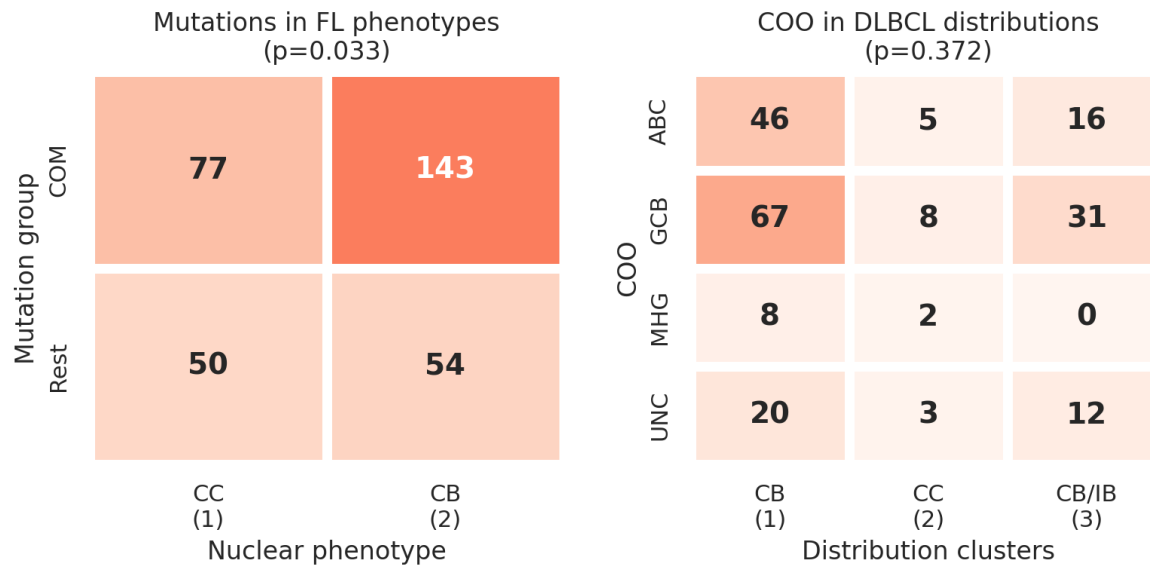
- ‘GCB’, a majority group displaying similarities in expression of BCL6, CD10 and

other markers of naive germinal centre B-cells[19]

- ‘MHG’, a ‘molecular high grade’ subgroup of GCB bearing similarities to Burkitt Lymphoma in expression and enrichment of *c-Myc* mutation[20]
- ‘ABC’, a subgroup with expression similarities to activated B-cells (plasmablasts), including expression of MUM1 and BCL2[19]
- ‘UNC’, a group consisting of cases otherwise unclassified

Significant differences were observed in molecular characteristics of FL nuclear phenotype groups (Figure 3.7, below, with molecular composition of clusters provided in Appendix Figure A.9). 73% of FL nuclear phenotype 2 patients (centroblastic) had COM mutation profiles, with respect to 61% in phenotype 1 (centrocytic; $p = 0.033$). No significant differences were observed between nuclear distribution groups (contingency of mutation vs. nuclear distribution in FL: χ^2 $p = 0.79$; not shown).

In DLBCL, contingency matrices of clusters against COO potentially evidenced decreased ABC expression in centrocytic and blastoid groups (nuclear phenotype 2 and distribution clusters 2 and 3) and increased MHG expression in centrocytic groups (phenotype and distribution cluster 2), though these were not significant (distribution χ^2 $p = 0.37$; phenotype χ^2 $p = 0.49$). Such effects were understandably not observed with binarisation of COO (GCB vs rest), with p-values above 0.9 for comparison with binarised expression and Hans algorithm prediction (a routine procedure for prediction of GCB COO against others[25]; not displayed). As with HMRN, DLBCL-MORPH Hans algorithm predictions were also not found to associate with clusters ($p > 0.9$; not shown). Since differences between DLBCL groups occurred primarily in ABC and MHG expression, absence of observable differences between groups in GCB expression (as with binarisation and the Hans algorithm) was expected.



(a) Follicular Lymphoma (FL)

(b) Diffuse Large B-cell Lymphoma (DLBCL)

Figure 3.7: Composition of best-performing clustering methods, with respect to supervised metric performance on COM mutation in FL (mutation[46]) and GCB cell-of-origin expression in DLBCL (COO;[56, 144–146]), where molecular data was available. Values represent number of slides out of all. P-values were calculated using χ^2 tests on the displayed $n \times m$ contingency tables of frequency, where n = number of molecular groups and m = number of patient clusters. No multiple testing correction was applied.

Finally, clustering methods were evaluated in the context of biologically-relevant patient subgroups using supervised clustering metrics (homogeneity, completeness and v-measure) on molecular features (Figure 3.8).

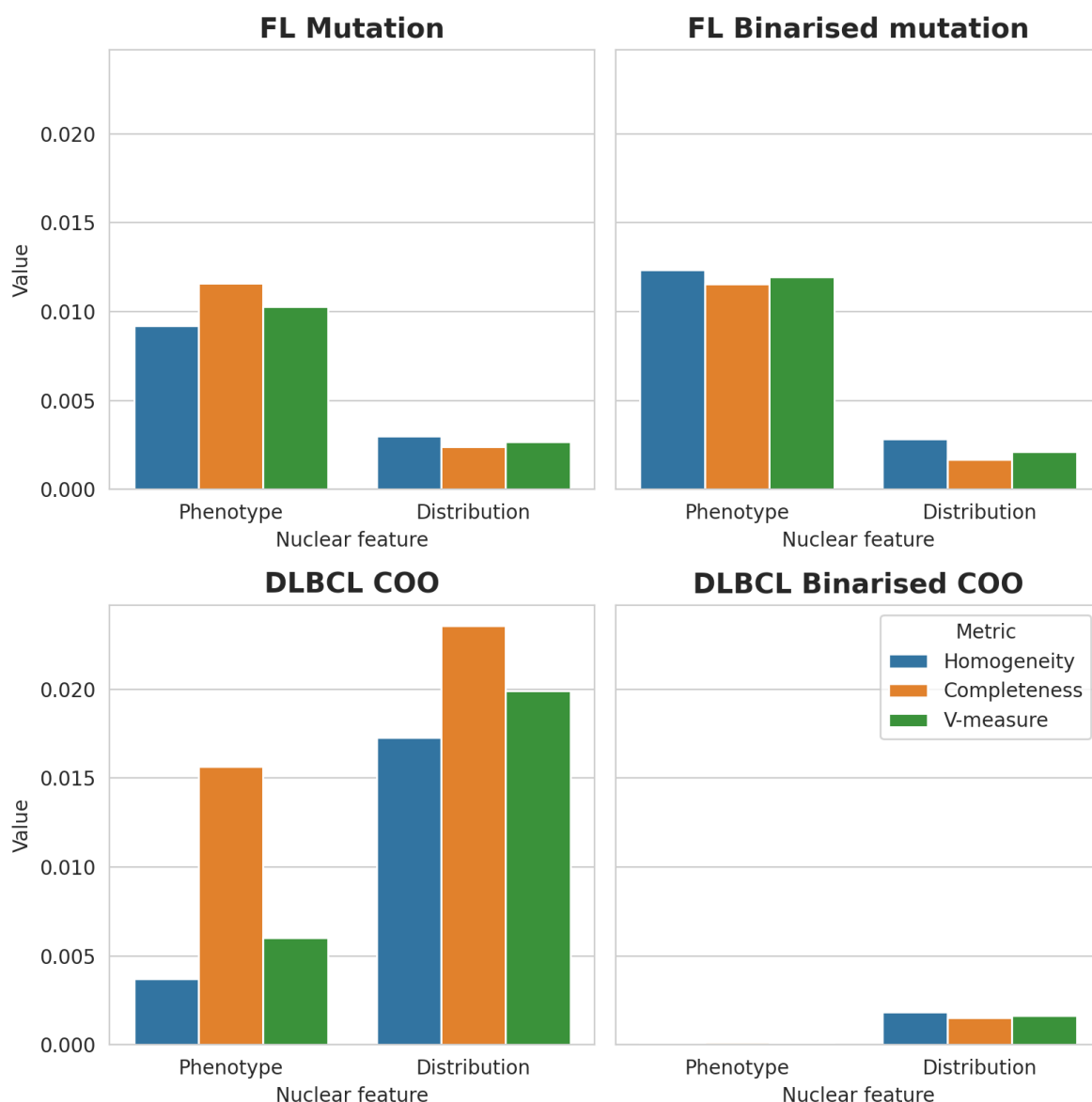


Figure 3.8: Evaluation of clustering methods using supervised metrics, with molecular subgroups as ground truth. In FL, mutation clusters were used (mutation[46]) and in DLBCL, cell-of-origin expression DLBCL (COO;[56, 144–146]). Calculated for samples with available molecular data. Two classification tasks were proposed:

Left column) Clustering against all molecular subgroups Right column) Clustering against binary subgroups, of the most prevalent molecular group vs. all other groups (rest)

For DLBCL, nuclear phenotypes were binarised to cluster 1 (centroblastic) vs. rest (2,3 and 4) as sample number in other classes was below 15 cases. Binarised groups: in FL, COM mutation vs. STAT6 & aSHM; in DLBCL: GCB expression vs. ABC & UNC & MHG.

In FL, binarisation of mutation clusters (COM vs. STAT6 & aSHM) led to modest in-

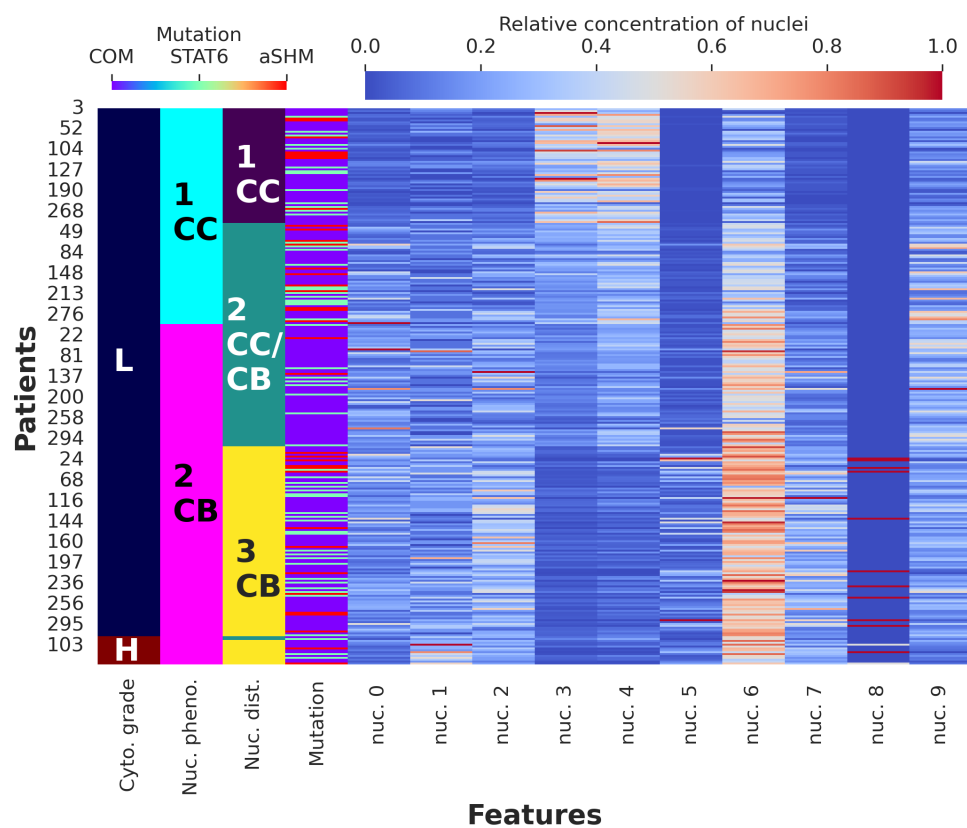
creases in homogeneity (and hence, V-measure; Figure 3.8 top left vs. right) for nuclear phenotype clusters and modest decreases in completeness for distribution groups. Nuclear phenotype clusters achieved higher performance than nuclear distribution clusters in all metrics.

In DLBCL, binarisation of COO (GCB vs. ABC & MHG & UNC) led to substantial decreases in all metric values, evidencing severe information loss. Binarisation was trialled with DLBCL nuclear distributions (distribution 2 vs 1 & 3) to assess performance when cluster number was the same between methods, but this lowered metric values further (not displayed).

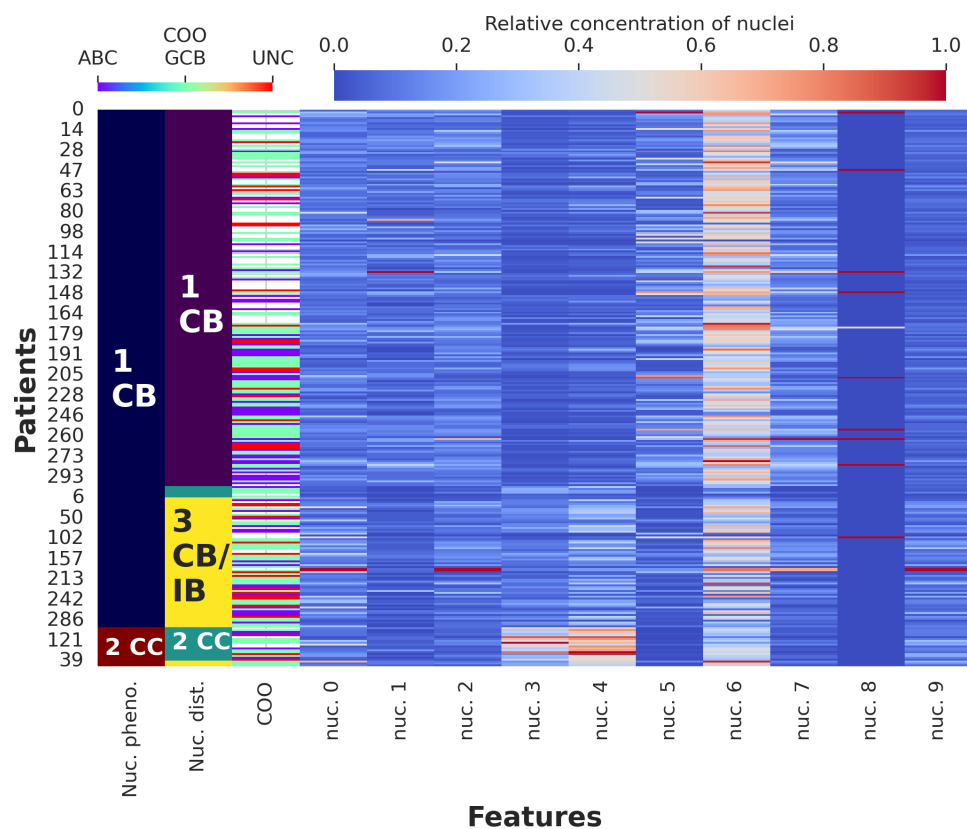
Nuclear distribution clusters outperformed phenotype clusters in all DLBCL metrics but especially in homogeneity. A major difference between methods was greater balance in DLBCL cluster sizes with distribution clusters, caused by distinction of patients with a diverse blastoid pattern of cells (cluster 3; CB/IB) from cluster 1, with a more homogeneous, centroblastic pattern.

Patient-level feature comparison

Relationships between cluster assignment, molecular groups (mutation clusters for FL and cell-of-origin expression clusters for DLBCL) and cell types involved can be observed in Figure 3.9, below.



(a) Follicular Lymphoma (FL)



(b) Diffuse Large B-cell Lymphoma clusters (DLBCL)

Figure 3.9: Bicluster plots of patient-level cell clusters, with respect to mutation (A; FL), expression (B; DLBCL) and relative proportion of discrete cell types. Assigned cluster numbers and associated cell types overlaid in the appropriate regions. Abbreviations: Cyto. grade = predicted cytological grade, using the methodology proposed in Chapter 2, as low (L) or high (H); Nuc. pheno. = Nuclear phenotype cluster; Nuc. dist. = Nuclear distribution cluster, with respect to centrocytic (CC), centroblastic (CB) and immunoblastic (IB) populations; COO = cell-of-origin expression; nuc. X = nuclei x, where x corresponds with the appropriate cells in Figure 3.3. Missing data masked using transparency.

In FL, all cases previously predicted to have high grade were assigned to centroblastic phenotype and distribution clusters. Centrocytic nuclear phenotype patients were observed to have low concentrations of nuclear class 6 (primarily centroblasts). Though there was considerable overlap between phenotype and distribution clusters associated with centrocytes (phenotype 1 and distribution 1) and centroblasts (phenotype 2 and distribution 3), distribution clusters were further enriched for nuclear classes 3 and 4 (mature lymphocytes/centrocytes) in the centrocytic group (1) and nuclear class 8 (light centrocytes) in the centroblastic group (3). Overall, this evidenced that nuclear phenotypes primarily distinguished patients on presence of centroblasts (nuclear class 6) while distribution clusters involved further cell types in clustering.

In DLBCL, nuclear phenotype and distribution clusters associated with centroblasts (phenotype 1 and distribution 1) and centrocytes (phenotype 2 and distribution 2) also overlapped, evidencing redundancy in clustering methods. Centrocytic clusters were observed to be enriched for nuclear classes 3 and 4 (mature lymphocytes/centrocytes) while displaying comparatively low concentrations of centroblasts (nuclear class 6). Distribution cluster 3 (centroblast/immunoblastic) was distinguished from cluster 1 (centroblastic) through moderately higher proportions of lymphocytes/centrocytes (nuclear classes 3 and 4). Otherwise, differences between these clusters appeared subtle.

4.5 Survival Analysis

Clinical utility of groups was assessed using comparisons of patient features (in FL) and survival.

Comparisons of patient features (such as performance status, involvement of extranodal sites and occurrence of B-symptoms) are presented in Appendix Figure A.4. In the HMRN dataset, effects were observed following multiple testing correction in the following:

- significantly lower number of ROIs assessed for nuclear distribution 1 as compared to the rest of clusters
- significantly lower numbers of nuclei counted per patient for distribution cluster 3, as compared to the rest of patients

The observation that distribution cluster 3 had significantly lower cell numbers per ROI may have been due to the larger size of the majority cell type (centroblasts), leading to lower number of cells filling the ROI. This finding therefore did not necessarily indicate unexpected properties of the group.

Clusters were compared with respect to patient overall survival using Cox model hazard ratios (HRs; Figure 3.10; full Cox model results included in Supplementary documents, Cox_model_results.xlsx). For FL, HRs were corrected with two existing gold standard measures of risk, FLIPI and PRIMA-PI. Kaplan-Meier estimators (KM plots) were used for visualisation of chemotherapy patient survival, which all datasets had (Figure 3.11, respectively). HRs displayed in KM plots were determined using univariate Cox models.

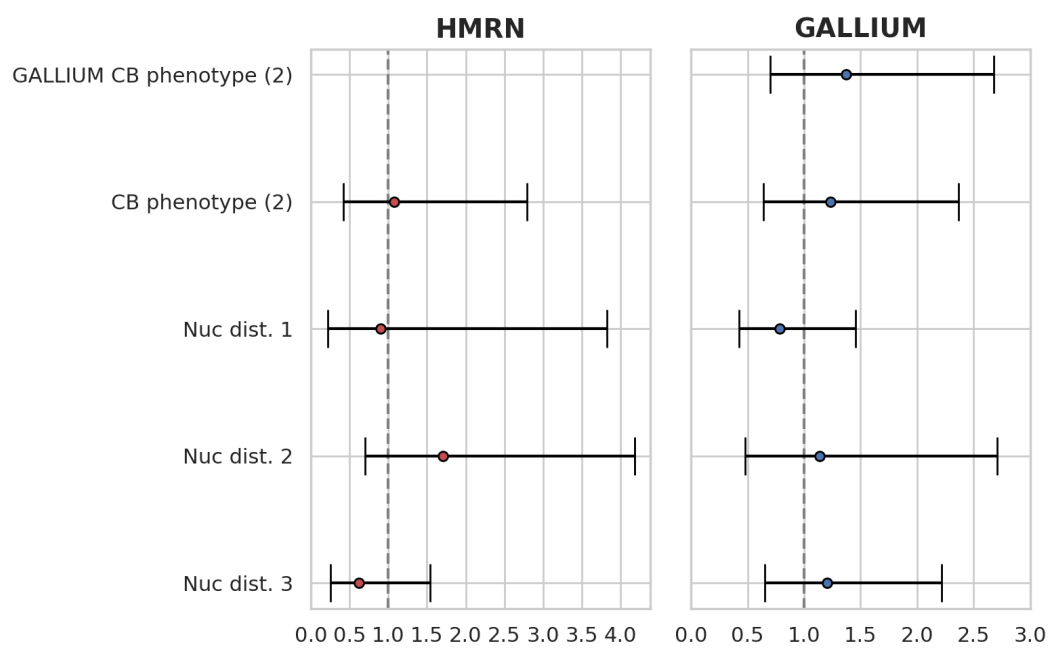
Follicular Lymphoma OS No significant effects were observed in FL survival, with all 95% confidence intervals overlapping a hazard ratio (HR) of 1. Yet, comparing effects between datasets revealed largely consistent HRs in immunochemotherapy patients. A centroblastic nuclear phenotype (CB phenotype 2) was associated with a slight increased risk in HMRN (HR CB (2) vs. CC (1) signature: 1.08; 95% CI: 0.42-2.79; $p = 0.88$) and greater risk in GALLIUM (HR CB (2) vs. CC (1) signature: 1.23; 95% CI: 0.64-2.37; $p = 0.53$). This risk was slightly higher still where the phenotype model was trained on

GALLIUM data (HR GALLIUM CB phenotype (2) vs. GALLIUM CC phenotype (1): 1.37; 95% CI: 0.70-2.68; $p = 0.36$). As expected, the centrocytic distribution cluster (1) displayed the opposite effect, with slight decreased risk.

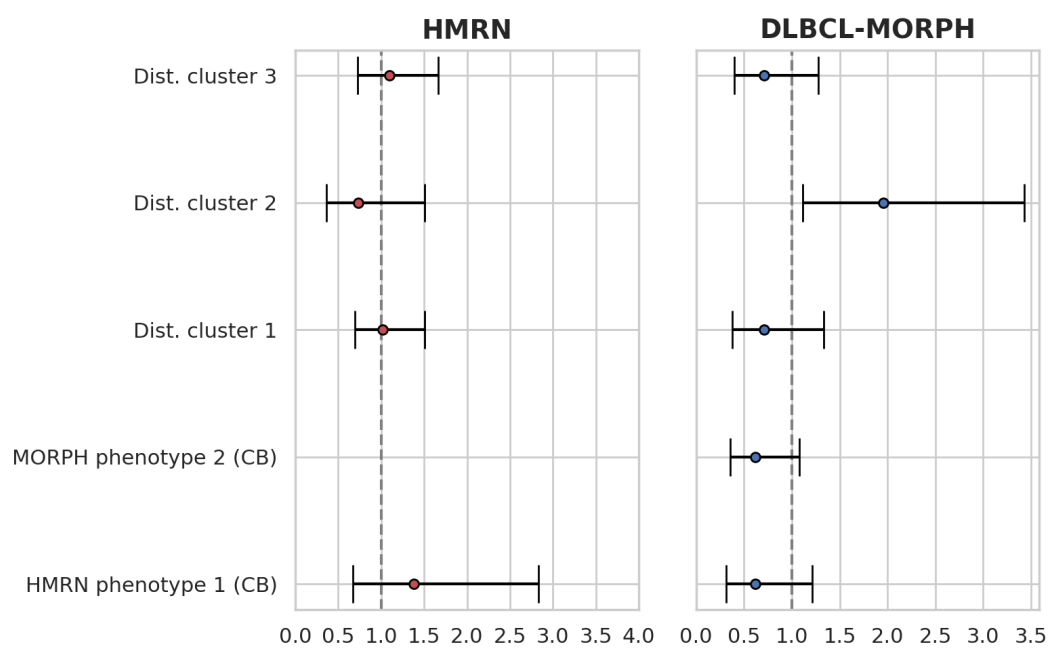
HRs were observed to be substantially higher in univariate Cox models (as displayed in KM plot titles; Figure 3.11) evidencing that effects were largely accounted for by clinical features considered by these indices (such as age, stage and blood β_2 -microglobulin concentration). Overall, these observations indicated potential small, general risk associated with patient clusters, not distinguishable with the available cohort sizes.

DLBCL OS No significant hazards were observed in HMRN DLBCL data. In DLBCL-MORPH, risk was found to be associated with centrocytic clusters: HR of 1.95 for distribution cluster 2 (95% CI: 1.11-3.43; $p = 0.02$). Centroblastic clusters were associated with reduced hazard, lowest in nuclear phenotypes fitted to DLBCL-MORPH data (HR MORPH phenotype 2 (CB) vs. 1 (CC/CB): 0.62; 95% CI: 0.35-1.08; $p = 0.09$). As can be seen in ‘at risk’ counts in KM plots, phenotype clusters were much more balanced in DLBCL-MORPH as compared to HMRN. This evidenced substantial differences in samples between these datasets. One potential cause of this may have been in confounding variables, such as staining or scanner variability between sites. Were systematic differences in digital pathology images the cause of differences observed in the datasets, cellular composition of nuclear phenotypes trained on each dataset would be expected to differ. They instead resembled corresponding HMRN phenotype compositions very closely (Appendix Figure A.8) evidencing that this was not the cause.

Another explanation of such differences is sampling error, since only 20-200 regions of interest out of many thousands (in some cases) were observed. Finally, it is possible that substantially different selection criteria existed for the two datasets given that HMRN was a population cohort while DLBCL-MORPH data underwent an extensive selection process for clinical trial.



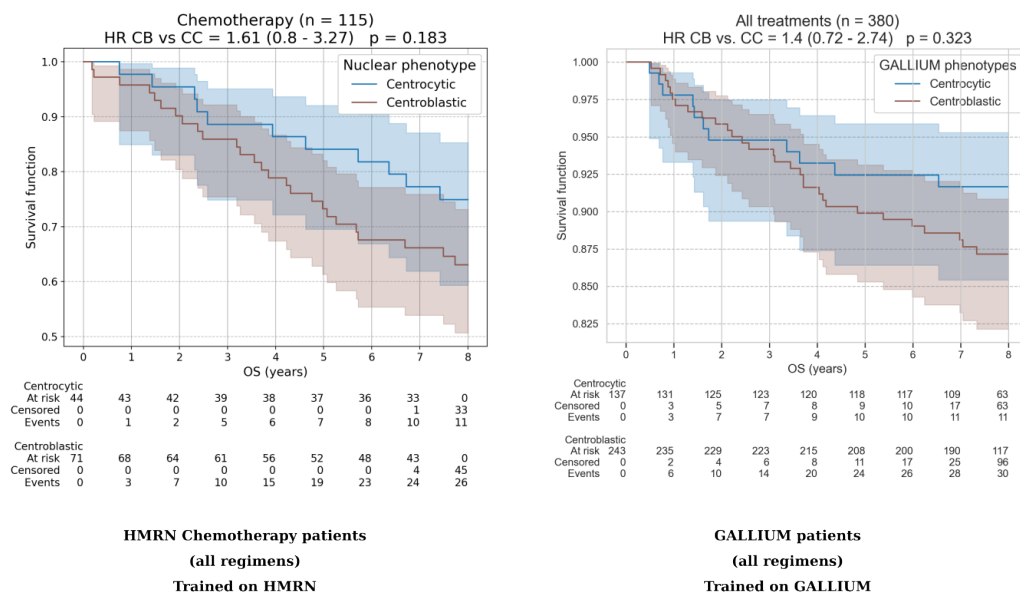
(a) Follicular Lymphoma (FL)
Immunochemotherapy



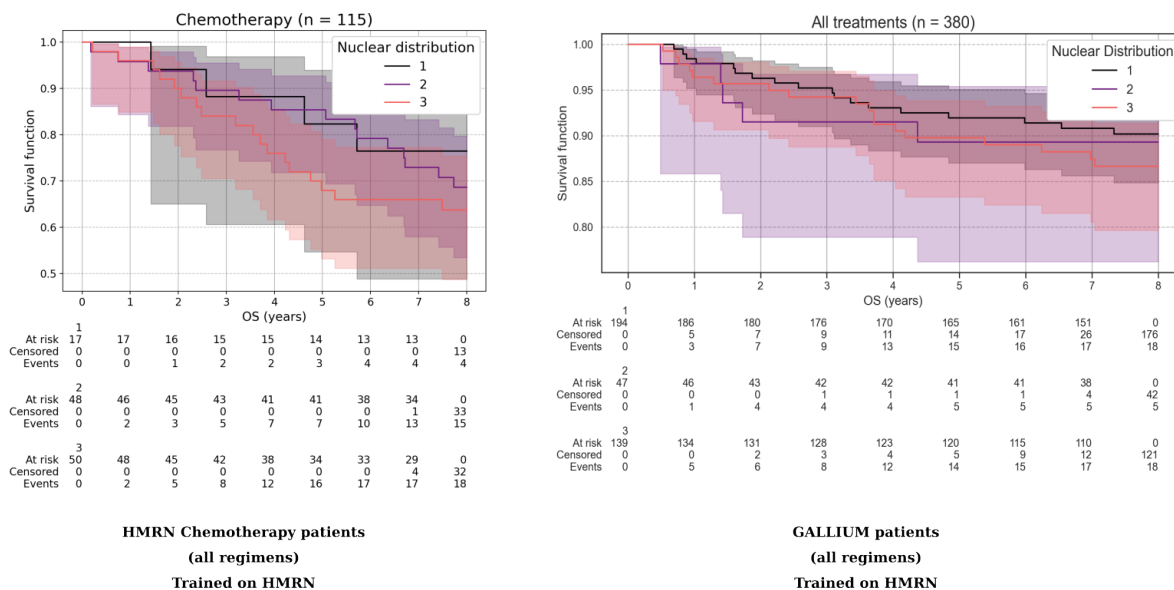
(b) Diffuse Large B-cell Lymphoma clusters (DLBCL)
All treatments

Figure 3.10: Cox model hazard ratios of patient clusters in HMRN and comparable external validation datasets. For FL, only HMRN results from the immunochemotherapy partition are shown for comparison against GALLIUM trial results. Error bars represent 95% confidence intervals. FL models trained in a multivariate configuration with PRIMA-PI quantile (low, medium, high) and FLIPI. DLBCL trained in a univariate configuration, due to high rates of missing IPI scores. Missing datapoints indicate insufficient data for fitting of Cox models.

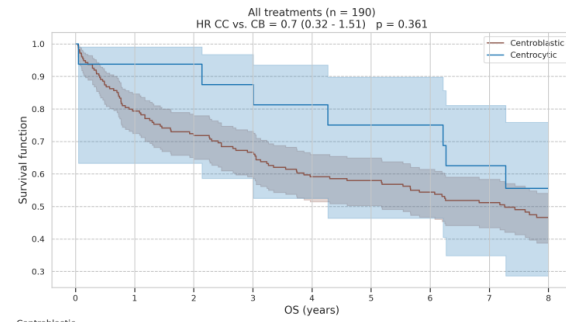
Nuclear phenotype



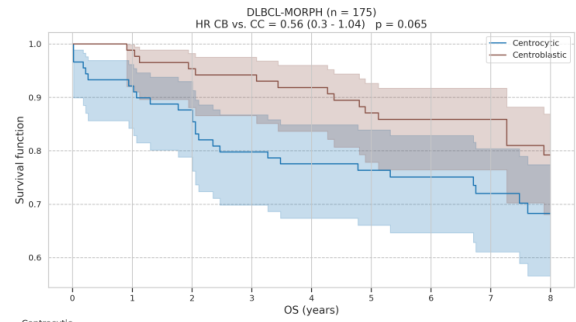
Nuclear distribution



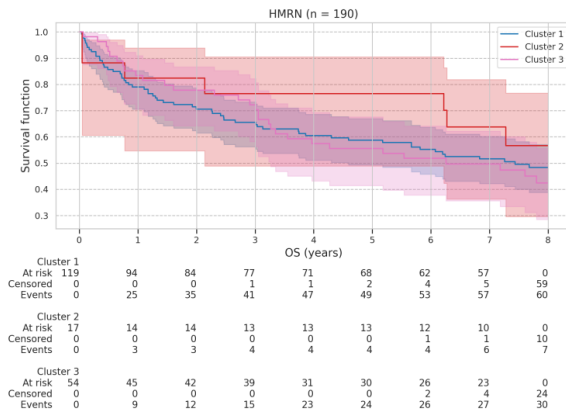
(a) Follicular Lymphoma (FL)

Nuclear phenotype

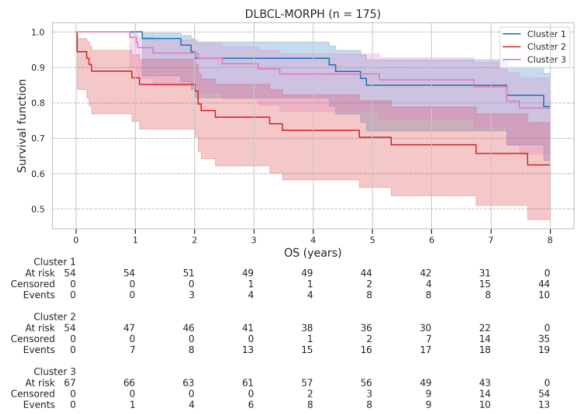
HMRN patients
(all regimens)
Trained on HMRN



DLBCL-MORPH patients
(all regimens)
Trained on DLBCL-MORPH

Nuclear distribution

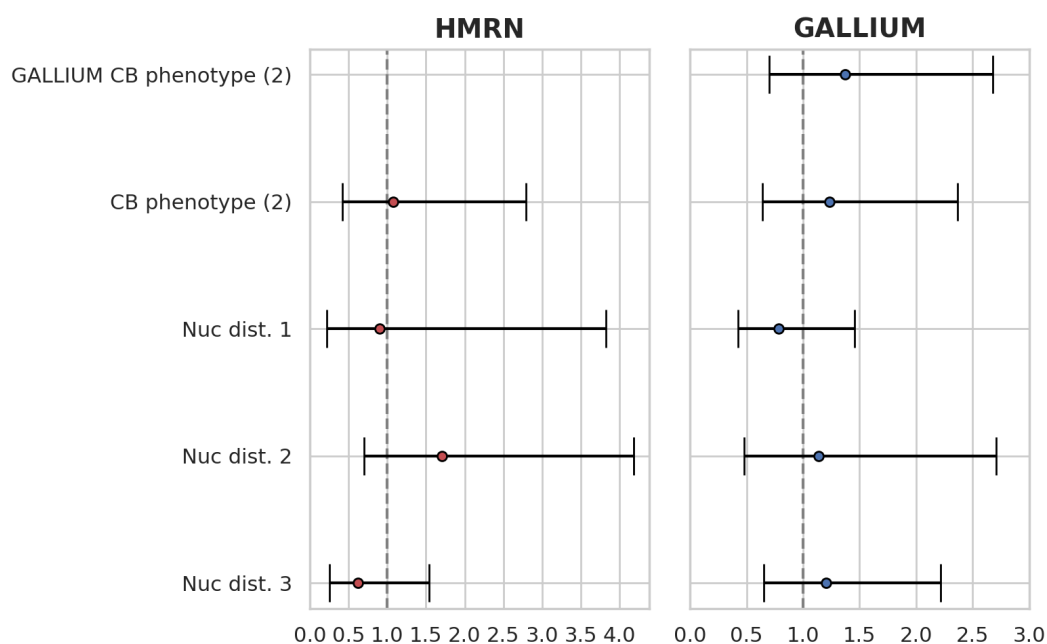
HMRN patients
(all regimens)
Trained on HMRN



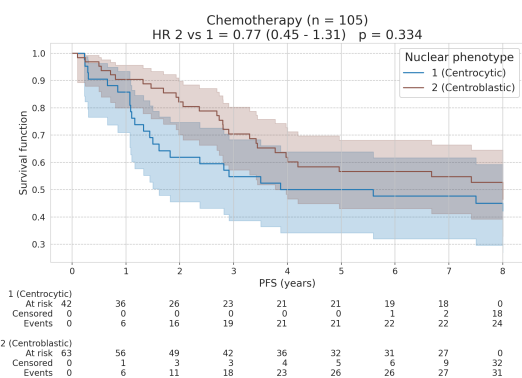
DLBCL-MORPH patients
(all regimens)
Trained on HMRN

(b) Diffuse Large B-cell Lymphoma clusters (DLBCL)

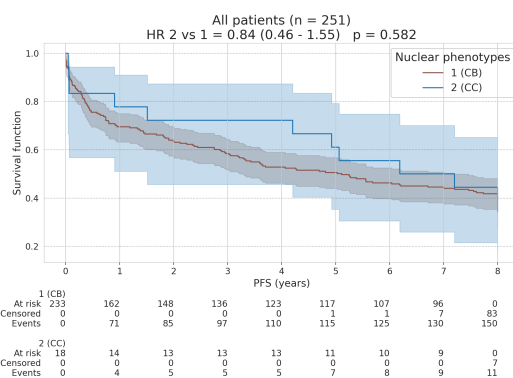
Figure 3.11: Kaplan-Meier plots of overall survival for patient clusters. Overall survival events defined as death by any cause. Number of OS and censoring events displayed below plots.



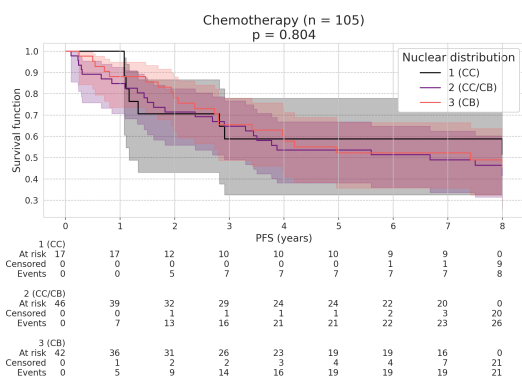
(a) Follicular Lymphoma (FL)
Immunochemotherapy



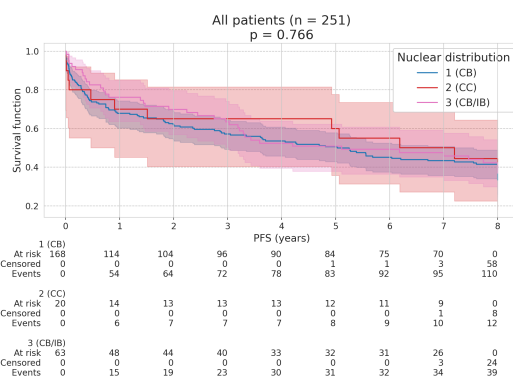
(a) Follicular Lymphoma
(FL)



(b) Diffuse Large B-cell Lymphoma clusters
(DLBCL)



(c) Follicular Lymphoma
(FL)



(d) Diffuse Large B-cell Lymphoma clusters
(DLBCL)

Figure 3.13: Kaplan-Meier plots of progression-free survival (PFS) for nuclear phenotype (top; a & b) and distribution (bottom; c & d) patient clusters. Number of PFS and censoring events displayed below plots. P-values from Wald test on Cox proportional hazards for nuclear phenotypes and multivariate logrank tests for distribution clusters.

FL PFS Nuclear phenotype clusters evidenced slight but insignificant reduced PFS hazard in patients with a centroblastic phenotype (HR. vs centrocytic = 0.77; $p = 0.334$; Figure 3.13a). These findings appeared to be a result of high event rate between years 1 and 2 in centrocytic patients. This differed to those for OS, in which hazard appeared higher for centroblastic patients. This effect evidenced potentially shorter overall survival but longer progression free survival in patients with a centroblastic phenotype. No PFS differences were observed in FL distribution clusters (Figure 3.13c).

DLBCL PFS PFS hazards mirrored OS in DLBCL, with apparent slight, insignificant reduced hazard with a centrocytic phenotype (HR vs. centroblastic = 0.84; $p = 0.582$; Figure 3.13b). No PFS differences were observed between distribution clusters (Figure 3.13d).

4.6 Overall discussion

Evaluation of clustering techniques

Clustering techniques were evaluated in two ways, focusing on distinction of:

1. Biological subgroups, defined as clustering of patients into groups with clear differences in cellular or molecular features. This was primarily measured using supervised clustering metrics using molecular subgroups as ground truth and further assessed using contingency matrices.
2. Distinction of clinically-relevant subgroups, with differences in survival. This was measured using hazard ratios and Kaplan-Meier estimators.

With respect to both mutation clusters and distinguishing overall survival risk, nuclear

phenotypes outperformed distribution clusters in FL. In this disease, cluster membership was more balanced (40% to centrocytic and 60% centroblastic).

In DLBCL, nuclear phenotype clustering led to highly imbalanced groups (91% centroblastic vs. 9% centrocytic/centroblastic), resulting in under-powered comparisons. In this situation, nuclear distribution clusters outperformed phenotypes in supervised metrics (with respect to COO expression) and survival (on DLBCL-MORPH data). Increased sensitivity to variation between patients that were otherwise indistinguishable with phenotype clusters was likely a result of the differing whole slide aggregation strategies of the clustering methods. Reliance on a single statistic (median number of each cell type) in nuclear phenotypes made it a computationally efficient technique but appeared to limit information required for effective clustering of a disease characterised by sheets of a single cell type, DLBCL. Distribution clusters, on the other hand, considered every cell observed. Though pairwise comparisons involving every cell were far less computationally efficient than clustering of median values, this process appears to have provided the necessary information to distinguish DLBCL patients.

Nuclear phenotype clusters had significantly different mutation profile composition in FL when comparing COM mutation against minor mutation profiles (STAT6 and aSHM). In DLBCL, potential differences were observed with modestly higher proportion of ABC expression in the centroblastic distribution cluster 1 (33% vs, 27-28% other). MHG expression was higher in the centrocytic cluster 2 (11% vs 0% cluster 3 and 6% cluster 1), albeit not significantly.

Finally, substantial and consistent survival differences were not observed between datasets. In FL, a small, consistent increased risk was observed with phenotype 2 (centroblastic) and nuclear distribution 1 (centroblast and immunoblastic) in immunochemotherapy patients. In DLBCL, no survival differences were observed in HMRN data but, clusters in this dataset were found to be highly imbalanced in size. In DLBCL-MORPH, the centrocytic distribution cluster 2 was associated with considerable increased risk (HR vs rest: 1.95; 95% CI: 1.11-3.43; $p = 0.02$). Centrocytic phenotype cluster 1 also demonstrated increased risk, albeit to a lesser extent (HR vs. cluster 2: 1.61; 95% CI: 0.93 - 2.86; $p =$

0.09).

Overall, these results evidenced that nuclear phenotypes separated FL patients on molecular features but not too well on risk. Distribution clusters appeared to separate DLBCL-MORPH patients on survival risk. Imbalance in HMRN DLBCL data and low incidence of MHG expression limited statistical power of comparisons and clustering efforts for that dataset.

Clustering of FL

The clustering methods presented identified two major groups of FL, with respect to cell types. A third cluster was evidenced with nuclear distribution clusters but, significantly lower ROI numbers per patient sample and similarities in cell types present evidenced that this group was not functionally dissimilar to the others. Instead, it was likely to have formed from low sample number, leading to poor estimations of the nuclear distribution and frequency estimations needed for clustering. Evidence for this was significantly lower number of observed regions with nuclear phenotype cluster 1 and similarities in cell types present between clusters 1&2 and 2&3.

40% of cases (nuclear phenotype cluster 1) were enriched for dark, cleaved centrocytes and had significantly higher incidence of infrequent mutation profiles (STAT6 & aSHM). 60% of cases (phenotype 2) had higher proportions of centroblasts, characterised by large nuclei bearing multiple nucleoli close to the nuclear envelope. This group had significantly higher incidence of the COM mutation profile, characterised by mutation of the KMT2D gene[46] and thus, associated with persistence of B-cell structures in lymph nodes [160].

In the most recent WHO classifications of FL, the disease was divided into 3 subtypes based on cell morphology and molecular characteristics (WHO-HAEM5; [23, 161, 162]). Similarities appeared to exist between these subgroups and those identified in this work:

Classical FL (cFL) characterised by canonical follicles, presence of centrocytes and centroblasts as well as significantly higher mutation targeting genes KMT2D[159] and specific regions of CREBBP[48], both encoding histone modifying proteins. In mutation

of KMT2D, this subtype resembled the COM mutation profile[46]. FL phenotype 2 and distribution 3 most closely resembled this group in mutation, with higher proportions of COM. In cellular composition, phenotype 1 more closely fitted the expected proportion of centrocytes than phenotype 2.

Follicular large B-cell Lymphoma (FLBL) characterised by homogenous regions of centroblasts (cytological grade 3B FL[161]) and treated like DLBCL. Such cases have previously been evidenced to have significantly lower proportions of the hallmark FL BCL2 mutation (t(14;18)) and significantly higher mutation of BCL6, with respect to cFL. Recent work identified DLBCL-like FL, with significantly higher mutation burden in regions associated with aberrant somatic hypermutation (aSHM)[48]. Centroblast-dominant groups identified, FL phenotype 2 and distribution cluster 3, closely fitted the cellular characteristics of FLBL, with low proportions of centrocytes. Neither cluster was found to have increased aSHM mutation, as expected from the literature.

FL with uncommon features (uFL) currently encompasses cases not classified into the two former categories due to uncommon cellular or follicular morphology[23, 161, 162]. In cellularity, this constituted cases with enrichment of non-centroblast blastoid cells (such as immunoblasts) and medium-sized cells resembling neoplastic Burkitt Lymphoma cells[161, 162]. Molecular details of blastoid uFL were not found in the literature.

In terms of follicularity, uFL included the so-called ‘diffuse pattern’. Here, follicles present back-to-back and cover large areas of the lymph node. These cases have been found to contain a mixture of neoplastic centrocytes and centroblasts, with no consistency in proportions[163]. They have been found to lack the hallmark FL BCL2 mutation (t(14;18)) and have significantly higher mutation of STAT6, as compared to cFL[157–159]. STAT6 mutation has been found to accompany a lack of the hallmark BCL2 mutation[164], observed in over 2/3rds of cases with STAT6 mutation. The relevance of these observations in the current work is that a majority of patients with a STAT6 mutation profile in the may have lacked typical FL molecular characteristics.

In summary, cases of uFL with clearly abnormal blastoid features have been reported

but a growing body of work has evidenced cases similar to cFL in cellularity, harboring mutations uncharacteristic of FL.

Given the range of cellular features considered in uFL, STAT6 mutation was the most reliable measure of centrocytic uFL while considerable non-centroblast blastoid populations were used to indicate blastoid uFL. Distribution cluster 2 (Figure 3.5 middle) contained considerable populations of blastoid cells (belonging to cell cluster 2 and UMAP component values close to -13). These cases appeared to match some of the classification guidelines for blastoid uFL, with cytology similar to published work[157, 162].

Phenotype 1 (centrocytic) nonetheless had higher proportions of STAT6/aSHM mutation, suggesting presence of uFL resembling classical FL in that group. Analyses of follicularity were not conducted in these groups but may aid in future distinction of diffuse uFL with STAT6 mutation and cFL.

Observed relationships between clusters, cellular features and molecular features are presented in Figure 3.14a, below.

Clustering of DLBCL

DLBCL not otherwise specified (DLBCL, NOS) comprises diseases sharing some morphological similarity but otherwise displaying immense diversity in mutation, expression and cell type presence[21, 23]. Both clustering methods identified a majority group composed of centroblasts, matching the unifying characteristic of DLBCL cases. This may have also been a result of nuclear features used for cell clustering, previously found to have utility in centroblast classification (Chapter 2 Section 3.8). Number of cases clustered into other groups (other than the majority) was ultimately too low in HMRN data to determine whether observed differences in centrocyte, mature lymphocyte or blastoid cell populations represented morphological subtypes of DLBCL with respect to survival or expression. Survival differences were previously reported in DLBCL-MORPH data with respect to ‘neoplastic’ cell morphology[66]. The nature of cells classified as neoplastic by the general nuclear classifier used, HoVerNet[103], was not investigated further but could be seen to include centrocytes and centroblasts in the publication (Figure 1c[66]). In the

clusters proposed in our work, we identified significantly increased overall survival risk with centrocytic phenotypes (in DLBCL-MORPH), potentially elucidating the nature of cellular markers of risk presented by Vrabac et al[66].

Uniformity in centroblast populations has been historically associated with the GCB cell-of-origin (COO). Expression largely resembles naive B-cells (centroblasts), with centrocytic expression in some cases, too[20]. MHG expression also indicates centroblastic disease, but without centrocyte involvement[20]. Finally, immunoblastic and T-cell rich cases have been found to have significantly higher ABC expression[165], supported by expression of activated B-cell markers (from the immunoblast and plasmablast stages of activation)[19, 20].

Nuclear phenotype 1 and distribution 1 appeared to fit descriptions of classical monomorphic DLBCL, expected to be enriched for GCB and MHG expression.

DLBCL phenotype 1 and distribution clusters 1 and 3 appeared to contain considerable populations of blastoid cells, matching cellular classification criteria for immunoblastic DLBCL with expected enrichment for ABC expression. Though potential enrichment for ABC was present in phenotype cluster 1 (31% of cases vs. 24% of cases in cluster 2), differences in nuclear distribution cluster ABC expression were more subtle. Higher proportions of GCB patients in nuclear phenotype and distribution cluster 2 (centrocytic) was not expected given that MHG expression has not been observed to involve centrocyte-associated genes. No statistically significant differences were observed at the HMRN dataset size used.

Immunoblast cell morphology has previously outperformed predicted cell-of-origin in stratifying patients on survival risk[67]. No comparable survival differences were observed in immunoblastic groups of either DLBCL dataset. Observed relationships between clusters, cellular features and molecular features are presented in Figure 3.14b, below.

4.7 Limitations

Limitations of the proposed methods potentially existed in the features used for cell clustering, nuclear embeddings in nuclear distribution clustering and methods of slide-level aggregation.

Omitted features

Despite high interpretability of nuclear features used, they were not sufficient for clustering of distinct cell types, which complicated interpretation of cell clusters. Recent publications in nuclear classification have used additional features to those used in this work. Success has been demonstrated in use of additional colour spaces, especially CIELAB and HSV[129, 139]. These alternative colour spaces differ from the RGB intensity features used here in the relationships between colours, with aims of colour space uniformity and progression more similar to the human eye for CIELAB and HSV, respectively. Comparison of 2-dimensional UMAP projections of cell features including features from those colour spaces[139] and ours revealed almost identical visualisations (Figure 3.3). Additionally, CIELAB and HSV intensity features (moments, maximum intensity, minimum intensity etc.) were found to correlate very highly with the respective RGB features in the unlabelled nuclei dataset (Pearson correlation > 0.95 ; not displayed). Therefore, features from alternative colour spaces did not appear to meaningfully increase the range of information extracted for each cell.

On the other, omission of contextual information may have limited efforts. By not considering surrounding cells, an assumption was made that function of cells was independent of their microenvironment and tissue region. As highlighted in Introduction section 2.1, this assumption appeared to be invalid in at least the context of FL.

Strategies that have demonstrated success in prediction of clinically useful features with integration of H&E cell context have included calculation of cell graphs to model interaction between cell neighbourhoods[166] and combination of region information with nuclear classifications[167].

Incorporation of contextual information about cells would likely allow for better distinction of molecular groups in FL and DLBCL, where cell density alone clustered patients with seemingly different molecular features (such as ABC and GCB expression in DLBCL) together. In DLBCL especially, such information may better resolve the majority centroblastic phenotype into more homogenous groups of balanced size.

Nuclear distribution clusters

Several limitations of the distribution cluster methodology were identified. Firstly, comparison of distributions was computationally expensive, with exponential growth in number of pairwise comparisons required as slide number was increased. Meanwhile, nuclear distribution clusters only noticeably differed at four points in the 1-dimensional UMAP embedding. Comparison of relative frequencies at these four cell types may have been sufficient, and far more computationally efficient for the clustering observed.

Secondly, the 1-dimensional UMAP embedding displayed a clear progression of cell types, with preservation of ‘distance’ evidenced by nuclei more similar in size, nucleoli number and colour intensity located more closely than dissimilar nuclei. Despite this, and most visibly in size, non-trivial variation did exist across the embedding, meaning that cell types at each point in the embedding were not perfectly homogenous. The impact of this on downstream clustering was not investigated, but it may conceivably have led to higher noise and complicated interpretation of clusters. A greater number of dimensions may have been required for improved distinction of cell types, though an alternative approach to measuring distance would also be required.

Thirdly, the validity of statistical measures of distance (Anderson-Darling) on a non-linear embedding (UMAP) was not clear. Comparison against linear feature reduction methods, such as NMF and PCA, would be required to determine the validity of results from statistical measures of distance.

Fourthly, as observed with significant differences in ROIs extracted per patients, this method was found to be sensitive to sampling error. Threshold ROI numbers were used

to filter for slides with sufficient tissue (and thus, cell number) for a representative number to be used. Fixed sample number for ROIs and cells in calculation of nuclear distributions may reduce the impact of confounding variables stemming from uneven sampling.

In summary, increase in dimension may have led to improved cell homogeneity in regions of the embedding space; reductions in scope to informative cell types may have led to improved efficiency in use of computational resources and finally, linear methods of dimensionality reduction may have had greater statistical validity for comparison of patients.

Slide level aggregation

Evidence exists that optimal methods for aggregation of information from the ROI level to a whole slide classification is task-dependent[168].

In the literature, computational methods for lymphoma have focused either on average cell densities in clinically-determined regions of slides[63] or on all observed regions[66]. The presented work followed the former strategy. Observations were only made in naive B-cell regions, following clinical recommendations in assessment of morphology and molecular features in tumour cell regions[23], to reduce potential noise from uninformative regions. The disadvantage of such an approach is potential loss of information from other areas, not yet accepted as clinically relevant.

Data for ROIs was aggregated to a slide level classification in different ways for the two clustering methods presented: median frequency of cells was used for nuclear phenotypes and a distribution of all cells in distribution clusters. Despite these differences, both techniques bore similarities in the assumed independence of cells and their context as well as equal importance to all regions. Aggregation methods operating on similar assumptions have been used to successfully stratify patients on risk in lymphoma[66] and wider cancers[139] despite the existence of aggregation methods in which such assumptions are not made[168]. Aggregation of features similar to those used in this work (cell type and morphology) using graphs has been shown to outperform other gold standard approaches in prediction of molecular features in breast cancer[85]. Such an approach may

not provide meaningful benefits over average and standard deviations of nuclear features in DLBCL[66], where tumour cells form largely homogenous sheets. On the other hand, the diversity of cell types present and heterogeneity in follicle (naive B-cell region), vascular and mature lymphocyte structures observed in FL mean that embedding of nuclear information in graphs may lead to improved overall distinction of patient subgroups.

In summary, best practice for aggregation of ROI information to the slide level appears to be disease-specific. Previous efforts have used summary statistics of clinically-derived regions in FL or random sampling in DLBCL. The work presented followed a clinically-derived methodology for region selection, potentially missing information in other areas and making an assumption of equal importance for all regions. Graph-based methods of nuclear feature aggregation have demonstrated success in prediction of molecular features in cancer. Given heterogeneity in cellular and tissue features in FL, aggregation of nuclear features using graph approaches may aid in appropriate weighting of informative features while also considering a larger portion of the slide. The value of such methods in DLBCL is unclear though, methods of distinguishing patients by features beyond cell density appear to be necessary for stratification of risk in population datasets such as HMRN.

5 Conclusions and future work

5.1 Patient groups

Patients diagnosed with follicular (FL) and diffuse large B-cell lymphoma (DLBCL) were clustered on thousands of unlabelled cells sampled from B-cell regions of lymph nodes.

In FL, two patient groups were identified (features summarised in Figure 3.14a). One was characterised by a centrocytic pattern of cells, associated with STAT6/aSHM mutation ($p = 0.03$). The other cluster demonstrated higher proportions of centroblasts, associated with the COM mutation profile. A slight increased risk was observed with a centroblastic phenotype in patients receiving immunochemotherapy (HRs of 1.08 and 1.37 vs. centrocytic phenotypes for HMRN and GALLIUM trained phenotypes, respec-

tively; following FLIPI and PRIMA-PI correction in HMRN and GALLIUM; $p = 0.88$ and 0.36 , respectively).

In DLBCL, centroblastic, centrocytic and blastoid patient clusters were identified (features summarised in Figure 3.14b). In DLBCL-MORPH data, patients in the centrocytic nuclear distribution cluster experienced substantially higher risk (HR vs rest = 1.95; $p = 0.02$). HMRN data evidenced potential differences in ABC and MHG expression between the centrocytic and other groups, though this observation did not fit with evidence from expression studies[20].

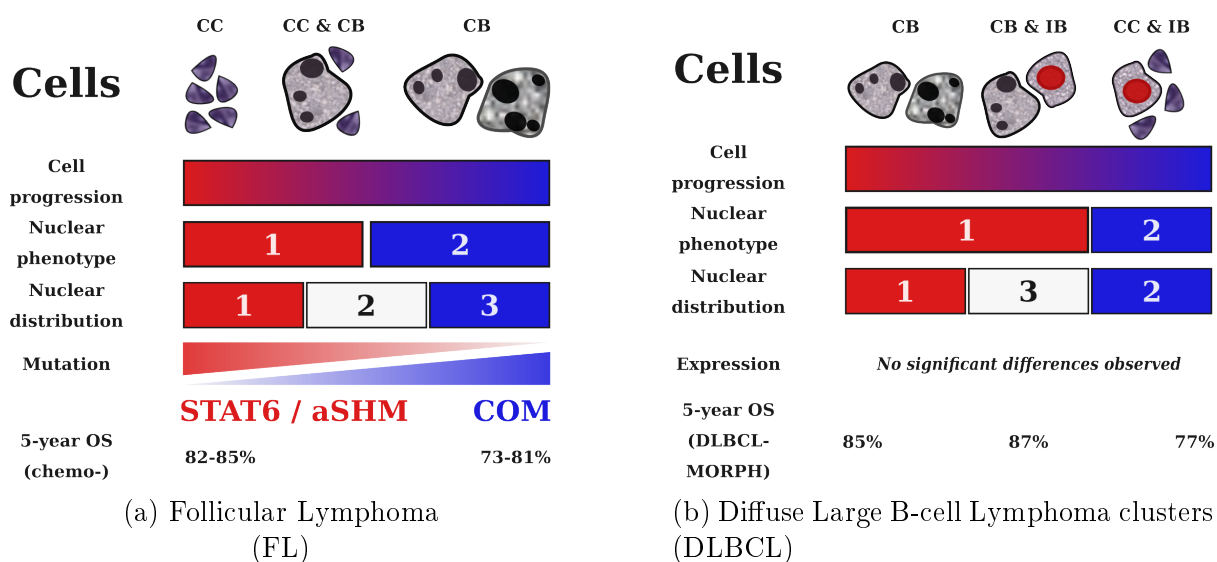


Figure 3.14: Proposed relationships between cell patterns, nuclear clusters and molecular features (mutation profile in FL and cell-of-origin expression in DLBCL). Cell type abbreviations: CC = centrocyte; CB = centroblast; IB = immunoblast

5.2 Clustering methods

This study was the first to combine objective, computational evaluation of cell types in DLBCL and FL H&E slides with molecular features such as expression and mutation profile. Over 43 million cells were considered, with identification of 2 major cellular phenotypes in both diseases and corresponding differences in molecular features.

Nuclear phenotype clusters outperformed distribution clusters in grouping patients on clinically-relevant features in FL, in which cellular heterogeneity is a classical feature. In

DLBCL, a condition characterised by sheets of a single cell type, distribution clusters appeared to better resolve patient clusters.

With these strengths in mind, optimal methods for clustering in other lymphomas may be:

- Distribution clusters in Burkitt Lymphoma, where neoplastic cells form sheets (like in DLBCL) with interspersed (tingible body) macrophages
- Phenotype clusters in Hodgkin Lymphoma, where neoplastic Reed-Sternberg cells classically form extensive interaction networks with immune cells

5.3 Summary of Findings

1. DLBCL patients with centrocytic cellularity experienced almost two times greater overall survival hazard, with respect to all other patients. Centrocytic patients may have also had lower proportions of ABC and higher MHG expression, though this did not agree with expression studies and requires validation
2. 60% of FL patients presented with a centroblastic cell phenotype, associated with common FL mutation (COM) and slightly increased risk, with respect to a centrocytic phenotype (hazard ratios between 1.08 and 1.37)

5.4 Limitations and future work

With respect to the aims, this work successfully clustered patients on clinically-relevant molecular features in FL. This was not achieved in DLBCL, possibly due to nuclear feature choice that did not effectively capture relevant differences between patients or inherent imbalance in the HMRN dataset. To improve upon this, future work can firstly utilise a greater range of morphological or colour space features to better resolve differences in DLBCL blastoid cell populations.

Future work may also utilise spatial information, with high biological relevance in lymphomas like FL and Hodgkin Lymphoma, where interaction of neoplastic and immune cells have been demonstrated to play a crucial part in pathogenesis.

Contributions

Labelling of the pathologist annotated dataset of nuclei was performed by Dr. Alireza Behzadnia of the Leeds Teaching Hospitals NHS Trust (LTHT). Professors David Westhead (UoL) and Andrew Janowczyk (Geneva University Hospitals, Switzerland and Emory University, USA) provided feedback on methodology throughout. All other work was conducted by Volodymyr Chapman (UoL).

Chapter 4

Immunohistochemistry clusters

1 Abstract

Immunohistochemical (IHC) analyses are a crucial part of diagnosis and subtyping in non-Hodgkin lymphoma but incur additional costs and time burden. We investigate utility of automated IHC analysis, patient subgrouping and adding value to routine H&E with prediction of IHC profile.

Over 6,000 IHC whole slide images (WSIs) of the HMRN dataset, from 330 Diffuse Large B-cell Lymphoma (DLBCL) and 185 Follicular Lymphoma (FL) biopsy samples were analysed. Macenko stain separation and DeepLIIF were used to quantify IHC. Samples were clustered into two groups within each disease. SlideGraph models were trained to predict these with corresponding H&E nuclear and image feature embeddings (UNI). Finally, patient characteristics were validated using routine H&E slides (DLBCL-MORPH $n = 190$; HMRN FL $n = 218$; HMRN DLBCL $n = 178$).

Predictions of IHC staining were found to correlate strongly to pathologist assessments of CD10 (mean Pearson $r = 0.7$ - 0.8) and moderately for BCL6 and MUM1 ($r = 0.4$ - 0.7). FL and DLBCL patients each clustered into 2 groups. FL cluster 2 was found to associate with increased COM mutation ($p = 0.275$ training / 0.035 validation) and slight reduction in overall survival risk (HR vs. cluster 1: 0.75 ; 95% CI: 0.43 - 1.31 ; $p = 0.31$). Image features successfully predicted clusters in H&E, with mean AUC-ROC between 0.7 and 0.8 (IHC quantified using DeepLIIF). Moderate to high performance in quantification of IHC and prediction of clusters with H&E provide promise of accurate automation of IHC analyses and extraction of added value in assessment of routine H&E.

2 Introduction

2.1 Value of IHC in Lymphoma

Haematoxylin and eosin (H&E) is a routine stain combination used in pathological analyses, visualising nuclear material (DNA/RNA) and proteins, respectively. While useful in diagnosis and subgrouping of disease based on distribution of cells and protein structures (e.g. extracellular matrix), distinction of cells is limited to those visually dissimilar in morphology, chromatin/protein distribution and in cell environment. Low agreement between expert pathologists[59, 61, 63] evidence limitations in reliable diagnosis and prognosis, at a cellular resolution.

Immunohistochemical (IHC) analysis involves targeting of specific proteins using enzyme-conjugated antibodies and subsequent staining of those regions of pathology samples. 3,3'-diaminobenzidine (DAB) is commonly used, undergoing oxidation on contact with enzyme-conjugated antibodies to form a brown/black precipitate. This process provides information on presence and distribution of clinically-relevant cells. It is highly valuable in classification of lymphomas, cancers of the lymphatic system, involving cells otherwise immensely challenging (if not, impossible) to distinguish with H&E staining. These include B- and T-lymphocytes (B- and T-cells), as well as macrophages, dendritic cells and centroblasts.

Key evidence for the utility of IHC in lymphoma include:

- Reduced staining for the naive B-cell marker CD10 and apoptosis escape protein BCL2 with high cytological grade in Follicular Lymphoma[130]
- Significantly greater odds of malignancy in bone marrow samples displaying B-cell phenotype over a T-cell phenotype[169]
- Hans algorithm in DLBCL, for classifying those with germinal centre-like expression[25]
- Recommended for differential diagnosis of classical and other Hodgkin Lymphoma subgroups[170]

- Prediction of MYC rearrangement, a hallmark of aggressive lymphomas such as Burkitt’s and double-hit (DHIT) DLBCL, in large cell non-Hodgkin lymphomas [171]

2.2 Limitations of IHC

While facilitating cell subtype-level pathological analyses, IHC also has considerable limitations. Firstly, reproducibility can be low with significantly different results between antibody clones used[69]; level of expert training in use of specific antibodies[71] and even staining apparatus[70]. These results evidence potential subjectivity in the thresholds used to classify cells as ‘positive’ for target proteins. This likely has limited impact in the clinic, where antibodies are typically only selected for use following rigorous validation on accuracy, robustness and consistency. Nonetheless, situations may exist where such stringent checks are not employed, such as in research settings.

Manual IHC quantification may also be vulnerable to sampling errors, as in H&E[61], as there are limits to the number of cells that can feasibly be evaluated of the millions present in biopsy samples. Poor reproducibility and restrictions in feasibility may limit the complexity of relationships that can be identified in IHC, with current algorithms limited to four stains in binary classification trees[25, 171, 172].

Besides these limitations, manual assessment is laborious for a population of highly skilled staff, histopathologists, whose numbers do not meet demand[2], in a time of growing waiting lists for diagnostic tests[3]. Despite evidenced clinical utility in quantification of specific cell types, such as T-cell and macrophage subtypes in Follicular Lymphoma[65, 135], additional cost, time burden and reproducibility concerns may prevent routine IHC analyses for these cells.

Given the above challenges, there is an unmet need for consistent and reproducible analysis of IHC in non-Hodgkin Lymphoma that does not increase the high existing histopathologist burden. Additionally, prediction of clinically-relevant features in H&E otherwise requiring protein-level analyses (such as IHC) may add considerable diagnostic

value where IHC is limited by resource restrictions.

2.3 Recent advances in IHC

Recent efforts have demonstrated promise in automation of IHC analyses, with clinical utility. These include:

- A convolutional neural network (CNN) trained to predict PD-L1 presence (IHC status) in breast cancer H&E slides, achieving area under the receiver-operating curve (AUC-ROC) scores of 0.92 on test data[173]
- A graph neural network (GNN) trained to predict HER2 involvement in breast cancer H&E slides, achieving AUC-ROC scores between 0.7-0.8 on test data[85]
- A CNN used to estimate staining (H-score) for AKT3 in pancreatic cancer and HER2 in breast cancer, achieving Pearson correlation against pathologist estimations of 0.78-0.98[174]
- A conditional generative adversarial network (cGAN) for creation of predicted IHC slides, with moderate agreement between real and synthetic Ki67 IHC data positive cell quantification (R^2 score of 0.56)[175].

The success of these efforts evidences potential for automated IHC quantification in non-Hodgkin Lymphoma and prediction of IHC features using routine H&E slides.

2.4 Aims

Key aims of the work presented were:

1. To automate IHC quantification in common lymphoma markers, with evaluation against pathologist estimates of CD10, BCL6 and MUM1
2. To cluster FL and DLBCL patients on IHC features, evaluated against clinical utility: mutation profiles, expression profiles and overall survival
3. To extend prediction of IHC clusters to routine H&E slides through training of deep

learning models on nuclear and self-supervised image features

4. To validate utility of predictions made on H&E slides, including external validation data, through comparison of clinically-relevant features with training data

2.5 Novel Contributions

Key novel contributions of the presented work include:

- Evaluation of automated IHC quantification methods against pathologist estimates of CD10, BCL6 and MUM1 staining
- Clustering of FL and DLBCL patients into IHC subgroups
- Comparison of overall survival and molecular features (mutation and expression profiles in FL and DLBCL, respectively) between IHC subgroups
- Prediction of IHC clusters using routine H&E slides in FL and DLBCL

3 Methods

3.1 Experimental workflow

Three major activities were undertaken in the presented work, as displayed in Figure 4.1, below:

1. First, immunohistochemical staining was quantified. This was performed using the DeepLIIF framework[176] at an individual cell level. A ‘basic’ slide level stain deconvolution method was used as a benchmark. Methods were evaluated against pathologist-assigned whole slide labels for three IHC markers in DLBCL.
2. DeepLIIF stain predictions were used to cluster patients in DLBCL and FL, to identify patterns of IHC staining. Clusters were compared to molecular features and overall survival to assess clinical utility.
3. IHC clusters were predicted using H&E. Two complementary strategies were evalu-

ated for patch-level feature extraction. These were nuclear features, including those describing morphology, intensity and texture and abstract foundation model features[82]. For slide level aggregation, a graph neural network incorporating spatial relationships[85] was used.

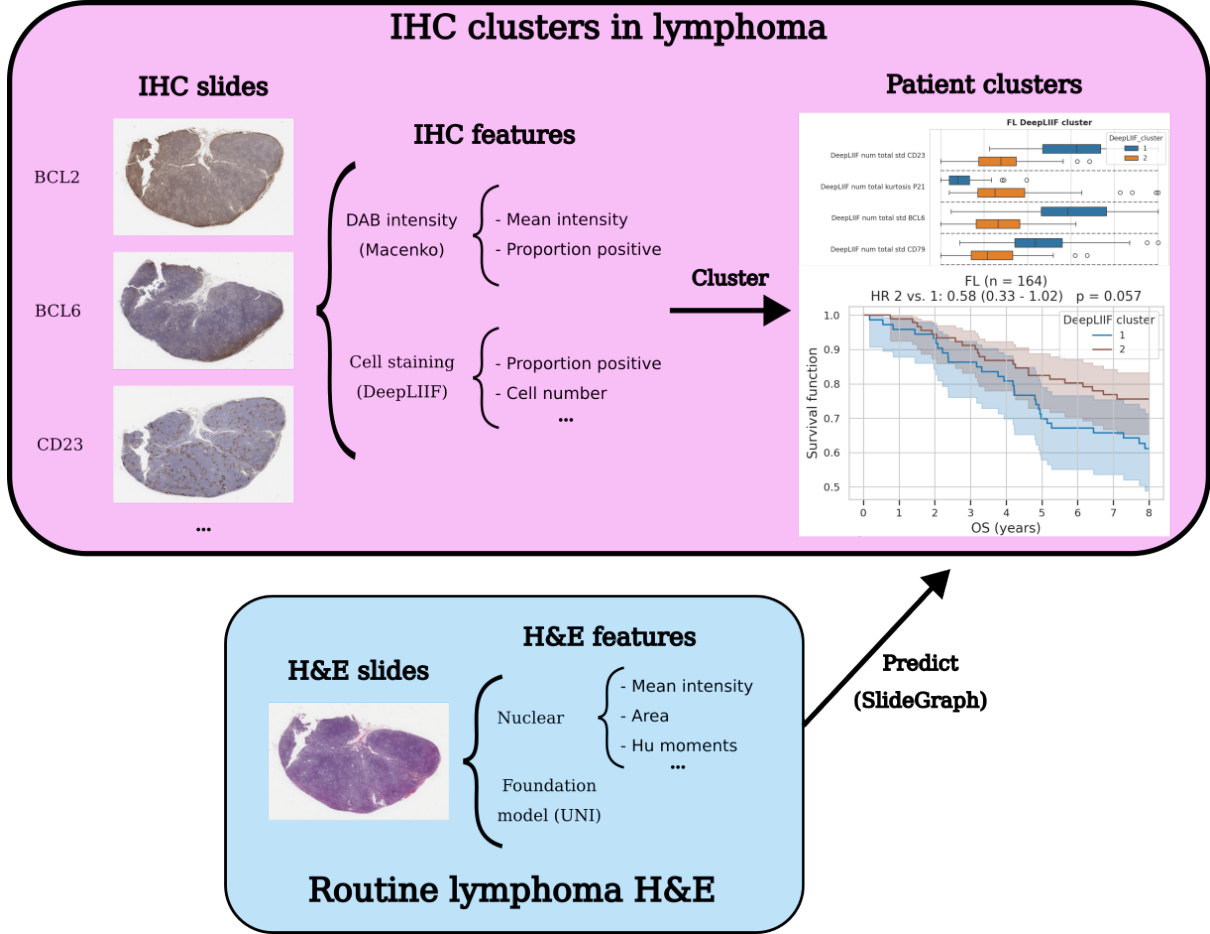


Figure 4.1: Overview of immunohistochemical stain quantification, subgrouping and prediction using H&E.

3.2 Data

Data used were from the UK population-based Haematological Malignancy Research Network (HMRN; <https://www.hmrn.org>[100]; Research Ethics Committee reference 04/Q1205/69) dataset, comprising data on all novel hematological malignancies from 14 hospitals in the Yorkshire & Humberside regions from the 1st September 2004 to the 31st August 2012. Clinical characterisation, including diagnosis, treatment choice and tracking of outcome until the 31st December 2018 was conducted by the Hematological Malignancy Diagnostic

Service (HMDS; www.HMDS.info). Mutation clusters and DLBCL cell-of-origin (COO) expression profiles were previously determined by the HMRN [46, 47, 145, 146]. Staining was performed on serial cuts of formalin-fixed paraffin-embedded (FFPE) tissue samples. Immunohistochemical (IHC) staining involved the stains haematoxylin and DAB (H-DAB). These were scanned at $40\times$ magnification using a Leica Aperio AT2 scanner at St. James’s University Hospital, Leeds. Pathologist estimates of IHC staining were available for Hans algorithm stains (CD10, BCL6 and MUM1) for some HMRN patients. These were encoded in clinical records as ‘+’, ‘+/-’ and ‘-’, with subsequent conversion to the numerical values 1, 0.5 and 0, respectively.

Whole slide, H&E-stained images of the DLBCL-Morph dataset[66, 142, 143], were used for external validation of DLBCL results. The dataset consisted of 204 WSIs from 149 patients, scanned at $40\times$ magnification.

For all datasets, multiple slides from the same patient were treated as separate data points.

3.3 Slide selection

HMRN Immunohistochemistry (IHC) slides were selected for use firstly on stain availability: those conducted on fewer than 80% of FFPE blocks in the respective disease partition were filtered off. For FL, 11 IHC stains met this criteria: BCL2, BCL6, CD10, CD20, CD23, CD3, CD5, CD79, MIB1, P21, P53. In DLBCL, FOXP1 and MUM1 additionally met the criteria (13 stains).

Next, slides were filtered on quality. Quality control (QC) was automated using HistoQC[114], a tool that can quantify slide features (such as fat content, colour intensity and contrast); predict regions of concern, such as those out of focus (blurry) and flag slides with substantial QC concerns. The v2.1 HistoQC configuration was selected, with masking of fat-like tissue (not deemed informative in this use case) and out-of-focus regions (deemed to likely severely impact DeepLIIF cell-based IHC quantification). HistoQC generated binary masks of whole slide images in which background, artefacts and fatty

tissue were indicated. These masks were used to guide sampling of quality regions for downstream tasks, avoiding background, blurry and fatty regions. Corresponding H&E images underwent an identical quality control and region sampling process.

Finally, IHC samples were excluded if the respective patient had less than 60% of required IHC stains (panel of 11 for FL and 13 for DLBCL).

3.4 Data Partitions

Handling missing IHC data Missing IHC quantities were interpolated using K-nearest neighbour (KNN) models for patients with more than 60% of required IHC (DLBCL $n = 162$, FL $n = 73$) but not all required IHC. Patients for which complete panels were available were not included in this interpolation step as it was deemed reasonable to assume that confounding variables existed between patient samples that were selected for full IHC panels and those selected for partial panels. Potential differences between these two groups led to the decision for interpolation to be conducted among patients with incomplete panels.

The threshold value for IHC availability (60%) was determined through random masking, (KNN) imputation of masked data and comparison to unmasked data. This was used to identify minimum required data availability before introduction of imputation errors. Data for which complete IHC panels were available, linearly scaled between 0 and 1, was used for this analysis. A range of missing data proportion between 0% and 90% imputed were assessed. Comparison of mean absolute error (MAE) and Welch's t-test p-values were used to assess divergence of imputed data from original data and so, introduction of imputation errors. MAE was found to increase linearly up to 40% of missing data, with MAE of 0.05. P-value of 0.66 implied that imputed data was largely similar to original data, with no evidence of being from different populations. Roughly exponential increases in MAE were observed with missing data proportions above 40% and so, threshold data availability of 60% was chosen for inclusion in imputed data sets (Appendix Figure A.10).

Imputed dataset Patients with imputed IHC values (DLBCL $n = 162$, FL $n = 73$) were used for grid searches of optimal clustering parameters, with the assumption that the parameters identified would also be suited for patients with complete IHC panel data. This was termed the ‘imputed’ dataset.

Complete IHC dataset Patients for which all required IHC was available were used for training and evaluation of IHC clustering models (DLBCL $n = 168$; FL $n = 112$) using 5-fold cross validation. This was termed the ‘Complete IHC’ dataset.

H&E training data Patients for whom IHC data was available ($>60\%$ of required IHC slides) and corresponding H&E slides were used in training and evaluation of models to predict IHC using H&E data. Training and evaluation were conducted simultaneously using 5-fold cross validation (as in IHC cluster training and evaluation). Where multiple H&E slides existed for the same biopsy sample (DLBCL $n = 6$; FL $n = 2$), these were assigned to the same fold to minimise data leakage between training and test partitions.

H&E Internal validation H&E slides corresponding to patients for whom IHC was not available (DLBCL $n = 178$; FL $n = 218$) were used for internal validation of IHC cluster prediction utility, with comparison against molecular features (mutation profiles in FL and cell-of-origin expression in DLBCL) and overall survival.

3.5 IHC quantification

Baseline deconvolution

The baseline method utilised synthetic separation of haematoxylin (H) and DAB components in the RGB image (termed stain deconvolution or separation) using ‘stain vectors’, the predicted influence of H and DAB on (R, G and B) intensity values in the resulting image. Separation of RGB into H and DAB components allowed prediction of positively-stained areas and the intensity of staining, as in previous efforts[85, 174]. Stain vectors were calculated using the Macenko approach[177], with singular value decomposition of pixel optical density (transformation of RGB pixel intensity; conducted

with HistomicsTK v1.3.14, <http://github.com/DigitalSlideArchive/HistomicsTK>). Stain deconvolution was performed on slides at $1\times$ magnification. $4\times$ magnification ($16\times$ larger images than at $1\times$ magnification) did not lead to noticeably different results (not shown). Pixels with DAB intensity values lower than an empirical threshold of 150 (darker regions) were classified as positive for the IHC stain. Percentage of positive pixels within acceptable HistoQC regions (using HistoQC binary mask) was used as a measure of IHC staining. To distinguish differences in density of staining, such as in a situation where regions classified as positive additionally have different volumes of positive cells, mean DAB pixel intensity within positive regions was also measured. These features were referred to as ‘Macenko IHC features’.

DeepLIIF

The DeepLIIF framework was used to approximate IHC staining in image patches. This tool was developed to predict multiplex immunofluorescence (mpIF) images using matched lung and bladder cancer IHC images. Following prediction of mpIF, cells are segmented using the input IHC and predicted mpIF images. IHC staining (positive or negative) is predicted for each cell. Finally, proportion of positively-stained cells and total number are quantified per input IHC image (patch)[176].

Proportion of positively-stained cells and estimated cell number were quantified in 20-500 RGB patches at $40\times$ magnification of size 512×512 pixels (number extracted depended on tissue availability). These were extracted from acceptable regions of IHC whole slide images (using HistoQC binary masks). DeepLIIF version 1.1.11 was used (<https://github.com/nadeemlab/DeepLIIF>). Mean, standard deviation and kurtosis of positive cell proportion was calculated for each whole slide image, aiming to capture stain intensity, variation in staining and staining ‘tailedness’, respectively. These features were referred to as ‘DeepLIIF IHC features’.

3.6 IHC clustering

Macenko and DeepLIIF IHC feature values were linearly-scaled to the range 0-1 for all downstream activities. Minimum and maximum feature values of all samples of the relevant disease were used in scaling.

Optimal parameters As outlined above (Section 3.4), optimal clustering parameters were determined using patients for whom 60-99% of IHC panels were available (the ‘imputed’ dataset). 5-fold cross validation grid searches were used to identify optimal parameters for each algorithm type (K-means clustering, hierarchical clustering, spectral clustering and Gaussian mixture models). Clustering methods conducive to prediction on unseen data (K-Means and Gaussian Mixture Models) underwent classical 5-fold cross validation, with 5 iterations of training utilising a different data fold (20% of data) for validation using Calinski-Harabasz (CH) score. Clustering methods constrained to input data (Hierarchical and Spectral clustering) used a masked ‘validation’ approach, where test sample data were scaled by uniformly-distributed random values between 0.5 and 1.5. Models were fitted to masked data and validated using CH score between unmasked data and clusters predicted on masked data. Finally, supervised metrics (homogeneity, completeness and V-measure) were used to select an optimal clustering algorithm with respect to molecular features (mutations clusters for FL and COO expression for DLBCL).

Model training Once an optimal clustering method and parameter set had been determined on the ‘imputed’ dataset, these clustering models were fitted to Macenko and DeepLIIF features of the corresponding ‘complete’ dataset, without missing or imputed IHC values.

Inference on imputed data Cluster membership was predicted for data of the ‘imputed’ dataset using clustering models trained on ‘complete’ data, above.

Interpretation of clusters Welch’s t-test was used for comparison of linearly-scaled input features between clusters, since sample numbers and hence, standard deviation were

not equal between clusters. Features were ranked on p-value and features with the 10 lowest p-values between clusters were used for interpretation, with respect to the greatest contrasts between clusters.

3.7 Survival Analysis

Univariate Cox proportional hazards models were trained on overall survival (OS) to obtain Hazard Ratios (HRs). OS was clipped at 8 years, to reduce impact of age (and associated variables such as co-morbidity prevalence) which violated the proportional hazard assumption on calculations (not shown). Kaplan-Meier estimators were used to visualise OS events between groups, termed Kaplan-Meier plots.

3.8 Validation

IHC cluster membership was predicted on H&E slides lacking associated IHC slides in the HMRN (FL $n = 218$; DLBCL $n = 178$) and DLBCL-MORPH[66] ($n = 190$) datasets. Molecular feature prevalence (mutation clusters in HMRN FL; COO expression in HMRN DLBCL and Hans algorithm[25] predictions of GCB expression in DLBCL-MORPH) and overall survival hazard ratios were compared between validation and training H&E data to validate general cluster characteristics. Cox models were fitted to probability of IHC cluster membership, as determined using disease-specific SlideGraph models. Kaplan-Meier estimators were fitted to maximum likelihood IHC clusters. Hazard ratios reported in Kaplan-Meier plots reflect this difference in use of binarised classifications and so differ from HRs reported in HR plots.

Prediction of IHC clusters in H&E

H&E feature extraction H&E feature extraction was performed using two competing methods, widely used in computational pathology efforts - UNI image features[82] and nuclear morphology features. Both methods were performed on identical 448×448 pixel patches, extracted at $40 \times$ magnification from acceptable HistoQC regions. Between 20 and 1,000 patches were extracted, depending on tissue availability.

UNI UNI features have been used across a range of histopathology domains and often found to outperform competing image feature representations[178–180]. The UNI feature extractor is a deep learning model trained on patches from over 100,000 WSIs to extract informative histopathology features. The model was trained in a self-supervised configuration, without explicit annotation of regions. In training, the objective was for the UNI model to ‘learn’ image features informative in predicting:

- Masked regions, thereby ‘learning’ patterns of tissue composition
- Similarities between transformed images, thereby ‘learning’ features with semantic value

The UNI model therefore aimed to ‘condense’ pathology image features while preserving semantic value and contextual information. The minute scale and abstract nature of relationships observed by UNI mean that feature embeddings (vectors of length 1024) present challenges for human interpretation but can provide substantial added value to histopathology analyses.

Embedding Patches were resized to 224×224 pixels using Pillow 10.2.0 to a final magnification of 20×(resolution of 0.50μm/pixel), as used by the developers[82]. Following colour normalisation using parameters of the original paper (for the R, G and B channels: mean=(0.485, 0.456, 0.406), std=(0.229, 0.224, 0.225)), 1024 dimension UNI features were extracted using the VIT/L-16 checkpoint(<https://github.com/mahmoodlab/UNI>).

Nuclear feature extraction Nuclear segmentation was performed on H&E patches using the ‘hoverfast_pretrained_pannuke’ model checkpoint of HoverFast[181]. Nuclei within 10 pixels of patch boundaries were filtered out, due to risk of fragmented nuclear images. 26 nuclear features were measured, selected on non-redundancy (defined as Pearson correlation lower than 0.9) in unlabelled lymphoma nuclei (Chapter 3 Section 3.5). A list of custom texture features is included in the Supplementary files, Chapter 1, Texture_features.xlsx; descriptions of remaining features can be found at <https://scikit-image.org/docs/stable/api/skimage.measure.html#skimage.measure.regionprops>.

Nuclear data was aggregated for each patch using mean feature values.

Graph construction

Features were extracted from H&E patches as described above, resulting in feature vectors for each H&E patch (length 26 for mean nuclear features and 1,024 for UNI features). Additional information on the location of the patch in the whole slide image (coordinates of the top left edge in $40\times$ magnification WSI) was used to construct what could be termed a ‘neighbourhood’ representation for each patch, taking into account other patches within range of the patch (dependent on a distance coefficient λ_d) and influence of their features (scaled with a coefficient λ_f). To reduce computational requirements, these ‘neighbourhood’ representations were clustered into types, taking into account patch features and influence of nearby patches, with clustering determined by a hierarchical clustering distance threshold λ_h . This effectively condensed a large number of H&E patches into a discrete number of tissue clusters, describing similar ‘neighbourhoods’. Finally, graph representations were constructed for whole H&E slides using average feature and coordinate (centroids) values of these tissue clusters, connected to nearby tissue centroids (with a maximum distance threshold between neighbours neighbour_d). The following parameters were used in graph construction, as in the SlideGraph publication[85]:

- $\lambda_d = 0.003$
- $\lambda_f = 0.001$
- $\lambda_h = 0.8$
- $\text{neighbour}_d = 4000$

Training

SlideGraph[85] models were used for prediction of IHC clusters using H&E data. In this model, features from tissue clusters described above as well as features influenced by successively more convoluted interactions in graph representations (Edge Convolution) were

used as input to small artificial neural networks (ANNs). In training, ANN parameters were optimised for prediction of IHC cluster membership (binary classification of cluster 2).

Models were trained and evaluated using 5-fold cross-validation. This involved splitting of H&E slides into 5 partitions, with cases where multiple slides were available for the same patient (FL $n = 2$; DLBCL $n = 6$) assigned to the same partition to reduce data leakage between partitions. Average partition size was 36 slides in FL and 63 in DLBCL. Training was repeated 5 times, with a different partition used for model evaluation in each repeat and the remaining 4 partitions used for model training. In training partitions, 20% of slides were randomly selected for use as a CV partition in each repeat. Performance of a model was evaluated on this CV partition at each epoch, for the purposes of tracking model training with respect to unseen data and determination of a suitable end point for training.

Training parameters used were:

- Maximum epochs = 1,000
- Early stopping tolerance = 250
- Optimiser = Adam
- Learning rate = 0.01
- Learning rate decay: $3\times$ every 100 epochs
- Adam weight decay = 0.0001
- Training dropout proportion = 0.3
- Dense layer perceptrons: 16 (base network), 16 (1st and 2nd Edge Convolution layers) and 8 (3rd Edge convolution layer)
- Pooling layer type = Mean
- WSI-level aggregation method = maximum pooling layer value

Evaluation

Classification metrics were used to compare model performance (Appendix section 4.1) on validation partitions (5-fold cross validation, above). Subsequently, these were used to assess optimal H&E features for prediction of IHC, and which IHC features were conducive to prediction with H&E.

Given limited dataset size, no additional hold-out dataset containing matched H&E and IHC was used to evaluate models in prediction of IHC in H&E. Instead, clinically-relevant patient features were compared between the matched IHC/H&E dataset used in training of models and H&E data without matched IHC. This consisted of 178 DLBCL and 218 FL slides of the HMRN dataset, as well as 190 DLBCL slides from the DLBCL-MORPH dataset.

Model interpretation

To interpret which features were used for successful prediction of IHC cluster in H&E, patches classified as highly predictive of each IHC cluster were compared with visual inspection. Heatmaps of WSIs were used to inspect distribution of these patches and determine which tissue regions were enriched for each. Additionally, GNNExplainer[182] was used to extract input feature importance values on internal validation data. This method aimed to identify features and subgraphs (small interaction networks) informative of the classifications made by SlideGraph models. Pearson correlation was calculated between feature importance for each slide and predicted (binary) cluster membership, to determine which input features were most informative of slide-level classifications over all data.

Significance testing

Observed Frequencies The χ^2 test of independence of variables was used to assess the probability that observed frequencies were independent between evaluated groups. This was used over Fisher’s exact test to accommodate matrices larger than 2×2 .

Comparisons of continuous variables Both Pearson (parametric) and Spearman (non-parametric) correlations were used to assess association between continuous features.

Paired samples The non-parametric Wilcoxon test was used to compare related samples, such as in comparison of models in prediction on identical data folds. A non-parametric approach was used over paired T-tests as fold number (5) was deemed too low for reliable estimation of distribution parameters.

Independent samples Where samples were independent, such as comparison of features between clusters, Welch’s t-test was used since group sample sizes (and hence, variance) were not equal between groups.

Hazard ratios Significance values assigned to Cox model hazard ratios were calculated using the Wald test of a single parameter, indicating probability of the observed or more extreme data, assuming a null hypothesis that hazard ratio equals 1 (no effect on risk).

Multiple testing correction Where indicated, the Bonferroni correction was used for multiple testing.

3.9 Computational specifications

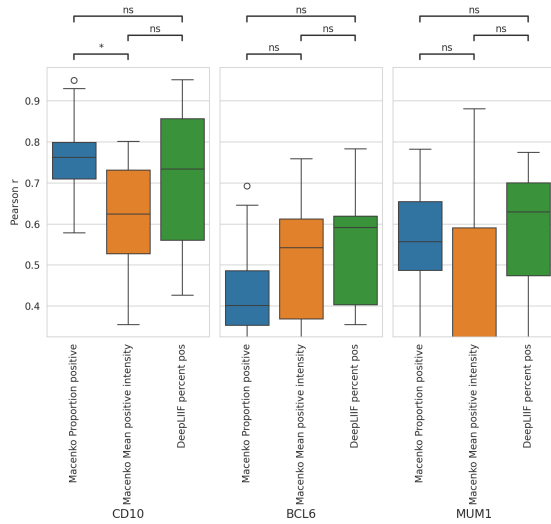
A workstation with an NVidia RTX 3060 Ti GPU, AMD Ryzen 5 3600 CPU and 16GB RAM was used for the majority of work presented. Virtual environment dependencies differed extensively for IHC feature extraction (DeepLIIF), H&E feature extraction (combined HoVerFast and UNI feature extraction) and H&E model training (SlideGraph) so are not reported here. Complete virtual environment specifications will be provided upon request.

4 Results and Discussion

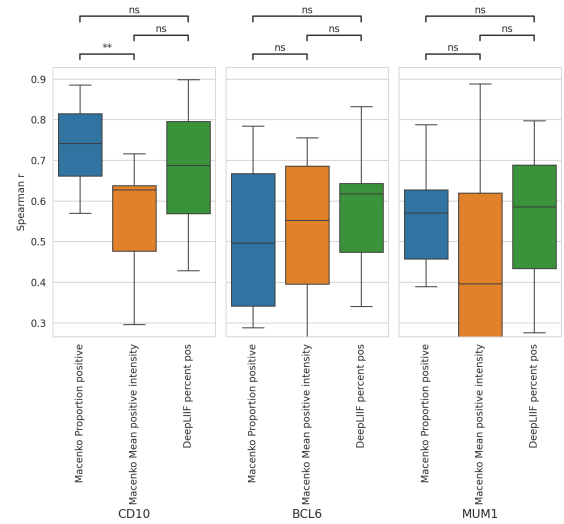
4.1 IHC quantification

Methods of IHC quantification were compared against pathologist assessments in the HMRN DLBCL dataset using correlation. In both parametric (Pearson) and non-parametric (Spearman) measures of correlation, positive IHC proportion predicted through Macenko stain deconvolution (Macenko Proportion Positive) significantly outperformed mean IHC intensity (Macenko Mean Positive Intensity) in CD10 quantification (Figure 4.2A and B). No significant differences were otherwise observed, including between Macenko and DeepLIIF predictions of positive IHC regions. Though not significant, percentage of positive cells, as estimated with DeepLIIF outperformed other methods in quantification of BCL6 and MUM1. Performance was least stable with Macenko Mean positive intensity, with visibly greater 95% confidence intervals in BCL6 and MUM1.

Correlation with pathologist estimations of proportion positive stain was not significantly different between target proteins with DeepLIIF but, Macenko method estimations of nuclear protein proportions (BCL6 and MUM1) were significantly worse than for the cell membrane protein CD10 (Figure 4.2C and D).



(a) Pearson correlation between quantification methods



(b) Spearman correlation between quantification methods

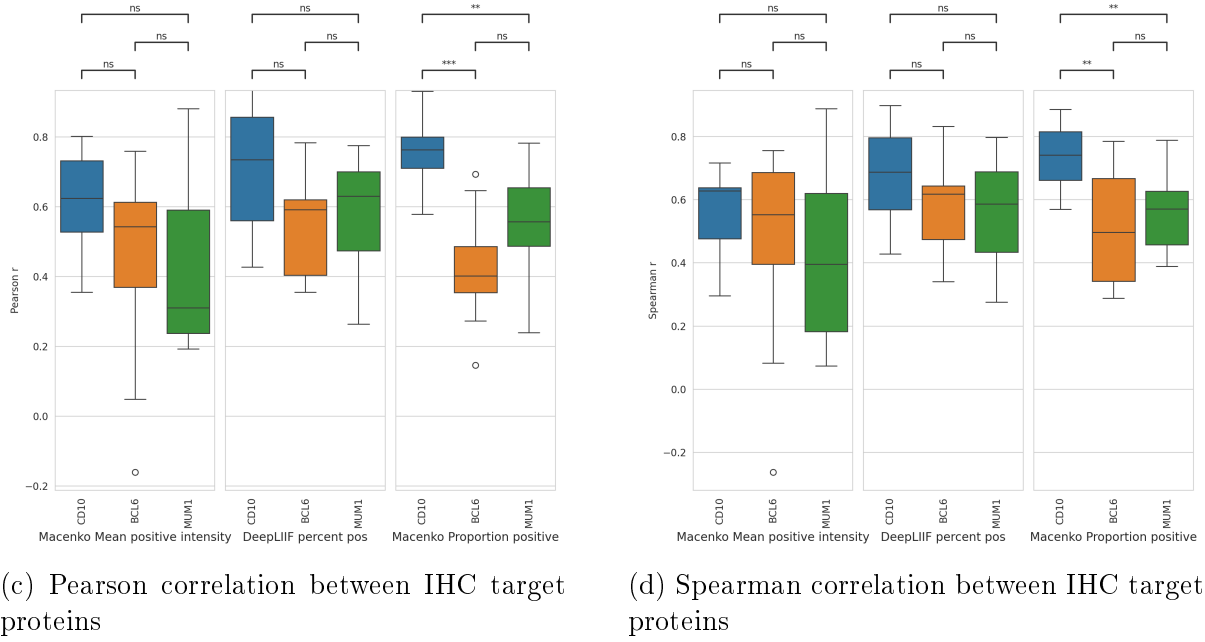


Figure 4.2: Performance of IHC quantification methods against pathologist assessment for CD10, BCL6 and MUM1.

A & B) Comparisons between IHC quantification methods, to assess optimal methods for each stain

C & D) Comparisons between target proteins within stains, to assess quantification method performance across stains

Notes: Confidence intervals were calculated using 10 partitions consisting of 10% of matched data. The Mann-Whitney test was used for significance testing. Significance key:

ns: not significant; *: $0.05 > p \geq 0.01$; **: $0.01 > p \geq 0.001$; ***: $0.001 > p \geq 0.0001$

To investigate the nature of disagreement between DeepLIIF and pathologist predictions of IHC staining, patch-wise estimations of BCL6 presence were compared over 10 random regions (Figure 4.3, below). DeepLIIF was found to primarily under-estimate BCL6, with respect to pathologist (Pearson $r = 0.57$). This appeared to originate from differences in threshold IHC presence required for classification of positive cells, especially clear in Figure 4.3 top, where all cells were predicted to be positive by a pathologist but the majority of cells with lower levels of staining were predicted to be negative for BCL6 by DeepLIIF (annotated with blue outlines and the letter ‘A’).

Uncommonly, DeepLIIF was found to slightly over-estimate staining. In these cases,

visual inspection identified false positive classifications of cells, potentially due to background IHC deposits (Figure 4.3 bottom).

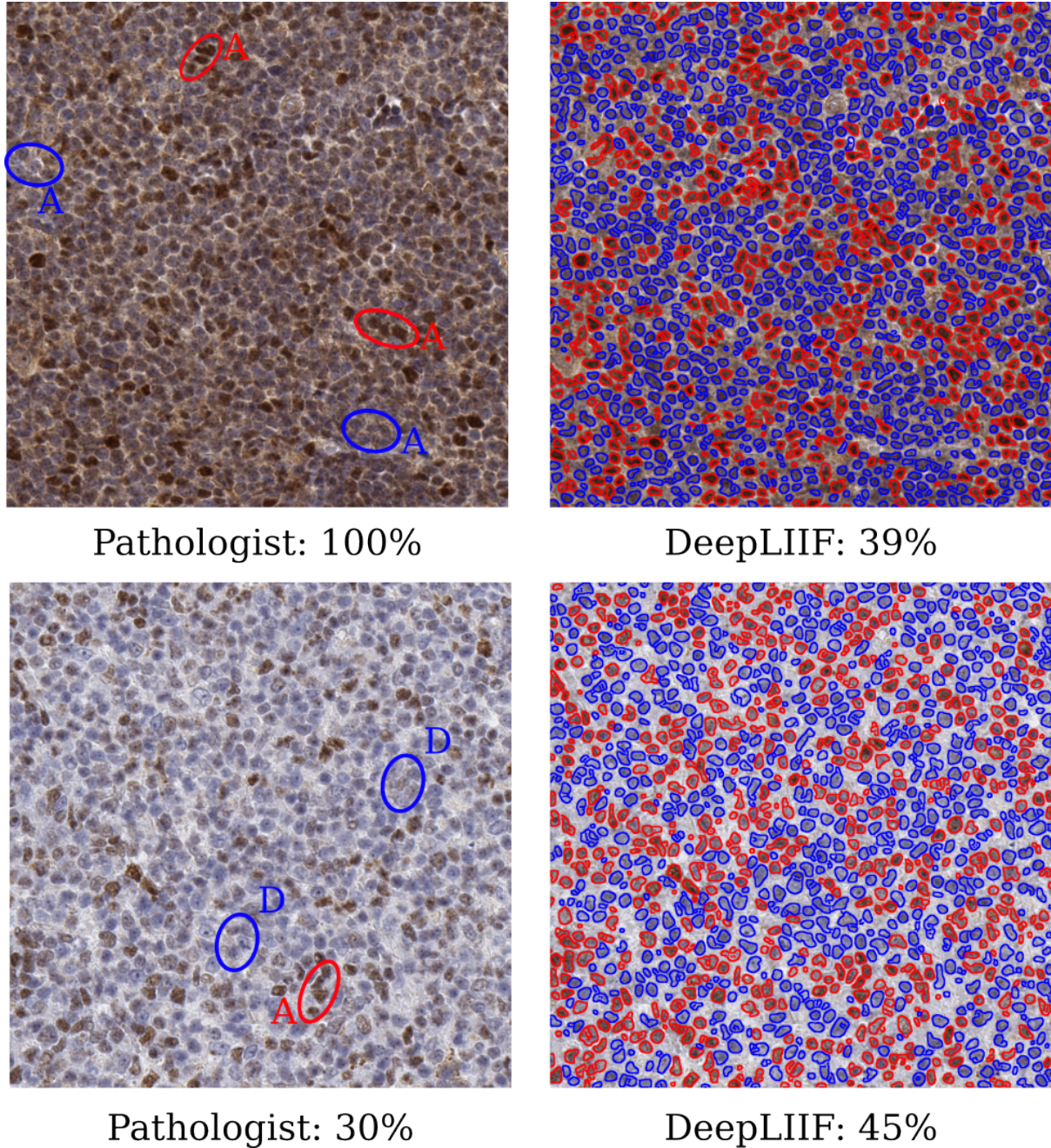


Figure 4.3: Comparison of patch-wise BCL6 quantification between DeepLIIF and pathologist. Left images = IHC patches extracted at $40\times$ magnification. Pathologist estimations of cell proportion positive for BCL6 below. The images were annotated with coloured ovals to indicate possible presence (red) or absence (blue) of IHC along with text to indicate agreement with DeepLIIF (A) or disagreement (D). Right images = DeepLIIF segmentation masks with classifications of BCL6 presence indicated by boundary colour of nuclei. Boundary colour key: Red = positive; Blue = negative. DeepLIIF estimations of cell proportion positive for BCL6 below.

Discussion

Macenko mean positive intensity displayed high uncertainty in almost all comparisons and did not achieve the highest correlation value for either of the IHC targets stained. The utility of this quantity, compared to the estimates of proportions of positive IHC used appeared low.

No significant differences were observed between Macenko and DeepLIIF proportions of IHC positive tissue. As such, performance in quantification of the cell membrane protein CD10 and nuclear protein MUM1 were comparable with Macenko and DeepLIIF. DeepLIIF appeared to outperform Macenko in quantification of the nuclear localised BCL6 transcription factor and showed greater consistency in quantification between stains. Macenko estimations of positive tissue proportions were significantly worse for nuclear proteins, compared to cell membrane CD10. Pearson correlation values between predicted staining proportion for the nuclear proteins BCL6 and MUM1 appeared consistent in DeepLIIF

Observations of DeepLIIF classification against pathologist estimates identified that DeepLIIF appeared to have more adaptive thresholds of staining than a pathologist, classifying comparable cells differently between images (Figure 4.3 top vs. bottom). Cells with high very IHC staining were classified as positive while those with moderate and weak staining, while above pathologist thresholds of positive staining, were classified as negative. This was likely due to stain normalisation procedures used as preprocessing steps by DeepLIIF.

Pearson correlation of 0.57 and visual inspection evidenced that DeepLIIF operated with greater specificity than pathologists, classifying only the most highly stained cells as positive for BCL6.

Overall, there were no significant performance differences between Macenko and DeepLIIF estimations of positive cells/tissue. Both methods displayed better performance in quantification of membrane (CD10) than nuclear proteins (significantly for Macenko but not for DeepLIIF). Disparities between pathologist and DeepLIIF IHC quantification were

largely in under-estimation of IHC staining, appearing to result from differences in thresholds for positive staining between the model and pathologist.

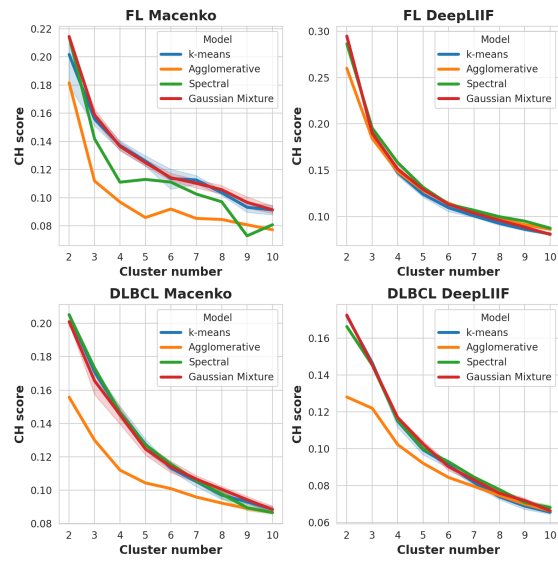
Despite considerable differences in appearance and pathologist estimations of staining (Figure 4.3 top vs. bottom), DeepLIIF estimations were at times similar, evidencing a need for suitable scaling to better resolve cases.

4.2 IHC clustering

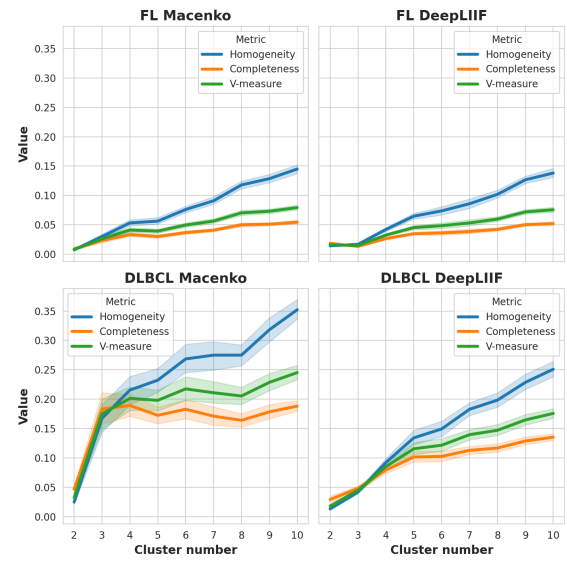
Optimal clustering

For both diseases and IHC feature sets used, Calinski-Harabasz score was maximised with a cluster number of two (Figure 4.4A). In FL, supervised metrics appeared to increase with increasing cluster number though, rate of increase appeared lower with more than 4 clusters for Macenko and 5 for DeepLIIF IHC (Figure 4.4B). Similar results were observed in DLBCL with lower rates of metric value increase from 3 clusters in Macenko and 5 clusters with DeepLIIF.

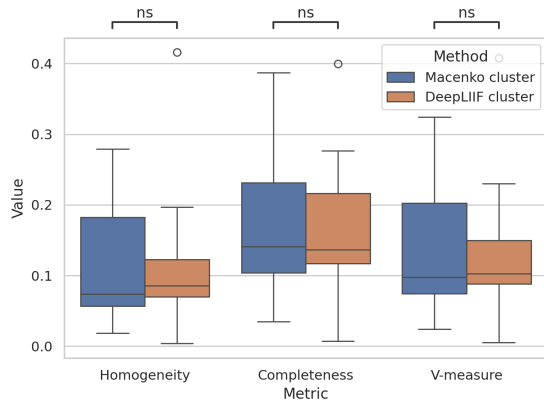
For optimal clustering methods with respect to CH score, no significant differences were observed between Macenko and DeepLIIF feature clusters in supervised metrics (Figure 4.4C and D). No noticeable differences were observed between DeepLIIF clusters in molecular feature distribution (FL $p = 0.879$; DLBCL $p = 0.662$) but small differences were observed with Macenko clusters (FL $p = 0.275$; DLBCL $p = 0.327$; Figure 4.4E).



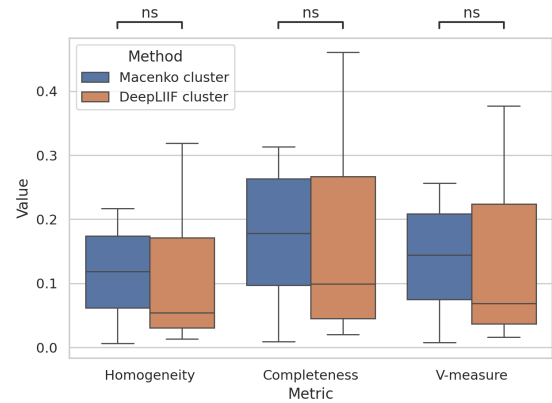
(a) Calinski-Harabasz score over cluster number



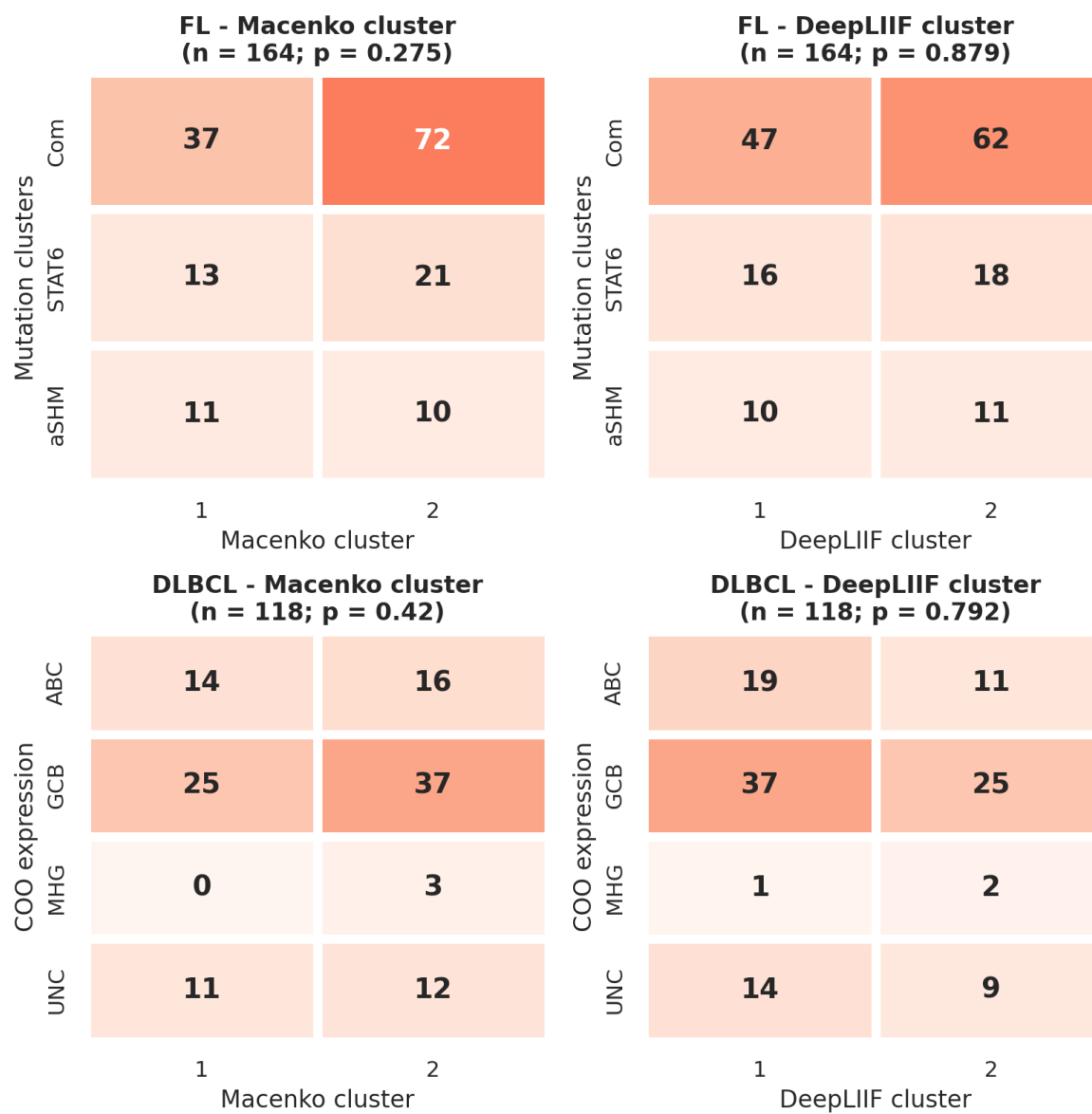
(b) Supervised metrics over cluster number



(c) Supervised metrics for 2 clusters, with respect to FL mutation groups



(d) Supervised metrics for 2 clusters, with respect to DLBCL cell-of-origin expression



(e) Contingency of clusters with molecular features

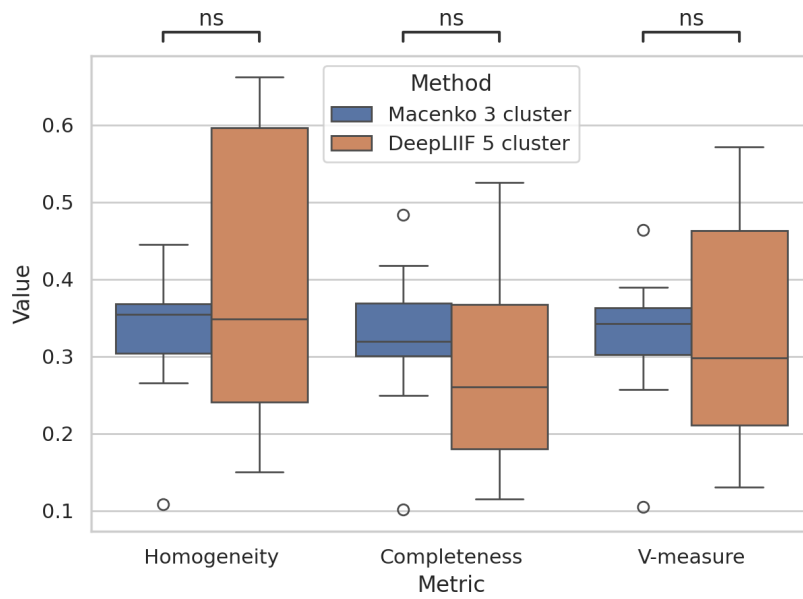
Figure 4.4: Evaluation of IHC clustering methods, with respect to Calinski-Harabasz score (CH) and supervised metrics against molecular features:

- A) CH score of optimal clustering models over cluster number (optimal parameters determined using 5-fold cross validation grid searches)
- B) Supervised clustering metric values, with respect to molecular features (mutation clusters in FL and COO expression in DLBCL) over cluster number. Data represented include all clustering models, using optimal parameters (determined through 5-fold cross validation grid searches using CH for scoring).
- C) FL model supervised clustering scores with respect to mutation groups (calculated over 10 folds consisting of 10% of data)
- D) DLBCL model supervised clustering scores with respect to Cell-of-Origin expression (calculated over 10 folds consisting of 10% of data)
- E) Contingency matrices of clustering against molecular features

Clustering DLBCL into more groups

Supervised metrics on DLBCL data evidenced considerable increases in utility with 3 clusters for Macenko features and 5 for DeepLIIF features. These alternative cluster numbers were compared against the two cluster method, above, and each other to evaluate clinical utility.

Evaluation against expression First, clusters were compared on COO expression. No significant difference was observed in supervised metric values though, Macenko clusters displayed far greater consistency than DeepLIIF clusters (Figure 4.5a). Analysis of frequency identified significant differences with Macenko clusters ($p < 10^{-5}$; Figure 4.5b), with enrichment of GCB expression in cluster 1 (75% of the cluster) and ABC in cluster 3 (58% of the cluster). Though DeepLIIF cluster 3 appeared enriched for GCB (63% of cluster), differences in frequency were not significant. Similarities appeared to exist between clusters 1 & 2 and clusters 3 & 4, respectively, with comparable proportions of constituent expression groups.



(a) Supervised metric values

DLBCL - Macenko 3 cluster (n = 118; p = 6e-06)				DLBCL - DeepLIIF 5 cluster (n = 118; p = 0.351)						
COO expression	ABC	2	14	14	ABC	5	5	9	6	5
	GCB	33	25	4	GCB	10	11	22	15	4
	MHG	3	0	0	MHG	1	0	1	1	0
	UNC	6	11	6	UNC	7	8	3	2	3
Macenko 3 cluster				DeepLIIF 5 cluster						
1 2 3				1 2 3 4 5						

(b) Contingency matrices

Figure 4.5: Comparison of higher cluster number techniques against cell-of-origin expression (COO) in HMRN data, for which COO was available:

A) With respect to supervised metric values

B) With respect to frequency of patients (p-values calculated using χ^2 test)

Discussion Similarities between DeepLIIF clusters 1/2 and 3/4, respectively, evidenced clustering into too many clusters. Significant differences in expression between 3 clusters (Macenko IHC features) further evidence this as a more appropriate number of clusters than 5. In summary, clustering of Macenko features into 3 clusters appeared to result in more clinically-useful clusters, with respect to COO expression.

Overall discussion on optimal clustering

Optimal cluster number, as indicated by maximum metric values, differed between the data-centric Calinski-Harabasz metric (maxima at $k = 2$; Figure 4.4a) and supervised metrics with a focus on mutation and gene expression profiles (Figure 4.4b). There did not appear to be a general optimum in clustering of IHC features used, with respect to agreement in maximum metric values. For this work, the data-centric optimum of 2 clusters was used for downstream work.

In FL, using higher number of clusters may have led to improved separation of patients on mutation profile (as evidenced by supervised metrics) but would have also led to lower cluster membership. Since this would have presented challenges in survival analyses, due to lower event numbers in the indolent disease, no further cluster numbers were assessed.

In DLBCL, far higher supervised metric values with 3 and 5 clusters, for Macenko and DeepLIIF IHC, respectively, evidenced far more meaningful clustering than with 2 clusters. Event number and patient sample sizes were higher in this aggressive lymphoma so, additional clustering was deemed more feasible than with FL. Therefore, DLBCL patients were also grouped into 3 (Macenko) and 5 (DeepLIIF) clusters to assess utility of higher cluster numbers in survival and separation of expression profiles. This was performed with a view of conducting further analyses if greater clinical utility was evidenced.

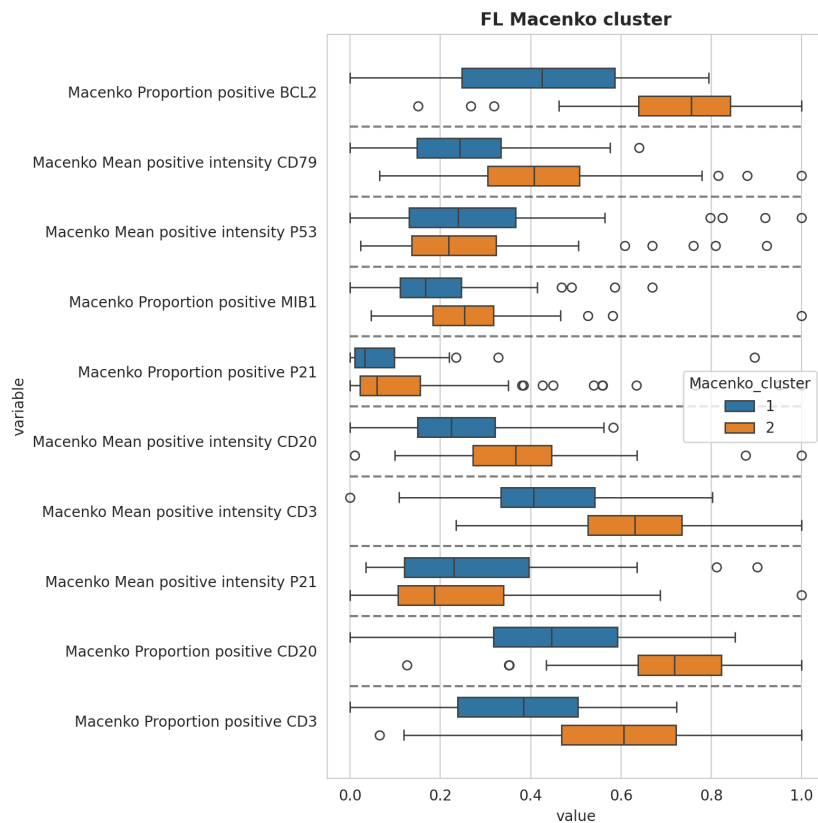
At a cluster number of 2, no significant differences were observed between Macenko and DeepLIIF features in clustering of molecular feature groups (Figure 4.4c and d). Nonetheless, Macenko features appeared to cluster patients on molecular features more effectively (Figure 4.4e).

In FL, V-measure for clustering of samples on mutation profile was found to be low at under 0.1 over the cluster range tested. In DLBCL COO expression, this was found to be considerably higher, especially with Macenko IHC quantification (V-measure value of 0.2 at $k = 4$). These results evidenced greater scope for prediction of DLBCL COO expression than FL mutation profile with the features used.

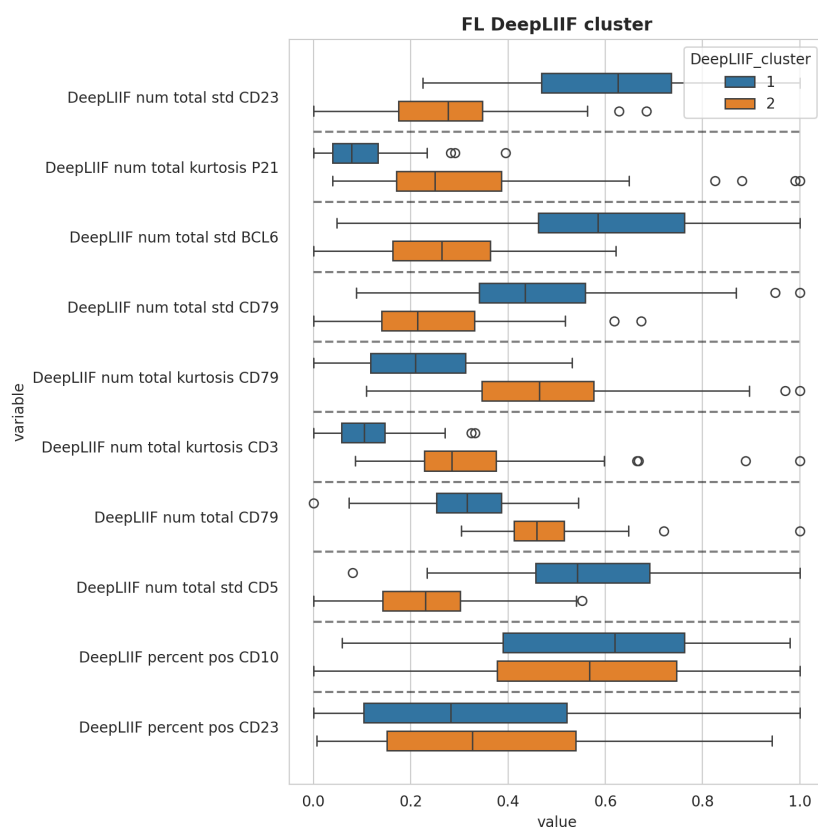
No feature filtering was used. Differences in supervised metric values between IHC quantification methods in DLBCL evidenced that feature choice may have led to considerable differences in clusters. This effect was far smaller in FL. Therefore, future feature selection efforts focusing on those informative of molecular characteristics or survival may lead to vastly different patient clusters than presented here.

IHC cluster Interpretation

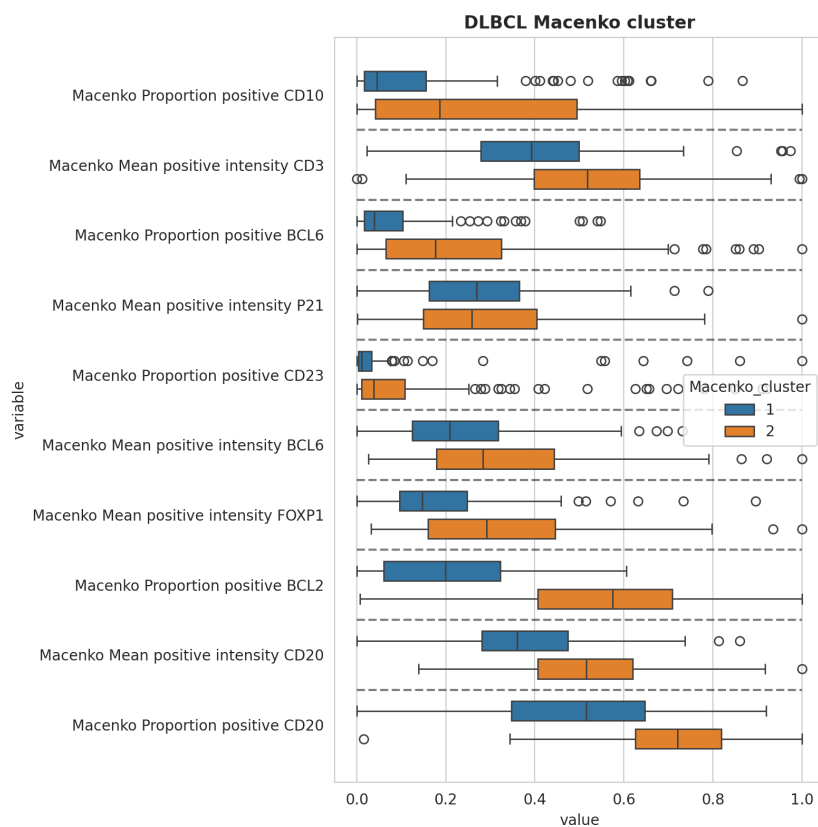
Boxplots of the most significantly different features between IHC clusters (Figure 4.6, below) were used as a means of interpreting tangible differences between clusters.



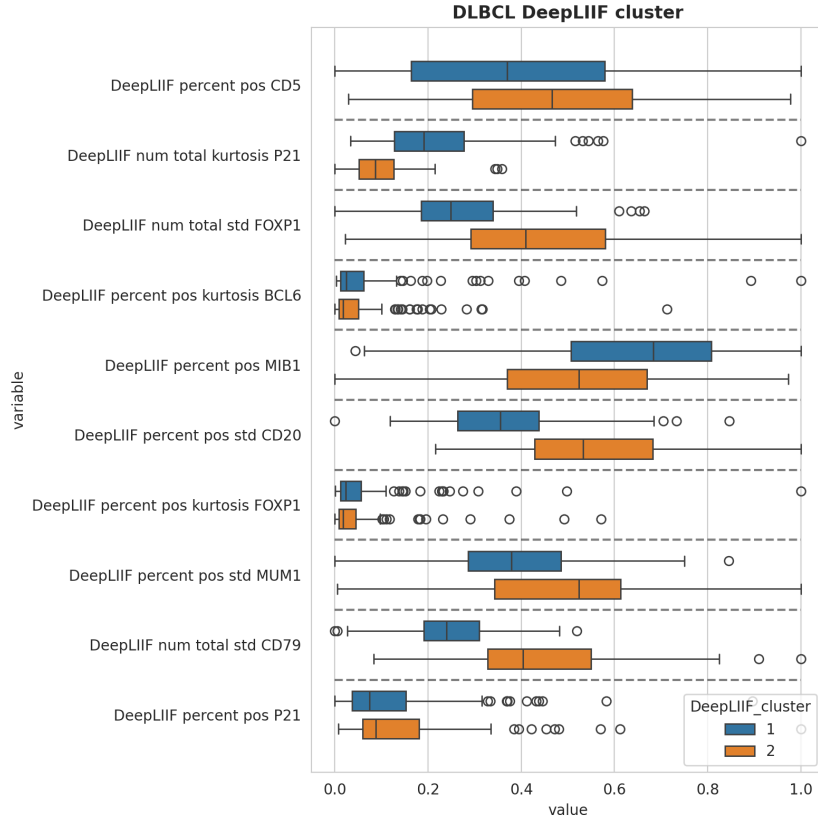
(a) FL Macenko clusters



(b) FL DeepLIIF clusters



(c) DLBCL Macenko clusters



(d) DLBCL DeepLIIF clusters

Figure 4.6: Boxplots of the most significantly different features between IHC clusters in FL and DLBCL, sorted from lowest p-value (top) to highest (bottom).

Interpretation of p-values directly was not deemed meaningful following clustering (which focused on separation of individuals by the greatest differences). Therefore, p-values were used here to deduce which features were most different between groups, not for the purposes of significance testing.

FL

Macenko clusters Cluster 2 displayed almost 2× higher BCL2 staining (0.75 in cluster 2 vs. 0.41 cluster 1) proportion and almost 50% higher CD20 and CD3 positive proportions. Also, slight differences were observed in CD79, CD3 and CD20 stain intensity.

DeepLIIF clusters Most differences related to total number of cells observed, with 2× greater standard deviation in CD23, CD5 and BCL6 cell numbers within cluster 1 (cluster 2 0.2-0.3 vs. cluster 1 0.55-0.65 for each).

Discussion BCL2 is produced by most cells in non-cancerous (reactive) lymph nodes, except naive B-cells undergoing somatic B-cell receptor mutations[183]. CD20 and CD3 IHC stain B- and T-lymphocytes, respectively. Lymph nodes are classically saturated with naive B-cells in germinal centres (GCs; follicles or sheets in the case of FL and DLBCL, respectively) and mature B-/T-cells in the inter-follicular regions[30, 183]. Far lower proportions of all cell markers expected to be present in lymph nodes - B-/T-lymphocyte markers (CD20 and CD3, respectively) and germinal centre B-cell markers (BCL2) - evidence low densities of all cells. This may in turn evidence classification of unintended features of samples, such as fibrous, fatty or necrotic tissue. Though HistoQC likely removed the vast majority of adipocyte (fat cell) masses from analyses, it is possible that small deposits contained within tissue regions (or fibrous regions) escaped the binary erosion strategy used for masking of fatty regions.

As with Macenko clusters, DeepLIIF clusters differed mainly in total cell count features (number, standard deviation and kurtosis), rather than on proportion of cells predicted to be positive for IHC stains. These observations evidenced that IHC clustering had classified FL samples by tissue types rather than on IHC subgroups of lymphoma patients, with the features used.

2 DLBCL clusters

Macenko clusters Cluster 2 displayed considerably higher proportions of tissue positive for CD10 (GC marker), BCL6 (transcription factor responsible for regulating GC B-cell activities), BCL2 (produced by most cells in lymph nodes) and CD20 (a general B-cell marker) than cluster 1[30]. Additionally, cluster 2 stain intensity was higher in CD3 (general T-cell marker), FOXP1 (produced by B-cells in the mantle zone, surrounding the GC[183]), BCL2 and CD20.

DeepLIIF clusters Cluster 1 displayed higher proportions of MIB1 staining (Ki67; a marker of rapid proliferation with classically strong staining in GCs[183]), lower CD5 staining (a marker of T-cells and some dendritic cells[184], expected to primarily stain T-cell rich regions towards the efferent lymphatic) and lower standard deviation in MUM1

staining (expected in B-cells[183]), with respect to cluster 2.

Discussion As with FL, DLBCL Macenko cluster 1 displayed far lower staining of a range of cell markers, including B-cell (CD20, CD10 and BCL6) and general cell markers (BCL2), evidencing lower general cell presence. Again, this indicated that differences in tissue characteristics (fibrous, fatty, vascular vs. lymph nodes densely-populated with B-/T-cells) led to the observed clusters.

Conversely, DeepLIIF cluster 1 displayed GC characteristics, with high proliferation rate (Ki67 staining) and low T-cell presence (CD5 staining). Lower standard deviation in MUM1 indicates homogeneity in B-cell phase in cluster 1. Multiple stages of consistent MUM1 presence exist, with high levels associated with early (centrocyte) and late stage (plasma cell lineage) B-cells[30] and non-GCB expression[25].

3-5 DLBCL clusters

Heatmaps of Macenko IHC features were used to interpret IHC staining characteristics of the clusters (Figure 4.7, below). Given lower utility of DeepLIIF clusters compared to Macenko clusters, these were not considered further.

The majority of patients assigned to cluster 1 demonstrated high proportions of CD20 and CD79 staining, with moderate BCL2, CD10 and MIB1 (Ki67) staining. This was accompanied by high stain intensity in CD20, CD3, CD5, CD79 and MIB1.

Cluster 2 patients generally displayed high proportions of CD20 and CD79 staining tissue, accompanied by moderately-high stain intensity in CD20, CD3, CD79 and MIB1.

Cluster 3 displayed very high proportions of BCL2, CD20 and CD79-positive tissue and moderate to high proportions of FOXP1, MIB1 and MUM1 in some patients. This was accompanied by high stain intensity in BCL2, CD20, CD79 and MIB1.

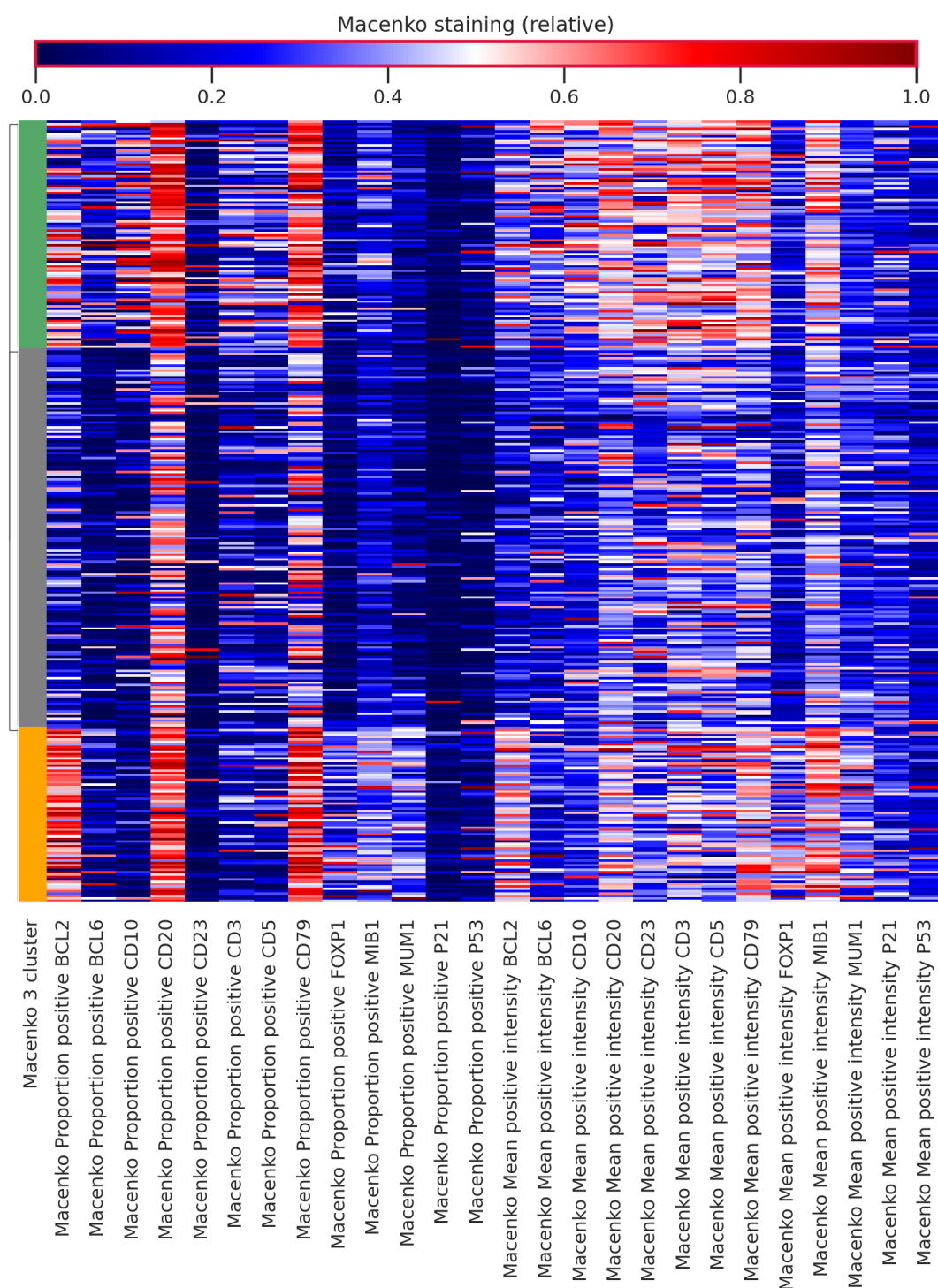


Figure 4.7: Heat map demonstrating scaled IHC values, ordered by Macenko IHC cluster (left column), with cluster colour matched to Kaplan-Meier estimators: green = cluster 1; grey = cluster 2; orange = cluster 3. Columns relate to Macenko IHC features while each row represents an individual patient.

Discussion All clusters displayed moderate to high levels of the B-cell markers CD20 and CD79, as expected for analysis of B-cell rich lymphoid tissue. Cluster 2 was dis-

tinguished by moderately-high presence of these markers (visibly higher in both clusters 1 and 3) and low proportions of other IHC markers, such as BCL2 and MIB1, present in the other clusters. This evidenced a general B-cell cluster, demonstrating non BCL2-mediated apoptosis escape but low proliferation.

Cluster 1 was distinguished through moderate presence of the naive B-cell marker CD10 (rarely observed in the other clusters), BCL2 (higher than in cluster 2 but lower than cluster 3), the T-cell marker CD3 (visibly higher than in the other clusters) and like cluster 3, high Ki67 levels. This evidenced a germinal centre B-cell pattern of pathogenesis, with BCL2-mediated apoptosis escape, presence of naive B-cell markers (CD10) and moderate proliferation rate (Ki67). The nature of T-cell involvement is unclear. Follicular helper T cells (T_{fh}) have been implicated in neoplastic B-cell survival in FL and GCB DLBCL but require BCL6[185], which was observed to be low in most patients of cluster 1. These T-cells could therefore be cytotoxic CD8 T-cells, mounting immune responses against DLBCL cells. It was not possible to deduce the lineage of these T-cells without further IHC markers (CD4, CXCR5, PD-1 and ICOS for T_{fh} and CD8 for cytotoxic T-cells).

Cluster 3 was distinguished by high BCL2 (low in cluster 2), low CD10 (high in cluster 1), moderate FOXP1 and MUM1 (largely absent in other clusters) and moderate MIB1 (Ki67; absent in cluster 2). This cluster therefore represented activated B-cells (FOXP1 and MUM1, and absence of CD10) with moderate proliferation.

Published algorithms for classification of COO using IHC include:

- Hans algorithm[25], operating on naive B-cell markers CD10, BCL6 and the activated B-cell marker MUM1 (IRF4)
- Theophraste algorithm[172], operating on CD10, MUM1, and the activated B-cell-associated FOXP1 and IgM proteins (IgM IHC not performed in this work)

Clear similarities existed between the presented, automated effort and these algorithms. As with the Hans algorithm, presence of CD10 distinguished clusters 2 and 3 (with considerable populations of non-GCB expression) from cluster 1. BCL6 was found to

be less informative for distinction of GCB as proportions of positive cases were quite low. As identified in Section 4.1, Macenko IHC estimation displayed relatively poor agreement with pathologists in estimation of BCL6. Disagreements in utility of this marker in GCB estimation may therefore have arisen from poor performance in BCL6 estimation. MUM1 also appeared informative in distinction of GCB-rich cluster 1 and cluster 3 but less informative in distinction from cluster 2, also less enriched in GCB. FOXP1 and MUM1, as proposed in the Theophraste algorithm, appeared to distinguish ABC-enriched cluster 3 well, with almost complete absence in the other clusters. These results support the Theophraste algorithm as a superior approach to distinction of GCB and ABC cases.

Neither of the previous algorithms distinguish the general B-cell cluster 2, heterogeneous in COO expression. Given the highly diverse range of neoplasms classified under the umbrella term of DLBCL, and the potential for observation bias from analysis of only 1 lymph node in a disease often diagnosed at stages III/IV (systemic disease), poor resolution of expression groups in this cluster is not unexpected. Bagging, the process of training machine learning classifiers to distinguish successively more challenging samples, may aid in resolving expression groups within this cluster.

Overall discussion on cluster characteristics

2 group clustering In all FL clusters and DLBCL Macenko clusters, the greatest differences between complementary clusters (1 vs. 2) appeared to relate to general cell density. This evidenced that features used led to clustering on extracellular tissue, such as fat, necrosis, muscle and fibrous tissue rather than on proportions of positively-stained cells. Samples may therefore have been clustered on the confounding variable of biopsy site, given differences in tissue composition between lymph nodes and even bone marrow samples included in the HMRN dataset.

Such information may have been clinically-relevant as locations of neoplastic lymph nodes can be associated to mechanisms of pathogenesis. This is especially valid with respect to the emerging classification of lymphoma of immune-privileged sites[23], where the

immune evasion mechanisms present in most cases of DLBCL[186] would likely be less necessary for tumour survival.

Information on biopsy site was not available at the time of writing. Given challenges in accurate, standardised and meaningful reporting of biopsy site, there were potential benefits to using the site naive IHC clustering approaches describe.

Finally, DLBCL DeepLIIF clusters appeared to distinguish cases on proliferation rates (high in cluster 1 and lower in cluster 2) and presence of T-cell subpopulations (low in cluster 1 and higher in cluster 2).

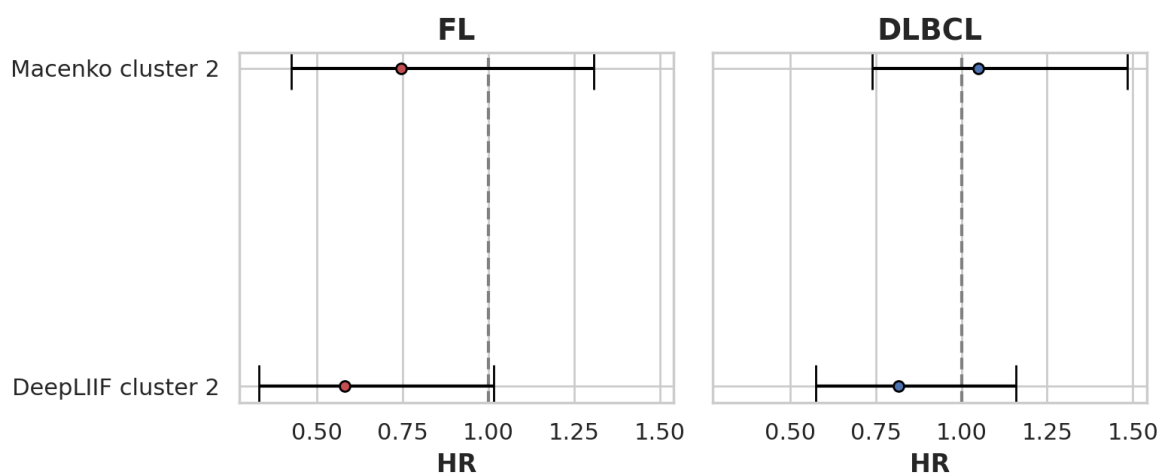
3-5 DLBCL clusters Overall, comparison of the 3 and 5 cluster approaches evidenced greater clinical utility in 3 clusters, with respect to expression and survival. Macenko IHC clusters 1 and 3 were associated with starkly different patient outcomes, with comparatively good 5-year OS (75%) with cluster 1, enriched for GCB expression, and exceptionally poor outcome with cluster 3 (5-year OS of 47%), enriched for ABC expression patients. These results agreed with observed survival for patients with GCB and ABC expression[19, 21]. Interpretation of IHC characteristics evidenced BCL2, CD10, FOXP1 and MUM1 as key distinguishing features, of which all but BCL2 have previously been proposed for classification of COO with IHC[25, 172]. These results evidenced utility of automated efforts in classification of IHC subgroups, potentially for triage of high risk, ABC expression patients.

DeepLIIF features appeared to perform less well than Macenko in clustering, despite higher agreement with pathologist estimates of positive staining (Section 4.1). Given that a major difference in Macenko and DeepLIIF clustering was inclusion of features irrelevant in IHC comparison with DeepLIIF, namely those relating to total number of cells detected, it appeared likely that insufficient feature selection contributed to poorer clustering with DeepLIIF. Targeted feature selection for those intuitive in IHC comparison (such as proportion positive cells) may have improved resolution of clinically-useful clusters with DeepLIIF features.

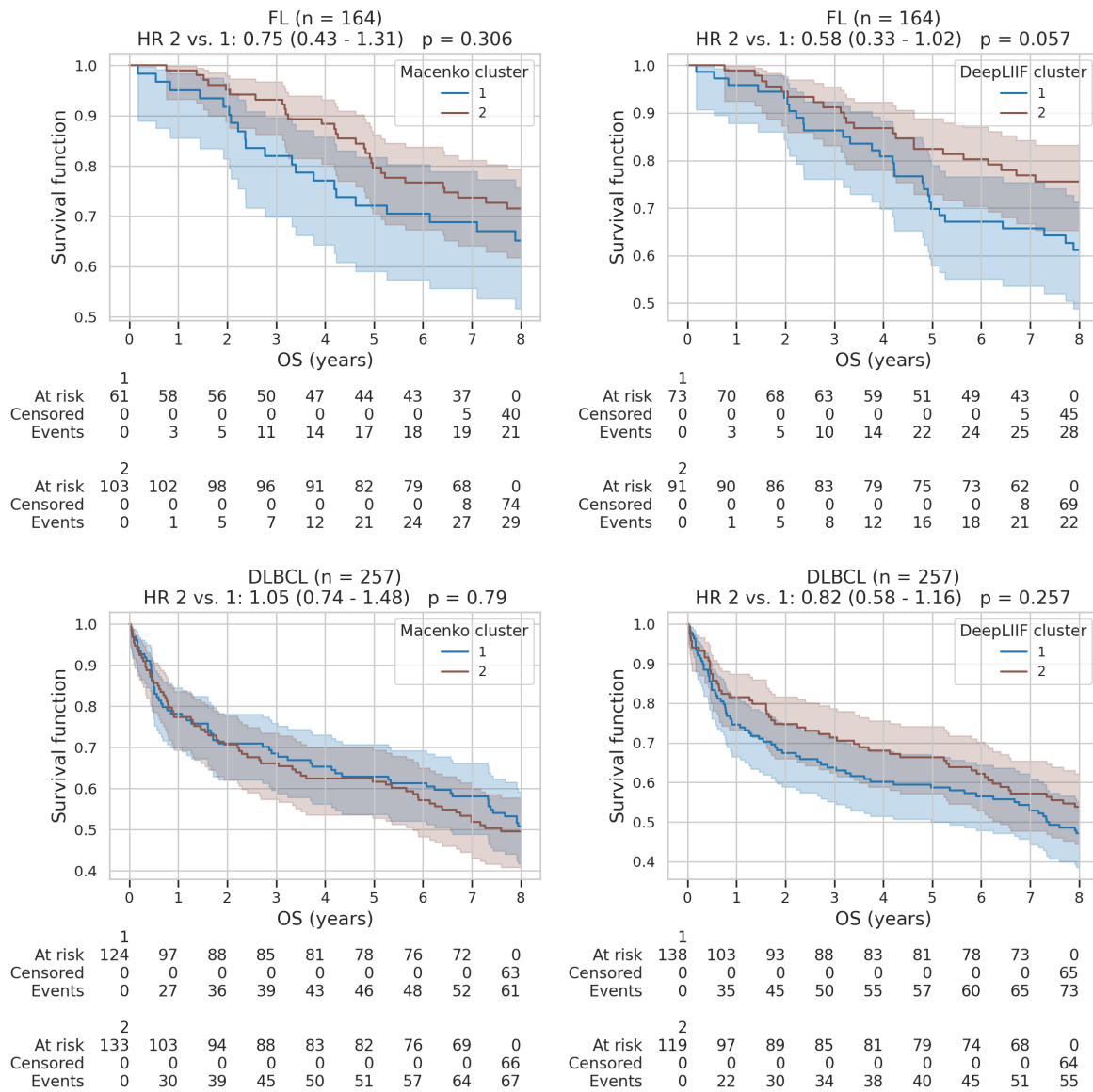
Finally, 74 patients were assigned to the smallest Macenko cluster (3). This was slightly higher than the smallest FL cluster in binary clustering work (FL Macenko cluster 2 $n = 64$) but, taking into account potential for missing or poor quality H&E slides, sample number was still low for 5-fold cross validation training/evaluation. Since HMRN slide scanning efforts were ongoing, the decision was made to await more slides to conduct training of classifiers to predict these IHC clusters using H&E.

4.3 Survival Analysis

Overall survival (OS) was compared between IHC clusters using Cox model hazard ratios (Figure 4.8A, below; complete hazard ratio data provided in Supplementary files, Chapter 3).



(a) Hazard ratios



(b) Kaplan-Meier plots

Figure 4.8: Survival analyses conducted on HMRN data, for which overall survival data was available:

A) Cox model hazard ratios for IHC clusters, fitted using a univariate approach (errorbars represent 95% confidence intervals)

B) Kaplan-Meier plots of overall survival

Notes: Cox models were fitted to binary cluster membership (cluster 1 vs. 2).

FL

Cluster 1, identified to have lower cell density above, displayed slight to moderate increased risk compared to cluster 1 in both Macenko (HR 1 vs. 2: 1.34; 95% CI: 0.76-2.35;

$p = 0.31$) and DeepLIIF IHC clusters (HR 1 vs. 2: 1.72; 95% CI: 0.98-3.01; $p = 0.06$). No differences were observed between patients initially undergoing immunochemotherapy and Watch & Wait (Supplementary files; section 3; Hazard ratios).

Kaplan-Meier estimators evidenced greatest survival differences from two to five years in Macenko clusters, with apparent gradual convergence in OS at 8 years (Figure 4.8B, top left). DeepLIIF clusters diverged at 5 years with abnormally high event number in cluster 1 (Figure 4.8B, top right).

Discussion Survival differences were slight between clusters. A chance increase in event number may have been responsible for apparent survival differences in DeepLIIF clusters, since event number was not noticeably different at any other time point.

2 cluster DLBCL

No apparent hazard was observed in Macenko clusters, with HRs close to 1. No difference was observed in Kaplan-Meier plots, either (Figure 4.8B, bottom left). DeepLIIF cluster 1 appeared to have slight increased risk, with respect to cluster 2 (HR 1.22: 95% CI: 0.86-1.74; $p = 0.26$). Kaplan-Meier plots evidenced that patient survival diverged at 1 year, maintaining survival function difference of 0.1 for the remainder of the 8-year period (Figure 4.8B, bottom right).

Discussion No survival differences were observed in Macenko clusters. Patients belonging to DeepLIIF cluster 1 appeared to experience greater hazard than cluster 2 after 1 year with visible separation between survival functions till the 5 year timepoint. This evidenced that cluster 1 patients, found to have higher Ki67 staining and thus, possibly higher neoplastic cell proliferation rate, may have been at greater risk in the first 5 years post-diagnosis.

3-5 cluster DLBCL

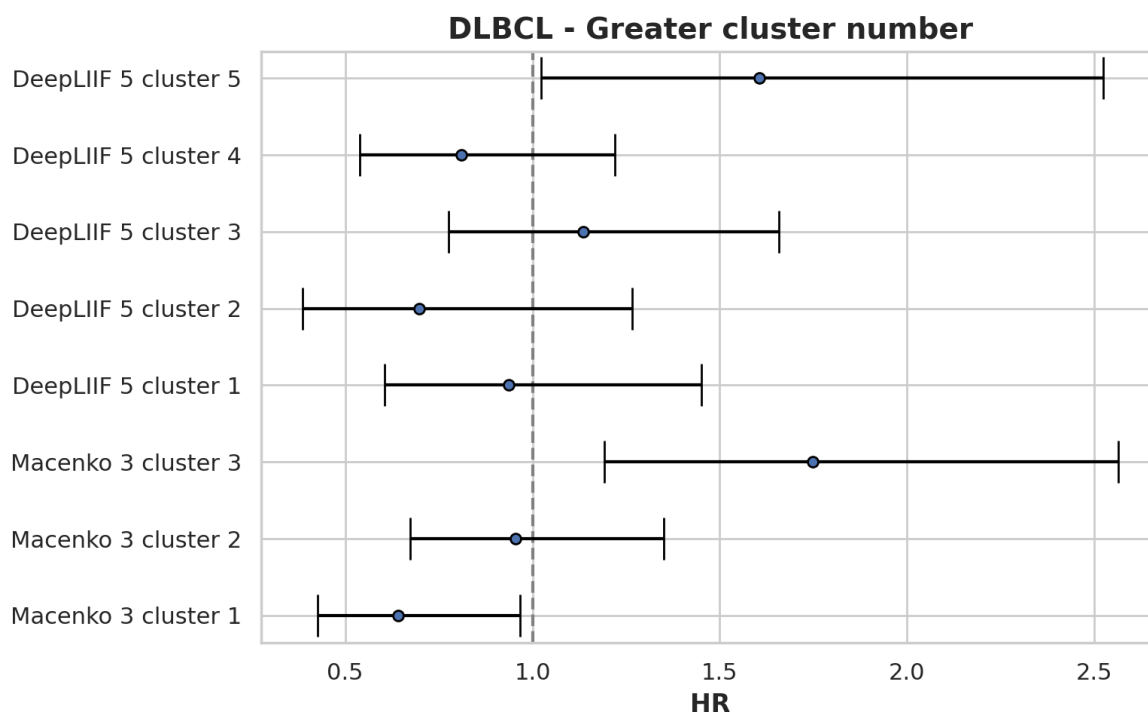
Analyses of Cox model hazard ratios (HRs) and Kaplan-Meier (estimator) plots were used to assess clinical utility of increased cluster number in DLBCL with respect to

overall survival (OS). Full HR data provided in Supplementary files, Chapter 3, HR data spreadsheet.

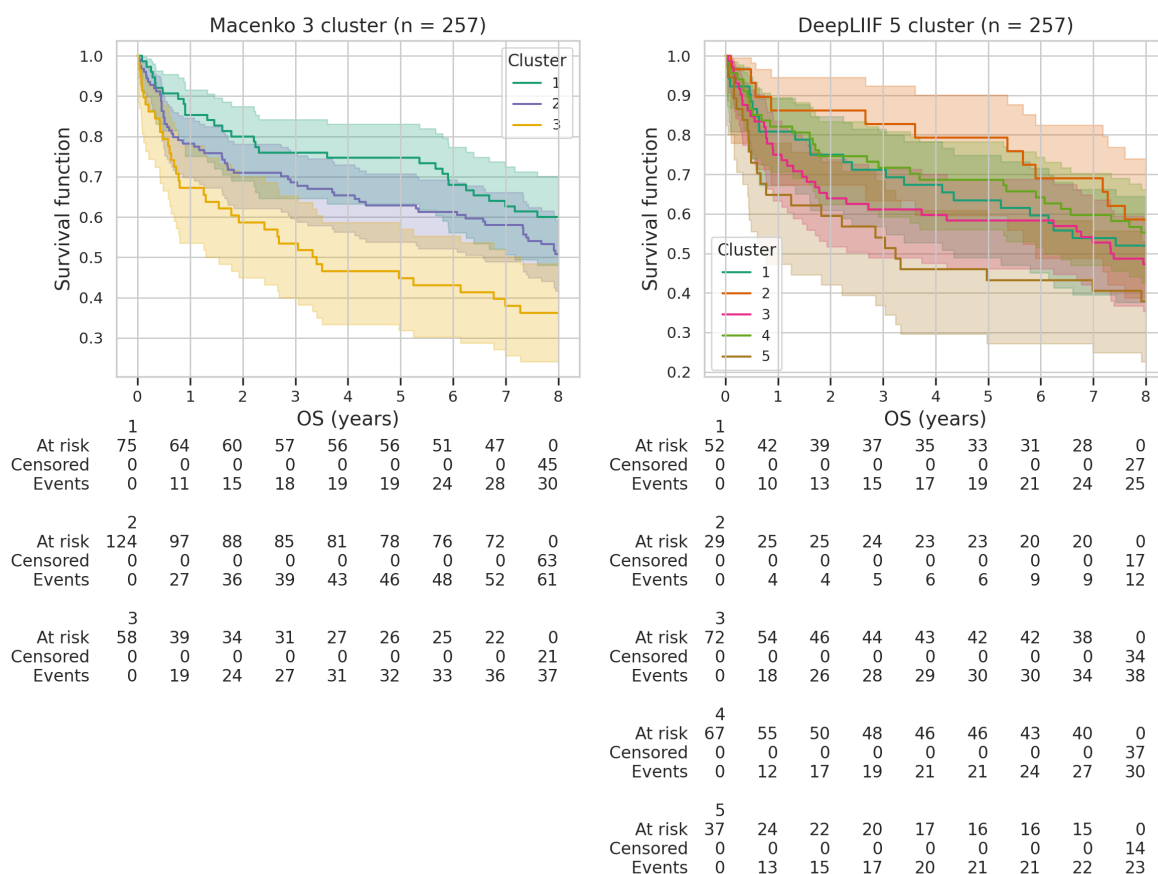
Macenko IHC feature cluster 1 was associated with significantly low hazard (HR vs. rest: 0.64; 95% CI: 0.43-0.97; $p = 0.03$); no hazard was observed with cluster 2 (HR vs. rest: 0.95; $p = 0.79$) and significantly high hazard was observed with cluster 3 (HR vs. rest: 1.75; 95% CI: 1.19-2.56; $p = 0.004$; Figure 4.9a).

Kaplan-Meier estimators showed highest OS event rates within 2 years in clusters 1 and 2, followed by relative stability in OS (Figure 4.9b). In cluster 3, high event rates were observed till 3.5 years after diagnosis. Event rates appeared approximately similar after 4 years.

DeepLIIF IHC clusters appeared to follow two general patterns of OS, with higher hazard observed in cluster 5 (HR vs. rest: 1.61; 95% CI: 1.02-2.52; $p = 0.04$) and slight hazard with cluster 3. Kaplan-Meier estimators showed that, like with Macenko cluster 3, event rate was comparatively high in the first 3 years but very low after this time.



(a) Hazard ratios



(b) Kaplan-Meier plots

Figure 4.9: Survival analyses conducted on HMRN data, for which overall survival data was available:

A) Cox model hazard ratios for IHC clusters, fitted using a univariate approach, against all other clusters (errorbars represent 95% confidence intervals)

B) Kaplan-Meier plots of overall survival

Discussion DeepLIIF cluster 5 had the highest proportion of patients with the high risk ABC expression profile (42%)[19]. Higher OS hazard therefore fits expected survival trends for this group. Cluster 3 was enriched for the, classically, lower risk GCB expression though, no reduction in OS risk was observed in this group. Cluster 2 was enriched for patients with unclassified expression (UNC) and had the best outcomes of all groups, though no significant reductions in hazard were observed. Overall, clusters 5 and 2 appear most useful with respect to patient outcomes. Analyses of mutation found that groups enriched for UNC expression had consistently good survival[21], evidencing that better OS in cluster 2 could also be attributed to constituent expression profiles.

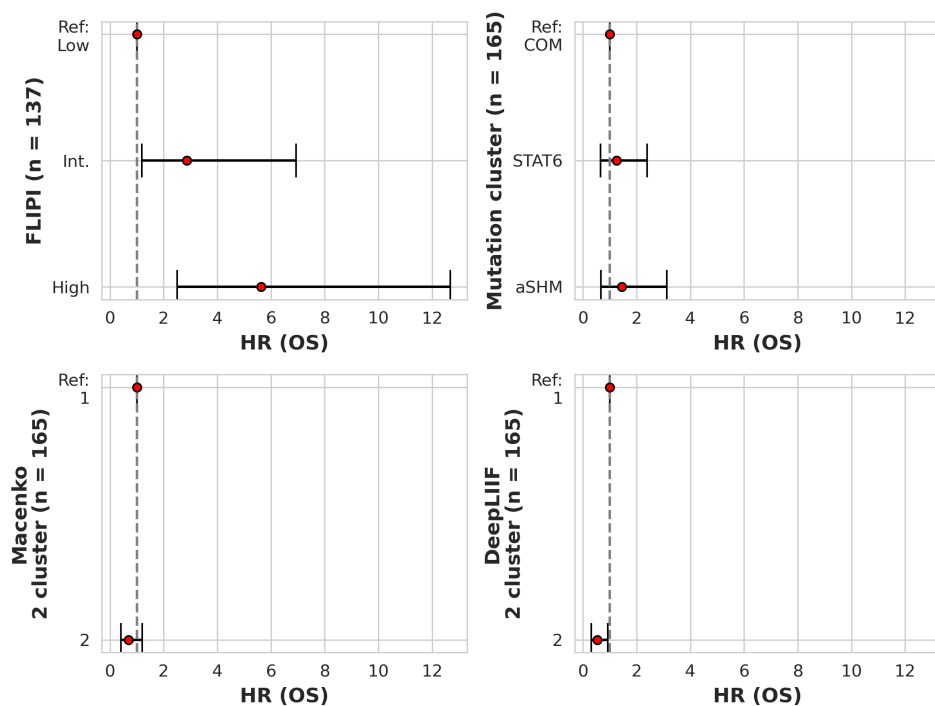
Macenko cluster 1 was found to be enriched for GCB while cluster 3 was enriched for ABC expression. GCB has been found to associate with better survival while ABC associates with worse outcome[19]. These effects appear to largely persist across mutation patterns (with the exception of better survival with a *BCL6* mutation group in ABC)[21]. These observations evidence that expression profile accounts for observed survival.

Overall, the utility of the 3 cluster Macenko method appeared far greater than the 5 cluster (DeepLIIF) approach, though survival in both appeared to be accounted for by ABC expression and partially by GCB (Macenko) and UNC (DeepLIIF).

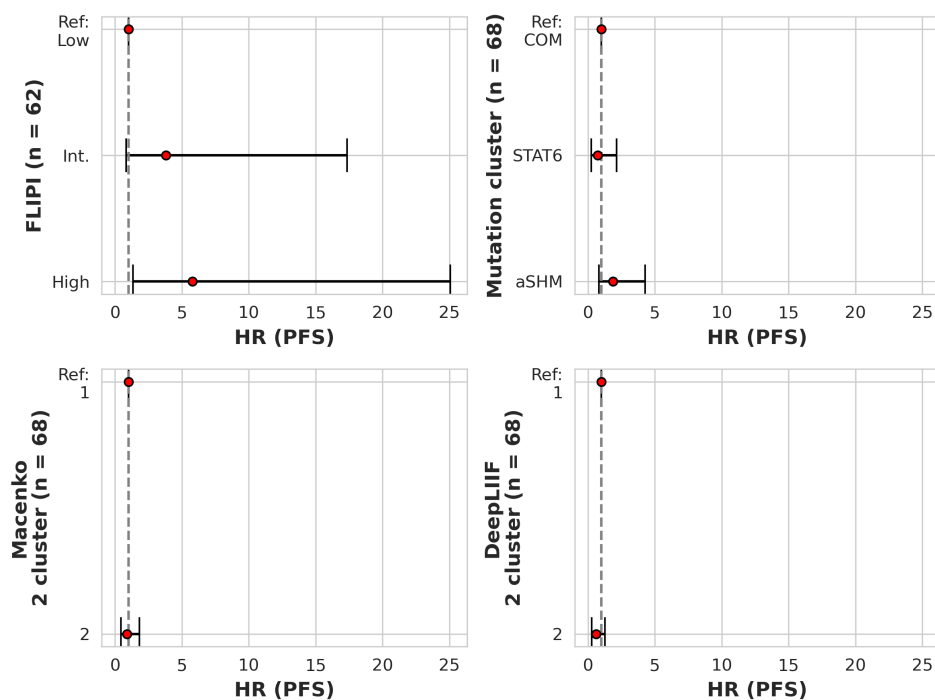
Comparison with existing predictors

Overall survival (OS) and progression-free survival (PFS) hazard of IHC clusters were compared against existing clinical markers of risk including clinical indices, IPI and FLIPI, expression profiles in DLBCL (cell-of-origin; COO) and mutation in FL.

Follicular Lymphoma



(a) OS Follicular Lymphoma

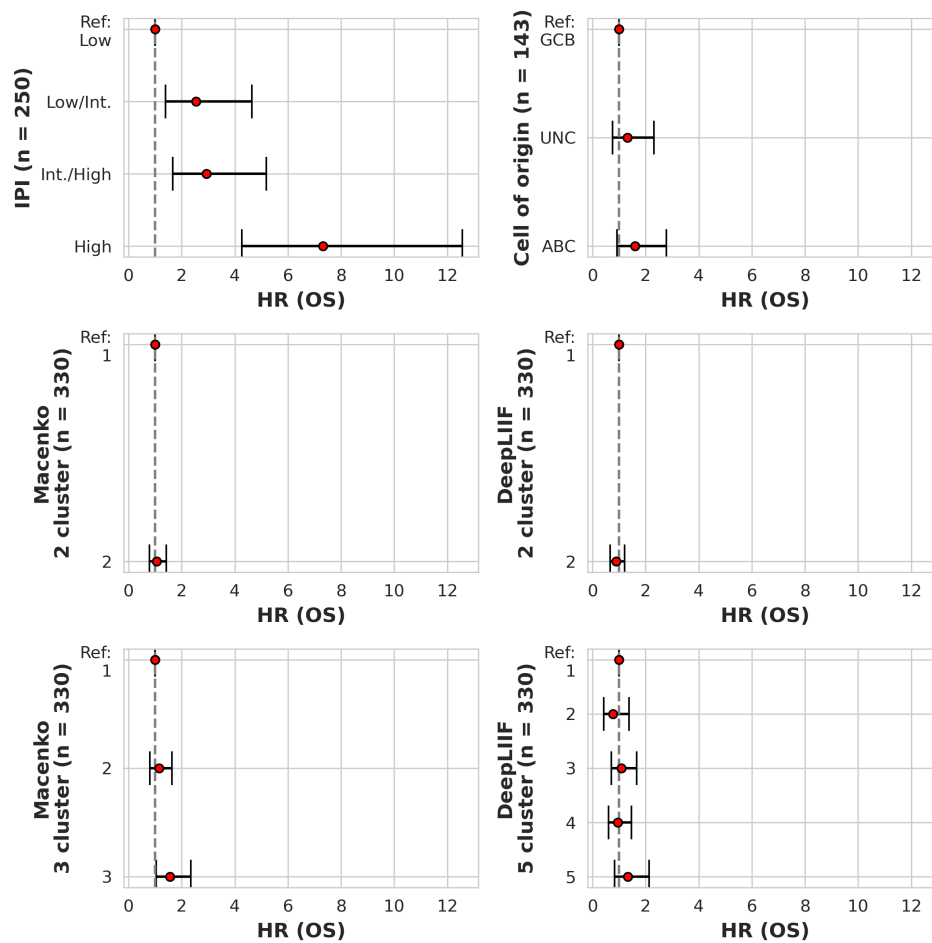


(b) PFS Follicular Lymphoma

Figure 4.10: Hazard ratios for established indices (FLIPI and mutation cluster) against IHC clusters on HMRN FL data. **Notes:** survival of groups was compared against reference groups for each index.

Intermediate and high FLIPI were associated with significant increased OS hazard of 2.87 ($p = 0.02$) and 5.63 ($p < 10^{-4}$), respectively (Figure 4.10a). No OS associations were observed between mutation profiles or Macenko IHC clusters. DeepLIIF IHC cluster 2 was associated with significantly decreased hazard compared to cluster 1 (HR 2 vs. 1 = 0.54; 95% CI: 0.31 - 0.93; $p = 0.03$). While this effect was far smaller than with high FLIPI, it was comparable in scale to the OS hazard of intermediate FLIPI. Due to low event number, PFS comparisons were underpowered in FL (Figure 4.10b). Significant hazard was only observed with respect to high FLIPI (HR vs. reference = 5.78; $p = 0.02$).

DLBCL



(a) OS DLBCL

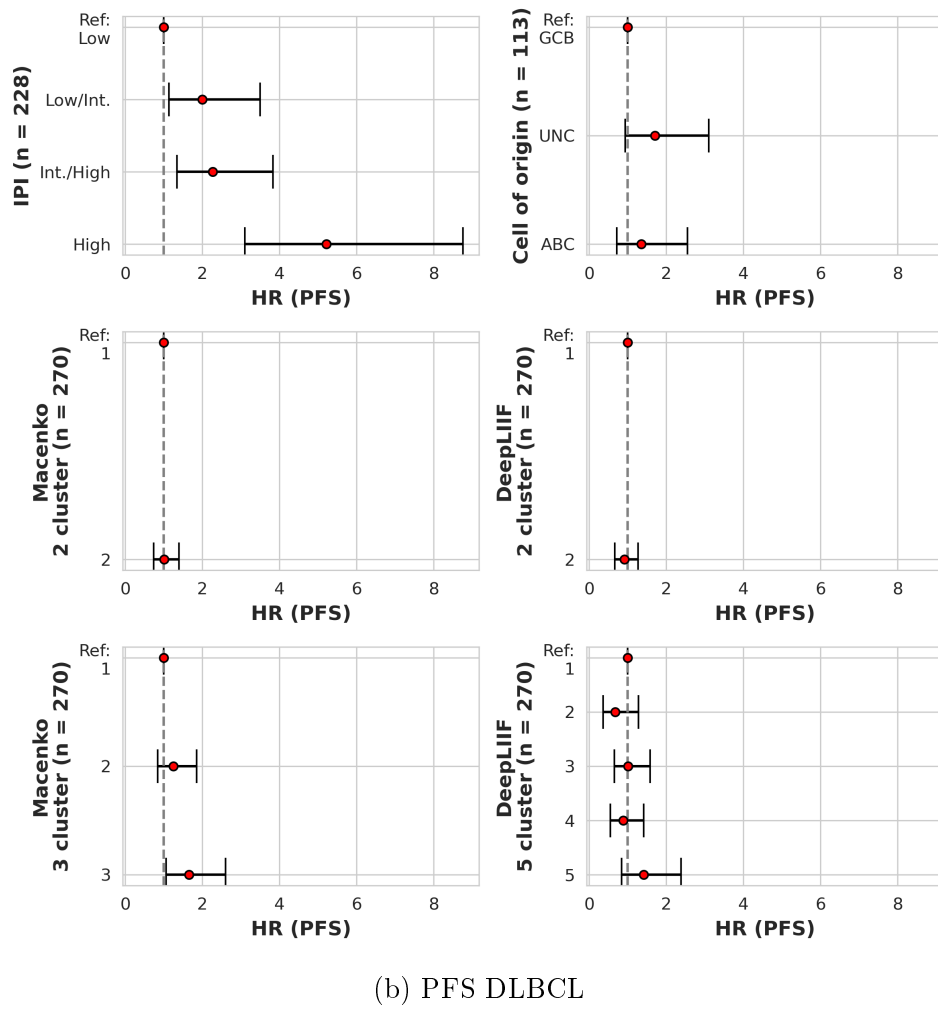


Figure 4.11: Hazard ratios for established indices (IPI and cell-of-origin, COO, expression) against IHC clusters on HMRN DLBCL data. **Notes:** survival of groups was compared against reference groups for each index.

Increased IPI was found to associate with increased OS hazard (4.11a), from a hazard ratio of 2.53 for the low/intermediate IPI against the reference group ($p < 0.01$) to 7.31 for high IPI ($p < 10^{-12}$). PFS hazards were comparable in scale (from 2.00 for low/intermediate IPI to 5.22 for high IPI) and also significant (4.11b). With respect to COO expression and mutation, significant survival differences were only observed in cluster 3 of the (3 cluster) Macenko IHC groups. OS hazard for this group was 1.56 against the reference (95% CI: 1.04 - 2.34; $p = 0.03$) while PFS hazard was 1.66 against the reference (95% CI: 1.06 - 2.60; $p = 0.03$). These results evidence a more subtle hazard associated with Macenko cluster 3 than observed with IPI and by extension, clinical characteristics of patients. Despite this, hazard ratios were of a comparable scale to those observed for

ABC expression, an established feature of high risk DLBCL.

4.4 Validation

Prediction of IHC clusters in H&E

Graph Neural Network (GNN) models[85] were trained to predict IHC clusters using routine H&E slide features. Nuclear features describing morphology, intensity and texture were evaluated as well as self-supervised pathology image embeddings(UNI model feature[82]).

Optimal feature types were identified using 5-fold cross validation. Performance of models trained on UNI and nuclear features were comparable, with overlapping interquartile ranges and no significant differences observed (Figure 4.12). Mean performance was higher for UNI features in prediction of FL DeepLIIF (AUC-ROC UNI 0.85 vs. 0.80 nuclear; Figure 4.12 row 2) and DLBCL Macenko (AUC-ROC UNI 0.68 vs. 0.62 nuclear; Figure 4.12 row 3) clusters. Nuclear features outperformed UNI features in prediction of FL Macenko clusters (AUC-ROC UNI 0.66 vs. 0.61 nuclear; Figure 4.12 row 1).

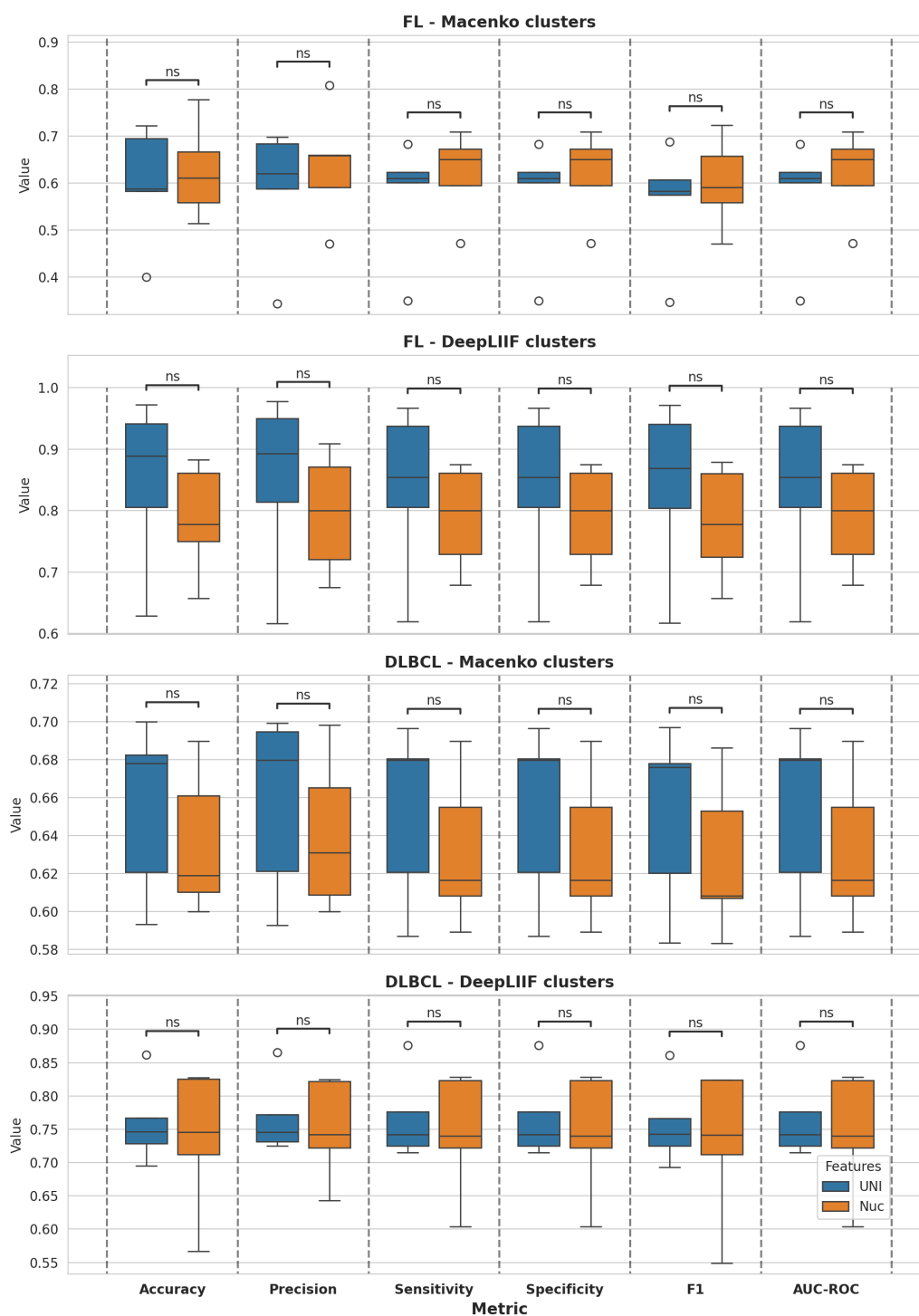


Figure 4.12: Supervised metric results for classification of IHC clusters with H&E slide features (errorbars represent 95% confidence intervals). IHC cluster types included those fitted with DAB intensity features extracted using Macenko stain separation (Macenko clusters) and cellular IHC features extracted using the DeepLIIF model[176]. H&E features included UNI model embeddings[82] and nuclear features (listed in Appendix Section 4). P-value key: ns = not significant (at p-value threshold of 0.05).

Models trained to predict DeepLIIF clusters using UNI features were found to perform better than those trained to predict Macenko clusters, achieving AUC-ROC of 0.86 in FL (vs. 0.61 for Macenko clusters; Figure 4.13 top) and 0.74 in DLBCL (vs. 0.68 for Macenko clusters; Figure 4.13 bottom). At the fold number used (5), these difference were not significant, despite no overlap in interquartile range (Figure 4.13).

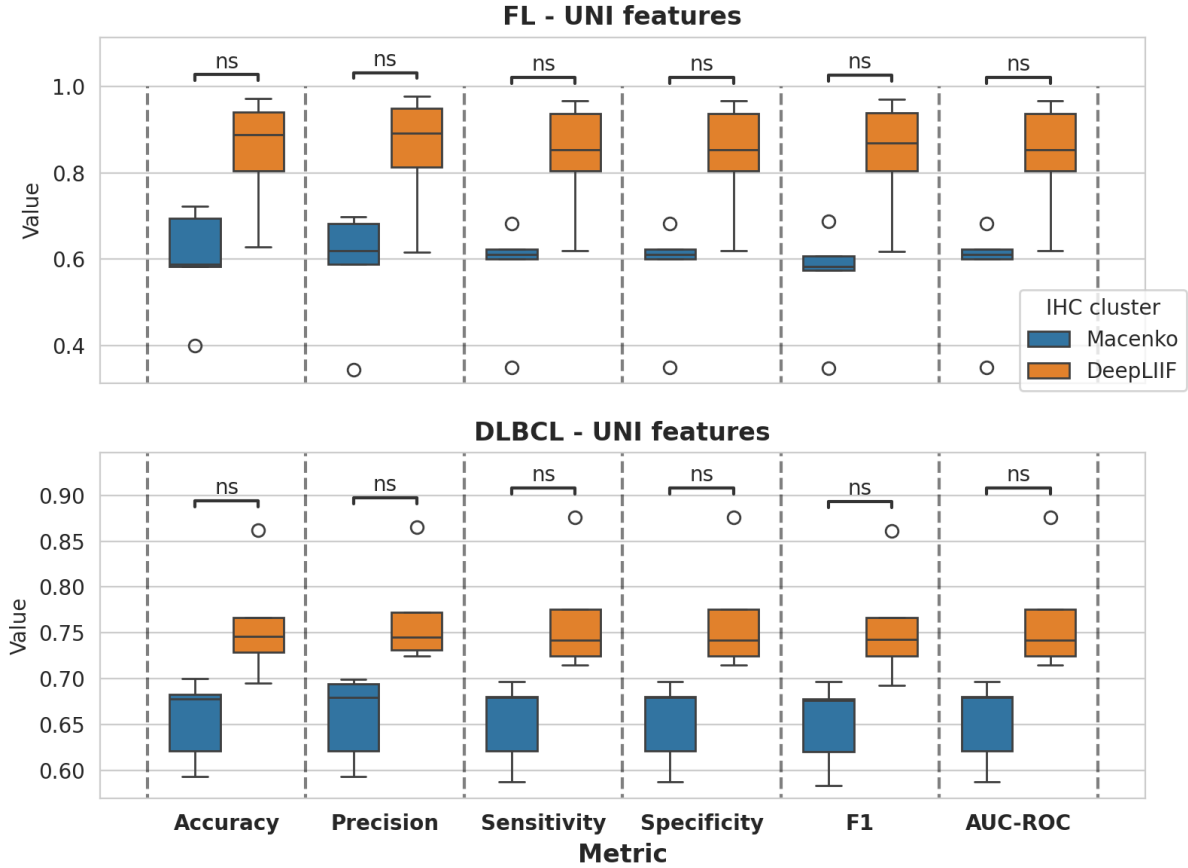


Figure 4.13: Supervised metric results for classification of IHC clusters with respect to the method used to quantify IHC (errorbars represent 95% confidence intervals). This plot used identical UNI performance data to Figure 4.12, above, with comparison of performance in predicting types of IHC cluster (quantification methods), with the purpose of assessing which method was more feasibly predicted in H&E. IHC cluster types included those fitted with DAB intensity features extracted using Macenko stain separation (Macenko clusters) and cellular IHC features extracted using the DeepLIIF model[176]. P-value key: ns = not significant (at p-value threshold of 0.05).

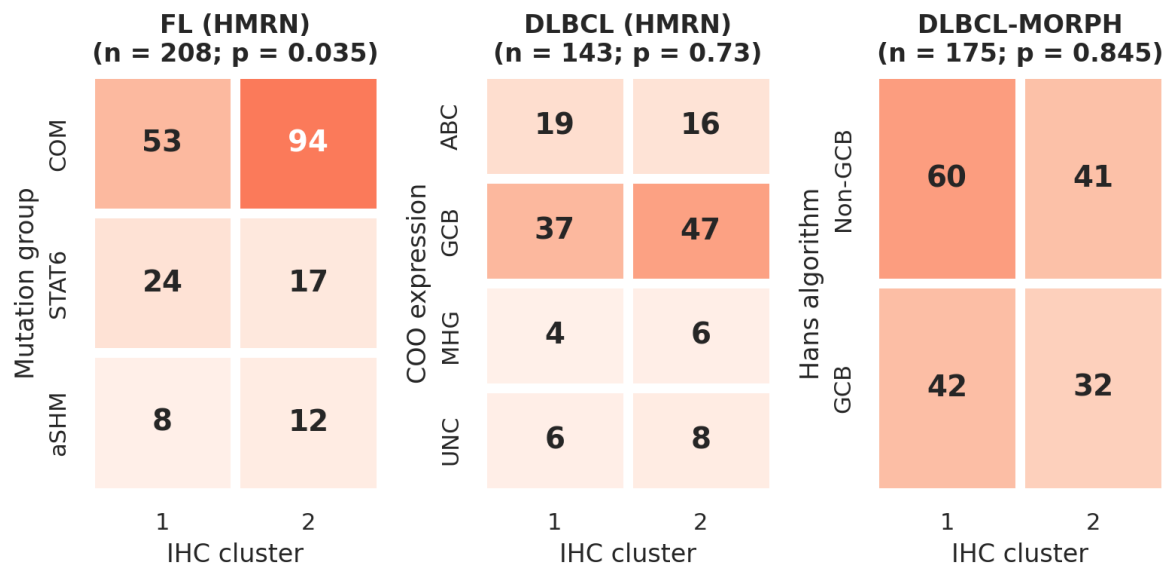
Discussion Nuclear features have been used extensively for classification in histopathology[66, 85, 103, 139]. An advantage is often in ease of interpretability, with differences

between groups clearly explained by differences in nuclear populations. The results observed in classification of IHC clusters evidenced slight advantages of patch-based approaches in some prediction tasks, supporting previous work in prediction of molecular features[82, 187].

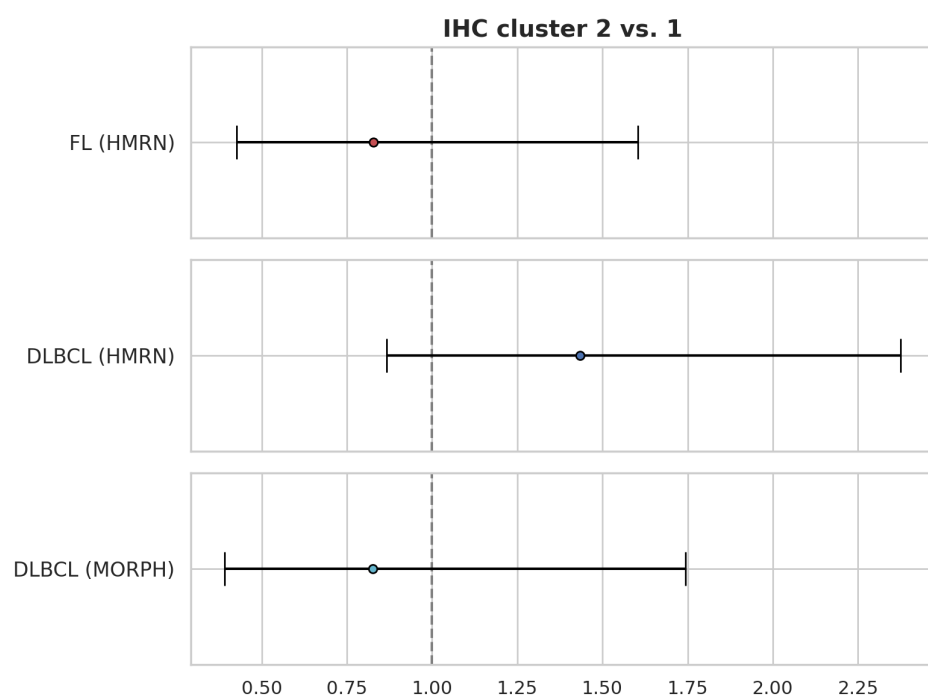
Several factors may have led to marginally lower performance, compared to state-of-the-art UNI features, in prediction of some IHC clusters. Firstly, nuclear approaches do not account for differences in cytoplasm, which are key for distinction of cells such as macrophages[188]. Secondly, by considering non-nuclear regions too, patch-level feature embeddings can inherently contain information on extracellular matrix and microenvironment, which have been evidenced to have involvement in FL[65, 189] and DLBCL[190].

Internal validation

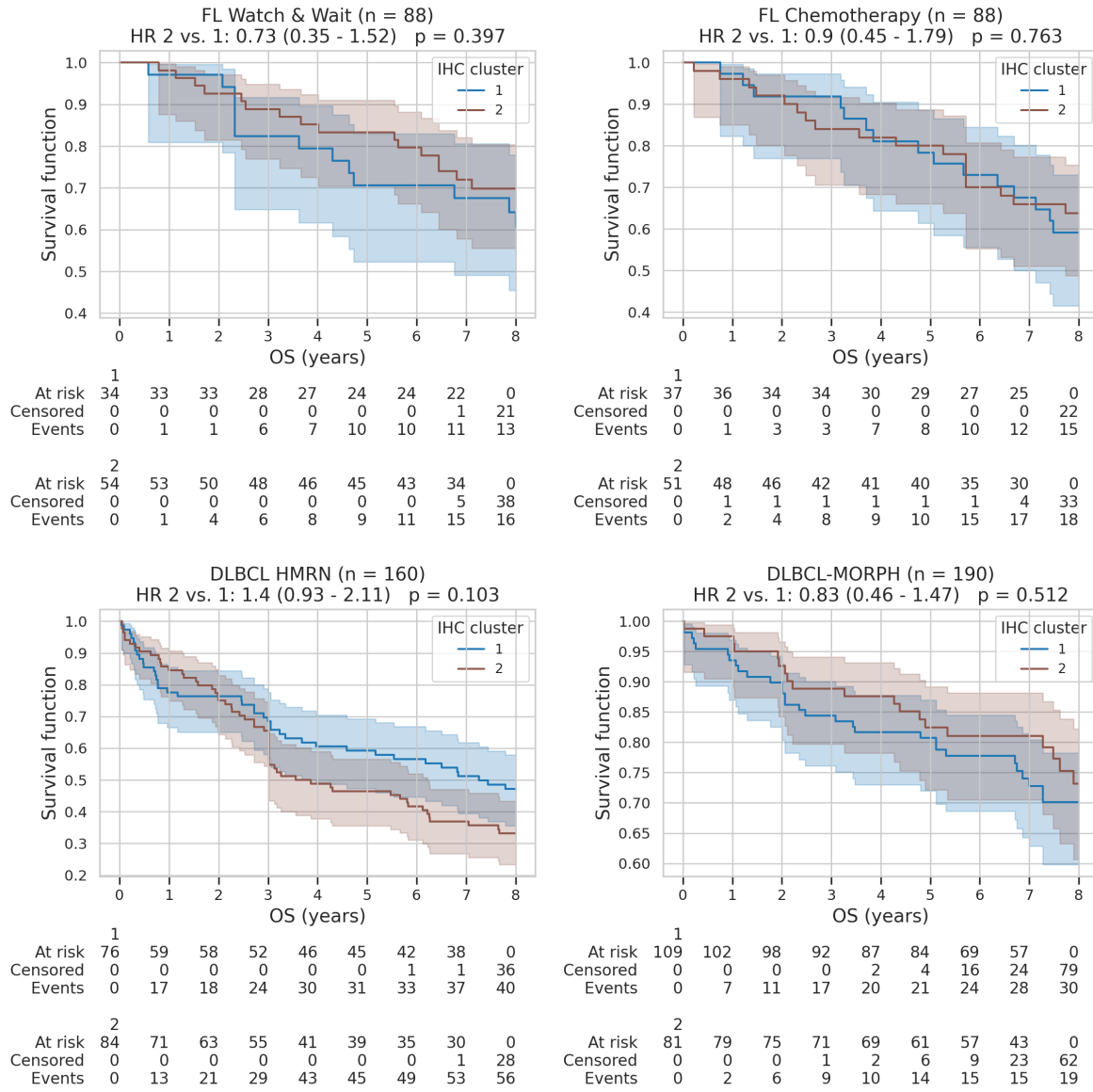
HMRN H&E slides without associated IHC data were used for internal validation of DeepLIIF IHC cluster patient characteristics: molecular features and overall survival (Figure 4.14A and B/C, respectively). H&E slides of the DLBCL-MORPH dataset[66] were used for external validation of DeepLIIF DLBCL cluster features.



(a) Molecular feature contingency



(b) Hazard ratios



(c) Kaplan-Meier plots

Figure 4.14: Clinically-relevant characteristics in IHC clusters (DeepLIIF), predicted with H&E slides. Predicted using UNI feature embeddings from H&E patches at 20 $\times$ magnification.

A) Contingency matrices of predicted IHC clusters against molecular features (p-values calculated using χ^2 contingency)

B) Cox model hazard ratios (errorbars represent 95% confidence intervals), fitted to probability of cluster 2 membership

C) Kaplan-Meier plots for predicted IHC clusters, fitted to maximum likelihood clusters

Molecular features

No significant differences were observed in DLBCL IHC cluster frequency with respect to cell-of-origin expression (predicted using the Hans algorithm in DLBCL-MORPH).

Significant differences were observed in FL cluster frequencies, with higher proportion of patients bearing COM mutations in cluster 2 and STAT6 in cluster 1.

Discussion Results for DLBCL were consistent with H&E-matched training data (Figure 4.4E), both not evidencing any differences in molecular features between groups. In FL, validation results were not consistent with training data in evidencing a difference in molecular feature frequency. IHC analyses were not typical for FL in HMRN data, with fewer than half referred by pathologists (185/403; 46% of all). This indicated that the cases with IHC data available were more challenging to classify in H&E, requiring differential diagnosis against DLBCL[23, 24]. FL with a diffuse pattern (discussed in Chapter 3 Section 4.6) has previously been found to have higher rates of STAT6 mutation[157–159]. A higher proportion of such cases in the training dataset, referred for IHC analysis, may have skewed frequency of cases bearing STAT6 mutations. This, in turn, may have ‘masked’ the effect observed in cases not referred for IHC analysis, for which pathologists were, presumably, sufficiently confident of an FL diagnosis.

Survival

Validation data hazard ratios for FL and DLBCL-MORPH were comparable in magnitude to those obtained in training data, while HMRN DLBCL validation data HRs were not. The following results were obtained for overall survival, fitted to SlideGraph model confidence of DeepLIIF cluster 2 membership (Figure 4.14B; complete HR data available in Supplementary files, Chapter 3):

- FL training data (all treatments) HR = 0.58 (95% CI = 0.33-1.02; p = 0.06) vs. validation data HR (all treatments) = 0.83 (95% CI = 0.43-1.60; p = 0.58)
- DLBCL training HR = 0.82 (0.58-1.16; p = 0.26) vs. DLBCL-MORPH HR = 0.83 (0.39-1.74; p = 0.62)

- DLBCL training HR = 0.82 (0.58-1.16; $p = 0.26$) vs. HMRN DLBCL validation HR = 1.43 (0.87-2.37; $p = 0.16$)

Kaplan-Meier estimators were fitted to predicted cluster membership (binarised into cluster 1 vs. cluster 2 using a confidence threshold of 0.5). Comparison of survival over time (Figure 4.14C) identified increased event number in FL Watch & Wait cluster 1 patients between the year 2 and 3 time points, leading to divergence in cluster survival. Survival in clusters did not appear to diverge for FL patients who initially underwent immunochemotherapy while DLBCL-MORPH cluster 1 survival appeared slightly, yet consistently, worse than cluster 2 over all time points. Finally, HMRN DLBCL patients belonging to cluster 2 experienced increased event number between years 2 and 4, leading to divergence in survival (with respect to cluster 2).

Discussion In FL, the increased event number observed in training data cluster 1 (at 5 years) was not observed in validation data. Survival was not observed to diverge between clusters. While somewhat consistent hazard ratio values evidence a potential slight increased hazard for cluster 1 patients, the sample sizes available were too small to determine magnitude with acceptable confidence.

DLBCL-MORPH and training data survival appeared consistent, with higher hazard for cluster 1 across all time points. HMRN DLBCL validation data appeared consistent with training data after 3 years, with similar event rates between clusters observed in both datasets. In both datasets, cluster 1 survival appeared slightly worse than cluster 2 in the first year, with similar difference in cluster survival functions (0.10). Validation and training datasets differed in event rate between one and two years, with an 8% increase in percentage of patients with events (25% in year 1 and 33% in year 2) in training data and only 2% in validation data (22% year 1 to 24% year 2).

Overall, limited consistency was found in FL survival between training and validation data, indicating that survival differences observed between clusters may not have reflected general survival differences in FL. On the hand, general consistency in DLBCL survival between datasets evidence a potential small but general increased hazard in cluster 1

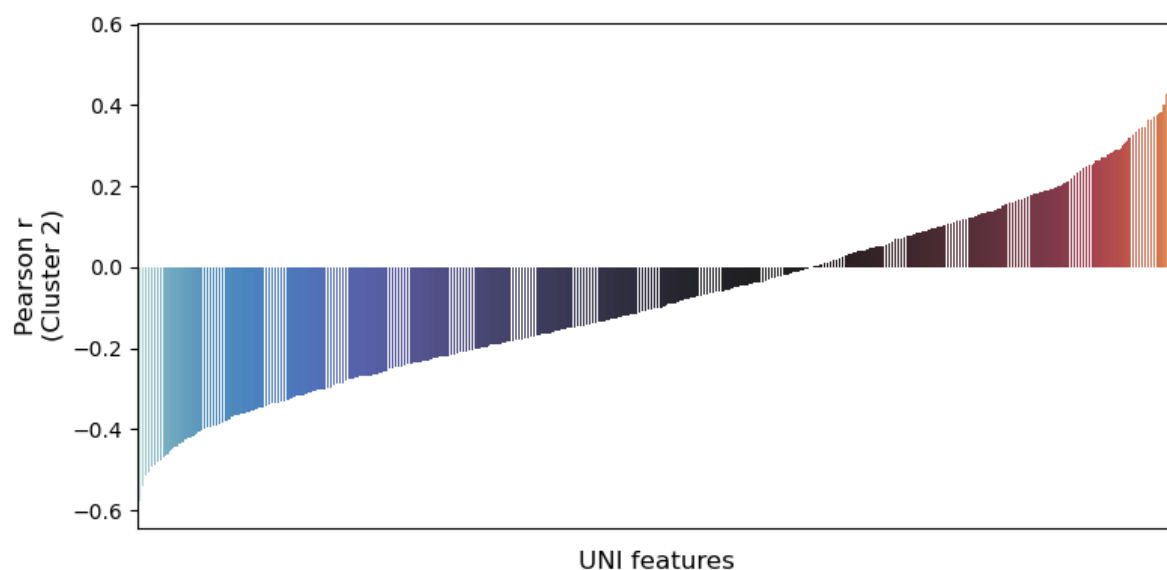
patients.

Interpretation of H&E models

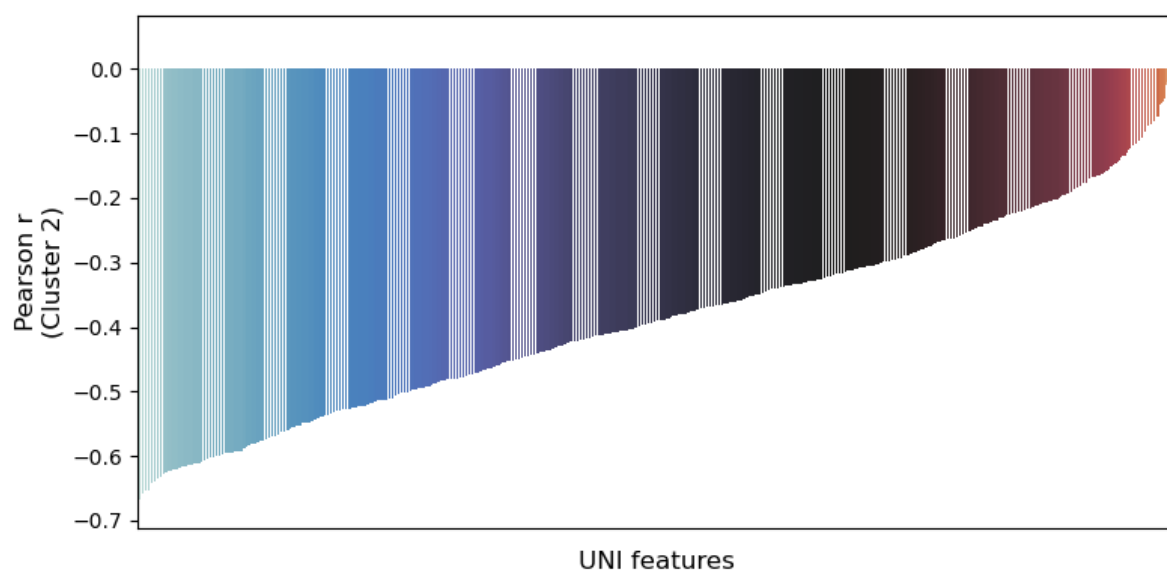
IHC cluster predictions potentially held valuable insights for use in routine pathologist analysis of H&E slides. Value was therefore identified in firstly, explanation of model decisions on H&E and subsequently, interpretation of which slide attributes could feasibly be used by pathologists.

Important features For model explanation, GNN Explainer[182] was used to determine UNI feature importance in slide-level classification (Figure 4.15, below). Graph subnets, as used in previous GNN explanation efforts[191] were not used due to high numbers of features with non-zero importance, leading to convoluted subnet representations.

Correlation (Pearson) was calculated between UNI feature importance and SlideGraph model confidence of cluster 2 membership over HMRN validation datasets.



(a) Follicular Lymphoma



(b) Diffuse Large B-cell Lymphoma

Figure 4.15: Association between feature importance and UNI features for prediction of IHC cluster 2 (in the respective diseases), extracted using GNN Explainer[182] over the HMRN H&E validation datasets (FL $n = 218$; DLBCL $n = 178$). UNI features ($n = 1,024$) on the x-axis are ordered from lowest to highest importance in prediction of cluster 2 (of the respective disease). Each bar represents an individual feature. Negative Pearson correlation values indicate association between a feature and classification of cluster 1. Colour used to emphasise correlation values.

FL A small number of features were found to have noticeably higher correlation with predicted IHC cluster. Top features in cluster 1 (negative correlation) and 2 prediction had correlation values 0.03 and 0.15 greater in magnitude than the next most informative features (Figure 4.15a; extreme right and left sides). Approximately half of features were found to correlate weakly with survival, with absolute correlation equal to or below 0.2.

DLBCL In DLBCL, the vast majority of features (over 1,000 of 1,024) were found to have negative correlation (prediction of cluster 1), with over 80% of features displaying moderate correlation (r between 0.3 and 0.6) with cluster 1 prediction (Figure 4.15b).

Discussion In FL, the vast majority of features did not appear to be informative, with fewer than 15% correlating moderately (above 0.3) with model predictions. This evidenced that a large portion of features used had little impact on classification, potentially negatively influencing training and complicating model interpretation. Future work may aim to improve classification performance, SlideGraph model efficiency and ease of classification interpretation through feature selection in data preprocessing.

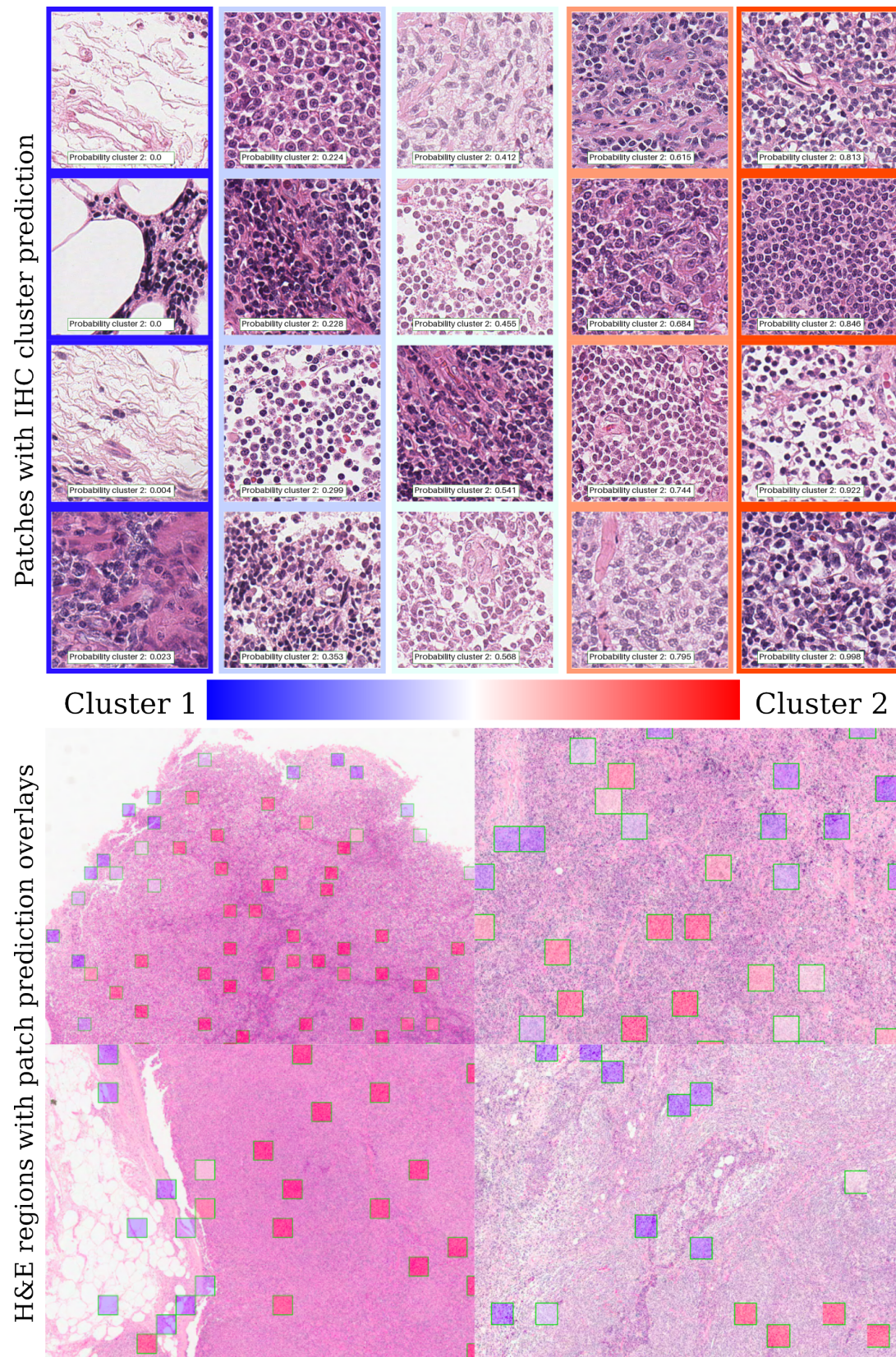
In DLBCL, a large portion of features displayed moderate correlation between calculated importance and model classification, possibly as a result of ‘dropout’ during training. In this process, SlideGraph model perceptrons were randomly masked to promote reliance on a wider feature range in prediction. As with FL, presence of a large number of features with similar correlation values with cluster prediction evidenced high levels of redundancy in features and potentially, unnecessary complication in the SlideGraph model trained. Combination of features using dimensionality reduction approaches may have aided in reduction of feature number, while conserving information. As with FL, elimination of features with low correlation values may have reduced noise and model complexity, also.

Interpretation of patch features

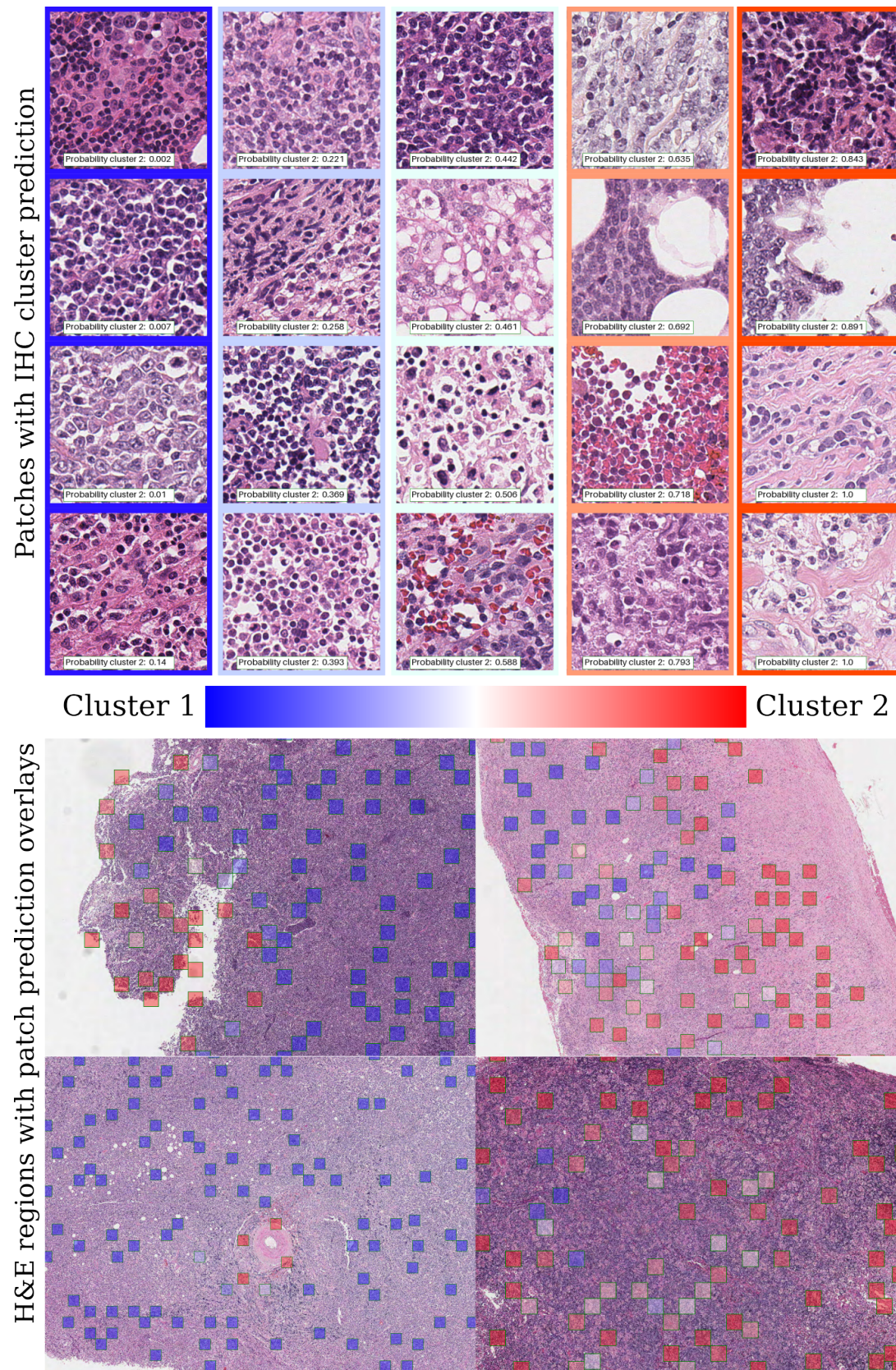
Visual inspection was used for interpretation of tangible H&E features in IHC cluster predictions (Figure 4.16).

FL Image patches predictive of cluster 2 were mostly distributed in the interior of tissue, while those classified as cluster 2 were largely observed in fibrous, fatty and tissue edge regions (Figure 4.16A, bottom). Comparison of patches over a range of confidence values identified that fibrous and fatty tissue was assigned high confidence values for cluster 1 (confidence cluster 2 < 0.2) while no noticeable differences were observed in patches with lower confidence (confidence cluster 2 ≥ 0.2 ; Figure 4.16A, top).

DLBCL Patches on tissue periphery, vasculature and fibrous regions were largely classified as cluster 2 (Figure 4.16B). Nonetheless, no visible distinctions were observed between clusters in patch images. Interior tissue region classification was observed to be more heterogeneous than in FL, with more subtle differences between clusters.



(a) Follicular Lymphoma



(b) Diffuse Large B-cell Lymphoma

Figure 4.16: Sample H&E patches (top) and regions (bottom) with corresponding IHC cluster predictions for the FL (A) and DLBCL (B) classifiers.

Top) Randomly sampled patches with corresponding IHC cluster prediction (binarised as probability cluster 2 in bottom left of patches) from highest probability cluster 1 in top left to cluster 2 in bottom right

Bottom) Sample H&E regions with predicted IHC cluster heatmaps overlaid. Colour key: Blue = cluster 1; White = uncertain; Red = cluster 2.

Discussion As indicated in interpretation of IHC clusters (Section 4.2), FL IHC cluster 1 appeared to be defined by low cell density and high proportions of fibrous, fatty and vascular tissue. In DLBCL, evidence existed that fibrous and vascular regions were largely classified as cluster 2, but these did not appear to be the sole determinants of classification. Further interpretation of DLBCL clusters was challenging for a number of reasons. Firstly, IHC clusters were themselves challenging to interpret, describing distribution features (kurtosis and standard deviation) not quantifiable, reliably, through visual inspection. Previous efforts with success in interpretability, such as the IGUANA model[191] used visually distinguishable, interpretable features including size and proportion of informative entities. Clustering of IHC data on interpretable features would likely have resulted in more readily interpretable classification models.

Secondly, UNI features were complex in nature (when compared to morphological features of nuclei, for example), with high dimensionality (1,024), describing seemingly abstract properties of images. A small number of these features were found to be more informative for the classification task than the remainder of features. Selection, training and model interpretation using that small subset of features would likely have been more efficient and feasible.

5 Conclusions

Quantification of IHC

Both tissue (Macenko proportion positive) and cell-level (DeepLIIF percentage positive cells) methods of IHC quantification were found to correlate highly with pathologist estimates of CD10, with Pearson correlation above 0.7. Performance was worse in quantification of nuclear proteins BCL6 and MUM1 (not significant for DeepLIIF; significant for Macenko proportion positive tissue, with $p < 0.01$). DeepLIIF was found to outperform Macenko (proportion positive) tissue-level quantification in BCL6 and MUM1 (though not significantly).

Discrepancies between DeepLIIF and pathologist estimates of IHC staining appeared to stem from disagreement in positive staining thresholds, especially in higher numbers of false negative classifications in regions saturated with IHC staining. Possible solutions to this disagreement may be in inclusion of unstained ‘control’ regions in DeepLIIF input data, to enforce appropriate stain normalisation.

Overall, these results evidence some utility of DeepLIIF in quantification of lymphoma IHC staining. Modifications in stain normalisation may lead to better performance against pathologist estimates still.

IHC clusters

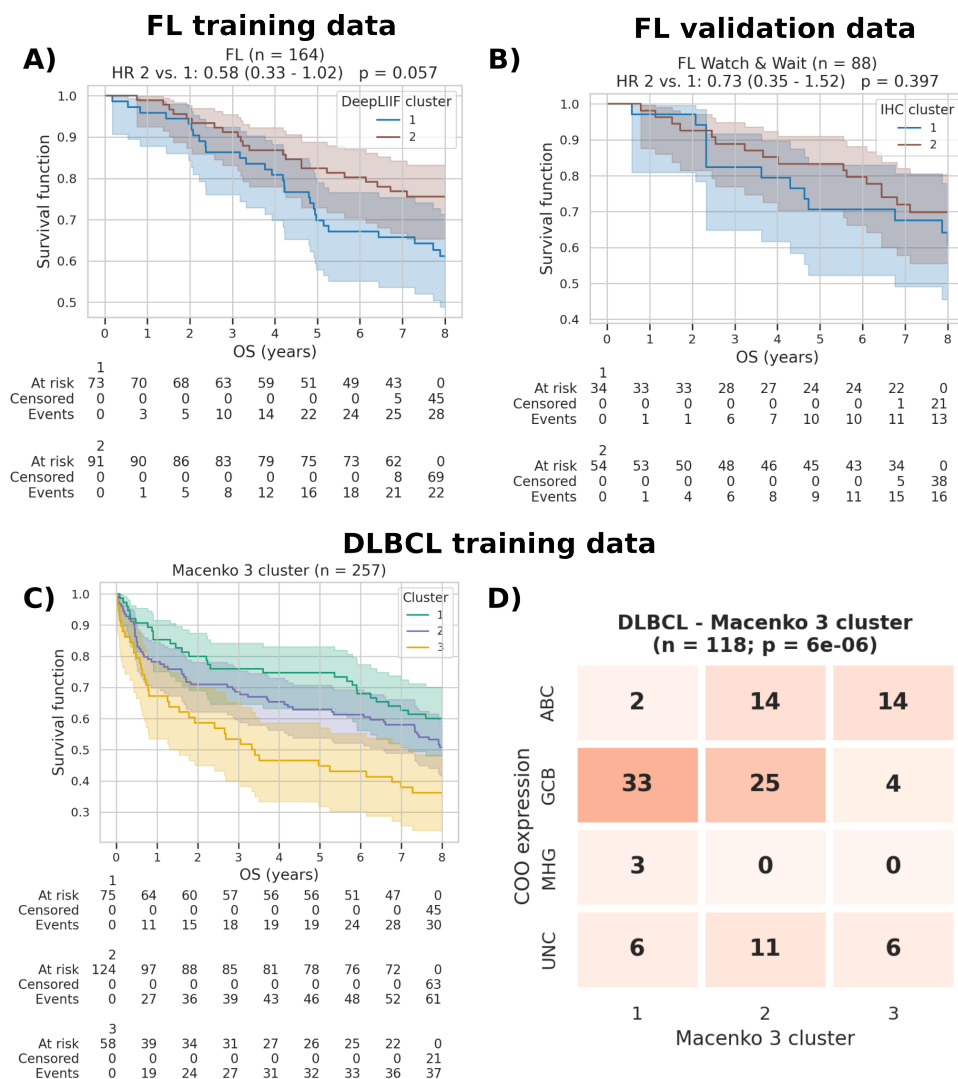


Figure 4.17: Summary of key cluster features, with respect to clinical utility:

A) Kaplan Meier estimators of FL training data, with respect to Macenko feature clusters. Hazard ratios calculated using univariate Cox proportional hazards models; significance testing performed using the Wald test.

B) Kaplan Meier estimators of FL validation data, with respect to Macenko feature clusters.

C) Kaplan Meier estimators of DLBCL training data, with respect to the 3 cluster Macenko groupings.

D) Contingency matrix of DLBCL training data, with respect to cell of origin expression and the 3 cluster Macenko groupings. Statistical significance calculated using the χ^2 test.

In both FL and DLBCL, unsupervised clustering metrics (Calinski-Harabasz score) evidenced that 2 clusters were optimal. In FL, cell-level (DeepLIIF) cluster survival patterns were found to be largely consistent between training data (Figure 4.17A) and validation data, predicted using H&E slides (Figure 4.17B). Survival hazard was found to be lower for cluster 2 against cluster 1, albeit not significantly at the sample size used.

With supervised metrics using mutation (in FL[46]) and Cell-of-origin expression (in DLBCL[47, 145, 146]) as ground truth identified optimal cluster numbers between 3 and 5. Three clusters (fitted to Macenko features) were found to separate patients well on survival, with significantly lower hazard in cluster 1 and higher hazard in cluster 3 (Figure 4.17C). These clusters were found to be enriched for specific expression profiles, with significant differences in expression profile frequency between groups (Figure 4.17D). On comparison of survival to established indices of risk, such as IPI (DLBCL), FLIPI (FL) and expression profiles (COO in DLBCL), FL DeepLIIF cluster 2 was found to associate with significantly lower OS hazard, comparable in scale to an intermediate FLIPI score. DLBCL Macenko cluster 3 was found to associate with significantly increased overall survival hazard, of a comparable scale to low/intermediate IPI score.

In FL, clustering was evidenced to associate with cell density, with lower proportions of all cells and higher proportions of fibrous tissue in cluster 1. Tissue level IHC clusters (Macenko stain deconvolution features) in DLBCL also appeared to cluster tissue on these properties, as opposed to staining of lymph-related tissues such as B-cell sheets. These results, as well as better performance of Macenko IHC (3 cluster approach) compared to DeepLIIF IHC (5 cluster approach) in stratifying patients evidenced need for more stringent IHC feature selection to avoid unintended clustering on tissue type.

On the other hand, cell-level IHC quantification (DeepLIIF) of DLBCL appeared to cluster samples on proliferation and T-cell subsets, with low CD5 staining and high Ki67 (MIB1) in cluster 1. This cluster was evidenced to have slightly but consistently worse survival than cluster 2 in training data and DLBCL-MORPH validation data. This was not observed in HMRN validation data, potentially due to low event number between years 1 and 2. Consistent performance between 2 of the datasets evidenced an

effect though, sample number was too low to determine hazard ratio within acceptable confidence intervals. Clustering of DLBCL patients into 3 clusters identified a cluster (1) enriched for GCB expression, with higher CD10 IHC staining and better survival (5-year OS of 75%). Cluster 3 was found to be enriched for ABC expression, accompanied by higher FOXP1 and MUM1 staining than other clusters and significantly worse outcome (5-year OS of 47%). Observed survival and IHC estimates agreed with published work on expression profiles[19, 21, 172].

H&E for prediction of IHC

Self-supervised pathology image embeddings (UNI[82]) were found to slightly outperform nuclear features in prediction of IHC clusters using H&E images. Graph Neural Network models used[85] to predict cell-level IHC clusters, those developed using DeepLIIF IHC features, outperformed those fitted to tissue level IHC clusters (Macenko stain separation features), though these differences were not significant. Additionally, molecular and survival characteristics of DLBCL patients were largely consistent between training (predicted using IHC data) and validation data (predicted on H&E slides).

These results evidenced feasibility in use of routine H&E for prediction of IHC features in non-Hodgkin lymphoma, potentially adding value where IHC analyses are not deemed necessary or feasible.

5.1 Limitations

Several limitations were identified in the presented method. Firstly, limited feature selection was performed. In both clustering of IHC and prediction of IHC clusters using H&E, use of all features complicated interpretation of groups and likely added unnecessary noise and redundancy.

Selection of interpretable features for clustering IHC would have improved explainability and interpretability of downstream processes. Likewise, exclusion of DeepLIIF features describing total cell number may have led to preferable clustering in FL, which was found to classify samples on non-cellular features such as fibrous tissue content.

Secondly, suboptimal clustering in FL may have been a result of insufficient filtration of acceptable data, as well as the clustering methods used. For the scale of biopsy slides assessed ($>6,000$), manual exclusion of abnormal slides or those without informative material was not feasible. More stringent exclusion of data on HistoQC features such as fat content may have led to improved clustering, on true cellular IHC staining. Additionally, use of clustering methods with inherent anomaly detection (such as DBSCAN) may have led to exclusion of abnormal samples, such as bone marrow biopsies, potentially improving clinical utility of clusters in lymph node biopsy analysis.

Thirdly, only one model architecture and parameter set was used for prediction of IHC clusters using H&E. Comparison of complementary approaches, such as CLAM[192], Chowder[193], DSMIL[192] or CTransPath[194] may have identified an architecture more suited to the presented task. Parameter optimisation may have also led to better performance in prediction of IHC clusters using H&E.

Finally, number of IHC and H&E patches sampled (for DeepLIIF quantification and prediction of IHC using H&E data, respectively) was chosen empirically. These sample sizes were assumed to be representative of slides. Lack of objective evidence for sample size used mean that it is uncertain whether clusters observed were a product of sampling errors, and whether sufficiently high sample number was used. Objective efforts to determine the minimum representative sample size required of a whole slide image may result in greater clustering consistency and optimal H&E classification performance.

5.2 Summary of Findings

- Tissue level quantification of IHC (Macenko proportion tissue) correlated with pathologist estimations of CD10 staining but significantly less well for BCL6 and MUM1. DeepLIIF IHC quantification correlated with pathologist estimations of CD10, BCL6 and MUM1, with only slight (not significant) decreases in performance observed for the latter 2 stains.
- Clustering of DLBCL IHC into 2 clusters identified a group with increased Ki67 staining (MUM1), decreased CD5 staining (T-cell and dendritic cell subsets) and

evidence of consistently worse survival (compared to the rest of patients).

- Clustering of DLBCL IHC into 3 clusters identified groups enriched for expression groups. Survival differences and IHC staining agreed with previous efforts that utilised direct gene expression or pathologist assessment of IHC for assignment of expression groups.
- UNI features were found to slightly outperform nuclear features in prediction of IHC clusters, achieving mean AUC-ROC scores of 0.73 in DLBCL and 0.92 in FL DeepLIIF cluster prediction, respectively.
- Patient characteristics between validation data (IHC clusters predicted using H&E slides) and training data (IHC clusters fitted using IHC slides) were largely consistent, evidencing adequate performance in classification of IHC groups with H&E.

5.3 Future Work

The presented IHC clusters were fitted using a wide range of features. Selection of features on clinical-relevance, such as on association with survival, mutation or expression profiles in future work will likely result in greater cluster utility.

Optimisation of IHC prediction using H&E was limited in the presented work. Future efforts with UNI feature selection, comparison of complementary IHC classification models and parameter optimisation may lead to improved performance in prediction of IHC clusters using routine H&E slides.

Contributions

All work was conducted by Volodymyr Chapman (UoL). Professor David Westhead (UoL) provided feedback on methodology throughout.

Conclusions and Future Work

Summary of Main Findings

In the work presented, analyses were conducted on the two most common classifications of non-Hodgkin Lymphoma, Diffuse Large B-cell Lymphoma (DLBCL) and Follicular Lymphoma (FL) with the aim of identifying clinically-useful subgroups in routine haematoxylin and eosin (H&E) pathology images. Using cellular approaches, efforts aimed to firstly, automate cytological grading in FL; secondly, to identify subgroups in FL and DLBCL on cell populations involved and thirdly, to identify subgroups of disease on presence of lymphoma-related proteins (via immunohistochemistry; IHC).

Cytological grading

Aims

Cytological grading of FL suffers from poor reproducibility, leading to uncertain utility and conflicting recommendations in use[23, 24]. It is a burdensome task, requiring quantification of centroblast cells within 10 affected naive B-cell regions (follicles). We aimed to develop an automated grading method, performing follicle region selection and centroblast quantification. Finally, we aimed to compare predicted grades with clinical features such as overall survival and mutation profiles[46] to evaluate clinical utility of grading.

Findings

Follicular Lymphoma (FL) patients undergoing immunochemotherapy treatment and determined to have high cytological grade (3) by pathologists experienced significantly increased risk of overall survival events, with hazard ratio of 2.62 when compared to similar patients with lower grade disease (95% confidence intervals: 1.29-5.32; $p = 0.008$). This

effect was not observed with classifications made using an automated route operating on H&E slides, in which survival appeared consistently better in the small group of patients classified.

Evaluation identified that the centroblast classification portion of the automated method operated well, with good agreement against pathologist estimations (R2 score of 0.68 in GALLIUM data centroblast counts and Cohen's κ of 0.55 in grading of defined regions at fixed magnification) and performance comparable to published methods (achieving sensitivity of 0.90 and F1 score of 0.77), using over $30\times$ more test data than best-performing published methods (13,558 nuclei vs. 447[62, 129]). Given these results supporting good performance in quantification of centroblasts, it was hypothesised that automated grading may classify patients with considerable non-neoplastic centroblast populations (possibly ignored in pathologist cytological grading), mounting immune responses against lymphoma cells.

Limitations and future work

Automated grading of whole slide images displayed poor agreement with pathologist estimates, achieving Cohen's κ scores close to 0. Despite moderately good performance of follicle region classification in H&E (mean DICE score of 0.70), models were found to perform far less well against IHC ground truth (naïve B-cell marker, CD10), achieving mean DICE score of 0.41. Poor overall performance in cytological grading was likely caused by discrepancies between pathologist and machine region selection, in turn possibly caused by unreliable ground truth.

Future efforts may use IHC (CD10 and BCL2) as more reliable ground truth for distinction of reactive (inflammatory) and neoplastic regions. In this way, it may be possible for automated procedures to grade similarly to pathologists, using neoplastic regions, and independently quantify reactive centroblasts as a novel biomarker.

Cellular phenotypes in non-Hodgkin Lymphoma

Aims

Evidence indicated utility of cell subgroups in FL and DLBCL prognosis[65, 135], with some evidence in H&E slides[66, 67]. This work aimed to characterise FL and DLBCL patterns of cell presence in H&E and assess clinical utility with respect to mutation, expression and overall survival.

Findings

FL and DLBCL patients each clustered into centrocytic and centroblastic subgroups. In FL, centroblastic phenotypes were associated with the COM mutation profile while those presenting with centrocytic disease had higher frequencies of STAT6 and aSHM mutation ($p = 0.03$). No survival differences were observed between clusters.

In DLBCL, patients presenting with centrocytic disease were found to be at almost two times higher risk in the external DLBCL-MORPH validation dataset (HR vs. rest: 1.95; 95% CI: 1.11-3.43; $p = 0.02$), though this effect was not observed in the HMRN dataset. No significant differences were observed in expression profile between clusters.

Limitations and future work

A limited subset of features, found to be informative in classification of centroblasts, were used for clustering of patients. This potentially led to overly-simplified patient subgroups, insufficiently representing variation in cell types other than centroblasts. Methods used to aggregate small scale measurements to whole slide images may have limited information and thus, complexity of features used in patient clustering. Future efforts may demonstrate improved performance in clustering of non-Hodgkin Lymphoma cells and patients upon use of a wider range of nuclear features, and methods of aggregation that maximise information preservation. Inclusion of kurtosis and standard deviation in aggregation, as well as nuclear features describing microenvironment and cells within range of interactions, may lead to more informative patient clusters.

Immunohistochemical subgroups

Aims

Immunohistochemistry (IHC) is routine in DLBCL[24] and may have clinical utility in FL[65, 135] yet, it can present an added burden to pathologist workloads and suffers from poor reproducibility[69–71], limiting utility. We aimed to evaluate existing methods of IHC quantification in FL and DLBCL and use these to subgroup patients. Finally, we trained classifiers of IHC subgroups to operate on H&E to assess utility in routine H&E.

Findings

DeepLIIF[176] estimations of IHC staining were found to correlate moderately with pathologist estimates of CD10 (mean Pearson $r = 0.74$), BCL6 (mean Pearson $r = 0.58$) and MUM1 (mean Pearson $r = 0.63$), evidencing utility in automation. IHC data optimally clustered into two to five groups, depending on metrics used. Clustering into two groups, FL patients were clustered on tissue types present while DLBCL patients clustered on Ki67 and CD5 staining. Clustering of DLBCL patients into three groups stratified patients on expression profiles well, with significant differences in survival and estimates of IHC staining agreeing with published work[19, 21, 172]. This work evidenced good agreement between effects observed using automated quantification of IHC and those observed with pathologist estimates.

Self-supervised pathology image features (UNI[82]) were found to outperform nuclear features in prediction of IHC clusters (generated using DeepLIIF estimates of staining) using H&E images, achieving AUC-ROC between 0.75-0.80. No significant and consistent differences were observed between datasets though, slight increased risk in DLBCL patients with low CD5 and high Ki67 staining was observed in training and external validation data (DLBCL-MORPH).

Limitations and future work

Confounding variables, namely presence of fat and fibrous tissue were evidenced to have involvement in FL clustering over staining of affected lymph regions. A majority of UNI

features used for prediction of IHC clusters in H&E were found have limited utility or high redundancy. Feature selection in clustering and prediction of IHC clusters, as well as stringent selection of regions with lymph involvement for IHC quantification may lead to improved model performance in subgrouping and classifying patients.

Findings in context

With respect to key questions raised in Introduction Key questions:

Is there clinical utility to cytological grading in FL? Comparison of pathologist-assigned low and high cytological grade groups evidenced significantly increased overall survival risk with high grade in patients undergoing immunochemotherapy. These findings supported clinical utility of cytological grading in FL patients undergoing immunochemotherapy. Yet, results of the automated classifier evidenced that not all centroblasts may be predictive of poor outcome. Distinction of neoplastic, lymphoma centroblasts and those involved in immune responses against lymphoma cells may reveal a novel biomarker for FL.

Can cellular subtypes be resolved in routine histopathology slides? With the features used, nuclei did not cluster into a clear optimum number of types. A greater range of texture and colour space features may have better resolved cell types in H&E.

Can molecular subtypes, originating from mutation or gene expression profiles, be resolved in histopathology slides? In both FL and DLBCL, nuclear features resolved patient subgroups to a degree. A wider range of nuclear features, and feature selection focused on clinically-useful classifications may have better resolved patient groups.

What is the clinical utility, with respect to OS, of histopathology clusters in FL and DLBCL? Differences between groups (in OS) were found to be subtle in H&E, but those efforts focused on use of nuclear features, which were found to underperform in

prediction of IHC clusters. Emerging trends in histopathology involve use of less intuitive but, empirically, highly informative feature representations, such as UNI features[82]. These may therefore have far greater value in H&E efforts focusing on OS in FL and DLBCL.

In DLBCL IHC, clustering of patients into three groups stratified patients on OS risk, but this could almost completely be attributed to expression profile. Existing classifiers appeared to stratify OS risk effectively. Focusing computational pathology efforts on classification of existing clusters (mutation and expression) may lead to good performance in stratification of risk.

Real-world merit of this thesis

Key novelties within this thesis were strategies to cluster patients in FL and DLBCL on nuclear features (Chapters 2 & 3) and cellular markers (Chapter 4) using routine H&E histopathology slides. Findings from these analyses have potential value in clinical practice. In this section, considerations for clinical deployment will be discussed briefly in order to highlight challenges and potential solutions for clinical adoption.

Technical

With respect to hardware requirements, all presented techniques were developed and validated on a medium tier consumer workstation. The presented methods are therefore relatively accessible. With further optimisation, for instance through substitution of HoVerNet[103] for the optimised HoVerFast[181] algorithm in nuclear phenotype inference, technical requirements may potentially be reduced further. In clinical settings where slides are routinely digitised, automated analysis of H&E lymphoma samples using the presented techniques would likely not require considerable changes to existing hardware and workflows. On the other hand, the digitisation of slides required for all methods presented is not universally adopted across healthcare providers (at the time of writing) due to cost and extent of changes required in clinical workflows. Technical

considerations for adoption of presented methods therefore largely depend on availability of slide scanning facilities.

Clinical

Clinical utility was evidenced in FL nuclear phenotypes, where associations were observed with mutation patterns and in prediction of DLBCL IHC clusters, observed to associate with high risk expression profiles, using H&E. In each of these cases, the methods presented may have utility in automated referral of slides for confirmatory mutation, gene expression or IHC analyses in the time that they would be queued for initial pathologist review in current practice. This may potentially save days in overall turnaround time, improve healthcare unit performance metrics and lead to faster diagnosis for patients.

Financial

The presented methods aimed to assist pathologist analysis of non-Hodgkin Lymphoma through automation of time-consuming processes (Chapter 2), determination of features otherwise not feasible for a human observer to calculate (Chapter 3) and make predictions otherwise requiring more expensive analyses(Chapter 4). Each of these aims presented potential financial benefits also, with substitution of expensive expert time in algorithmic, repetitive processes with electricity costs; added value in clinical decision-making with additional features, reducing costs through administration of more effective interventions and thirdly, potential substitution of IHC for H&E in classification of high risk DLBCL. Finally, if used for reduction in turnaround times as discussed above, higher rates of compliance with health authority targets may also have a positive financial impact for users.

Regulatory and Ethical

Current classification

Under the most recent MHRA guidelines[195], the authority responsible for approval of medical products in the UK, tools such as those presented in this thesis would be classified as Class II in vitro diagnostic devices (Class II IVD) as they aim to influence diagnostic decisions. For instance, IHC clusters (Chapter 4) would ideally inform clinicians of the likelihood of a patient's death within 2 years of diagnosis, with the standard of care. The equivalent body in the US, the FDA recently approved an algorithm, similarly operating on histopathology images, as a Class II Medical Device[196]. This evidences that a similar classification will likely apply in the US and more broadly. These classifications require devices to undergo rigorous validation through a 'clinical study', performed on a dataset representative of the intended patient (as outlined in the joint MHRA / FDA machine learning principles[197]). The tools presented have been validated on subgroups of the population, including a population dataset limited to a region of the UK (HMRN dataset) and global population limited to narrow clinical criteria (GALLIUM clinical trial dataset). This scope would not likely meet the criteria of a representative dataset, as outlined by both the FDA and MHRA. So, clinical use of the presented tools (except under the in-house exclusion[195] within the Leeds Teaching Hospital NHS Trust) would currently not be advised.

For clinical approval of the presented tools, they would require validation in geographically dispersed centres with sufficient representation of all patient groups outlined in the intended use. For instance, approval of the DLBCL IHC clusters classifier for general use in the UK would likely require participation of clinics throughout the devolved nations as well as representation of patients receiving each of the common interventions (R-CVP, Pola-R-CHP and R-CHOP).

Future approaches

Though not appropriate for clinical decision-making yet, there is scope for more immediate use of presented tools for purely administrative purposes, where potential harm is far

less severe. In a more general context or a far narrower scope of utility (not making clinical decisions), requirements for a representative dataset would likely be more manageable. One such use case may be in arranging cases for pathologist review by cell phenotype. In this example, the model may ensure that samples with high centroblast presence (typically treated with immunochemotherapy) are analysed by a pathologist more quickly, allowing faster intervention while not explicitly influencing clinician decisions.

Additionally, the regulatory landscape with respect to use of artificial intelligence in the clinic is in flux. A recent focus of the authority responsible for determining if and how interventions are used within the NHS (National Institute for Health and Care Excellence; NICE) is the development of efficient processes for validation and approval of AI medical devices[198]. As such, processes for validating and approving the reported tools will likely change in the near future. One such recent change has been the establishment of a National Secure Data Environment, for collation of medical data across dozens of NHS trusts[199]. Validation of the presented tools on this dataset, representative of the population of the UK, would likely meet UK clinical study criteria of Class II medical devices.

Bibliography

1. Of Pathologists, T. R. C. *Meeting pathology demand Histopathology workforce census* 2018. <https://www.rcpath.org/static/952a934d-2ec3-48c9-a8e6e00fcdca700f/Meeting-Pathology-Demand-Histopathology-Workforce-Census-2018.pdf>.
2. Walsh, E. & Orsi, N. M. The current troubled state of the global pathology workforce: a concise review. *Diagnostic Pathology* **19**, 1–10. ISSN: 17461596. <https://diagnosticpathology.biomedcentral.com/articles/10.1186/s13000-024-01590-2> (2024).
3. Association, B. M. *NHS diagnostics data analysis* 2023. <https://www.bma.org.uk/advice-and-support/nhs-delivery-and-workforce/pressures/nhs-diagnostics-data-analysis>.
4. Nica, A., Sutradhar, R., Kupets, R., Covens, A., Vicus, D., Li, Q., Ferguson, S. E. & Gien, L. T. Pre-operative wait times in high-grade non-endometrioid endometrial cancer: Do surgical delays impact patient survival? *Gynecologic Oncology* **164**, 333–340. ISSN: 10956859. [http://www.gynecologiconcology-online.net/article/S0090825821016218/fulltext%20http://www.gynecologiconcology-online.net/article/S0090825821016218/abstract%20https://www.gynecologiconcology-online.net/article/S0090-8258\(21\)01621-8/abstract](http://www.gynecologiconcology-online.net/article/S0090825821016218/fulltext%20http://www.gynecologiconcology-online.net/article/S0090825821016218/abstract%20https://www.gynecologiconcology-online.net/article/S0090-8258(21)01621-8/abstract) (2022).
5. Denault, M.-H., Labbé, C., St-Pierre, C., Fournier, B., Gagné, A., Morillon, C., Joubert, P., Simard, S. & Martel, S. Wait Times and Survival in Lung Cancer Patients across the Province of Quebec, Canada. *Current Oncology* **29**, 3187–3199. ISSN: 1718-7729. <https://www.mdpi.com/1718-7729/29/5/259> (2022).
6. *Report of the Review of NHS Pathology Services in England* 2006. <https://webarchive.nationalarchives.gov.uk/ukgwa/20130123210254/http://www>.

- dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4137606.
7. UK, C. R. *Testing times to come? An evaluation of pathology capacity across the UK 2016* 2016. https://www.cancerresearchuk.org/sites/default/files/testing_times_to_come_nov_16_cruk.pdf.
 8. *NHS Long Term Workforce Plan* 2023. <https://www.england.nhs.uk/long-read/nhs-long-term-workforce-plan-2/>.
 9. Armitage, J. O., Gascoyne, R. D., Lunning, M. A. & Cavalli, F. Non-Hodgkin lymphoma. *The Lancet* **390**, 298–310. ISSN: 1474547X (2017).
 10. Digital, N. *Cancer Incidence and Mortality* 2024. https://nhsd-ndrs.shinyapps.io/incidence_and_mortality/.
 11. Digital, N. *SCDI breakdown by Cancer Alliance, 2014-2015* 2024. https://nhsd-ndrs.shinyapps.io/scdi_ar_geography/.
 12. Lee, S., Fujita, K., Negoro, E., Morishita, T., Yamauchi, H., Oiwa, K., Ueda, T. & Yamauchi, T. The impact of diagnostic wait time on the survival of patients with diffuse large B-cell lymphoma: effect modification by the International Prognostic Index. *British Journal of Haematology* **187**, 195–205. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/bjh.16078>. <https://onlinelibrary.wiley.com/doi/abs/10.1111/bjh.16078> (2019).
 13. Lamb, M., Painter, D., Howell, D., Barrans, S., Cargo, C., de Tute, R., Tooze, R., Burton, C., Patmore, R., Roman, E. & Smith, A. Lymphoid blood cancers, incidence and survival 2005-2023: A report from the UK's Haematological Malignancy Research Network. *Cancer Epidemiology* **88**, 102513. ISSN: 1877-7821. <https://www.sciencedirect.com/science/article/pii/S1877782123001935> (2024).

14. Network, H. M. R. *Survival* <https://hmrn.org/statistics/survival> (2024).
15. Rodríguez-Sevilla, J. J., Fernández-Rodríguez, C., Bento, L., Diez-Feijóo, R., Pinzón, S., Gibert, J., Fernández-Ibarrondo, L., Lafuente, M., Ferrer, A., Sánchez-González, B., Gimeno, E., Sainz, J., Ramos, R., García, J. F., Colomo, L., Bellosillo, B., Gutiérrez, A. & Salar, A. Evaluation of 4 prognostic indices in follicular lymphoma treated in first line with immunochemotherapy. *Blood Advances* **7**, 1606–1614. ISSN: 2473-9529. eprint: https://ashpublications.org/bloodadvances/article-pdf/7/8/1606/2045733/blooda_adv-2022-007949-main.pdf. <https://doi.org/10.1182/bloodadvances.2022007949> (2023).
16. Enemark, M. H., Hemmingsen, J. K., Andersen, M. D., Hybel, T. E., Bjørn, M. E., Josefsson, P. L., Pedersen, L. M., Juul, M. B., Pedersen, R. S., Thorsgaard, M., Sillesen, I. B., Plesner, T. L., Hamilton-Dutoit, S. J., Jensen, P., Madsen, C. & Ludvigsen, M. Progression of disease within 24 months (POD24) in follicular lymphoma in the rituximab era: incidence, clinicopathological risk factors, and outcome in a population-based Danish cohort. *Blood Cancer Journal* 2024 14:1 **14**, 1–4. ISSN: 2044-5385. <https://www.nature.com/articles/s41408-024-01150-3> (2024).
17. Casulo, C., Dixon, J. G., Le-Rademacher, J., Hoster, E., Hochster, H. S., Hidemann, W., Marcus, R., Kimby, E., Herold, M., Sebban, C., Gyan, E., Foon, K., Nielsen, T., Vitolo, U., Salles, G. A., Shi, Q. & Flowers Christopher R., f. t. F. (L. A. o. S. H. G. Validation of POD24 as a robust early clinical end point of poor survival in FL from 5225 patients on 13 clinical trials. *Blood* **139**, 1684–1693. ISSN: 0006-4971. eprint: <https://ashpublications.org/blood/article-pdf/139/11/1684/1881984/bloodbld2020010263.pdf>. <https://doi.org/10.1182/blood.2020010263> (2022).
18. Sehn, L. H. & Salles, G. Diffuse Large B-Cell Lymphoma. *New England Journal of Medicine* **384**, 842–858. eprint: <https://www.nejm.org/doi/pdf/10.1056/>

- NEJMra2027612. <https://www.nejm.org/doi/full/10.1056/NEJMra2027612> (2021).
19. Alizadeh, A. A., Elsen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marü, G. E., Moore, T., Hudson Jr., J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O. & Staudt, L. M. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511. <https://www.nature.com/articles/35000501> (2000).
 20. Sha, C., Barrans, S., Cucco, F., Bentley, M. A., Care, M. A., Cummin, T., Kennedy, H., Thompson, J. S., Uddin, R., Worrillow, L., Chalkley, R., Hoppe, M. V., Ahmed, S., Maishman, T., Caddy, J., Schuh, A., Mamot, C., Burton, C., Tooze, R., Davies, A., Du, M. Q., Johnson, P. W. & Westhead, D. R. Molecular high-grade B-cell lymphoma: Defining a poor-risk group that requires different approaches to therapy. *Journal of Clinical Oncology* **37**, 202–212. ISSN: 15277755. <https://ascopubs.org/doi/pdf/10.1200/JCO.18.01314> (2019).
 21. Wright, G. W., Huang, D. W., Phelan, J. D., Coulibaly, Z. A., Roulland, S., Young, R. M., Wang, J. Q., Schmitz, R., Morin, R. D., Tang, J., Jiang, A., Bagaev, A., Plotnikova, O., Kotlov, N., Johnson, C. A., Wilson, W. H., Scott, D. W. & Staudt, L. M. A Probabilistic Classification Tool for Genetic Subtypes of Diffuse Large B Cell Lymphoma with Therapeutic Implications. *Cancer Cell* **37**, 551–568.e14. ISSN: 18783686. [http://www.cell.com/article/S1535610820301550/fulltext%20http://www.cell.com/article/S1535610820301550/abstract%20https://www.cell.com/cancer-cell/abstract/S1535-6108\(20\)30155-0](http://www.cell.com/article/S1535610820301550/fulltext%20http://www.cell.com/article/S1535610820301550/abstract%20https://www.cell.com/cancer-cell/abstract/S1535-6108(20)30155-0) (2020).

22. Dabrowska-Iwanicka, A. & Nowakowski, G. S. DLBCL: who is high risk and how should treatment be optimized? *Blood* **144**, 2573–2582. ISSN: 0006-4971. <https://www.sciencedirect.com/science/article/pii/S0006497123142686> (2024).
23. Alaggio, R., Amador, C., Anagnostopoulos, I., Attygalle, A. D., de Oliveira Araujo, I. B., Berti, E., Bhagat, G., Borges, A. M., Boyer, D., Calaminici, M., Chadburn, A., Chan, J. K., Cheuk, W., Chng, W. J., Choi, J. K., Chuang, S. S., Coupland, S. E., Czader, M., Dave, S. S., de Jong, D., Du, M. Q., Elenitoba-Johnson, K. S., Ferry, J., Geyer, J., Gratzinger, D., Guitart, J., Gujral, S., Harris, M., Harrison, C. J., Hartmann, S., Hochhaus, A., Jansen, P. M., Karube, K., Kempf, W., Khoury, J., Kimura, H., Klapper, W., Kovach, A. E., Kumar, S., Lazar, A. J., Lazzi, S., Leoncini, L., Leung, N., Leventaki, V., Li, X. Q., Lim, M. S., Liu, W. P., Louissaint, A., Marcogliese, A., Medeiros, L. J., Michal, M., Miranda, R. N., Mitteldorf, C., Montes-Moreno, S., Morice, W., Nardi, V., Naresh, K. N., Natkunam, Y., Ng, S. B., Oschlies, I., Ott, G., Parrens, M., Pulitzer, M., Rajkumar, S. V., Rawstron, A. C., Rech, K., Rosenwald, A., Said, J., Sarkozy, C., Sayed, S., Saygin, C., Schuh, A., Sewell, W., Siebert, R., Sohani, A. R., Tooze, R., Traverse-Glehen, A., Vega, F., Vergier, B., Wechalekar, A. D., Wood, B., Xerri, L. & Xiao, W. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid Neoplasms. *Leukemia* 2022 36:7 **36**, 1720–1748. ISSN: 1476-5551. <https://www.nature.com/articles/s41375-022-01620-2> (2022).
24. Campo, E., Jaffe, E. S., Cook, J. R., Quintanilla-Martinez, L., Swerdlow, S. H., Anderson, K. C., Brousset, P., Cerroni, L., de Leval, L., Dirnhofer, S., Dogan, A., Feldman, A. L., Fend, F., Friedberg, J. W., Gaulard, P., Ghia, P., Horwitz, S. M., King, R. L., Salles, G., San-Miguel, J., Seymour, J. F., Treon, S. P., Vose, J. M., Zucca, E., Advani, R., Ansell, S., Au, W.-Y., Barrionuevo, C., Bergsagel, L., Chan, W. C., Cohen, J. I., d'Amore, F., Davies, A., Falini, B., Ghobrial, I. M., Goodlad, J. R., Gribben, J. G., Hsi, E. D., Kahl, B. S., Kim, W.-S., Kumar, S., LaCasce, A. S., Laurent, C., Lenz, G., Leonard, J. P., Link, M. P., Lopez-Guillermo, A.,

- Mateos, M. V., Macintyre, E., Melnick, A. M., Morschhauser, F., Nakamura, S., Narbaitz, M., Pavlovsky, A., Pileri, S. A., Piris, M., Pro, B., Rajkumar, V., Rosen, S. T., Sander, B., Sehn, L., Shipp, M. A., Smith, S. M., Staudt, L. M., Thieblemont, C., Tousseyn, T., Wilson, W. H., Yoshino, T., Zinzani, P.-L., Dreyling, M., Scott, D. W., Winter, J. N. & Zelenetz, A. D. The International Consensus Classification of Mature Lymphoid Neoplasms: a report from the Clinical Advisory Committee. *Blood* **140**, 1229–1253. ISSN: 0006-4971. eprint: <https://ashpublications.org/blood/article-pdf/140/11/1229/1921035/bloodbld2022015851c.pdf>. <https://doi.org/10.1182/blood.2022015851> (2022).
25. Hans, C. P., Weisenburger, D. D., Greiner, T. C., Gascoyne, R. D., Delabie, J., Ott, G., Müller-Hermelink, H. K., Campo, E., Braziel, R. M., Jaffe, E. S., Pan, Z., Farinha, P., Smith, L. M., Falini, B., Banham, A. H., Rosenwald, A., Staudt, L. M., Connors, J. M., Armitage, J. O. & Chan, W. C. Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray. *Blood* **103**, 275–282. ISSN: 0006-4971. eprint: <https://ashpublications.org/blood/article-pdf/103/1/275/1693896/zh800104000275.pdf>. <https://doi.org/10.1182/blood-2003-05-1545> (2004).
26. Choi, W. W., Weisenburger, D. D., Greiner, T. C., Piris, M. A., Banham, A. H., Delabie, J., Braziel, R. M., Geng, H., Iqbal, J., Lenz, G., Vose, J. M., Hans, C. P., Fu, K., Smith, L. M., Li, M., Liu, Z., Gascoyne, R. D., Rosenwald, A., Ott, G., Rimsza, L. M., Campo, E., Jaffe, E. S., Jaye, D. L., Staudt, L. M. & Chan, W. C. A New Immunostain Algorithm Classifies Diffuse Large B-Cell Lymphoma into Molecular Subtypes with High Accuracy. *Clinical Cancer Research* **15**, 5494–5502. ISSN: 1078-0432. eprint: <https://aacrjournals.org/clincancerres/article-pdf/15/17/5494/1985597/5494.pdf>. <https://doi.org/10.1158/1078-0432.CCR-09-0113> (2009).
27. Tavakkoli, M. & Barta, S. K. 2024 Update: Advances in the risk stratification and management of large B-cell lymphoma. *American Journal of Hematology* **98**,

- 1791–1805. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajh.27075>. <https://onlinelibrary.wiley.com/doi/abs/10.1002/ajh.27075> (2023).
28. Moser, E. C., Noordijk, E. M., van Leeuwen, F. E., le Cessie, S., Baars, J. W., Thomas, J., Carde, P., Meerwaldt, J. H., van Glabbeke, M. & Kluin-Nelemans, H. C. Long-term risk of cardiovascular disease after treatment for aggressive non-Hodgkin lymphoma. *Blood* **107**, 2912–2919. ISSN: 0006-4971. eprint: <https://ashpublications.org/blood/article-pdf/107/7/2912/1338532/zh800706002912.pdf>. <https://doi.org/10.1182/blood-2005-08-3392> (2006).
29. Carter, R. H. B cells in health and disease. *Mayo Clinic Proceedings* **81**, 377–384. ISSN: 00256196. [http://www.mayoclinicproceedings.org/article/S0025619611614663/fulltext%20http://www.mayoclinicproceedings.org/article/S0025619611614663/abstract%20https://www.mayoclinicproceedings.org/article/S0025-6196\(11\)61466-3/abstract](http://www.mayoclinicproceedings.org/article/S0025619611614663/fulltext%20http://www.mayoclinicproceedings.org/article/S0025619611614663/abstract%20https://www.mayoclinicproceedings.org/article/S0025-6196(11)61466-3/abstract) (2006).
30. Silva, N. S. D. & Klein, U. Dynamics of B cells in germinal centres. *Nature Reviews Immunology* 2015 15:3 **15**, 137–148. ISSN: 1474-1741. <https://www.nature.com/articles/nri3804> (2015).
31. Roghanian, A. & Newman, R. *B Cells* 2021. <https://www.immunology.org/public-information/bitesized-immunology/cells/b-cells>.
32. Smith, T. & Cunningham-Rundles, C. Primary B-cell immunodeficiencies. *Human Immunology* **80**. Forefronts of B cell Biology in Transplantation, 351–362. ISSN: 0198-8859. <https://www.sciencedirect.com/science/article/pii/S0198885918302544> (2019).
33. Elsner, R. A. & Shlomchik, M. J. Germinal Center and Extrafollicular B Cell Responses in Vaccination, Immunity, and Autoimmunity. *Immunity* **53**, 1136–1150. ISSN: 1074-7613. <http://www.cell.com/article/S1074761320304933/>

- fulltext%20http://www.cell.com/article/S1074761320304933/abstract%20https://www.cell.com/immunity/abstract/S1074-7613(20)30493-3 (2020).
34. Sepich-Poore, G. D., Zitvogel, L., Straussman, R., Hasty, J., Wargo, J. A. & Knight, R. The microbiome and human cancer. *Science* **371**, eabc4552. eprint: <https://www.science.org/doi/pdf/10.1126/science.abc4552>. <https://www.science.org/doi/abs/10.1126/science.abc4552> (2021).
 35. Botía-Sánchez, M., Alarcón-Riquelme, M. E. & Galicia, G. B Cells and Microbiota in Autoimmunity. *International Journal of Molecular Sciences* **22**. ISSN: 1422-0067. <https://www.mdpi.com/1422-0067/22/9/4846> (2021).
 36. Kunath, B. J., Rudder, C. D., Laczny, C. C., Letellier, E. & Wilmes, P. The oral–gut microbiome axis in health and disease. *Nature Reviews Microbiology* 2024 *22:12* **22**, 791–805. ISSN: 1740-1534. <https://www.nature.com/articles/s41579-024-01075-5> (2024).
 37. LeBien, T. W. & Tedder, T. F. B lymphocytes: how they develop and function. *Blood* **112**, 1570–1580. ISSN: 0006-4971. eprint: <https://ashpublications.org/blood/article-pdf/112/5/1570/1485355/zh801708001570.pdf>. <https://doi.org/10.1182/blood-2008-02-078071> (2008).
 38. Chi, X., Li, Y. & Qiu, X. V(D)J recombination, somatic hypermutation and class switch recombination of immunoglobulins: mechanism and regulation. *Immunology* **160**, 233–247. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/imm.13176>. <https://onlinelibrary.wiley.com/doi/abs/10.1111/imm.13176> (2020).
 39. Zhang, Y., Zhang, X., Dai, H. Q., Hu, H. & Alt, F. W. The role of chromatin loop extrusion in antibody diversification. *Nature Reviews Immunology* 2022 *22:9* **22**,

- 550–566. ISSN: 1474-1741. <https://www.nature.com/articles/s41577-022-00679-3> (2022).
40. Nutt, S. L., Hodgkin, P. D., Tarlinton, D. M. & Corcoran, L. M. The generation of antibody-secreting plasma cells. *Nature Reviews Immunology* 2015 15:3 **15**, 160–171. ISSN: 1474-1741. <https://www.nature.com/articles/nri3795> (2015).
 41. Horton, M. B., Cheon, H. C., Duffy, K. R., Brown, D., Naik, S. H., Alvarado, C., Groom, J. R., Heinzel, S. & Hodgkin, P. D. Lineage tracing reveals B cell antibody class switching is stochastic, cell-autonomous, and tuneable. *Immunity* **55**, 1843–1855.e6. ISSN: 10974180. [http://www.cell.com/article/S1074761322003971/fulltext%20http://www.cell.com/article/S1074761322003971/abstract%20https://www.cell.com/immunity/abstract/S1074-7613\(22\)00397-1](http://www.cell.com/article/S1074761322003971/fulltext%20http://www.cell.com/article/S1074761322003971/abstract%20https://www.cell.com/immunity/abstract/S1074-7613(22)00397-1) (2022).
 42. Basso, K. & Dalla-Favera, R. Germinal centres and B cell lymphomagenesis. *Nature Reviews Immunology* 2015 15:3 **15**, 172–184. ISSN: 1474-1741. <https://www.nature.com/articles/nri3814> (2015).
 43. Dhanasekaran, R., Deutzmann, A., Mahauad-Fernandez, W. D., Hansen, A. S., Gouw, A. M. & Felsher, D. W. The MYC oncogene — the grand orchestrator of cancer growth and immune evasion. *Nature Reviews Clinical Oncology* 2021 19:1 **19**, 23–36. ISSN: 1759-4782. <https://www.nature.com/articles/s41571-021-00549-2> (2021).
 44. Mato, S., Castrejón-de-Anta, N., Colmenero, A., Carità, L., Salmerón-Villalobos, J., Ramis-Zaldivar, J. E., Nadeu, F., Garcia, N., Wang, L., Verdú-Amorós, J., Andrés, M., Conde, N., Celis, V., Ortega, M. J., Galera, A., Astigarraga, I., Perez-Alonso, V., Quiroga, E., Jiang, A., Scott, D. W., Campo, E., Balagué, O. & Salaverria, I. MYC-rearranged mature B-cell lymphomas in children and young adults are molecularly Burkitt Lymphoma. *Blood Cancer Journal* 2024 14:1 **14**, 1–10. ISSN: 2044-5385. <https://www.nature.com/articles/s41408-024-01153-0> (2024).

45. Rabkin, C. S., Hirt, C., Janz, S. & Dölken, G. t(14;18) Translocations and Risk of Follicular Lymphoma. *JNCI Monographs* **2008**, 48–51. ISSN: 1052-6773. eprint: <https://academic.oup.com/jncimono/article-pdf/2008/39/48/2797253/lgn002.pdf>. <https://doi.org/10.1093/jncimonographs/lgn002> (2008).
46. Crouch, S., Painter, D., Barrans, S. L., Roman, E., Beer, P. A., Cooke, S. L., Glover, P., Van Hoppe, S. J., Webster, N., Lacy, S. E., Ruiz, C., Campbell, P. J., Hodson, D. J., Patmore, R., Burton, C., Smith, A. & Tooze, R. M. Molecular subclusters of follicular lymphoma: a report from the United Kingdom's Haematological Malignancy Research Network. *Blood Advances* **6**, 5716–5731. ISSN: 2473-9529. eprint: https://ashpublications.org/bloodadvances/article-pdf/6/21/5716/2054222/blooda_adv-2021-005284-main.pdf. <https://doi.org/10.1182/bloodadvances.2021005284> (2022).
47. Lacy, S. E., Barrans, S. L., Beer, P. A., Painter, D., Smith, A. G., Roman, E., Cooke, S. L., Ruiz, C., Glover, P., Hoppe, S. J. V., Webster, N., Campbell, P. J., Tooze, R. M., Patmore, R., Burton, C., Crouch, S. & Hodson, D. J. Targeted sequencing in DLBCL, molecular subtypes, and outcomes: a Haematological Malignancy Research Network report. *Blood* **135**, 1759–1771. ISSN: 15280020. <https://ashpublications.org/blood/article/135/20/1759/452715/Targeted-sequencing-in-DLBCL-molecular-subtypes> (2020).
48. Dreval, K., Hilton, L. K., Cruz, M., Shaalan, H., Ben-Neriah, S., Boyle, M., Collinge, B., Coyle, K. M., Duns, G., Farinha, P., Grande, B. M., Meissner, B., Pararajalingam, P., Rushton, C. K., Slack, G. W., Wong, J., Mungall, A. J., Marra, M. A., Connors, J. M., Steidl, C., Scott, D. W. & Morin, R. D. Genetic subdivisions of follicular lymphoma defined by distinct coding and noncoding mutation patterns. *Blood* **142**, 561–573. ISSN: 0006-4971. <https://www.sciencedirect.com/science/article/pii/S0006497123009825> (2023).

49. Ruppert, A. S., Dixon, J. G., Salles, G., Wall, A., Cunningham, D., Poeschel, V., Haioun, C., Tilly, H., Ghesquieres, H., Ziepert, M., Flament, J., Flowers, C., Shi, Q. & Schmitz, N. International prognostic indices in diffuse large B-cell lymphoma: a comparison of IPI, R-IPI, and NCCN-IPI. *Blood* **135**, 2041–2048. ISSN: 0006-4971. eprint: <https://ashpublications.org/blood/article-pdf/135/23/2041/1743304/bloodbld2019002729.pdf>. <https://doi.org/10.1182/blood.2019002729> (2020).
50. Bankhead, P., Loughrey, M. B., Fernández, J. A., Dombrowski, Y., McArt, D. G., Dunne, P. D., McQuaid, S., Gray, R. T., Murray, L. J., Coleman, H. G., James, J. A., Salto-Tellez, M. & Hamilton, P. W. QuPath: Open source software for digital pathology image analysis. *Scientific Reports* **2017 7:1** **7**, 1–7. ISSN: 2045-2322. <https://www.nature.com/articles/s41598-017-17204-5> (2017).
51. Russler-Germain, D. A., Krysiak, K., Ramirez, C., Mosior, M., Watkins, M. P., Gomez, F., Skidmore, Z. L., Trani, L., Gao, F., Geyer, S., Cashen, A. F., Mehta-Shah, N., Kahl, B. S., Bartlett, N. L., Alderuccio, J. P., Lossos, I. S., Ondrejka, S. L., Hsi, E. D., Martin, P., Leonard, J. P., Griffith, M., Griffith, O. L. & Fehniger, T. A. Mutations associated with progression in follicular lymphoma predict inferior outcomes at diagnosis: Alliance A151303. *Blood Advances* **7**, 5524–5539. ISSN: 2473-9529. eprint: https://ashpublications.org/bloodadvances/article-pdf/7/18/5524/2079819/blooda_adv-2023-010779-main.pdf. <https://doi.org/10.1182/bloodadvances.2023010779> (2023).
52. Shelton, V., Detroja, R., Liu, T., Isaev, K., Silva, A., Passerini, V., Bakhtiari, M., Calvente, L., Hong, M., He, M. Y., Modi, S., Hershenfeld, S. A., Ludvigsen, M., Madsen, C., Hamilton-Dutoit, S., d'Amore, F. A., Brodtkorb, M., Johnson, N. A., Baetz, T., LeBrun, D., Tobin, J. W., Gandhi, M. K., Mungall, A. J., Xu, W., Ben-Neriah, S., Steidl, C., Delabie, J., Tremblay-LeMay, R., Jegede, O., Weigert, O., Kahl, B., Evens, A. M. & Kridel, R. Identification of genetic subtypes in follicular

- lymphoma. *Blood Cancer Journal* 2024 14:1 **14**, 1–13. ISSN: 2044-5385. <https://www.nature.com/articles/s41408-024-01111-w> (2024).
53. Shen, R., Fu, D., Dong, L., Zhang, M. C., Shi, Q., Shi, Z. Y., Cheng, S., Wang, L., Xu, P. P. & Zhao, W. L. Simplified algorithm for genetic subtyping in diffuse large B-cell lymphoma. *Signal Transduction and Targeted Therapy* 2023 8:1 **8**, 1–11. ISSN: 2059-3635. <https://www.nature.com/articles/s41392-023-01358-y> (2023).
54. Riedell, P. A. & Smith, S. M. Double hit and double expressors in lymphoma: Definition and treatment. *Cancer* **124**, 4622–4632. ISSN: 10970142. <https://acsjournals.onlinelibrary.wiley.com/doi/full/10.1002/cncr.31646> (2018).
55. Ren, W., Wan, H., Own, S. A., Berglund, M., Wang, X., Yang, M., Li, X., Liu, D., Ye, X., Sonnev, K., Enblad, G., Amini, R. M., Sander, B., Wu, K., Zhang, H., Wahlin, B. E., Smedby, K. E. & Pan-Hammarström, Q. Genetic and transcriptomic analyses of diffuse large B-cell lymphoma patients with poor outcomes within two years of diagnosis. *Leukemia* 2023 38:3 **38**, 610–620. ISSN: 1476-5551. <https://www.nature.com/articles/s41375-023-02120-7> (2023).
56. Sha, C., Barrans, S., Care, M. A., Cunningham, D., Tooze, R. M., Jack, A. & Westhead, D. R. Transferring genomics to the clinic: Distinguishing Burkitt and diffuse large B cell lymphomas. *Genome Medicine* **7**, 1–13. ISSN: 1756994X. <https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-015-0187-6> (2015).
57. Kotlov, N., Bagaev, A., Revuelta, M. V., Phillip, J. M., Cacciapuoti, M. T., Antysheva, Z., Svekolkina, V., Tikhonova, E., Mihecheva, N., Kuzkina, N., Nos, G., Tabbo, F., Frenkel, F., Ghione, P., Tsiper, M., Almog, N., Fowler, N., Melnick, A. M., Leonard, J. P., Inghirami, G. & Cerchietti, L. Clinical and biological subtypes of B-cell lymphoma revealed by microenvironmental signatures. *Cancer*

- Discovery*. ISSN: 2159-8274. <https://cancerdiscovery.aacrjournals.org/content/early/2021/02/02/2159-8290.CD-20-0839> (2021).
58. Belkacem-Boussaid, K., Pennell, M., Lozanski, G., Shana'Ah, A. & Gurcan, M. Computer-aided classification of centroblast cells in follicular lymphoma. *Analytical and quantitative cytology and histology / the International Academy of Cytology [and] American Society of Cytology* **32**, 254. ISSN: 08846812. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3078581/> (2010).
 59. Belkacem-Boussaid, K., Pennell, M., Ah, A. S., Gerwitz, A., Zhao, W., Racke, F., Hsi, E., Lozanski, G. & Gurcan, M. Inter-Reader Variability In Identifying Centroblast Cells From Digital Follicular Lymphoma Cases. *Blood* **116**, 5095. ISSN: 0006-4971. <https://ashpublications.org/blood/article/116/21/5095/66471/Inter-Reader-Variability-In-Identifying> (2010).
 60. Sertel, O., Lozanski, G., Shana'ah, A. & Gurcan, M. N. Computer-Aided Detection of Centroblasts for Follicular Lymphoma Grading Using Adaptive Likelihood-Based Cell Segmentation. *IEEE Transactions on Biomedical Engineering* **57**, 2613–2616. <https://ieeexplore.ieee.org/document/5497103> (2010).
 61. Lozanski, G., Pennell, M., Shana'ah, A., Zhao, W., Gewirtz, A., Racke, F., Hsi, E., Simpson, S., Mosse, C., Alam, S., Swierczynski, S., Hasserjian, R. P. & Gurcan, M. N. Inter-reader variability in follicular lymphoma grading: Conventional and digital reading. *Journal of Pathology Informatics* **4**, 30. ISSN: 21533539. [/pmc/articles/PMC3869955/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3869955/) [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3869955/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3869955/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3869955/) (2013).
 62. Kornaropoulos, E. N., Niazi, M. K. K., Lozanski, G. & Gurcan, M. N. Histopathological image analysis for centroblasts classification through dimensionality reduction approaches. *Cytometry Part A* **85**, 242–255. ISSN: 1552-4930. <https://onlinelibrary.wiley.com/doi/full/10.1002/cyto.a.22432> <https://onlinelibrary.wiley.com/doi/full/10.1002/cyto.a.22432> (2013).

- [//onlinelibrary.wiley.com/doi/abs/10.1002/cyto.a.22432](https://onlinelibrary.wiley.com/doi/abs/10.1002/cyto.a.22432) <https://onlinelibrary.wiley.com/doi/10.1002/cyto.a.22432> (2014).
63. Fauzi, M. F. A., Pennell, M., Sahiner, B., Chen, W., Shana'Ah, A., Hemminger, J., Gru, A., Kurt, H., Losos, M., Joehlin-Price, A., Kavran, C., Smith, S. M., Nowacki, N., Mansor, S., Lozanski, G. & Gurcan, M. N. Classification of follicular lymphoma: The effect of computer aid on pathologists grading Clinical decision-making, knowledge support systems, and theory. *BMC Medical Informatics and Decision Making* **15**, 1–10. ISSN: 14726947. <https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-015-0235-6> (2015).
 64. Xian, R. R., Xie, Y., Haley, L. M., Yonescu, R., Pallavajjala, A., Pittaluga, S., Jaffe, E. S., Duffield, A. S., McCall, C. M., Gheith, S. M. & Gocke, C. D. CREBBP and STAT6 co-mutation and 16p13 and 1p36 loss define the t(14;18)-negative diffuse variant of follicular lymphoma. *Blood Cancer Journal* **2020 10:6** **10**, 1–12. ISSN: 2044-5385. <https://www.nature.com/articles/s41408-020-0335-0> (2020).
 65. Tsakiroglou, A.-M., Astley, S., Dave, M., Fergie, M., Harkness, E., Rosenberg, A., Sperrin, M., West, C., Byers, R., Linton, K., RichardByers, R. B. & KimLinton, K. L. Immune infiltrate diversity confers a good prognosis in follicular lymphoma. *Cancer Immunology, Immunotherapy* **70**, 3573–3585. ISSN: 14320851. <https://doi.org/10.1007/s00262-021-02945-0> (2021).
 66. Vrabac, D., Smit, A., Rojansky, R., Natkunam, Y., Advani, R. H., Ng, A. Y., Fernandez-Pol, S. & Rajpurkar, P. DLBCL-Morph: Morphological features computed using deep learning for an annotated digital DLBCL image set. *Scientific Data* **2021 8:1** **8**, 1–8. ISSN: 2052-4463. <https://www.nature.com/articles/s41597-021-00915-w> (2021).
 67. Ott, G., Ziepert, M., Klapper, W., Horn, H., Szczepanowski, M., Bernd, H.-W., Thorns, C., Feller, A. C., Lenze, D., Hummel, M., Stein, H., Müller-Hermelink, H.-K., Frank, M., Hansmann, M.-L., Barth, T. F. E., Möller, P., Cogliatti, S., Pfre-

- undschuh, M., Schmitz, N., Trümper, L., Loeffler, M. & Rosenwald, A. Immunoblastic morphology but not the immunohistochemical GCB/nonGCB classifier predicts outcome in diffuse large B-cell lymphoma in the RICOVER-60 trial of the DSHNHL. *Blood* **116**, 4916–4925. ISSN: 0006-4971. eprint: <https://ashpublications.org/blood/article-pdf/116/23/4916/1333209/zh804910004916.pdf>. <https://doi.org/10.1182/blood-2010-03-276766> (2010).
68. Onaindia, A., Santiago-Quispe, N., Iglesias-Martinez, E. & Romero-Abrio, C. Molecular Update and Evolving Classification of Large B-Cell Lymphoma. *Cancers* **13**. ISSN: 2072-6694. <https://www.mdpi.com/2072-6694/13/13/3352> (2021).
69. Rimm, D. L., Han, G., Taube, J. M., Yi, E. S., Bridge, J. A., Flieder, D. B., Homer, R., West, W. W., Wu, H., Roden, A. C., Fujimoto, J., Yu, H., Anders, R., Kowalewski, A., Rivard, C., Rehman, J., Batenchuk, C., Burns, V., Hirsch, F. R. & Wistuba, I. I. A Prospective, Multi-institutional, Pathologist-Based Assessment of 4 Immunohistochemistry Assays for PD-L1 Expression in Non-Small Cell Lung Cancer. *JAMA Oncology* **3**, 1051–1058. ISSN: 2374-2437. eprint: https://jamanetwork.com/journals/jamaoncology/articlepdf/2608280/jamaoncology_rimm_2017_oi_170002.pdf. <https://doi.org/10.1001/jamaoncol.2017.0013> (2017).
70. Koens, L., van de Ven, P. M., Hijmering, N. J., Kersten, M. J., Diepstra, A., Chamuleau, M. & de Jong, D. Interobserver variation in CD30 immunohistochemistry interpretation; consequences for patient selection for targeted treatment. *Histopathology* **73**, 473–482. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/his.13647>. <https://onlinelibrary.wiley.com/doi/abs/10.1111/his.13647> (2018).
71. Pang, J. M. B., Castles, B., Byrne, D. J., Button, P., Hendry, S., Lakhani, S. R., Sivasubramaniam, V., Cooper, W. A., Armes, J., Millar, E. K., Raymond, W., Roberts-Thomson, S., Kumar, B., Burr, M., Selinger, C., Harvey, K., Chan, C.,

- Beith, J., Clouston, D., O'Toole, S. A. & Fox, S. B. SP142 PD-L1 Scoring Shows High Interobserver and Intraobserver Agreement in Triple-negative Breast Carcinoma but Overall Low Percentage Agreement with Other PD-L1 Clones SP263 and 22C3. *American Journal of Surgical Pathology* **45**, 1108–1117. ISSN: 15320979. https://journals.lww.com/ajsp/fulltext/2021/08000/sp142_pd_l1_scoring_shows_high_interobserver_and.10.aspx (2021).
72. Taqi, S. A., Sami, S. A., Sami, L. B. & Zaki, S. A. A review of artifacts in histopathology. *Journal of Oral and Maxillofacial Pathology* **22**, 279. ISSN: 1998393X. https://journals.lww.com/jpat/fulltext/2018/22020/a_review_of_artifacts_in_histopathology.26.aspx (2018).
73. Williams, B. J., Bottoms, D. & Treanor, D. Future-proofing pathology: the case for clinical adoption of digital pathology. *Journal of Clinical Pathology* **70**, 1010–1018. ISSN: 0021-9746. eprint: <https://jcp.bmj.com/content/70/12/1010.full.pdf>. <https://jcp.bmj.com/content/70/12/1010> (2017).
74. Evans, A. J., Vajpeyi, R., Henry, M. & Chetty, R. Establishment of a remote diagnostic histopathology service using whole slide imaging (digital pathology). *Journal of Clinical Pathology* **74**, 421–424. ISSN: 0021-9746. eprint: <https://jcp.bmj.com/content/74/7/421.full.pdf>. <https://jcp.bmj.com/content/74/7/421> (2021).
75. Sarker, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science* **2**, 1–21. ISSN: 26618907. <https://link.springer.com/article/10.1007/s42979-021-00592-x> (2021).
76. Yu, B., Zhang, L., Ye, X., Wang, Y., Ying, H., Zhu, W., Xiang, W. & Wu, X. State-of-the-art review on various applications of machine learning techniques in materials science and engineering. *Chemical Engineering Science*, 121147. ISSN: 0009-2509. <https://www.sciencedirect.com/science/article/pii/S0009250924014477> (2024).

77. Khoei, T. T., Slimane, H. O. & Kaabouch, N. Deep learning: systematic review, models, challenges, and research directions. *Neural Computing and Applications* **35**, 23103–23124. ISSN: 14333058. <https://link.springer.com/article/10.1007/s00521-023-08957-4> (2023).
78. McGenity, C., Clarke, E. L., Jennings, C., Matthews, G., Cartlidge, C., Freduah-Agyemang, H., Stocken, D. D. & Treanor, D. Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy. *npj Digital Medicine* 2024 7:1 **7**, 1–19. ISSN: 2398-6352. <https://www.nature.com/articles/s41746-024-01106-8> (2024).
79. Ronneberger, O., Fischer, P. & Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation* in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (eds Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F.) (Springer International Publishing, Cham, 2015), 234–241. ISBN: 978-3-319-24574-4. https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28.
80. Sheng, B., Zhou, M., Hu, M., Li, Q., Sun, L. & Wen, Y. A blood cell dataset for lymphoma classification using faster R-CNN. *Biotechnology and Biotechnological Equipment* **34**, 413–420. ISSN: 13102818. <https://doi.org/10.1080/13102818.2020.1765871> (2020).
81. Alom, Z., Asari, V. K., Parwani, A. & Taha, T. M. Microscopic nuclei classification, segmentation, and detection with improved deep convolutional neural networks (DCNN). *Diagnostic Pathology* **17**, 1–17. ISSN: 17461596. <https://diagnosticpathology.biomedcentral.com/articles/10.1186/s13000-022-01189-5> (2022).
82. Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A. H., Shaban, M., *et al.* Towards a General-Purpose Founda-

- tion Model for Computational Pathology. *Nature Medicine*. <https://www.nature.com/articles/s41591-024-02857-3> (2024).
83. Yao, W., Bai, J., Liao, W., Chen, Y., Liu, M. & Xie, Y. From CNN to Transformer: A Review of Medical Image Segmentation Models. *Journal of Imaging Informatics in Medicine* **37**, 1529–1547. <https://link.springer.com/article/10.1007/s10278-024-00981-7> (2024).
84. Khemani, B., Patil, S., Kotecha, K. & Tanwar, S. A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. *Journal of Big Data* **11**, 1–43. ISSN: 21961115. <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00876-4> (2024).
85. Lu, W., Toss, M., Dawood, M., Rakha, E., Rajpoot, N. & Minhas, F. SlideGraph+: Whole slide image level graphs to predict HER2 status in breast cancer. *Medical Image Analysis* **80**, 102486. ISSN: 1361-8415. <https://www.sciencedirect.com/science/article/pii/S1361841522001335> (2022).
86. Meng, X. & Zou, T. Clinical applications of graph neural networks in computational histopathology: A review. *Computers in Biology and Medicine* **164**, 107201. ISSN: 0010-4825. <https://www.sciencedirect.com/science/article/pii/S0010482523006662> (2023).
87. Institute, N. C. *Follicular Lymphoma — Cancer Stat Facts* 2020. <https://seer.cancer.gov/statfacts/html/follicular.html>.
88. McNamara, C., Montoto, S., Eyre, T. A., Ardeschna, K., Burton, C., Illidge, T., Linton, K., Rule, S., Townsend, W., Wong, W. L. & McKay, P. The investigation and management of follicular lymphoma. *British Journal of Haematology* **191**, 363–381. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/bjh.16872>. <https://onlinelibrary.wiley.com/doi/abs/10.1111/bjh.16872> (2020).

89. Jacobsen, E. Follicular lymphoma: 2023 update on diagnosis and management. *American Journal of Hematology* **97**, 1638–1651. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajh.26737>. <https://onlinelibrary.wiley.com/doi/abs/10.1002/ajh.26737> (2022).
90. Horn, H., Schmelter, C., Leich, E., Salaverria, I., Katzenberger, T., Ott, M. M., Kalla, J., Romero, M., Siebert, R., Rosenwald, A. & Ott, G. Follicular lymphoma grade 3B is a distinct neoplasm according to cytogenetic and immunohistochemical profiles. *Haematologica* **96**, 1327–1334. ISSN: 1592-8721. <https://haematologica.org/article/view/6070> (2011).
91. Wahlin, B. E., Yri, O. E., Kimby, E., Holte, H., Delabie, J., Smeland, E. B., Sundström, C., Christensson, B. & Sander, B. Clinical significance of the WHO grades of follicular lymphoma in a population-based cohort of 505 patients with long follow-up times. *British Journal of Haematology* **156**, 225–233. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2141.2011.08942.x>. <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2141.2011.08942.x> (2012).
92. Rimsza, L. M., Li, H., Braziel, R. M., Spier, C. M., Persky, D. O., Dunlap, J., Leblanc, M., Bartlett, N., Leonard, J. P., Smith, S. M., Press, O. W. & Friedberg, J. W. Impact of histological grading on survival in the SWOG S0016 follicular lymphoma cohort. *Haematologica* **103**, e151–e153. ISSN: 1592-8721. <https://haematologica.org/article/view/8431> (2018).
93. Dong, N., Saeed, H., Isenalumhe, L., Tokumori, F. C., Shah, B., Gaballa, S., Bello, C., Sokol, L., Ibarz, J. P. & Chavez, J. C. Initial Treatment Vs Watch and Wait in Advanced-Stage Follicular Lymphoma in the Rituximab Era - an Analysis of the National Cancer Database (NCDB). *Blood* **136**, 44–44. ISSN: 0006-4971. <https://ashpublications.org/blood/article/136/Supplement%201/44/472538/Initial-Treatment-Vs-Watch-and-Wait-in-Advanced> (2020).

94. Kim, D., Pantanowitz, L., Schüttler, P., Yarlagadda, D. V. K., Ardon, O., Reuter, V. E., Hameed, M., Klimstra, D. S. & Hanna, M. G. (Re) Defining the High-Power Field for Digital Pathology. *Journal of Pathology Informatics* **11**, 33. ISSN: 2153-3539. <https://www.sciencedirect.com/science/article/pii/S2153353922002620?via%3Dihub> (2020).
95. Samsi, S., Lozanski, G., Shanarah, A., Krishanmurthy, A. K. & Gurcan, M. N. Detection of Follicles From IHC-Stained Slides of Follicular Lymphoma Using Iterative Watershed. *IEEE Transactions on Biomedical Engineering* **57**, 2609–2612. <https://ieeexplore.ieee.org/document/5510114> (2010).
96. Oger, M., Belhomme, P. & Gurcan, M. N. A general framework for the segmentation of follicular lymphoma virtual slides. *Computerized Medical Imaging and Graphics* **36**, 442–451. ISSN: 0895-6111. <https://www.sciencedirect.com/science/article/pii/S0895611112001000> (2012).
97. Trust, M. U. N. F. *Immuno-histochemistry and Research* 2017. <https://mft.nhs.uk/the-trust/other-departments/laboratory-medicine/histopathology/immunohistochemistry-and-research>.
98. Koch, K., Hoster, E., Ziepert, M., Unterhalt, M., Ott, G., Rosenwald, A., Hansmann, M. L., Bernd, W., Stein, H., Pöschel, V., Dreyling, M., Trümper, L., Löffler, M., Schmitz, N., Hiddemann, W., Pfreundschuh, M. & Klapper, W. Clinical, pathological and genetic features of follicular lymphoma grade 3A: a joint analysis of the German low-grade and high-grade lymphoma study groups GLSG and DSHNHL. *Annals of Oncology* **27**, 1323–1329. ISSN: 0923-7534. [https://www.annalsofoncology.org/article/S0923-7534\(19\)35701-1/fulltext](https://www.annalsofoncology.org/article/S0923-7534(19)35701-1/fulltext) (2016).
99. Koch, K., Richter, J., Hanel, C., Huttmann, A., Duhrsen, U. & Klapper, W. Follicular lymphoma grade 3B and diffuse large B-cell lymphoma present a histopathological and molecular continuum lacking features of progression/ transformation.

- Haematologica* **107**, 2144–2153. <https://haematologica.org/article/view/haematol.2021.279351> (2022).
100. Smith, A., Howell, D., Crouch, S., Painter, D., Blase, J., Wang, H. I., Hewison, A., Bagguley, T., Appleton, S., Kinsey, S., Burton, C., Patmore, R. & Roman, E. Cohort Profile: The Haematological Malignancy Research Network (HMRN): a UK population-based patient cohort. *International Journal of Epidemiology* **47**, 700–700g. ISSN: 0300-5771. <https://dx.doi.org/10.1093/ije/dyy044> (2018).
101. Michail, E., Kornaropoulos, E. N., Dimitropoulos, K., Grammalidis, N., Koletsa, T. & Kostopoulos, I. *Detection of centroblasts in H&E stained images of follicular lymphoma* in *2014 22nd Signal Processing and Communications Applications Conference (SIU)* (2014), 2319–2322. <https://ieeexplore.ieee.org/document/6830728>.
102. Klein, U. & Dalla-Favera, R. Germinal centres: role in B-cell physiology and malignancy. *Nature Reviews Immunology* *2008 8:1* **8**, 22–33. ISSN: 1474-1741. <https://www.nature.com/articles/nri2217> (2008).
103. Graham, S., Vu, Q. D., Raza, S. E., Azam, A., Tsang, Y. W., Kwak, J. T. & Rajpoot, N. Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis* **58**, 101563. ISSN: 1361-8415. <https://www.sciencedirect.com/science/article/pii/S1361841519301045> (2019).
104. Freedman, A. & Jacobsen, E. Follicular lymphoma: 2020 update on diagnosis and management. *American Journal of Hematology* **95**, 316–327. ISSN: 1096-8652. <https://onlinelibrary.wiley.com/doi/full/10.1002/ajh.25696>
<https://onlinelibrary.wiley.com/doi/abs/10.1002/ajh.25696>
<https://onlinelibrary.wiley.com/doi/10.1002/ajh.25696> (2020).

105. Constance, A. F., Andrew, P., Kanagaratnam, J., Leslie, S., Parkin, S. D. M. & Whelan, S. *World Health Organization Geneva* 2012. <https://www.who.int/publications/i/item/international-classification-of-diseases-for-oncology>.
106. *Study Details | A Study of Obinutuzumab (RO5072759) Plus Chemotherapy in Comparison With Rituximab Plus Chemotherapy Followed by Obinutuzumab or Rituximab Maintenance in Patients With Untreated Advanced Indolent Non-Hodgkin's Lymphoma (GALLIUM) | ClinicalTrials.gov* 2011. <https://clinicaltrials.gov/study/NCT01332968>.
107. Marcus, R., Davies, A., Ando, K., Klapper, W., Opat, S., Owen, C., Phillips, E., Sangha, R., Schlag, R., Seymour, J. F., Townsend, W., Trněný, M., Wenger, M., Fingerle-Rowson, G., Rufibach, K., Moore, T., Herold, M. & Hiddemann, W. Obinutuzumab for the First-Line Treatment of Follicular Lymphoma. *New England Journal of Medicine* **377**, 1331–1344. ISSN: 0028-4793. <https://www.nejm.org/doi/full/10.1056/NEJMoa1614598> (2017).
108. Schmidt, U., Weigert, M., Broaddus, C. & Myers, G. *Cell Detection with Star-Convex Polygons in Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II* **11071 LNCS** (Springer Verlag, 2018), 265–273. ISBN: 9783030009335. <https://arxiv.org/abs/1806.03535v2>.
109. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods* 2020 18:1 **18**, 100–106. ISSN: 1548-7105. <https://www.nature.com/articles/s41592-020-01018-x> (2020).
110. Vu, Q. D., Graham, S., Kurc, T., To, M. N. N., Shaban, M., Qaiser, T., Koohbanani, N. A., Khurram, S. A., Kalpathy-Cramer, J., Zhao, T., Gupta, R., Kwak, J. T., Rajpoot, N., Saltz, J. & Farahani, K. Methods for Segmentation and Classification of Digital Microscopy Tissue Images. *Frontiers in Bioengineering and Biotechnology*

7. ISSN: 2296-4185. <https://www.frontiersin.org/journals/bioengineering-and-biotechnology/articles/10.3389/fbioe.2019.00053> (2019).
111. Gamper, J., Koohbanani, N. A., Benes, K., Khuram, A. & Rajpoot, N. *PanNuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification* in *European Congress on Digital Pathology* (2019), 11–19. https://link.springer.com/chapter/10.1007/978-3-030-23937-4_2.
112. Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A. & Sethi, A. A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology. *IEEE Transactions on Medical Imaging* **36**, 1550–1560. <https://ieeexplore.ieee.org/document/7872382> (2017).
113. Verma, R., Kumar, N., Patil, A., Kurian, N. C., Rane, S., Graham, S., Vu, Q. D., Zwager, M., Raza, S. E. A., Rajpoot, N., Wu, X., Chen, H., Huang, Y., Wang, L., Jung, H., Brown, G. T., Liu, Y., Liu, S., Jahromi, S. A. F., Khani, A. A., Montahaei, E., Baghshah, M. S., Behroozi, H., Semkin, P., Rassadin, A., Dutande, P., Lodaya, R., Baid, U., Baheti, B., Talbar, S., Mahbod, A., Ecker, R., Ellinger, I., Luo, Z., Dong, B., Xu, Z., Yao, Y., Lv, S., Feng, M., Xu, K., Zunair, H., Hamza, A. B., Smiley, S., Yin, T.-K., Fang, Q.-R., Srivastava, S., Mahapatra, D., Trnavska, L., Zhang, H., Narayanan, P. L., Law, J., Yuan, Y., Tejomay, A., Mitkari, A., Koka, D., Ramachandra, V., Kini, L. & Sethi, A. MoNuSAC2020: A Multi-Organ Nuclei Segmentation and Classification Challenge. *IEEE Transactions on Medical Imaging* **40**, 3413–3423. <https://ieeexplore.ieee.org/document/9446924> (2021).
114. Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M. & Madabhushi, A. HistoQC: An Open-Source Quality Control Tool for Digital Pathology Slides. *JCO Clinical Cancer Informatics*, 1–7. ISSN: 24734276. <https://ascopubs.org/doi/10.1200/CCI.18.00157> (2019).

115. Walt, S. V. D., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E. & Yu, T. Scikit-image: Image processing in python. *PeerJ* **2014**. ISSN: 21678359. <https://peerj.com/articles/453/> (2014).
116. Miao, R., Toth, R., Zhou, Y., Madabhushi, A. & Janowczyk, A. Quick Annotator: an open-source digital pathology based rapid image annotation tool. *The Journal of Pathology: Clinical Research* **7**, 542–547. ISSN: 2056-4538. <https://onlinelibrary.wiley.com/doi/full/10.1002/cjp2.229> <https://onlinelibrary.wiley.com/doi/abs/10.1002/cjp2.229> <https://onlinelibrary.wiley.com/doi/10.1002/cjp2.229> (2021).
117. Wodzinski, M. & Skalski, A. Multistep, automatic and nonrigid image registration method for histology samples acquired using multiple stains. *Physics in Medicine & Biology* **66**, 025006. <https://dx.doi.org/10.1088/1361-6560/abcad7> (2021).
118. Breiman, L. Random forests. *Machine Learning* **45**, 5–32. ISSN: 08856125. <https://link.springer.com/article/10.1023/A:1010933404324> (2001).
119. Walker, C., Talawalla, T., Toth, R., Ambekar, A., Rea, K., Chamian, O., Fan, F., Berezowska, S., Rottenberg, S., Madabhushi, A., Maillard, M., Barisoni, L., Horlings, H. M. & Janowczyk, A. PatchSorter: a high throughput deep learning digital pathology tool for object labeling. *npj Digital Medicine* **2024 7:1** **7**, 1–7. ISSN: 2398-6352. <https://www.nature.com/articles/s41746-024-01150-4> (2024).
120. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Int. Res.* **16**, 321–357. ISSN: 1076-9757. <https://www.jair.org/index.php/jair/article/view/10302> (2002).
121. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. *Densely Connected Convolutional Networks* 2018. arXiv: 1608.06993 [cs.CV]. <https://arxiv.org/abs/1608.06993>.

122. Tan, M. & Le, Q. V. *EfficientNetV2: Smaller Models and Faster Training* 2021. arXiv: 2104.00298 [cs.CV]. <https://arxiv.org/abs/2104.00298>.
123. Davidson-Pilon, C. lifelines: survival analysis in Python. *Journal of Open Source Software* **4**, 1317. <https://doi.org/10.21105/joss.01317> (2019).
124. Rummel, M. J., Niederle, N., Maschmeyer, G., Banat, A., Gruenhagen, U. v., Losem, C., Heil, G., Welslau, M., Balser, C., Kaiser, U., Ballo, H., Weidmann, E., Duerk, H. A., Kofahl-Krause, D., Roller, F., Barth, J., Hoelzer, D., Hinke, A. & Brugger, W. Bendamustine Plus Rituximab Is Superior in Respect of Progression Free Survival and CR Rate When Compared to CHOP Plus Rituximab as First-Line Treatment of Patients with Advanced Follicular, Indolent, and Mantle Cell Lymphomas: Final Results of a Randomized Phase III Study of the StiL (Study Group Indolent Lymphomas, Germany). *Blood* **114**, 405–405. ISSN: 0006-4971. <https://doi.org/10.1182/blood.V114.22.405.405> (2009).
125. Flinn, I. W., van der Jagt, R., Kahl, B., Wood, P., Hawkins, T., MacDonald, D., Simpson, D., Kolibaba, K., Issa, S., Chang, J., Trotman, J., Hallman, D., Chen, L. & Burke, J. M. First-Line Treatment of Patients With Indolent Non-Hodgkin Lymphoma or Mantle-Cell Lymphoma With Bendamustine Plus Rituximab Versus R-CHOP or R-CVP: Results of the BRIGHT 5-Year Follow-Up Study. *Journal of Clinical Oncology* **37**. PMID: 30811293, 984–991. eprint: <https://ascopubs.org/doi/pdf/10.1200/JCO.18.00605>. <https://ascopubs.org/doi/abs/10.1200/JCO.18.00605> (2019).
126. Khan, A., Janowczyk, A., Müller, F., Blank, A., Nguyen, H. G., Abbet, C., Studer, L., Lugli, A., Dawson, H., Thiran, J.-P. & Zlobec, I. Impact of scanner variability on lymph node segmentation in computational pathology. *Journal of Pathology Informatics* **13**, 100127. ISSN: 2153-3539. <https://www.sciencedirect.com/science/article/pii/S2153353922007210> (2022).

127. Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J. M., Ciompi, F. & van der Laak, J. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis* **58**, 101544. ISSN: 1361-8415. <https://www.sciencedirect.com/science/article/pii/S1361841519300799> (2019).
128. van Rijthoven, M., Balkenhol, M., Siliņa, K., van der Laak, J. & Ciompi, F. HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. *Medical Image Analysis* **68**, 101890. ISSN: 1361-8415. <https://www.sciencedirect.com/science/article/pii/S1361841520302541> (2021).
129. Saxena, P., Goyal, A., Bivi, M. A., Singh, S. K. & Rashid, M. Segmentation of Nucleus and Cytoplasm from H&E-Stained Follicular Lymphoma. *Electronics* **12**. ISSN: 2079-9292. <https://www.mdpi.com/2079-9292/12/3/651> (2023).
130. Marafioti, T., Copie-Bergman, C., Calaminici, M., Paterson, J. C., Shende, V. H., Liu, H., Baia, M., Ramsay, A. D., Agostinelli, C., Brière, J., Clear, A., Du, M.-Q., Piccaluga, P. P., Masir, N., Nacheva, E. P., Sujobert, P., Shanmugam, K., Grogan, T. M., Brooks, S. P., Khwaja, A., Ardeshtna, K., Townsend, W., Pileri, S. A., Haioun, C., Linch, D., Gribben, J. G., Gaulard, P. & Isaacson, P. G. Another look at follicular lymphoma: immunophenotypic and molecular analyses identify distinct follicular lymphoma subgroups. *Histopathology* **62**, 860–875. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/his.12076>. <https://onlinelibrary.wiley.com/doi/abs/10.1111/his.12076> (2013).
131. Rummel, M. J., Niederle, N., Maschmeyer, G., Banat, G. A., von Grünhagen, U., Losem, C., Kofahl-Krause, D., Heil, G., Welslau, M., Balser, C., Kaiser, U., Weidmann, E., Dürk, H., Ballo, H., Stauch, M., Roller, F., Barth, J., Hoelzer, D., Hinke, A. & Brugger, W. Bendamustine plus rituximab versus CHOP plus rituximab as first-line treatment for patients with indolent and mantle-cell lymphomas: an open-

- label, multicentre, randomised, phase 3 non-inferiority trial. *The Lancet* **381**, 1203–1210. ISSN: 0140-6736. [https://doi.org/10.1016/S0140-6736\(12\)61763-2](https://doi.org/10.1016/S0140-6736(12)61763-2) (2013).
132. Luminari, S., Manni, M., Galimberti, S., Versari, A., Tucci, A., Boccomini, C., Farina, L., Olivieri, J., Marcheselli, L., Guerra, L., Ferrero, S., Arcaini, L., Cavallo, F., Kovalchuk, S., Skrypets, T., del Giudice, I., Chauvie, S., Patti, C., Stelitano, C., Ricci, F., Pinto, A., Margiotta Casaluci, G., Zilioli, V. R., Merli, A., Ladetto, M., Bolis, S., Pavone, V., Chiarenza, A., Arcari, A., Anastasia, A., Dondi, A., Manina, D., Federico, M. & null null, n. Response-Adapted Postinduction Strategy in Patients With Advanced-Stage Follicular Lymphoma: The FOLL12 Study. *Journal of Clinical Oncology* **40**. PMID: 34709880, 729–739. eprint: <https://ascopubs.org/doi/pdf/10.1200/JCO.21.01234>. <https://ascopubs.org/doi/abs/10.1200/JCO.21.01234> (2022).
133. Locke, F. L., Filosto, S., Chou, J., Vardhanabhuti, S., Perbost, R., Dreger, P., Hill, B. T., Lee, C., Zinzani, P. L., Kröger, N., López-Guillermo, A., Greinix, H., Zhang, W., Tiwari, G., Budka, J., Marincola, F. M., To, C., Mattie, M., Schupp, M., Cheng, P., Bot, A., Shen, R., Bedognetti, D., Miao, H. & Galon, J. Impact of tumor microenvironment on efficacy of anti-CD19 CAR T cell therapy or chemotherapy and transplant in large B cell lymphoma. *Nature Medicine* 2024 30:2 **30**, 507–518. ISSN: 1546-170X. <https://www.nature.com/articles/s41591-023-02754-1> (2024).
134. Blaker, Y. N., Spetalen, S., Brodtkorb, M., Lingjærde, O. C., Beiske, K., Østenstad, B., Sander, B., Wahlin, B. E., Melen, C. M., Myklebust, J. H., Holte, H., Delabie, J. & Smeland, E. B. The tumour microenvironment influences survival and time to transformation in follicular lymphoma in the rituximab era. *British Journal of Haematology* **175**, 102–114. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/bjh.14201>. <https://onlinelibrary.wiley.com/doi/abs/10.1111/bjh.14201> (2016).

135. Hagos, Y. B., Akarca, A. U., Ramsay, A., Rossi, R. L., Pomplun, S., Ngai, V., Moiola, A., Gianatti, A., Mcnamara, C., Rambaldi, A., Quezada, S. A., Lynch, D., Gritti, G., Yuan, Y. & Marafioti, T. High inter-follicular spatial co-localization of CD8+FOXP3+ with CD4+CD8+ cells predicts favorable outcome in follicular lymphoma. *Hematological Oncology*. ISSN: 1099-1069. <https://onlinelibrary.wiley.com/doi/full/10.1002/hon.3003><https://onlinelibrary.wiley.com/doi/abs/10.1002/hon.3003> (2022).
136. Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermesen, M., van de Loo, R., Vogels, R., Manson, Q. F., Stathonikos, N., Baidoshvili, A., van Diest, P., Wauters, C., van Dijk, M. & van der Laak, J. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience* **7**, giy065. ISSN: 2047-217X. eprint: https://academic.oup.com/gigascience/article-pdf/7/6/giy065/60709038/gigascience_7_6_giy065.pdf. <https://doi.org/10.1093/gigascience/giy065> (2018).
137. Graham, S., Vu, Q. D., Jahanifar, M., Weigert, M., Schmidt, U., Zhang, W., Zhang, J., Yang, S., Xiang, J., Wang, X., Rumberger, J. L., Baumann, E., Hirsch, P., Liu, L., Hong, C., Aviles-Rivero, A. I., Jain, A., Ahn, H., Hong, Y., Azzuni, H., Xu, M., Yaqub, M., Blache, M.-C., Piégu, B., Vernay, B., Scherr, T., Böhlend, M., Löffler, K., Li, J., Ying, W., Wang, C., Snead, D., Raza, S. E. A., Minhas, F. & Rajpoot, N. M. CoNIC Challenge: Pushing the frontiers of nuclear detection, segmentation, classification and counting. *Medical Image Analysis* **92**, 103047. ISSN: 1361-8415. <https://www.sciencedirect.com/science/article/pii/S1361841523003079> (2024).
138. Yu, W.-H., Li, C.-H., Wang, R.-C., Yeh, C.-Y. & Chuang, S.-S. Machine Learning Based on Morphological Features Enables Classification of Primary Intestinal T-

- Cell Lymphomas. *Cancers* **13**. ISSN: 2072-6694. <https://www.mdpi.com/2072-6694/13/21/5463> (2021).
139. Abel, J., Jain, S., Rajan, D., Padigela, H., Leidal, K., Prakash, A., Conway, J., Nercessian, M., Kirkup, C., Javed, S. A., Biju, R., Harguindeguy, N., Shenker, D., Indorf, N., Sanghavi, D., Egger, R., Trotter, B., Gerardin, Y., Brosnan-Cashman, J. A., Dhoot, A., Montalto, M. C., Parmar, C., Wapinski, I., Khosla, A., Drage, M. G., Yu, L. & Taylor-Weiner, A. AI powered quantification of nuclear morphology in cancers enables prediction of genome instability and prognosis. *npj Precision Oncology* 2024 8:1 **8**, 1–14. ISSN: 2397-768X. <https://www.nature.com/articles/s41698-024-00623-9> (2024).
 140. Holm, S. Handle with care: Assessing performance measures of medical AI for shared clinical decision-making. *Bioethics* **36**, 178–186. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/bioe.12930>. <https://onlinelibrary.wiley.com/doi/abs/10.1111/bioe.12930> (2022).
 141. MHRA. *Project Glass Box (interpretability): Software and AI as a Medical Device Change Programme* 2023. <https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme/software-and-ai-as-a-medical-device-change-programme-roadmap#wp-10-project-glass-box-ai-interpretability>.
 142. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L. & Prior, F. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging* **26**, 1045–1057. ISSN: 08971889. <https://link.springer.com/article/10.1007/s10278-013-9622-7> (2013).
 143. Fernandez-Pol, S., Natkunam, Y., Vrabac, D., Rojansky, R., Advani, R., Rajpurkar, P. S. & Ng, A. Y. H&E and immunohistochemical stain images of 209 cases of diffuse large B-cell lymphoma linked with cytogenetic features and clinical

- outcomes (Version 1) [Data set]. *The Cancer Imaging Archive*. <https://www.cancerimagingarchive.net/collection/dlbcl-morphology/> (2022).
144. Care, M. A., Barrans, S., Worrillow, L., Jack, A., Westhead, D. R. & Tooze, R. M. A Microarray Platform-Independent Classification Tool for Cell of Origin Class Allows Comparative Analysis of Gene Expression in Diffuse Large B-cell Lymphoma. *PLoS ONE* **8**, e55895. ISSN: 19326203. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0055895> (2013).
145. Painter, D., Barrans, S., Lacy, S., Smith, A., Crouch, S., Westhead, D., Sha, C., Patmore, R., Tooze, R., Burton, C. & Roman, E. Cell-of-origin in diffuse large B-cell lymphoma: findings from the UK's population-based Haematological Malignancy Research Network. *British Journal of Haematology* **185**, 781–784. ISSN: 1365-2141. <https://onlinelibrary.wiley.com/doi/full/10.1111/bjh.15619><https://onlinelibrary.wiley.com/doi/abs/10.1111/bjh.15619><https://onlinelibrary.wiley.com/doi/10.1111/bjh.15619> (2019).
146. Cucco, F., Barrans, S., Sha, C., Clipson, A., Crouch, S., Dobson, R., Chen, Z., Thompson, J. S., Care, M. A., Cummin, T., Caddy, J., Liu, H., Robinson, A., Schuh, A., Fitzgibbon, J., Painter, D., Smith, A., Roman, E., Tooze, R., Burton, C., Davies, A. J., Westhead, D. R., Johnson, P. W. & Du, M. Q. Distinct genetic changes reveal evolutionary history and heterogeneous molecular grade of DLBCL with MYC/BCL2 double-hit. *Leukemia* *2019 34:5* **34**, 1329–1341. ISSN: 1476-5551. <https://www.nature.com/articles/s41375-019-0691-6> (2019).
147. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. <https://arxiv.org/abs/1802.03426v3> (2018).
148. Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W., Ng, L. G., Ginhoux, F. & Newell, E. W. Dimensionality reduction for visualizing single-cell data using

- UMAP. *Nature Biotechnology* 2018 37:1 **37**, 38–44. ISSN: 1546-1696. <https://www.nature.com/articles/nbt.4314> (2018).
149. Packer, J. S., Zhu, Q., Huynh, C., Sivaramakrishnan, P., Preston, E., Dueck, H., Stefanik, D., Tan, K., Trapnell, C., Kim, J., Waterston, R. H. & Murray, J. I. A lineage-resolved molecular atlas of *C. Elegans* embryogenesis at single-cell resolution. *Science* **365**. ISSN: 10959203. <https://www.science.org/doi/10.1126/science.aax1971> (2019).
150. Hirschberg, J. B. & Rosenberg, A. V-Measure: A conditional entropy-based external cluster evaluation, 410–420. <https://academiccommons.columbia.edu/doi/10.7916/D80V8N84> (2007).
151. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Cell Reports Resource Deciphering Signatures of Mutational Processes Operative in Human Cancer. <http://dx.doi.org/10.1016/j.celrep.2012.12.008> (2013).
152. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P. & SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17**, 261–272. <https://www.nature.com/articles/s41592-019-0686-2> (2020).
153. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**. ISSN: 15324435. <https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html> (2011).

154. Solal-Céligny, P., Roy, P., Colombat, P., White, J., Armitage, J. O., Arranz-Saez, R., Au, W. Y., Bellei, M., Brice, P., Caballero, D., Coiffier, B., Conde-Garcia, E., Doyen, C., Federico, M., Fisher, R. I., Garcia-Conde, J. F., Guglielmi, C., Hagenebeck, A., Haïoun, C., LeBlanc, M., Lister, A. T., Lopez-Guillermo, A., McLaughlin, P., Milpied, N., Morel, P., Mounier, N., Proctor, S. J., Rohatiner, A., Smith, P., Soubeyran, P., Tilly, H., Vitolo, U., Zinzani, P.-L., Zucca, E. & Montserrat, E. Follicular Lymphoma International Prognostic Index. *Blood* **104**, 1258–1265. ISSN: 0006-4971. eprint: <https://ashpublications.org/blood/article-pdf/104/5/1258/1701112/zh801704001258.pdf>. <https://doi.org/10.1182/blood-2003-12-4434> (2004).
155. Bachy, E., Maurer, M. J., Habermann, T. M., Gelas-Dore, B., Maucourt-Boulch, D., Estell, J. A., Van den Neste, E., Bouabdallah, R., Gyan, E., Feldman, A. L., Bargay, J., Delmer, A., Slager, S. L., Gomes da Silva, M., Fitoussi, O., Belada, D., Maisonneuve, H., Intragumtornchai, T., Ansell, S. M., Lamy, T., Dartigues, P., Link, B. K., Seymour, J. F., Cerhan, J. R. & Salles, G. A simplified scoring system in de novo follicular lymphoma treated initially with immunochemotherapy. *Blood* **132**, 49–58. ISSN: 0006-4971. eprint: <https://ashpublications.org/blood/article-pdf/132/1/49/1406502/blood816405.pdf>. <https://doi.org/10.1182/blood-2017-11-816405> (2018).
156. Khanlari, M. & Chapman, J. R. Follicular lymphoma: updates for pathologists. *Journal of Pathology and Translational Medicine* **56**, 1. ISSN: 23837845. /pmc/articles/PMC8743801/%20/pmc/articles/PMC8743801/?report=abstract%20<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8743801/> (2022).
157. Behery, R. E., Laurini, J. A., Weisenburger, D. D., Smith, L. M., Dave, B. J., Yuan, J., Fu, K., Chan, W. C., Nathwani, B. N., Bierman, P. J., Bociek, R. G., Vose, J. M., Armitage, J. O., Greiner, T. C. & Aoun, P. Follicular large cleaved cell (centrocytic) lymphoma: an unrecognized variant of follicular lymphoma. *Human*

- Pathology* **72**, 180–190. ISSN: 0046-8177. <https://www.sciencedirect.com/science/article/pii/S0046817717304082> (2018).
158. Zamò, A., Pischmarov, J., Horn, H., Ott, G., Rosenwald, A. & Leich, E. The ex-omic landscape of t(14;18)-negative diffuse follicular lymphoma with 1p36 deletion. *British Journal of Haematology* **180**, 391–394. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/bjh.15041>. <https://onlinelibrary.wiley.com/doi/abs/10.1111/bjh.15041> (2018).
 159. Nann, D., Ramis-Zaldivar, J. E., Müller, I., Gonzalez-Farre, B., Schmidt, J., Egan, C., Salmeron-Villalobos, J., Clot, G., Mattern, S., Otto, F., Mankel, B., Colomer, D., Balagué, O., Szablewski, V., Lome-Maldonado, C., Leoncini, L., Dojcinov, S., Chott, A., Copie-Bergman, C., Bonzheim, I., Fend, F., Jaffe, E. S., Campo, E., Salaverria, I. & Quintanilla-Martinez, L. Follicular lymphoma t(14;18)-negative is genetically a heterogeneous disease. *Blood Advances* **4**, 5652–5665. ISSN: 2473-9529. eprint: <https://ashpublications.org/bloodadvances/article-pdf/4/22/5652/1790410/advancesadv2020002944.pdf>. <https://doi.org/10.1182/bloodadvances.2020002944> (2020).
 160. Ortega-Molina, A., Boss, I. W., Canela, A., Pan, H., Jiang, Y., Zhao, C., Jiang, M., Hu, D., Agirre, X., Niesvizky, I., Lee, J. E., Chen, H. T., Ennishi, D., Scott, D. W., Mottok, A., Hother, C., Liu, S., Cao, X. J., Tam, W., Shaknovich, R., Garcia, B. A., Gascoyne, R. D., Ge, K., Shilatifard, A., Elemento, O., Nussenzweig, A., Melnick, A. M. & Wendel, H. G. The histone lysine methyltransferase KMT2D sustains a gene expression program that represses B cell lymphoma development. *Nature Medicine* **21**, 1199–1208. ISSN: 1546-170X. <https://www.nature.com/articles/nm.3943> (2015).
 161. Kurz, K. S., Kalmbach, S., Ott, M., Staiger, A. M., Ott, G. & Horn, H. Follicular Lymphoma in the 5th Edition of the WHO-Classification of Haematolymphoid

- Neoplasms-Updated Classification and New Biological Data. *Cancers* **15**. ISSN: 2072-6694. <https://pubmed.ncbi.nlm.nih.gov/36765742/> (2023).
162. Karube, K., Satou, A. & Kato, S. New classifications of B-cell neoplasms: a comparison of 5th WHO and International Consensus classifications. *International Journal of Hematology*, 1–11. ISSN: 18653774. <https://link.springer.com/article/10.1007/s12185-024-03781-5> (2024).
 163. Siddiqi, I. N., Friedman, J., Barry-Holson, K. Q., Ma, C., Thodima, V., Kang, I., Padmanabhan, R., Dias, L. M., Kelly, K. R., Brynes, R. K., Kamalakaran, S. & Houldsworth, J. Characterization of a variant of t(14;18) negative nodal diffuse follicular lymphoma with CD23 expression, 1p36/TNFRSF14 abnormalities, and STAT6 mutations. *Modern Pathology* **29**, 570–581. ISSN: 15300285. [http://www.modernpathology.org/article/S0893395222017549/fulltext%20http://www.modernpathology.org/article/S0893395222017549/abstract%20https://www.modernpathology.org/article/S0893-3952\(22\)01754-9/abstract](http://www.modernpathology.org/article/S0893395222017549/fulltext%20http://www.modernpathology.org/article/S0893395222017549/abstract%20https://www.modernpathology.org/article/S0893-3952(22)01754-9/abstract) (2016).
 164. Kalmbach, S., Grau, M., Zapukhlyak, M., Leich, E., Jurinovic, V., Hoster, E., Staiger, A. M., Kurz, K. S., Weigert, O., Gaitzsch, E., Passerini, V., Engelhard, M., Herfarth, K., Beiske, K., Micci, F., Möller, P., Bernd, H. W., Feller, A. C., Klapper, W., Stein, H., Hansmann, M. L., Hartmann, S., Dreyling, M., Holte, H., Lenz, G., Rosenwald, A., Ott, G. & Horn, H. Novel insights into the pathogenesis of follicular lymphoma by molecular profiling of localized and systemic disease forms. *Leukemia* 2023 37:10 **37**, 2058–2065. ISSN: 1476-5551. <https://www.nature.com/articles/s41375-023-01995-w> (2023).
 165. Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltneane, J. M., Hurt, E. M., Zhao, H., Averett, L., Yang, L., Wilson, W. H., Jaffe, E. S., Simon, R., Klausner, R. D., Powell, J., Duffey, P. L., Longo, D. L., Greiner, T. C., Weisen-

- burger, D. D., Sanger, W. G., Dave, B. J., Lynch, J. C., Vose, J., Armitage, J. O., Montserrat, E., López-Guillermo, A., Grogan, T. M., Miller, T. P., LeBlanc, M., Ott, G., Kvaloy, S., Delabie, J., Holte, H., Krajci, P., Stokke, T. & Staudt, L. M. The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma. *New England Journal of Medicine* **346**, 1937–1947. eprint: <https://www.nejm.org/doi/pdf/10.1056/NEJMoa012914>. <https://www.nejm.org/doi/full/10.1056/NEJMoa012914> (2002).
166. Lu, C., Koyuncu, C., Corredor, G., Prasanna, P., Leo, P., Wang, X., Janowczyk, A., Bera, K., Lewis Jr., J., Velcheti, V. & Madabhushi, A. Feature-driven local cell graph (FLoK): New computational pathology-based descriptors for prognosis of lung cancer and HPV status of oropharyngeal cancers. *Medical Image Analysis* **68**, 101903. ISSN: 1361-8415. <https://www.sciencedirect.com/science/article/pii/S136184152030267X> (2021).
167. Shephard, A. J., Bashir, R. M. S., Mahmood, H., Jahanifar, M., Minhas, F., Raza, S. E., McCombe, K. D., Craig, S. G., James, J., Brooks, J., Nankivell, P., Mehanna, H., Khurram, S. A. & Rajpoot, N. M. A fully automated and explainable algorithm for predicting malignant transformation in oral epithelial dysplasia. *npj Precision Oncology* **2024** 8:1 **8**, 1–12. ISSN: 2397-768X. <https://www.nature.com/articles/s41698-024-00624-8> (2024).
168. Bilal, M., Jewsbury, R., Wang, R., AlGhamdi, H. M., Asif, A., Eastwood, M. & Rajpoot, N. An aggregation of aggregation methods in computational pathology. *Medical Image Analysis* **88**, 102885. ISSN: 1361-8415. <https://www.sciencedirect.com/science/article/pii/S1361841523001457> (2023).
169. Johnston, A., Brynes, R. K., Naemi, K., Reisian, N., Bhansali, D., Zhao, X. & Rezk, S. A. Differentiating Benign From Malignant Bone Marrow B-Cell Lymphoid Aggregates: A Statistical Analysis of Distinguishing Features. *Archives of Pathology & Laboratory Medicine* **139**, 233–240. ISSN: 0003-9985. eprint: <https://meridian>.

- allenpress.com/aplm/article-pdf/139/2/233/1614468/arpa_2013-0678-
oa.pdf. <https://doi.org/10.5858/arpa.2013-0678-0A> (2015).
170. Falini, B., Martino, G. & Lazzi, S. A comparison of the International Consensus and 5th World Health Organization classifications of mature B-cell lymphomas. *Leukemia* 2022 37:1 **37**, 18–34. ISSN: 1476-5551. <https://www.nature.com/articles/s41375-022-01764-1> (2022).
 171. Papaleo, N., Molina-Alvarez, A., Tapia, G., Onieva, R., Salido, M., Lome-Maldonado, C., Ara-Mancebo, X., Puiggros, A., Espinet, B., Blazquez, C., Fuertes, D., Sanchez-Gonzalez, B., Yelamos, J., Calvo, X. & Colomo, L. Combined immunohistochemical profile CD10/LMO2/MYC is a useful tool to screen MYC rearrangements in aggressive large B-cell lymphomas. *Virchows Archiv*, 1–9. ISSN: 14322307. <https://link.springer.com/article/10.1007/s00428-024-03945-y> (2024).
 172. Theophraste, H., Bruneau, J., Copie, C., Ruminy, P., Tony, P., Briere, J., Delarue, R., Parrens, M., Haioun, C., Canioni, D., Casasnovas, O., Fabiani, B., Morschhauser, F., Emile, J.-F., Tilly, H., Figeac, M., Gaulard, P., Salles, G. A., Jardin, F., Leroy, K., Jais, J.-P. & Molina, T. J. A New Immunohistochemical Algorithm Using Four Antibodies Has a High Concordance with Gene Expression Profiling Defined ABC and GCB Diffuse Large B-Cell Lymphoma. *Blood* **130**, 416–416. ISSN: 0006-4971. https://dx.doi.org/10.1182/blood.V130.Suppl_1.416.416 (2017).
 173. Shamaï, G., Livne, A., Polónia, A., Sabo, E., Cretu, A., Bar-Sela, G. & Kimmel, R. Deep learning-based image analysis predicts PD-L1 status from H&E-stained histopathology images in breast cancer. *Nature Communications* 2022 13:1 **13**, 1–13. ISSN: 2041-1723. <https://www.nature.com/articles/s41467-022-34275-9> (2022).
 174. Wen, Z., Luo, D., Wang, S., Rong, R., Evers, B. M., Jia, L., Fang, Y., Daoud, E. V., Yang, S., Gu, Z., Arner, E. N., Lewis, C. M., Soto, L. M. S., Fujimoto, J., Behrens, C., Wistuba, I. I., Yang, D. M., Brekken, R. A., O'Donnell, K. A., Xie, Y. & Xiao,

- G. Deep Learning-Based H-Score Quantification of Immunohistochemistry-Stained Images. *Modern Pathology* **37**. ISSN: 15300285. [http://www.modernpathology.org/article/S0893395223003034/fulltext%20http://www.modernpathology.org/article/S0893395223003034/abstract%20https://www.modernpathology.org/article/S0893-3952\(23\)00303-4/abstract](http://www.modernpathology.org/article/S0893395223003034/fulltext%20http://www.modernpathology.org/article/S0893395223003034/abstract%20https://www.modernpathology.org/article/S0893-3952(23)00303-4/abstract) (2024).
175. Martino, F., Ilardi, G., Varricchio, S., Russo, D., Di Crescenzo, R. M., Staibano, S. & Merolla, F. A deep learning model to predict Ki-67 positivity in oral squamous cell carcinoma. *Journal of Pathology Informatics* **15**, 100354. ISSN: 2153-3539. <https://www.sciencedirect.com/science/article/pii/S2153353923001682> (2024).
176. Ghahremani, P., Li, Y., Kaufman, A., Vanguri, R., Greenwald, N., Angelo, M., Hollmann, T. J. & Nadeem, S. Deep learning-inferred multiplex immunofluorescence for immunohistochemical image quantification. *Nature Machine Intelligence* **4**, 401–412. <https://www.nature.com/articles/s42256-022-00471-x> (2022).
177. Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., Schmitt, C. & Thomas, N. E. *A method for normalizing histology slides for quantitative analysis* in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro* (2009), 1107–1110. <https://ieeexplore.ieee.org/document/5193250>.
178. Patkar, S., Chen, A., Basnet, A., Bixby, A., Rajendran, R., Chernet, R., Faso, S., Kumar, P. A., Desai, D., El-Zammar, O., Curtiss, C., Carello, S. J., Nasr, M. R., Choyke, P., Harmon, S., Turkbey, B. & Jamaspishvili, T. Predicting the tumor microenvironment composition and immunotherapy response in non-small cell lung cancer from digital histopathology images. *npj Precision Oncology* **2024 8:1** **8**, 1–15. ISSN: 2397-768X. <https://www.nature.com/articles/s41698-024-00765-w> (2024).

179. Valieris, R., Martins, L., Defelicibus, A., Bueno, A. P., de Toledo Osorio, C. A. B., Carraro, D., Dias-Neto, E., Rosales, R. A., de Figueiredo, J. M. B. & da Silva, I. T. Weakly-supervised deep learning models enable HER2-low prediction from H&E stained slides. *Breast Cancer Research* **26**, 1–11. ISSN: 1465542X. <https://breast-cancer-research.biomedcentral.com/articles/10.1186/s13058-024-01863-0> (2024).
180. Wang, R., Gunesli, G. N., Skingen, V. E., Valen, K.-A. F., Lyng, H., Young, L. S. & Rajpoot, N. Deep learning for predicting prognostic consensus molecular subtypes in cervical cancer from histology images. *npj Precision Oncology* 2025 9:1 **9**, 1–13. ISSN: 2397-768X. <https://www.nature.com/articles/s41698-024-00778-5> (2025).
181. Liakopoulos, P., Massonnet, J., Bonjour, J., Mizrakli, M. T., Graham, S., Cuen-det, M. A., Seipel, A. H., Michielin, O., Merkler, D. & Janowczyk, A. HoverFast: an accurate, high-throughput, clinically deployable nuclear segmentation tool for brightfield digital pathology images. *Journal of Open Source Software* **9**, 7022. <https://doi.org/10.21105/joss.07022> (2024).
182. Ying, R., Bourgeois, D., You, J., Zitnik, M. & Leskovec, J. *GNNEExplainer: Generating Explanations for Graph Neural Networks* 2019. arXiv: 1903.03894 [cs.LG].
183. Gars, E., Butzmann, A., Ohgami, R., Balakrishna, J. P. & O'Malley, D. P. The life and death of the germinal center. *Annals of Diagnostic Pathology* **44**, 151421. ISSN: 1092-9134. <https://www.sciencedirect.com/science/article/pii/S1092913419303508> (2020).
184. He, M., Roussak, K., Ma, F., Borcharding, N., Garin, V., White, M., Schutt, C., Jensen, T. I., Zhao, Y., Iberg, C. A., Shah, K., Bhatia, H., Korenfeld, D., Dinkel, S., Gray, J., Antonova, A. U., Ferris, S., Donermeyer, D., Arlehamn, C. L., Gubin, M. M., Luo, J., Gorvel, L., Pellegrini, M., Sette, A., Tung, T., Bak, R., Modlin, R. L., Fields, R. C., Schreiber, R. D., Allen, P. M. & Klechevsky, E. CD5 expression

- by dendritic cells directs T cell immunity and sustains immunotherapy responses. *Science* **379**, eabg2752. eprint: <https://www.science.org/doi/pdf/10.1126/science.abg2752>. <https://www.science.org/doi/abs/10.1126/science.abg2752> (2023).
185. Mintz, M. A. & Cyster, J. G. T follicular helper cells in germinal center B cell selection and lymphomagenesis. *Immunological Reviews* **296**, 48–61. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/imr.12860>. <https://onlinelibrary.wiley.com/doi/abs/10.1111/imr.12860> (2020).
 186. Upadhyay, R., Hammerich, L., Peng, P., Brown, B., Merad, M. & Brody, J. D. Lymphoma: Immune Evasion Strategies. *Cancers* **7**, 736–762. ISSN: 2072-6694. <https://www.mdpi.com/2072-6694/7/2/736> (2015).
 187. Wang, J. M. *et al.* Deep learning integrates histopathology and proteogenomics at a pan-cancer level. *Cell Reports Medicine* **4**, 101173. ISSN: 2666-3791. [https://www.cell.com/cell-reports-medicine/fulltext/S2666-3791\(23\)00326-9?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS2666379123003269%3Fshowall%3Dtrue](https://www.cell.com/cell-reports-medicine/fulltext/S2666-3791(23)00326-9?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS2666379123003269%3Fshowall%3Dtrue) (2023).
 188. Moran CORE Moran Eye Centre, U. *Cellular Histopathology* 2015. <https://morancore.utah.edu/section-04-ophthalmic-pathology/cellular-histopathology/>.
 189. Perrett, M., Edmondson, C. & Okosun, J. Biology of follicular lymphoma: insights and windows of clinical opportunity. *Hematology* **2022**, 688–694. ISSN: 1520-4391. eprint: <https://ashpublications.org/hematology/article-pdf/2022/1/688/2022038/688perrett.pdf>. <https://doi.org/10.1182/hematology.2022000361> (2022).
 190. Cerchietti, L. Genetic mechanisms underlying tumor microenvironment composition and function in diffuse large B-cell lymphoma. *Blood* **143**, 1101–1111. ISSN:

- 0006-4971. <https://www.sciencedirect.com/science/article/pii/S0006497124001642> (2024).
191. Graham, S., Minhas, F., Bilal, M., Ali, M., Tsang, Y. W., Eastwood, M., Wahab, N., Jahanifar, M., Hero, E., Dodd, K., Sahota, H., Wu, S., Lu, W., Azam, A., Benes, K., Nimir, M., Hewitt, K., Bhalerao, A., Robinson, A., Eldaly, H., Raza, S. E. A., Gopalakrishnan, K., Snead, D. & Rajpoot, N. Screening of normal endoscopic large bowel biopsies with interpretable graph learning: a retrospective study. *Gut* **72**, 1709–1721. ISSN: 0017-5749. eprint: <https://gut.bmj.com/content/72/9/1709.full.pdf>. <https://gut.bmj.com/content/72/9/1709> (2023).
 192. Lu, M. Y., Williamson, D. F., Chen, T. Y., Chen, R. J., Barbieri, M. & Mahmood, F. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering* **5**, 555–570. ISSN: 1361-8415. <https://www.sciencedirect.com/science/article/pii/S136184152030267X> (2021).
 193. Courtiol, P., Tramel, E. W., Sanselme, M. & Wainrib, G. Classification and Disease Localization in Histopathology Using Only Global Labels: A Weakly-Supervised Approach. arXiv: 1802.02212 [cs.CV]. <https://arxiv.org/abs/1802.02212> (2020).
 194. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J. & Han, X. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis* **81**, 102559. ISSN: 1361-8415. <https://www.sciencedirect.com/science/article/pii/S1361841522002043> (2022).
 195. MHRA. *Medical devices: software applications* 2023. <https://www.gov.uk/government/publications/medical-devices-software-applications-apps>.
 196. FDA. *Software Algorithm Device To Assist Users In Digital Pathology* 2025. <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMN/pmn.cfm?ID=K241232>.

197. FDA, M.
bibinitperiod. *Good Machine Learning Practice for Medical Device Development: Guiding Principles* 2021. <https://www.gov.uk/government/publications/good-machine-learning-practice-for-medical-device-development-guiding-principles/good-machine-learning-practice-for-medical-device-development-guiding-principles>.
198. NICE. *NICE statement of intent for artificial intelligence (AI)* 2024. <https://www.nice.org.uk/corporate/ecd12/chapter/objectives>.
199. Digital, N. *Secure Data Environment* 2025. <https://digital.nhs.uk/services/secure-data-environment-service>.
200. Kirillov, A., He, K., Girshick, R., Rother, C. & Dollar, P. *Panoptic Segmentation* 2019. arXiv: 1801.00868. <https://arxiv.org/abs/1801.00868>.
201. Vu, Q. *HoVerNet* 2022. https://github.com/vqdang/hover_net/archive/67e2ce5.zip.
202. Giavarina, D. Understanding Bland Altman analysis. *Biochemia Medica* **25**, 141. ISSN: 13300962. <https://pmc.ncbi.nlm.nih.gov/articles/PMC4470095/> (2015).

Appendix

1 Chapter 1: Introduction

Magnification in digital slides

Magnification in traditional pathology relates to the ratio between the perceived size of an object through a magnifying instrument (image size) and the object's true size, with the term 'size' referring to a length measure. Common magnifications include $400\times$, which would create an image of width 4cm from a biopsy sample of width 0.1mm. Digital imaging is far less restricted in magnification than microscopes - there is a physical limit to the magnification of a microscope while a digital image can be scaled infinitely - but there is a limit to the 'information' presented. This is limited in large part by the resolution - the minimum distance that two points can be identified as being distinct - of the image. In traditional microscopy, resolution is mostly limited by the wavelength of light used and an observer's eyesight. In digital images, this limit is set by the density of display units, pixels, which are the minimum unit of distinct colour. Scaling a digital image many times will increase image size and therefore, magnification but will not provide more information about what is between pixels. 'Magnification' is commonly used in digital pathology to communicate the image type and so it is also used in this thesis. For clarity on the amount of information presented within images, the resolution is reported in terms of physical distance represented by the side length of a single pixel, μm per pixel. Throughout the thesis, magnifications used and respective resolutions are:

- $40\times$, relating to resolution of $0.25\mu\text{m}$ per pixel
- $20\times$, relating to resolution of $0.50\mu\text{m}$ per pixel
- $1\times$, relating to resolution of $10\mu\text{m}$ per pixel

2 Chapter 2: Cytological Grading

Feature categor- ies	Dataset	HMRN		GALLIUM					
	Part- ition	All		All suff.		Len.		Pass	
		Y (%)	n	Y (%)	n	Y (%)	n	Y (%)	n
Scan- ning	Hama- matsu	0 (0)	308	498 (96.0)	519	380 (100)	380	269 (100)	269
	Aperio	308 (100)	308	21 (4.0)	519	0 (0)	380	0 (0)	269
Discrete features	Male	158 (51.3)	308	232 (44.7)	519	157 (41.3)	380	108 (40.1)	269
	Night sweats	53 (18.1)	293	135 (26.0)	519	98 (25.8)	380	72 (26.8)	269
	Fever	0 (0)	297	18 (3.5)	519	15 (3.9)	380	12 (4.5)	269
	Weight loss	44 (14.8)	297	89 (17.1)	519	63 (16.6)	380	44 (16.4)	269
	Extra- nodal sites	90 (30.1)	299	347 (66.9)	519	248 (65.3)	380	175 (65.1)	269
	Marrow involv- ement	87 (32.2)	270	263 (51.7)	509	197 (52.8)	373	142 (54.2)	262
	Elevated LDH	55 (19.6)	280	162 (31.2)	519	120 (31.6)	380	84 (31.2)	269
	POD24	27 (25.5)	106	73 (14.1)	519	52 (13.7)	380	34 (12.6)	269

Immuno-chemo-therapy	Chemo- (any)	117 (38.0)	308	519 (100)	519	380 (100)	380	269 (100)	269
	Benda- must- ine	1 (0.3)	305	338 (65.1)	519	244 (64.2)	380	175 (65.1)	269
	CVP	75 (24.6)	305	55 (10.6)	519	33 (8.7)	380	17 (6.3)	269
	CHOP	28 (9.2)	305	126 (24.3)	519	103 (27.1)	380	77 (28.6)	269
	Other chemo-	11 (3.6)	305	0 (0)	519	0 (0)	380	0 (0)	269
	Imm- uno (any)	107 (35.1)	305	519 (100)	519	380 (100)	380	269 (100)	269
	Rit- uxi- mab	107 (35.1)	305	251 (48.4)	519	192 (50.5)	380	129 (48.0)	269
Other treat- ments	Palli- ative	3 (1.0)	308	N/A					
	Radio- therapy	67 (21.8)	308	33 (6.4)	519	26 (6.8)	380	15 (5.6)	269
	Watch & Wait	121 (39.3)	308	N/A					
Contin- uous features		Median (IQR)	n	Median (IQR)	n	Median (IQR)	n	Median (IQR)	n
	Age	62.9 (53.6 -73.0)	308	58.0 (48.0-66.0)	519	57.0 (48.0-66.0)	380	57.0 (48.0-66.0)	269

ECOG	0.0 (0.0-1.0)	307	1.0 (1.0-1.0)	519	1.0 (1.0-1.0)	380	1.0 (1.0-1.0)	269
Stage	3.0 (2.0-4.0)	273	4.0 (3.0-4.0)	517	4.0 (3.0-4.0)	379	4.0 (3.0-4.0)	268
FLIPI	2.0 (1.0-3.0)	299	2.0 (2.0-3.0)	519	2.0 (2.0-3.0)	380	2.0 (2.0-3.0)	269
PRIMA-PI quantile	2.0 (1.0-2.0)	158	2.0 (1.0-3.0)	519	2.0 (1.0-3.0)	380	2.0 (1.0-3.0)	269

Table A.1: Patient characteristics in the HMRN and external validation GALLIUM datasets. IQR = interquartile range.

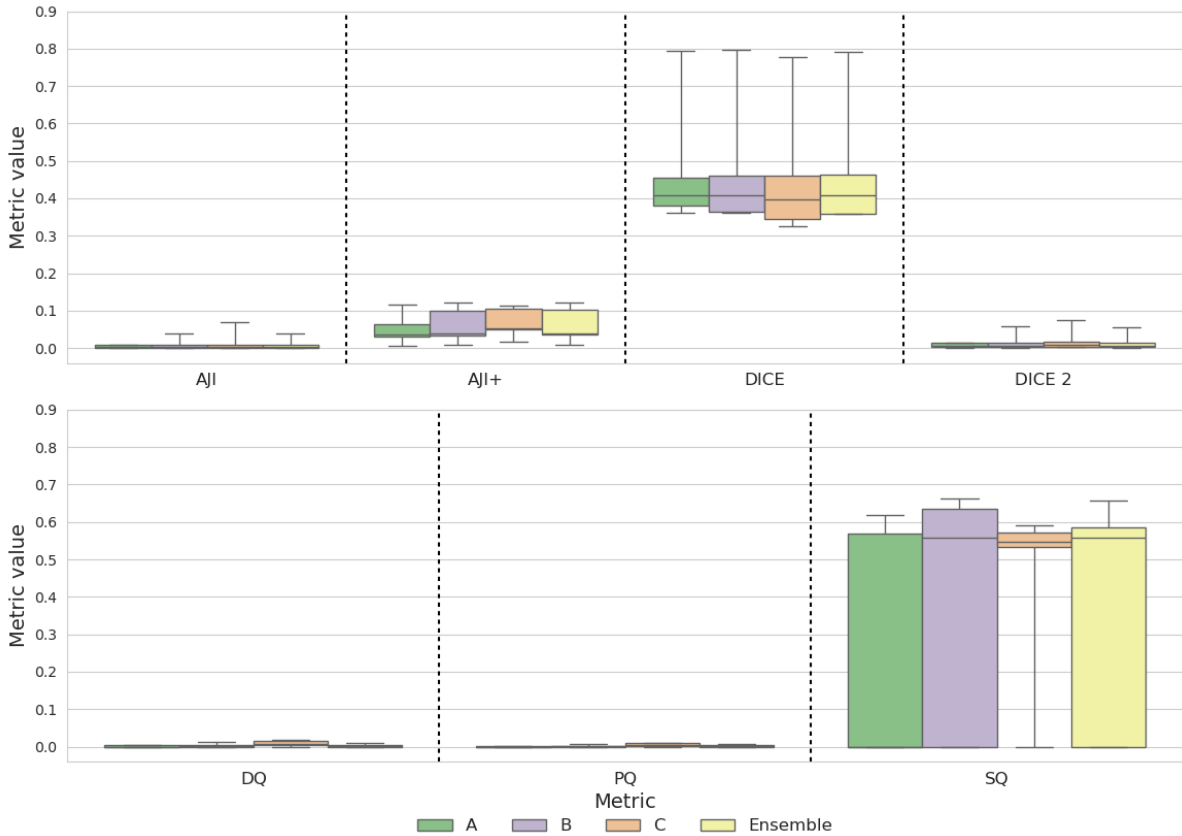
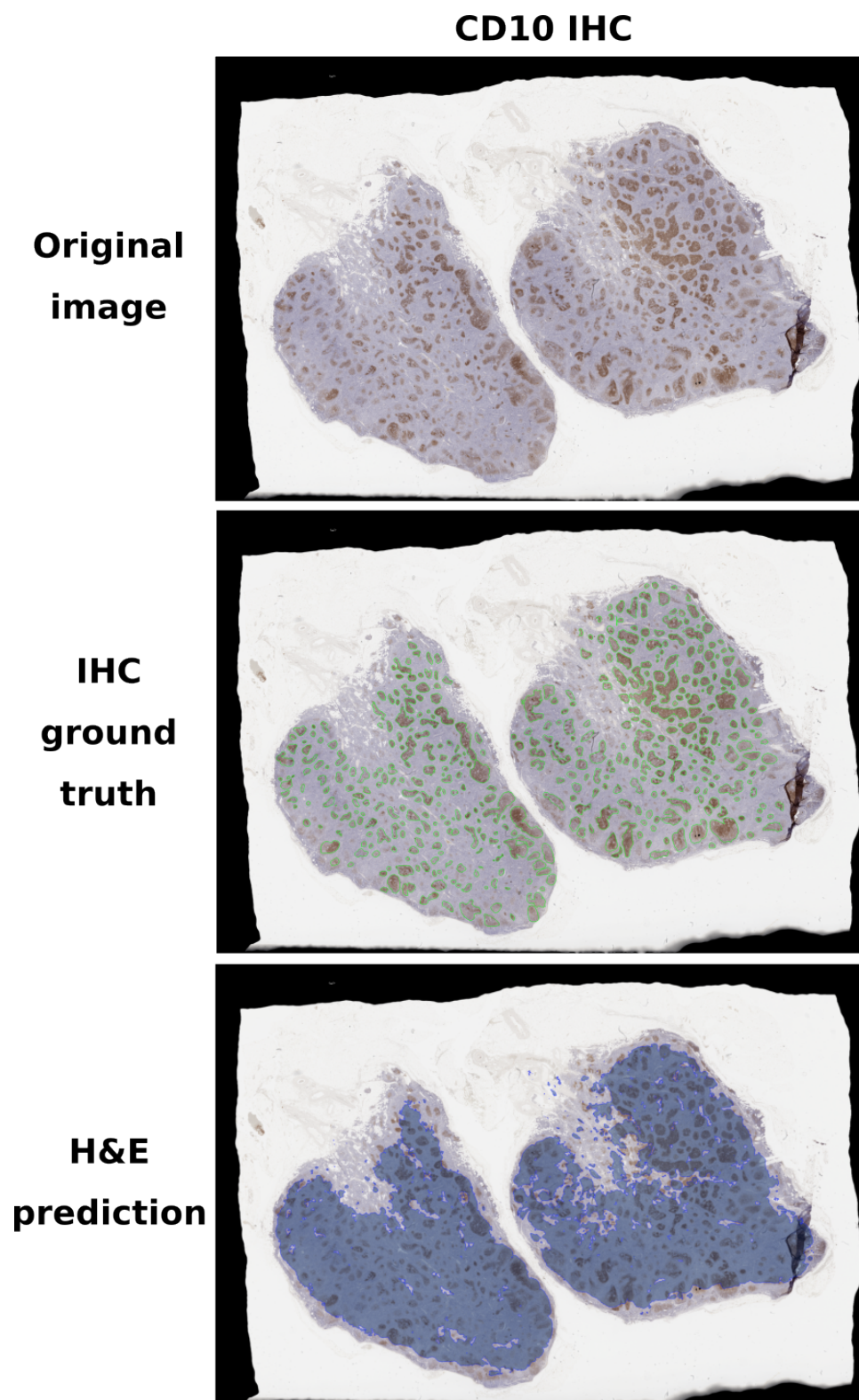
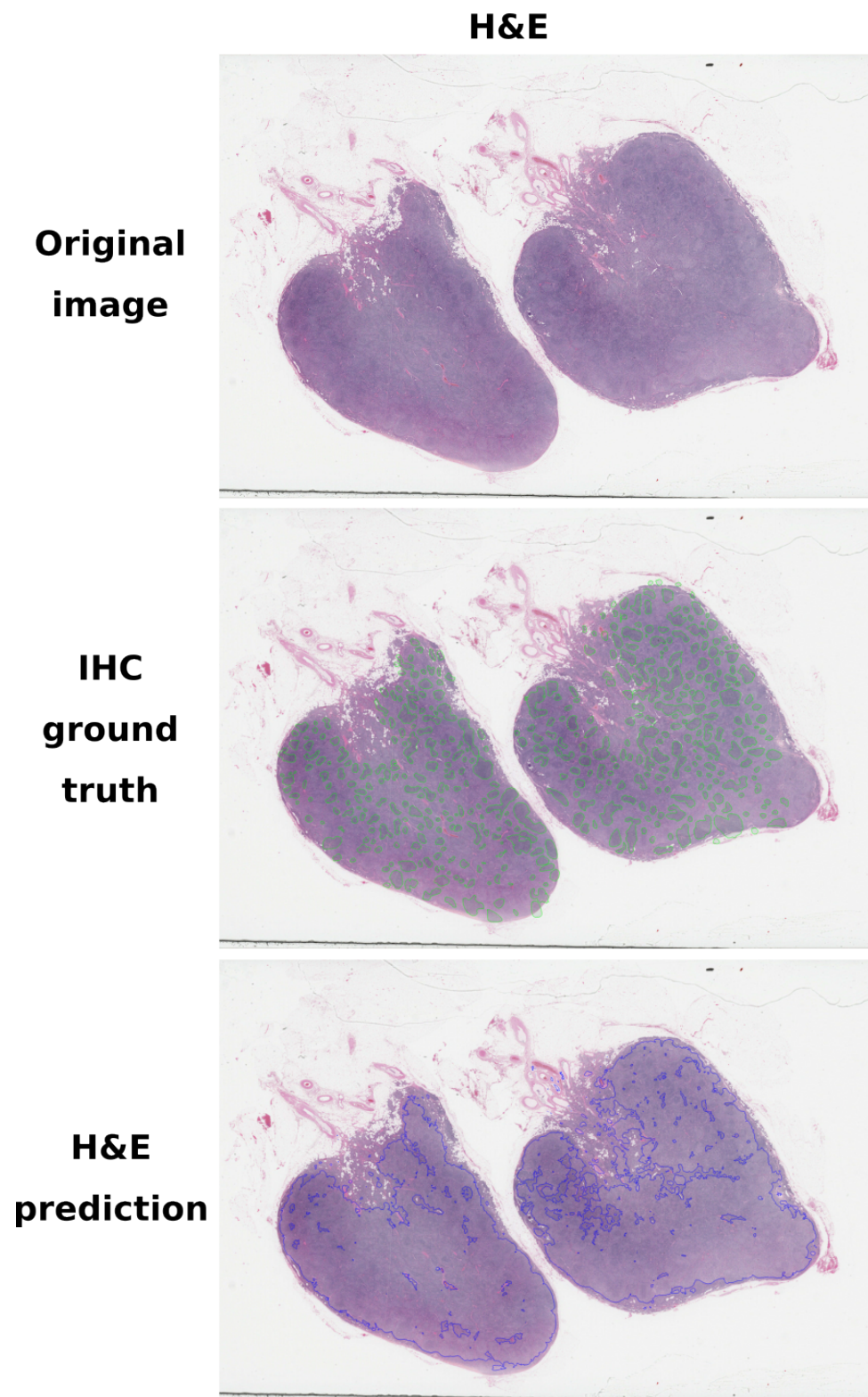


Figure A.1: Performance of the H&E Follicle Segmentation Algorithm with respect to an immunohistochemistry ground truth. The naive B-cell marker CD10 was used for annotation of ground truth. AJI = Aggregated Jaccard Index. AJI+ = Aggregated Jaccard Index Plus. DICE2 = Ensemble DICE score. DQ = Detection Quality. PQ = Panoptic Quality. SQ = Segmentation Quality. Metrics are defined in Chapter 2, Section 4.1, Segmentation.



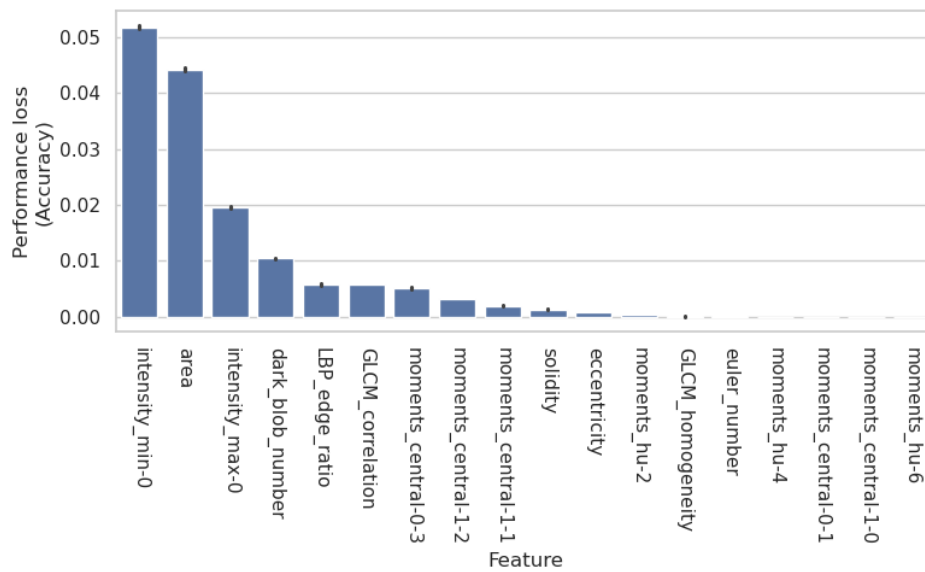
(a) Registered CD10 IHC slide



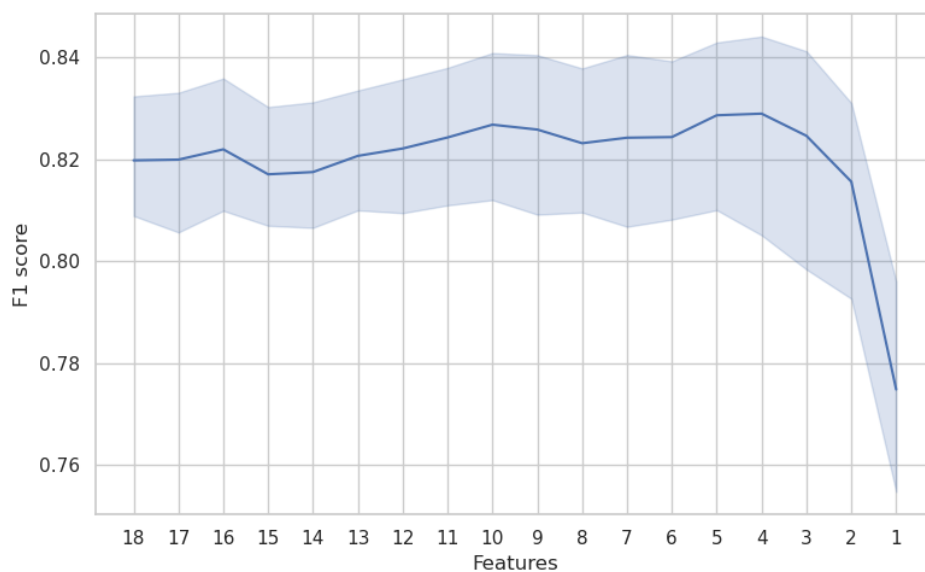
(b) H&E slide

Figure A.2: Performance of the H&E Follicle Segmentation Algorithm with respect to a CD10 immunohistochemistry ground truth. Details of the development of the algorithm are provided in Chapter 2, Section 4.1, Naive B-cell Region Classification. Green = IHC ground truth annotations. Blue = Algorithm predictions on H&E data.

- a) CD10 ground truth and H&E algorithm predictions superimposed on CD10 IHC slide, used for annotation of IHC ground truth.
 b) CD10 ground truth and H&E algorithm predictions superimposed on H&E slide, used for algorithm prediction.



(a) Permutation importance of nuclear features in classification of cell type



(b) Nuclear classifier performance with successive Recursive Feature Elimination

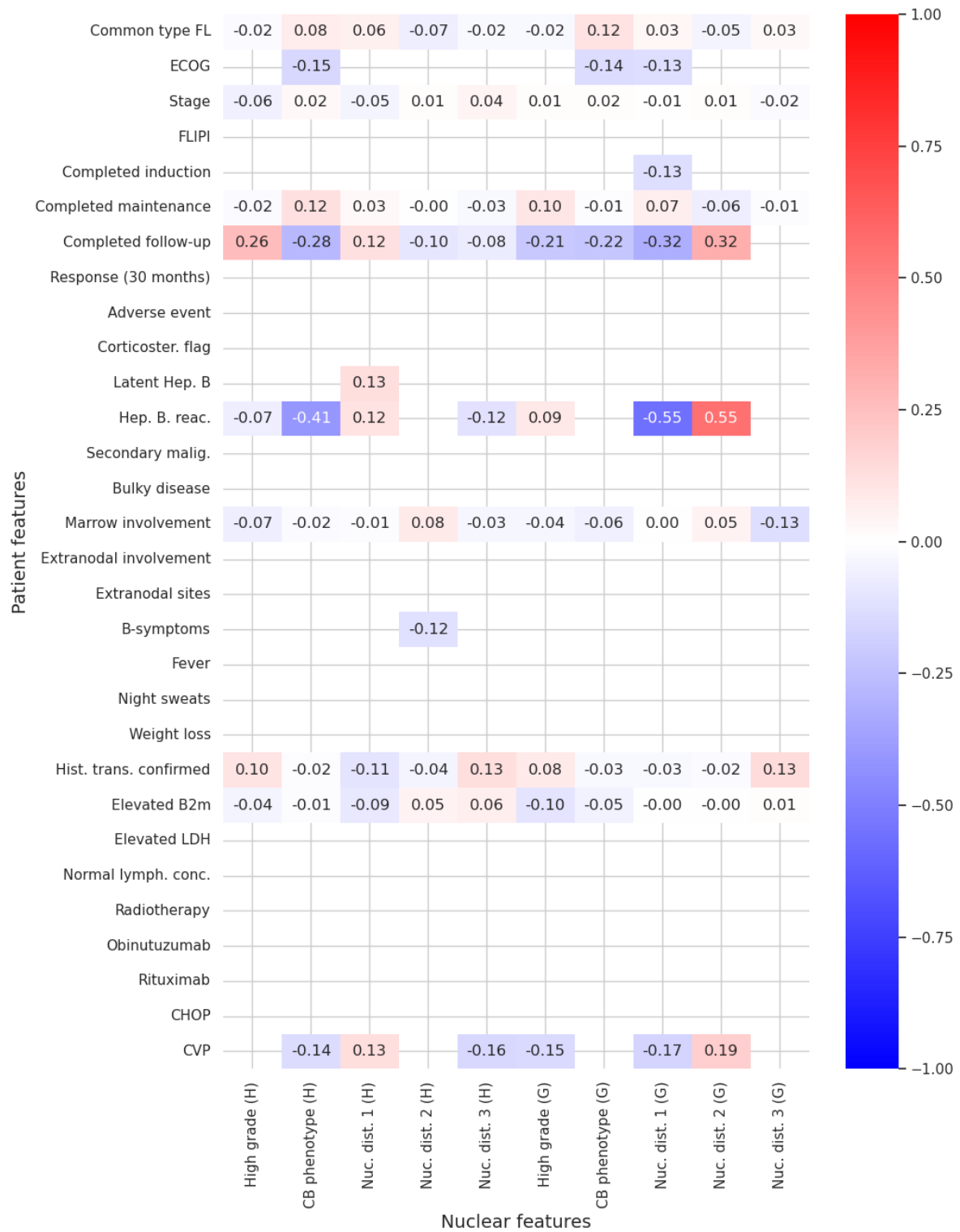
Figure A.3: Results of nuclear feature selection approaches in predictions of cell type using nuclear features. Details of the development of the algorithm are provided in Chapter 2, Section 3.8, Machine Learning approach.

a) Permutation importance of nuclear features, measured using decrease in accuracy upon random shuffling (permutation) of feature values. Used to gauge utility of features in the classification task.

b) Random Forest Classifier performance (macro-averaged F1-score) with successive rounds of recursive feature elimination (RFE), used to identify optimal feature number and gauge the most informative features for the classification task. Solid line represents mean value. Shaded area indicates 95% confidence intervals of performance, determined using 5-fold cross validation.



(a) HMRN FL dataset



(b) GALLIUM FL dataset

Figure A.4: Pearson correlation between patient features and binary nuclear features (membership resulting in a value of 1). Correlation values with p-value greater than 0.05 between groups masked. P-values calculated in a one vs. rest configuration where multiple classes existed (AD clusters).

Abbreviations:
ROIs = regions of interest; conc.= concentration, in blood; LDH = Lactate dehydroge-
nase; Hist. Trans. = Histologic transformation; DLBCL = Diffuse Large B-cell Lym-
phoma; HM = Haematological Malignancy; NHL = Non-Hodgkin Lymphoma; CB =
Centroblast; AD = Anderson-Darling cluster (also referred to as nuclear distribution
cluster).
* = Effect observed following Bonferroni multiple testing correction.

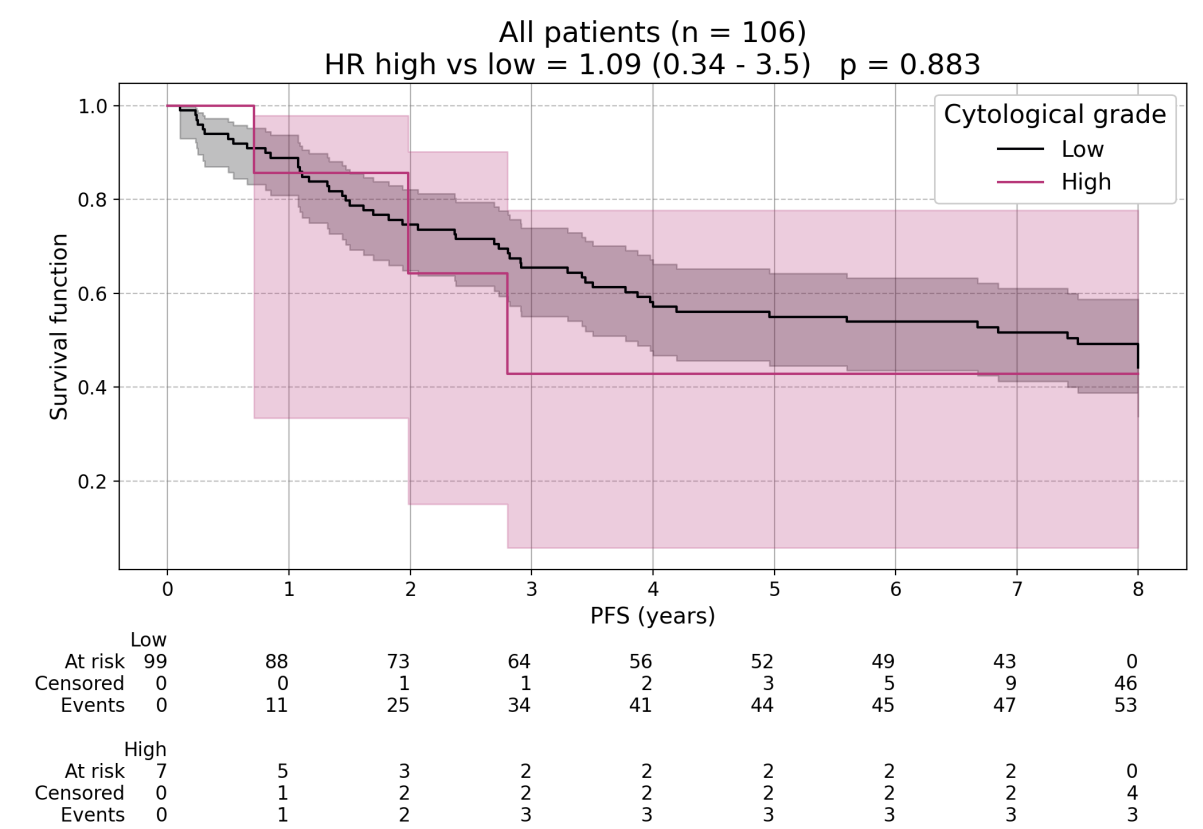
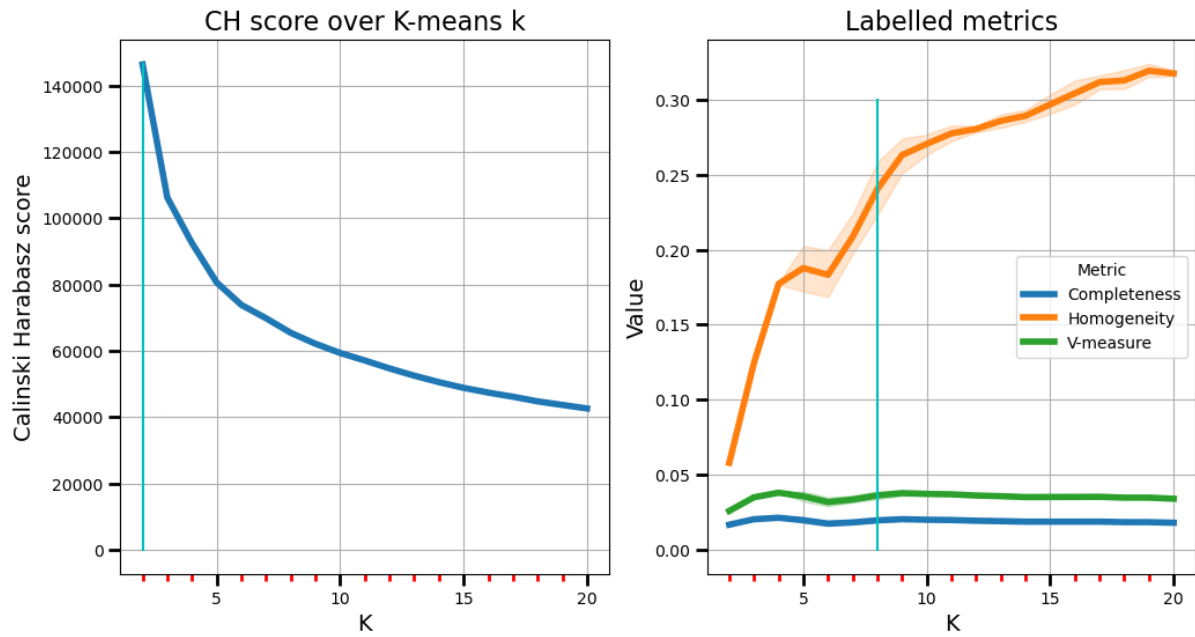
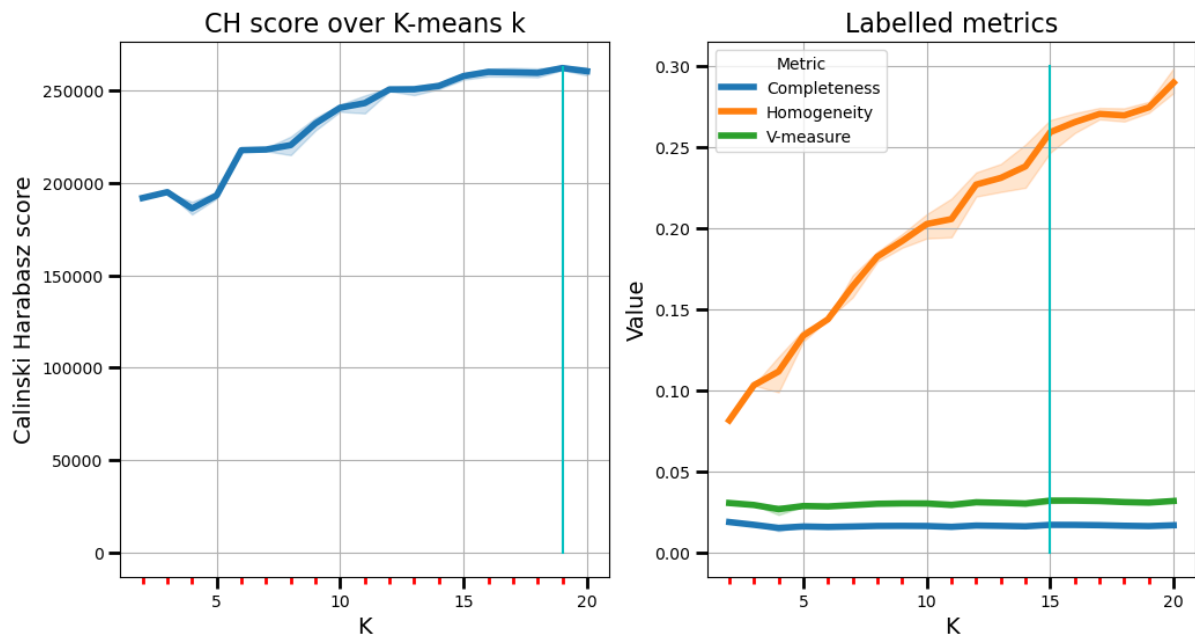


Figure A.5: Kaplan-Meier estimator fitted to HMRN dataset progression-free survival (PFS) events with up to 8 years of follow up. Estimators not displayed for individual treatment partitions due to low event number.

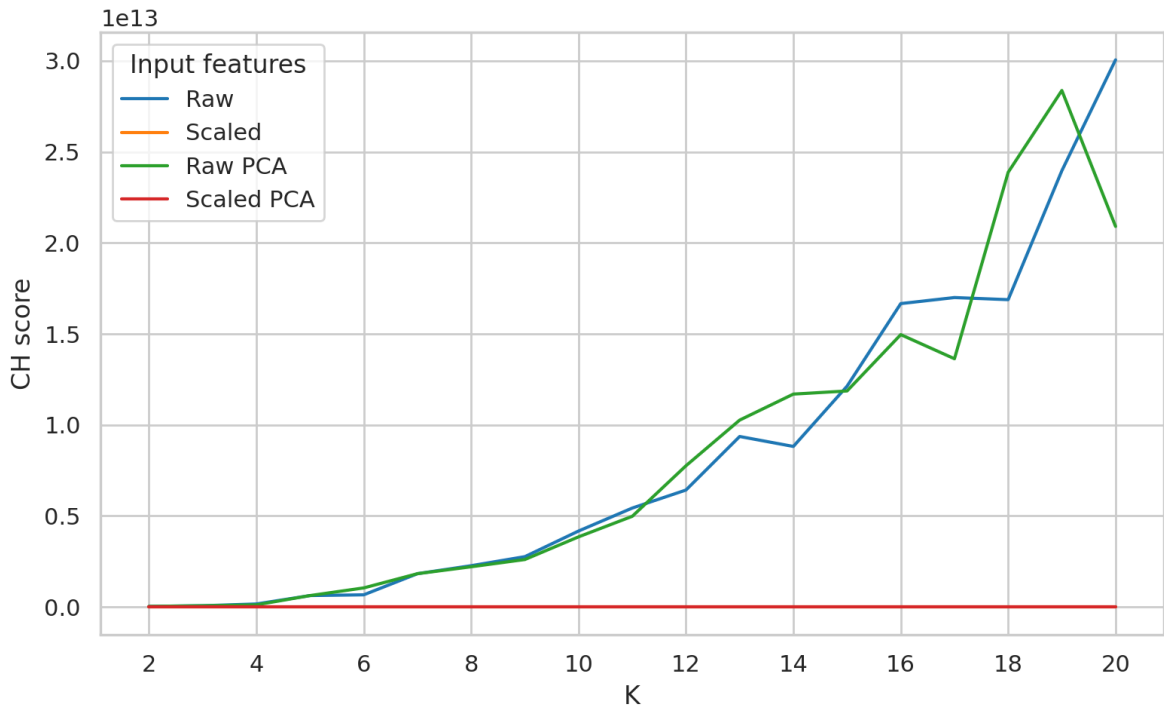
3 Chapter 3: Nuclear clusters



(a) Cell clustering on nuclear features



(b) Cell clustering on UMAP embedded nuclear features



(c) Clustering of FL/DLBCL nuclei on 66 nuclear features

Figure A.6: Assessment of clustering over a range of cluster numbers (K). Clustering models fitted to linearly scaled nuclear features (a), 2-dimensional UMAP embedded features (b) and 66 nuclear features transformed using a range of pre-processing methods (c). Metric values for K-means clustering models fitted on 177,000 unlabelled nuclei from a range of Non-Hodgkin Lymphomas (a and b) and 4.6 million nuclei sampled from naive B-cell regions of FL and DLBCL slides in the HMRN dataset (c).

Homogeneity, completeness and V-measure (a and b, right) calculated on 67,000 nuclei labelled by a pathologist (A.B.) as centroblast, non-centroblast or non-nuclear. Calinski-Harabasz (CH) score calculated using input feature values and predicted cluster assignment (a and b, left, and c).

Input feature types (c) key: Raw = unscaled nuclear features; Scaled = linearly scaled features (to within range of 0 and 1); Raw PCA = unscaled features reduced to 2 principal components; Scaled PCA = linearly scaled features reduced to 2 principal components. Solid lines indicate mean performance metric values, over 10 repeats of model fitting. Shaded areas indicate 95% confidence intervals. Red, vertical lines in plots a and b indicate the cluster number at which Calinski-Harabasz score (left plots) or V-measure (right plots) were maximised.

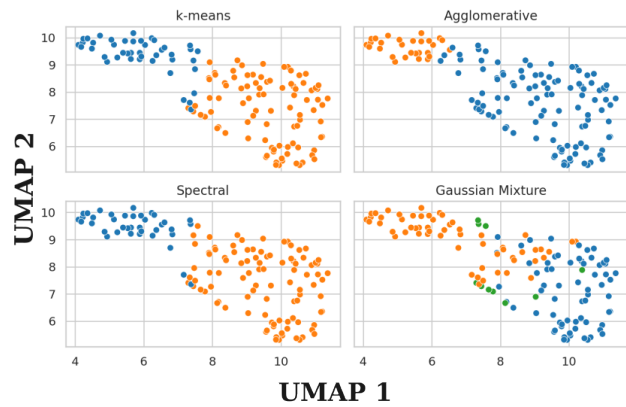
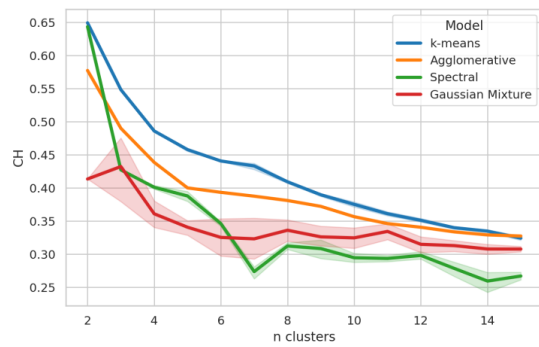
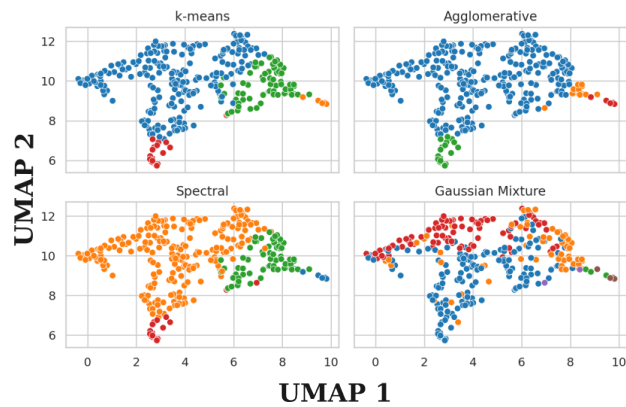
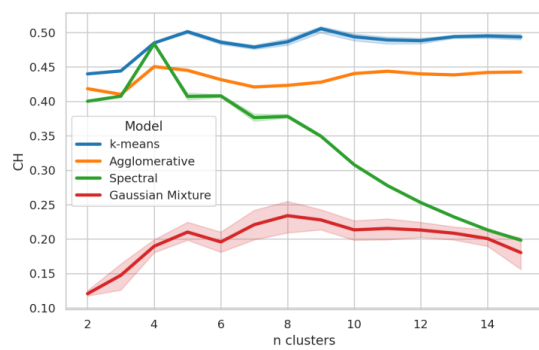
FL**DLBCL**

Figure A.7: Quantitative results of patient clustering in FL(top) and DLBCL (bottom), using a range of clustering methods.

Left) Calinski-Harabasz score (CH; scaled to dataset size) over cluster number. Solid lines indicate mean performance metric values, over 10 repeats of model fitting. Shaded areas indicate 95% confidence intervals.

Right) Optimal clustering of slides, as measured by CH score for each method in a 2D UMAP space. Arbitrary cluster membership denoted by colour.

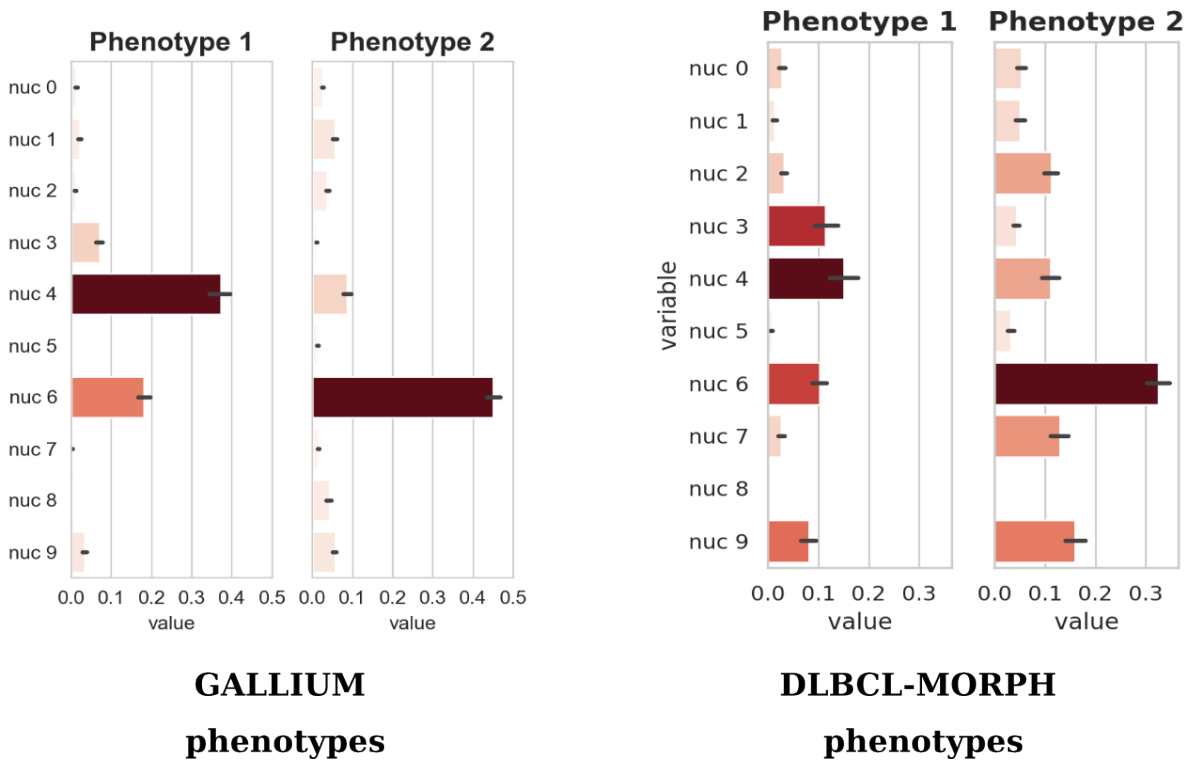
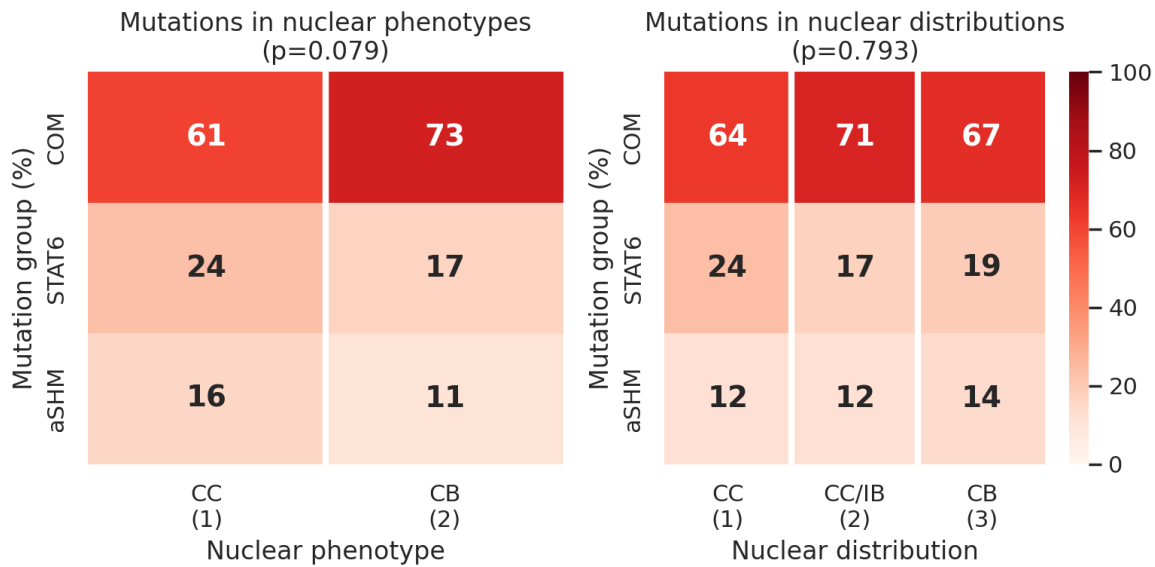
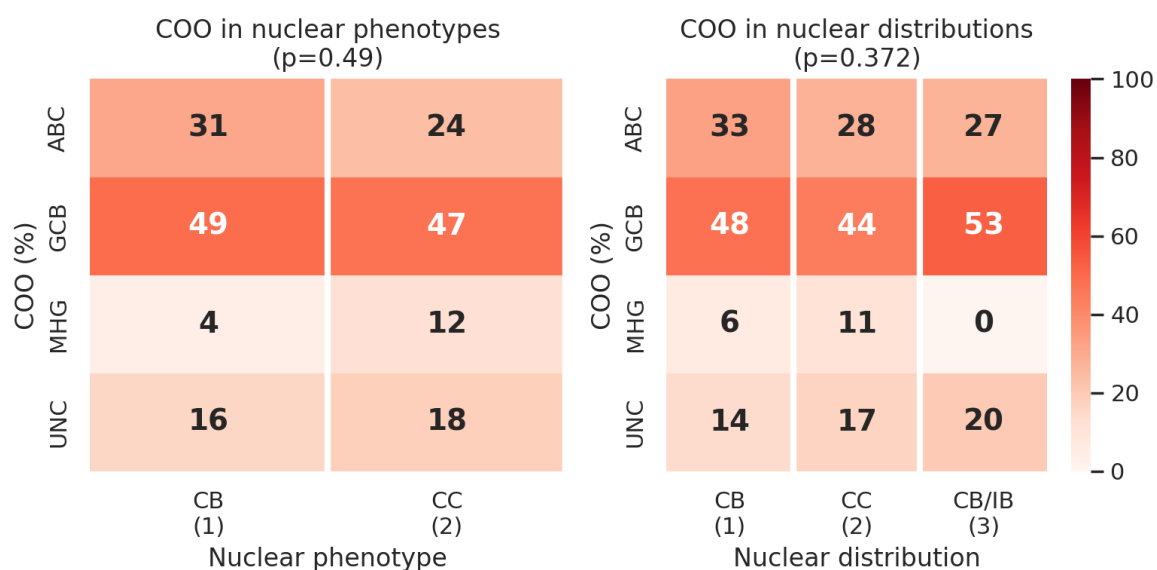


Figure A.8: Cellular composition of nuclear phenotypes trained on external datasets. Bars represent mean values for patients belonging to each cluster, with 95% confidence intervals (errorbars). Values on the x-axis correspond to linearly-scaled values of cell proportions (scaled to minimum/maximum values of respective datasets). Y-axis values relate to predicted cell types, as determined using the pan non-Hodgkin classifier.



(a) Follicular Lymphoma (FL)



(b) Diffuse Large B-cell Lymphoma (DLBCL)

Figure A.9: Composition of nuclear groups with respect to published molecular groups in FL (mutation[46]) and DLBCL (cell-of-origin expression; COO[56, 144–146]). Values represent percentage of each nuclear cluster (x-axis), for samples with available molecular data. P-values were calculated using χ^2 tests on $n \times m$ contingency tables of frequency, where n = number of molecular groups and m = number of nuclear groups. No multiple testing correction was applied.

4 Chapter 4: Immunohistochemistry clusters

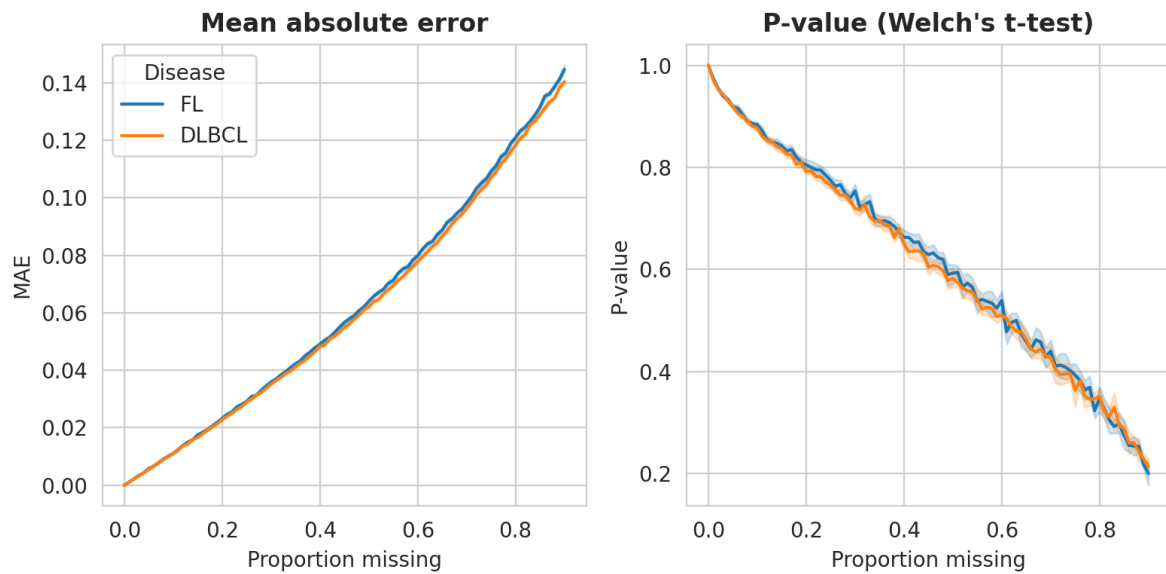


Figure A.10: Divergence of imputed data from original data as measured using Mean Absolute Error (MAE) and p (Welch's t-test). Patient samples for which complete IHC data was available were used. Proportions of data, as indicated on the x-axis were randomly masked and values of masked data were imputed (K-Nearest Neighbours model). MAE and p-values (mean over all feature columns) were calculated between imputed and original data over 10 repeats. Disease key: FL = Follicular Lymphoma; DLBCL = Diffuse Large B-cell Lymphoma.

Nuclear features

- area
- Eccentricity
- Orientation
- Solidity
- Euler number
- Maximum intensity (red channel)
- Mean intensity (red channel)
- Minimum intensity (red, green and blue channels)

- Hu moments (1st, 3rd, 4th, 5th and 7th)
- The following central mixed moments up to the 4th order:
 - 1st-2nd
 - 1st-4th
 - 2nd-1st
 - 2nd-2nd
 - 2nd-3rd
 - 2nd-4th
 - 3rd-2nd
 - 3rd-4th
 - 4th-1st
 - 4th-2nd
 - 4th-3rd

4.1 Evaluation metrics

Segmentation

Segmentation algorithms - those tasked with predicting the boundary of an object - were used to classify centroblast-rich regions (naive B-cell regions; section 3.7) and the interiors of nuclei (section 3.6), to guide feature extraction. Segmentation algorithms performance was evaluated using two complementary types of metric, as previously performed[103]. Firstly, metrics that compared ground truth and predicted segmentation area only, regardless of the number of distinct objects present or predicted - DICE score, the Aggregated Jaccard index (AJI) and Segmentation Quality (SQ)[200]. Secondly, metrics designed to describe quality of individual objects segmented - Ensemble DICE

(DICE2)[110], Aggregated Jaccard Index Plus (AJI+)[201], F_1 score of segmented objects, referred to as Detection or Recognition Quality (DQ)[200] and Panoptic Quality Score (PQ)[200].

A limitation of the first group of metrics in the proposed use case was that they did not distinguish between models on ability to segment individual nuclei. This was a core requirement as bias in distinction of nuclei could have presented a severe confounding variable in downstream classification of nuclei. Equations for all metrics used are presented below. For all metrics, higher values indicate better performance with a value of 1 indicating perfect segmentation performance and 0, completely imperfect performance.

DICE:

$$DICE(X, Y) = \frac{2 \times |X \cap Y|}{|X| + |Y|} \quad (A.1)$$

DICE is a pixel-wise comparison of a predicted segmentation X , with respect to ground truth segmentation Y . Where $|X \cap Y|$ represents the number of pixels within the intersection of predicted segmentation X and ground truth segmentation Y and $|X| + |Y|$ represents the sum of pixels in X and Y . The classification metric equivalent of DICE is F1-score, the harmonic mean of precision and sensitivity.

DICE 2:

$$\begin{aligned} &\text{Ground truth objects } Y_n; \ n = 1, \dots, p \\ &\text{Predicted segmentation objects } X_m; \ m = 1, \dots, q \\ DICE\ 2(X, Y) &= 2 \times \frac{\sum_{n, m, Y_n \cap X_m \neq \emptyset} |X_m \cap Y_n|}{\sum_{n, m, Y_n \cap X_m \neq \emptyset} (|X_m| + |Y_n|)} \end{aligned} \quad (A.2)$$

DICE 2 can be interpreted as an object-wise DICE score designed to penalise algorithms that segment a large number of small objects to fill a space. It is calculated between all distinct objects n in the ground truth segmentation Y and the corresponding subset of predicted objects m that overlap n .

$|Y_n \cap X_m|$ represents the number of pixels within the intersection of ground

truth object n and predicted object m . $|Y_m| + |X_n|$ represents the sum of the areas of m and n , in number of pixels.

Aggregated Jaccard Index:

$$AJI(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \quad (\text{A.3})$$

Aggregated Jaccard Index (AJI) is a pixel-wise comparison of a predicted segmentation X , with respect to ground truth segmentation Y otherwise known as Intersection-over-Union (IoU). Where $|X \cap Y|$ represents the number of pixels within the intersection of predicted segmentation X and ground truth segmentation Y and $|X \cup Y|$ represents the number of pixels within the union of X and Y .

Aggregated Jaccard Index Plus:

For ground truth objects Y_n ; $n = 1, \dots, p$

Predicted segmentation objects X_m ; $m = 1, \dots, q$

Unmatched Ground truth objects Y_{n^C} ; $n^C \notin Y_n$

Unmatched Predicted objects X_{m^C} ; $m^C \notin X_m$

$$AJI + (X, Y) = \frac{\sum_{n, m, Y_n \cap X_m \neq \emptyset} |X_m \cap Y_n|}{\sum_{n, m, Y_n \cap X_m \neq \emptyset} |X_m| \cup |Y_n| + \sum_{n^C} |Y_{n^C}| + \sum_{m^C} |X_{m^C}|} \quad (\text{A.4})$$

Aggregated Jaccard Index Plus (AJI+) can be interpreted as an object-wise AJI, aggregating AJI across overlapping pairs of ground truth and predicted objects. AJI+ penalises algorithms which predict more than one segmentation per ground truth object through addition of the areas of unpaired objects to the denominator. AJI+ is calculated across all distinct objects m in the predicted segmentation X . Objects m are paired to a maximum of 1 ground truth object n using maximum pairwise AJI value. $|X_n \cap Y_m|$ represents the number of pixels in the intersection of predicted object X_n and matched ground truth object Y_m . $|X_n \cup Y_m|$ represents the number of pixels within

the union of objects X_n and Y_m . The areas of unmatched ground truth and predicted objects Y_{n^c} and X_{m^c} , respectively, are added to the denominator to penalise prediction of too few or too many objects, with respect to number of objects in the ground truth.

Segmentation Quality:

$$SQ(X, Y) = \frac{\sum_{n=0}^{(n,m) \in TP} AJI_{n,m}}{|TP| + \alpha} \quad (\text{A.5})$$

Where true positive segmentations (TP) are defined as predicted objects n that have Aggregated Jaccard Index (AJI) values above 0.5 with a matched ground truth object m . The numerator is therefore the sum of the AJI between matched pairs of n and m . The denominator is the number of true positives $|TP|$. α is a small value (10^{-6}) to prevent zero division.

Detection Quality:

$$DQ(X, Y) = \frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \quad (\text{A.6})$$

Where true positive segmentations (TP) are defined as predicted objects in the predicted segmentation X that have Aggregated Jaccard Index (AJI) values above 0.5 with a matched ground truth object in the ground truth segmentation Y . False positives (FP) are defined as objects predicted in X with no corresponding objects to be matched against (no objects with $AJI \geq 0.5$) in Y . False Negatives FN are defined as objects in Y with no corresponding objects to be matched against (no objects with $AJI \geq 0.5$) in X .

Panoptic Quality:

$$PQ(X, Y) = SQ \times DQ \quad (\text{A.7})$$

Panoptic Quality (PQ) is the product of Segmentation Quality (SQ) and Detection Quality (DQ) for predicted segmentation X , with respect to ground truth segmentation Y .

Classification

Classification algorithms - those tasked with predicting categories of objects - were used to predict cell type (section 3.8). Cell counts were used to assign cytological grades to patients (section 3.9). These tasks were evaluated using Accuracy, Precision, Sensitivity, Specificity, F1 and Area under the Receiver operating curve (AUC-ROC).

Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (A.8)$$

Accuracy is the proportion of all objects to be correctly classified. TP = True positive. TN = True negative. FP = False positive. FN = False negative.

Precision:

$$Precision = \frac{TP}{TP + FP} \quad (A.9)$$

Precision is the proportion of all objects predicted to belong to a class that are correctly classified. TP = True positive. FP = False positive.

Sensitivity:

$$Sensitivity = \frac{TP}{TP + FN} \quad (A.10)$$

Sensitivity (also known as Recall) is the proportion of all objects of a class to be correctly classified. TP = True positive. FN = False negative.

Specificity:

$$Specificity = \frac{TN}{TN + FP} \quad (A.11)$$

Specificity is the proportion of all objects that do not belong to a class that are correctly classified as not belonging to the class. TN = True negative.

FP = False positive.

F1:

$$F1 = \frac{2 \times Precision \times Sensitivity}{Precision + Sensitivity} = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (A.12)$$

F1 score is the harmonic mean of Precision and Sensitivity. The equivalent segmentation metric is DICE score. TP = True positive. FP = False positive.

FN = False negative.

Area under the Receiver Operating Curve:

$$AUC-ROC = \int_0^1 Sensitivity \ d(1 - Specificity) \quad (A.13)$$

Area under the receiver operating curve (AUC-ROC) is defined as the area contained under a curve of *Sensitivity* over *False Positive Rate* (FPR; equivalent to $(1 - Specificity)$), formed by all classifications.

Agreement

Agreement is defined as level of similarity in exact values, differing to correlation which instead compares *change* in value over a range of conditions. This was deemed a valuable comparison in cytological grade, where exact grade assigned (especially grade 3) can be used in informing treatment choice[24]. Agreement between pathologist and machine estimates of continuous features (such as number of cells and cytological grade) were assessed using the coefficient of determination (R^2) and Bland-Altman (B&A) plots.

R^2 : R^2 provided details on how close predicted values were to pathologist estimates, weighted by variation in pathologist estimates.

The metric value was calculated as follows:

$$SS_{residual} = \sum_{i=1}^i (y_i - h_i)^2$$

$$SS_{true} = \sum_{i=1}^i (y_i - \bar{y})^2 \quad (\text{A.14})$$

$$R^2 = 1 - \frac{SS_{residual}}{SS_{true}}$$

R^2 is a measure of the difference between ground truth measurements, y , and predicted values, h , over all matched datapoints i , weighted by variation observed in ground truth. $SS_{residual}$ = sum of the squares of residuals (differences) between true and predicted values. SS_{true} = sum of the square differences between true measurements and the mean of true values, $\bar{y}$. Since it is possible for the value of $SS_{residual}$ to be higher than SS_{true} , R^2 values can be in the interval $(-\infty, 1]$. R^2 is not defined if all values of y_i equal $\bar{y}$.

Bland-Altman plots

In B&A plots, residuals of measurements (pathologist estimate - machine prediction) were plotted on the Y-axis against pathologist estimates on the X-axis. These plots provided information on[202]:

- Systematic differences between pathologist and machine estimates; for instance, a general under-estimation of grade by 1 unit, compared to pathologists. Horizontal red, dotted lines were plotted at mean residual values, used to indicate systematic differences in estimates
- Consistency in agreement over a range of measurement values; for instance, demonstrating consistent agreement in cell counts up to 50 cells but increased disagreement at higher values

- 95% confidence intervals of ground truth (pathologist) estimates were displayed using grey, dotted horizontal lines on plots. These indicated acceptable limits for agreement, with proportion of machine predictions falling outside of these limits used as a measure of agreement quality (higher values demonstrating better agreement).

Clustering

Clustering performance was evaluated using complementary strategies which operated with respect to:

1. Underlying data structure - Calinski-Harabasz index (CH score)
2. Ground truth labels, assigned by a pathologist - Homogeneity, Completeness and their harmonic mean, V-measure[150]

For all metrics, higher values indicated better quality of fit. While homogeneity, completeness and V-measure values were in the range 0-1 and could be directly compared between experiments, CH score was not bounded and so could only be interpreted through comparison with other clustering schema on the same data.

In the implementation of Homogeneity and Completeness used[153], only valid combinations of cluster assignment and ground truth object labels α_{yk} are considered - those with at least one member.

Equations for metrics:

Calinski-Harabasz index:

For data points: $\{x_1, x_2, \dots, x_n\}$; their respective assignments to m clusters: $\{k_1, k_2, \dots, k_m\}$; centroids of clusters: $\{c_1, c_2, \dots, c_m\}$ and the global centroid $\mathbf{c}$:

$$CH = \frac{\frac{1}{m-1} \sum_{i=1}^m n_i \|c_i - \mathbf{c}\|^2}{\frac{1}{n-m} \sum_{i=1}^m \sum_{x \in k_i} \|x - c_i\|^2} \quad (\text{A.15})$$

The numerator represents the mean square Euclidean distance between the cluster centroid and global centroid, $\mathbf{c}$, with respect to number of clusters m and weighted by the number of objects assigned to each cluster n . This is also known as ‘Between-Cluster Sum of Squares’ (BCSS).

The denominator represents the mean square Euclidean distance between data points x and their respective cluster centroid k , with respect to number of data points. This is also known as ‘Within-Cluster Sum of Squares’ (WCSS). The CH score is a measure of cluster dispersion (BCSS), with respect to dispersion within clusters (WCSS).

Homogeneity: The total number of data points is represented by N , number of objects in cluster k , N_k , and number of objects of class y , N_y .

For ground truth labels belonging to p classes: $\{y_1, y_2, \dots, y_p\}$;

assignment of data points to m clusters: $\{k_1, k_2, \dots, k_m\}$

and total number of data points of each class y belonging to cluster k , α_{yk} : $\{\alpha_{1,1}, \alpha_{1,2}, \dots, \alpha_{p,m}\}$:

$$\text{Homogeneity} = 1 - \frac{H(Y|K)}{H(Y)}$$

$$H(Y|K) = - \sum_{k=1}^m \sum_{y=1}^p \frac{a_{yk}}{N} \log \frac{a_{yk}}{N_k} \quad (\text{A.16})$$

$$H(Y) = - \sum_{y=1}^p \frac{N_y}{p} \log \frac{N_y}{p}$$

Homogeneity is a measure of cluster purity with respect to ground truth labels. The value of homogeneity approaches 1 as a greater proportion of data points assigned to cluster k belong to the same class y .

The numerator represents information encoded in ground truth labels Y , given

clustering K ($H(Y/K)$). It is a measure of the number of objects of class y belonging to cluster k , with respect to the total number of objects of class y , scaled using the total number of objects.

The denominator represents information encoded in Y ($H(Y)$) and has the effect of scaling ($H(Y/K)$) to account for imbalance in class labels. It is a measure of the total number of objects within each class y , scaled by the total number of objects.

Completeness:

$$Completeness = 1 - \frac{H(K|Y)}{H(K)}$$

$$H(K|Y) = - \sum_{y=1}^p \sum_{k=1}^m \frac{a_{yk}}{N} \log \frac{a_{yk}}{N_y} \quad (\text{A.17})$$

$$H(K) = - \sum_{k=1}^m \frac{N_k}{m} \log \frac{N_k}{m}$$

Completeness is a measure of data fragmentation. Completeness approaches 1 as a greater proportion of data points belonging to class y are assigned to the same cluster k .

The numerator represents information encoded in clustering K , given ground truth labels Y ($H(K/Y)$). It is a measure of the number of objects of class y belonging to cluster k , with respect to the total number of objects of class y , scaled using the total number of objects.

The denominator represents information encoded in clustering K ($H(K)$) and has the effect of scaling ($H(K/Y)$) to account for imbalance in cluster sizes. It is a measure of the total number of objects within each cluster k , scaled by the total number of objects.

V-measure:

Where β is a weight coefficient, leading to greater contribution from Completeness (C) as $\beta \rightarrow 0$, and greater contribution from Homogeneity (H) as $\beta \rightarrow \infty$. A default β value of 1 led to equal weighting between H and C.

$$V - measure = \frac{(1 + \beta) * H * C}{(\beta * H) + C}$$

$$\text{Default value of } \beta = 1 : \tag{A.18}$$

$$V - measure = \frac{2 * H * C}{H + C}$$

V-measure is the harmonic mean of Homogeneity and Completeness. Since these metrics oppose one another - an increase in cluster number often leads to a decrease in Homogeneity and increase in Completeness - V-measure can provide a convenient aggregation.