École Doctorale Biologie Santé Lille
Université de Lille

PhD THESIS
To obtain the degree of Doctor in Molecular and cellular aspects of biology from the University of Lille

Visualizing the molecular interaction network-level heterogeneity of malignant tumors by mass spectrometry imaging

PhD thesis submitted by LAURINE LAGACHE

Defended in Lille, the December the 19th, 2024

Under the supervision of Pr. Michel Salzet,
Co-supervision of Dr. Nawale Hajjaji

Defense committee :

President of the jury	Professor	Pierre Chaurand	Université de Montreal
Reporter	Research Director	Sarah Cianferani	Université de Strasbourg
Reporter	Research Director	Thomas Daubon	UMR5095 Université de Bordeaux
Examinator	Professor	Isabelle Fournier	U1192 Université de Lille
Co-supervisor	Doctor	Nawale Hajjaji	Centre Oscar Lambret de Lille
Thesis Director	Professor	Michel Salzet	U1192 Université de Lille

École Doctorale Biologie Santé Lille
Université de Lille

THESE DE DOCTORAT
En vue de l'obtention du grade de Docteur en Science de l'Université de Lille en aspects moléculaires et cellulaires de la biologie

Visualisation de l'hétérogénéité au niveau des réseaux d'interactions moléculaires des tumeurs malignes par imagerie par spectrométrie de masse

Thèse de doctorat soumise par LAURINE LAGACHE

Soutenue à Lille, le 19 Décembre 2024

Sous la direction du Pr. Michel Salzet,
Co-direction du Dr. Nawale Hajjaji

Composition du jury :

Président du jury	Professeur	Pierre Chaurand	Université de Montréal
Rapporteure	Directrice de recherche	Sarah Cianferani	Université de Strasbourg
Rapporteur	Directeur de recherche	Thomas Daubon	UMR5095 Université de Bordeaux
Examinateure	Professeure	Isabelle Fournier	U1192 Université de Lille
Co-encadrante	Docteure	Nawale Hajjaji	Centre Oscar Lambret de Lille
Directeur de thèse	Professeur	Michel Salzet	U1192 Université de Lille

“Le talent, ça n'existe pas.

Le talent, c'est d'avoir envie de faire quelque chose.

Je suis convaincu que ce qu'on appelle talent,

c'est principalement du courage, de l'envie et de la sueur.

Tout le reste, c'est de la sueur.”

Jacques Brel

Acknowledgments

À l'issue de ce travail, je souhaite exprimer ma profonde gratitude envers toutes les personnes qui ont contribué à sa réalisation. Leur soutien, leurs précieux conseils et leur accompagnement ont été essentiels pour mener à bien ce projet.

Je tiens tout d'abord à exprimer ma reconnaissance au Pr Michel Salzet, dont l'accompagnement en tant que directeur de thèse a été d'une aide précieuse tout au long de cette aventure singulière. Merci pour ton soutien constant et ta bienveillance, qui m'ont aidé à trouver ma place au sein du laboratoire. Je me souviendrai de nos nombreux échanges scientifiques et des conseils avisés qui ont grandement contribué à la réalisation de ce travail. Je remercie également le Pr Isabelle Fournier pour sa confiance et pour les bons moments que nous avons partagés. Vous m'avez tous deux offert des opportunités précieuses qui m'ont permis de grandir à la fois sur le plan personnel et professionnel. Je n'oublierai jamais l'impact que vous avez eu sur mon parcours, et j'espère pouvoir mettre à profit tout ce que j'ai appris dans mes futures expériences.

Je tiens à remercier ma co-directrice, le Dr Nawale Hajjaji, de m'avoir offert l'opportunité de travailler sur ce sujet si passionnant. Je vous suis reconnaissante pour la confiance que vous m'avez accordée, le temps que vous m'avez consacré, ainsi que pour le partage de vos connaissances qui ont enrichi ce travail. J'en profite pour exprimer ma gratitude aux patients qui, malgré les épreuves traversées, ont généreusement accepté de contribuer à des projets de recherche tels que celui-ci.

Je suis également reconnaissante envers tous les membres de mon comité de thèse pour le temps qu'ils ont consacré à cette évaluation. Je remercie tout particulièrement le Dr Thomas Daubon, le Dr Sarah Cianferani, le Pr Pierre Chaurand et le Pr Isabelle Fournier pour leur expertise et précieux commentaires constructifs.

Je tiens à remercier l'Université de Lille pour son soutien financier, ainsi que les sociétés savantes (SFSM, FPS et GDRmsi), qui m'ont permis de participer à de nombreux congrès. Ces expériences ont été essentielles pour enrichir ma recherche et élargir mes réseaux professionnels.

Un grand merci à Soulaimane. Cette aventure a commencé avec toi, alors que je n'étais que stagiaire. Depuis, tu as toujours veillé sur moi, que ce soit pour me soutenir, m'encourager ou me conseiller. Au début de ma thèse, tu m'as dit : « Tu vas voir, ça va passer très vite, ta thèse, tu la soutiens demain. », pour moi c'était hier. Tu sais toujours trouver les mots justes au moment opportun. Je te remercie également pour nos discussions, que ce soit à propos de tout ou de rien. Enfin, merci d'être devenu une personne sur qui je peux compter. J'espère que nous aurons l'occasion de collaborer à nouveau sur de beaux projets à l'avenir et de maintenir cette complicité.

Antonella et Tristan, je vous remercie pour votre gentillesse et votre bonne humeur. Votre aide et vos conseils ont été des éléments précieux pendant ces trois années. J'ai aussi eu la chance de vous découvrir sur un plan plus personnel et de partager de merveilleux moments en votre compagnie. J'espère que nous aurons encore l'occasion de vivre de belles expériences ensemble à l'avenir (avec un peu d'ananas, tout est toujours meilleur) !

Merci à Lucie, Etienne, Diala, Maheul, Marius, Sarah, Adel, Angèle, Jeanne, Chloé, Lucille, Paul, David, Léna, Dona , Blandine , Jean-Pascal et François pour toutes nos discussions. Une pensée spéciale pour Alice, Kamel, Lucas et Diego. J'ai été ravie de partager ces instants avec vous et je vous souhaite une pleine réussite dans tous vos projets futurs.

Je tiens à remercier mes nouveaux collègues enseignants, Marie, Julien, Maxence, Christophe et Franck pour vos conseils avisés et les moments passés en tant que binômes de TP.

À Yanis et Alexandre, merci pour votre esprit d'équipe et l'excellente ambiance de travail que nous avons su créer ensemble. Travailler avec vous a rendu chaque défi plus simple et constructif. Yanis, ta précieuse aide a été déterminante dans ce projet; il n'aurait pas eu autant d'impact sans toi. Je suis convaincue que tu iras loin. Alexandre, on peut travailler sur n'importe quoi, mais pas avec n'importe qui. Merci d'être la personne que tu es, c'est un vrai plaisir de collaborer avec toi.

Léa et Lydia, vous êtes les amies que je ne m'attendais pas à rencontrer, mais dont je suis tellement heureuse d'avoir croisé le chemin. Vous avez toujours été là, que ce soit dans les moments de joie ou dans les épreuves les plus difficiles. Chaque instant passé avec vous est inoubliable. Léa, ta générosité et ton grand cœur te mèneront loin, j'en suis convaincue. Lydia, je garderai toujours en mémoire ta chaleur et ton rire communicatif, qui apportent la bonne humeur partout où tu vas. J'espère de tout cœur que nous resterons en contact toutes les trois.

Je ne m'attendais pas à ce que les remerciements soient la partie la plus difficile à écrire, mais c'est en pensant à vous, ma famille, que l'émotion se fait le plus ressentir.

Maman, Papa, il m'est difficile de trouver les mots justes, mais je veux simplement vous dire merci. Merci pour tout ce que vous m'avez donné et pour tout ce que vous continuez de m'apporter, jour après jour. Votre amour et votre soutien m'ont en grande partie donné la force, le courage et l'ambition d'être la personne que je suis et de parvenir là où je suis aujourd'hui. Cette réussite est aussi la vôtre. Je suis fière d'être votre fille, et j'espère que vous êtes tout aussi fiers de moi.

Chère petite sœur. Nous avons pris le temps avant de vraiment bien nous entendre, mais, sache que tu fais partie intégrante de ma vie. Merci d'avoir mis sur ma route Quentin, qui est devenu pour moi à la fois un ami et un frère. Je suis reconnaissante de vous avoir tous les deux à mes côtés.

A mes grands-parents. Il y a tant de choses que j'aimerais vous dire. Merci pour votre soutien indéfectible, pour les gâteaux, les repas partagés et tous ces moments précieux qui me tiennent tant à cœur. Votre présence a toujours été synonyme de réconfort et de motivation dans ma vie. Je suis profondément reconnaissante de pouvoir vivre encore ces instants ensemble. Merci à la vie de nous offrir l'opportunité de créer encore de merveilleux souvenirs.

Parrain, tu as toujours été un modèle pour moi. J'espère avoir atteint, à ma manière, un peu de ton niveau. Je te suis profondément reconnaissante pour tous tes conseils et ton aide dans tous domaines. Un grand merci également à ma marraine pour nos discussions et pour m'aider à rattraper mes cheveux à chaque fois. C'est toujours un plaisir de passer du temps avec toi !

Antoine. Je pense que tu as vécus cette thèse tout autant que moi. Ce n'a pas toujours été facile (Ok, je n'ai pas toujours été facile) mais tu trouves toujours une manière de me faire sourire. Merci d'être mon pilier et ma force au quotidien, pour les rires et la tendresse qui rendent ma vie plus légère. Je n'ai pas besoin d'en dire plus, car tu sais déjà tout. Partout où nous sommes, ensemble, nous sommes chez nous, et c'est plus que suffisant.

Un grand merci à Marylène et Jean-Louis pour votre aide quotidienne, votre bienveillance et votre compréhension. Merci également pour tous les petits plats et l'esprit de famille que vous m'offrez, qui m'ont souvent remonté le moral. Je tiens aussi à remercier Juju, qui pense toujours à moi, même de loin.

A ma famille de cœur, Cloé, Théo et June. À vous, âmes exceptionnelles, je vous remercie d'avoir toujours été là et de nous faire partager ces moments de bonheur précieux. Vous êtes une source de réconfort, d'inspiration et de motivation constante pour moi. J'espère que nous vous apportons autant que ce que vous nous offrez. Merci d'avoir mis sur notre chemin la petite Juju, qui est un véritable rayon de soleil.

Je tiens également à remercier Sandrine et Damien, des amis qui me sont chers. Dès notre première rencontre, notre amitié a été tellement évidente. Je garde en mémoire nos discussions, nos séances de yoga et vos encouragements, qui me font toujours du bien.

Enfin, je ne pouvais pas terminer ces remerciements sans évoquer mon équipe de basket. Les filles, je vous remercie pour votre compréhension lors des moments difficiles, pour la bonne énergie que vous transmettez à chaque rencontre, et pour me faire vibrer chaque semaine.

Curriculum Vitae

Formations

Scientific Baccalaureate – with honors. Margueritte de Flandres High School in Gondecourt. Obtained in July 2015.

Preparatory Class for Biology, Chemistry, Physics, and Earth Sciences (BCPST). Chatelet High School in Douai. September 2015 – July 2017.

Bachelor's Degree in Biology, specializing in Biochemistry. University of Lille, Science Campus, Villeneuve d'Ascq. September 2017 – July 2019.

Master's in Chemistry and Life Sciences, specializing in Bioanalytics – with honors. University of Lille, Science Campus, Villeneuve d'Ascq. September 2019 – July 2021.

Ph.D. in Health Biology. University of Lille, Science Campus, Villeneuve d'Ascq. October 2021 – October 2024

Experiences

Second-year internship (L2) at the Biology Department Laboratory of Lille, specializing in glycobiology. University of Lille, Science Campus, Villeneuve d'Ascq. Duration: 1 month. Deciphering biosynthesis mechanisms of O-acetylated GD2 in breast cancer.

Third-year internship (L3) at the Pasteur Institute of Lille, Inserm U1167. Duration: 3 months. Involvement of AKT, FOXO3A, and PGC1α proteins in the regulation of SOD2 acetylation in the heart.

Master's internship (M2) at the PRISM Laboratory, Inserm U1192. University of Lille, Science Campus, Villeneuve d'Ascq. Duration: 6 months. Identification of protein markers in breast cancer.

Scientific contributions

Publications

Under review

Roussel Lucas, Zirem, Yanis, **Lagache Laurine**, Ledoux Lea, Meresse Bertand, Delbecke Marie and Leblanc Eric, Yagnik Gargey, Lim Marc J, Rothschild Kenneth J, Robin Yves-Marie, Pasquesoone Camille, Lemaire Anne-Sophie, Bertin Delphine, Narducci Fabrice, Hudry Dephin, Salzet Michel and Fournier Isabelle. Spidermass and Machine Learning-Based Lipids Immunoscoring Forwards Real Time Ovarian Cancer Diagnosis and Prognosis in Surgery. Available at SSRN: https://ssrn.com/abstract=4979013 or http://dx.doi.org/10.2139/ssrn.4979013. (Revisions submitted on October 3rd 2024 in iScience).

Yanis Zirem, Léa Ledoux, Nina Ogrinc, **Laurine Lagache**, Roland Bourette, Chann Lagadec, Paul Chaillou, Michel Salzet and Isabelle Fournier. Development of Molecular Digital Twins Based on Ambient Ionization Mass Spectrometry Imaging for Application in Cancer Surgery. (Revisions submitted on September 27th 2024 in npj Digital Medicine)

Accepted

Laurine Lagache, Yanis Zirem, Émilie Le Rhun, Isabelle Fournier and Michel Salzet. Heterogeneity Assessment and Protein Pathway Prediction via Spatial Lipidomic and Proteomic Correlation: Advancing Dry Proteomics concept for Human Glioblastoma Prognosis. (Accepted on December 4th 2024 in MCP journal).

Daniel Simon, Gabriel Stefan Horkovics-Kovats, Yuchen Xiang, Ronan Battle, Yu Wang, Julia Abda, Dimitris Papanastasiou, Stefania Maneta, Hui-Yu Ho, Haixing Wang, Richard Schäffer, Tamas Karancsi, Anna Mroz, Istvan Pap, **Laurine Lagache**, Julia Balog, Isabelle Fournier, Robert Murray, Josephine Bunch and Zoltan Takats. Enabling Cellular Resolution Molecular Pathology for Surgical Interventions Using Laser Desorption – Rapid Evaporative Ionization Mass spectrometry. ChemRxiv. 2024; doi:10.26434/chemrxiv-2023-p2g9h-v3.

In preparation

Alexandre Goossen, **Laurine Lagache**, Christophe Biot, Nawale Hajjaji, Cédric Lion, Michel Salzet and Isabelle Fournier. Click-&-Detect: Enhancing Multiplex IHC with MALDI MSI Through Innovative Bioorthogonal Chemistry, digging into breast cancer microenvironment.

Laurine Lagache, Yanis Zirem, Michel Salzet and Nawale Hajjaji. Organoids for luminal breast cancer therapy guidance including molecular heterogeneity.

Laurine Lagache, Yanis Zirem, Michel Salzet and Nawale Hajjaji. 4D longitudinal proteomics tracking of breast cancer heterogeneity community response to therapeutics.

Grants

Financial aid for a European conference of 600 euros from FPS. **ProteoAix 2023** – Aix en Provence, France. June 20-23, 2023.

Financial aid for an international conference of 500 euros from GDR-MSI. **IMSIS 2023** – Montreal, Canada. October 23-25, 2023.

Financial aid for an international conference of 800 euros from SFSM. **ASMS 2024** – Anaheim, USA. June 2-6, 2024.

Oral communications

Laurine Lagache, Yanis Zirem, Alexandre Goossen, Nawale Hajjaji, Zoltan Takats, Isabelle Fournier, Michel Salzet. **GDR-MSI workshop 2023 – Paris, France. May 9 - 12, 2023**. Development of an optimized diagnostic method for tumor heterogeneity in cancer.

Laurine Lagache, Yanis Zirem, Alexandre Goossen, Nawale Hajjaji, Zoltan Takats, Isabelle Fournier, Michel Salzet. **GDR-MSI 2023 – Lille, France. June 12 - 13, 2023**. New method machine process to unravel tissue biomarkers and validate it by multiplex MALDI IHC.

Laurine Lagache, Yanis Zirem, Alexandre Goossen, Nawale Hajjaji, Zoltan Takats, Isabelle Fournier, Michel Salzet**. ProteoAix 2023 – Aix en Provence, France. June 20 - 23, 2023**. New method machine process to unravel tissue biomarkers and validate it by multiplex MALDI IHC.

Laurine Lagache, Yanis Zirem, Alexandre Goossen, Nawale Hajjaji, Zoltan Takats, Isabelle Fournier, Michel Salzet. **JFSM 2023 – Marseille, France. September 5 - 8, 2023**. New method machine process to unravel tissue biomarkers and validate it by multiplex MALDI IHC.

Laurine Lagache, Yanis Zirem, Alexandre Goossen, Nawale Hajjaji, Zoltan Takats, Isabelle Fournier, Michel Salzet. **IMSIS 2023** – **Montreal, Canada. October 23 - 25, 2023**. New method machine process to unravel tissue biomarkers and validate it by multiplex MALDI IHC.

Laurine Lagache, Yanis Zirem, Alexandre Goossen, Nawale Hajjaji, Zoltan Takats, Isabelle Fournier, Michel Salzet. **JAV 2023 – Lille, France. September 5, 2023**. Visualizing the molecular interaction network-level heterogeneity of malignant tumors by mass spectrometric imaging.

Laurine Lagache, Yanis Zirem, Isabelle Fournier, Michel Salzet. **ASMS 2024 – Anaheim, USA. June 2-6, 2024.** Spatial multi-omics guided by SVD *k*-means ++ clustering and statistical estimation of heterogeneity : Towards dry proteomic guided by lipids MALDI MSI.

Laurine Lagache, Yanis Zirem, Emilie Le Rhun, Isabelle Fournier, Michel Salzet**. GDR-MSI 2024 – Strasbourg, France. June 24-25, 2024.** Predicting protein pathways via spatial lipidomic and proteomic correlation: toward Dry Proteomics in human Glioblastoma.

Posters

Laurine Lagache, Nawale Hajjaji, Delphine Bertin, Isabelle Fournier, Michel Salzet, Zoltan Takats. **Strategic Day of the SFR-TSM – Lille, France. March 29, 2022**. Analysis of the spatial heterogeneity of the alternative proteome in breast cancer.

Laurine Lagache, Nawale Hajjaji, Delphine Bertin, Isabelle Fournier, Michel Salzet, Zoltan Takats. **Analytics 2022 – Nantes, France. October 5 - 8, 2022**. Visualizing the molecular heterogeneity of breast cancer by mass spectrometry imaging for therapeutics.

Laurine Lagache, Yanis Zirem, Alexandre Goossen, Nawale Hajjaji, Zoltan Takats, Isabelle Fournier, Michel Salzet**. Doctoriales 2023 – Belle Dune, France. June 5 - 9, 2023**. Visualizing the molecular interaction network-level heterogeneity of malignant tumors by mass spectrometric imaging.

Laurine Lagache, Nawale Hajjaji, Delphine Bertin, Isabelle Fournier, Michel Salzet, Zoltan Takats. **Strategic Day of the SFR-TSM – Lille, France. December 5, 2023**. Analysis of the spatial heterogeneity proteome in a temporal point of view in breast cancer case.

Laurine Lagache, Yanis Zirem, Alexandre Goossen, Nawale Hajjaji, Zoltan Takats, Isabelle Fournier, Michel Salzet**. Euron 2024 – Lille, France. February, 2024.** Spatial multi-omics guided by SVD *k*-means ++ clustering and statistical estimation of heterogeneity: RB sections as playground.

Laurine Lagache, Yanis Zirem, Emilie Le Rhun, Isabelle Fournier, Michel Salzet**. Canceropole 2024 – Deauville, France. May 2024.** Towards dry proteomic guided by lipid MALDI MSI.

Laurine Lagache, Yanis Zirem, Emilie Le Rhun, Isabelle Fournier, Michel Salzet**. SMAP 2024 – Lille, France. September 2024.** Predicting protein pathways via spatial lipidomic and proteomic correlation: toward Dry Proteomics in human Glioblastoma.

Summary

Figure List

Tables List

Abbreviations

2D: Two-dimensional

3D: Three-dimensional

9-AA: 9-aminoacridine

µm: micrometer

ACN: Acetonitrile

AI: Artificial Intelligence

AIMS: Ambient Ionisation Mass Spectrometry

ANOVA: Analysis of Variance

AR: Androgen Receptor

BC: Breast Cancer

CCC: Clear cell carcinoma

CD8: Lymphocyte TCD8+

CD4: Lymphocyte TCD4+

CD68: Macrosialin protein in the lysosome of macrophages

CHIP: Chemical Inkjet Printer

CI: Chemical Impact

Ck: Cytokeratin

CPDT: Cancer patient digital twins

DAN: 1,5-diaminonaphthalene

DCIS: Ductal Carcinoma in situ

DHB: 2,5-Dihydroxybenzoic acid

DIA: Data Independent Analysis

DNA: Desoxyribose Nucleic Acid

DT: Digital Twins

ECM: Extracellular matrix

EGFR: Epithelial Growth Factor Receptor

EI: Electronic Impact

ELISA : Enzyme-Linked Immunosorbent Assays

EMT: Epithelial-mesenchymal transition

ER : Estrogen receptor

ESI: Electro Spray Ionization

EtOH: Ethanol

FA: Formic acid

FAB: Fast Atom Bombardment

FD: Field Desorption

FDR: False Discovery Rate

FF: Fresh Frozen

FFPE: Formalin Fixed Paraffin Embedded

FI: Flow Injection

FTE: Fallopian tube epithelium

GL: Granular Layer

GM: Grey Matter

GO: Gene Ontology

HCCA: 4-hydroxy-α-cyanocinnamic acid

HER2: Human Epidermal Growth Factor Receptor-2

HGSOC: High Grade Serous Ovarian Carcinoma

HPS : Hématoxyline Phloxine Safran

HR: Hormonal Receptor

IC50: Half maximal inhibitory concentration

IDC: Invasive Ductal Carcinoma

IHC: Immunohistochemistry

ILC: Invasive Lobular Carcinoma

IQR: Interquartile range

IR: Infra-Red

ITO: Indium Tin Oxide

Ki67: Proliferation marker protein Ki-67

LC: Liquid Chromatography

LCIS: Lobular Carcinoma in situ

LCM: Laser Capture Microdissection

LDA: Linear discriminant analysis

LDI: Laser Desorption Ionization

LESA: Liquid Extraction Surface Analysis

LFQ: Label-Free Quantification

M1: Macrophages M1-like phenotype

M2: Macrophages M2-like phenotype

m/z: weight on charge ratio

MALDI: Matrix Assisted Laser Desorption Ionization

MC: Mucinous ovarian cancer

MET: Mesenchymal-epithelial transition

ML: Molecular Layer

MS: Mass Spectrometry

MS/MS: Tandem Mass spectrometry

MSI: Mass Spectrometry Imaging

NEDC: N- (1-naphthyl) ethylenediamine dihydrochloride

NK: Natural killers

NNMF: Non-Negative Matrix Factorization

OC: Ovarian cancer

OPO: Optical Parametric Oscillator

OS: Overall survival

nLC-MS/MS : nano Liquid Chromatography / Mass Spectrometry

PA: Phosphatidic Acid

PC: Phosphatidyl choline

PCA: Principal Component Analysis

PCR: Polymerase Chain Reaction

PDO: Patient-Derived Organoid

PDX: Patient-Derived Xenograft

PE: Phosphatidyl ethanolamine

PI: Phosphatidyl inositol

PR : Progesterone Receptor

PS: Phosphatidyl serine

PTMs: Post-Translational Modifications

RA: Androgen Receptor

RB: Rat brain

RMS: Root Mean Square

ROI: Region of interest

SA: Sinapinis Acid

SA-ani: Sinapinic Acid – aniline

SBL: Serous borderline carcinoma

SBR: Scraff Bloom and Richardson

SOPs: Standard Operating Procedures

SUMO: Small Ubiquitin like Modifier

SVD: Singular Value Decomposition

t-SNE: t-distributed Stochastic Neighbor Embedding

TG: Triglyceride

TIC: Total Ion Current

TNM: Tumor, Nodes, Metastasis

TOF: Time Of Flight

UV: Ultra Violet

WM: White Matter

General Introduction

Cancer is the second leading cause of death globally. In 2020, there were approximately 10 million cancer-related deaths and 19 million new cancer cases reported worldwide (Globocan 2020). Among these, breast cancer is the most prevalent in women, with 2.26 million cases and nearly 685,000 deaths in 2020. This underscores the critical importance of directing research efforts toward breast cancer.

Breast cancer arises from the uncontrolled proliferation of abnormal cells, forming tumors within the lobules or milk ducts of the mammary gland. This multi-step process involves successive genetic mutations and interactions with environmental factors, leading to malignant transformation and phenotypic alterations of the cells.

Diagnosis currently relies on the examination of biopsy samples through histopathological analysis. This process identifies the morphological and molecular characteristics of the cells to confirm the presence of cancer and determine its subtype. Breast cancers are classified based on the expression of hormone receptors, oncoproteins, epidermal growth factor receptors, cytokeratins, and cell proliferation markers. This information, along with clinical characteristics such as patient age, medical history, environmental factors, tumor size, and disease stage, is crucial for developing an effective therapeutic strategy.

Despite comprehensive treatments, 30% of breast cancer patients experience local recurrence or distant metastases. This phenomenon can be attributed to molecular heterogeneity within the tumor, which results in varying sensitivity to treatments. Recent advances in genomic sequencing have unveiled a higher degree of heterogeneity than previously anticipated. Numerous mutations within a single cell lead to subsequent clonal expansions, influenced by factors such as the tumor microenvironment and the exposome, including the effects of treatment. Additionally, non-genetic mechanisms contribute to tumor heterogeneity. Transcriptional, translational, and metabolic adaptations can result in the emergence of molecular subpopulations of clonal cancer cells. Consequently, a patient's breast tumor consists of a complex mosaic of genetically related clones, each unique to the individual.

This heterogeneity poses a significant challenge in developing effective cancer treatments, as each tumor comprises various cell phenotypes, each with its specific sensitivity or resistance to treatments. This diversity partly explains the occurrence of recurrences. Conventional treatments are typically based on genetic markers derived from bulk analyses, which do not account for the molecular heterogeneity within the tumor. As a result, these treatments may be effective for some molecular subpopulations but not for others.

Objectives of the Research Project

To address this issue, it is essential to refine tumor analysis by characterizing their heterogeneity to identify tailored markers and actionable targets at the subpopulation level. This study aims to enhance the analysis of tumor functional heterogeneity by integrating proteome analysis using mass spectrometry to better understand molecular heterogeneity. Improved understanding of tumors can lead to better predictions of cancer progression and recurrence risks through biomarkers, ultimately providing more personalized and effective treatments for patients. To arise this goal, the project was structured around four main objectives (**Figure 1**).

Evaluating the Feasibility of Developing Therapeutic Guidelines Based on Breast Cancer Tumor Patient Proteomic Heterogeneity.

The primary objective was to explore the intra- and inter-tumoral heterogeneity of breast cancer (BC) for therapeutics, using matrix-assisted laser desorption/ionization mass spectrometry imaging (MALDI MSI). The identification of distinct molecular subpopulations and clones within and between patient tissues highlighted the complexity of characterizing BC tumors, which is crucial for proposing effective treatments. Spatial proteomic methods were employed to analyze the proteome of different tumor clones in depth. The resulting protein data identified potential druggable targets across all tumor clones, enabling the development of treatments tailored to tumor heterogeneity. To validate this process, patient-derived organoids paired with original tumors were treated with either conventional therapies recommended by pathologists or treatments tailored based on the clonal proteomic data. Preliminary results, obtained from a limited number of patient-derived tumoral tissues, showed that treatments guided by proteomic data provided superior anti-tumoral efficacy compared to conventional treatments. Furthermore, these experiments helped identify potential biomarkers of drug resistance, based on proteomic profiles and organoid drug responses, which are critical for guiding therapy. However, this approach requires substantial biological material and is time-consuming, presenting challenges for routine clinical application.

Development of a Machine Learning Model for Predicting Protein Pathways from Lipid Analyses.

The following step of the study was to develop a machine learning model capable of directly predicting biological pathways and protein information of heterogeneous clusters from lipid analyses via MALDI MSI, without conducting separate spatial proteomic experiments. The concept of "dry proteomics" was introduced, focusing on the spatial localization of identified clusters in both lipid and protein imaging. By establishing consistency in cluster appearance across omics images, it becomes feasible to link them to specific lipid and protein pathways initially identified through rigorous spatial lipidomic and proteomic analysis. This approach allows for the integration of lipid

and protein pathways within these clusters, forming the core of dry proteomics. The concept was first optimized using rat brain tissues before being applied to glioblastoma tissues for validation on complex and heterogeneous clinical samples. Thus, by predicting protein information from tumor mass spectrometry images, dry proteomics could serve as a promising tool for clinical use, significantly reducing the time required for tumor characterization.

Understanding Spatial and Temporal Heterogeneity of Breast Cancer.

The third objective aimed to understand the spatial and temporal heterogeneity of breast cancer by applying the dry proteomics workflow in clinical settings and taking into account the notion of clonal expansion of the tumor during its trajectory in time course, sub-types tumors and the treatment nature (chemotherapy, hormone therapy, radiotherapy, immunotherapy combine or alone). This approach was used to identify actionable targets and potential treatments, which were then validated *in vitro* using organoids derived from the tumors.

Development of MALDI IHC Technique Based on Tag Mass Technology.

The primary objective was to explore the use of MALDI multiplex immunohistochemistry (IHC) based on tag mass technology, which enables the rapid and sensitive identification of predicted protein targets for therapeutic intervention, while simultaneously mapping their spatial localization. This technique was applied in supplementary studies to assess its potential and identify areas for improvement. Additionally, by employing specific immune cell markers, such as immune checkpoints, the method allowed for the spatial delineation of intricate interactions between tumor and immune cells, along with their phenotypic expression. This innovative approach aimed to enhance the precision and efficiency of identifying critical protein targets for personalized cancer therapies.

Together, these objectives aimed to expand the functional heterogeneity analysis of tumors, improve the prediction of cancer progression and recurrence risk through biomarkers, and provide more personalized and effective treatments for patients.

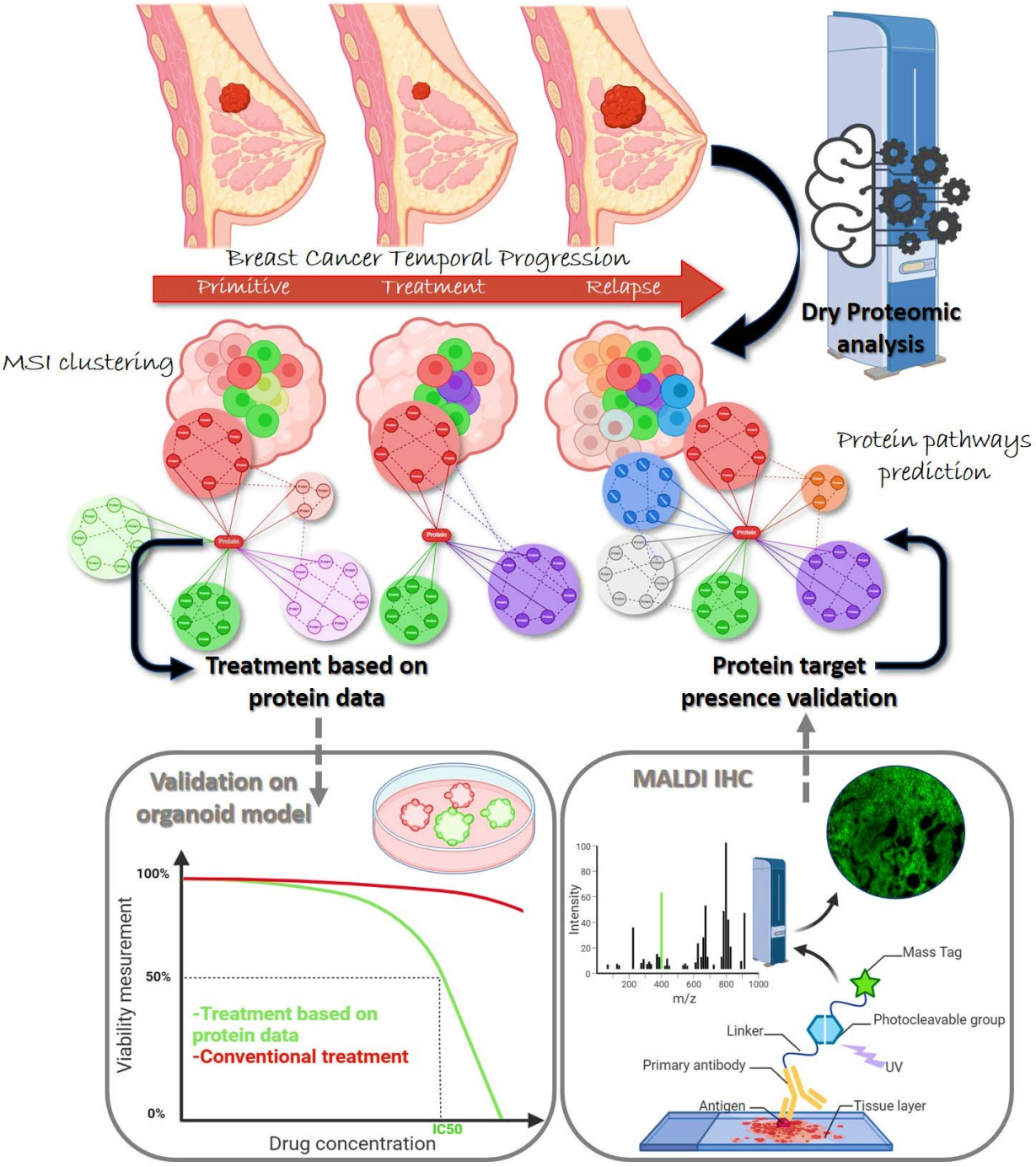
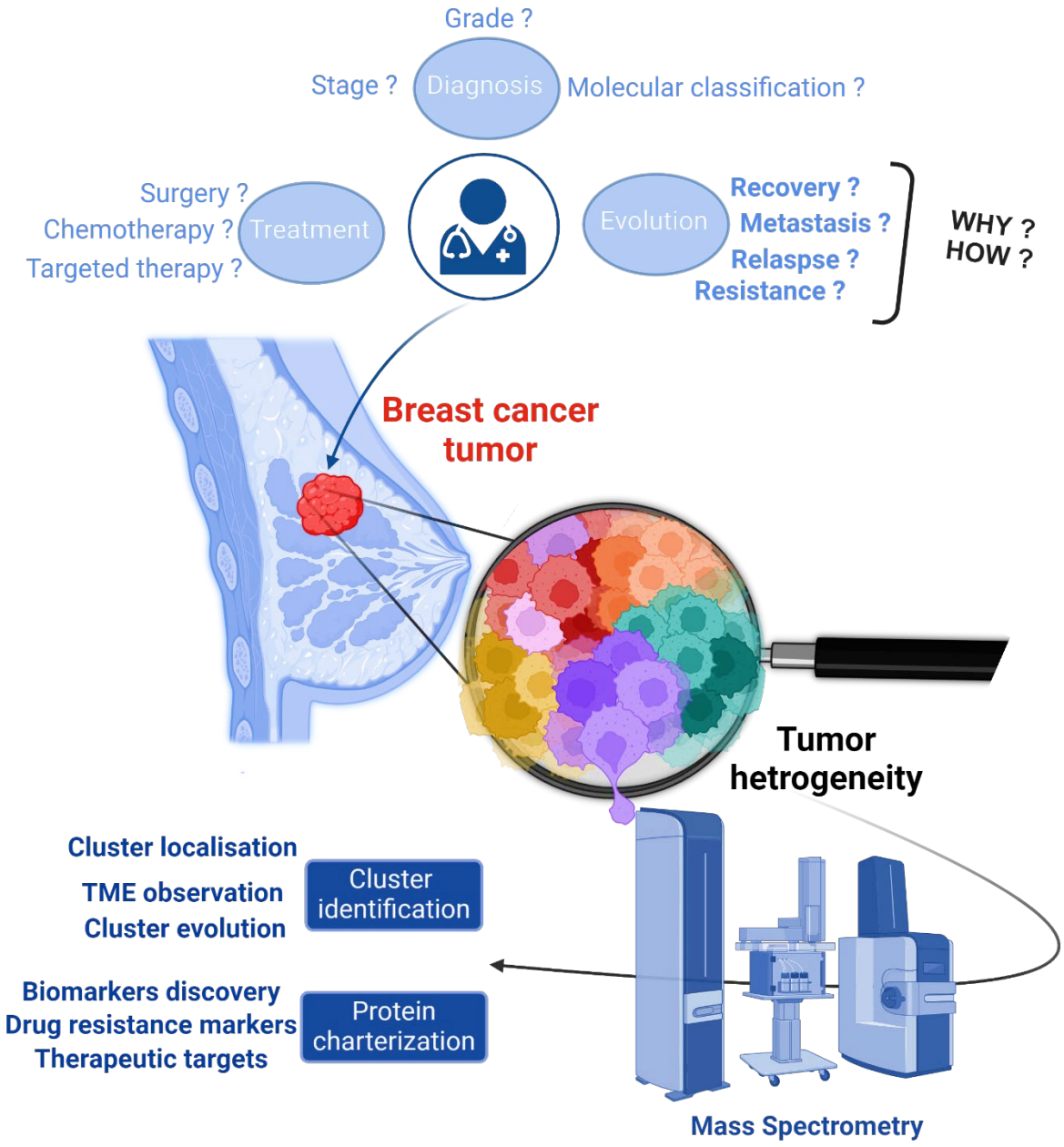


Figure 1: Thesis aims overview. The figure illustrates a personalized breast cancer treatment strategy integrating dry proteomic analysis, network modeling, and experimental validation using organoid models and MALDI IHC to guide targeted therapies and prevent relapse.

CHAPTER 1

State of the Art



CHAPTER 1: State of the Art

Clinical Background

Breast Anatomy and Physiology

The breast anatomy structure reflects its specific biological function, which is the milk production for lactation further to new-born birth. This gland sits on top of the upper ribs and chest muscles. There is a left and right breast. Each breast is made up of a nipple surrounded by the areola skin, a mammary gland and a connective tissue that contains vessels, fibers and fat (Bazira et al., 2022). The mammary gland contains fifteen to twenty compartments, each separated by fat tissue, made of lobules (milk production place) and ducts (small canals coming from the lobules and carry the milk to the nipple) (**Figure 2**). Its development and function are allowed thanks to two sexual hormones, the estrogen and the progesterone, produced by the ovaries. The estrogen allows in the breast development during the puberty and the pregnancy, whereas the progesterone has a role in the differentiation of breast cells and in the menstrual cycle.

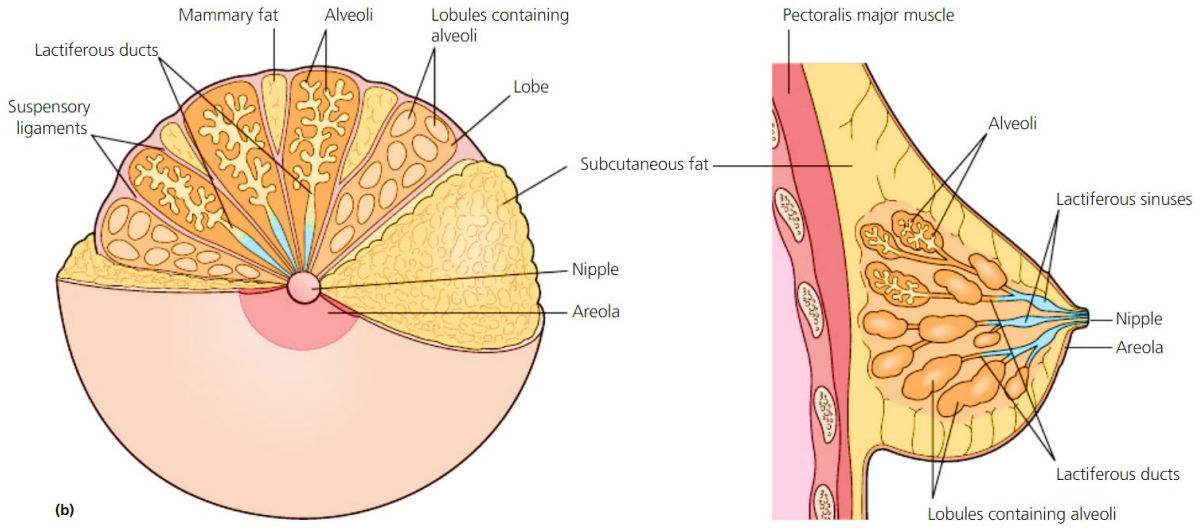


Figure 2: Human breast anatomy. Illustration of internal breast anatomy, showing lobes, ducts, and alveoli involved in lactation, along with surrounding fat, muscles, and external structures like the nipple and areola.

The vessels that run through the breast are blood and lymphatic in nature. Lymphatic vessels lead to many lymph nodes, thus forming the lymphatic system (**Figure 3**). This latter is a part of the body's immune system, working at a network level with lymph nodes, ducts, vessels and organs, to collect and carry clear lymph fluid through the body tissues to the blood. The clear lymph fluid circulates from the breast to the nodes and can therefore contains waste material produced by the tissues, as well as immune system cells. The breast lymphatic nodes are mainly localized: in the armpit (axillary lymph nodes), above the collarbone (supraclavicular nodes), below the collarbone

(sub clavicular or infraclavicular nodes), and inside the chest, around the sternum (internal mammary nodes).

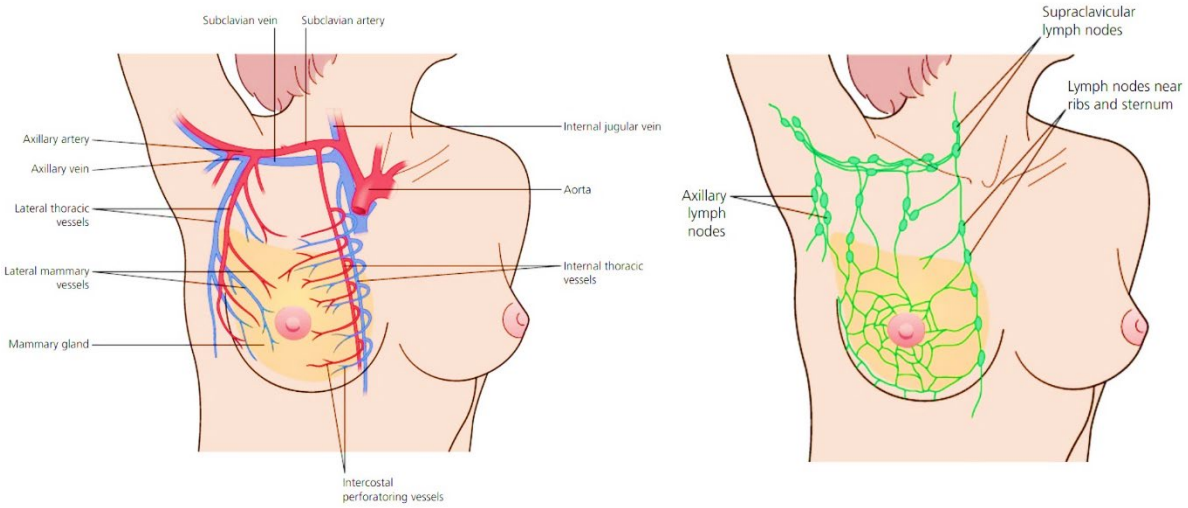


Figure 3: Lymphatic vessels communicating with breast. Illustration showing bloody vessels and lymph nodes networks linked to breast.

Breast Cancer

Incidence

Breast cancer (BC) is the most diagnosed cancer in the world with 2.26 million new cases in 2020, as well as 685 000 deaths, making it the second deadliest cancer in woman. Breast cancer, predominantly associated with women, stands as one of the most prevalent and devastating forms of cancer worldwide (**Figure 4**). While it primarily affects women, it's crucial to acknowledge that men can also be affected, albeit less frequently. In France, the Foundation for Medical Research conducted estimations in 2018, revealing a concerning statistic: the risk of developing breast cancer for women is approximately 1 in 8, with over 47% of cases diagnosed in women under the age of 65. The incidence of breast cancer exhibits an alarming upward trajectory over time, transcending age demographics.

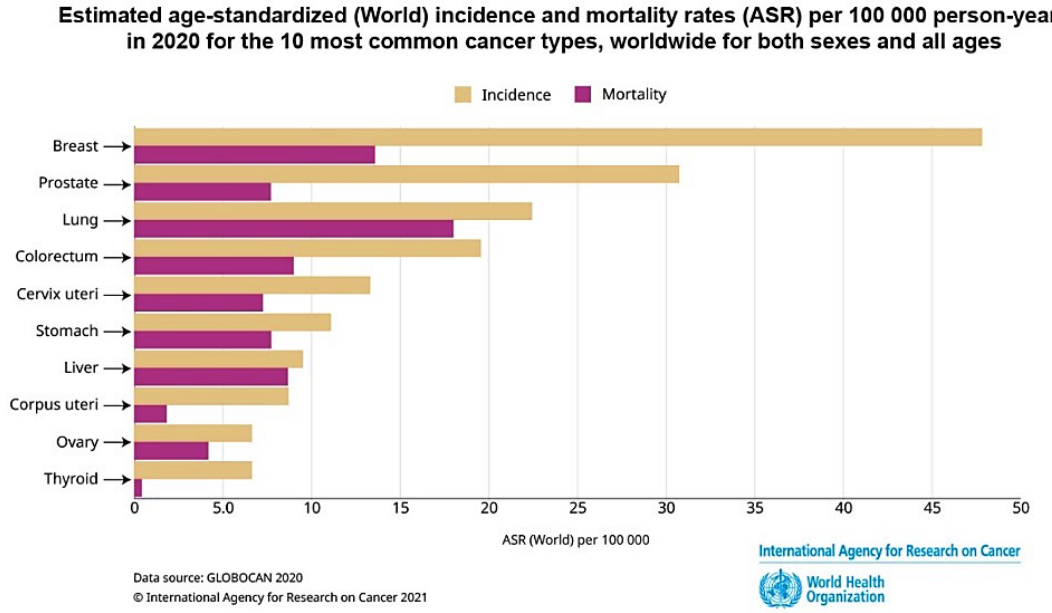


Figure 4: Incidence and mortality rates for the 10 most common cancer types worldwide in 2020. The diagram highlights that breast cancer has the highest incidence rate and is the second leading cause of cancer-related mortality worldwide, across all sexes and age groups.

Counterintuitively, amidst this surge in diagnoses, there's been a noteworthy decline in mortality rates (**Figure 5**). Enhanced survival outcomes have become increasingly attainable, particularly when cancer is detected at its nascent stages (Cronin et al., 2018). This decline in mortality can be attributed to significant therapeutic advancements, including hormonal therapy, taxanes, and targeted treatments tailored to the molecular profile of tumors. Additionally, the uptick in early-stage diagnoses, facilitated by screening programs, has played a pivotal role. Despite these strides, breast cancer remains a formidable adversary, with its far-reaching impact still making it the foremost cause of cancer-related deaths among women in France. The overall prognosis, however, has significantly improved (**Figure 5**), with a commendable 88% survival rate over five years for cases diagnosed between 2005 and 2010 (Cowppli-Bony et al., 2017). Nonetheless, the sheer magnitude of annual diagnoses underscores the ongoing urgency in tackling this formidable disease. Indeed, the incidence of breast cancer increases over the time, at any age of the population. The French Public Health Agency measured an increase of new cases diagnosed by 95% between 1990 and 2018. This augmentation can be explained, on part by the increase in risk and half to the increase and decrease aging of the population (+26% and +21% respectively), as well as the development and implementation of screening programs.

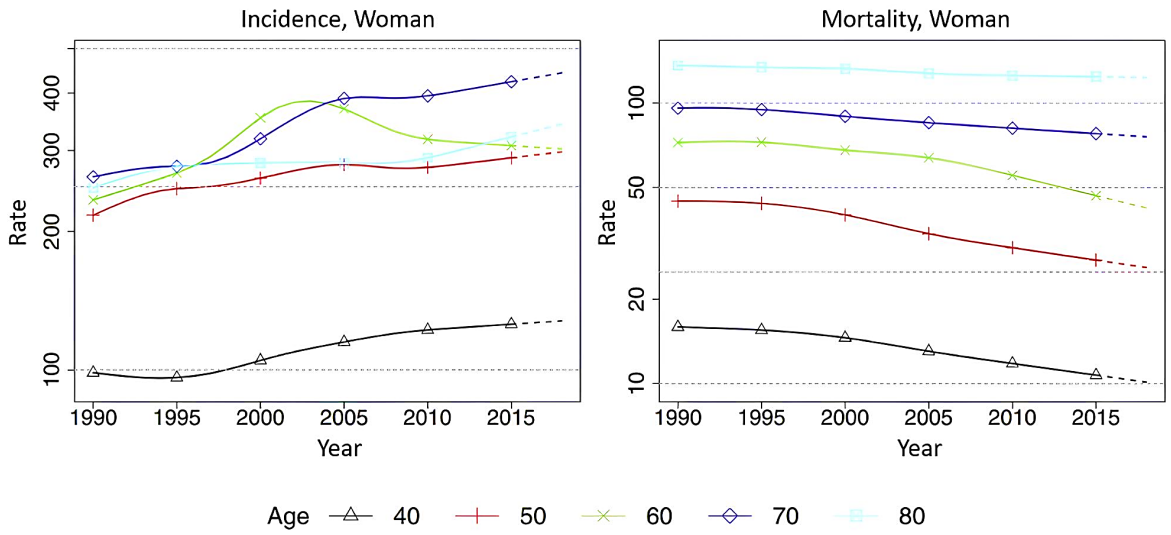


Figure 5: Incidence and mortality evolution of breast cancer over time for woman highlighting a decrease of mortality due to medicine and diagnostic improvements while incidence is increasing.

Breast Cancer Types

Breast cancer results in an unregulated proliferation of epithelial cells at the level of the mammary gland, following a multi-step process that is caused by consecutive genetic mutations and interaction with environmental factors (Polyak, 2007). These cells divide and proliferate more rapidly than healthy cells, hence the accumulation of cells forming a mass called a tumor. These tumors elicit phenotypical alterations further to the malignant transformation.

Biologically speaking, a factor that increases cell proliferation will increase also the risk of cancer. In this way, the high rate of breast cancer in woman may be explain by the impact of woman hormones (estrogen and progesterone) on the breast cell division rate.

There are many different types of breast cancer, all determined by the kind of breast cells affected and the technic used, the gold standard is based on the pathological examination. The most often, breast cancer starts in the milk ducts (forming ductal carcinoma type) or in the lobules (forming lobular carcinoma type). It can also begin with the areola skin cells of the nipple (called Paget disease of the breast), the fat tissue cells (called phyllodes tumor) or the lining of the blood and lymph vessels (called angiosarcoma), but these later are fewer common types of breast cancer. The type of breast cancer can differ whatever the cancer has spread or not. There is a pre-cancer type, called in situ breast cancer (ductal carcinoma in situ or DCIS, lobular carcinoma in situ or LCIS), where the cancer has not spread into the rest of the breast tissue. At the opposite, an invasive/infiltrating breast cancer (invasive ductal carcinoma or IDC, and invasive lobular carcinoma or ILC) describe a type of breast cancer that has spread into the surrounding breast tissue (**Figure 6**). This invasive type represents about 70-80% of all breast cancer diagnosed (Shehata et al., 2019)

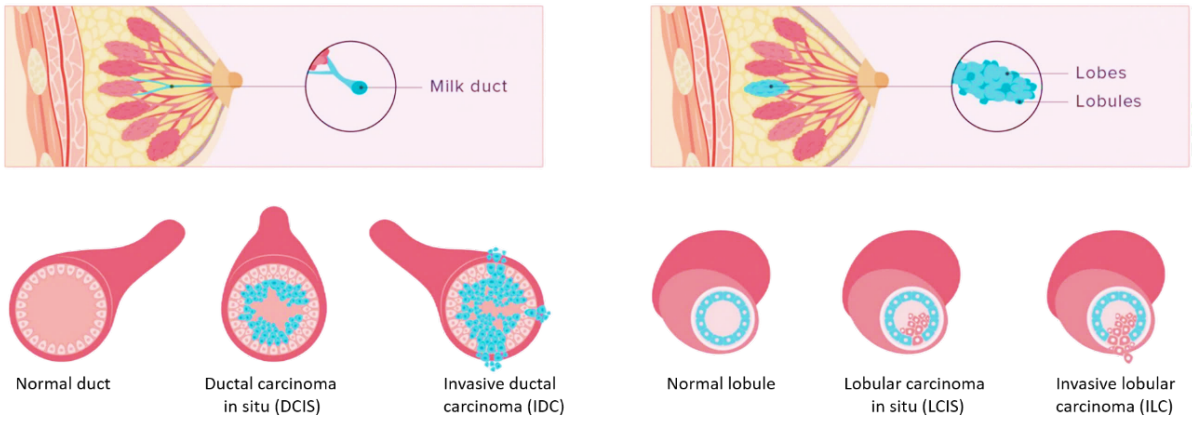


Figure 6: Representation primary origins of breast cancer in milk ducts and the lobules. In both cases, the progression of ductal carcinoma is shown. The stages demonstrate the growth of abnormal cells, with potential spread to surrounding tissue.

Breast cancer can also progress to metastatic cancer when cancer cells escape and spread elsewhere via hematological and lymphatic routes. Indeed, like mentioned previously, blood and lymph vessels run through the mammary gland and can therefore carry cancer cells to other parts of the body to metastasize, often in preferential distant sites (lung, liver, bones, brain). In this case, the lymph nodes can develop metastases (adenocarcinoma), just as other organs may develop metastases without any cancer cells in the lymph nodes.

Breast Cancer Signs and Diagnosis

Breast Cancer Symptoms

It is possible that breast cancer doesn't cause any symptoms during the first stages of the disease. The first signs appear when the tumor is big enough to feel a lump in the breast, or in the lymph nodes when the cancer has spread to nearby tissues.

Other symptoms of ductal or lobular breast cancer may include:

- mass in the armpit (axillary hollow);
- change in breast size or shape;
- nipple changes, such as a nipple that suddenly starts to point inward (inverted nipple);
- discharge from the nipple without being compressed or that is tinged with blood.

Breast Cancer Diagnosis

The diagnosis of a breast cancer mainly begins when a lump is detected by palpation, or/and imaging by mammography, echography or IRM. If there is a suspicious lesion in the breast, a biopsy sample from this lesion is performed by a percutaneous incision, allowing an anatomopathological examination.

This examination is very important since it makes it possible to determine with certainty whether it is a cancerous lesion or not, as well as the characteristics of the tumor. Indeed, in addition to the type of cancer, a breast cancer is commonly classified according to a stage, grade and molecular subtype. This information is useful to know more about the cancer and helps guide treatment options for the patients.

Stage classification

The first step of an anatomopathological examination consist in defining the propagation of the cancer. This is carried out by macroscopic analysis, based on the TNM classification (Tumor, Nodes, Metastasis) (https://www.uicc.org/):

- The T designates the size of the tumor, the rating ranging from T0 to T4. Indeed, the size and infiltration of the tumor gives an indication of the disease progression.
- The N refers to node's affectation. The tumor presence, or not, in nodes as well as how many there are and where they are located can inform about cancer propagation.
- The M refers to the spread of the cancer, with the absence (note M0) or the presence (note M1) of metastases in the body.

The combination of these criteria allows to evaluate the global stage of breast cancer, classified from 0 to IV, depending on whether the cancer is at an early, advanced or metastatic stage (**Table 1**).

Table 1: Breast cancer stage classification according to tumor size, lymph nodes cancer affectation and spreading.

Stage		*Tumor size*	*Lymph nodes*	*Spreading*
Stage 0	Non invasive		No cancer	Confined to the breast
Stage I	Early stage	< 2cm	No cancer	Confined to the breast
Stage II	Localized	2-5cm	Affected by cancer	Confined to the breast
Stage III	Regional spread	>5cm	Affected by cancer	Confined to the breast
Stage IV	Distant spread		Affected by cancer	Metastases outside the breast area

Grade classification

The examination of breast cancer is then specified thanks to a microscopic analysis of tumor cells, based on the SBR (Scarff Bloom and Richardson (Bloom & Richardson, 1957; Le Doussal et al., 1989)) grade prognosis, to characterize the aggressiveness of the cancer. The grade is defined according to a scoring system considering the amount of gland formation (the cell differentiation), the nuclear features (the degree of pleomorphism) and the mitotic activity (the tumor cells proliferation). Each of these features is scored from 1 to 3, and then added together to give a total score corresponding to a grade; higher is the score, more aggressive is the cancer (**Table 2** and **Figure 7**).

Table 2: Breast cancer grade classification according to glandular differentiation, nuclear pleomorphism and mitotic count.

Grade	*Glandular differentiation*	*Nuclear pleomorphism*	*Mitotic count*
Score 1	>75% of tumor forms glands	Uniform cells with small nuclei similar in size to normal breast epithelial cells	< 7 mitoses per 10 high power fields
Score 2	10% to 75% of tumor forms glands	Cells larger than normal with open vesicular nuclei, visible nucleoli, and moderate variability in size and shape	8-15 mitoses per 10 high power fields
Score 3	<10% of tumor forms glands	Cells with vesicular nuclei, prominent nucleoli, marked variation in size and shape	> 16 mitoses per 10 high power fields

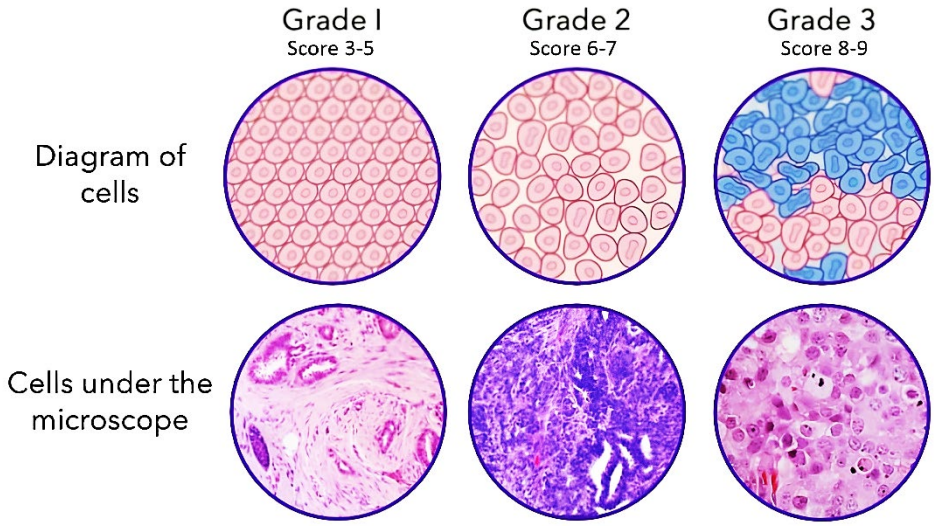


Figure 7: Histological grading of breast tumor cells. Illustration of the histological grading of tumor cells based on their structural differentiation, with diagrams of cells and corresponding microscope images. Grade 1 cells, with a score between 3 and 5, are well-differentiated and closely resemble normal tissue in both structure and organization. Grade 2 cells, scoring between 6 and 7, are moderately differentiated, showing irregularities in shape and arrangement, with less resemblance to normal cells. Grade 3 cells, which score between 8 and 9, are poorly differentiated and highly disorganized, indicating a more aggressive and abnormal tumor structure.

Molecular classification

Breast cancer intrinsic molecular subtype is the transcriptomic-based classification that most precisely defines this cancer. This approach (Sørlie et al., 2001) classifies breast cancer in the following subtypes: luminal A, luminal B, HER2-enriched and basal which have different prognosis. In practice, the pathologists use a simpler method to estimate this classification. They take into account the genes expressed in cancer cells, which control the behavior of cells. It is mainly based on the presence or absence of three major receptors on the cell surface. These receptors are hormone receptors (HR) such as Human Epidermal growth factor-2-Receptor (HER2), estrogen receptor (ER), and progesterone receptor (PR). Some other proteins can also specify the cancer subtype, like the Ki-67 nuclear antigen, an active cell proliferation marker.

In this way, three breast cancer subtypes are defined (**Figure 8**):

➔ **The luminal breast cancer** which can be splited in three subclasses:

- **The luminal A breast cancer** is ER positive, PR positive, and HER2 negative, with a low level of Ki-67.
- **The luminal B breast cancer** is ER positive, PR negative and HER2 negative, with a high level of Ki-67.
- **Luminal B-like breast cancer** is ER positive, HER2 positive and PR negative or positive, with any level of Ki-67.

➔ **The HER2 breast cancer** is ER negative, PR negative, and HER2 positive.

➔ **The triple negative, or basal-like breast cancer** is ER negative, PR negative, and HER2 negative.

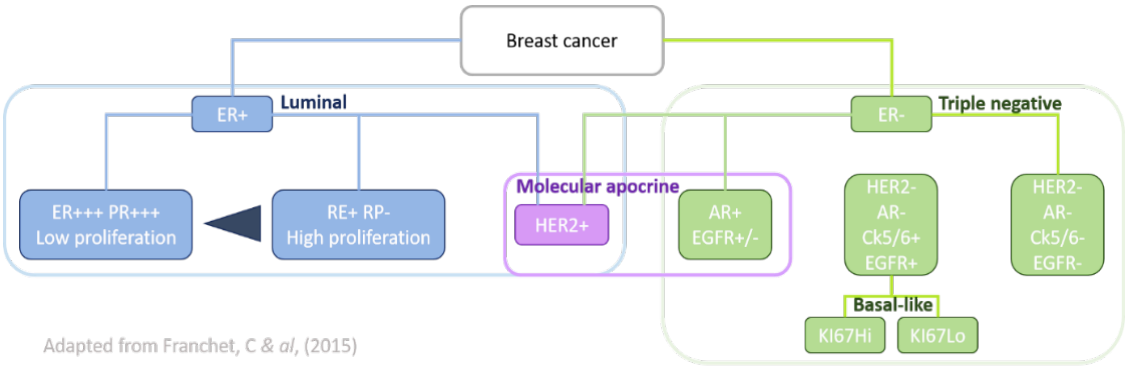


Figure 8: Breast cancer major molecular subtypes classification groups. Breast cancer luminal, HER2 enriched and triple negative categories according to receptor status and proliferation markers: human epidermal growth factor-2-Receptor (HER2), estrogen receptor (ER), progesterone receptor (PR), Ki-67, cytokeratin (Ck) and androgen receptors (AR) expression.

Breast Cancer Treatment Options

Breast cancer treatment guidelines depends on elements following the pathologist analysis as well as the personal history of the patient:

- The type of cancer (invasive ductal, invasive lobular or in situ)
- The stage and the grade
- Cancer hormone receptor status
- HER2 status
- Recurrence risks
- Overall health
- Menopausal status
- Case of cancer in the family
- Personal environment

According to these features, breast cancer treatments are divided in two options: a local treatment or a systemic treatment. The local option includes the surgery and radiation therapy, whereas a systemic treatment regroups hormonal therapy, chemotherapy, targeted therapy, immunotherapy, given as adjuvant (after surgery) or neoadjuvant (before surgery) systemic therapy.

Local Treatments

Surgery

Surgery is the first type of treatment for early BC. It depends of the type of breast cancer (most often stages I, II or III) and is usually followed by chemotherapy or radiotherapy in case of high risk of recurrence, and by endocrine therapy if tumors express estrogen or progesterone receptors, or targeted therapy if indicated. This part of the treatment aims to remove as much of the cancer as possible, to determine whether the cancer has spread to the axillary lymph nodes under the arm or to symptoms of advanced cancer.

There are two main protocols of breast cancer surgery: the lumpectomy, a breast-conserving surgery where the cancerous lump is removed, and the mastectomy, where the whole breast is removed with or without breast reconstruction.

In the case where the cancer could spread, a sentinel lymph node biopsy or an axillary node dissection can be performed on one or a few axillary lymph nodes, to evaluate the evolution of the cancer. This step can be carried on during the breast cancer removal, or as a separate operation (Czajka & Pfeifer, 2022).

Radiotherapy

Radiotherapy is a treatment given when treating early-stage BC, often after surgery. The goal of this procedure is to destroy any cancer cells that my remain after surgery. The radiation therapy can also be used in cases where breast cancer can't be removed with surgery, or in metastatic breast cancer which has spread to other organs of the body.

This method uses high-energy X-rays, or particles radiations, to damage cancer cells DNA. Indeed, when cell's DNA is damaged, it cannot divide and dies. However, cancer cells divide and proliferate more rapidly and are less organized than healthy cells, thus cancer cells are more affected than normal cells, which are better able to repair themselves and survive to the radiations.

Systemic Treatments

Chemotherapy

Chemotherapies for breast cancer use anti-cancer drugs which can be injected by intravenous or orally administered (**Table 3**). The drug will therefore travel through the bloodstream around the body to target and destroy cancer cells. This type of treatment is frequently used in the

therapeutic strategy along with other treatments such as surgery, radiation or hormone therapy. Chemotherapy may be given after surgery (adjuvant therapy), to reduce any risk of cancer recurring or spreading by destroying undetected cancer cells, or it may be given before surgery (neoadjuvant therapy), to shrink larger cancers and enable the surgeon to remove the entirely tumor with less invasive surgery. Neoadjuvant chemotherapy is often used for inflammatory breast cancers, HER2-positive BC, Triple negative BC, High grade BC, cancers that have spread to the lymph nodes, and larger BC.

Table 3: Drugs for chemotherapy in breast cancer. (Wind & Holen, 2011)

Class of drug	*Drug*	*Mechanism of action*
Anthracycline	Doxorubicin	Acts by intercalating DNA, resulting in complex formation which inhibits DNA and RNA synthesis. Triggers DNA cleavage by topoisomerase II resulting in cell death.
	Epirubicin	Acts by intercalating DNA.
Taxan	Paclitaxel	Mitotic inhibitor; interferes with the normal function of microtubule breakdown. Also induces apoptosis.
	Docetaxel	Interferes with microtubule breakdown.
Anti-metabolites	5-Fluorouracil	Metabolized to cytotoxic metabolites which are incorporated into DNA and RNA, inducing cell cycle arrest and apoptosis.
Alkylation agent	Cyclophosphamide	Inhibits DNA synthesis.

Hormone therapy

Hormone therapy consists in slowing or stopping the growth of hormone-sensitive tumors by interfering the body's ability to produce hormones or act on breast cancer cells (**Table 4**).

Table 4: Drugs for hormone therapy in breast cancer. (Journé et al., 2008)

Class of drug	*Drug*	*Breast cancer treatment*	*Mechanism of action*
Selective estrogen receptor modulator	Tamoxifen	- approved for premenopausal and postmenopausal women - after having surgery for early-stage ER-positive breast cancer have reduced risks of breast cancer recurrence - approved to treat metastatic breast cancer	Competes with estrogens for binding to the estrogen receptor antagonizing the proliferative effects of estrogens. Have partial estrogen agonist actions on other organs (e.g., uterus, bone).

Selective estrogen receptor degrader	Fulvestrant	- approved for postmenopausal women - metastatic ER-positive breast cancer that has spread after treatment with other antiestrogens - HR-positive, HER2-negative locally advanced or metastatic breast cancer who have not previously been treated with hormone therapy	Competes with estrogens for binding to the estrogen receptor antagonizing the proliferative effects of estrogens. Has no partial estrogen agonist effects. Inactivates and destroys estrogen receptor.
Aromatase inhibitor *Steroidal*	Exemestane	Postmenopausal women with advanced breast cancer whose disease has worsened after treatment with tamoxifen.	Inhibit estrogen biosynthesis by inhibiting aromatase, the enzyme that catalyzes conversion of androgens to estrogen.
Aromatase inhibitor *Non-steroidal*	Anastrozole	Postmenopausal women as initial therapy for metastatic or locally advanced hormone-sensitive breast cancer.	
Aromatase inhibitor *Non-steroidal*	Letrozole	Post menopausal women with advanced disease.	

Immunotherapy

Immunotherapy has emerged as a revolutionary approach in breast cancer treatment, offering a new dimension of care, particularly for patients with difficult-to-treat subtypes like triple-negative breast cancer. Unlike conventional treatments such as surgery, chemotherapy, and radiation, which directly target cancer cells, immunotherapy leverages the body's own immune system to recognize and eliminate cancer cells (Debien et al., 2023).

One mechanism in immunotherapy is the targeting of immune checkpoints. These are proteins, such as PD-1 or PD-L1, that normally act as brakes on the immune system, preventing it from attacking healthy cells. Drugs known as checkpoint inhibitors, like pembrolizumab and atezolizumab, block these proteins, allowing immune cells (particularly T-cells) to find and destroy cancer cells more effectively (Debien et al., 2023).

In addition to checkpoint inhibitors, anti-tumor vaccines are another form of immunotherapy being explored for breast cancer. These vaccines are designed to train the immune system to recognize

specific proteins or mutations present on cancer cells. By doing so, the immune system can mount a stronger and more targeted attack against the cancer. Beyond immediate tumor destruction, these vaccines can also create immune memory, a long-lasting immune response that helps prevent the cancer from returning. For HER2+ BC, oncologists generally use monoclonal antibodies such as trastuzumab and pertuzumab in combination with chemotherapy. These antibodies specifically target the HER2 protein on the surface of cancer cells, helping the immune system recognize and destroy them.

A more experimental but highly promising avenue in immunotherapy is CAR-T cell therapy. This treatment involves extracting a patient's T-cells and genetically modifying them to express chimeric antigen receptors (CARs), which are engineered to target specific antigens on cancer cells (Dey et al., 2023). Once these enhanced T-cells are infused back into the patient, they are better equipped to locate and destroy cancer cells. While CAR-T cell therapy has shown remarkable success in blood cancers, ongoing research is exploring its potential in solid tumors like breast cancer. Early results in preclinical studies, particularly in animal models, have been encouraging, showing the potential for this approach to combat even the most resilient forms of breast cancer.

Targeted therapy

Targeted therapy is a type of cancer treatment that uses drugs or substances to specifically identify and attack cancer cells while sparing healthy cells. In breast cancer, these therapies target specific molecules, such as proteins or genes, that are responsible for cancer growth. By blocking the function of these targets, targeted therapies can slow or stop the cancer's progression. They are often used alongside other treatments like chemotherapy, hormone therapy, or surgery.

Different types of targeted therapies are available depending on the breast cancer subtype:

- HER2+ BC: In cases where the HER2 protein is overexpressed, therapies like trastuzumab and pertuzumab block the HER2 protein to prevent cancer growth. These therapies fall under immune therapies for HER2-positive breast cancer.
- HR+ (ER+ or PR+) BC: These cancers depend on hormones like estrogen or progesterone to grow. Targeted therapies in this subtype work by blocking hormone receptors or reducing hormone production. For instance, CDK4/6 inhibitors (e.g., Palbociclib, Ribociclib, Abemaciclib) block proteins that promote cell division in hormone receptor-positive cancers.
- BRCA-Mutated BC: Women with BRCA1 or BRCA2 gene mutations have a higher risk of developing breast cancer, as these mutations impair the cell's ability to repair DNA. PARP

inhibitors (e.g., Olaparib, Talazoparib) exploit this weakness by inhibiting the PARP enzyme, leading to the accumulation of DNA damage and the death of cancer cells.

- PI3K/AKT/mTOR Pathway: Some breast cancers have mutations in this signaling pathway, which drives cell growth. Drugs targeting this pathway can effectively inhibit tumor growth.

The benefits of targeted therapies in breast cancer treatment lie in their precision and reduced toxicity compared to traditional chemotherapy. These therapies are designed to specifically attack cancer cells by targeting molecules or pathways involved in cancer growth, which helps to minimize damage to healthy cells and reduces the severity of side effects like hair loss, fatigue, and nausea often associated with chemotherapy. Additionally, targeted therapies can be more effective for certain breast cancer subtypes, such as HER2+ or BRCA-mutated cancers, by directly addressing the biological mechanisms driving tumor growth. However, despite these advantages, there are limitations. One major challenge is the development of resistance over time, where cancer cells adapt and become less responsive to the treatment. Furthermore, not all patients have the specific molecular targets needed for these therapies to be effective, meaning they may not benefit from these treatments.

Breast Cancer Heterogeneity Drawbacks

It was noticed that 30% of patients treated for breast cancer relapse locally or with distant metastases, despite appropriate multidisciplinary treatments. This can be explained by molecular heterogeneity within the tumor generating variable sensitivity to treatments (**Figure 9**). Indeed, recent advances in genomic sequencing have revealed greater than expected heterogeneity (Burrell et al., 2013). The occurrence of numerous point alterations at a genetic and epigenetic level within a single cell are subsequently associated with subsequent clonal expansions. These can vary according to many parameters such as the tumor microenvironment of each tumor, the patient's environment, or the impact of the treatment imposed on the latter (Caiado et al., 2016). Non-genetic mechanisms also contribute to heterogeneity within tumors; transcriptional, translational, and metabolic adaptations can lead to the emergence of molecular subpopulations of clonal cancer cells. Ultimately, a given patient's breast tumor is made up of a complex mosaic of genetically related clones, which is specific to him (Burrell et al., 2013).

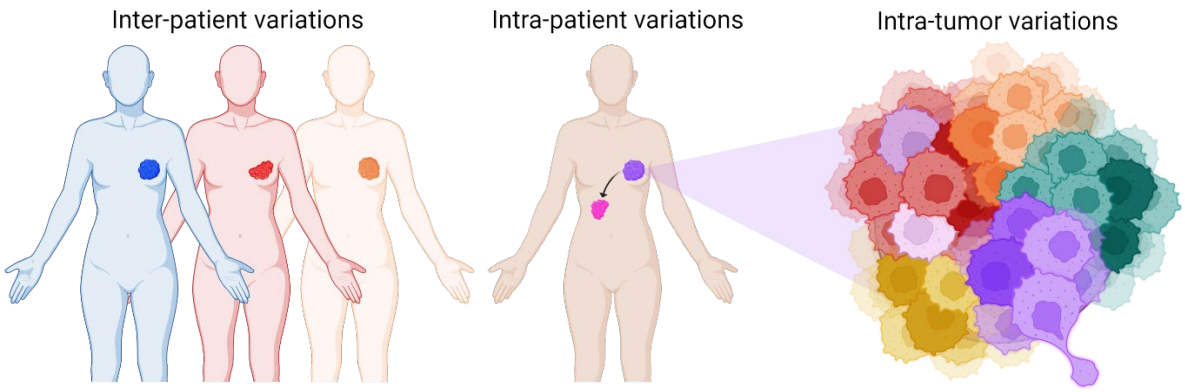


Figure 9: Breast cancer inter- and intra-tumor heterogeneity.

This heterogeneity is a major drawback when developing a treatment for cancer because there is both a genetic and functional difference between tumors and within tumors (Burrell et al., 2013), due to variabilities between diverse cellular subpopulations. The challenge is all the greater as this heterogeneity can evolve over time and across space, thus in part explaining recurrences caused by patient predisposition to treatment resistance *(Dagogo-Jack & Shaw, 2017)*. Indeed, the conventional treatments are mainly based on genetic markers which do not consider the molecular heterogeneity inside the tumor. In this way, they may as well be effective on some molecular subpopulations as not at all. It is, therefore, necessary to refine the analysis of tumors by integrating their heterogeneous subpopulations, to identify tailored markers and potentially actionable targets at a subpopulation level.

Cancer Heterogeneity and Preclinical Target Discovery Challenges

Biomarkers Discovery Principle

In cancer context, a biomarker can be defined as a "biological molecule produced by the tumor cell or human tissues in response to cancer that is objectively measured and evaluated as an indicator of cancerous processes within the body" (Füzéry et al., 2013), which can also inform on cancer progression regarding therapy. Biomarkers can include molecules such as proteins, nucleic acids, metabolites, and imaging-based parameters. There are three major types of cancer biomarkers (Mordente et al., 2015):

- Prognostic markers can predict the natural progression of cancer like aggressiveness or malignancy, without treatment impact. It allows us to guide on the best treatment choice possible. These markers can also predict the patients' survival.
- Predictive markers, allow promising cancer behavior facing some treatment. They are found clinically according to "responder" or "non-responder" patients. In this way, the treatment can be readapted to the tumor sensibility. Some biomarkers can also predict anti-cancer drugs' toxicity.

- Pharmacodynamic markers, inform the effect of the drug on the patient's body (absorption, target inhibition, metabolite pathway, elimination). It also warns about the toxicity of the treatment, in which case the doses must be appropriate.

Their use in the clinic makes it possible to predict the risk of development of cancer, the likelihood of tumor response, early cancer detection, and an improved diagnosis for an optimal therapy selection.

The pipeline of biomarker development includes 5 steps (**Figure 10**) (Kenner et al., 2017):

1. The first one, preclinical exploratory studies, is the main issue of this thesis, and will be the one detailed in this manuscript. It consists in discovering tumor biomarkers following a hypothesis-driven approach (a target method measuring the involvement of candidate biomarkers in cancer biology) or a discovery-based approach (Füzéry et al., 2013), which is exploited in this study. This later is an untargeted method of identifying candidate biomarkers thanks to experimental analysis at different stages of the pathology, using high-throughput omics technologies. In this way, the evolution of discovered markers can be developed by measuring their presence as well as their differential expressions at different stages of cancer.
2. Once the biomarker is identified, it will pass some validation tests in clinical assay development to measure its analytical validity, clinical validity, and clinical utility in clinical laboratories.
3. The third step of the marker development is a retrospective longitudinal repository study. The biomarker will be evaluated in samples collected from research cohorts.
4. Then the biomarker is tested by screening the patient to validate his diagnostic function, which can lead to a treatment.
5. Finally, the biomarker is tested on a large population to observe if the use of this later helps to reduce the mortality rate (validation of clinical utility).

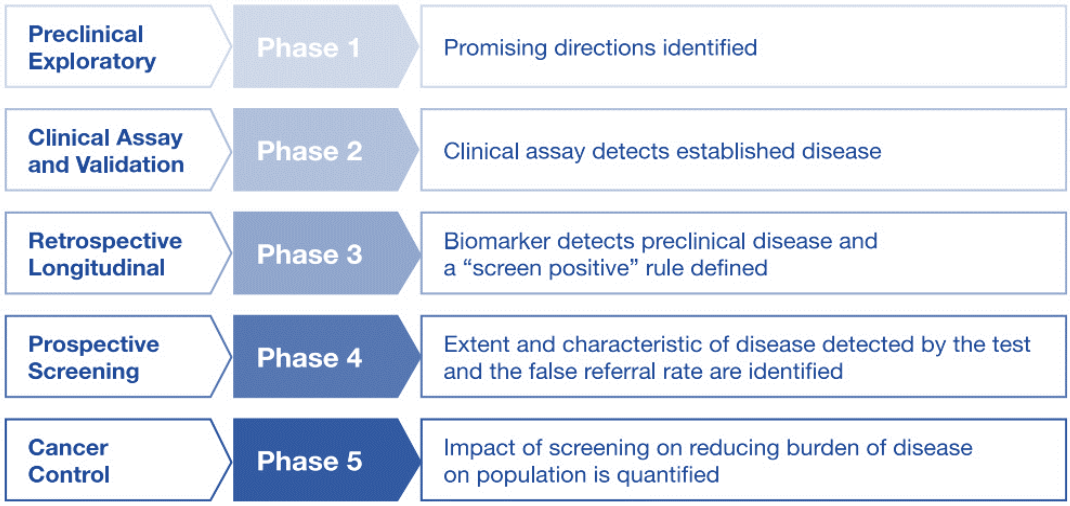


Figure 10: Biomarker discovery and validation phases. Five phases of biomarker development in disease screening are outlined. Phase 1 involves preclinical exploratory studies to identify promising research directions. In Phase 2, clinical assays are validated for detecting established diseases. Phase 3 focuses on retrospective longitudinal studies where biomarkers detect preclinical disease, and screening criteria are established. Phase 4 involves prospective screening to assess the extent and characteristics of disease detection and to identify false referral rates. Finally, Phase 5 quantifies the impact of screening on reducing the disease burden within the population.

Preclinical Biomarker Qualification Process

In Vitro Model Impact

In the realm of preclinical drug development for cancer, a significant challenge arises from the absence of suitable cell culture model systems. These systems play a pivotal role in assessing the effectiveness of potential anti-cancer drugs before they are tested on patients. The cumulative effect of inadequate in vitro model has significant implications for drug development. An estimated 96% of drugs fail to progress from the discovery stage to clinical trials. This high failure rate can, in part, be attributed to the fact that the preclinical models used are not sufficiently aligned with the biological intricacies of tumors in patients (Heem Wong et al., 2019). Consequently, promising drug candidates that perform well in these imperfect models might not demonstrate the same efficacy or safety profiles in human clinical trials. This observation is all the truer when it comes to finding therapeutic targets considering the context of molecular heterogeneity of cancer. However, the existing models are continually in development in replicating the complex conditions found within actual tumors. Among all the models used in research, the most relevant are: the two-dimensional cell lines, the patient derived xenograft models and the patient derived tumor organoid models.

Two-Dimensional (2D) cancer cell lines

Two-dimensional (2D) cancer cell lines are frequently employed in the initial stages of biological research, including drug development, to study cell behavior, molecular mechanisms, and responses to external factors such as drugs or other treatments.

Concretely, 2D cell lines are established cell lines derived from cancerous primary tumor tissues, growing in a flat, monolayer arrangement in culture dishes. They can be maintained and propagated in culture, with the required nutrient-rich medium environment to grow and survive.

In context of drug development for cancer, 2D cell culture models are often used in the early stages of screening potential drug candidates:

- Drug Screening: expose the cultured cancer cells to various drug compounds, either individually or in combination, to assess their effects on cell growth, viability, and proliferation. This step helps identify promising drug candidates that exhibit favorable anti-cancer activity in a controlled laboratory setting.
- Mechanism of Action: molecular mechanisms studies underlying a drug's effects on cancer cells by analyzing changes in gene expression, protein levels, and signaling pathways within the cells.
- Dose-Response Curve: by testing different concentrations of a drug on the cells, researchers can generate a dose-response curve, which provides insights into the drug's potency and effectiveness at different doses.

While 2D cell culture models offer certain advantages, such as simplicity and ease of use, they also come with significant limitations (**Figure 11**). A critical limitation of 2D cancer cell lines is their failure to accurately replicate the interactions that occur between cancer cells and their surrounding micro-environment within a tumor. The tumor micro-environment encompasses a diverse range of components, including neighboring cells including immune cells, blood vessels, and the extracellular matrix. The absence of this complex interplay in 2D models leads to an incomplete understanding of how drugs interact with the tumor. The deficient representation of the tumor micro-environment in 2D cell cultures has significant implications for drug resistance. The tumor micro-environment plays a crucial role in influencing how cancer cells respond to treatment. By disregarding these interactions, 2D models can overlook important mechanisms of drug resistance, potentially leading to inaccurate predictions of drug efficacy.

In summary, while 2D cell culture models have been valuable tools for early-stage drug screening and mechanistic studies, they have inherent limitations that must be considered when interpreting results and making predictions about the effectiveness of potential cancer therapies in clinical settings. In this way, more advanced models are in development, such as three-dimensional (3D) cell cultures, organoids, and organ-on-a-chip systems, to better recapitulate the complexities of tumors and improve the accuracy of preclinical drug development studies.

Patient-Derived Xenograft model

A Patient-Derived Xenograft (PDX) model is an advanced preclinical research technique used in cancer drug development. It involves the transplantation of tumor tissue obtained directly from a cancer patient into immunodeficient mice. This model allows to study the growth and behavior of the patient's tumor in a living organism, providing a more accurate representation of human cancer biology compared to traditional cell culture or genetically engineered mouse models.

Experimentally, the process begins with the surgical removal or biopsy of a tumor from a cancer patient. This tumor tissue contains the complex genetic, molecular, and cellular characteristics of the patient's cancer. The patient tumor sample is then divided into small fragments and implanted under the skin or within an organ of immunodeficient mice. These mice lack a functional immune system, which prevents rejection of the human tumor tissue. The implanted tumor fragments establish a PDX model in the mice, where they grow and replicate the characteristics of the original patient tumor. PDX models can be propagated across multiple generations of mice to ensure consistency and reproducibility.

Thus, PDX models closely mimic the genetic and cellular characteristics of the patient's tumor, offering a more accurate representation of human cancer biology compared to other model systems. These models also maintain the diverse cell populations and genetic heterogeneity present in the patient's tumor, allowing for the study of different subpopulations within the same model.

As 2D cells, PDX models have several valuable applications in cancer drug development:

- Drug Screening: to test the effectiveness of potential cancer drugs. This allows for more accurate prediction of how a drug will interact with the patient's tumor, providing insights into its potential efficacy and toxicity.
- Personalized Medicine: PDX models can be used to develop personalized treatment approaches. By studying how an individual patient's tumor responds to different drugs, clinicians can make more informed decisions about the most suitable treatment strategy.
- Understanding Tumor Biology: PDX models provide a platform for studying the biology of different cancer types. Changes in gene expression, protein profiles, and cellular interactions within the tumor microenvironment can be analyzed.
- Resistance Mechanisms: PDX models enable the investigation of mechanisms underlying drug resistance. This allows to study how tumors evolve and adapt in response to treatment, offering insights into strategies to overcome resistance.

Finally, PDX models represent a powerful tool in cancer drug development, offering a more clinically relevant and personalized approach to understanding tumor biology and evaluating

potential therapies. These models bridge the gap between traditional in vitro cell cultures and complex human tumors, contributing to the advancement of precision medicine and improved patient outcomes.

However, some limitations of these models must be noticed (**Figure 11**). The use of immunocompromised mice hinders the study of immune system interactions, an important aspect of cancer research. Additionally, the time-consuming and expensive nature of PDX model maintenance makes them less suitable for high-throughput screenings, which require rapid and efficient testing of multiple compounds or treatments (Marshall et al., 2014). These limitations must be considered when choosing the appropriate model system for their specific research goals .

Patient-derived tumor organoids model

A Patient-Derived Organoid (PDO) model is an innovative and advanced in vitro system used in cancer research and drug development. PDOs are 3D structures that closely resemble miniature versions of human organs or tissues, created from patient-derived cells. These models aim to capture the complex architecture and cellular interactions present in real organs, making them valuable tools for studying disease biology, drug responses, and personalized medicine hence their interest in this study.

The conception of PDOs begins with obtaining a small tissue sample, such as a biopsy, from a patient's tumor or healthy tissue. This tissue contains a mixture of different cell types and maintains the genetic and molecular characteristics of the original organ. Specific cells are isolated and cultured in a specialized environment that encourages them to self-organize and form 3D structures resembling the original organ's architecture. The isolated cells continue to proliferate and differentiate, eventually forming multicellular 3D structures known as organoids. These organoids can be propagated and expanded over multiple generations, maintaining their genetic and molecular features. However, it requires careful optimization and standardization to ensure reproducibility and consistency across experiments (Raffo-Romero et al., 2023).

PDO models have several important applications in cancer research and drug development:

- Tumor Biology Studies: PDOs provide a platform for studying the behavior, growth, and interactions of cancer cells within the context of the original organ. This offers insights into tumor biology, progression, and heterogeneity.
- Drug Screening: Researchers can use PDOs to test the effects of various drugs on patient-derived tumor tissue. This allows for more accurate assessment of drug responses and the identification of potential treatment options tailored to individual patients.

- Personalized Medicine: PDO models enable the testing of multiple treatment options on a patient's own tumor tissue. This allows clinicians to select the most effective treatment strategy based on the response of the PDOs, potentially leading to more personalized and targeted therapies.
- Understanding Drug Resistance: PDO models can help uncover mechanisms of drug resistance by studying how tumor cells respond to treatment and how they adapt over time.
- Mechanistic Studies: PDOs offer a controlled experimental platform for investigating the molecular and cellular mechanisms underlying cancer development and progression.

PDOs models offer a promising avenue for advancing cancer research and drug development by providing a more accurate and patient-relevant platform for studying tumor biology and treatment responses. These models bridge the gap between traditional cell culture systems and complex animal models, even if the conceptual problem of the representation of intra-tumoral heterogeneity by the in vitro organoid model is critically discussed. PDOs are still offering a valuable tool for advancing our understanding of cancer and improving therapeutic approaches (**Figure 11**).

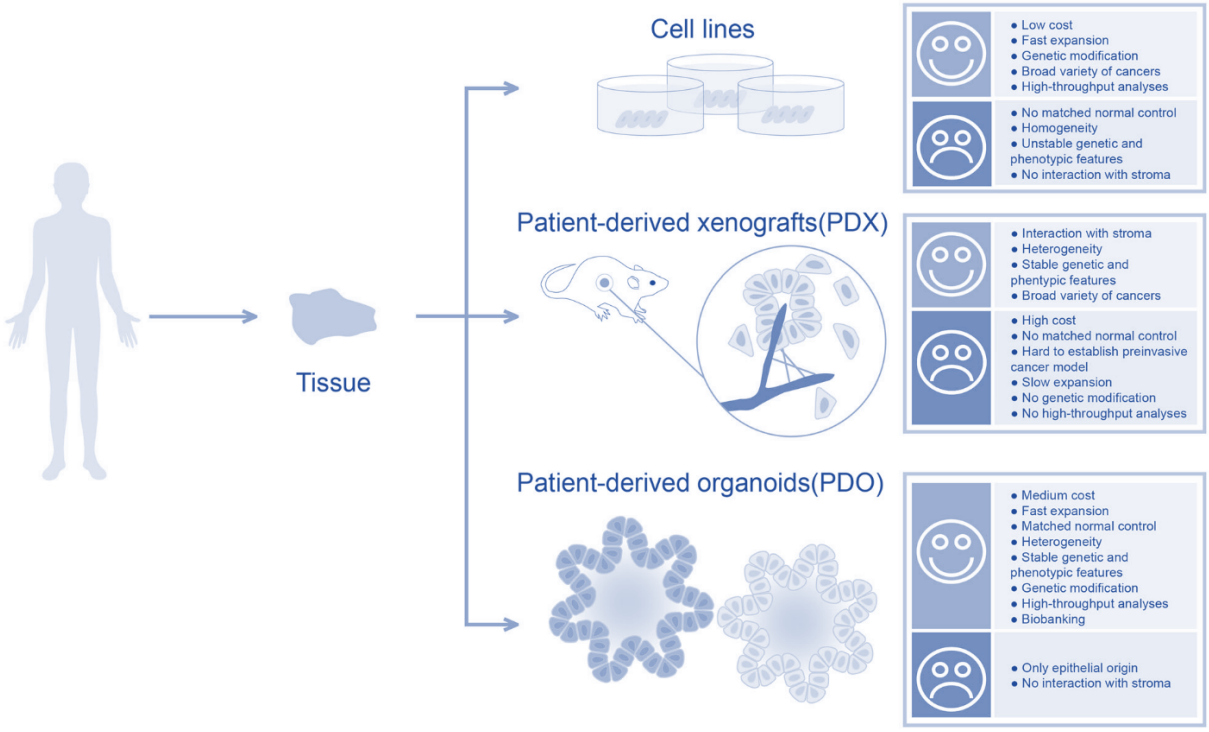


Figure 11: Comparison of Cancer Models Derived from Patient Tissues. Comparison of three cancer modeling systems: cell lines, patient-derived xenografts (PDX), and patient-derived organoids (PDO), highlighting their pros and cons in terms of cost, genetic features, and biological relevance. (Y. Li et al., 2020)

Experimental Design and Protocol Development

The design of experiments and development of protocols play a critical role in biomarker qualification. Variables such as sample size, control groups, and experimental time points must be carefully considered. Longitudinal studies might be necessary to capture dynamic changes in

biomarker levels over time. Protocols for sample collection, storage, and processing must be standardized to ensure consistency across experiments. Detailed documentation of experimental procedures is essential for reproducibility and comparability.

Analytical Methods and Assay Development

Developing robust analytical methods and assays is essential for accurate biomarker measurement. Immunoassays, such as enzyme-linked immunosorbent assays (ELISAs), are commonly used for protein biomarkers, while polymerase chain reaction (PCR) and next-generation sequencing are employed for nucleic acid-based biomarkers. Mass spectrometry offers high sensitivity and specificity for detecting small molecules and peptides. Assay validation should include parameters such as linearity, precision, accuracy, and limit of detection. Quality control samples and calibration curves are essential components of assay development.

Data Analysis and Interpretation

Data analysis involves processing raw data into meaningful results. Statistical methods, such as ANOVA (analysis of variance), t-tests, or regression analysis, are used to determine significant differences between groups. Bioinformatics tools are employed for omics data analysis, enabling the identification of patterns and correlations. Data interpretation involves relating biomarker levels to the underlying biological processes, disease progression, and treatment effects. Visualization tools, such as heatmaps, scatter plots, and pathway analysis, aid in conveying complex data to a broader audience.

Reproducibility and Robustness

Ensuring the reproducibility and robustness of biomarker qualification studies is vital for their credibility. Quality control measures, such as using validated reagents and calibrators, help minimize assay variability. Standard operating procedures (SOPs) outline step-by-step protocols for sample handling, assay procedures, and data analysis. Collaborative efforts, such as inter-laboratory studies, can assess the robustness of biomarker assays across different research settings. The use of reference materials with known biomarker concentrations aids in standardizing measurements and comparing results across laboratories.

Introduction to Mass Spectrometry Imaging for Tumors Characterization

Contextualization

A range of traditional methods, such as biopsies, histology and genetic sequencing, have been instrumental in providing insight into the nature of tumors and guiding treatment decisions. However, they have certain limitations that hinder a comprehensive understanding of the complexity inherent in cancer, particularly regarding tumor heterogeneity. One of the primary challenges of these conventional methods lies in their inability to capture the full spectrum of diversity present

within tumors. As previously evocated, cancer consists of a mosaic of different cell types, genetic mutations, and molecular characteristics that vary widely, even within the same tumor. Traditional techniques often sample only a small portion of the tumor, potentially missing crucial information about the entire tumor landscape. Moreover, while these methods excel in identifying specific genetic or molecular markers, they don't consider the spatial context, due to their inability to reveal untargeted molecular changes within the tumor. Understanding the spatial distribution of various molecules, proteins, and cell types within a tumor is pivotal as the location of these components often plays a significant role in determining treatment responses and disease progression. This is where the significance of spatial information becomes evident in cancer. Different areas of a tumor might exhibit distinct characteristics, leading to variations in how they respond to treatments or progress over time.

The limitations of traditional cancer analysis methods in capturing the complexity and spatial arrangement of tumor components highlight the critical need for advanced technologies like Mass Spectrometry Imaging (MSI). MSI has emerged as a powerful tool capable of providing spatially resolved molecular information within tumors, enabling to visualize the intricate landscape of targeted or untargeted molecules within tissues, and understand the heterogeneous nature of cancer.

In essence, while traditional methods have been invaluable in cancer research, their limitations in revealing the spatial and heterogeneous nature of tumors underscore the importance of embracing innovative techniques like MSI to unlock the deeper complexities of cancer biology and pave the way for more personalized and effective treatments.

Mass Spectrometry Generalities

Mass spectrometry (MS) is a technique, widely used as an analytical method, allowing rapid and direct analysis of various biomolecules from diverse origins (mineral or organic), complexity (pure or complex mixture) and sizes (from single atoms to protein complexes). It makes it possible to detect, identify and characterize the chemical structure of molecules of interest by mass measurement and fragmentation. The principle of MS is based on the instrument's ability to transform gas-phase and charged molecules under high vacuum, to separate them according to their mass-to-charge ratio (*m/z*), thanks to electric or magnetic fields. Thus, a mass spectrometer is composed of 3 parts combining an ionization source, followed by one or more analyzers, and a detector (**Figure 12**). In gas phase, positively or negatively charged ions are generated within the ionization source, then guided and separated within the analyzer before being detected by

converting ion current into an electrical signal that can be processed either analogically or digitally. This results in a mass spectrum which reports the ion current as a function of the *m/z* ratio.

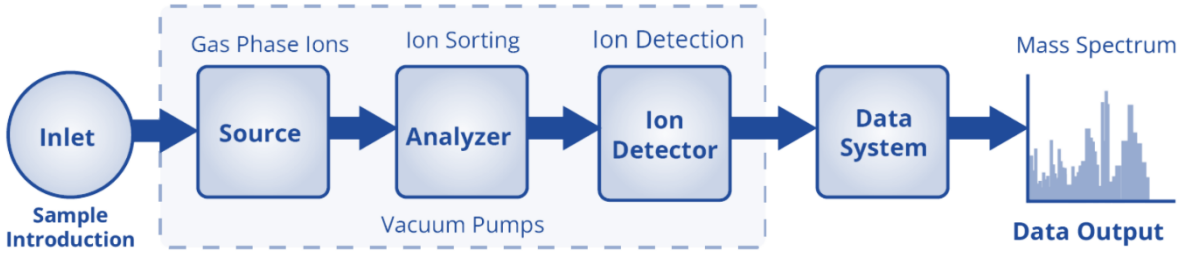


Figure 12: Mass spectrometer instrumentation. This figure represents the steps in mass spectrometry analysis. A sample is introduced through the inlet, followed by ionization in the source to produce gas-phase ions. These ions are then separated in the analyzer based on their mass-to-charge ratio. The ion detector captures the sorted ions, and the data system processes the signal to generate a mass spectrum, providing the final data output for analysis.

Numerous technological advancements have been necessary to enhance the diversity of samples analyzable by MS. Since the creation of gas phase ions is required for the measurement, diverse type of ionization source was developed and adapted according to the variety of samples. For example, the electronic impact (EI) or chemical ionization (CI) sources, involving sample collision with high-energy electrons, are limited to the small molecular weight and polarity compounds analysis. In the same idea, the flow injection (FI) and the field desorption (FD) ionization methods, consisting in injecting a small and constant flow sample into the mass spectrometer with or without the application of a strong electric field, are not suitable non-volatile compounds. At the opposite, some ionization sources overcome this limit, such as the fast atom bombardment (FAB), the plasma desorption, or the thermospray ionization, however it needs important protein concentrations. The emergence of the matrix assisted laser desorption ionization (MALDI) developed by Karas and Hillenkamp (Hillenkamp et al., 1991), as well as the electrospray ionization (ESI), developed by Fenn team (Fenn et al., 1989), has gained major interest in the field of chemical analysis since 1980.

The advantages of MALDI resides in its high spatial resolution and sensitivity, conferring to the instrument the ability to analyses a wide range of molecules such as metabolites, lipids, peptide and proteins. This technique has the capability to generate gas-phase ions from molecules that have previously crystallized in low concentration with an aromatic matrix (**Figure 13**). This matrix serves to lower the ionization energy emitted by a UV or IR laser, thereby producing more stable ions for the intact analysis of compounds with minimal fragmentation. Indeed, photons emitted by the laser are absorbed by the matrix components, elevating them to a higher electronic energy state. The transfer of this energy, during relaxation, leads to the desorption of molecules that subsequently desolvate and ionize under the influence of the applied high vacuum within the instrument. This refers to the laser desorption/ionization phenomenon (LDI), introduced by Koichi Tanaka which earned him the

Nobel Prize in Chemistry in 2002. Thus, this ionization source, ultimately regarded as a soft ionization method, thus facilitates the analysis of solid samples, such as biological surfaces.

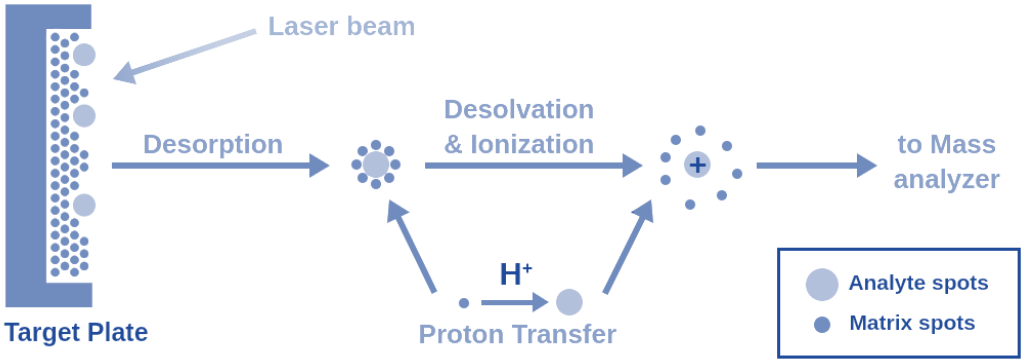


Figure 13: MALDI source ionization. Analyte spots are embedded in matrix spots on a Target Plate. A laser beam irradiates the matrix, causing desorption of the analyte-matrix mixture. During desolvation and ionization, analyte molecules are ionized through proton transfer (H^+), leading to the formation of charged analyte ions. These ions are directed to the Mass Analyzer for further analysis.

Regarding the ESI source, the electrospray ionization involves the creation of ions from a liquid sample solution, based on the electro nebulization phenomenon at atmospheric pressure. Basically, the sample dissolved in a volatile solvent is introduced in a capillary, or needle. A high voltage is applied to the solution as it emerges from the capillary, resulting in the formation of the Taylor cone: a fine aerosol of charged droplets. As these charged droplets travel through the ESI source, solvent molecules evaporate, leaving behind highly charged analyte ions (**Figure 14**). The process generates ions with various charge states, creating a distribution of ions corresponding to different molecular weights and charges. These ions are then directed into the mass spectrometer for analysis. Due to its compatibility with a wide range of analytes (including large biomolecules like proteins, peptides, nucleic acids, and small organic molecules), the ESI source is most coupled downstream of separation techniques, such as liquid chromatography. The advantage of these setups allows for the analysis of highly complex liquid sample mixtures, by coupling liquid chromatography with mass spectrometry.

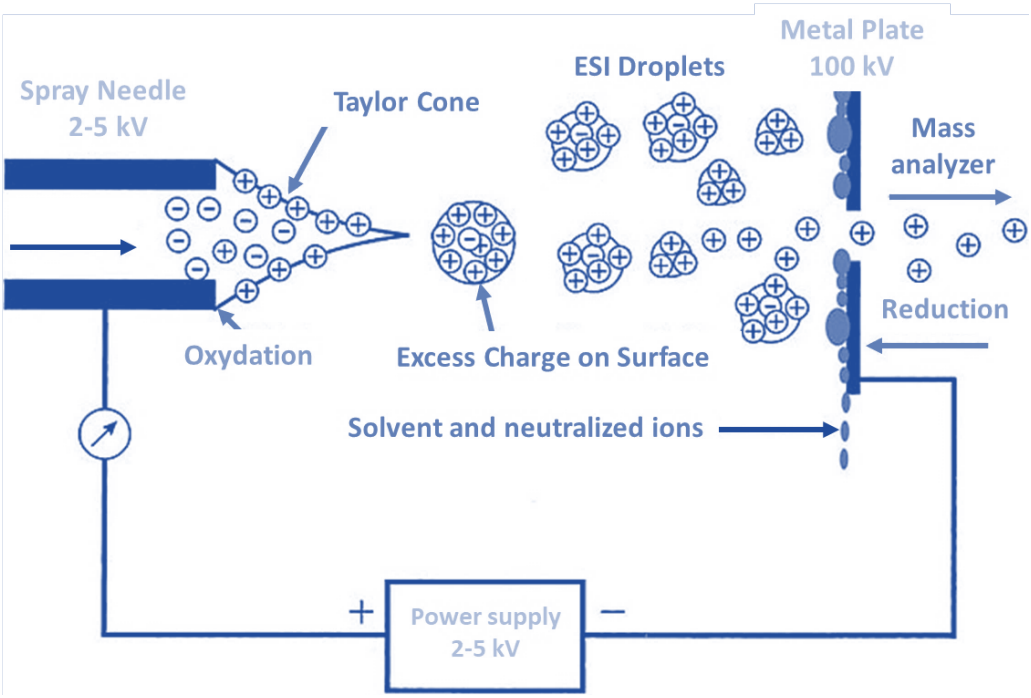


Figure 14: Representation of ESI source ionization technique used in mass spectrometry to ionize analytes from a solution. A high voltage (2-5 kV) is applied to the Spray Needle, which causes the liquid to form a Taylor Cone, ejecting charged droplets. These ESI Droplets contain excess charge on their surface as the solvent evaporates. The droplets are drawn towards a Metal Plate (100 kV), reducing in size through evaporation, leaving charged analyte ions. The Mass Analyzer detects and separates the ions based on their mass-to-charge ratio for analysis.

Introduction to Mass Spectrometry Imaging

Mass spectrometry imaging (MSI) is a powerful analytical technique that combines the capabilities of mass spectrometry with spatial information, allowing the visualization and analysis of the distribution of molecules within a histologic complex tissue sample. Indeed, the principle lies in a pixel-by-pixel analysis of the entire sample surface. A mass spectrum is recorded for each position, containing the *m/z* ratios of each detected ion, with its abundance represented by an intensity, as well as its x and y coordinates. Thus, the MSI technique allows to provide the molecular cartography of a whole histologic sample to perform targeted or untargeted analysis. The ionization sources used for MSI depend on their ability to generate charged ions from the molecules present in the sample without significantly disrupting their spatial arrangement. MSI finds applications across numerous scientific fields, including biology, medicine, pharmacology, forensics, and material science. It has been used for various purposes, such as biomarker discovery (identifying and localizing specific molecules associated with diseases or biological processes within tissues), drug distribution studies (mapping the distribution and metabolism of pharmaceutical compounds within biological tissues to understand drug efficacy and toxicity), or for molecular histology (distinguishing tumoral or healthy tissue areas), for example. Indeed, taking the biological environment into account is essential information for understanding the functional relevance of molecules, especially in the study of cancer heterogeneity, which is covered here.

Many mass spectrometry sources are available for use in imaging. Among these, the MALDI imaging technique will be the only one discussed here. Indeed, MALDI imaging is the main technique that was used throughout this work, due to its compatibility with multi-omics analyses directly on a solid surface (such as biological cancer tissues) and with use in routine.

MALDI MSI

The first mass spectrometry imaging method using a MALDI was developed by Caprioli's group in 1997 (Caprioli et al., 1997). The first MALDI imaging software (MS imaging tool) was developed by Stoeckli (Stoeckli et al., 1999), then MITICS by PRISM (Jardin-Mathé et al., 2008). As presented previously, this ionization source is based on the desorption/ionization process caused by a UV or IR laser shot on a solid sample having previously crystallized with an inorganic matrix. The matrix then absorbs the laser energy to ionize the compounds, without fragmentation, allowing the desorption and ionization of the molecules in the sample.

More experimentally, sample preparation is a limiting step in the quality of the results obtained by imaging. MALDI imaging requires a series of essential preliminary steps for data acquisition, which will depend on the type of molecules to be analyzed, as well as the nature and conservation of the sample. Carrying out these preparatory steps will have a significant influence on the definition of MALDI spectra, and therefore on the accuracy of the image visualization. In the context of this study aiming to analyze the molecular heterogeneity of tumor tissues, the sample preparation steps include tissue preparation for preservation, sectioning tissues to be mounted on a slide for instrumental analysis, selection and application of the matrix, data acquisition in imaging mode, and data processing enabling their visualization as images (**Figure 15**). Over the years, numerous technical advancements have been made to this process, aiming to enhance the performance of MALDI imaging in terms of analysis quality, spatial resolution, as well as data acquisition and processing times.

Every stage of sample preparation plays a crucial role in achieving high-quality imaging outcomes. This starts with the preparation of the sample. There are two primary methods used to preserve a biological sample, both of which have advantages and disadvantages depending on the type of analysis required. The first preservation method involves freezing fixation. This method is carried out immediately after the sample is taken in the operating room. It entails immersing the sample in a tissue freezing medium, which is then cooled in liquid nitrogen between -45°C and -70°C. The sample, thus frozen, can be stored at -80°C until sectioning using a cryostat at -20°C. Fresh frozen (FF) samples remain in a stable state close to the living condition for about a year before the initial molecular degradation is observed. This stabilization offers the advantage of being directly compatible with various MALDI imaging strategies to analyze the full range of omics molecules. To

enhance the stability of molecules, the paraformaldehyde fixation method is the most used technique, particularly for samples from hospital environments. This fixation, combined with paraffin embedding, allows preservation for several decades, which is useful for retrospective studies. Tissue stabilization, known as FFPE (Formalin-Fixed Paraffin-Embedded), relies on the creation of chemical cross-links between proteins. However, these cross-links pose a challenge for the analysis of these samples in MALDI imaging. Special treatments are required prior to matrix deposition after sectioning with a microtome (**Figure 15**). The various treatments used aim to remove the paraffin and chemical cross-links to facilitate the ionization of molecules for MALDI imaging analysis. During these steps, the loss of small molecules (lipids and metabolites) and tissue degradation are often observed, but still analyzable. Nevertheless, this doesn't hinder the attainment of good-quality and reproductible proteomic and glycan imaging analyses.

The second key point in sample preparation is matrix deposition (**Figure 15**). Indeed, the choice of matrix and deposition procedure are crucial, depending on the analysis and the quality of the desired results.

Regarding the deposition method, two solutions are commonly used nowadays. The first involves matrix deposition by an automatic sprayer. This tool optimizes spray parameters (temperature, drying time, number of passes, spray speed, etc.) for each matrix deposition to achieve a homogeneous coating and small-sized crystals, promoting better ionization of molecules. The second option is a matrix deposition method by sublimation. In this method, the matrix is heated to its sublimation temperature (where it transitions from solid to gas without passing through a liquid phase) and then recondenses on the sample plate, forming a thin, even layer of matrix crystals. Sublimation generally produces more well-defined, high-quality crystals, which can enhance the ionization efficiency and reproducibility. However, it can be a slower process compared to spray deposition, which is more adapted for routine analyses.

The choice of matrix is a critical aspect in conducting MALDI imaging experiments. There are several matrices, each characterized for detecting a specific class of molecules, with varying spectrum quality. In the case of metabolomic or lipidomic approaches, using a matrix with a low number of matrix peaks is preferable to minimize overlap with potential low mass biomarker peaks. For the analysis of positively charged small molecules like lipids, the most used matrix is 2,5-dihydroxybenzoic acid (DHB). Conversely, the analysis of metabolites or lipids in negative mode is favored by using 9-aminoacridine (9-AA), 1,5-diaminonaphthalene (DAN), or N- (1-naphthyl) ethylenediamine dihydrochloride (NEDC) matrices. More recently, Norharman has emerged as a new matrix with good spectral and imaging quality, enabling analysis of metabolites and lipids in both

negative and positive modes. Regarding protein detection, the primary matrix used is synapinic acid (SA), or the SA-Aniline matrix developed at the PRISM laboratory (Franck et al., 2009). Similarly, peptides can be detected using specific matrices such as α-cyano-4-hydroxycinnamic acid (HCCA) or HCCA-Aniline matrix (Bonnel et al., 2013). The emergence of the aforementioned ionic matrices is due to a significant increase in performance compared to organic matrices, achieved by incorporating an ionic compound like aniline. A summary of the matrices mentioned in this section is presented in the **Table 5**.

Table 5: Exhaustive list of matrices with characteristics.

Name	*Chemical formula*	*Polarity*	*MW*	*Detectable molecules*
9-AA	$C_{13}H_{10}N_2$	Negative	194.23	Metabolites/Lipids
DAN	$C_{10}H_6\ (NH_2)_2$	Both	158.20	Metabolites/Lipids
DHB	$(OH)_2C_6H_3CO_2H$	Both	154.12	Lipids
HCCA	$OHC_6H_4CH{=}C(CN)CO_2H$	Positive	189.17	Peptides
HCCA-Ani	HCCA-$C_6H_5NH_2$	Positive	282.30	Peptides
NEDC		Negative		Metabolites/Lipids
Norharman	$C_{11}H_8N_2$	Both	168.17	Metabolites/Lipids
SA	$C_{11}H_{12}O_5$	Positive	224.21	Proteins
SA-Ani	SA-$C_6H_5NH_2$	Positive		Proteins

In conclusion, MALDI MSI is a versatile and valuable technique that enables the spatially resolved analysis of biomolecules in tissues and biological samples. Its ability to provide molecular maps of diverse compounds without the need for specific labelling makes it a powerful tool in advancing our understanding of biological systems, diseases, and drug responses. Continued advancements in instrumentation, methodology, and data analysis techniques further enhance the potential and applicability of MALDI MSI in various research and clinical settings.

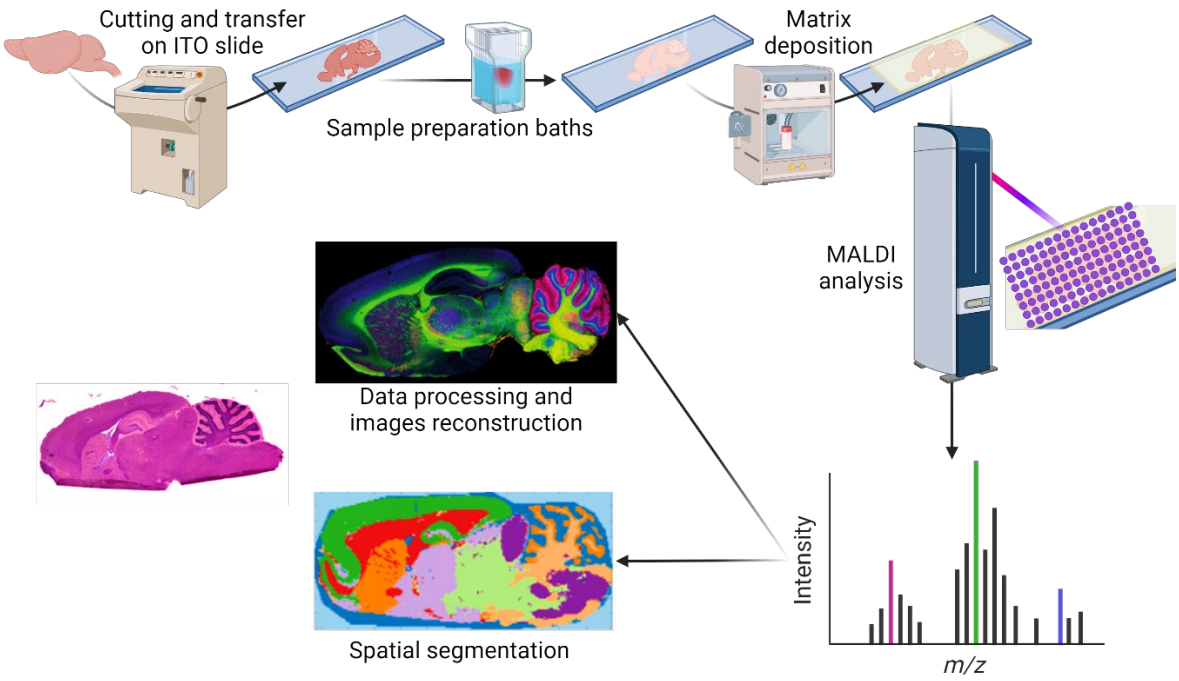


Figure 15: MALDI MSI general procedure. After preparing tissue sections and mounting them onto conductive slides, different tissue preparation washes are possible according to the molecule of interest to be analyzed before the matrix application. Mass spectra are recorded for each coordinate on the tissue. The recorded mass spectra, along with their spatial coordinates, are processed to generate molecular images that illustrate the localization of molecules within the tissue.

Machine Learning for MSI Data Processing

MALDI Imaging Data Processing Challenge

MALDI MSI is indeed a sophisticated technique that produces a three-dimensional data structure known as a hyperspectral image or data cube. This data cube integrates spatial information (x, y dimensions), representing the physical area or pixels of the sample, with spectral data (z dimension) corresponding to mass-to-charge ratio (*m/z*) values. The dimensions of an MSI image can vary considerably, ranging from hundreds of megabytes to several terabytes. Within this data, each individual pixel encapsulates a complex chemical composition and provides specific spatial information about the sample. Each pixel in the MSI dataset contains a raw spectrum showing intensity measurements over a broad range of *m/z* intervals, typically covering a wide range of 10,000 to 100,000 *m/z* bins, revealing information about potentially hundreds of different molecules. The complexity arises from the fact that each spectrum represents the observed intensities for numerous *m/z* bins, collectively characterizing the chemical profile of the sample at a particular pixel. Extracting meaningful insights from such complex datasets requires sophisticated computational approaches and bioinformatic analysis (Alexandrov, 2012).

The analysis of MALDI MSI data typically involves several stages: pre-processing and processing, also known as multivariate statistical analysis. Raw spectra often contain significant noise due to electrical background signal from the instrument and contaminants in the sample, making

pre-processing crucial. The primary objective of pre-processing is to refine and cleanse these spectra, reducing experimental variation within the dataset and preparing them for subsequent statistical analysis. Errors during pre-processing can significantly impact the final statistical results.

Pre-Processing and Processing Steps

Pre-processing is a crucial stage in data analysis, involving meticulous optimization and consideration of dataset characteristics and analytical objectives. Typically, it comprises three key steps: binning, baseline correction, and normalization:

- Binning, the initial step, involves grouping nearby *m/z* values into bins. This process reduces data dimensionality and enhances the signal-to-noise ratio, simplifying subsequent analysis while retaining relevant spectral information.
- Baseline correction is equally vital, aimed at eliminating background noise and instrumental artifacts from the spectra. By doing so, it heightens the accuracy of peak detection and quantification, ensuring that only genuine signal peaks influence further analysis.
- Normalization stands out as a fundamental pre-processing step, facilitating meaningful comparisons between spectra by adjusting intensity values. This adjustment mitigates the risk of conflating differences in overall signal intensity with true biological variations. Common normalization techniques include total ion current (TIC), median normalization, and root mean square (RMS), each with its own nuances.

Furthermore, depending on the analytical objectives, additional steps may prove beneficial. For instance, feature extraction or peak picking algorithms, as employed in this study, identify and extract discriminant peaks corresponding to molecular ions, contributing to a more refined analysis.

Following pre-processing, dimensionality reduction becomes essential for visualizing and exploring MSI datasets for biological interpretation. PCA (Principal Component Analysis) is a widely employed technique that transforms data into orthogonal components, capturing maximum variance (Bonnel et al., 2011; Fonville et al., 2012; Trim et al., 2008). Another valuable tool for visualizing spatial distribution in MALDI MSI data is non-linear t-SNE (t-distributed Stochastic Neighbor Embedding), preserving local structures while reducing high-dimensional data to 2D or 3D (Abdelmoula et al., 2018; Wang et al., 2022). NNMF (Non-negative Matrix Factorization) is also effective in pre-processing MSI data, generating a part-based representation aiding interpretation (Leuschner et al., 2019; Nijs et al., 2021). In the case of this study, SVD (Singular Value Decomposition) data reduction algorithm is the one used. SVD is a powerful technique offering a more general decomposition, which is beneficial for large datasets.

However, selecting appropriate methods poses challenges due to the lack of consensus and standardized methodologies across instrument vendors and platforms, complicating robust and reproducible data analysis in MALDI MSI and multi-omics MSI (Brunelle & Laprévote, 2012; Deininger et al., 2011).

Machine Learning and MSI Heterogeneity Interpretation

Machine learning methods are increasingly used to unlock valuable insights from data. These methods generally fall into three main categories: supervised learning, unsupervised learning, and reinforcement learning, as shown in **Figure 16**.

Supervised learning is like teaching a model using a well-organized set of examples where each input is paired with an output. This approach helps the model learn how to predict outcomes and classify new data based on what it has learned from the past. Common tools in supervised learning include linear regression, logistic regression, support vector machines, decision trees, and neural networks.

On the other hand, unsupervised learning deals with data that doesn't come with labels or predefined outcomes. Instead, it tries to find hidden patterns or structures within the data on its own. This method is particularly useful for exploring data and finding clusters or groupings without knowing in advance what to look for. Examples of unsupervised learning techniques include *k*-means clustering, hierarchical clustering, principal component analysis, and independent component analysis.

Semi-supervised learning is a middle ground that uses both labeled and unlabeled data. It's handy when you have a small amount of labeled data and a lot of unlabeled data. This approach can help improve model performance by making the most of all available data. Methods in this category include self-training, co-training, and multi-view learning.

Reinforcement learning is a bit different. It involves teaching an agent to make decisions through trial and error, based on rewards or penalties from its actions. The agent learns to maximize long-term rewards by interacting with its environment. This type of learning is great for tasks where decisions need to be made sequentially, like in robotics, game playing, or autonomous driving.

In essence, supervised learning helps with predictions and classifications using labelled data, unsupervised learning finds patterns in unlabeled data, semi-supervised learning makes the most of both labelled and unlabeled data, and reinforcement learning focuses on decision-making through feedback and rewards.

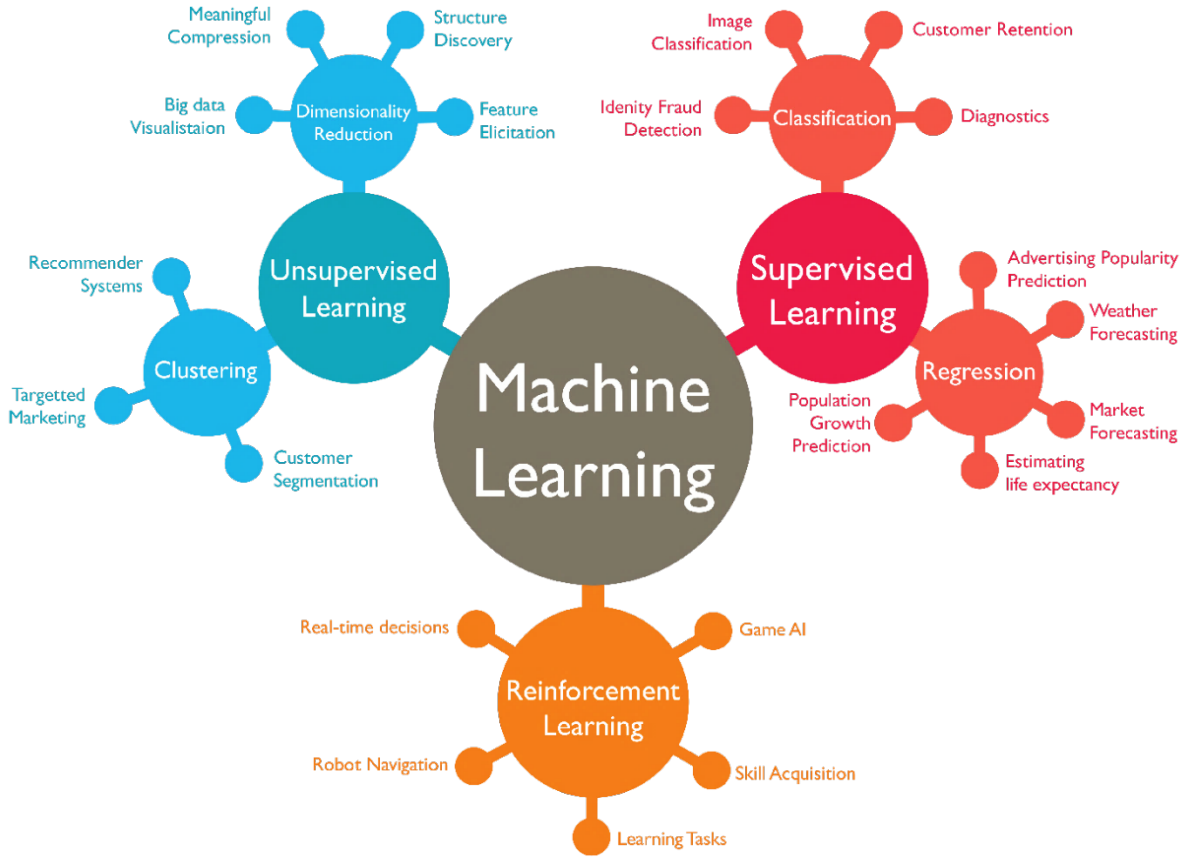


Figure 16: Machine learning types. This diagram provides an overview of Machine Learning and its three main subcategories: Supervised Learning, Unsupervised Learning, and Reinforcement Learning. Each subcategory is represented by a different color and contains examples of techniques and applications within it.

In context of heterogeneous sample MSI, machine learning and clustering are closely connected, with clustering serving as a fundamental technique within the broader field of machine learning. Clustering involves grouping similar data points based on spectral similarity, without prior knowledge of label classes. Consequently, clustering techniques are invaluable for MS image segmentation, where pixels with similar spectra are grouped to delineate regions of interest. The integration of machine learning into clustering enables unsupervised learning, allowing algorithms to identify patterns within datasets without the need for labelled training examples.

Segmenting methods address sample heterogeneity challenges, with unsupervised clustering algorithms like bisecting *k*-means, hierarchical clustering, and *k*-means commonly employed for insight extraction (**Figure 17**) (Arthur & Vassilvitskii, n.d.-a; Duda & Peter, 2012). Clustering partitions tissue samples based on molecular information, aiming to delineate distinct regions corresponding to biological features or molecular species. Mathematically, *k*-means clustering divides data into k clusters, iteratively assigning pixels to the nearest cluster centroid based on mass spectra similarity until convergence. Hierarchical clustering forms dendrograms by merging or splitting clusters based on similarity, aiding visualization of molecular similarity levels. Bisecting *k*-means, a variation, iteratively divides a single cluster into two smaller ones. Hierarchical methods are beneficial when

the optimal number of clusters is unknown, although they present challenges in omics MSI interpretation due to manual determination and spatial connectivity limitations.

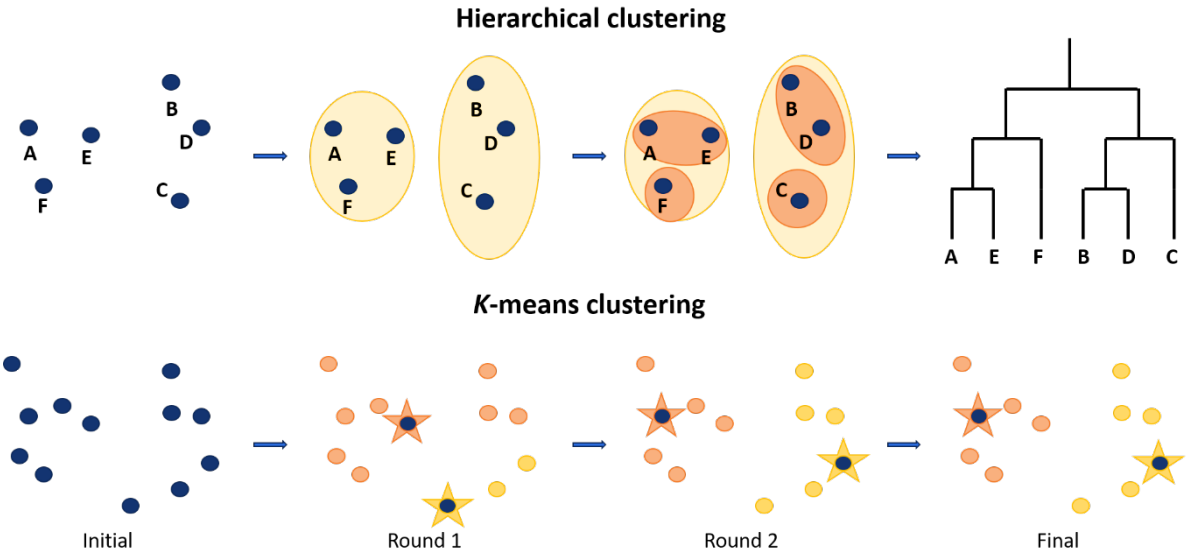


Figure 17: Hierarchical clustering and *k*-means clustering techniques for data analysis. In case of hierarchical clustering, data points (A-F) are iteratively grouped based on their similarity, forming a tree-like structure that merges clusters step-by-step. At the opposite, *k*-means clustering assigns data points to clusters based on the proximity to a central point or "centroid." The centroids adjust iteratively as data points are grouped, leading to more refined cluster assignments over time.

Choosing the right number of k-clusters is non-trivial. Too many clusters yield fragmented data, while too few may oversimplify patterns. In this way, the optimal k can be determined by evaluating intra-class and inter-class distances across different k values using recent statistical indices called criterion. Silhouette criterion (Rousseeuw, 1987) is the function used for heterogeneity assessment based on MALDI MSI in this study. The silhouette plot displays a measure of the proximity of each point in a cluster. This measure has a range (-1, 1). A value close to 1 indicates that the cluster is distant from neighboring clusters (the spectra are very compact within the cluster to which it belongs and distant from other clusters). A value of 0 indicates that the sample is very close to the decision boundary between two neighboring clusters (overlapping clusters). Negative values indicate that these samples may have been assigned to the wrong cluster. This approach minimizes intra-class distances while maximizing inter-class distances, ensuring meaningful pattern discovery without over- or under-interpretation.

Furthermore, machine learning is gaining increasing attention for various purposes in cancer analysis. Indeed, through clustering results obtained from MALDI MSI analysis, machine learning techniques can effectively identify molecular features or ion patterns that correlate with specific biological processes, disease states, or clinical outcomes. By comparing spectra from different tissue regions or experimental conditions, sophisticated machine learning can develop models for patient

stratification, prediction of treatment response, and personalized medicine. These models can identify molecular signatures that predict individual patient outcomes (Zirem et al., 2024) or response to specific treatments, thereby guiding clinical decision-making and therapy selection.

Spatial Micro-Proteomic Technic Correlated with MALDI MSI

As mentioned earlier, MALDI imaging enables the elucidation of heterogeneous molecular subpopulations. The observation of these distinct clusters implies diverse subpopulation phenotypes, linked to specific physio-pathological pathways that influence varied responses to treatment sensitivity, drug resistance mechanisms, patient recurrence outcomes, and survival prognosis. However, MALDI MSI does not directly provide in-depth identification of proteome pathways within these clusters, which could offer crucial insights into potential protein targets and resistance markers, among other factors.

MS spatial proteomic strategies on tissue enable comprehensive insights into the spatial distribution and abundance of proteins within the tissue microenvironment. These methodologies offer the mapping of protein expression within specific regions of interest within tissues. Two primary methodologies are employed: bottom-up and top-down approaches. The top-down approach directly analyzes intact proteins without prior digestion. Within the mass spectrometer, intact proteins undergo ionization and fragmentation, enabling the detection and characterization of intact protein ions along with their associated post-translational modifications (PTMs). However, due to the intricacies of intact protein spectra, top-down proteomics demands sophisticated instrumentation and data analysis techniques. Conversely, the bottom-up approach, chosen for this project, begins with the enzymatic digestion of proteins into smaller peptide fragments using proteolytic enzymes like trypsin. These peptides are subsequently separated using techniques such as liquid chromatography (LC) and analyzed via mass spectrometry with MS/MS fragmentations. Bottom-up proteomics boasts high sensitivity and is well-suited for identifying numerous peptides in complex samples.

A strategy called spatial micro-proteomics has been developed in the laboratory to enable large-scale identification of proteins from a tumor microenvironment in bottom up (**Figure 18**). This technique was extensively used in this study to deeply analyze the protein heterogeneity of tumors. The method involves performing micro digestions on a tissue area of interest, at a scale of approximately 500 µm. Following this micro digestion, the peptides are extracted using micro junction liquid extraction with appropriate solvents. The extracted peptides are then analyzed by LC-MS/MS for their identification and LFQ (Label Free Quantification) quantification. Recently coupled with EVOSEP chromatography and Bruker's TimsTOF Flex mass spectrometer, this technique

currently allows the identification of around 6000 proteins using the DIA (Data Independent Analysis) Pages method.

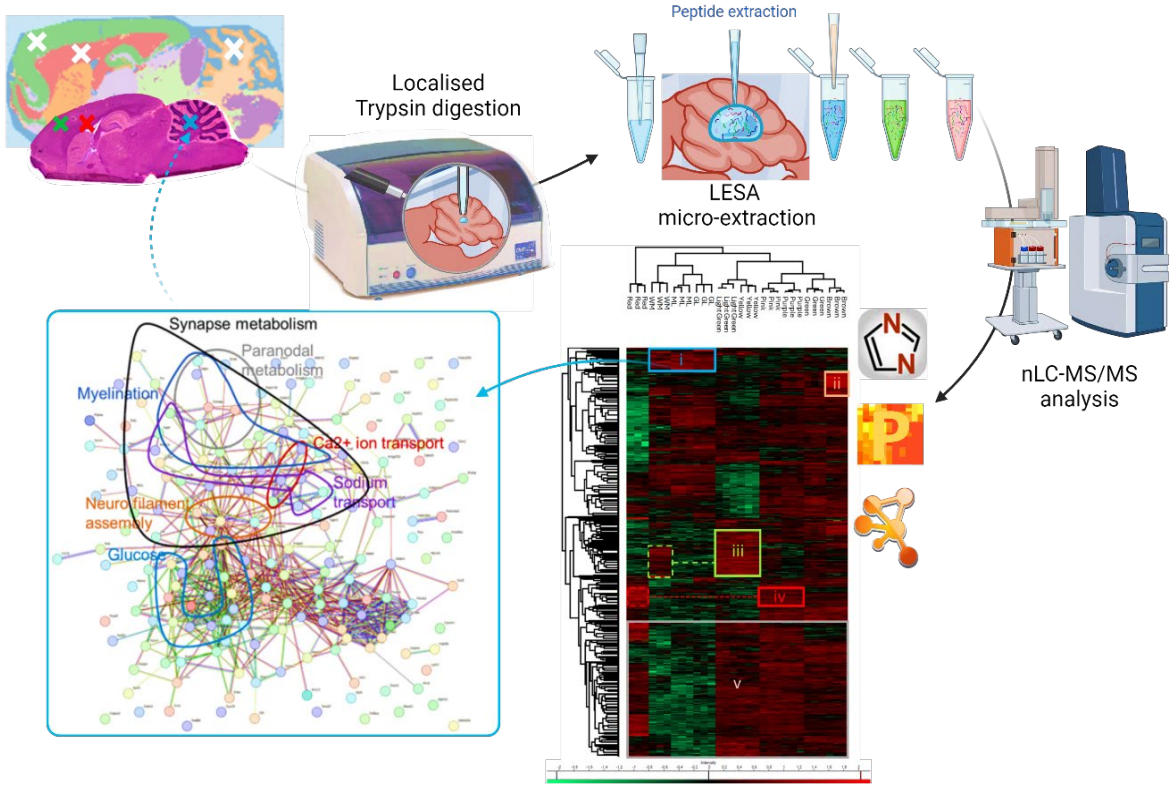


Figure 18: Spatial proteomic workflow using CHIP 1000 for trypsin localized micro digestion followed by peptide micro-extraction with LESA and nLC-MS/MS sample analysis in dia-PASEF mode. Resulting data are then analyzed through Perseus for statistical analysis and Cytoscape for ClueGO and biological pathway analysis

The limiting point when performing spatial proteomics is the extraction resolution size, which varies according to the extraction on tissue procedure used. Laser capture microdissection (LCM) is technically the better method suited. LCM systems can achieve spatial resolutions on the order of a few micrometers, allowing for the precise isolation of individual cells or small tissue structures. However, LCM remains technically somehow difficult to setup whatever the used technology, and very expensive.

SpiderMass is an advanced and cost-effective technology designed to identify approximately 5,000 proteins from a 250 µm laser spot on tissue samples through backside experiments. The technique involves firing a laser at the back of a slide where the tissue is mounted, allowing for precise tissue ablation and subsequent protein analysis. This innovative approach offers high-resolution insights into tissue proteomics. However, it's important to note that this technology is still in the developmental phase at the PRISM Laboratory.

Multiplex MALDI Immunohistochemistry and TAGmass Technology

Immunohistochemistry

Immunohistochemistry (IHC) is a widely utilized technique in biology and pathology for characterizing patient tissues by identifying specific proteins at the cellular level. This method enables the detailed mapping of the structural organization of biological tissues, allowing distinct heterogeneous areas within the tissues to be spatially identified and analyzed (Katikireddy & O'Sullivan, 2011; Stack et al., 2014). Primarily, IHC serves as a crucial tool in diagnostic pathology by detecting specific markers such as Ki67, which is associated with cell proliferation in cancer diagnosis. Additionally, IHC is invaluable in research, aiding in the study of the distribution and localization of biomarkers across various tissue types. In the field of pharmacology, IHC assists in evaluating drug diffusion and efficacy by targeting specific tissues, thereby providing insights into therapeutic effects and mechanisms.

The principle of IHC relies on visualizing specific targets within tissue cells by using antibodies linked to enzymatic reporters, chromogenic substrates, or fluorescent labels. This allows for the identification of single or multiple proteins simultaneously within the same tissue section. Multiplex IHC, which involves detecting multiple biomarkers at once, is particularly crucial in analyzing heterogeneous tissues, such as cancer tissues, where colocalization of potential biomarkers is important. However, the capacity for multiplex IHC using fluorescent microscopy is typically limited to detecting 3-5 different biomarkers simultaneously due to the broad excitation and emission spectra of fluorophores, which can cause spectral overlap (Katikireddy & O'Sullivan, 2011; Parra et al., 2017; Stack et al., 2014). Advanced techniques, such as hyperspectral or multispectral imaging, can extend this capability, allowing for the detection of up to 8 biomarkers by differentiating between the overlapping spectral properties of multiple fluorophores. However, even with these advanced methods, multiplexing remains somewhat limited (Tsurui et al., 2000). Recent innovations in IHC technology aim to overcome these limitations, including the development of novel fluorophores with narrower spectral properties and improved imaging systems with enhanced resolution and sensitivity. These advancements hold the potential to significantly expand the applications of IHC in both clinical and research settings, providing deeper insights into tissue architecture and disease mechanisms.

MALDI Immunohistochemistry

Over the past two decades, MS has emerged as a highly effective method for mapping antigen distributions in tissues, offering superior multiplexing capabilities. A significant advancement in this field is the combination of photocleavable mass-tags (Tag-Mass) with MSI, such as MALDI MSI (Lemaire et al., 2007; Stauber et al., 2010). This technology allows for the simultaneous imaging of

multiple biomarkers within biological tissues. Tag-Mass technology involves chemically modifying antibodies to include a photocleavable reporter that carries a peptide tag. When exposed to UV light, this tag is cleaved and subsequently detected by MALDI during the MSI experiment (**Figure 19**). Mass spectrometry excels in multiplexing capabilities because it can detect a wide range of mass components with a resolution of less than 1 Da. This precision makes it an ideal tool for multiplex IHC, which requires the analysis of multiple tags with distinct masses. For example, recent advances based on the Tag-mass technology by Ambergen have enabled up to 100-plex MALDI IHC using polypeptide mass-tags (Yagnik et al., 2021). However, this approach involves complex chemical synthesis processes, which can be challenging for biological applications. To address this, efforts are being made to commercialize complete probes that consist of photocleavable mass-tag linked antibodies, simplifying their use in various biological contexts.

During this thesis, the objective was to experiment with the multiplex IHC MALDI MSI technique using Tag-Mass technology. The aim was to gain a deeper understanding of its potential and identify methods to optimize and enhance the technique for future applications. While the core thesis project remained focused on specific research goals, this innovative multiplex IHC MALDI-MSI approach was applied to a range of complementary, independent projects. These exploratory applications revealed the capacity for generating highly detailed spatial maps of proteins and biomarkers within tissue sections. The use of Tag-Mass technology allowed for precise multiplexing, where multiple proteins could be detected and visualized simultaneously in a single tissue section, something not achievable with traditional IHC techniques. Moreover, the ability to detect protein expression with higher sensitivity allowed the identification of low-abundance biomarkers which would have otherwise been difficult to detect, like immune cells markers. As a result, the technique provided more accurate and reliable insights into protein distribution, localization, and expression patterns within the tissue, contributing valuable data for both biological research and potential clinical applications.

In this way, tag mass technology could be a powerful tool in the thesis subject for exploring cancer heterogeneity, offering detailed insights into the molecular diversity within tumors. This technology could also enhances our understanding of cancer biology, aiding in the identification of new biomarkers and therapeutic targets, and supports the development of more effective, personalized treatments.

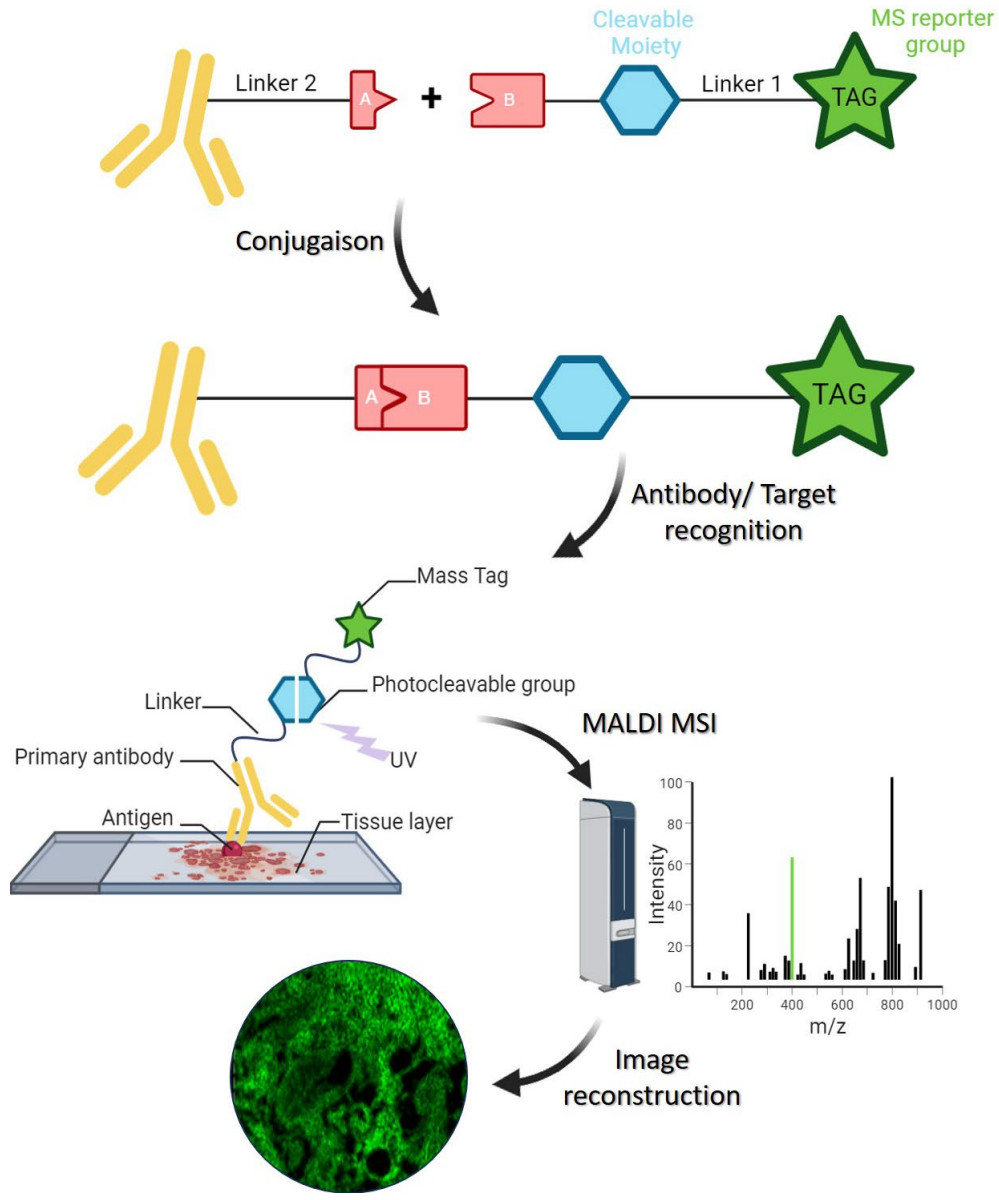


Figure 19: Tag mass workflow for analyzing protein expression in tissue samples. The first step involves assembling a bifunctional linker system composed of Linker 1 and Linker 2, which connect antibodies to a mass reporter group (TAG) and a cleavable moiety. This system is then applied to tissue samples, allowing primary antibodies to bind to specific antigens. Upon exposure to UV light, the photolytic cleavage of the linker releases mass tags, facilitating the identification of multiple targets. Following this, MALDI MSI is performed, where mass spectra are collected and analyzed to visualize the distribution of biomolecules within the tissue. The inset provides imaging of the tissue sample, highlighting the distribution of tags across the tissue.

Thesis Objectives and Results Overview

This research project aimed to address the challenge of tumor heterogeneity in cancer treatment, particularly breast cancer, by refining tumor analysis and identifying more personalized therapeutic targets. The project was structured around four main objectives, each addressing different aspects of tumor complexity and treatment response.

<u>Evaluating Proteomic Heterogeneity in Breast Cancer for Therapeutic Guidelines</u>

The first objective, described in **CHAPTER 2: Organoids for Luminal Breast Cancer Therapy Guidance Including Molecular Heterogeneity**, was to explore the intra- and inter-tumoral heterogeneity of breast cancer using MALDI MSI, which enabled the identification of distinct molecular subpopulations within tumors, offering a more detailed understanding of their complexity. In parallel, patient-derived organoids were employed to evaluate the efficacy of conventional therapies versus treatments tailored to proteomic data (**Figure 1**). The results showed that proteomic-based treatments significantly outperformed standard therapies, demonstrating superior anti-tumor efficacy. Additionally, the study identified key biomarkers of drug resistance, providing valuable insights for enhancing personalized treatment strategies. While this proteomic analysis holds great promise for tailoring optimal therapies, it is time-consuming and requires specialized expertise, presenting challenges for routine clinical use. To address these limitations, the "dry proteomics" concept, based on machine learning, was developed to streamline the process and make it more accessible for clinical applications.

Developing a Machine Learning Model to Predict Protein Pathways from Data MSI

The second objective focused on creating a machine learning model to predict biological pathways and protein information directly from lipid MSI analyses, termed "dry proteomics." By integrating lipid and protein data from MALDI MSI, this method reduced the need for separate proteomic experiments, saving time and resources (**Figure 1**). Initial validation on rat brain tissues and glioblastoma samples showed that this model could successfully predict key molecular targets, making it a promising tool for clinical applications. This advancement has the potential to significantly accelerate tumor characterization and treatment planning. Applied to breast cancer heterogeneity, the model could help uncover valuable therapeutic insights by identifying actionable molecular targets, aiding in the development of more efficient diagnosis or personalized treatment strategies. Results relative to this part are presented in **CHAPTER 3: Dry Proteomic Concept Based on Lipid MALDI MSI**.

Understanding Spatial and Temporal Heterogeneity of Breast Cancer

The **CHAPTER 4: 4D Longitudinal Proteomics Tracking of Breast Cancer Heterogeneity Community Response to Therapeutics** is focused on exploring the spatial and temporal heterogeneity of breast cancer, considering how tumors evolve over time and respond to different therapies (**Figure 1**). This involved tracking the molecular changes that occur as tumors progress and adapt to various treatments. By applying the dry proteomics workflow to clinical samples, molecular changes that occurred as tumors progress were identified, allowing for the identification of more adapted therapeutic guideline, and potential targets. Additionally, the study underscored the critical role of the tumor microenvironment in shaping both tumor progression and response to treatment.

The interactions between tumor heterogeneous clusters and surrounding immune, stromal, and other microenvironmental components significantly influence how the tumor evolves and reacts to therapies. Therefore, the research also included an analysis of tumor heterogeneity at the community level, focusing on the diverse clusters of tumor and microenvironmental implication.

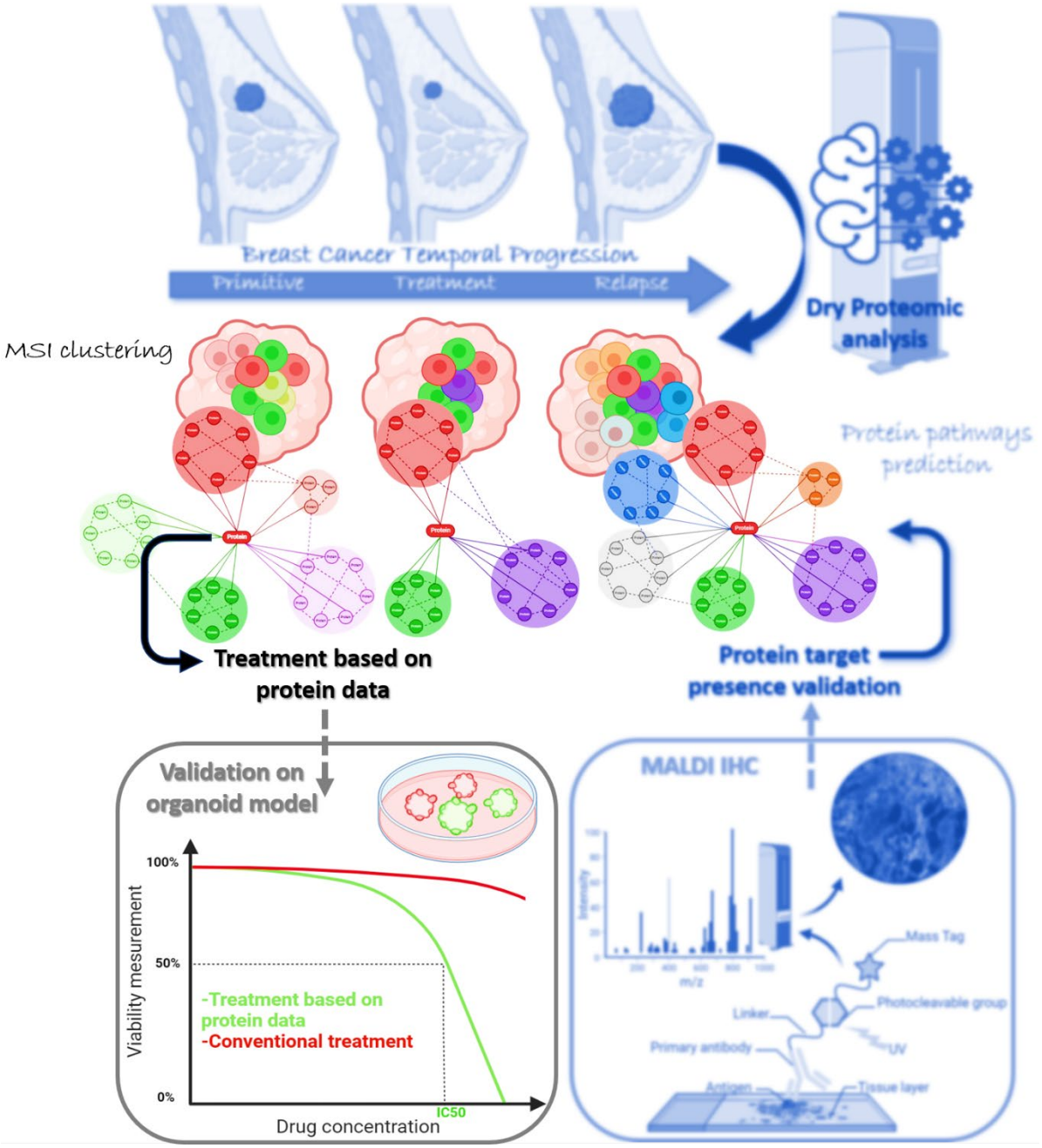
Overall, the project enhanced the analysis of tumor heterogeneity by leveraging advanced proteomic techniques, machine learning, and innovative imaging methods. These efforts offered valuable insights into cancer progression and treatment resistance, supporting the development of more personalized and effective treatment strategies for breast cancer patients. The detailed results of each objective are presented in the subsequent chapters, showcasing the potential impact of these approaches on cancer therapy. The general conclusion and future perspectives on this topic are discussed in detail in **Chapter 5 : General Conclusion and Perspectives**.

Developing a Multiplex IHC Technique Based on MALDI

The final objective aimed to experiment with the MALDI-based multiplex IHC technique using tag mass technology to gain a deeper understanding and identify potential improvements for future projects. This innovative approach was designed to enable rapid and precise identification of protein targets for therapeutic intervention, significantly improving the speed and accuracy of protein analysis. By employing mass-tagged antibodies, this technique allowed for the simultaneous detection of multiple proteins in a single tissue sample, providing a detailed spatial map of their distribution within the tumor. The results of this work demonstrated a significant enhancement in the precision and efficiency of identifying relevant protein targets and immune markers, thus enabling more personalized and targeted cancer treatments. Furthermore, this technique has the potential to serve as a powerful validation tool for confirming the presence of predicted proteins identified through dry proteomic analysis or other molecular profiling methods. By quickly verifying these proteins in clinical samples, the MALDI-based multiplex IHC technique could accelerate the transition from molecular discovery to therapeutic application, making it a promising tool for both research and clinical settings, as developed in **CHAPTER 6: Annex Contributions Involving Tag Mass Technology for MALDI IHC Applications.**

CHAPTER 2

Organoids for Luminal Breast Cancer Therapy Guidance Including Molecular Heterogeneity



CHAPTER 2: Organoids for Luminal Breast Cancer Therapy Guidance Including Molecular Heterogeneity

Introduction

It has been observed that approximately 30% of patients treated for breast cancer experience a recurrence, which may manifest as either localized or metastatic disease. Metastatic recurrences arise when cancer cells from the original breast tumor migrate to distant organs and subsequently re-emerge after a period of dormancy. Despite various treatment protocols aimed at achieving optimal outcomes, the complete eradication of all cancer cells remains challenging. This difficulty is largely due to the inherent heterogeneity among cancer cells within the tumor

Tumor heterogeneity is a major obstacle in developing effective cancer treatments. This heterogeneity manifests both inter-tumoral (between different tumors) and intra-tumoral (within the same tumor. These differences contribute to variability in treatment response, as cancer cells may have distinct biological behaviors, genetic mutations, and functional properties that affect their sensitivity to therapies. The presence of diverse cell populations within a tumor can lead to differential responses to treatment, with some cells being more resistant to therapy. This variability complicates the development of targeted treatments and contributes to the challenge of achieving sustained remission or cure. Therefore, a deeper understanding of the molecular and cellular heterogeneity at a clonal level within tumors is essential for designing more effective and personalized treatment strategies.

MALDI MSI and spatial micro-proteomic LC-MS/MS analyses, on BC luminal tissues, have been previously shown to be effective in observing the spatial heterogeneity on a tissue while characterizing in depth the molecular subpopulations (Hajjaji et al., 2021). The later present a robust strategy to identify potential actionable targets for specific tumors treatment development. In addition, the emergence of patient-derived tumor organoids (PDOs) as patient specific *in vitro* models, would allow to continue the therapeutic optimizations on faithful reconstruction of patient tumors (Campaner et al., 2020). Indeed, cell culture models are the major obstacle in drug development in the problematic of cancer heterogeneity. PDOs results in three-dimensional (3D) culture of tumor cells from BC biopsies, while conserving the histopathological characteristics of the original tumor, as well as his molecular heterogeneity (Drost & Clevers, 2018). It's, therefore, a promising model for drug screening.

The study provides experimental evidence that the molecular heterogeneity of a tumor is an important element in determining an effective treatment. To this end, the anti-tumor activity of alternative treatments based on tumor heterogeneity profiles will be compared *in vitro* with conventional chemotherapy. The model used is FFPE breast tumors with paired organoids of luminal breast cancer subtype collected from two clinical studies coordinated by Dr. Hajjaji (Oscar Lambret Cancer Center). For each tumor, the FFPE tissue is submitted to peptide MALDI MSI analysis, and spatial proteomics analysis (Hajjaji et al., 2021), to observe the spatial heterogeneity and identify actionable targets at a tumor subpopulation level. In this way, the drug discovery method used take in account the molecular heterogeneity of the entire tumor for a personalized and more effective treatment proposition. The fresh paired tumor sample of the same patient is cultivated in 3D to grow organoids for specific treatment tests, to compare the antitumor activity of therapy guided by MSI with conventional treatment, by measuring the PDOs viability.

Experimental Procedures

Chemical Products and Material

Water (H_2O), ethanol (EtOH), acetic acid, acetonitrile (ACN), DPBS (no calcium, no magnesium) and methanol (MeOH) were obtained from Thermo Fisher Scientific (Courtaboeuf, France). 99% pure trifluoroacetic acid (TFA), α-cyano-4-hydroxycinnamic acid (HCCA), aniline, formic acid (FA), Dimethyl sulfoxide (DMSO), forskolin, hyaluronidase, HEPES, iodoacetamide (IAA), nicotinamide, SB202190, Y-27632, Nomifensin, Cerulenin, Palbociclib, Bovine Serum Albumin (BSA), Paclitaxel and ammonium bicarbonate (NH_4HCO_3) were purchased from Sigma-Aldrich (Saint-Quentin Fallavier, France). Porcine Trypsin Sequencing Grade and CellTiter®-GLO 3D cell viability assay was from Promega (Charbonnières, France). A83-01 was obtained from Tocris. Advanced DMEM/F1, GlutaMax 100X, HEPES, Penecillin/Streptomycin and Promocin were from Invitrogen. B27 supplement , enzyme 1X TrypLE™ express and fetal bovine serum were purchased from Gibco. Neuregulin, EGF, FGF7, FGF10 and Noggin were from Peprotech. DL-Dithiothreitol (DTT) was purchased from VWR Life Science.

Tissues were cut on a microtome (Leica Microsystems, Nanterre, France). Indium Tin Oxide slides were purchased from LaserBio Labs (Valbonne, France), whereas the poly-lysine coated slides were from Epredia™ (Braunschweig, Germany). The MALDI matrices and the trypsin were deposited on the tissue sections using the HTX M5-Sprayer™ (HTX Technologies, Carboro, NC, USA). Mass spectrometry imaging analyses were performed using the MALDI-TOF Rapiflex Tissuetyper (Bruker Daltonics, Bremen, Germany) equipped with the Smart Beam 3D laser. Spatial proteomic analysis were carried out through the utilization of chemical printer (CHIP-1000, Shimadzu, Kyoto, Japan) and the TriVersa Nanomate device (Advion Biosciences Inc, Ithaca, NY, USA). Samples were dried in a SpeedVac

(SPD13DPA, Thermo Fisher Scientific, Waltham, Massachusetts, USA). nLC-MS/MS analysis were performed with TimsTOF Flex (Bruker) coupled to an EVOSEP One (EVOSEP).

Sample Preparation

Tumor BC tissues used for this study were collected from patients from the Oscar Lambret Center in Lille the study was approved by the local research committee of Oscar Lambret Cancer center and a French Ethical Committee (study IdRCB 2021-A00670-41). The. All patients gave their informed consent. The patients (n=4) participating in the present study were women with luminal breast cancer, whose FFPE tissue was available following surgery or a fine needle biopsy. The FFPE tissues obtained were cut with a microtome as follows: a polylysine glass slide with 5 µm thick tissue for HPS staining and pathological analysis; an ITO conductive slide with 8 µm thick tissue, which was fixed with 2% ovalbumin, for imaging molecular by mass spectrometry; and a polylysine glass slide with 20 µm thick tissue for spatial proteomics .

Hemalum-Phloxin-Safran (HPS) Coloration

The HPS coloration of the slides was carried out in an automated way via instrument, using this following process (**Table 6**):

Table 6: HPS coloration steps.

Step	*Solution*	*Time*
Drying		15 min
Paraffin removal	Xylen	5 min
	Xylen	5 min
Rehydration	EtOH 100%	5 min
	EtOH 95%	5 min
	H2O	2 min
Nucleus coloration	Hemalum	5 min
Rince	H2O	2 min
Differentiation	HCl 0,8%	10 sec
Rince	H2O	2 min
Cytoplasm coloration	Phloxin	20 sec
Rince	H2O	5 min
Dehydration	EtOH 95%	1 min 30
	EtOH 100%	2 min 30
Collagen coloration	Safran	2 min
Rince	EtOH100%	20 sec

	EtOH 100%	20 sec
Lightening	Xylen	1 min

The histopathological analysis were then performed by pathologist at the Oscar Lambret Center.

Peptide MALDI Mass Spectrometry Imaging

The slides intended were subjected to a deparaffinization step, using two successive baths of xylene during 5 min; followed by a rehydration step, using three baths of decreasing degree of ethanol: 2x90°, 1x30°, and two baths of 10 mM NH4HCO3 buffer during 5 min each. An antigen unmasking step is necessary so that the tissue becomes more accessible for the enzymatic digestion that follows. For this, the slides are soaked in a bath of 20 mM Tris buffer at pH 9 for 20 minutes at 90°C, then in two washing baths containing 10 mM of NH4HCO3 for 1 minute, before being dried under vacuum.

The tryptic digestion was performed by applying trypsin (40µg/mL, dissolved in NH4HCO3 50mM) via an HTX M5-SprayerTM. Once the enzyme was deposited, the slides were incubated overnight in a box containing MeOH/H2O placed in an oven set at 56°C. The slides are then dried under vacuum the next day.

An HCCA-aniline matrix was deposited by the HTX M5-Sprayer automaton. Briefly, 43,2 µL of aniline were added to 5 mL of a solution of 10 mg/mL HCCA dissolved in ACN/TFA0,1% (7:3, v/v).

Slides were analyzed on a MALDI-TOF Rapiflex instrument, equipped with a Smart Beam 3D laser. The spectra were obtained in the positive delayed extraction reflectron mode analysis, with a mass range of 360-3200 *m/z*, and averaged from 200 laser shots per pixel for a spatial resolution of 60µm. The laser energy was set around 40 %. The voltages of the ion source were 20 kV and 11 kV for the lens.

MALDI MSI Data Processing and Analysis

Dry proteomic pipeline was used for MSI clustering. The raw MALDI MSI data were initially converted into the imzML format (Römpp et al., 2011a) using SCiLS lab software . Subsequently, the imzML converter, version 1.3.3, was used to import these datasets into MATLAB R2019a. It's worth noting that MSI data is characterized by high dimensionality, often reaching sizes of up to 100 GB per image. This size makes it impractical to analyze such data. To overcome this problem and to prevent data loss through peak list generation, SVD data compression was implemented as a pre-processing step prior to segmentation. The *k*-means++ algorithm, implemented as the '*k*-means' function in the MATLAB Statistics Toolbox, was used for the segmentation process. *K*-means++ provides improved centroid initialization, which improves the quality of the clustering (Arthur & Vassilvitskii, n.d.-a). The

cosine distance metric was used to calculate the cosine angle between two spectra to quantify similarity. For visualization, the pixels of each cluster are uniformly assigned a specific color to facilitate the creation of a segmentation map. This map delineates the cluster, or region of interest, to which each pixel spectrum belongs. The silhouette criterion was used to estimate the correct number of clusters. Once the number of clusters was determined, the silhouette plot method was used to assess the stability of the clusters. The silhouette plot gives a measure of the proximity of each point in a cluster. This measure has a range of (-1, 1). A value close to 1 indicates that the cluster is distant from neighboring clusters (the spectra are very compact within the cluster to which it belongs and distant from other clusters). A value of 0 indicates that the sample is very close to the decision boundary between two neighboring clusters (overlapping clusters). Negative values indicate that these samples may have been assigned to the wrong cluster (Rousseeuw, 1987). he silhouette plot was calculated using the silhouette function in MATLAB. Each centroid within these clusters is then carefully exported in CSV format, ready for further in-depth analysis and exploration.

Spatial Proteomics Extraction

The different clusters identified by the segmentation process were submitted to spatially resolved proteomics. Each cluster was analyzed in triplicate from the same tissue section as describe bellow. A localized digestion was carried out by deposing a trypsin solution (40 µg/mL in NH_4HCO_3 50mM), on a region of 500 μm^2 of tissue (4 x 4 droplets of 200 µm in diameter), using CHIP-1000. The deposition method comprises approximately 1205 cycles per digestion spot, i.e., 3 hours of deposition, with a drop volume of 150 pL. Finally, each spot was digested with 0.112 µg of trypsin. Following the micro-digestion, each spot was extracted by liquid micro-junction using the TriVersa Nanomate device, with LESA (Liquid Extraction and Surface Analysis) parameters (Quanico et al., 2013; Wisztorski et al., 2016). The tryptic peptides were extracted by performing 2 consecutive extraction cycles for three different solvents mixtures (TFA 0.1%; ACN/0.1% TFA (8:2, *v/v*); and MeOH/0.1% TFA (7:3, *v/v*)) for a total of 6 extractions. For each cycle, 2 µL of solvent was drawn into the tip of the pipette, of which 0.8 pL was brought into contact with the surface. 15 back and forth movements were performed to extract the peptides before collecting the solution in a recovery tube. All extracts were pulled in one tube and 50 µL of ACN were finally added before drying the samples in a SpeedVac. The samples were then stored at -20°C prior to nLC-MS/MS analysis.

NanoLC-MS/MS Analysis

All sample analysis was performed on a timsTOF fleX mass spectrometer online coupled to an Evosep One nano-flow liquid chromatography system. Peptides were separated using an 8 cm x 150 µm C18 column with 1.5 µm beads and the 60 samples per day method from Evosep One. The mobile phases comprised 0.1% FA in water as solution A and 0.1% FA in ACN as solution B. To perform DIA

analysis in PASEF mode (Meier et al., 2020), one MS1 scan was followed by 10 dia-PASEF scans from *m/z* 100 to 1700. The ion mobility range was set to 1.42 and 0.65 V.s/cm^{-2}. The accumulation and ramp times were specified as 100 ms. As a result, each MS1 scan and each MS2/dia-PASEF scan last 100 ms plus additional transfer time, and a dia-PASEF method with 22 dia-PASEF scans has a cycle time of 1.06s. The mass spectrometer was operated in high sensitivity mode, with a collision energy ramped linearly as a function of the ion mobility from 59 eV at $1/K_0$=1.6Vs.cm^{-2} to 20 eV at $1/K_0$=0.6Vs.cm^{-2}. The ion mobility was calibrated with three Agilent ESI Tuning Mix ions (*m/z*, $1/K_0$: 622.02, 0.98 V.cm^{-2}, 922.01, 1.19 V.cm^{-2}, 1221.99, and 1.38 V.cm^{-2}).

Data Analysis

DIA-NN version 1.8.1 was used to search DIA raw files and dia-PASEF files. A Human library was generated with the software parameters set as following: complete proteome of Homosapiens from UniProt database (Release January 2024, 92958 entries), Trypsin protease with 2 missed cleavages and a maximum number of variable modification at 3, methionine oxidation as variable, peptide length range from 7 to 30, precursor charge range from 1 to 4, precursor *m/z* range comprised between 100 and 1700, fragment ion *m/z* range between 200 and 1700, 0.1% precursor FDR, protein inference set on 'genes', neural network classifier on single-pass mode, quantification strategy set on robust LC (high accuracy), RT-dependent cross-run normalization, and library generation fixed on the 'IDs, RT & IM profiling' ruban. Samples were interrogated according to the resulting Rattus library with the same options. Statistical analyses were carried out using Perseus software v2.0.5.0. ANOVA tests were performed with p-value ≤ 0.01 to be statistically significant and generate heat maps of differentially expresses proteins across sample. STRING (Szklarczyk et al., 2015) and Gene Ontology analysis were performed using ClueGO (Bindea et al., 2009) with GO term database, on Cytoscape v3.10.2 (Shannon et al., 2003).

Drug Targeting

The draggability level of the targets was classified according to the Illuminating the Druggable Genome Knowledge Management Center (IDG-KMC) definitions. These definitions categorize targets into four target development levels (TDLs):

→ Tclin: Targets with activities in DrugCentral (i.e., approved drugs) and a known mechanism of action. DrugCentral (http://drugcentral.org) is an online drug information resource created and maintained by the Division of Translational Informatics at University of New Mexico in collaboration with the IDG Illuminating the Druggable Genome (IDG) (https://druggablegenome.net/index) (Sheils et al., 2020). DrugCentral provides information on active ingredients, chemical entities, pharmaceutical products, the mode of action of drugs, indications, and pharmacologic action. Data is monitored on FDA, EMA, and PMDA for

new drug approval on regular basis. Supported target search terms are HUGO gene symbols, Uniprot accessions and target names, and Swissprot identifiers. The WHO anatomical therapeutic chemical (ATC) classification was used to categorize drugs.

- → Druggability
- → Tchem: Targets with activities in ChEMBL or DrugCentral that meet the activity thresholds specified at [druggablegenome.net/ProteinFam] (https://druggablegenome.net/ProteinFam).
- → Tbio: Targets with no known drug or small molecule activities that meet the activity thresholds and criteria detailed at [druggablegenome.net/ProteinFam] (https://druggablegenome.net/ProteinFam).
- → Tdark: Targets with virtually no known drug or small molecule activities that meet the criteria defined by IDG-KMC.

Organoid Culture

Each fresh biopsy tissue was digested 2h at 37°C in 2mL Hank's Balanced Salt Solution (HBSS, Gibco) containing antibiotic and anti-fungal (1X Penicillin/Streptomycin, 1X Amphoteromicin), with 1 mg/mL collagenase type IV (Sigma) and 5 U/mL hyaluronidase (Sigma). In order to help the digestion, the later mixture was mixed every 15 minutes. After digestion, 7 mL of HBSS with antibiotics were added to the cell suspension before to be filtered over a 100µm filter (Dutcher). The result was centrifugated et 300 G for 5 minutes at 4°C. In case of presence of erythrocytes, visible through a red pellet, a lyse was performed with 1 mL of red blood cell lysis buffer (RBC, Invitrogen) for 5 minutes at room temperature. The mixture was then completed with 6 mL of cold PBS and centrifugated at 300 G for 5 minutes at 4°C. The cell pellet was resuspended in a reduced growth factor solubilized basement membrane matrix for organoid culture (Matrigel , Corning) and plated as a 30 µL drop in 24-well plate. After 30 minutes in the incubator for Matrigel solidification, 500 µL of medium were carefully pour. The medium culture was composed of Advanced DMEM supplemented with 1X Glutamax, 10 mM Hepes, 1X Penicillin/Streptomycin, 1X Amphoteromicin, 50µg/mL Primocin, 1X B27 supplement, 5 mM Nicotinamide, 1.25 mM N-Acetylcystein, 250 ng/mL R-spondin 1, 5 nM Heregulinβ-1, 100 ng/mL Noggin, 20 ng/mL FGF-10, 5 ng/mL FGF-7, 5 ng/mL EGF, 500 nM A83-01, 500 nM SB202190 and 5µM Y-27632.

Organoids were multiplicated when confluent. To do so, cold PBS was used to harvest tumoroids from the Matrigel and collected in a 15 mL falcon, pre-coated with PBS – BSA 1%. The resulting solution was centrifugated at 300 G for 5 minutes at 4°C. The pellet was digested with 1 mL of TrypLE solution during 5 minutes at 37°C. Then, 6 mL of PBS were added to the mixture before a centrifugation at 300 G for 5 minutes at 4°C. The pellet was finally resuspended with Matrigel and reseeded as explained above.

Organoid Protein Analysis

Organoids were collected in triplicate before treatment. The tumor sections and organoid pellets were lysed with RIPA buffer (150 mM NaCl, 50 mM Tris, 5 mM EGTA, 2 mM EDTA, 100 mM NaF, 10 mM sodium pyrophosphate, 1% NP40, 1 mM PMSF, and 1X protease inhibitors) for total protein extraction. The samples underwent three cycles of 30-second sonication at 50% amplitude on ice. Cell debris was removed by centrifugation at 16,000 × g for 10 minutes at 4°C. The supernatants were collected, and protein concentrations were measured using a Bio-Rad Protein Assay Kit according to the manufacturer's instructions. To normalize the protein quantities of the organoids and tumors, 100 µg of each sample was used for protein digestion and subsequent shotgun proteomics analysis.

Protein digestion was carried out using the FASP method. First, a reduction solution (100 mM DTT in 8 M urea, 0.1 M Tris/HCl, pH 8.5) was added to the sample and incubated at 95°C for 15 minutes. The protein solution was then applied to 10 kDa Amicon filters, supplemented with 200 µL of UA buffer, and centrifuged at 14,000 g for 30 minutes. This step was repeated with another 200 µL of UA buffer. Next, 100 µL of alkylation solution (0.05 M iodoacetamide in UA buffer) was added, and the mixture was incubated in the dark for 20 minutes, followed by centrifugation at 14,000 g for 30 minutes. The filters were then washed three times with 50 mM ammonium bicarbonate (AB) solution, each followed by centrifugation at 14,000 g for 30 minutes. For digestion, 50 µL of LysC/Trypsin solution (20 µg/mL in AB buffer) was added, and the samples were incubated overnight at 37°C. The resulting peptides were collected by centrifugation at 14,000 g for 30 minutes. The filters were washed twice with 100 µL of AB buffer, each time followed by centrifugation at 14,000 g for 30 minutes. Finally, the eluted peptides were acidified with 10 µL of 0.1% trifluoroacetic acid (TFA) and dried under vacuum.

Final samples were analyzed through nLC-MS/MS in DIA-PASEF mode.

Organoid Response to Drugs

For the organoid culture and drug response analysis, an equal number of organoids was dissociated with cold PBS. The pellet was digested with TrypLE solution (Gibco) for 5 minutes at 37°C. The organoids were then diluted in HBSS and passed through a 100 µm filter (Dutcher) to remove larger tumoroids. The resulting organoids were centrifuged at 300 g for 5 minutes and resuspended in 2% Matrigel/organoid culture medium (3,000-5,000 tumoroids/mL). For the drug response analysis, 100 µL of the organoid's suspension was placed into wells of 96-well plates coated with 1.5% agarose. The organoids were allowed to form over 72 hours and then treated for 7 days before performing the viability test. Cell viability was assessed using CellTiter-Glo 3D (Promega) according to

the manufacturer's instructions, and the results were normalized to the controls. Drug concentrations ranged from 0.01 µM to 100 µM across five concentrations, with DMSO controls included. After 7 days, 100 µL of CellTiter-Glo 3D reagent (Promega, Madison, WI, USA) was added to each well, and the plate was shaken at room temperature for 25 minutes. Luminescence was measured using a TriStar2 S LB 942 Multimode Microplate Reader, and the data were analyzed using GraphPad Prism 6.

Results

This chapter aimed to highlight the project's key issues related to breast cancer heterogeneity by examining various hypotheses:

- Is combining MALDI MSI with spatial proteomic analysis an effective approach for addressing the complexity of breast cancer heterogeneity?
- Can spatial proteomic analysis of diverse breast cancer tissues lead to more personalized drug treatments compared to conventional methods?
- Can targeting the cooperative networks among tumor clusters enhance treatment outcomes and reduce the risk of resistance?
- Are organoids a sufficiently reliable in vitro model for validating the efficacy of proposed treatments and correlating findings with the patient's tumor?

The strategy employed for studying the proteomic and spatial heterogeneity of tumors involves combining MALDI MSI with spatial proteomics to identify and characterize tumor clones. The study utilized FFPE biopsies from patients with early stage luminal breast cancer, for whom derived organoid were available. Serial tissue sections were prepared for each tissue. One section was designated for HPS staining, which allowed the characterization of tumor regions through detailed histological analysis performed by a pathologist. Another section was subjected to MALDI MSI analysis to generate a peptide map of the tumor, which was then processed dry proteomic clustering pipeline. This processing defines the optimal segmentation of the section by assigning colors based on spectral similarity. Only the tumor regions identified by histological analysis were selected for further segmentation. This segmentation distinguished proteomic sub-populations within the tumor, corresponding to different tumor clones exhibiting intra-tumoral heterogeneity. These clones were further analyzed to uncover potential therapeutic targets and identify tailored treatments. Additionally, organoids were used to validate the proposed treatments according paired tissue proteomic data (**Figure 20**). This integrated approach ensured that the spatial proteomic data correlates with the in vitro organoid models, providing a robust

framework for identifying and validating personalized therapeutic strategies for luminal breast cancer.

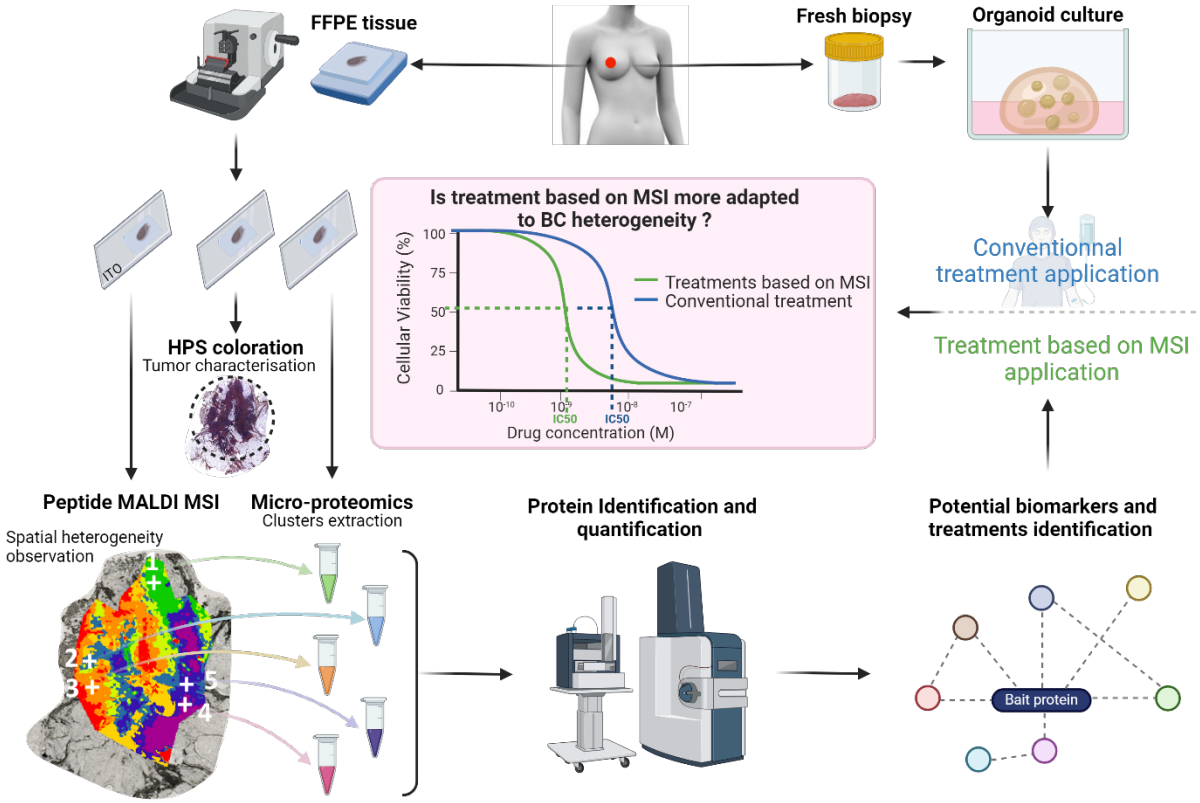


Figure 20: Theragnostic approach for breast cancer treatment, integrating MSI to address tumor heterogeneity. The workflow begins with tissue sampling and imaging, followed by MSI analysis to map molecular variations in the tumor. The approach leverages proteomics to identify key protein interactions, leading to more precise, tailored cancer treatments. These data guide personalized treatment strategies. The effectiveness of MSI-based treatments are compared to conventional methods on tumor paired organoids, by measuring organoid viability.

Intra-Tumor Heterogeneity Observation Through MALDI MSI

The described workflow was applied to four FFPE tumor samples to achieve a spatially resolved, unsupervised, and unlabeled visualization of breast cancer heterogeneity, along with in-depth proteomic profiling.

MALDI MSI on-tissue was first performed to map peptide compositions on patient tumor tissue. The resulting spectral data were then preprocessed with RMS normalization and SVD data compression before to be clustered using the *k*-means ++ method. Image clustering process assigned color-coded groups to different tumor areas based on the similarity of their proteomic signatures. Silhouette criterion was also used to predict the optimal number of clusters corresponding to tissue heterogeneity. Imaging revealed distinct proteomic clones, as demonstrated by the representative MALDI MS images of a primary tumor in **Figure 21**. Overlaying these results with histological analyses enabled the selection of clonal tumor regions for spatial proteomics analysis.

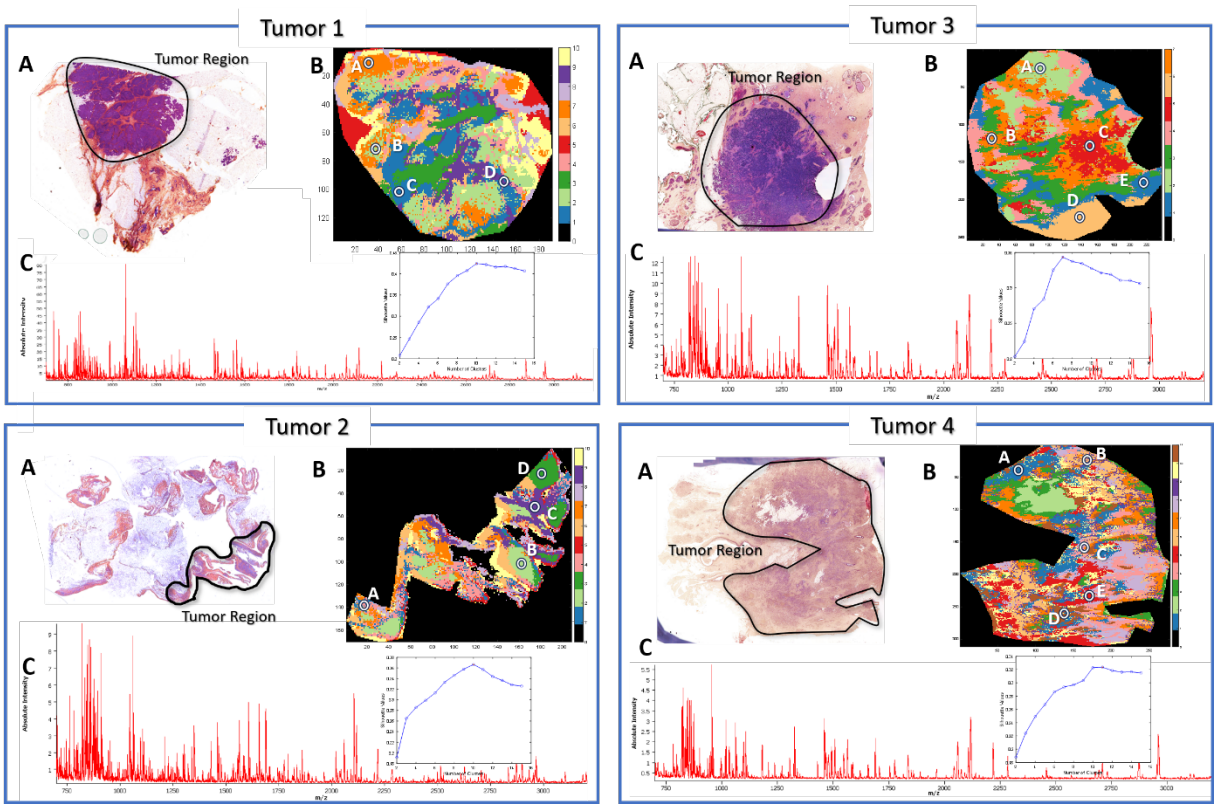


Figure 21: Intra-tumors analysis of tumor 1, 2, 3 and 4 with A) histologic coloration, B) MS image segmentation with spatial proteomic regions of interest and C) peptide MSI mean spectra.

Clonal Proteome Analysis and Drug Target

Proteomic analysis of each tumoral clone confirmed their distinctiveness, as previously observed with MSI (**Figure 21**). In the clonal proteome dataset for each tumor, proteins were either shared among clones, specific to individual clones, or differentially expressed (**Figure 22**). These proteins may be associated with biological pathways intrinsic to breast cancer stage, tumor microenvironment, or progression. Consequently, these proteins could be potential targets for drug development.

For example, in tumor 1, MSI identified three distinct tumoral clones (A,B and D) and a stroma clone C, which were then analyzed using spatial proteomics to investigate their protein expression profiles. This comprehensive analysis identified a total of 5554 proteins across the clones. Of these, 4685 proteins were shared among all three tumoral clones, indicating a core set of proteins common to the tumor. However, significant differences were also observed: 18 proteins were uniquely expressed in clone A, 21 in clone B, and 13 in clone C. These findings underscore the proteomic heterogeneity among the clones. To further elucidate these differences, a multiple sample ANOVA test was employed to identify differentially expressed proteins among the tumoral clones, using a stringent p-value < 0.01. This analysis revealed 1221 proteins with significant expression

differences across the clones. The resulting heatmap (**Figure 22B**) clearly displayed distinct clusters of over-expressed proteins, highlighting the unique and shared proteomic signatures of each clone.

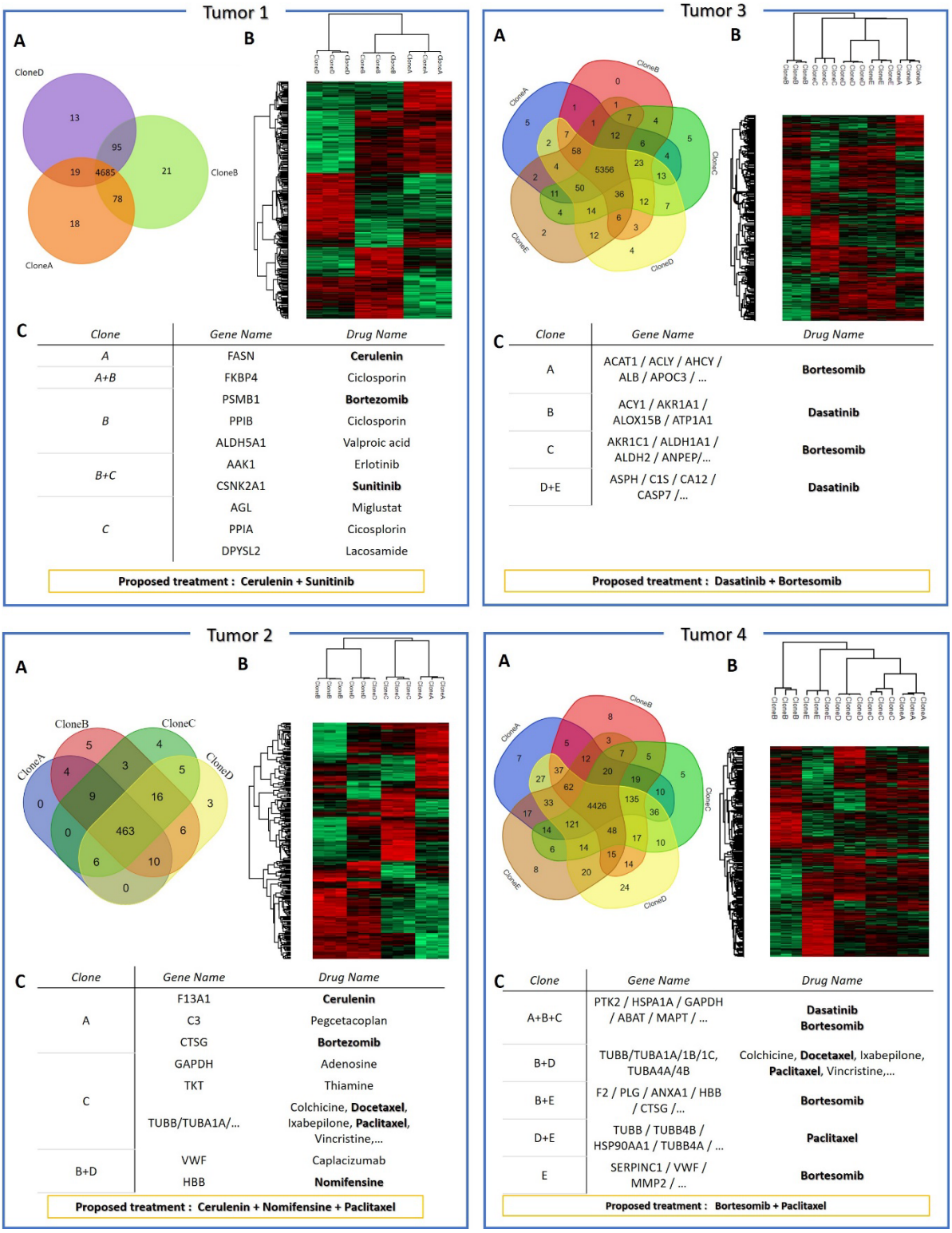


Figure 22: Intra-tumor spatial proteomic analysis with A) Venn diagram, B) Heatmap of over-expressed proteins after ANOVA p-value < 0.01, and C) potential protein target and associated drug.

Subsequent ClueGo analysis of these protein clusters identified the biological pathways associated with the over-expressed protein clusters (**Appendix A, Figure 62**). Interestingly, each over-expressed protein cluster was involved in distinct biological pathways, underscoring the tumoral heterogeneity and the complexity of therapeutic targeting. For instance for tumor 1 (**Figure 23**), the pathway related to the packaging of telomere ends was actively involved in clone A, driven by specific over-expressed proteins. The telomerase enzyme, which adds repetitive nucleotide sequences to the ends of chromosomes, is known to be highly expressed in breast cancer cells and is associated with poor sensitivity to therapies (Xu & Goldkorn, 2016). This makes the telomere ends pathway a potential target for cancer therapies, as inhibiting telomerase activity could limit the replicative potential of cancer cells (Judasz et al., 2022). In clones B and C from tumor 1, integrin surface and extracellular matrix (ECM) proteoglycans pathways were prominently involved. Integrins are crucial surface adhesion receptors that mediate interactions between the ECM and cells, playing essential roles in cell migration, adhesion, and the maintenance of tissue homeostasis. Aberrant integrin activation has been shown to promote initial tumor formation, growth, and metastasis. Recent evidence indicates that integrins are highly expressed in numerous cancer types and perform multiple functions in tumorigenesis, including influencing cell survival, proliferation, and migration. Consequently, integrins have emerged as attractive targets for the development of cancer therapeutics, with various integrin inhibitors currently being explored in clinical trials (Liu et al., 2023).

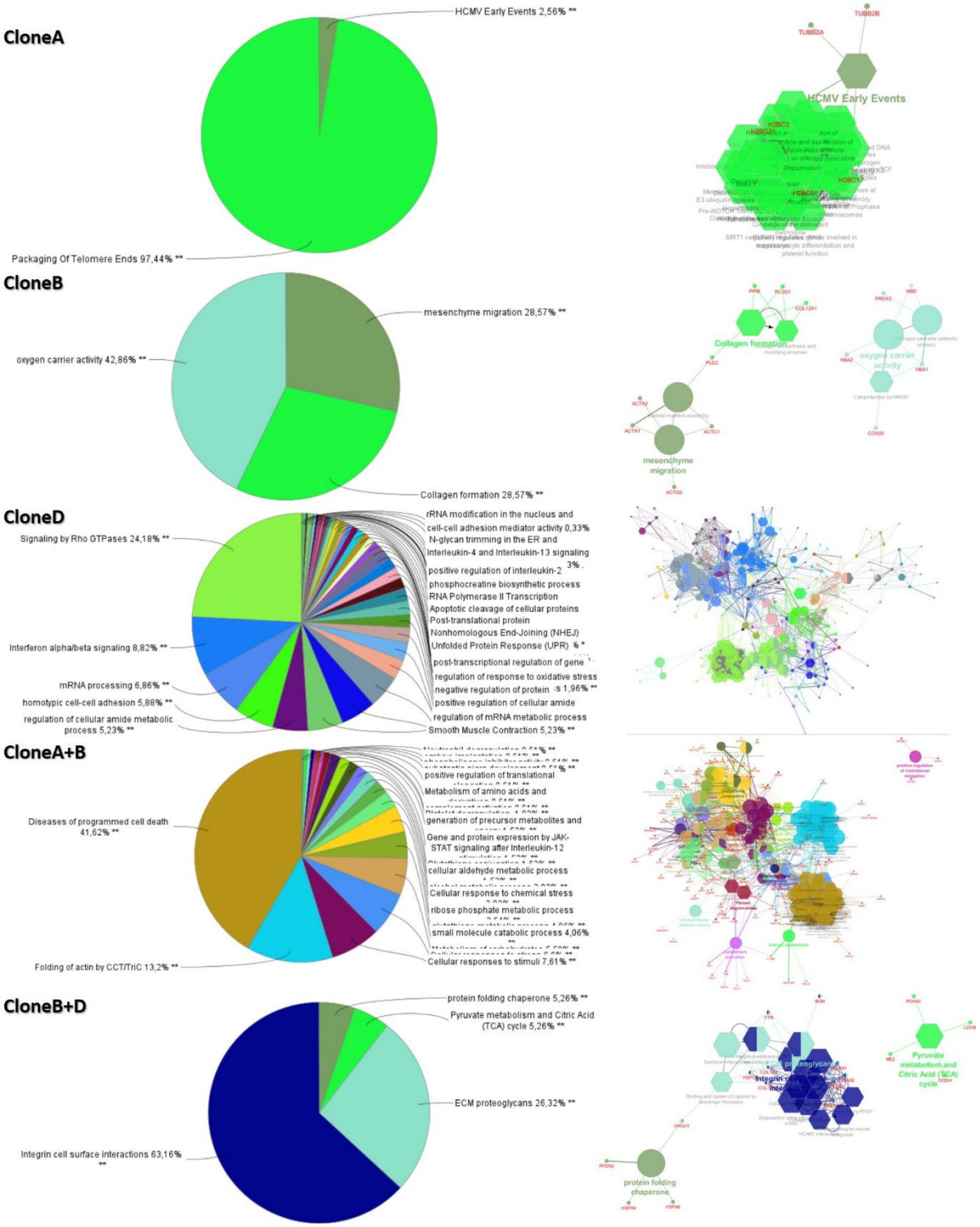


Figure 23: Biological pathways associated with over-expressed clusters involved in tumor 1 clones according to ClueGo analysis.

These findings highlight the complex proteomic landscape within tumors and emphasize the need for targeted therapeutic strategies that address the specific biological pathways active in different tumor clones. By understanding the unique proteomic and pathway signatures of each clone, more effective and personalized treatments can be developed for breast cancer patients. Through enrichment analyses correlated with the DrugCentral database, targetable proteins were

identified and associated with drugs in each tumor clonal proteome dataset. Relevant candidate clonal tumor targets and associated drugs were selected and combined as potential treatments. This proteomic data-driven approach ensures that treatments are more suitable and adapted to the tumor, impacting each clone effectively.

For instance, in tumor 1 (**Figure 22**), the FASN protein was strongly enriched in clone A. FASN is an enzyme critical for de novo lipogenesis, which is often upregulated in cancer cells to support rapid growth and proliferation. Cerulenin, a drug that inhibits FASN, could effectively target this metabolic pathway, disrupting tumor growth in clone A. Similarly, protein PSMB1 was identified in clone B. PSMB1 is a component of the 20S proteasome, which is involved in protein degradation and regulation of various cellular processes. Bortezomib, a proteasome inhibitor, can block this pathway, leading to the accumulation of pro-apoptotic factors and induction of cell death in clone B. Protein CSNK2A1 was over-expressed in both clones B and C. CSNK2A1 is a serine/threonine kinase involved in regulating various cellular functions, including cell cycle progression and apoptosis. Sunitinib, a multi-kinase inhibitor, can target CSNK2A1 and other kinases, offering therapeutic benefits for both clones B and C. Hence sunitinib should be effective in both clone B and C, the drug was preferred to bortezomib which impacted only clone B. Therefore, a combination of cerulenin and sunitinib should effectively address the heterogeneity of tumor 1 (**Table 7**).

Interestingly, each treatment proposed following this strategy is markedly different from the conventional treatment plans typically recommended by oncologists (**Table 7**). This innovative approach, based on the detailed proteomic profiles of tumor clones, offers a more tailored and potentially more effective alternative to standard therapies. By focusing on the specific molecular characteristics of each clone, this method aims to overcome the limitations of one-size-fits-all treatments and address the unique challenges posed by tumor heterogeneity.

Table 7: Patient breast tumor treatments according to oncologist versus proteomic analysis.

Tumor	*Clone analyzed*	*Conventional treatment*	*Proposed treatment based on tumor heterogeneity*
Tumor 1	3	Paclitaxel	Sunitinib + Cerulenin
Tumor 2	4	Palbociclib	Cerulenin + Paclitaxel + Nomifensin
Tumor 3	5	Epirubicin + Cyclophosphamide + Paclitaxel	Dasatinib + Bortezomib
Tumor 4	5	Epirubicin + Cyclophosphamide + Paclitaxel	Paclitaxel + Bortezomib

Treatment Guideline Comparison on Organoids

To validate the proposed tumor treatments based on MSI and spatial proteomics, tumor-paired organoids were treated with either conventional therapies or those derived from proteomic data. The response of the organoids to these treatments varied depending on the case.

Following the example of Tumor 1, proteomic analysis of paired organoids before treatment confirmed the presence of the protein targets FASN, PSMB1, and CSNK2A1, which were previously identified in Tumor 1 tissue through spatial proteomic analysis. FASN was highly overexpressed in both the tissue from clone A and in the organoid proteomic data, PSMB1 was enriched in clone B and the organoid, while CSNK2A1 was enriched in clone D and the organoid (**Figure 24A**). These findings validate the organoid model's reliability in representing the primary tumor and support the accuracy of the drug response results. Interestingly, the treatment regimen proposed based on proteomic data demonstrated significantly greater effectiveness against cancer cells compared to conventional therapies (**Figure 24B**). The IC50 values, which indicate the concentration of drug required to inhibit 50% of cancer cell growth, were notably higher for conventional treatments, suggesting they were less effective. Specifically, the IC50 for paclitaxel was 37.728 µM (**Figure 24C**), indicating a high concentration was needed to achieve therapeutic efficacy. In contrast, the IC50 values for the proteomic-based treatment options were substantially lower: 25.188 µM for cerulenin (**Figure 24D**), 11.460 µM for sunitinib (**Figure 24E**), and 11.781 µM for the combination of sunitinib and cerulenin (**Figure 24F**). These lower IC50 values reflect a more potent anti-cancer effect of the proteomic-based regimen, suggesting that this approach offers a more effective strategy for targeting tumor 1's specific proteomic profile.

Moreover, it is noteworthy that the cell viability curve for paclitaxel displayed a more linear response compared to the other treatments, suggesting a potential resistance to paclitaxel. Deep proteomic analysis of tumor tissue clone effectively identified proteins associated with paclitaxel resistance. For instance, EDIL3 and CA12 proteins, which were exclusively found in clone B, may account for the observed resistance. EDIL3 is known to promote epithelial-mesenchymal transition and paclitaxel resistance by interacting with integrin αVβ3 in cancer cells. This interaction is also reflected in the biological pathway analysis of proteins associated with clone B and C. Additionally, research on CA12 has demonstrated that silencing the Carbonic Anhydrase 12 gene can restore paclitaxel sensitivity in drug-resistant breast cancer cells. Other markers of paclitaxel resistance identified across the three clones include PGK1, a well-known predictor of poor survival and a novel prognostic biomarker for chemoresistance to paclitaxel, as well as CapG proteins. These findings highlight the utility of proteomic analysis in uncovering mechanisms of drug resistance and in guiding the development of more effective, targeted therapies. Proteomic analysis of organoids derived from Tumor 1 before

treatment also confirmed the presence paclitaxel resistance markers (EDIL3, CA12, PGK1, and CapG proteins), which could explain the observed drug response patterns.

Results for other patient derived organoids are still in progress and won't be presented in this manuscript.

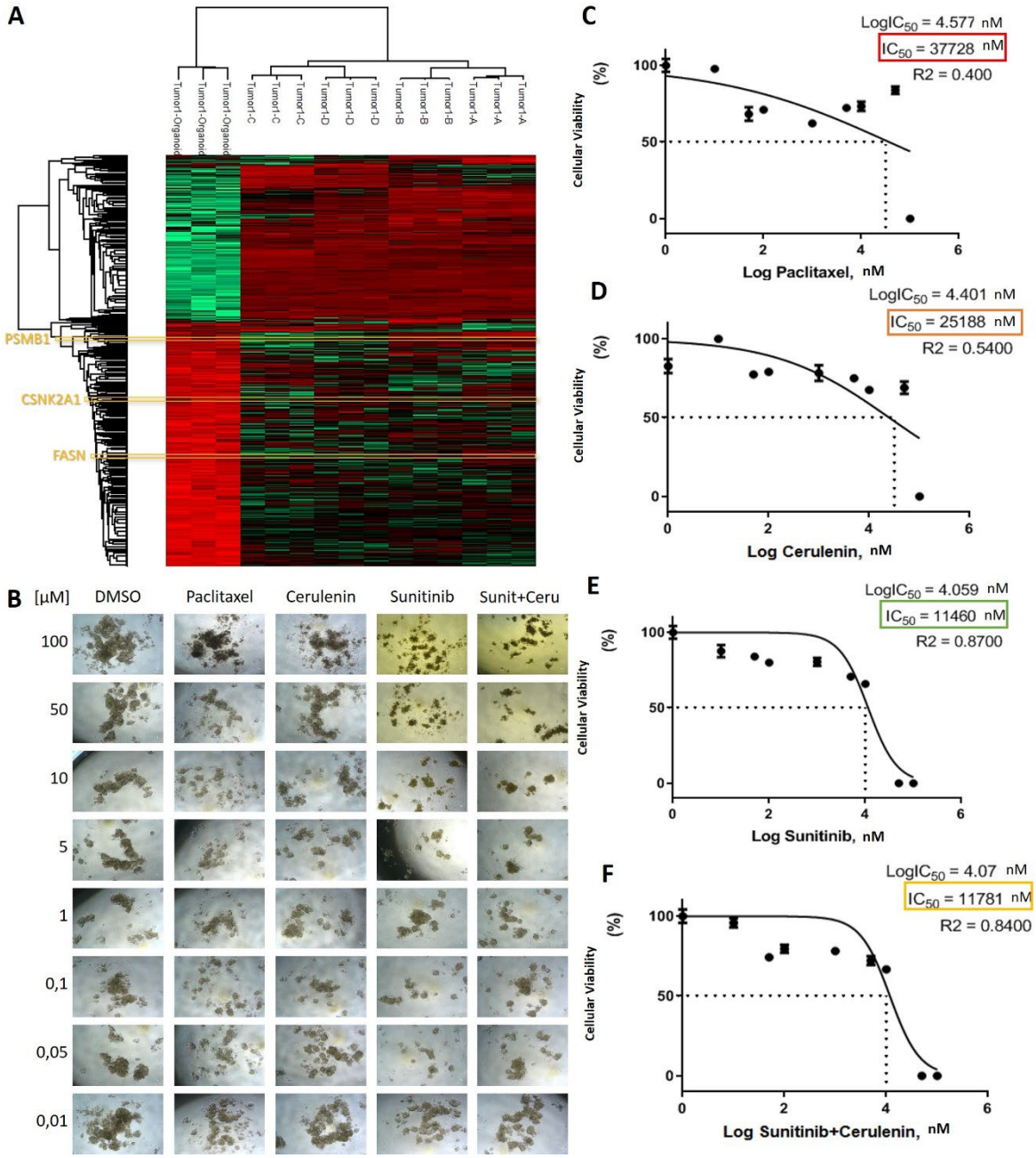


Figure 24: Proteomic Analysis and Drug Response of Tumor 1 Organoids. A) Heatmap showing over-expressed proteins in Tumor 1 and its derived organoid. Comparison of Tumor 1 organoid treatment with B) organoid images at varying drug concentrations and cell viability data for C) Paclitaxel, D) Cerulenin, E) Sunitinib, and F) the Sunitinib-Cerulenin combination.

Inter-Patient Tumor Heterogeneity Analysis

Some similarities in conventional treatment and proteomic proposed treatment between the four tumors was an intriguing finding. This suggests that specific clones may be shared by tumors

despite originating from different patients. Consequently, tumor clone heterogeneity can be observed across different patients. This information could then be integrated into machine learning concepts for the quick proposal of treatment guidelines, the prediction of drug resistance, or the evaluation of drug efficacy for individual patient tumors from tumor MALDI MSI.

A co-segmentation analysis was performed on the four tumors, resulting in an image with nine distinct clusters (**Figure 25A**). This approach revealed distinct molecular patterns, or clones, associated with each tumor. Notably, Tumor 1 displayed significant differences compared to Tumors 2, 3, and 4, specifically characterized by the presence of cluster 5 (red). Similarly, Tumor 2 exhibited marked dissimilarity from Tumors 1, 3, and 4, primarily due to the presence of cluster 2 (light green). In contrast, while Tumors 3 and 4 were distinct from Tumors 1 and 2, they closely resembled one another, sharing several clones represented by clusters 1, 7, and 9 (blue, orange, and purple). Nevertheless, they could still be differentiated by the unique presence of clusters 3 (green) and 6 (light orange) in Tumor 4, as well as cluster 8 (light purple) in Tumor 3. This finding suggests that the conventional treatments administered for Tumors 3 and 4 were identical, indicating their closely related molecular profiles, as evidenced by the co-segmentation analysis. Additionally, t-SNE visualization of the clusters **in Figure 25C**, further illustrated these relationships, highlighting the distinct molecular landscapes of each tumor type.

To deepen the analysis, the proteomes of each clone were combined per patient to enable a comparative proteomic assessment across the four patients. A resulting Venn diagram (**Figure 25D**) demonstrated that each tumor tissue harbored a specific proteomic profile due to the presence of distinct clones. For example, Tumor 4 exhibited 19 unique proteins, while Tumor 2 had 12 exclusive proteins. An ANOVA test, applying a stringent p-value threshold of < 0.01, identified 4767 differentially expressed proteins across the four tumors. The corresponding heatmap (**Figure 25E**) clearly delineated distinct clusters of over-expressed proteins, underscoring both the unique (framed in yellow) and shared (framed in blue) proteomic signatures among the tumors, directly linked to the specific clones within each sample.

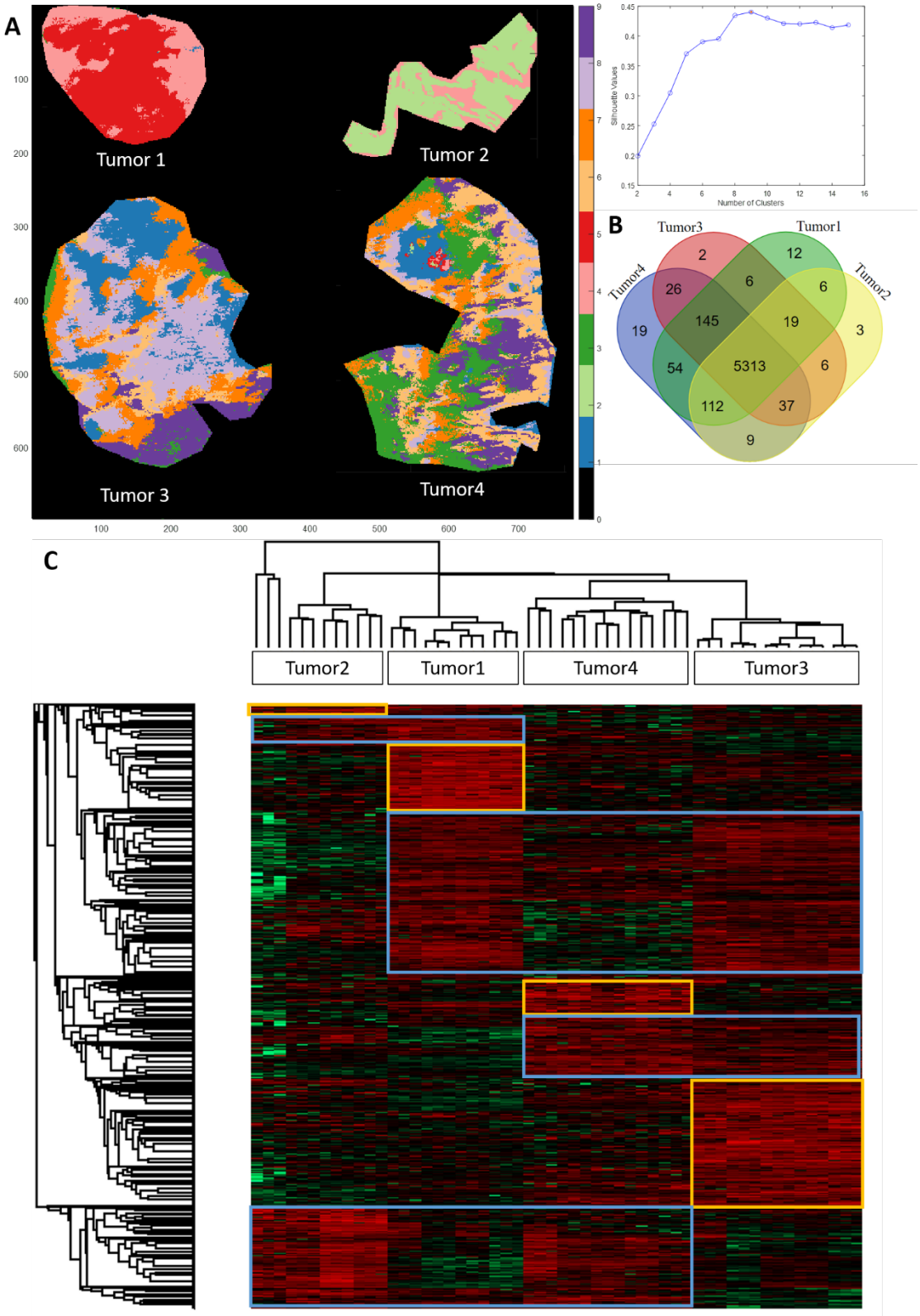


Figure 25: Inter-tumor heterogeneity analysis between tumors 1, 2, 3 and 4 according to: A) MALDI images co-segmentation with 9 clusters following silhouette criterion, spatial proteomic analysis represented through B) Venn diagram and C) over expressed proteins heatmap .

Notably, several druggable proteins were differentially expressed across the clusters, as shown in **Figure 22** and **Appendix A, Figure 63**. For example, in Tumor 2, proteins HBB and F13A1 were distinctly over-expressed, suggesting that the drugs nomifensin and cerulenin could be promising therapeutic options for this tumor. On the other hand, Paclitaxel-targetable proteins, such as TUBB and TUBA1A, were under-expressed in Tumor 1, potentially explaining the lack of response to Paclitaxel in Tumor 1-derived organoids. In contrast, these proteins were highly expressed in Tumors 2, 3, and 4, indicating that Paclitaxel might be more effective in those cases. Furthermore, the protein PSMB1 was exclusively over-expressed in Tumor 1, which may indicate that Clone B, originating from Tumor 1, is uniquely associated with this patient's tumor tissue. This finding highlights the molecular heterogeneity of the tumors and suggests that targeted therapies could be personalized based on the specific proteomic and clonal landscape of each patient's tumor.

A ClueGo analysis was first performed for tumor specific over-expressed protein clusters associated genes (**Figure 26**), supporting results observed through the heatmap.

Vpu mediated degradation of CD4 was the major biological process associated to tumor 1 over-expressed proteins cluster. Vpu is a protein encoded by HIV-1 known for degrading CD4 receptors on immune cells, leading to immunosuppression. In the context of breast cancer, Vpu-mediated CD4 degradation could have several consequences (Cong et al., 2021). By reducing CD4+ T-cell activity, the immune system's ability to recognize and eliminate cancer cells may be weakened, allowing tumor cells to evade immune detection . This could also promote an immunosuppressive tumor microenvironment, reducing pro-inflammatory cytokine production and potentially increasing regulatory T-cell activity, which further supports tumor growth. Additionally, weakened CD4+ T-cell responses may reduce the effectiveness of immunotherapies or cancer vaccines. While Vpu's role is specific to HIV, similar mechanisms of CD4 downregulation could theoretically impact breast cancer progression and immune evasion strategies.

In another hand, autophosphorylation of SRC (Sarcoma proto-oncogene) played a significant role in the context of tumor 2 (**Figure 26**). SRC is a non-receptor tyrosine kinase belonging to the SRC family kinases, which are crucial regulators of various cellular processes such as proliferation, migration, and survival (Elsberger, 2014). SRC activation has been implicated in numerous cancers, including breast cancer, where it drives oncogenic signaling pathways. In luminal breast cancer, where estrogen receptor (ER) signaling is the primary driver of tumor growth, SRC plays a key modulatory role (Elsberger, 2014). Over time, however, some tumors may develop resistance to endocrine therapies, such as tamoxifen or aromatase inhibitors, which are designed to block ER signaling. SRC signaling has been directly linked to these resistance mechanisms, as it can facilitate alternative pathways that allow cancer cells to bypass the need for ER activation, promoting uncontrolled

growth despite anti-estrogen treatments. Due to its involvement in both oncogenic signaling and resistance to endocrine therapies, SRC has emerged as a promising therapeutic target in luminal breast cancer (Kohale et al., 2022; Luo et al., 2022).

In tumor 3, SUMOylation pathway was the biological process mainly highlighted (**Figure 26**). SUMOylation is a post-translational modification where Small Ubiquitin-like Modifier (SUMO) proteins are covalently attached to target proteins, is increasingly recognized for its critical role in tumor initiation and progression. As one of the most common regulatory modifications in cells, SUMOylation can influence a variety of cellular processes by altering protein stability, localization, and activity. Recent studies have revealed that large numbers of SUMOylated and deSUMOylated proteins act as key players in the adaptation of cancer cells to the tumor microenvironment (Gu et al., 2023). These proteins mediate essential responses such as hypoxia adaptation, metabolic reprogramming, inflammatory regulation, and immune evasion processes that are crucial during tumor development and the formation of the TME. Moreover, SUMOylation has been shown to influence inflammatory and immune responses within the TME. By modifying key transcription factors and signaling molecules, SUMOylation can either promote or suppress inflammation and modulate the immune system's ability to recognize and attack cancer cells. This dual role in immune evasion makes SUMOylation a highly attractive target for therapeutic intervention (Gu et al., 2023). In summary, SUMOylation holds promise as both a diagnostic and therapeutic target.

Finally, chylomicron assembly emerged as a relevant biological pathway due to its central role in lipid metabolism of tumor 4 (**Figure 26**). Chylomicrons are lipoproteins produced by enterocytes in the small intestine, transport dietary triglycerides and cholesterol through the lymphatic system to other tissues. While chylomicrons are not directly linked to cancer formation, alterations in lipid metabolism, including their assembly and utilization, are believed to promote tumor growth and metastasis in breast cancer (Pandrangi et al., 2022). Breast cancer cells, particularly in aggressive subtypes, reprogram their metabolism to depend heavily on lipids for energy, membrane synthesis, and signaling. Chylomicron-derived fatty acids and cholesterol can fuel this metabolic shift, supporting cancer cell proliferation and spread. Given the connection between chylomicron assembly, lipid metabolism, and breast cancer progression, targeting these pathways could provide new therapeutic opportunities. Inhibiting lipid transport or synthesis and disrupting lipid signaling may slow tumor growth or improve sensitivity to existing treatments. Additionally, changes in chylomicron levels or function could serve as biomarkers for tracking disease progression (I. Sinha et al., 2023) or treatment response.

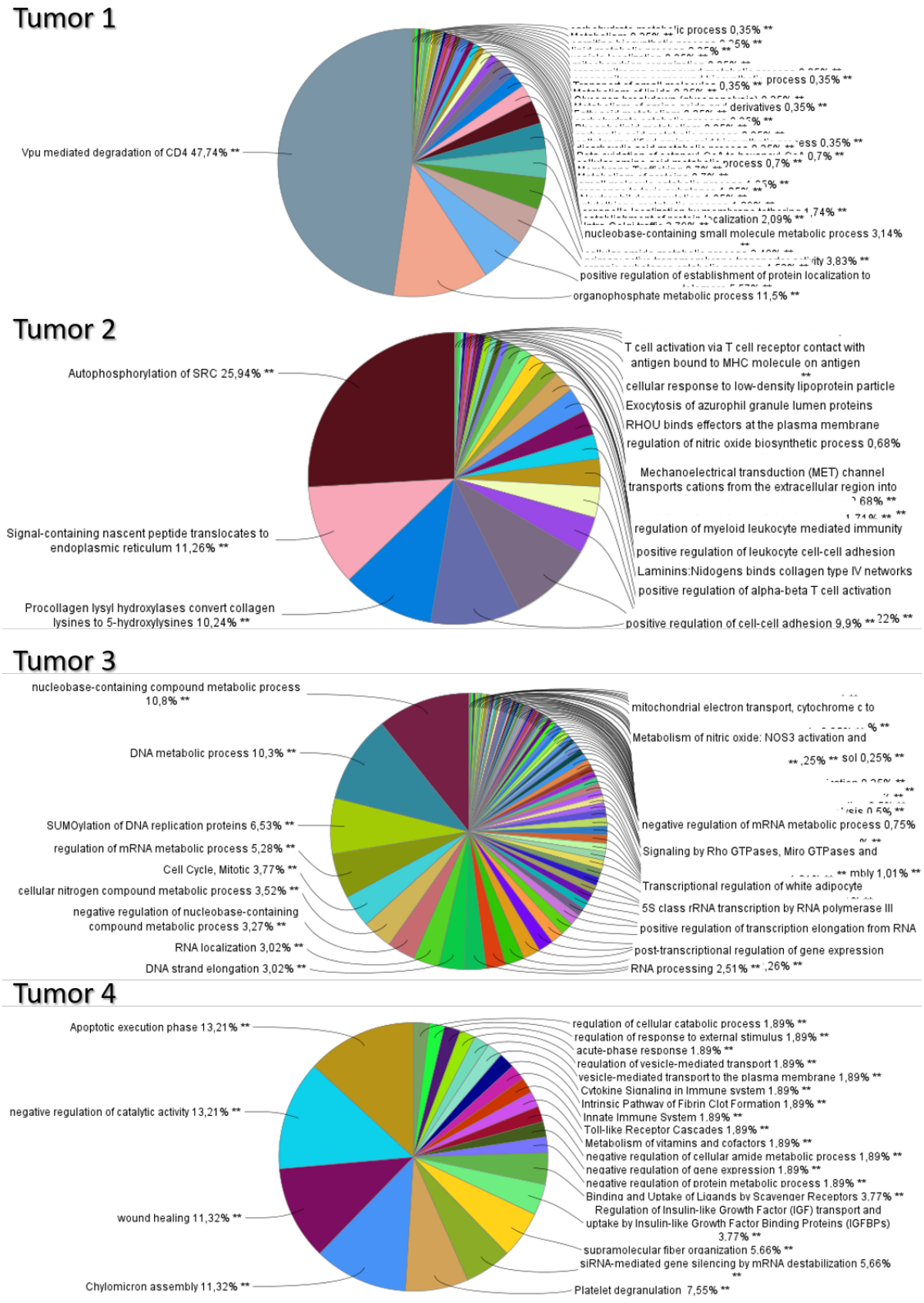


Figure 26: ClueGo biological pathways associated to over-expressed clusters specifically involved in tumors 1, 2, 3 or 4 according to inter-tumoral protein analysis.

To further analyze the data, a ClueGo analysis was conducted on the over-expressed protein clusters shared across multiple tumors, highlighted in blue in **Figure 25**. The purpose of this analysis was to identify potential therapeutic pathways that are common to different tumors (**Figure 27**).
Four main over-expressed protein clusters were identified, showing common proteins between tumor pairs: tumor 1 and tumor 2, tumor 2 and tumor 4, tumor 1 and tumor 3, and tumor 4 and tumor 3.

Interestingly, the overexpressed proteins found in both Tumor 1 and Tumor 2 were linked to the SRP-dependent co-translational protein targeting to the membrane pathway (**Figure 27**). The Signal Recognition Particle (SRP) is a ribonucleoprotein complex that plays a crucial role in recognizing and targeting specific proteins to the endoplasmic reticulum membrane during their translation. In the context of cancer, SRP is believed to influence cell signaling pathways that affect growth and survival. Ongoing investigations into SRP proteins have suggested their overexpression in various cancer tissues, including BC (Kellogg et al., 2022), positioning them as potential therapeutic targets for both Tumor 1 and Tumor 2.

For Tumor 1 and Tumor 3 (**Figure 27**), the primary pathway identified was related to the cellular response to stimuli. This finding suggests that these tumors may activate specific signaling networks that enable tumor cells to adapt to their changing environments, potentially promoting tumor progression. Moreover, the notable interaction with HIV factors, particularly through the Vpu-mediated degradation of CD4 pathway, lends further weight to the hypothesis that Tumor 1 is influenced by HIV infection. This interaction might not only affect immune evasion but also alter the tumor microenvironment, providing a survival advantage to the cancer cells.

In the case of Tumor 2 and Tumor 4, numerous pathways were identified, including the EPH-ephrin signaling pathway (**Figure 27**). This signaling pathway plays a significant role in various biological processes, including cell migration, adhesion, and proliferation. In the context of carcinogenesis, EPH-ephrin signaling has been implicated in promoting tumor growth and metastasis by modulating interactions within the tumor microenvironment. Given its established role in multiple cancer types, including breast cancer, EPH-ephrin signaling represents a promising target for the development of new anticancer therapies (Psilopatis et al., 2022).

Regarding the shared overexpressed proteins between Tumor 3 and Tumor 4, several pathways related to RNA processing were involved, in addition to MET (Mesenchymal-epithelial transition) signaling (**Figure 27**). The MET receptor, which binds to hepatocyte growth factor (HGF), plays a critical role in promoting cell proliferation, survival, and migration. The activation of RNA processing pathways in conjunction with MET signaling may enhance tumor cell division and growth, providing additional mechanisms that contribute to tumor progression (Famta et al., 2024).

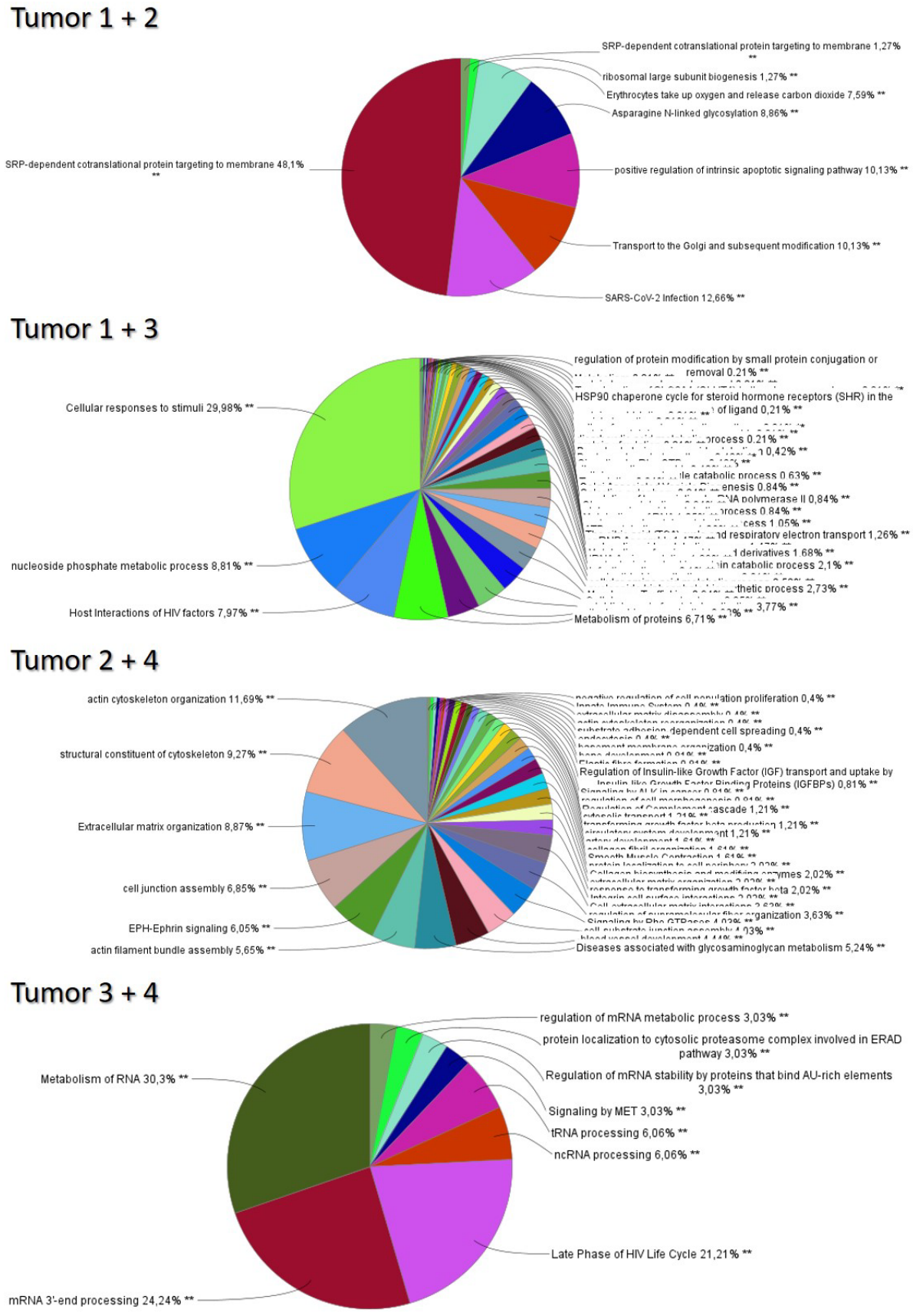


Figure 27: ClueGo biological pathways associated to over-expressed clusters shared between tumors 1, 2, 3 or 4 according to inter-tumoral protein analysis.

Conclusion and Perspectives

By leveraging MALDI MSI alongside advanced proteomic profiling techniques, the study revealed the intricate proteomic landscape of tumor tissues, identifying distinct tumor clones based on their proteomic signatures. In addition, the application of a spatially resolved, proteomics-driven workflow to the analysis of four FFPE breast cancer tumor samples provided profound insights into the molecular complexity and heterogeneity of breast cancer. These clones, characterized by differential protein expression and involvement in specific biological pathways, underscore the multifaceted nature of breast cancer progression and the varied responses to treatment that arise from intra-tumor heterogeneity.

Importantly, these proteomic variations were linked to key biological pathways involved in breast cancer progression, such as telomere end packaging, integrin surface interactions, and extracellular matrix remodeling. The identification of such pathways offers critical insights into the molecular mechanisms driving tumor growth, metastasis, and resistance to conventional treatments. The study emphasized the importance of targeting these specific pathways for more effective therapeutic interventions. For instance, in tumor 1, pathways like telomerase activity in clone A and integrin-related functions in clones B and C were identified as key drivers of cancer cell behavior, pointing to potential drug targets such as telomerase inhibitors and integrin modulators.

One of the most significant outcomes of the study was the proposal of personalized treatment strategies based on the proteomic profiles of each tumor clone. Unlike conventional one-size-fits-all treatments, which often overlook the heterogeneity within and between tumors, the proteomic data allowed for the identification of precise drug targets, such as FASN in clone A of tumor 1 and PSMB1 in clone B, leading to tailored drug combinations like cerulenin and sunitinib. This tailored approach has the potential to significantly improve treatment efficacy by addressing the specific vulnerabilities of each clone, thereby overcoming the limitations of traditional therapies. The validation of these proteomic-based treatment regimens in organoid models further supported their effectiveness, as evidenced by lower IC50 values compared to conventional drugs like paclitaxel, which showed potential resistance due to the presence of proteins like EDIL3 and CA12 in resistant clones.

Additionally, the study's inter-patient tumor analysis revealed that some clones shared similar proteomic profiles across different patients, suggesting common molecular patterns that could be leveraged for broader therapeutic applications. This cross-tumor comparison opens up possibilities for the development of machine learning models that could predict treatment efficacy or drug resistance based on tumor proteomic data. By identifying common protein targets across

patients, personalized treatments could be rapidly designed for new cases based on previously observed molecular patterns, significantly enhancing the precision and speed of clinical decision-making.

Furthermore, the biological pathways enriched within each tumor's clones, ranging from the telomerase pathway in tumor 1 to SRC activation and lipid metabolism in tumors 2 and 4, highlight the relevance of targeting these distinct processes in breast cancer treatment. These findings underscore the complexity of tumor biology and the critical need for a nuanced, multi-targeted therapeutic approach. The involvement of pathways such as chylomicron assembly and SUMOylation further exemplifies how tumor cells adapt their metabolic and signaling pathways to sustain growth, evade immune surveillance, and resist conventional therapies. In addition, the ClueGO analysis of overexpressed protein clusters across multiple tumors has provided valuable insights into potential therapeutic pathways relevant to multiple BC tissues. By identifying four significant protein clusters shared between tumor pairs, we highlighted critical pathways such as SRP-dependent co-translational protein targeting, cellular response to stimuli, EPH-ephrin signaling, and MET signaling. The involvement of SRP proteins in Tumor 1 and Tumor 2 suggests they could serve as promising therapeutic targets. Additionally, the activation of cellular signaling networks in Tumor 1 and Tumor 3 underscores the adaptive mechanisms that tumors utilize to thrive in changing environments, while the EPH-ephrin pathway's relevance to Tumor 2 and Tumor 4 indicates its potential as a target for new cancer therapies. Furthermore, the interplay of RNA processing and MET signaling in Tumor 3 and Tumor 4 emphasizes the complexity of tumor biology and the need for multifaceted therapeutic strategies. Overall, these findings pave the way for future investigations aimed at developing targeted interventions that could improve patient outcomes in breast cancer.

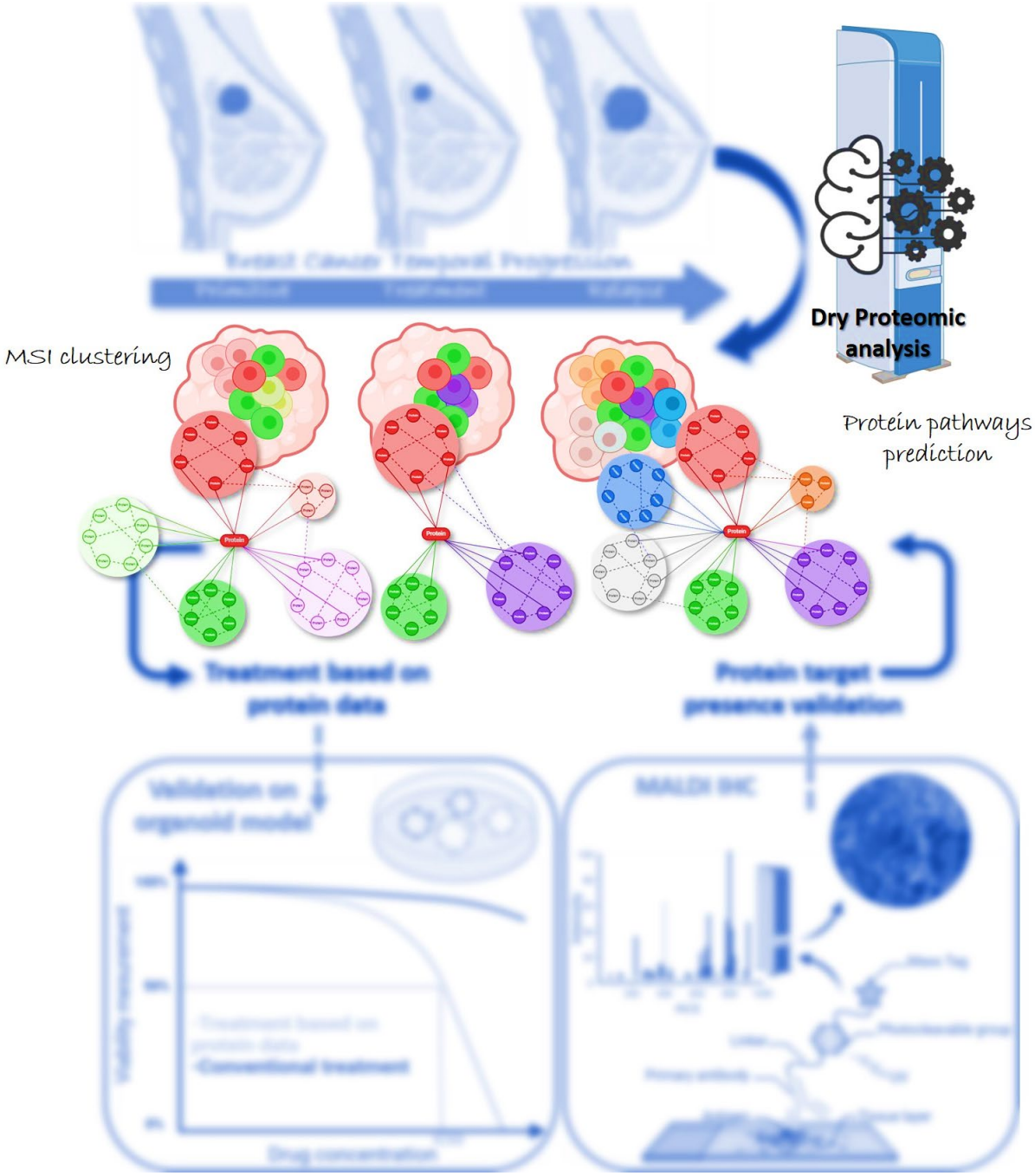
To resume, this study illustrates the immense potential of proteomic-driven analysis for breast cancer treatment, providing a clear pathway toward more effective, personalized therapies. By focusing on the unique proteomic landscapes of tumor clones, this approach not only enhances our understanding of tumor heterogeneity but also presents a powerful tool for overcoming drug resistance and optimizing therapeutic outcomes. The integration of spatial proteomics with clinical treatment strategies offers a promising avenue for the future of precision oncology, where individualized treatment plans can be tailored to the specific molecular characteristics of each patient's tumor, ultimately improving survival rates and quality of life for breast cancer patients.

However, it is important to recognize that this approach is quite complex and time-consuming, making it challenging to implement in routine clinical practice. To advance this study, it will be crucial to expand the dataset by including more patient clones, as well as associated

biomarkers and treatment options. This expanded dataset would enable the development of a machine learning model capable of automatically analyzing a patient's tumor and proposing personalized, optimized treatment strategies tailored to the tumor's specific heterogeneity. Such a model could significantly streamline the therapeutic decision-making process, making it more efficient and precise. Indeed, by integrating spatial proteomics with machine learning, this approach has the potential to transform cancer treatment, offering a more adaptive and individualized therapy plan for each patient. Moreover, it could improve the prediction of drug resistance and treatment efficacy, ultimately paving the way for a more personalized and dynamic form of cancer care that can rapidly adjust to evolving tumor profiles.

CHAPTER 3

Dry Proteomic Concept Based on Lipid MALDI MSI



CHAPTER 3: Dry Proteomic Concept Based on Lipid MALDI MSI

Introduction

Since the gap between mass spectrometry imaging (MSI) and proteomics has been bridged by the development of spatially resolved proteomics guided by MALDI MSI (Delcourt et al., 2018; Kertesz et al., 2015; Lee et al., 2008; Quanico et al., 2013; Wisztorski et al., 2013), the next challenge was to perform multi-omics analyses at the spatial level (Dewez et al., 2020; Donnarumma & Murray, 2016; Lamont et al., 2017; Mezger et al., 2021; Quanico, Franck, Wisztorski, et al., 2017; Sun et al., 2023). Nevertheless, there are still developments to be performed to correlate from lipid MSI data, proteins and lipid networks to retrieved functions. Multi-omics MSI is particularly valuable for the analysis of heterogeneous biological samples, such as brain or tumors, which consist of different cell types and regions with distinct molecular composition and function (Delcourt et al., 2018). Indeed, tumor heterogeneity is a significant and growing area in cancer research. An overview on tumoral heterogeneous proteome is subsequently linked to therapeutic, allowing drug resistance analysis and optimized treatment guideline proposal, tending to personalize medicine strategy. However, the complex nature of protein annotation and the lack of standardized methodologies pose challenges to the effectiveness of MALDI-MSI data analysis, especially in multi-omics clinical research. The interpretation and integration of the vast amount of data generated by these technologies remains a significant limitation (Quanico, Franck, Cardon, et al., 2017). Extracting meaningful insights from complex datasets therefore requires sophisticated computational approaches and bioinformatic analysis (Alexandrov, 2012). MALDI MSI data analysis involves pre-processing and processing stages, preparing them for subsequent statistical analysis. Reduction techniques, like PCA (Principal Component Analysis) (Fonville et al., 2012; Trim et al., 2008), t-SNE (t-distributed Stochastic Neighbour Embedding) (Abdelmoula et al., 2018; Wang et al., 2022), or NNMF (Non-negative matrix factorisation)(Leuschner et al., 2019; Nijs et al., 2021), are particularly useful for exploring the spatial distribution of molecular features in MALDI MSI data (Brunelle & Laprévote, 2012; Deininger et al., 2011). In addition, the combination of MSI and machine learning methods is widely used in the processing step to effectively extract the essential information contained in complex MSI data. The emergence of segmentation methods, such as bisecting *k*-means, hierarchical clustering and *k*-means clustering (Arthur & Vassilvitskii, n.d.-a; Duda & Peter, 2012), provides valuable insights from complex data like meaningful regions corresponding to biological features in heterogeneous sample. However, choosing the right number of *k*-clusters is not straightforward, limiting biological conclusions (Arthur & Vassilvitskii, n.d.-a; Duda & Peter, 2012). The common method involves

performing *k*-means clustering for different k values ($2<k<k_{max}$) and calculate the distances between clusters. The aim is to find the optimal *k* that minimizes intra-class distances while maximizing inter-class distances. Several statistical indices, called criterions, have been developed for this purpose (Nardecchia et al., 2020; Nguyen et al., 2015) .

Here, we introduce the concept of dry proteomics, an automated procedure capable of identifying heterogeneous clusters of biological samples according to their lipid signature, thought lipid MALDI MSI, and automatically providing their associated protein data without any proteomic experiments. The development of this machine learning method required overcoming several challenges. The central hypothesis was that if a cluster appeared identical in both lipid and protein images, it should possess lipids and paired proteins related to a specific biological pathway, like a unique barcode that allows one cluster to be distinguished from others (**Figure 28**). Thus, the correlation between lipids and proteins in a biological network, within different clusters, forms the basis of dry proteomics. The data processing workflow was first developed on lipid, protein, and peptide MSI datasets performed on rat brain (RB) tissue. We succeeded in building a segmentation pipeline, consisting of Singular Value Decomposition (SVD) data compression pre-processing and *k*-means++ segmentation processing steps. The integration of the silhouette criterion allowed to optimize and automate the optimal number of clusters finding for MSI analysis, corresponding to the sample heterogeneity. The next step was to develop a prediction model that could blindly identify the different RB clusters from a lipid MS image according to their spectral fingerprint. The prediction model was complemented by discriminative lipid and protein identifications for each cluster, forming a dry proteomic reference dataset for RB tissue section.

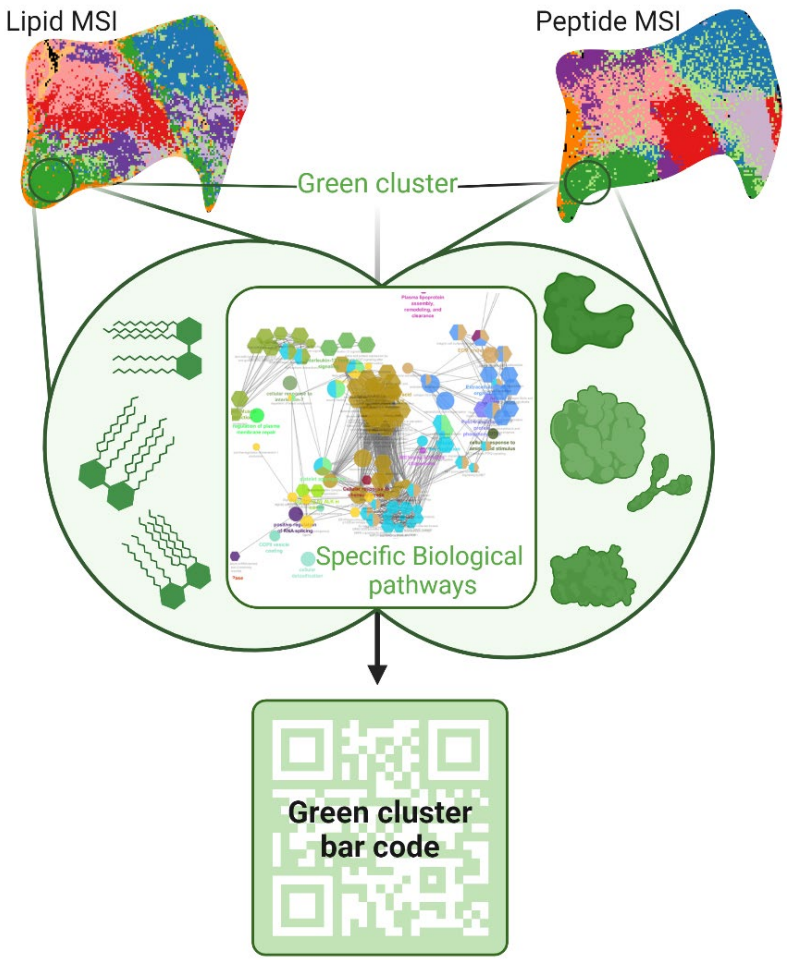


Figure 28: Basics of dry proteomics. Clusters appearing identical in both lipid and protein images should contain lipids and corresponding proteins linked to specific biological pathways.

Finally, the dry proteomics concept is a simple and rapid procedure, as the user only needs to perform lipid MALDI MSI to automatically identify the heterogeneous clusters present in a sample and obtain their specific proteome (**Figure 29**). The development of this tool is aimed at clinical application for patient therapeutic guidance. Indeed, the protein information provided by the dry proteomics process can be related to drug resistance, potential therapeutic target or patient survival, which could help the oncologist to propose a therapeutic guideline adapted to the patient's tumour. In this way, the ultimate phase of presented research involved the application of this innovative concept to intricate and heterogeneous pathology samples, particularly human Glioblastoma (Bikfalvi et al., 2023; Duhamel et al., 2022; Zirem et al., 2024). In addition, by applying the dry proteomics workflow, correlation between predicted protein and patient survival outcome information allowed to establish a robust model for glioblastoma patient survival prediction. This crucial validation step not only enhances confidence in the reliability of this approach but also holds significant promise for advancing personalized medicine strategies in the management of this challenging disease. Indeed, the assessment of heterogeneity, whether intra or interpatient, is pivotal in personalized medicine, as

it allows for the identification of unique molecular profiles that can inform tailored treatment strategies for individual patients.

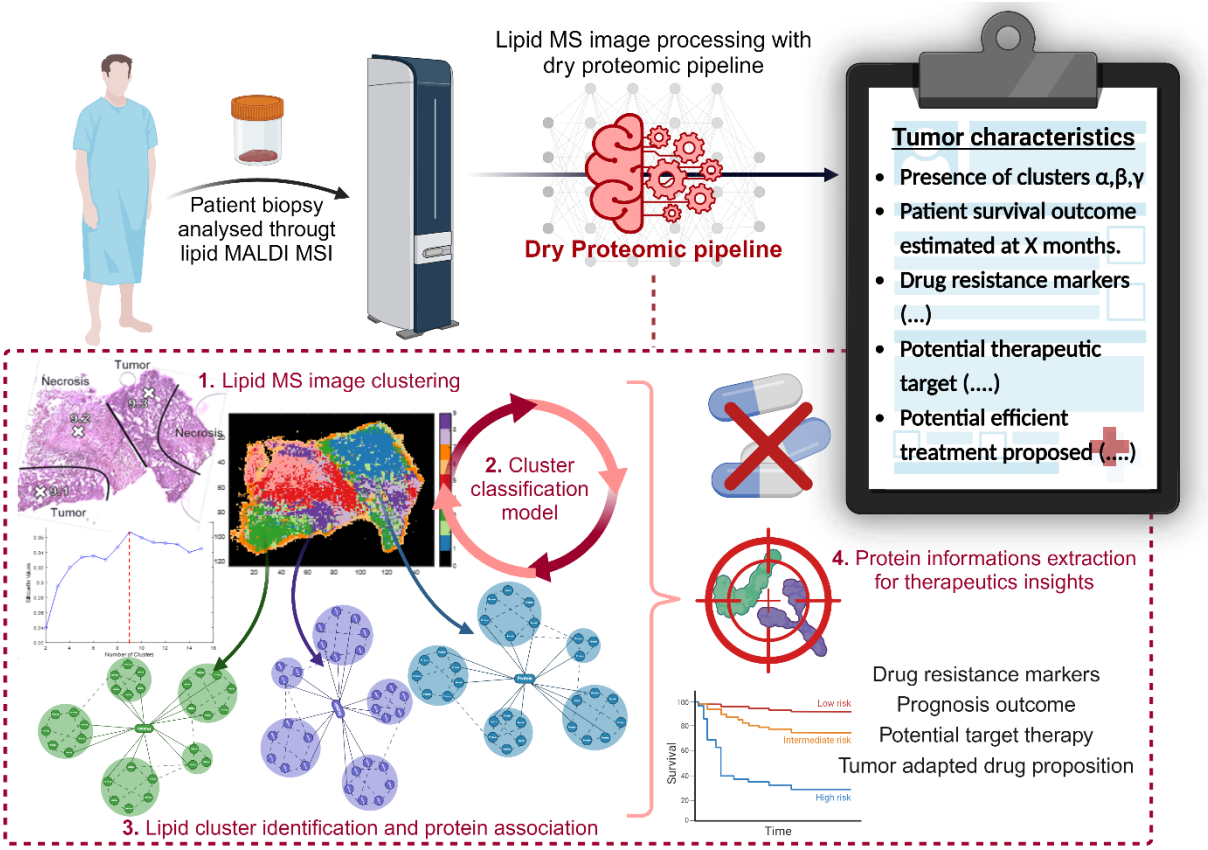


Figure 29: Dry proteomic concept general concept. This workflow illustrates the use of MALDI MSI and machine learning for personalized cancer treatment. Tissue samples are processed through MALDI MSI, and lipid data is fed into a segmentation pipeline utilizing clustering methods (e.g., SVD, *k*-means) to identify spatial proteomic patterns. These patterns highlight intra-tumor heterogeneity. A machine learning model then integrates cluster associated proteomic data to highlight with drug resistance markers, predict patient prognosis, and suggest potential targeted therapies based on the tumor's molecular profile, aiding in personalized treatment strategies.

Material and Method

Experimental Design and Statistical Rationale

For MALDI imaging and spatial omics development studies n = 3 male Wistar rats were sacrificed. All the experiments were performed in biological triplicate to ensure data reproducibility. For the proteomic statistical analysis of conditioned media, as a criterion of significance, an ANOVA significance threshold of p-value ≤ 0.01 was applied, and heat maps were generated. Normalization was achieved using a Z-score with matrix access by rows. To assess the statistical significance of biomarkers for lipids MSI biomarkers, a non-parametric Kruskal-Wallis test was employed. Bonferroni corrections were applied to adjust p-values for multiple comparisons. Values are presented as medians and visualized through scatter boxplots.

A retrospective cohort of 50 fresh frozen (FF) glioblastoma tissues was obtained from the Pathology department of Lille Hospital, France. A prospective cohort of 50 FF glioblastoma tissues

were also included in this study. 50 patients with newly diagnosed glioblastoma were prospectively enrolled between September 2014 and November 2018 at Lille University Hospital, France (NCT02473484). This research complies with all relevant ethical regulations. Approval of the study protocol was obtained from the Lille Hospital research ethics committee (ID-RCB 2014-A00185-42) before the initiation of the study. The study adhered to the principles of the Declaration of Helsinki and the Guidelines for Good Clinical Practice and is registered at NCT02473484. Informed consent was obtained from patients. Participants did not receive any compensation. According to the French Public Health Code and in application of the General Data Protection Regulations, all patients had been informed at the time of care that their standard clinical and biological data could be used for research purposes regarding the retrospective analysis of FF samples, and none had expressed his opposition. Regarding the prospective collection of samples, each patient's informed consent for the collection and publication of clinical and biological data was obtained at the time of hospitalization prior to surgical intervention (Duhamel et al., 2022; Zirem et al., 2024). Tissue sections were subject to H&E coloration for histopathological analysis. The regions annotations were made by a pathologist.

Chemical Products and Material

Water (H_2O), ethanol (EtOH), acetic acid, acetonitrile (ACN) and methanol (MeOH) were obtained from Thermo Fisher Scientific (Courtaboeuf, France). 99% pure trifluoroacetic acid (TFA), α-cyano-4-hydroxycinnamic acid (HCCA), sinapinic acid (SA), 2,5-dihydroxybenzoic acid (2,5-DHB), aniline, formic acid (FA) and ammonium bicarbonate (NH_4HCO_3) were purchased from Sigma-Aldrich (Saint-Quentin Fallavier, France). The chloroform ($CHCl_3$) was obtained from Carlo Erba Reagents (Val-de-Reuil, France). Porcine Trypsin Sequencing Grade was from Promega (Charbonnières, France).

Tissues were cut on a cryostat (Leica Microsystems, Nanterre, France). Indium Tin Oxide slides were purchased from LaserBio Labs (Valbonne, France), whereas the poly-lysine coated slides were from Epredia™ (Braunschweig, Germany). The MALDI matrices and the trypsin were deposited on the tissue sections using the HTX M5-Sprayer™ (HTX Technologies, Carboro, NC, USA). Mass spectrometry imaging analyses were performed using the MALDI-TOF Rapiflex Tissuetyper (Bruker Daltonics, Bremen, Germany) equipped with the Smart Beam 3D laser. Spatial proteomic analysis were carried out through the utilization of chemical printer (CHIP-1000, Shimadzu, Kyoto, Japan) and the TriVersa Nanomate device (Advion Biosciences Inc, Ithaca, NY, USA). Samples were dried in a SpeedVac (SPD13DPA, Thermo Fisher Scientific, Waltham, Massachusetts, USA). nLC-MS/MS analysis were performed with TimsTOF Flex (Bruker) coupled to an EVOSEP One (EVOSEP).

Sample Preparation

Rat brains were obtained from our collaborator Dr. Dasa Cizkova (Institute of Neuroimmunology, Slovak Academy of Science, Bratislava). Male Wistar rats of adult age were

sacrificed by CO_2 asphyxiation and dissected. Brain tissues were frozen in isopentane at -50 °C and stored at -80 °C until use. Experiments on animals were carried out according to institutional animal care guidelines conforming to international standards and were approved by the State Veterinary and Food Committee of Slovak Republic (Ro-4081/17-221), and by the Ethics Committee of the Institute of Neuroimmunology, Slovak Academy of Science, Bratislava. For this study, FF rat brain tissues were cut using a cryostat at -20°C. All sections were obtained at the same time and stored at -80°C until their use. Rat brain sagittal 12 µm sections were prepared, to finally reach 22 batch of 4 consecutive sections. Tissues were mounted on ITO slides and respectively intended to: back-up, lipid in negative and positive mode imaging, protein imaging and peptide imaging in positive mode (Caprioli et al., 1997; Hajjaji et al., 2022a).

10 others consecutive rat brain sagittal sections of 12 µm were mounted on poly-lysine coated slide for lipid analysis carried out by SpiderMass technology. Another three consecutive 20 µm sections were mounted on poly-lysine coated slide for spatial proteomic analysis.

Finally, 3 different rat brain sagittal 12 µm section were fixed onto ITO coated slide as a validation cohort for the lipid predictive model.

For the analysis of horizontal rat brain tissues, 4 consecutives sections were prepared for multi-omics MSI analysis as describe bellow, followed by another consecutive sections for spatial proteomic analysis. This schema was repeated on 4 different rat brains.

Lipid MALDI MS Imaging

Tissues were dried in a desiccator before a matrix deposition. Norharmane was used as MALDI matrix for positive and negative lipid imaging. The matrix was deposited at 7 mg/mL in $CHCl_3$: MeOH (2:1, *v/v*). The HTX parameters for norharmane spray were: spray at 30°C with 10 psi pressure, a pattern CC, a flow rate of 0.1 mL/min, a velocity of 1200 mm/min, for 12 passages with 2 mm track spacing. Lipid images were performed on the MALDI-TOF Rapiflex Tissuetyper mass spectrometer. The spectra were acquired within the *m/z* 200-1200 range in positive ion mode and the *m/z* 400-1500 range in negative ion mode. All data were performed in the delayed extraction reflectron mode with an average of 300 laser shots per pixel for a spatial resolution of 50 µm. The laser energy was set around 60 % and the voltages of the ion source were 20 kV and 11 kV for the lens. Same protocol was applied for 10 µm lipid imaging.

Other images were performed with DHB matrix in positive ion mode. The matrix was deposited at 10 mg/mL in MeOH: TFA 0.1% (7:3, *v/v*). The HTX parameters for DHB spray were: spray at 75°C, tray at 55°C, with 10 psi pressure, a pattern CC, a flow rate of 0.1 mL/min, a velocity of 1200

mm/min, for 8 passages with 2 mm track spacing. Lipid images were performed on the MALDI-TOF Rapiflex Tissuetyper mass spectrometer. The spectra were acquired within the *m/z* 200-1200 range in positive ion mode. All data were performed in the delayed extraction reflectron mode with an average of 300 laser shots per pixel for a spatial resolution of 50 µm. The laser energy was set around 85 % and the voltages of the ion source were 20 kV and 11 kV for the lens.

Protein MALDI MS Imaging

Tissues were vacuum dried before being subjected to delipidation using sequential baths of EtOH: H_2O (70:30, *v/v*) for 30 s, EtOH 100% for 30 s, Carnoy solution (EtOH/Chloroform/Acetic acid, 3:6:1, *v/v/v*) for 2 min, EtOH 100% for 30 s, TFA 0.1%/H_2O for 30 s and EtOH 100% for 30 s. After drying the sections, SA-Aniline (SA-ANI) MALDI matrix was deposited on tissue. SA-Aniline was prepared by dissolving sinapinic acid matrix at 10 mg/mL in ACN/TFA 0.1% (50:50, *v/v*) and adding 24.3 µL of aniline. The HTX parameters included a temperature of spray at 75°C with 10 psi pressure, a pattern CC, a flow rate of 0.1 mL/min, a velocity of 1100 mm/min, a temperature of tray at 55°C, for 8 passages with 2 mm track spacing. The slides were analyzed on the MALDI-TOF Rapiflex Tissuetyper mass spectrometer. MS spectra were acquired in the positive linear delayed extraction mode, on the *m/z* 2400-30,000 range with an average of 700 laser shots per pixel and at a spatial resolution of 50 µm. The laser energy was set around 90 %. The voltages of the ion source were 20 kV and 9 kV for the lens.

Peptide MALDI MS Imaging

For peptide imaging, the slides were dried and delipidated using a similar protocol as for protein MS Imaging. The tissue sections were then submitted to trypsin digestion. The tryptic digestion was performed by applying trypsin (40 µg/mL in NH_4HCO_3 50 mM). The HTX parameters included a temperature of spray at 65°C with 10 psi pressure, a pattern CC, a flow rate of 0.1 mL/min, a velocity of 1100 mm/min, for 12 passages with 2 mm track spacing. Once the trypsin was deposited the slides were incubated overnight at 56°C in a humidified box containing MeOH/H_2O. The slides were then dried under vacuum over the next day. An HCCA-aniline matrix was deposited by the HTX M5-Sprayer. Briefly, 43.2 µL of aniline were added to 5 mL of a solution of 10 mg/mL HCCA dissolved in ACN/TFA 0.1% (7:3, *v/v*). Slides were analyzed on a MALDI-TOF Rapiflex. Spectras were obtained in the positive delayed extraction reflector mode analysis, with a mass range of 700-3200 *m/z*, and averaged from 500 laser shots per pixel for a spatial resolution of 50 µm. The laser energy was set around 40 %. The voltages of the ion source were 20 kV and 11 kV for the lens.

Multi-Omics MSI Segmentation

The raw MALDI MSI data for lipids in both ionization modes, peptide and protein data were initially converted into the imzML format (Römpp et al., 2011b) using SCiLS lab software.

Subsequently, the imzML converter, version 1.3.3, was employed to import these datasets into MATLAB R2019a. It's worth noting that MSI data is characterized by high dimensionality, often reaching sizes of up to 100 GB per image. This magnitude makes it infeasible to analyze such data. To address this issue and prevent data loss using peak list generation, data compression was implemented as a preprocessing step before segmentation. Several data reduction (compression) algorithms were explored, including t-SNE (t-distributed Stochastic Neighbor Embedding), NNMF (Non-Negative Matrix Factorization) and SVD (Singular Value Decomposition). For the segmentation process, the *k*-means++ algorithm was utilized, implemented as the '*k*-means' function in the MATLAB Statistics Toolbox. *K*-means++ offers an improved initialization of centroids, enhancing the quality of clustering (Arthur & Vassilvitskii, n.d.-b). The cosine distance metric was employed to calculate the cosine angle between two spectra for quantifying the similarity. For visualization, each cluster's pixels are uniformly assigned a specific color, facilitating the creation of a segmentation map. This map delineates the cluster or region of interest to which each pixel (spectrum) belongs. To estimate the right numbers of clusters, the Silhouette criterion was used. After predefining the number of clusters, the silhouette plot method was used to assess the stability of the clusters. The silhouette plot displays a measure of the proximity of each point in a cluster. This measure has a range (-1, 1). A value close to 1 indicates that the cluster is distant from neighboring clusters (the spectra are very compact within the cluster to which it belongs and distant from other clusters). A value of 0 indicates that the sample is very close to the decision boundary between two neighboring clusters (overlapping clusters). Negative values indicate that these samples may have been assigned to the wrong cluster (Rousseeuw, 1987). Silhouette plot was calculated using the function silhouette in MATLAB. Subsequently, each centroid within these clusters is thoughtfully exported in CSV format, ready for further in-depth analysis and exploration.

Differential Analysis Between Clusters

The centroids generated from the image segmentation were imported into Python using the panda's library. All centroid data was structured into a data frame. A custom script was developed to automate the execution of a statistical test. This script iterates over all *m/z* variables, identifying ions that exhibited statistical significance between the regions of interest (ROIs). To enhance data quality, a peak picking algorithm was employed. Specifically, the find_peaks_cwt function from the Scipy library was utilized to effectively remove instrument noise. A non-parametric statistical test, the Kruskal-Wallis test with Bonferroni correction, was conducted. Only features deemed statistically significant, with a p-value equal to or less than 0.05, were retained. A manual step is added to isolate and retain only the mono-isotopic peaks. The seaborn library was utilized to generate corresponding box plots.

Prediction Model Based on Lipid MALDI Imaging and Associated Proteins Pathways

The previously developed pipeline served as the foundation for constructing the optimal model adapted to the dataset based on highest accuracy and F1-score. These predictive models are designed to classify new MSI-lipid samples pixel by pixel, or the centroid of clusters after segmentation. While models cannot directly predict protein pathways, clusters previously associated with detected proteins using spatially resolved proteomics can indicate these pathways. Therefore, a logical algorithm was integrated into the prediction process. When a model predicts clusters, it also highlights the associated pathways and the corresponding list of proteins.

The three selected models for both rat brain optimization and glioblastoma applications were Stochastic Gradient Descent (SGD) (Ketkar, 2017), RidgeClassifier (Dijkstra, 2014) and Light Gradient Boosting Machine (LGBM) (Ke et al., n.d.). The **Table 8** summarize the performance of each model in both rat brain and glioblastoma analysis. In addition, LIME (Local Interpretable Model-agnostic Explanations) was used for each model to understand the decision-making process of the models and thus identify the molecules that contribute most to predicting each cluster. The highest-contributing molecules are considered potential biomarkers.

Table 8: Model algorithms implication.

Model	Algorithm	F1 score
RB cerebellum clusters lipid (-)	SGD	94%
RB cerebellum clusters lipid (+)	RidgeClassifier	98%
GBM lipid classification	LGBM	97%
GBM protein classification	RidgeClassifier	96%

Lipid annotation by SpiderMass Technology

The basic design of the instrument setup has been described in detail elsewhere (Saudemont et al., 2018). In addition, here, the laser system used was an Opolette 2940 laser (OPOTEK Inc., Carlsbad, California, USA). The infrared laser microprobe was turned at 2.94 µm to excite the most vibrational band of water (O-H). The laser beam was injected into a 1 m reinforced jacketed fiber of 450 µm inner core diameter equipped at its extremity with a handheld including a focusing lens with 4 cm focal distance to get a 500 µm spot on the tissue. To aspirate and analyze the ablated material, a Tygon® tubing (Akron, OH, USA) is directly connected to Q-TOF mass spectrometer (Xevo, Waters, UK) through a REIMS interface. Each rat brain cerebellum clusters, observed by MSI, were analyzed by SpiderMass with four independent biological repetitions. Briefly, the laser was directly placed above the region of interest at the 4 cm focal distance. The laser energy was fixed to 4 mJ/pulse(Ledoux et al., 2023). On each spot, three acquisitions of 10 repetitive laser shots (10 Hz) were performed which resulted in 3 individual MS spectra. The data were acquired in both negative and positive polarities, in the sensitivity mode over a *m/z* 100-2000 range. The previously identified discriminative ions were

selected for MS/MS analysis with 0.1 *m/z* isolation window. MS/MS was performed using collision induced dissociation (CID) with argon as collision gas and a collision energy of 25 eV.

Spatially Resolved Proteomics Extraction

The different clusters identified by the segmentation process were submitted to spatially resolved proteomics. Each cluster was analyzed in triplicate from the same tissue section as describe bellow. A localized digestion was carried out by deposing a trypsin solution (40 µg/mL in NH_4HCO_3 50mM), on a region of 500 μm^2 of tissue (4 x 4 droplets of 200 µm in diameter), using CHIP-1000. The deposition method comprises approximately 1205 cycles per digestion spot, i.e., 3 hours of deposition, with a drop volume of 150 pL. Finally, each spot was digested with 0.112 µg of trypsin. Following the micro-digestion, each spot was extracted by liquid micro-junction using the TriVersa Nanomate device, with LESA (Liquid Extraction and Surface Analysis) parameters (Quanico et al., 2013). The tryptic peptides were extracted by performing 2 consecutive extraction cycles for three different solvents mixtures (TFA 0.1%; ACN/0.1% TFA (8:2, *v/v*); and MeOH/0.1% TFA (7:3, *v/v*)) for a total of 6 extractions. For each cycle, 2 µL of solvent was drawn into the tip of the pipette, of which 0.8 pL was brought into contact with the surface. 15 back and forth movements were performed to extract the peptides before collecting the solution in a recovery tube. All extracts were pulled in one tube and 50 µL of ACN were finally added before drying the samples in a SpeedVac. The samples were then stored at -20°C prior to nLC-MS/MS analysis.

nLC-MS/MS Bottom-Up Analysis

All sample analysis was performed on a timsTOF fleX mass spectrometer online coupled to an Evosep One nano-flow liquid chromatography system. Peptides were separated using an 8 cm x 150 µm C18 column with 1.5 µm beads and the 60 samples per day method from Evosep One. The mobile phases comprised 0.1% FA in water as solution A and 0.1% FA in ACN as solution B. To perform DIA analysis in PASEF mode(Meier et al., 2020), one MS1 scan was followed by 10 dia-PASEF scans from *m/z* 100 to 1700. The ion mobility range was set to 1.42 and 0.65 $V.s/cm^{-2}$. The accumulation and ramp times were specified as 100 ms. As a result, each MS1 scan and each MS2/dia-PASEF scan last 100 ms plus additional transfer time, and a dia-PASEF method with 22 dia-PASEF scans has a cycle time of 1.06s. The mass spectrometer was operated in high sensitivity mode, with a collision energy ramped linearly as a function of the ion mobility from 59 eV at $1/K_0$=1.6$Vs.cm^{-2}$ to 20 eV at $1/K_0$=0.6$Vs.cm^{-2}$. The ion mobility was calibrated with three Agilent ESI Tuning Mix ions (*m/z*, $1/K_0$: 622.02, 0.98 $V.cm^{-2}$, 922.01, 1.19 $V.cm^{-2}$, 1221.99, and 1.38 $V.cm^{-2}$).

Proteomic Data Analysis

DIA-NN version 1.8.1 was used to search DIA raw files and dia-PASEF files. A Rattus library was generated with the software parameters set as following: complete proteome of Rattus

norvegicus from UniProt database (Release January 2024, 92958 entries), Trypsin protease with 2 missed cleavages and a maximum number of variable modification at 3, methionine oxidation as variable, peptide length range from 7 to 30, precursor charge range from 1 to 4, precursor *m/z* range comprised between 100 and 1700, fragment ion *m/z* range between 200 and 1700, 0.1% precursor FDR, protein inference set on 'genes', neural network classifier on single-pass mode, quantification strategy set on robust LC (high accuracy), RT-dependent cross-run normalization, and library generation fixed on the 'IDs, RT & IM profiling' ruban. Samples were interrogated according to the resulting Rattus library with the same options. Data are available via ProteomeXchange with identifier PXD054488**.** Statistical analyses were carried out using Perseus software v2.0.5.0. ANOVA tests were performed with p-value ≤ 0.01 to be statistically significant and generate heat maps of differentially expresses proteins across sample. Gene Ontology (GO) analysis were performed using ClueGO (Bindea et al., 2009) with GO term database, on Cytoscape v3.10.2 (Shannon et al., 2003).

Results

The main goal of this study was to develop a machine learning pipeline capable of automatically identifying tissue heterogeneity clusters from lipid MSI data and providing associated protein networks without requiring additional protein experiments. To this end, the first challenge was to demonstrate that identified clusters are specifically spatially localized by MSI, regardless of whether lipid or protein imaging is used. Following this idea, if a cluster is identical on these omics images, it should possess specific lipid and protein pathways, tending to the basis of the dry proteomics concept.

Segmentation Workflow Development on RB Cerebellum Omics MSI.

Clustering Multi-omics MALDI MSI Workflow Optimization

The machine learning clustering processing was the first development to adapt a workflow for multi-omics MALDI MSI. This step was focused on RB cerebral lump, a model whose anatomical and molecular characteristics are already well referenced. For the latter, four main clusters are described (**Figure 30**): the white matter (WM) and the grey matter (GM), composed of the molecular layer (ML), Purkinje cells and the granular layer (GL). The first aim was to demonstrate that these clusters could be observed with the same spatial localization in each omics image, using an adapted segmentation process script.

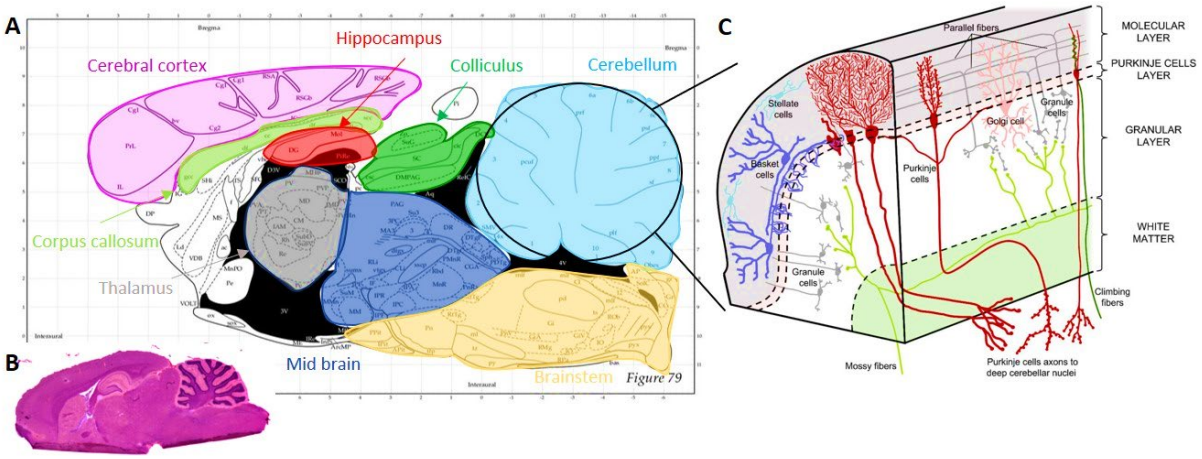


Figure 30: Rat brain anatomy of sagittal section, with A) Atlas annotations B) HPS coloration and C) cerebellum layers (Marcos et al., 2023).

For that, 22 RB sagittal sections were analyzed for lipid in negative (-) and positive (+) ion mode, while 12 slides were analyzed for protein and peptide, focusing on the RB cerebellum area. First, the MS spectra revealed different molecular fingerprints regarding WM, GL, and ML clusters for each molecular MSI (**Appendix B, Figure 64**), confirmed for lipid (-) and protein data by t-SNE revealing clear separations of the different ROIs. On the contrary, the t-SNE obtained for lipid (+) and peptides did not show a clear separation of the different ROIs, which could predict difficulties for data processing of the latter.

To generate the most relevant segmented images, the image data was first analyzed on SCiLS lab software using RMS (Root Mean Square) normalization. The SCiLS software allows to play with different clustering parameters. Several segmentation methods were tested, including bisecting *k*-means, hierarchical clustering and *k*-means segmentation using correlation or Euclidean distance metrics. As shown in **Appendix B, Figure 65**, the use of bisecting *k*-means and hierarchical clustering were ruled out due to the difficulty of interpreting the results for several reasons. First, the complexity of manually determining the desired number of clusters, which can be difficult in the case of a complex and unknown image. In addition, the spatial connectivity limitations of bisecting *k*-means do not adequately account for the connectivity between pixels in an image. This oversight can lead to segmentation discontinuities that undermine the overall accuracy and coherence of the segmentation process. *k*-means segmentation appeared to be more user-friendly, with multiple clusters defined subjectively. Unfortunately, it seems that poor centroid initialization led to insufficient clustering performance, rendering the segmentations of lipid, protein, and peptide images incomparable despite using the same number of clusters.

To find a more transparent and robust strategy, data from SCiLS was imported into MATLAB software. To improve the previous clustering performance, the *k*-means++ segmentation algorithm with cosine distance metric was used. This algorithm ensures more intelligent centroid initialization,

thereby improving the overall quality of the clustering. Beyond the initialization step, the rest of the algorithm remains consistent. To overcome the high dimensionality MSI data problem and to avoid data loss due to peak list generation, a pre-processing step involving data compression was introduced prior to segmentation.

For this purpose, several data reduction algorithms were investigated, including t-SNE, NNMF and SVD. As for PCA, t-SNE and NNMF are common preprocessing methods used for MSI data processing, compared to SVD. PCA and SVD are known to be suitable for linear dimensionality reduction and preserving global structure, NNMF is useful for non-negative data and part-based representation, while t-SNE excels in visualizing high-dimensional data. As shown in **Figure 31C**, SVD compression was found to be optimal. Indeed, even if t-SNE presented good segmentations for lipid (-) and peptide images, it was difficult to distinguish the GL from the ML and WM in lipid (+) and protein cases. The results using NNMF (**Figure 31B**) and SVD (**Figure 31A**) were correct, observing the three RB areas in each omics image. It can be added that the images generated by the latter have a better resolution and are more looking alike. Therefore, the SVD compression was kept for the future to obtain the best possible segmentations.

It is noteworthy that within the context of this investigation, only three out of the four primary cerebellum clusters were discernible using a 50 µm MSI spatial resolution: the ML, WM, and GL in conjunction with Purkinje cells. Additionally, lipid cerebellum images were captured at a finer resolution of 10 µm (**Figure 31D and E**) and subsequently processed, thereby confirming the distinct visualization of all four cerebellum clusters. This underscores the crucial role that spatial resolution plays in the generation and differentiation of clusters, yet the spatial resolution was fixed at 50 µm since proteins imaging needs a higher resolution to get enough signals. Despite the potential for finer resolution to improve cellular component discrimination within the cerebellum, it was pragmatically determined that the 50 µm resolution sufficed for the objectives of this study, given the constraints and goals at hand.

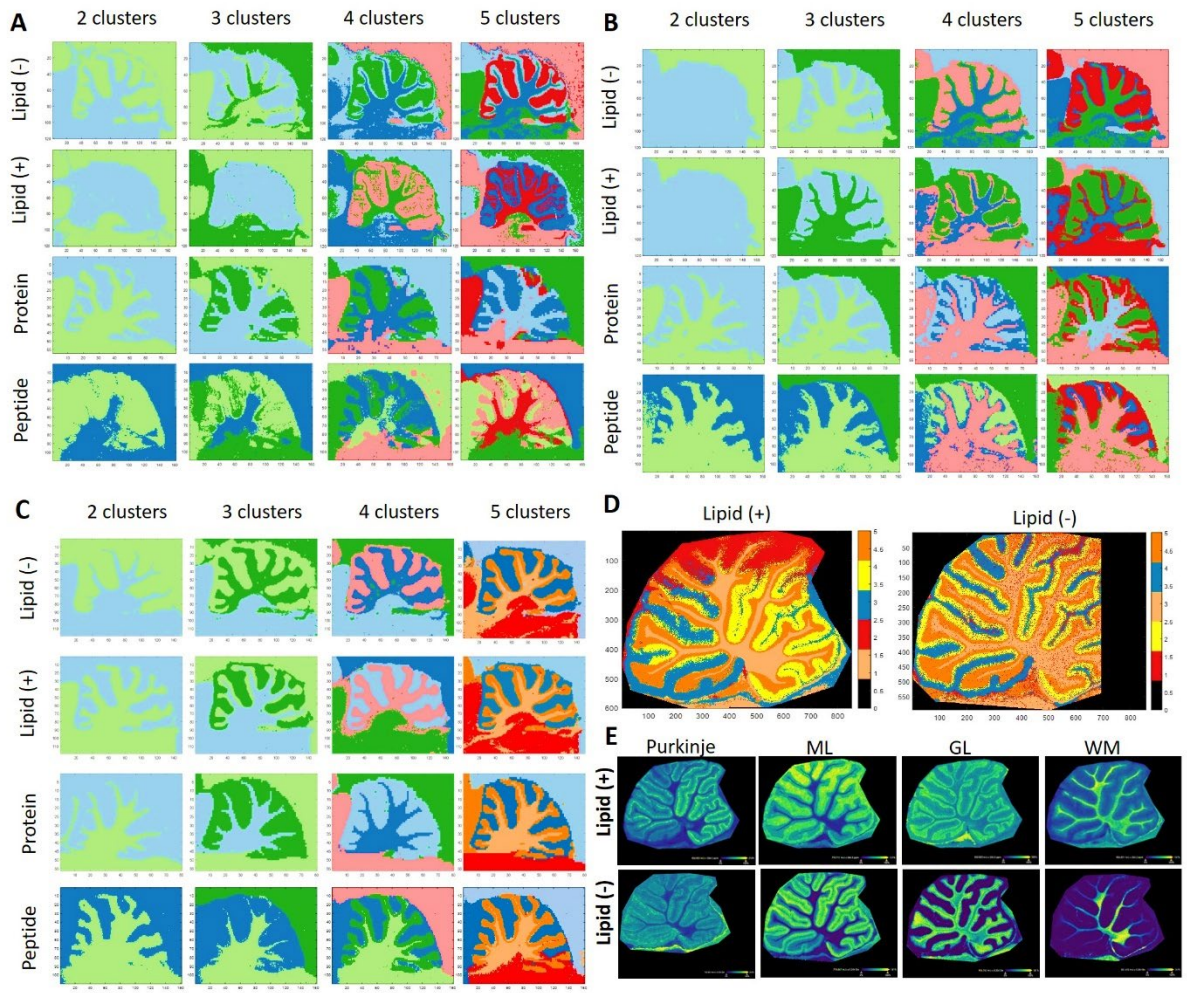


Figure 31: Omics MALDI MSI clustering procedure optimization on rat brain cerebellum. Comparison of A) t-SNE, B) NNMF and C) SVD data compression followed by *k*-means++ segmentation for 2 to 5 clusters applied to lipid negative mode, lipid positive mode, protein, and peptide MSI. D) Lipid MALDI MSI in negative and positive mode with 10 µm spatial resolution with image segmentation composed by 5 clusters, and ion spatial distribution specific to Purkinje cells, ML, GL and WM. E) Distribution of lipid (-) and lipid (+) ions with specific spatial distribution in Purkinje cells, ML, GL and WM from lipid MALDI MSI with 10 µm spatial resolution.

Unsupervised Cluster Number Estimation

It was demonstrated that the three regions of the cerebellum can be observed in a similar way on lipid or protein image constructed with five clusters. However, the choice of the number of clusters was made in a semi-supervised manner. To automate the process of lipid-based proteomics, it was necessary to implement a tool capable of estimating the optimal number of clusters. To estimate the correct number of clusters in a non-subjective way, the silhouette criterion was used. The advantage was that it can be used multiple times, both to find the optimal number of clusters and to assess their stability and compactness.

As shown in **Figure 32A**, Silhouette estimated the optimal number of clusters at 5 for the lipid images, which was a coherent result with respect to the previously selected semi-supervised number. Furthermore, the fact that the same results were obtained for the lipids in negative or positive mode was expected due to their identical nature and metabolism. The 5 estimated clusters included 4

corresponding to the ML, GL, WM and brainstem regions of the rat brain, while 1 cluster represented a tissue-free area containing only matrix. These clusters were also observable for protein and peptide images with 5 clusters.

The same was true for predictions using protein and peptide data sets, which yielded a consistent and identical number of clusters. The heterogeneity predicted for peptides and proteins was more important, with a cluster number between 9 and 10 (**Figure 32A**). This can be explained by the diversity of proteins compared to lipids. Proteins are made up of a combination of 20 different amino acids, which may explain the presence of more protein clusters in the depth of the tissue compared to what is observed by lipid imaging or immunohistochemistry**.** Moreover, it is suggested that the over-segmentation in protein and peptide data may result from artifacts, particularly in the tissue-free regions. While a single cluster was expected to represent the matrix, as observed in the lipid data, three distinct clusters were instead identified, likely due to inhomogeneous crystallization resulting from the nature of the matrix (i.e., Norharman for lipids and HCCA-aniline vs. SA-aniline for proteins). HCCA-aniline and SA-aniline are ionic matrices based on two components, which explains the presence of three clusters instead of just one (corresponding to HCCA, HCCA-aniline, and aniline clusters, or for SA, SA-aniline and aniline clusters). Considering that the nature of the ionic matrix in proteins and peptides results in the formation of three additional clusters, these can be removed, leading to a final total of seven clusters related to the tissue. Subdivisions were also observed in two clusters for the molecular layer (possibly linked to the presence of Purkinje cells in some pixels) and the brainstem, which were also identified in lipid images containing ten clusters. Thus, a total of seven clusters were found for lipids and seven clusters for peptides and proteins, as it can be seen in the **Figure 32A** for the 10 clusters images, still suggesting a degree of concordance between lipid and protein clustering in tissue regions.

For a simpler process, dry proteomics on lipid images was used because lipid imaging does not require additional sample preparation steps, protects the tissue from artifacts and potential degradation, and is less time consuming for routine analysis. Consequently, the clusters identified in lipid images are more representative of the RB cerebellum anatomy.

In the present case, the principle of dry proteomics through lipid imaging was relied upon, and therefore, the five cluster omics images were selected for the remainder of the study.

Finally, the optimal segmentation workflow developed was a MATLAB script (**Figure 32B**), integrating a SVD compression of data with 10 principal components, combined with a *k*-means++ segmentation using a cosine distance with a silhouette criterion. This approach allowed the visualization of the three main clusters of the RB cerebellum (ML in blue, GL in orange, WM in light orange), in an identical and specific spatial localization, from the 5 cluster images respectively

generated for: lipid (-) and (+) MSI with Norharman matrix , lipid (+) MSI with DHB matrix , protein MSI and peptide MSI , with semi-supervised observation.

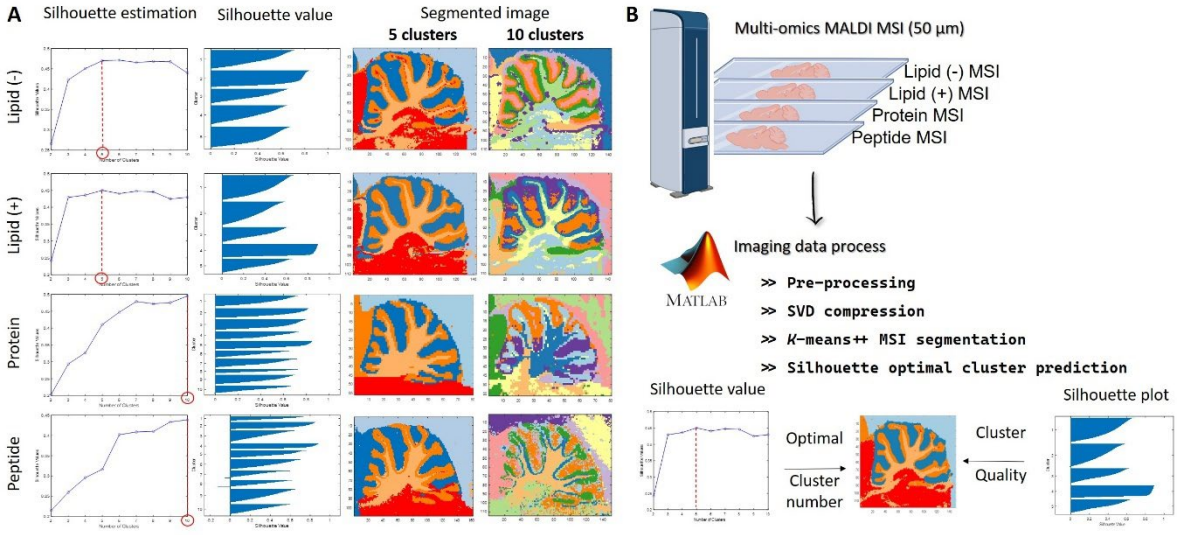


Figure 32: Silhouette Analysis and Multi-Omics MALDI MSI Segmentation Using *k*-means++ Clustering. A)Use of Silhouette criterion for the number of cluster estimation and each cluster value determination applied to lipid negative mode, lipid positive mode, protein, and peptide imaging. B) Optimal segmentation workflow developed on MATLAB integrating a SVD compression data with 10 principal components, combined with a *k*-means++ segmentation using a cosine score with a Silhouette criterion.

Prediction Model on Lipid MALDI Imaging

To automatically identify each cluster present in a tissue from a lipid image, a machine learning algorithm was trained on the 22 positive and negative lipid imaging datasets previously obtained. The ML, GL and WM centroids were extracted from the 5-cluster segmented lipid images and imported into Python. The datasets were subjected to peak picking and a non-parametric Kruskal-Wallis test to compare the significance of each ion between each ROI. Only features with a p-value equal to or less than 0.05 were retained as discriminant ions . After isotope filtering, a final list of 36 lipid (-) and 19 lipid (+) discriminant ML, GL and WM ions were identified (**Figure 33A-B**). The spatial distribution of each ion also confirms its specificity to its assigned cluster (**Figure 33D**). The annotation of the discriminant lipid ions was performed by SpiderMass MS/MS experiments, as its highest lipidomic similarities with the MALDI (Ledoux et al., 2023). All the specific ions of the lipids in a region are listed in an internal database, which is used to predict regions on MALDI lipid images. Various prediction models were evaluated, respectively for each lipid mode analysis, taking in account the discriminant ions previously set out. In case of lipid (-) datasets, the SGD (Drucker, 1997; Hastie et al., 2009) algorithm was selected as optimal model and was validated using a 5-fold cross-validation (Kohavi, n.d.) with an accuracy of 94%. The Ridge classifier model was the one adapted to the lipid (+) datasets, with 98% accuracy after 5-fold cross-validation. The robustness of the developed models were subsequently evaluated by blind cohort validation, which included three

different datasets of cerebellum RB lipid images for both polarity modes. Notably, in each instance, the model achieved 100% accuracy in its classifications.

The same data processing was performed on the protein imaging datasets to provide the corresponding discriminant protein ions for each RB cerebellum clusters (**Figure 33C**). The list of discriminant protein ions was then added to discriminant lipid ions for each cluster in order to create discriminant protein and lipid ions dataset reference for each RB cerebellum area.

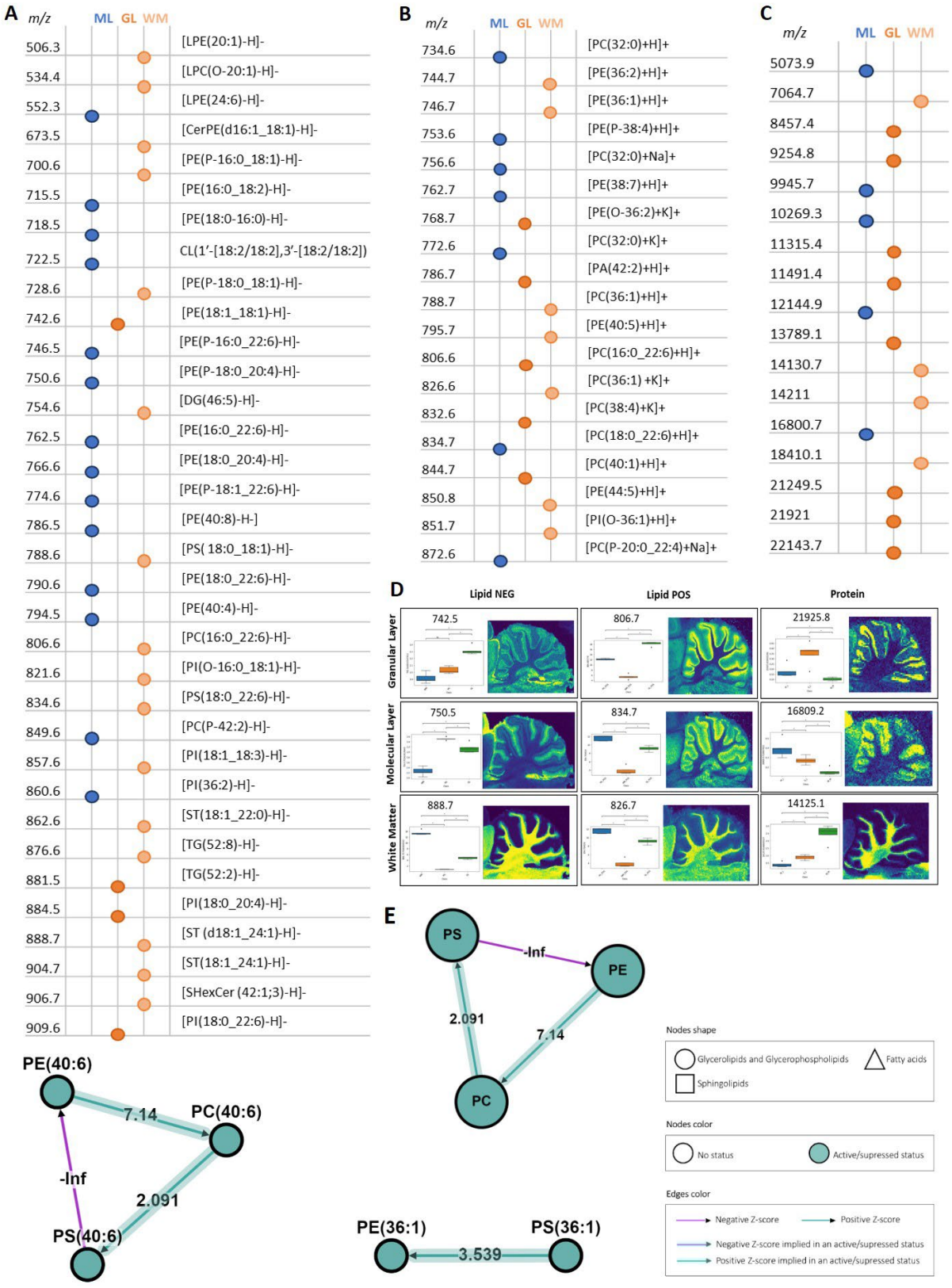


Figure 33: Discriminant lipid and protein ions present in RB cerebellum with BioPAN lipid pathways. Exhaustive list of A) 36 lipid (-), B) 19 lipid (+) and C) protein discriminant ML, GL, and WM cerebellum ions. D) Distribution of lipid (-) and lipid (+) discriminant ions with specific spatial distribution in ML, GL and WM. E) BioPAN biological lipid pathways involved in white matter represented according to lipid species and lipid classes, with nodes legend.

Lipids Biological Network Analysis

Based on the compilation of annotated lipids, a greater number of lipids have been specifically identified in WM compared to GM. This observation is in line, considering that the WM is predominantly comprised of myelin, a substance containing a higher lipid content (78-81%) than both white (49-66%) and grey matter (36-40%) (O'brien et al., 1965). In the same way, myelin is composed of a high percentage of galactoceroboside and cholesterol compared to GM, which is why more diglycerides (DG), triglycerides (TG) and fatty compounds are identified in the latter. On the other hand, GM presents a higher percentage of phosphatidylethanolamines (PE) and phosphatidylcholines (PC), which correlate with presented annotations (O'brien et al., 1965).

To highlight the biological process involved by lipid data, a comparison between WM and GM discriminant lipid was performed on BioPAN (Gaud et al., 2021). The results, shown in **Figure 33E**, revealed PC biosynthesis as the most active pathway in WM (with the involvement of PEMT predicted gene), whereas PE biosynthesis was observed as the most active pathway in GM (with the involvement of PISD predicted gene). These results clearly indicate discriminants lipids involved in specific biological pathways associated to distinct cerebellum regions.

Biologically, phosphatidylcholine is an essential choline reservoir for brain function (Blusztajn et al., 1987). In fact, choline is an important molecule for neurotransmission in neurons, which may explain the high activation of PC biosynthesis in WM. Phosphatidylethanolamine's biological function is more due to its small chemical structure, which allows fluidity of the neuronal membrane(Lohner, 1996). The hypothesis is that this facilitates vesicle budding and membrane fusion (Glaser & Gross, 1995), a key step in synaptic transmission in GM. Finally, biological pathway based on lipids analysis showed that PC may be involved in the neurotransmission process in WM, whereas PE is more involved in synaptic transmission in GM (Tracey et al., 2021). These conclusions were further corroborated by Reactome analysis of the lipid dataset, which demonstrated their involvement in the neural system, signal transduction, small molecule transport or metabolism of protein and vesicle-mediated transport pathways.

Consolidation Method by Protein Pathway Analysis

As discriminant biological pathways were defined for different regions of the cerebellum RB based on lipid species, the proteomes of these regions were analyzed to validate the hypothesis of a correlation between lipids and proteins within the same biological network. This analysis aimed to consolidate the dry proteomics processes.

The ML, GL and WM regions observed in the multi-omics MALDI MSI were therefore subjected to spatial proteomics using the micro-proteomics workflow on three different RB sections (Mallah et al., 2019, 2023). By regrouping the triplicates for each cluster, a total of 5270 proteins were identified for

WM, 5390 for GL and 5354 for ML. This study confirmed the spatial heterogeneity of proteins previously observed in imaging. The results showed that discriminant lipid species for each ROIs are consistently linked to specific proteins in the same ROI, thereby forming region-specific pathways and functions.

Indeed, the Venn diagram, shown in **Figure 34A**, considers the protein diversity between each region by the presence of proteins exclusive to each of them. In total, 85 proteins were exclusive, of which 7 were specific for WM, 11 for ML and 67 for GL. It must be noted it was found among the 11 specific proteins in ML, two important enzymes involved in lipids metabolism *e.g.* Phosphoinositide phospholipase C and Inositol monophosphatase 1 whereas in GL, the Gamma-butyrobetaine dioxygenase know to catalyze the formation of L-carnitine and the Plasmolipin in WM, a main component of the myelin sheath involved in intracellular transport, lipid raft formation, and Notch signaling were identified (Shulgin et al., 2021). The GL contain several neuropeptides or neuropeptide hormone activity such as Corticotropin-like intermediary peptide, Somatostatin-14; Pro-thyrotropin-releasing hormone, cholecystokinin-12; Neurokinin-B, Cocaine- and amphetamine-regulated transcript protein; Pituitary adenylate cyclase-activating polypeptide 27 or Ephexin-1 (Z. H. Li et al., 2024). In ML, among the identified protein the lamin B-binding protein (BAF: Barrier-to-autointegration factor) and Myogenin are of particular interest. In fact, BAF is required during brain development as a regulator of nuclear migration during neurogenesis of the CNS (Evangelisti et al., 2022). Myogenin is also detected in Allan brain atlas and is linked to motor neurons (Ayasoufi et al., 2023). Similarly, in WM among the specific proteins identified, the Lymphocyte specific 1 is recently known to be correlated with tissue resident memory T cells (Ayasoufi et al., 2023) and T cell infiltration (Batterman et al., 2021). Interestingly, Phosphatidylserine decarboxylase proenzyme (PISD) was also found in both WM and ML regions and was a predicted gene previously reported in BioPAN GM lipid pathway. The presence of PISD protein may explain the amount of PE identified in the ML region. In this context, PE may contribute to the integrity and function of neuronal membranes, influence synaptic transmission, and participate in signaling events. This again demonstrated the relevance of the different clusters by MSI, which predicted their own lipid/protein pathway and therefore biological heterogeneity.

To go further, the common proteins were subjected to an ANOVA test (p-value < 0.01) and showed that 2204 out of 5465 proteins have a significant variability of expression. According to Allan brain Atlas, based on transcriptomic analyses, 196 genes are Cerebellum enriched gene and 59 out of those genes show highest expression levels in cerebellum. 90% of their corresponding proteins have been identified such as CBLN1 and CBLN3. Among them, some are known to be specifically located to the Purkinje layer which was regrouped with the GL after clustering**.** Specific proteins were identified from the Purkinje cells (MYH10, HOMER3, KIT, QKI, MX1, PCP-2, PP1R17, ARGEF33). For example, QKI

protein expressed by radial astrocytes (Bergmann glia) with processes through the molecular layer all the way to the pial surface of the cerebellar cortex has been identified. MX1 is known to be in the dendritic processes of Purkinje cells. Moreover, other specific proteins of granular layer, GABRB2, TMEM6 and KCNIP4, markers of synaptic glomeruli from granular cells are also detected.

This was reflected by the presence of different clusters of over- or under-expressed proteins between each RB cerebellum area (**Figure 34B**). The gene lists corresponding to over-expressed protein clusters were analyzed using ClueGO software to identify the biological pathways associated with the significant proteins identified in each distinct cluster. It turns out that the overexpressed proteins in the WM are mainly involved in myelination, glucose and neurofilament metabolism (**Figure 34C**), which is a consistent result according to the bibliography (Gianola et al., 2003). In fact, WM consists of myelinated axons, so it's involved in the transmission of nerve impulses by axons. The presence of glucose metabolism is also interesting when correlated with the galactoceroboside myelin composition previously suggested by lipid WM analysis. Furthermore, iron metabolism is another important biological process in the white matter, e.g. for myelin formation, redox reactions or neuronal development and synaptic plasticity (Hulet et al., 1999, 2002; Kirilina et al., 2020). This information can be linked to biological pathways previously found by lipids analysis, which also highlighted the neurotransmission pathway in WM. Regarding the GL, the neuropeptide hormone activity pathway was found to play a role in the processing and regulation of peptides that influence synaptic transmission, neural signaling and modulation of neuronal activity (**Figure 34D**). Purine metabolism also plays a crucial role thereby influencing various physiological processes such as neurotransmission, synaptic plasticity, and energy metabolism. Dysregulation of purine metabolism in the brain has been implicated in several neurological disorders, including epilepsy, Parkinson's disease, and neurodegenerative diseases. Similarly, the relevance of synaptic organization and sodium ion transport pathways involved in the molecular layer (**Figure 34E**) were expected results given their role in neurotransmission and synaptic signaling between these cell types.(Ma et al., 2020). It's interesting to remember that the biological processes of synaptic transmission, vesicle transport and signaling were also predominant pathways in the previous lipid study. Thus, it has been shown that the ML, WM, and GL have their own specific proteome that can be correlated with specific lipid associated to distinct biological pathways.

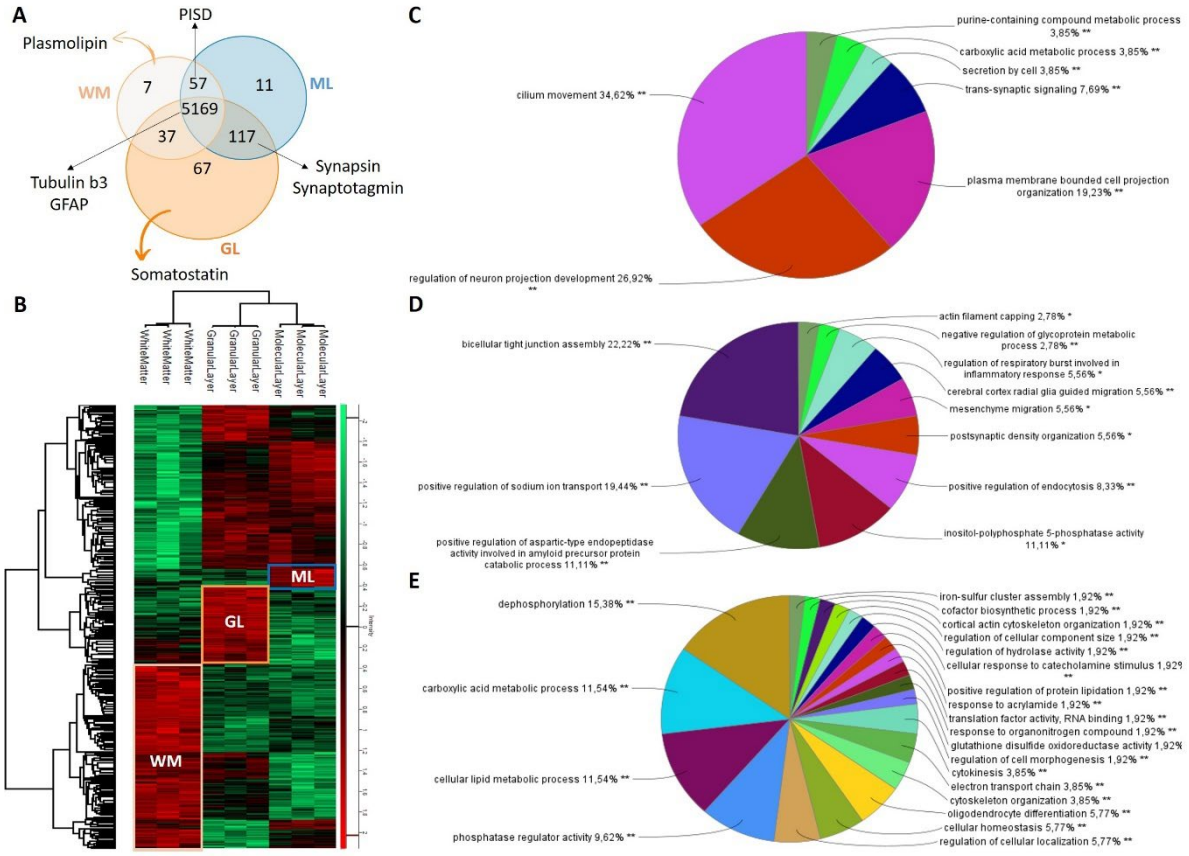


Figure 34: Rat brain cerebellum regions spatial proteomic analysis. A) Venn diagram of the specific proteins per layer. B) Heatmap after ANOVA (p-value <0.01) analysis demonstrated the presence of different of overexpressed proteins. ClueGO biological pathways involving the significant proteins found in C) granular layer, D) white matter, and E) molecular layer of the cerebellum.

Dry Proteomics Based on RB Horizontal Lipid Imaging Application

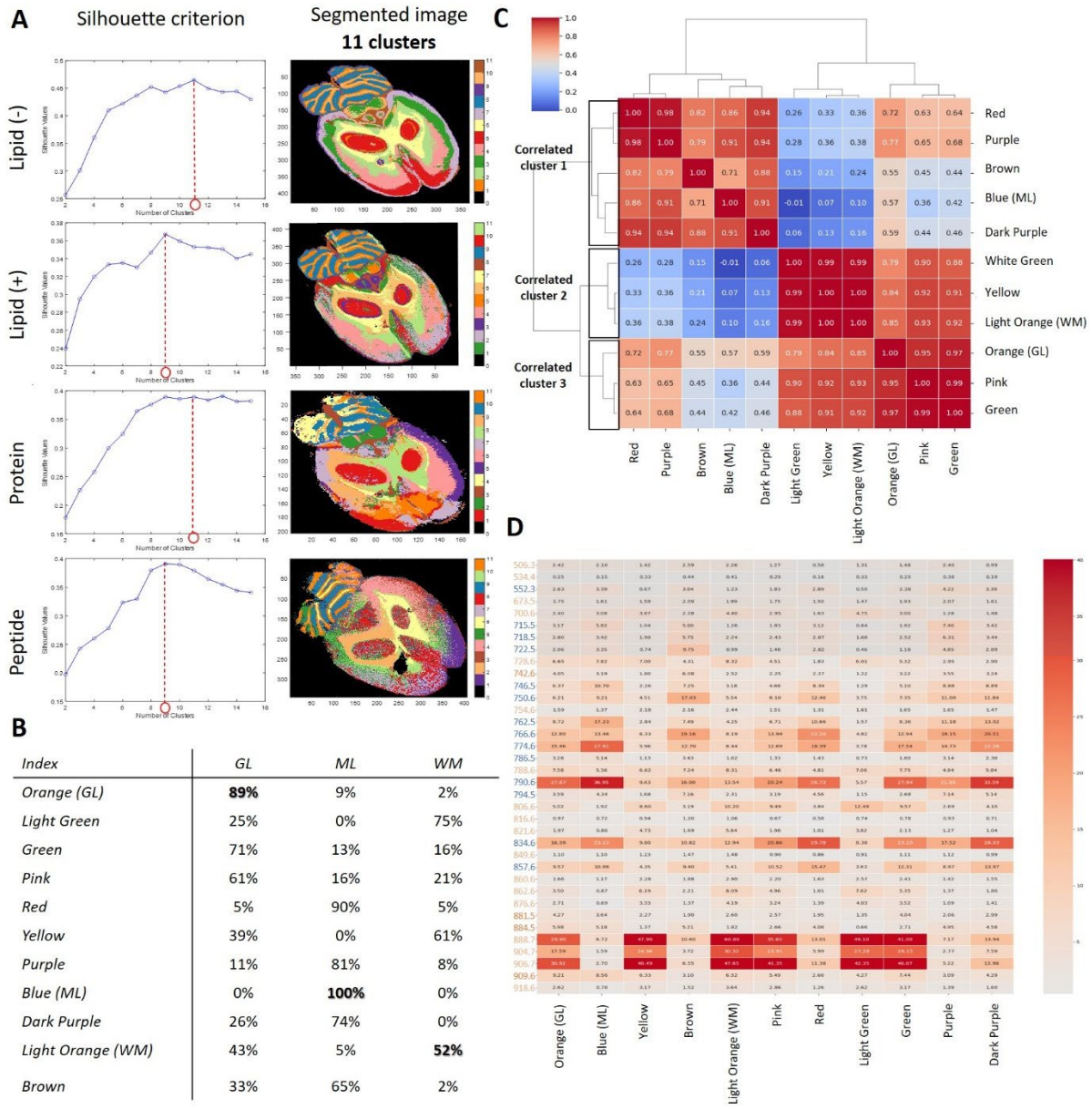
Multi-omics RB Horizontal Sections Generation

To validate the dry proteomics workflow to more complex tissue, the analysis was widened to total horizontal RB sections. As previously, multi-omics MALDI MSI were performed on 4 different sets of consecutive horizontal RB sections, and resulting data were submitted to the imaging data processing workflow, excluding matrix signal. The Silhouette criterion was around 11 for each lipid replicate, leading to multi-omics images composed of 11 clusters (**Figure 35A**). A similar spatial clustering shape was observed for each lipid image, including the well-known areas of the cerebellum, as well as other specific areas such as: the corpus callosum subdivided into clusters white, green and yellow, the cerebral cortex and thalamus in purple, red and pink, and the ventricular system in brown. These specific regions were also observed on the protein and peptide images built with 11 clusters, again confirming the lipid/protein pathway cluster appurtenance.

RB Cerebellum Lipid Classification Model: Prediction on Horizontal Sections

Four replicate lipid (-) horizontal RB images were blindly analyzed using the pre-built classification model trained on 22 RB cerebellum lipid (-) MSI datasets. The model returned a confidence score for predicting each ROI. Since the model was trained on three ROIs, the default

confidence score to predict an ROI (region of interest) was >33%. The model successfully predicted the ML area with a mean confidence score of 100%, WM with a confidence score of 52%, and GL with 89% (**Figure 35B**). For WM, although 52% is significantly higher than 33%, the lower confidence score may be due to the discrepancy in surface area between the sagittal and horizontal brain slices of the rats, with the former showing a significantly greater extent of WM. Other clusters were also analyzed using the predictive model (**Figure 35B**) with interesting results. The light green and yellow clusters (corpus callosum region) were predicted as WM with confidence scores of 75% and 61%, respectively. Similarly, the green cluster (colliculus regions) was predicted as GL with a confidence score of 71%. A Pearson's correlation of the discriminant lipid negative ions, shown in **Figure 35C**, further validated these predictions. Two main clustering branches were identified: one leading to correlated cluster 1 associated with ML, and another leading to two separate clusters, correlated cluster 2 associated to GL and correlated cluster 3 associated to WM. In correlated cluster 1, dark purple and brown ROIs were grouped with ML, sharing the 774.6 and 790.6 lipid (-) ions (**Figure 35D**). In correlated cluster 2, WM was grouped with the yellow and light green ROIs, as predicted by the model, with the main involvement of the 888.7 and 906.7 lipid (-) ions (**Figure 35D**). Biologically, these results were expected. The corpus callosum (light green and yellow clusters) forms the largest commissural WM bundle in the brain, which has a distinct molecular composition due to its significant size and role, explaining the observed clustering (Yamazaki et al., 2016). Similar observations were valuable for the colliculus (green cluster), which also contains a superficial grey layer (Yamazaki et al., 2016). This explains the presence of orange color in both granular layer and colliculus clusters, corresponding to GM, and accounts for the 71% confidence score prediction explaining similarity.



Index	GL	ML	WM
Orange (GL)	**89%**	9%	2%
Light Green	25%	0%	75%
Green	71%	13%	16%
Pink	61%	16%	21%
Red	5%	90%	5%
Yellow	39%	0%	61%
Purple	11%	81%	8%
Blue (ML)	0%	**100%**	0%
Dark Purple	26%	74%	0%
Light Orange (WM)	43%	5%	**52%**
Brown	33%	65%	2%

Figure 35: Horizontal rat brain section omics MALDI MSI analysis. A) Lipid (-), lipid (+), protein and peptide MSI segmentation images with 11 clusters and Silhouette criterion. B) Clusters mean scores prediction based on rat brain cerebellum lipid (-) model. C) Clusters Pearson's correlation. D) Prediction lipid (-) model peaks involvement.

With the aim to justify the images segmentation, discriminant lipid (-) ions were identified for different cluster observed on the horizontal RB section lipid (-) image (**Figure 36**). Many peaks were spatially distributed regrouping multiple clusters. For example, common ions were spatially distributed in ML, cerebral cortex and hypothalamus regions, like *m/z* 790.4, 834.4 and 886.5. The ion *m/z* 599.4 was collocated in GL and green cluster. Same observations for both the WM and the corpus collosum, for which ions as *m/z* 701.6, 889.6 and 904.7 were also spatially present. These ions could explain the correlation clusters highlighted by the Pearson's correlation graph in **Figure 35D**. However, discriminant ions were also extracted for specific regions, explaining their segmentation as single cluster during MSI data processing. The ventricular region (brown cluster) possessed various

discriminant ions such as *m/z* 473.2 and 615.1. Specific ions were also discriminant for the red cluster, regrouping cerebral cortex and hypothalamus regions (*m/z* 746.6, 766.6, and 834.5), as well as for the corpus callosum like *m/z* 806.6. This ions list was added to the prediction model, to refine predictions.

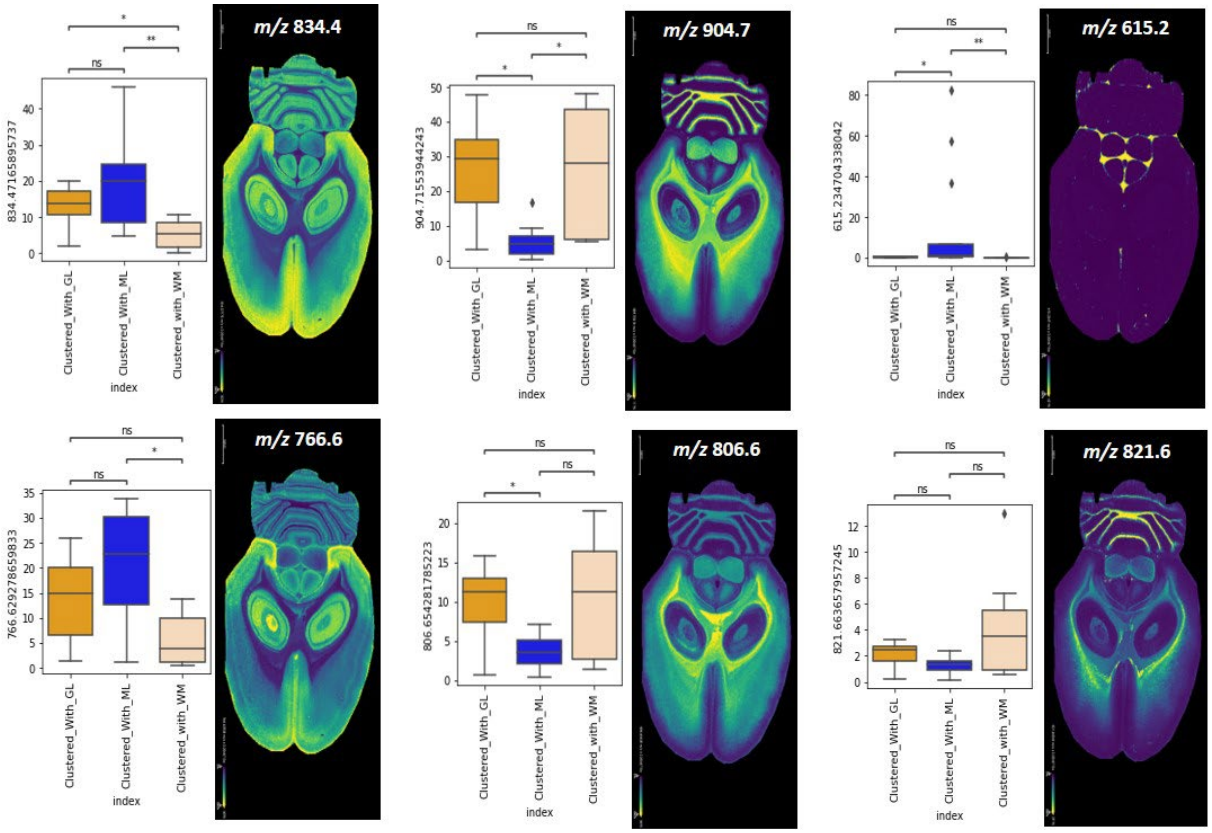
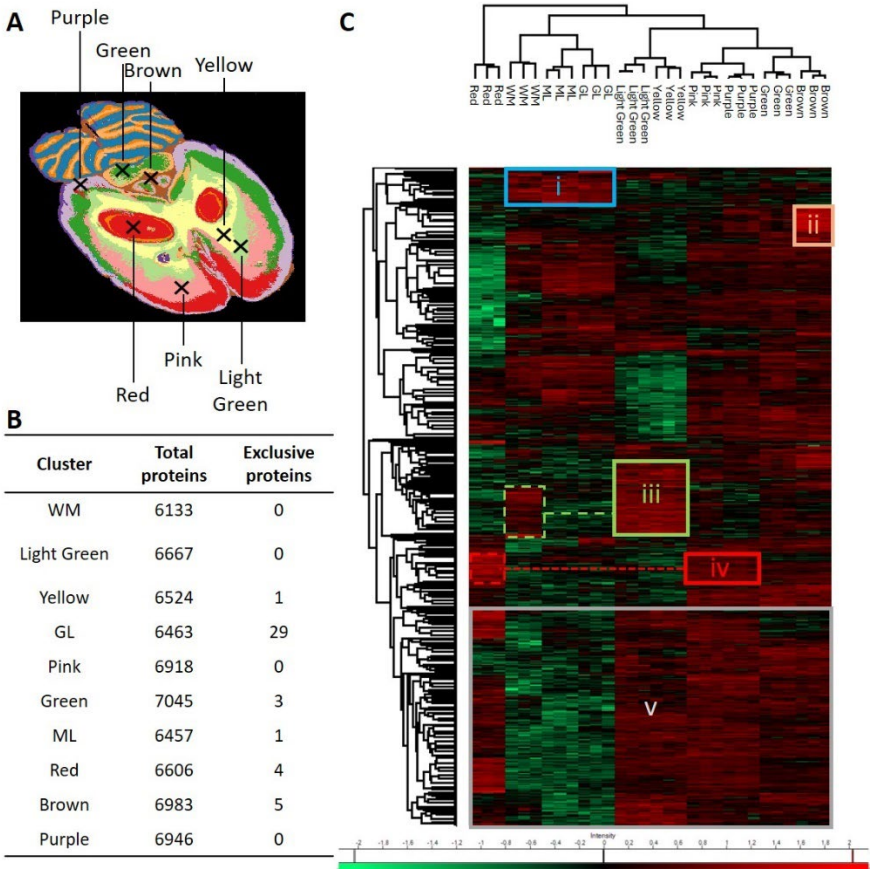


Figure 36: Spatial ion distribution across rat brain regions on horizontal section. Common ions (e.g., *m/z* 834.4, 904.7) were found in multiple areas, including the cerebral cortex and hypothalamus, while specific ions (e.g., *m/z* 615.1, 806.6) were unique to regions like the ventricular system.

Proteome Horizontal RB Section Cluster Comparison

To have a look at the proteome specificity of the RB horizontal section clusters, spatial proteomic analysis was also performed on the 7 clusters observed from the rat brain 11-cluster segmentation image (excluding the cerebellum cluster, already analyzed) (**Figure 37A**). Proteins from red cluster were extracted from hypothalamus region, while proteins from purple and pink clusters were extracted from cerebral cortex. The green cluster was extracted from colliculus area, brown cluster from ventricular system and yellow and light green clusters from corpus callosum. Experiments were performed in biological triplicate. Data were processed with ML, WM and GL previous data in DIA-NN software for protein identification, quantification and correlation. By regrouping the triplicates for each cluster, more than 17243 proteins were identified, among them 5498 were proteotypics (**Figure 37B**). Common proteins were subjected to an ANOVA test (p-value < 0.0001) and showed that 4481 out of 7223 proteins had a significant variability in expression. This was represented by the presence of different clusters of over- or under-expressed proteins between

each extracted region (**Figure 37C**). The resulting heatmap highlighted different clusters of overexpressed proteins. First, cluster i consisted of proteins overexpressed in the cerebellum regions (ML, GL, and WM), while cluster v consisted of proteins overexpressed in the other regions. Specific overexpressed protein clusters were also highlighted for the ventricular system in cluster ii and for the corpus callosum in cluster iii. It was also observed that cluster iii was involved in WM, confirming their correlation in the previous lipid Pearson's analysis (**Figure 35C**). The overexpressed protein cluster iv was involved in the cerebral cortex and hypothalamus brain regions, explaining their similar image segmentation in the red cluster (**Figure 37A**), Pearson's correlation and prediction model using lipid (-) data (**Figure 35C**).



Cluster	Total proteins	Exclusive proteins
WM	6133	0
Light Green	6667	0
Yellow	6524	1
GL	6463	29
Pink	6918	0
Green	7045	3
ML	6457	1
Red	6606	4
Brown	6983	5
Purple	6946	0

Figure 37: Spatial proteomic analysis of rat brain horizontal clusters. A) 10 different clusters identified thanks to lipid (-) lipid MSI and spatial proteomic extraction points. B) Protein Venn diagram. C) Heatmap after ANOVA (p-value 0.0001) analysis demonstrated the presence of different of overexpressed proteins.

Biological pathway analysis of these later over-expressed protein clusters also confirms this observation. Indeed, biological pathways involved in cerebellum (cluster i) were mainly concentered around synapse metabolism, with myelination, paranodal metabolism, neurofilament assembly, and calcium/sodium (**Appendix B, Figure 66A**) transport. This biological process also resumed the one's independently found for ML, GL and WM (**Figure 34**). At the opposite, biological process involved in the cerebral cortex (cluster v) contributed to at least NMDA selective glutamate receptor signaling, regulation of neurotransmitter receptor transport (endosome to postsynaptic membrane) (**Appendix**

B, Figure 66B). This distinction of biological process well defined and distinguished the cerebellum and cerebral cortex regions of the brain. Indeed, the cerebellum is primarily involved in coordinating motor movements, maintaining posture and balance, and motor learning (Harvey, 1980), whereas the cerebral cortex is responsible for higher cognitive functions including perception, memory, attention, language, and consciousness (Javed et al., 2023).

Same conclusions were observable analyzing biological pathways specifically involved in cerebral cortex and hypothalamus, in cluster iv, where mains pathways regrouped vocal and auditory learning, memory and feeling process with serotonin metabolism (**Appendix B, Figure 66D**). Likewise, myelination and neurofilament pathways were involved in cluster iii, for corpus callosum RB area, which was linked to WM biological pathways (**Appendix B, Figure 66E**). The biological pathways for cluster ii, specific of ventricular system RB region, was also analyzed. It turned out that cholesterol, triglyceride, and blood coagulation regulation were the most relevant pathways (**Appendix B, Figure 66C**). These results fit with the neuroanatomy of ventricular system, where cerebrospinal fluid flows in the regions thanks to blood pulsations in surrounding blood vessels (Firdaus, 2020). Furthermore, triglycerides cross the blood-brain barrier and are found in cerebrospinal fluid helping in satiety and cognition mechanisms (Banks et al., 2018).

In this way, it was demonstrated from a protein pathway point of view that cerebellum regions are distinct from the cerebral cortex regions, which itself consists of several specific areas. Their proteomes were also integrated into the model with their paired lipid clusters. In addition, proteomic data of this study were in line with previous analysis already performed on RB regions from published studies. This allowed to add more information to the RB dry proteomics model. First, proteins identified here in bottom-up were compared with proteins identified by top-down in the hippocampus and corpus callosum RB areas, presented in a previous study (Delcourt et al., 2018). According to Delcourt et al., 2018, 16 over 22 proteins identified in top-down for the corpus callosum were also identified and over-expressed in this area according to presented protein dataset. Same observations for 15 proteins over the 20 identified in top-down for the hippocampus.

Workflow Robustness

The robustness of the dry proteomics workflow was thoroughly assessed by examining the redundancy of spectral lipidome and proteome identifications within each cluster across independent triplicates. To ascertain clustering repeatability, the spectral lipid (-) dataset from each cluster was compared among triplicates, as illustrated. Impressively, an average of 99% of common lipid (-) ions was consistently identified across replicates within clusters. Similarly, an in-depth analysis of the spatial proteomic dataset, with a specific focus on distinct clusters, revealed a remarkable consistency, with 93% of the proteins consistently identified across each replicate extraction point within a cluster.

This robustness succinctly summarized the percentage of common protein identifications in replicates for each cluster. Notably, it's worth mentioning that proteins involved in cluster-specific pathways, as previously depicted in **Figure 37** were fully recovered at a 100% rate in subsequent analyses. This underscores the reliability and reproducibility of the methodology employed in capturing proteomic signatures associated with distinct cellular clusters. This reproducibility is the essence of dry proteomics. For future analyses, there's no need to redo spatially resolved proteomics. Simply start with a lipid image and query the dry proteomics model to reliably determine the cluster type, associated proteins, and relevant biological pathways.

Glioblastoma Tumoral Heterogeneity Analysis

Lipid and Peptide MSI Segmentation Correlation

Finally, the dry proteomics workflow was performed on a prospective and retrospective cohort of glioblastoma (GBM), re-using collected data from Duhamel et al., 2022; Zirem et al., 2024 study. The previous study performed patient's stratification based on spatial proteomic and spatial lipidomic guided by MALDI MSI associated to patient survival (Duhamel et al., 2022; Zirem et al., 2024). The cohort consisted of 50 GBM patient tissues, referenced to P1 to P53. Peptide MALDI MSI was performed for all samples, and lipid MSI was conducted for 13 of these tissues. Thus, peptide and paired lipid images were collected for these 13 patients and were processed through developed data imaging workflow. Initially, each tissue was analyzed individually to assess its heterogeneity using Silhouette criterion and generate segmented images. Subsequently, peptide and lipid images were created with 8 to 13 clusters each. The findings of this study revealed an intriguing correlation between lipid and peptide distributions in samples labeled P1 to P14, as evidenced by the generation of highly similar numbers of clusters in both types of images. This correlation underscores the inherent link between the spatial heterogeneity of peptides and lipids within the tissue microenvironment (refer to **Figure 38**). Furthermore, segmentation analysis effectively mirrored histological annotations, enabling the delineation of distinct regions of tumoral proliferation from necrotic or inflammatory areas (as depicted in **Figure 38**), as it was also evocated by Duhamel, M. *et al*., 2022 in previous studies. Prior investigations have primarily relied on lipid and protein to differentiate between these three main tissue types based on specific molecular signatures. In contrast, dry proteomics segmentation workflow offers a more detailed representation of the intricate composition of biological tissues. This enhanced segmentation, not only facilitates the precise identification of pathological features but also reveals previously undetected levels of heterogeneity within tumor, necrotic, and inflammatory regions. This not only achieved improved delineation between annotated areas but also unveiled a greater-than-expected level of heterogeneity within these regions. This heightened resolu-

tion enhances understanding of tissue composition and offers valuable insights into the underlying biological processes driving tumor progression and response to treatment (**Figure 38**).

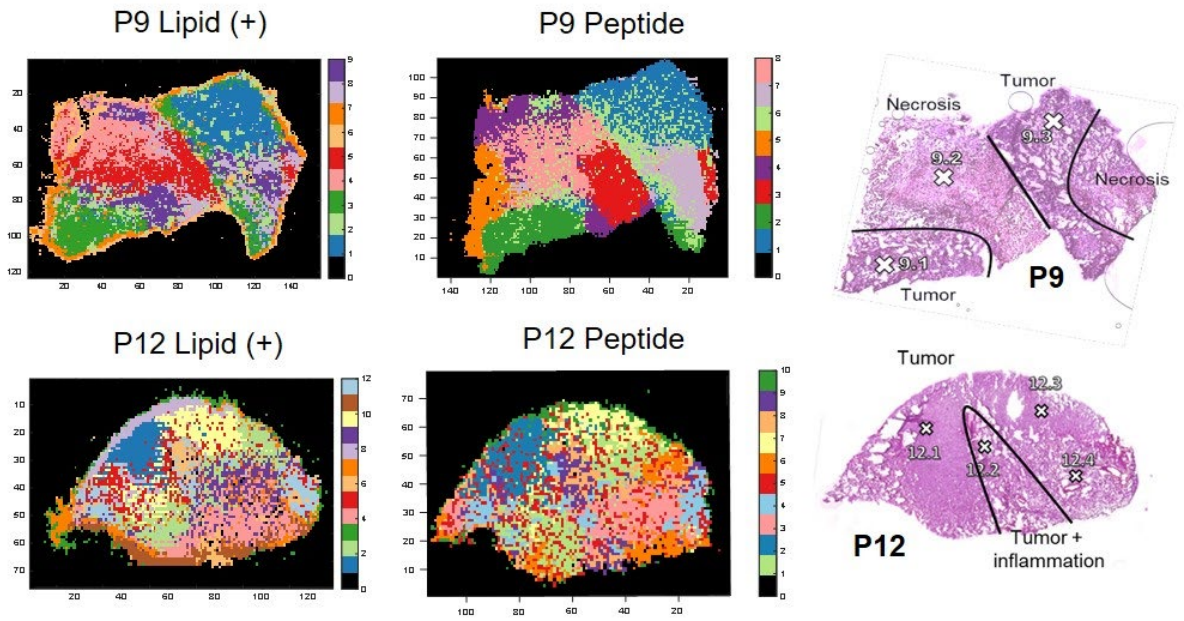


Figure 38: Glioblastoma patient lipid and protein heterogeneity analysis. MSI segmentation examples of patients P9 and P12 lipid and peptide MSI with histopathological annotations.

Lipid-MSI Clusters Classification and Proteomic Correlation

To have a large view on the general heterogeneity on the whole cohort, a co-segmentation was performed on 9 lipid images dataset. It turned out that 13 different clusters were shared between these 9 patients' tissues (**Figure 39A**). Some clusters were correlated to biological specific tissues regions according to histopathological annotations. In this way, clusters 4 (light pink) and 9 (dark purple) were identified as necrosis tissues, clusters 1 (blue), 2 (light green) and 7 (orange) seemed to be specific tumors, whereas clusters 3 (green) and 5 (red) were tumoral areas near to inflammation and clusters 6 (light orange), 8 (light purple), 10 (yellow), 12 (light blue) and 13 (pink) were tumoral areas with necrosis. Clusters were predominantly identified within specific tissues, such as cluster 9 primarily present in P9, or shared across multiple tissues, as observed with cluster 3 in P1, P2, and P13. Once more, the segmentation underscored the molecular diversity within necrotic and tumoral regions, revealing a mosaic of numerous clusters.

A t-SNE representation of tissue lipid imaging clusters allowed to distinguish two mains groups of clusters based on lipid MSI (**Figure 39B**): group A was regrouping clusters 6, 8, 9, 10, 12 and 13, while group B regrouped clusters 1, 2, 3, 4, 5, and 7. Cluster 11 was shared between the two groups. A correlation heatmap, presented in **Figure 40A**, also highlighted the correlation between lipid clusters regrouped in group A and B.

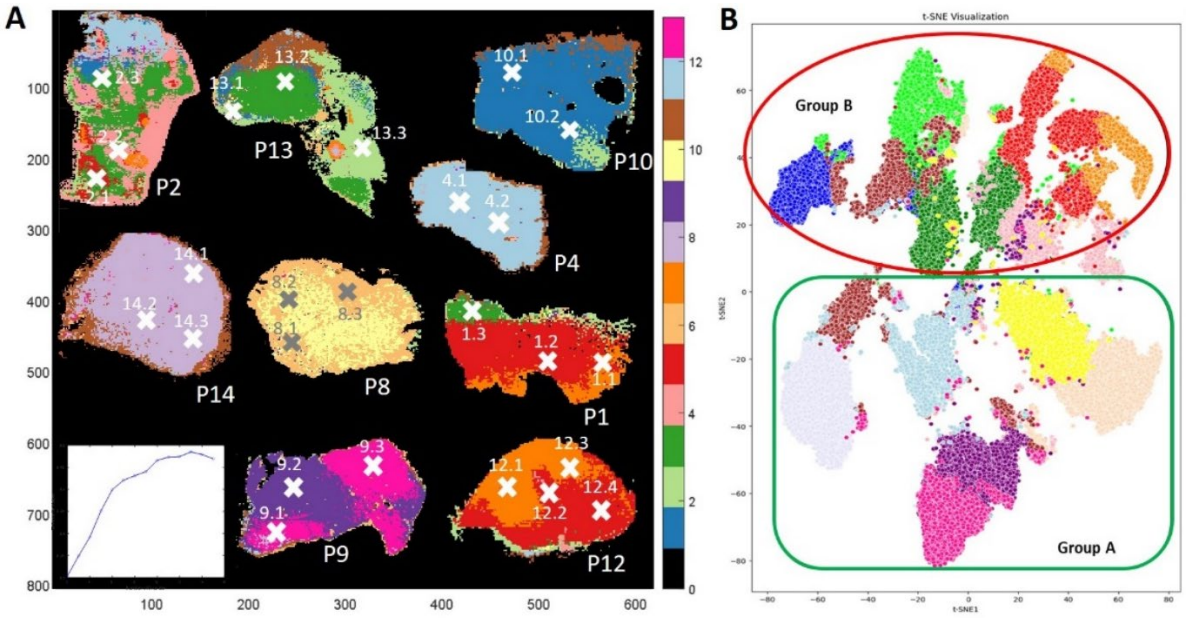


Figure 39: Glioblastoma patient lipid MSI heterogeneity analysis. A) Co-segmentation of 9 tissues previously analyzed by lipid MALDI MSI. B) t-SNE representation of each cluster identified through lipid co-segmentation.

The proteomic data obtained from nine distinct tissue samples were leveraged to conduct a comparative analysis of the various clusters identified through lipid imaging segmentation. Notably, specific extraction points analyzed in this study correlated with clusters identified in lipid imaging (**Figure 40A**). Through statistical analysis, employing an ANOVA test with a significance threshold set at $p < 0.01$, 373 out of 3616 proteins were identified, exhibiting significant variability in expression levels. First, biological pathways were identified for each cluster through ClueGo analysis, based on the overexpressed proteins present in each. Interestingly, some pathways were specific to particular lipid clusters. For example, the RAC3 GTPase cycle pathway was unique to cluster 7 (**Appendix B, Figure 67F**), playing an important role in neuronal development and tumor progression (de Curtis, 2019). L1CAM expression was particularly found in cluster 1 (**Appendix B, Figure 67A**), underscoring the tumor aggressiveness of this cluster. This pathway is a focal point of active investigation in GBM due to its profound implications for tumor aggressiveness, invasion, therapeutic resistance, and poor prognosis. Similarly, overexpressed proteins in cluster 5 were specifically involved in the axon guidance pathway (Chédotal et al., 2005), which is currently a therapeutic area of research for the treatment of malignancy. On the other hand, some biological pathways were common across multiple clusters. Notably, the interleukin-12 family signaling pathway (Cirella et al., 2022), a current therapeutic target in cancer immunotherapy, was identified in clusters 6, 9, 10, and 13 (**Appendix B, Figure 67E, H and J**). Similarly, the ECM proteoglycans pathway (Schönthal et al., 2023) associated with tumor development in GBM was found in clusters 4 and 9 (**Appendix B, Figure 67C and H**). Finally, the biological pathway analysis of each cluster revealed distinct characteristics: some clusters exhibited a

more aggressive GBM pattern, whereas others showed a less aggressive pattern and identified potential therapeutic targets.

Further investigation on protein data allowed to compare the proteome of each cluster and identify correlations between them. It revealed the presence of 2 distinct clusters of over-expressed proteins, namely protein cluster A and B (**Figure 40A**)**.** Of particular interest, protein cluster A was found to correspond to regions of necrotic tissue, encompassing the imaging clusters 9 and 4 previously described. To gain deeper insights into the biological processes associated with these necrotic regions, ClueGO analysis was performed on protein cluster A, utilizing GOterms and Reactome databases (**Figure 40B**). This analysis unveiled a multitude of signaling pathways implicated in necrosis processes. Notably, pathways such as platelet degranulation, blood coagulation, MyD88 deficiency, and IRE1 chaperone activation emerged as significant contributors in modulating cell death processes, including necrosis and can influence tissue damage and disease progression in various pathological conditions such as GBM. In the same way, an intriguing correlation in protein cluster A was observed among protein extracted from lipid imaging clusters 6, 8, 10, 12, and 13, as depicted in **Figure 40A**. The later result confirmed the lipid image cluster classification in group A, proposed previously according to lipid MSI co-segmentation analysis (**Figure 39B**). This cluster notably encompassed tumoral clones characterized by the presence of necrotic regions. Through ClueGO analysis, the significant implication of selenoamino acid metabolism within this cluster was unveiled, shedding light on its pivotal role in the pathogenesis of glioblastoma. This pathway was also individually identified previously in **Appendix B, Figure 67E and J** in lipid cluster 6, 10 and 13. Selenoamino acids, such as selenocysteine and selenomethionine, are fundamental constituents of selenoproteins, where selenium, an essential trace element, is incorporated. These selenoproteins orchestrate a myriad of cellular processes, including antioxidant defense, redox regulation, and DNA synthesis and repair. The dysregulation of selenoamino acid metabolism has been implicated in the intricate progression of GBM through various mechanisms, contributing to disease aggressiveness and resistance to therapy. Similarly, the over-expressed proteins identified within protein cluster B, primarily comprising lipid imaging clusters 3, 4, 5, and 7, yielded significant insights, particularly regarding the involvement of L1CAM interactions (**Figure 40B**) from cluster 1 (**Appendix B, Figure 67A**). Protein cluster B suggested a more aggressive tumor phenotype compared to those within protein cluster a, with implications for poor prognosis or short survival prediction. The intricate interplay between selenoamino acid metabolism and L1CAM interactions underscored the multifaceted nature of GBM pathogenesis, highlighting potential avenues for targeted therapeutic interventions and personalized treatment strategies aimed at mitigating tumor progression and improving patient outcomes.

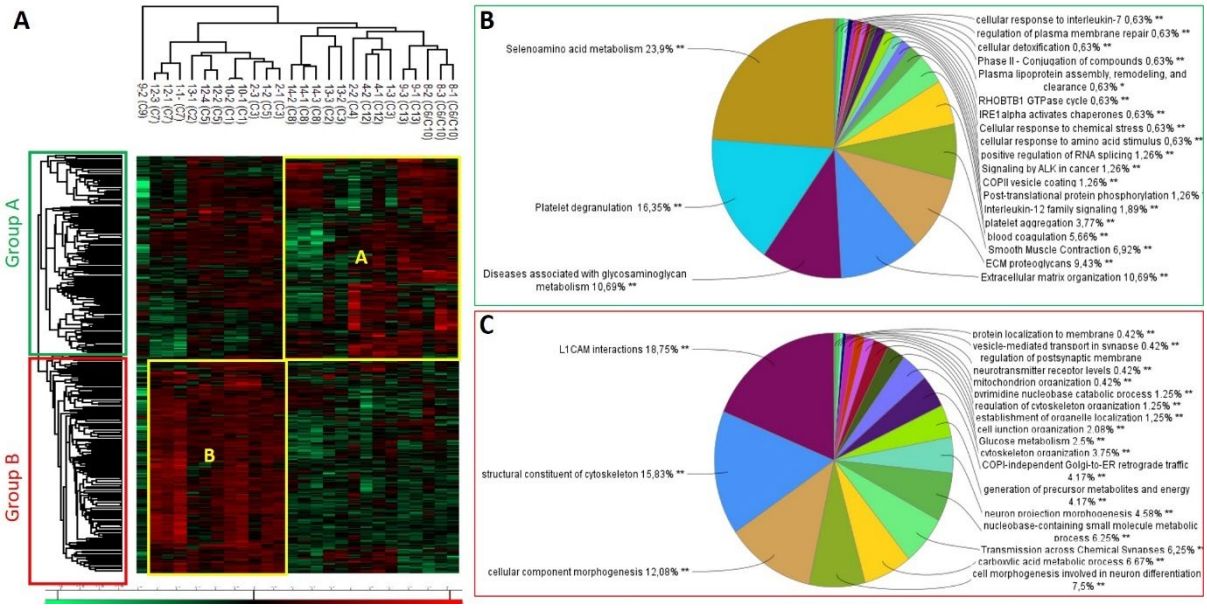


Figure 40: Glioblastoma patient protein heterogeneity analysis. A) Protein heat map after ANOVA (p-value 0.01) analysis demonstrating the presence of different of over-expressed proteins according to lipid clusters. B) Group A and C) group B over-expressed protein clusters ClueGO biological pathways analysis.

Finally, two distinct classification groups, labeled group A and group B, were highlighted and cross-validated between lipid MSI and proteomic analysis. Proteins from the over-express protein cluster A were associated to lipid cluster A, resulting in group A. Thus, group A was associated to the lipid clusters 6, 8, 9, 10, 12, 13, and protein, involving specific protein pathways with a pivotal role in GBM, such as selenoamino acid metabolism. In another hand, group B regrouped lipid clusters 1, 2, 3, 4, 5, 7, and the over-expressed protein cluster B, which possessed more aggressive protein pathways with the implication L1CAM interactions.

Patient Proteome Blind Prediction Based on Lipid Cluster Classification

To predict patient proteome through group A and B, two distinct classifications models were developed. Firstly, a model was trained on the lipid-MSI data from the 13 clusters comprising groups A and B. The aim of this classification model was to classify patient tissue according to lipid images, and associate their paired protein pathway. The resulting model was built with LGBM algorithm with an accuracy of 97% after 5-fold cross validation with an individual accuracy up to 95% for each cluster. Specific lipid ions involved in the model were extracted and identified in specifics clusters using LIME algorithm. The top lipid biomarkers implicated to classify each cluster with 82.3% of contribution.

Thus, the classification of all 9 patients was carefully reviewed according to the patient group A or B classification model, based on the presence of specific lipid clusters in tumoral tissue (**Figure 41A**). In scenarios where tissue samples exhibited clusters overlapping both group A and B, they were unequivocally classified into group B, prioritizing the presence of markers indicative of unfavorable outcomes. This approach ensured a rigorous and systematic evaluation, wherein each case was subjected to thorough examination, with particular emphasis placed on identifying and prioritizing

markers associated with poorer prognostic indicators. By adhering to the following patient classification method, the prognostic assessment process maintained an exemplary level of precision and consistency, empowering clinicians to render well-informed decisions regarding patient management and treatment strategies. As depicted in **Figure 41A**, 4 patients were classified in group B and 5 patients in group A. It's noteworthy that previous investigations have emphasized the importance of assessing patient classification based on the expression levels of key proteins (Duhamel et al., 2022). In 9 patient's cohort (outlined in **Figure 41A**), prior studies classified 2 patients in group B and 7 patients in group A using this protein panel (Duhamel et al., 2022). However, resulting analysis unveiled a nuanced disparity in group classification for patients P10 and P12. This discrepancy can be attributed to the incorporation of molecular heterogeneity into analysis, offering additional insights into survival prediction. Furthermore, upon scrutinizing the co-segmentation analysis illustrated in **Figure 39A**, it became evident that P12 and P1 shared significant cluster composition. Given P1's association with group B, it was reasonable to surmise that P12, sharing similar cluster characteristics, would also be classified within group B. This observation underscores the importance of integrating molecular heterogeneity and comprehensive data analysis techniques to refine classification assessments and enhance clinical decision-making processes.

The 4 last patient tissues, for which lipid-MSI and protein data were available (P3, P5, P6 and P11), were blindly interrogated, pixel per pixel, in classification model based on lipid-MSI clusters. P3 and P11 presented the IDH1 mutation and were not considered in the studies of (Duhamel et al., 2022; Zirem et al., 2024). Upon blind interrogation of the lipid cluster images, patients 3 and 6 harbored a non-negligible percentage of lipid clusters 4 and 2, leading to the prediction to proteins associated to wound healing, or ECM proteoglycans biological pathways for example (**Figure 41B**). The presence of the latter lipid cluster and biological pathways in patient 3 and 6 were thus indicative of the group B classification. Conversely, patients 5 and 11 mainly predicted with high percentage of lipid cluster 6 and 8, allowed the prediction of proteins associated to biological pathways such as Interleukin-12 family signaling, peptide chain elongation or RHO GTPase active ROCKS (**Figure 41B**). In this way, patient 5 and 11 were classified in group A. This result also correlated with a lipid MSI co-segmentation performed on the 13 tissues (**Figure 41C**). The resulting image was composed of 14 clusters according to Silhouette criterion. Interestingly, P6 and P3 were segmented apart of the rest of the cohort, suggesting a possible new lipid class. P5 and P11 tissue associated to group A were sharing specific clusters with P8 and P14, already previously classified in group A (**Figure 41A**).

A

Patient	Classification group according heterogeneity	Classification group according original study
P1	Group B	Group B
P2	Group B	Group B
P4	Group A	Group A
P8	Group A	Group A
P9	Group A	Group A
P10	Group B	Group A
P12	Group B	Group A
P13	Group A	Group A
P14	Group A	Group A

B

Image lipid clusters		Cluster prediction according lipid (-) model and associated proteins pathways			
		Patient 3	Patient 5	Patient 6	Patient 11
C1 Blue	B	/	/	/	/
C2 Light green	B	/	/	**22%**	3%
C3 Green	B	/	/	/	4%
C4 Light pink	B	**39%**	5.4%	**25%**	1.3%
C5 Red	B	/	/	/	2%
C6 Light orange	A	34%	**75%**	50%	/
C7 Orange	B	/	/	/	/
C8 Light purple	A	0.1%	0.2%	/	**51%**
C9 Purple	A	/	/	/	0.3%
C10 Yellow	A	/	/	/	/
C11 Brown	NA	6.7%	19%	3%	36%
C12 Light blue	A	/	0.3%	/	1.5%
C13 Pink	A	21%	/	/	1.1%
Main biological associated pathways		-Wound healing -Regulation of protein-containing complex assembly -ECM proteoglycans -Selenoamino acid metabolism	-Serine family amino acid metabolic -Interleukin-12 family signaling	-Establishment of spindle localization -Wound healing -ECM proteoglycans -Interleukin-12 family signaling	-Peptide chain elongation -RHO GTPase active ROCKS - ATP dependent protein folding chaperone
Patient classification		B	A	B	A

C

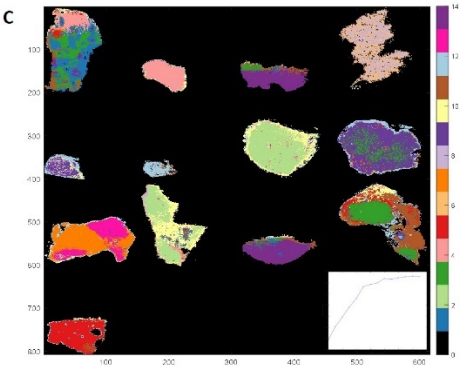


Figure 41: Patient classification and co-segmentation analysis using lipid model in MALDI MSI. A)Patient classification group A and B prediction according to lipid and protein model. B) Lipid cluster and associated protein blind prediction on patient P3, P5, P6 and P11. C) Co-segmentation of 13 tissues previously analyzed by lipid MALDI MSI.

Complementary, a second classification model was constructed using RidgeClassifier with group A and B protein data. The objective was first to intricately cross-validate the lipid MSI-based classification model. The resulting model had an accuracy of 96%, with a 5-fold cross validation. Specific proteins involved in the model decision-making were identified in specifics clusters. Among them, group A and group B biomarkers were distinct, referring to selenoamino acid metabolism or L1CAM interaction pathway for instance. This sophisticated approach underscored the synergy between lipidomic and proteomic analyses in refining group A and B classification for glioblastoma patients, thus paving the way for personalized therapeutic interventions tailored to individual risk profiles.

Thus, previous finding was further reinforced by the protein classification model, which concurred in its classification assessment, designating patients P3 and P6 to group B, whereas P11 and P5 were classified to group A. Hence, both the lipid-MSI clusters and protein models converged in classifying these patients within group A, or B. This alignment serves to authenticate the reliability and validity of the classification model, as well as enhancing the dry proteomics concept on clinical study as GBM.

Groups Classification and Patient Outcome Correlation

The dry proteomics developed pipeline, in both using lipid-MSI data and proteins classification models, led to the discernment of two distinct classes in GBM study, labeled as classification group A and B, illustrated in **Figure 40A**. Leveraging patient survival data, prognostic outcomes were correlated with specific lipid-MSI clusters. The clinical characteristics of the patient, evocated in studies (Duhamel et al., 2022; Zirem et al., 2024), revealed that patients involved in group A, through

lipid-MSI clusters 6, 8, 9, 10, 12, and 13 implication, were upper the survival interquartile range with a survival outcome surpassing 32 months. In the same logic, patients associated to group B, with the presence of lipid clusters 1, 2, 3, 4, 5, and 7, had a poorer survival prognosis of less than 30 months.

Indeed, some of lipid biomarkers involved in lipid-MSI classification model were already recognized as prognostic markers in previous research (Zirem et al., 2024). For instance, lipid ions with *m/z* of 864.7, 866.7, and 881.7 were identified in both studies as markers for survival outcomes exceeding 36 months, primarily present in clusters 8 and 9 from group A. Conversely, lipid ions such as *m/z* 760.6, 788.6, and 810.6 were associated with shorter survival durations, less than 30 months, and were distinctly present in clusters 2 and 5 from group B. Moreover, these significant findings were consistent with prior investigations, reinforcing the notion that protein group B typically correlates with a poorer prognosis compared to group A. Particularly notable was the identification of over-expressed proteins ANXA6 and GPHN within group B, both previously implicated as unfavorable prognostic indicators (Duhamel et al., 2022). Conversely, group A exhibited elevated expressions of proteins RPS14 and MTDH, associated with more favorable prognostic outcomes (Duhamel et al., 2022). Thus, the identification of group A and B lipid features by MALDI MSI, would automatically provide the paired protein pathways, associated to short or long survival patient outcome.

As the left 37 patients were only analyzed through peptide MALDI MSI and spatial proteomics due to the data reuse, the later were interrogated through classification model with proteomics data, to predict their appurtenance to group A and group B, and thus their protein networks and prognosis. Finally, among the cohort of 50 patients, 11 patients were classified in group A with a prognosis survival outcome >32 months, whereas 39 patients were classified in group B with a survival outcome <30 months. The latter results correlated with the clinical characteristics of the patient evocated in study (Duhamel et al., 2022). Indeed, 4 patients with IDH mutation were excluded, 12 patients were upper the survival interquartile range (IQR) set at 13.5 and 32 months, 23 patients were intermediate IQR, and 11 patients were lower IQR.

Dry Proteomics Limitations

Although the dry proteomics model is robust, fast, and simplifies the analysis of complex heterogeneous tissues, it has some experimental and predictive limitations.

Technically, it is impossible to obtain identical consecutive tissue sections due to the z-dimensional factor related to tissue depth during cryostat sectioning. For example, less structural changes between consecutive sections of the cerebellum were observed. However, in horizontal sections, where the anatomy is more complex and variable, differences between consecutive sections are noticeable. These differences affect the imaging of lipids, proteins and peptides due to anatomical changes with depth. To address this issue, spatially resolved proteomics was performed on the

same section used for lipid imaging. Once the model is trained, dry proteomics becomes a useful tool because only one lipid image is needed to assess the heterogeneity, identify the clusters, and associate the proteome, avoiding issues related to anatomical changes in consecutive sections. The second limitation concerns the predictive ability of the model, which is based on experimental data of clusters obtained by segmentation of lipid images. A reliable and accurate model requires a large cohort with representative replicates of the studied population. Building a generalizable model is challenging because some tissue-specific clusters may not be represented in our analyses. When the model encounters an unknown cluster that it hasn't been trained on, it will likely misclassify it by approximating a known cluster. There are two ways to address this problem. First, by checking the approximation of an unknown cluster by the unsupervised *k*-means++ and t-SNE models. This involves plotting the matrix of this cluster on the *k*-means++ and t-SNE axes to see which known cluster it is close to, thereby confirming or disproving the model's predicted approximation. Second, consider the use of self-training algorithms in the future (Hu & Laskin, 2022). This involves retraining our model with known labeled clusters and new unknown and unlabeled clusters to improve and update the model specifically for clinical routine use. In this case, it will also be necessary to update the proteomic data for the new unknown clusters.

To extrapolate the strategy of dry proteomic to other tissue types or diseases, different learning model approaches are possible. The first one consists in a specific model for a specific tissue type or disease. In this case, the model would be trained on clusters specific to a particular tissue type or disease. While this approach is limited to the heterogeneity of that single tissue or disease, it offers greater accuracy by focusing on fewer clusters, which reduces the risk of false positive predictions (fewer classes in a multi-class classification task). This results in a more targeted and precise model. The second possibility is to improve the model in an agnostic model. This is a global model designed to work across multiple tissue types or diseases. To improve its performance, the model would need to be trained on clusters from various diseases and tissue types. Such a model would be capable of predicting and identifying clusters specific to particular tissues or diseases, while also recognizing common clusters across different tissue types. This approach could be especially useful for large-scale studies, such as PAN-cancer research. However, agnostic models are typically less accurate and require sophisticated feature engineering to enhance their performance. Another strategy to improve agnostic models is to use a transfer learning approach, where specific models are trained on individual diseases and then adapted for broader applications. Once refined, this type of agnostic model could also be applied to study metastasis and help trace the origin of cancers.

Conclusion and Perspectives

An automated dry proteomics method utilizing lipid MALDI MSI was designed to tackle various challenges in correlating cluster-specific lipids and proteins for imaging and pathway analyses. This method's segmentation pipeline was enhanced with SVD data compression, *k*-means++ segmentation, and the Silhouette criterion, allowing for effective multi-omics MALDI MSI data correlation. The Silhouette criterion proved vital in detecting tissue heterogeneity and determining the optimal number of clusters automatically and unsupervised.

The RB cerebellum tissue model was used to validate the workflow's capability for imaging lipids, proteins, and peptides, demonstrating superior performance compared to other segmentation algorithms. The robustness of the MS image processing model was confirmed through multiple experimental replications. Multi-omics segmented images identified RB cerebellum clusters ML, GL, and WM, each with unique spatial distributions, lipid and protein compositions, and biological pathways. A predictive model was developed based on these lipid fingerprints and complemented by specific protein compositions and pathways for each cluster, validating the dry proteomics approach for GL, ML, and WM.

When the predictive model was applied to more complex tissues beyond cerebellar regions, such as RB horizontal slices, multi-omics MALDI MSI analysis successfully identified clusters with unique spatial localizations, including cerebellum clusters. The model, trained with cerebellar lipid datasets, effectively annotated these areas and provided insights into their specific lipids, proteins, and pathways. Further refinements improved the model's accuracy in predicting complex tissue compositions, highlighting the potential of dry proteomics in revealing intricate biological processes.

In glioblastoma patient cohorts, the dry proteomics workflow provided significant insights into the spatial heterogeneity of peptides and lipids within the tumor microenvironment. By integrating previous research data with advanced imaging techniques, previously unknown complexities in GBM tissues were revealed. The segmentation accurately delineated pathological features and highlighted subtle variations within tumor, necrotic, and inflammatory regions, offering a comprehensive tissue composition profile.

The relationship between lipid and peptide distributions suggests their potential as reliable tumor biomarkers. Distinct molecular signatures were identified in various tumor regions, reflecting diverse biological processes and cellular compositions. Co-segmentation revealed 13 unique clusters among patients, each corresponding to specific biological tissue regions. The integration of proteomic data deepened the understanding of the molecular landscape within these clusters. Statistical analysis showed significant protein expression variability across different clusters,

identifying distinct biological pathways. ClueGO analysis emphasized the role of signaling pathways, such as selenoamino acid metabolism and L1CAM interactions, in GBM pathogenesis.

By integrating the dry proteomics method with prognosis, a sophisticated classification model for GBM patients was developed. This model identified cluster types and corresponding proteomic data with region-specific pathways and functions, categorizing patients into distinct prognostic groups. Two patient groups, A and B, were predicted using a model based on GBM lipid-MSI that captured molecular heterogeneity within tumor tissues. This model was also adapted into a proteomic model distinguishing groups A and B based on protein data alone, validating the lipid MSI data-based model and enabling patient classification with only proteomic data. Group A's protein networks were then associated with survival outcomes over 32 months, while Group B's networks were linked to outcomes under 30 months. This classification provided insights into tumor protein networks related to survival prognosis. Overall, this project highlights the importance of multi-omics approaches for comprehensive prognostic assessments in GBM. By exploring the connections between molecular characteristics and clinical outcomes, the developed model offers valuable insights for personalized treatment strategies and improved patient management in the complex GBM landscape.

Finally, the dry proteomics approach, which first identifies tissue heterogeneity and distinct clusters through lipid imaging and then automatically associates specific proteins and biological pathways with each cluster, is crucial for clinical applications. These insights help identify potential therapeutic targets or prognostic markers, as demonstrated in the glioblastoma study, paving the way for better patient outcomes and personalized treatment strategies (**Figure 42**).

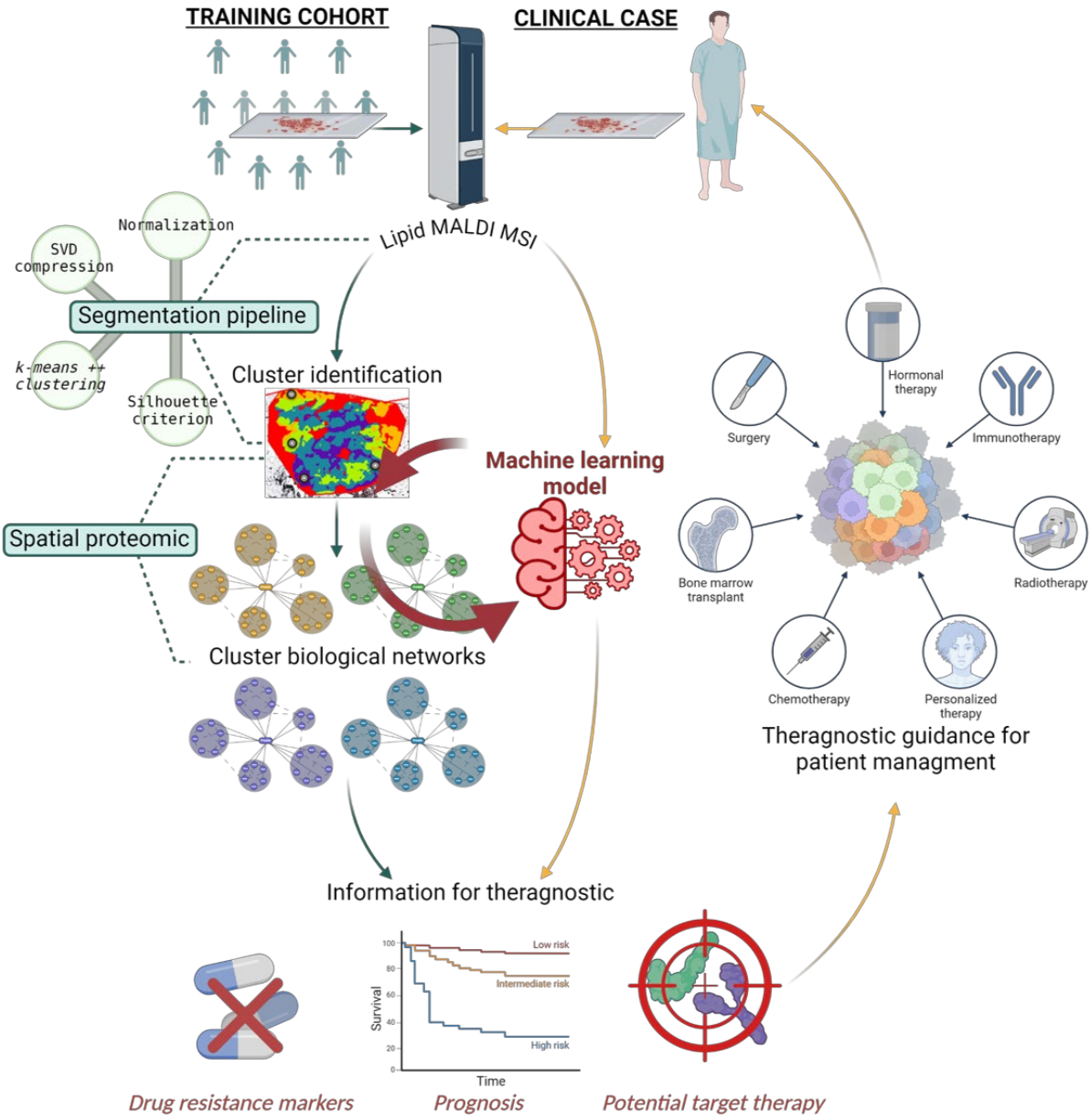


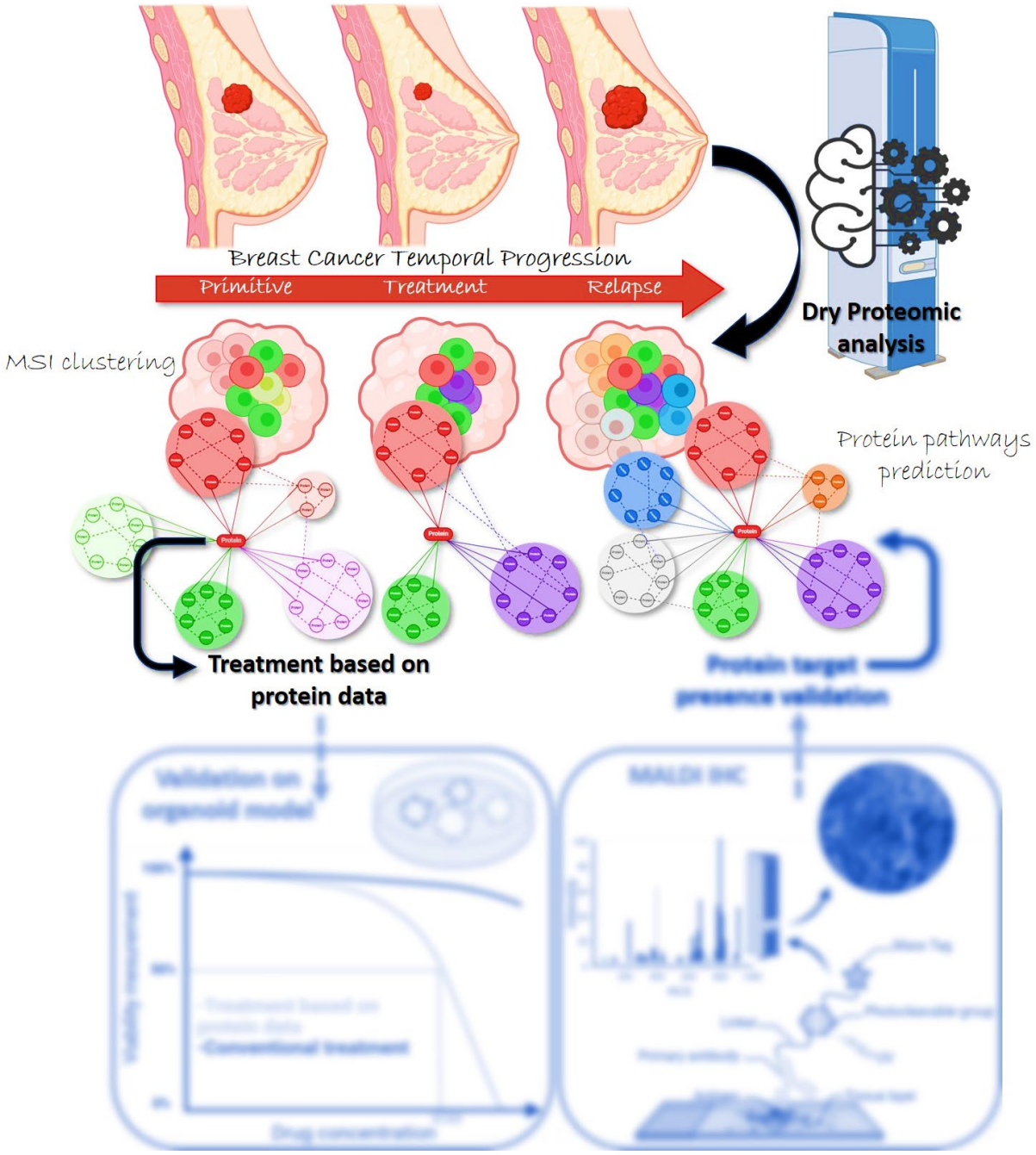
Figure 42: Dry Proteomic Workflow for Tumor Characterization. This workflow illustrates the use of MALDI MSI and machine learning for personalized cancer treatment. Tissue samples are processed through MALDI MSI, and lipid data is fed into a segmentation pipeline utilizing clustering methods (e.g., SVD, *k*-means) to identify spatial proteomic patterns. These patterns highlight intra-tumor heterogeneity. A machine learning model then integrates cluster associated proteomic data to highlight with drug resistance markers, predict patient prognosis, and suggest potential targeted therapies based on the tumor's molecular profile, aiding in personalized treatment strategies.

This workflow was then applied to breast cancer study with the aim to comprehensively understand the spatial and temporal heterogeneity of the pathology within clinical settings. This approach took into account the clonal evolution of the tumor over time, the diversity of tumor subtypes, and the impact of various treatment modalities, including chemotherapy, hormone therapy, radiotherapy, and immunotherapy, both in combination and as standalone treatments. By integrating these factors, the dry proteomics workflow was used to identify actionable molecular

targets and potential therapeutic strategies tailored to the specific characteristics of the tumor heterogeneity at different stages.

CHAPTER 4

4D Longitudinal Proteomics Tracking of Breast Cancer Heterogeneity Community Response to Therapeutics



CHAPTER 4: 4D Longitudinal Proteomics Tracking of Breast Cancer Heterogeneity Community Response to Therapeutics

Introduction

Breast cancer continues to be a leading cause of mortality among women globally, with 2.26 million cases and nearly 685,000 deaths reported in 2020, according to Globocan 2020. Despite advancements in treatment, the challenge of relapse persists, with approximately 30% of breast cancer cases experiencing recurrence. This troubling reality highlights the critical need to understand the dynamic progression of breast cancer and its response to therapy over time. Central to this understanding is the influence of heterogeneous tumor clusters, which play a significant role in shaping drug response and treatment outcomes.

Breast cancer's complexity is significantly driven by the presence of heterogeneous tumor clusters, which not only vary within and between patients but also interact dynamically, influencing drug response and disease progression. These clusters form a complex community of genetically related yet functionally distinct subpopulations that constantly interact, compete, and evolve within the tumor microenvironment (Burrell et al., 2013; Yates, 2017). This interaction among clusters profoundly impacts the pathology of breast cancer, as the collective behavior of these subpopulations can dictate the overall trajectory of the disease and its response to therapy. The presence of such heterogeneous clusters introduces a level of unpredictability in treatment outcomes. Each cluster may exhibit unique molecular signatures, resistance mechanisms, and metabolic adaptations, leading to a spectrum of responses to the same therapeutic agent (Marine et al., 2020). Importantly, the interactions between clusters can worsen resistance by allowing them to communicate and strengthen survival mechanisms or create conditions that help resistant groups grow. This means that even if one cluster is successfully treated, other clusters may still survive due to their interactions, leading to treatment failure, cancer recurrence, and further tumor growth (Dagogo-Jack & Shaw, 2017; Pogrebniak & Curtis, 2018). Understanding and addressing the impact of these interacting clusters is crucial for developing effective treatment strategies. It requires a paradigm shift from viewing tumors as homogeneous masses to recognizing them as dynamic ecosystems were subpopulation interactions shape drug response and pathology evolution. By targeting these clusters in a coordinated manner, considering not just their individual characteristics but also their collective behavior and interactions, there is potential to disrupt the adaptive mechanisms that drive resistance and disease progression.

MALDI MSI, has already proven effective in analyzing breast cancer tissues. When integrated with spatial proteomics and wet proteomics techniques, MALDI MSI provides detailed spatial information about the distribution of molecules within tumor subpopulations, allowing for a deeper understanding of the molecular architecture of heterogeneous clusters and identifying potential druggable targets (Hajjaji et al., 2022b; Quanico et al., 2013).

In this study, a longitudinal, prospective, and retrospective analysis of the spatiotemporal evolution of breast cancer heterogeneity in response to therapy is presented, with a focus on luminal, triple-negative, and HER2-low breast cancer subtypes. MALDI MSI enabled spatially resolved, label-free imaging of diverse molecular classes, particularly proteins, within their histological context. The MSI data were processed using an unsupervised segmentation method, allowing for the automatic assessment of tumor heterogeneity and revealing functionally distinct subpopulations. Further analysis was conducted using a spatial proteomics to extract and profile selected subclones in situ via LC-MS/MS. This approach facilitated the identification of reference proteins specific to each clone, deepening our understanding of tumor biology and informing targeted therapeutic strategies. By integrating MSI-specific clusters with their proteomic profiles in a machine learning framework, a database of breast cancer heterogeneous clones was constructed, and integrated in dry proteomics model for BC study. This tool enables the tracking of cancer heterogeneity evolution over time, evaluates drug efficacy, identifies potential druggable protein targets, and correlates treatment responses with recurrence rates in a heterogeneous tumor environment.

This study's 4D longitudinal proteomics approach offers a novel way to explore these interactions over time, revealing how cluster communities evolve and respond to therapeutic pressures. By mapping the spatial distribution and molecular profiles of these clusters using advanced techniques like MALDI MSI and integrating this data with machine learning, critical insights were gained into the complex interplay within the tumor. This comprehensive understanding allows for the identification of new therapeutic targets that can disrupt the harmful interactions between clusters, paving the way for more adaptive and personalized treatment strategies that can overcome the tumor’s evolving resistance mechanisms.

Results

This section focused on exploring key questions about BC tumoral heterogeneity:

- Are heterogeneous tumor clusters cooperating within the breast cancer microenvironment, and if so, how does this cooperation impact disease progression and treatment response?
- How can understanding tumor microenvironment dynamics improve breast cancer treatment strategies?

The strategy employed for studying the proteomic and spatial heterogeneity of tumors involves combining MALDI MSI with spatial proteomics to identify and characterize tumor clones. The study utilized FFPE biopsies from 16 patients with luminal, triple-negative, or HER2-low breast cancer, for whom multiple biopsies were available over time (**Table 9**). Serial tissue sections were prepared for each tissue. One section was designated for HPS staining, which allowed the characterization of tumor regions through detailed histological analysis performed by a pathologist. Another section was subjected to MALDI MSI analysis to generate a peptide map of the tumor, which was then processed dry proteomic clustering pipeline. This processing defines the optimal segmentation of the section by assigning colors based on spectral similarity. Only the tumor regions identified by histological analysis were selected for further segmentation. This segmentation distinguished proteomic sub-populations within the tumor, corresponding to different tumor clones exhibiting intra-tumoral heterogeneity. These clones were individually analyzed by spatial proteomics. Imaging and protein data from each tissues were then compared together in order to observe BC heterogeneity evolution overtime and treatments.

Table 9: Breast cancer patient tissues clinical data.

N° Patient	Sample Overtime	Age At Collection	Sampling	Tissue Type	Histology	Subtype	Setting	Previous Treatments
2	T1	47	Surgery	Breast	Ductal	HR+ Her2 Low	Primary BC	0
	T2	55	Surgery	Breast	Ductal	HR- Her2-	Relapse Loco-regional	Chimio EC-taxol
	T3	55	Surgery	Breast	Ductal	HR- Her2-	Relapse Loco-regional	Rt-capECitabine
3	T1	59	Microbiopsy	Breast	Ductal	HR+ Her2+	Relapse Loco-regional	Chimio 3EC-3TXT+trastuz
	T2	66	Surgery	Node	Ductal	HR+ Her2+	Relapse Loco-regional	Hormono Tamoxifen
4	T1	54	Microbiopsy	Lung	Ductal	HR+ Her2+	Metastatic	Hormono Exemestane
	T2	58	Microbiopsy	Liver	Ductal	HR+ Her2+	Metastatic	Hormono Tamoxifen
5	T1	72	Microbiopsy	Breast	Ductal	HR- Her2 Low	Primary BC	0
		72	Microbiopsy	Node	Ductal	HR- Her2 Low	Primary BC	0
	T2	73	Surgery	Breast	Ductal	HR- Her2-	Post Neoadjuvant	Chimio 3EC-3TXT
6	T2	73	Microbiopsy	Breast	Ductal	HR- Her2-	Primary BC	0
	T2	73	Surgery	Breast	Ductal	HR- Her2-	Post Neoadjuvant	Chimio TXT Cyclophosphamide
	T3	74	Microbiopsy	Breast	Ductal	HR+ Her2-	Metastatic	Taxol - Rt
	T4	74	Microbiopsy	Skin	Ductal	HR- Her2-	Metastatic	Chimio - FEC50
6b	T1	57	Microbiopsy	Breast	Ductal	HR+ Her2+	Metastatic	Hormono Tamoxifen
	T2	57	Surgery	Breast	Ductal	HR+ Her2+	Metastatic	Pertuz Trastuz
7	T1	37	Surgery	Breast	Ductal	HR- Her2-	Post Neoadjuvant	Chimio 3EC-3TXT
	T2	41	Microbiopsy	Node	Ductal	HR- Her2-	Metastatic	0
7b	T1	72	Microbiopsy	Breast	Ductal	HR- Her2+	Relapse Loco-regional	0
	T2	72	Surgery	Breast	Ductal	HR- Her2+	Post Neoadjuvant	Chimio 3EC-3TXT+trastuz
8	T1	62	Surgery	Breast	Ductal	HR- Her2-	Relapse Loco-regional	Hormono
	T2	65	Surgery	Breast	Ductal	HR- Her2-	Relapse Loco-regional	Chimio 3EC-3TXT
10	T1	57	Microbiopsy	Node	Ductal	HR- Her2 Low	Relapse Loco-regional	Hormono Arimidex
	T2	61	Microbiopsy	Node	Ductal	HR- Her2 Low	Relapse Loco-regional	Chimio - FEC50
11	T1	64	Microbiopsy	Breast	Ductal	HR- Her2+	Metastatic	Trastuzumab
	T2	65	Surgery	Breast	Ductal	HR- Her2+	Metastatic	Trastuzumab
	T3	67	Surgery	Breast	Ductal	HR+ Her2 Low	Metastatic	Trastuzumab
14	T1	54	Surgery	Breast	Ductal	HR+ Her2+	Primary BC	0
	T2	59	Microbiopsy	Breast	Ductal	HR+ Her2+	Relapse Loco-regional	Hormono
	T3	59	Surgery	Breast	Ductal	HR+ Her2+	Relapse Loco-regional	Hormono
17	T1	49	Microbiopsy	Breast	Ductal	HR+ Her2-	Primary BC	0
	T2	50	Surgery	Breast	Ductal	HR+ Her2-	Post Neoadjuvant	Chimio 3EC-3TXT
	T3	53	Microbiopsy	Breast	Ductal	HR+ Her2-	Metastatic	Hormono Anastrozole
	T4	53	Microbiopsy	Node	Ductal	HR+ Her2-	Relapse Loco-regional	Hormono Anastrozole
18	T1	40	Microbiopsy	Breast	Ductal	HR- Her2-	Metastatic	Fulvestrant Palbociclib
	T2	40	Surgery	Breast	Ductal	HR- Her2 Low	Metastatic	Fulvestrant Palbociclib
	T3	41	Microbiopsy	Liver	Ductal	HR- Her2 Low	Metastatic	Fulvestrant Palbociclib
	T4	42	Microbiopsy	Liver	Ductal	HR- Her2 Low	Metastatic	Atezo+bdb001
19	T1	41	Surgery	Breast	In Situ	In Situ	Primary BC	0
	T2	43	Microbiopsy	Breast	Ductal	HR+ Her2+	Relapse Loco-regional	Surgery
	T3	44	Surgery	Breast	Ductal	HR+ Her2+	Post Neoadjuvant	Chimio 3EC-3TXT+trastuz
20	T1	49	Surgery	Node	Lobular Pleomorphe	HR+ Her2-	Primary BC	0
		49	Surgery	Breast	Lobular Pleomorphe	HR+ Her2-	Primary BC	0
	T2	51	Microbiopsy	Breast	Lobular Pleomorphe	HR+ Her2 Low	Metastatic	Chimio EC-taxol - Rt - Hormono Letrozole
	T3	52	Surgery	Breast	Lobular Pleomorphe	HR+ Her2 Low	Metastatic	Fulvestrant Palbociclib
		52	Surgery	Breast	Lobular Pleomorphe	HR+ Her2 Low	Metastatic	Fulvestrant Palbociclib
		52	Surgery	Breast	Lobular Pleomorphe	HR+ Her2 Low	Metastatic	Fulvestrant Palbociclib
	T4	53	Microbiopsy	Liver	Lobular Pleomorphe	HR- Her2-	Metastatic	Fulvestrant Palbociclib

Analysis of Breast Cancer Heterogeneity in Individual Tissues

Following the same strategy from **CHAPTER 2**, initial analyses were conducted separately for each tissue sample in the cohort to investigate intra-tumor heterogeneity. These analyses utilized peptide data from BC tissues, acquired through MALDI MSI, which were individually segmented within the tumor regions for each patient using the dry proteomic image clustering pipeline, as described in **CHAPTER 3**. This approach allowed a detailed examination of the molecular architecture of each tumor sample. In this manuscript, we focused on three representative patients from the primary subtypes of breast cancer present in our cohort. Specifically, we analyzed Patient 2, who had the HR+/HER2- or low subtype; Patient 11, who was HER2+; and Patient 18, who had the HR-/HER2- or low subtype (**Figure 43**), providing a deeper look into their tumor heterogeneity.

As demonstrated in **CHAPTER 2**, the intra-tumoral analysis of these samples revealed distinct molecular subpopulations, each represented as clusters in the MALDI MSI segmentation images. These clusters reflect the underlying heterogeneity within the tumors, where each cluster corresponds to unique molecular features, indicating a complex network of tumor biology. The identification of these clusters is crucial because they represent potential variations in tumor behavior, which could influence prognosis and treatment response. To better understand the clone heterogeneity observed in these clusters, the spectral centroid of each cluster was extracted for comparative analysis across the entire patient cohort. This allowed to perform future comparative study of tumor clones, highlighting how certain molecular subtypes or clones might recur or evolve across different patients.

The correlation between segmented images and tissue anatomopathological annotations facilitated the selection of regions of interest for spatial proteomic analysis (represented with circles in **Figure 43**). These areas were selected based on their clinical significance and histopathological features, ensuring that the molecular findings were directly relevant to tumor behavior and patient outcomes. In total, nearly 200 patient clones were extracted from the entire cohort using this technique, providing a rich dataset for further investigation. Protein analysis of each intra-tumor cluster further reinforced the molecular differences identified through imaging. Each cluster displayed a distinct proteomic profile, reflecting different pathways, processes, and biological activities occurring within the tumor.

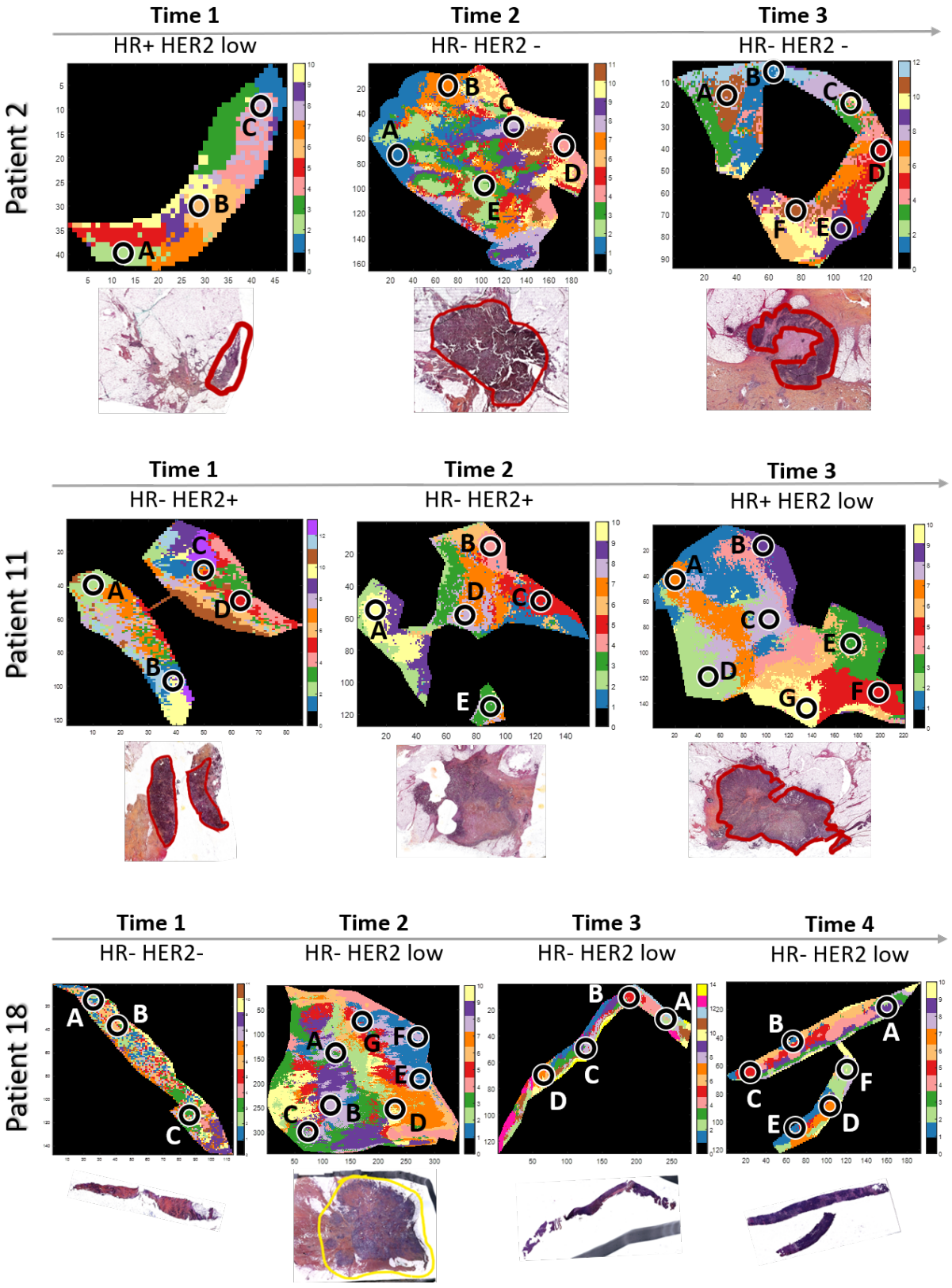


Figure 43:Intra tumor breast cancer segmentation based on peptide MALDI MSI for patients 2, 11 and 18, with HPS coloration and spatial proteomic extraction points annotations (circles).

As highlighted in **CHAPTER 2**, protein analysis could uncover potential protein markers with clinical importance as therapeutic targets, indicators of drug resistance, or factors linked to patient

prognosis. Moreover, a co-segmentation of MSI data and proteomic analysis enabled the identification of specific and shared clusters, or tumor clones, across different tissues from the same patient. This ability to detect both common and distinct molecular signatures within different regions of the tumor offers insights into tumor evolution. Shared clusters might suggest a core tumor biology that remains consistent, while unique clusters could indicate new subpopulations emerging over time, possibly driven by treatment pressure or disease progression.

Based on these results, the central objective of this chapter was to explore how tumor heterogeneity evolves over time, particularly in response to therapy. By comparing the molecular landscape of tumors at different stages of disease and under various treatments, we sought to unravel the dynamic nature of tumor evolution. This approach sheds light on how tumors adapt or respond to therapeutic interventions and how these changes can affect treatment efficacy and outcomes. Understanding the temporal evolution of tumor heterogeneity provides crucial insights into the mechanisms driving tumor progression, treatment resistance, and relapse, offering opportunities for more targeted and effective therapeutic approaches in the future.

Dynamic Analysis of Breast Cancer Heterogeneity in Individual Patients Over Time

To assess and compare tumor heterogeneity over time for each patient, tissues from different time points were co-segmented. For protein analysis, proteomic data from individual tissue samples were merged, allowing for a comprehensive comparison of the entire proteome across all tissues.

Patient 2 Tumoral Heterogeneity Over Time Analysis

Interesting findings were observed for patient 2. Three time point tissues were available for Patient P2: a first tissue P2T1 from the primary BC tumor HR+/HER2+, a second P2T2 from a loco-regional relapse HR-/HER2- treated with EC-Taxol, and a third one P2T3 from a loco-regional relapse HR-/HER2- treated with RT-capecitabine (**Table 9**). **Figure 44A** illustrates the co-segmentation of tissue samples from three different time points (P2T1, P2T2, and P2T3), clustered in 10 clusters according to Silhouette criterion. This co-segmentation highlighted substantial differences between them, with very few shared clusters across the time points. For instance, cluster 8 (purple) was highly abundant in the tissue from the first time point (P2T1), signifying a dominant tumor subpopulation at that stage. However, this cluster nearly disappeared in the second time point (P2T2), where only faint traces were observed, suggesting a significant reduction or transformation of that subpopulation. P2T2 also marked an increase in heterogeneity, shown by a wider variety of clusters. Between time points 2 (P2T2) and 3 (P2T3), the only shared cluster was cluster 6, which appeared in both tissues but in limited amounts in the third time point. This indicates that while some tumor subpopulations persisted, their presence diminished over time. The absence of other common

clusters and the overall reduction in shared subpopulations suggest a radical shift in tumor composition between the three time points. This dynamic change in tumor heterogeneity points to the evolving nature of the tumor environment in patient 2, likely influenced by disease progression or treatment. Indeed, clinical data reveal that P2T1 was characterized as a primary BC that was HR+/HER2 low and untreated. In contrast, P2T2 represented a relapse, with the tumor evolving to an HR-/HER2- status and having been treated with EC-Taxol. By P2T3, the tumor had received treatment with RT-capecitabine, highlighting the impact of treatment strategies on tumor evolution over time. HR+/HER2low tumors are typically less aggressive because they express hormone receptors, which allow for effective treatment options with hormone therapies like tamoxifen or aromatase inhibitors. Conversely, the emergence of the HR-/HER2- subtype significates a shift toward a more aggressive form of breast cancer (TNBC). This subtype lacks both hormone receptors and HER2 expression, which is associated with a faster growth rate and a higher likelihood of metastasis.

These findings were further supported by proteomic data analysis. As shown in the Venn diagram in **Figure 44B**, each tissue from the different time points exhibited a distinct proteome, with several exclusive proteins. FUNRICH analysis of these exclusive proteins enabled investigation of the biological pathways involved (**Figure 44C**). Notably, tissues from P2T2 and P2T3 showed a higher percentage of genes involved in glypican pathways compared to P2T1. Glypicans, a family of heparan sulfate proteoglycans attached to the cell membrane, play crucial roles in cell signaling, growth, and development. In cancer, glypicans interact with growth factors and cytokines, modulating signaling pathways that influence tumor growth, metastasis, and the tumor microenvironment (Grillo et al., 2021). By regulating these pathways, glypicans affect cell proliferation and survival. Previous studies have already shown that higher glypican levels are associated with advanced breast cancer grades and larger tumor sizes (Alshammari et al., 2021; Grillo et al., 2021), making them a promising therapeutic target in breast cancer. This observation aligns with the high percentage of genes associated with mesenchymal to epithelial transition (MET) in P2T1 compared to P2T2 and P2T3. Glypicans are known to modulate MET and epithelial-mesenchymal transition (EMT), both of which are critical processes in tumor growth, invasion, and metastasis (Lambert et al., 2017). By influencing these transitions, glypicans interact with the tumor microenvironment (TME) to facilitate tumor progression and growing (Famta et al., 2024; Lambert et al., 2017). This may help explain the tumor's evolution from HR+/HER2 low in P2T1 to HR-/HER2- in P2T2, indicating increased tumor aggressiveness over time. The proteomic data also revealed an increase in genes associated with the ErbB1 internalization and vascular endothelial growth factor receptor (VEGFR) signaling pathways between time points. ErbB1, also known as EGFR, internalization could explain the loss of hormone receptors on the cell surface, contributing to the progression to an HR- tumor type. Similarly, the VEGFR signaling pathway promotes angiogenesis, supporting tumor growth and metastasis (Ceci et

al., 2020). These results explained the evolution of treatment over-time. As the tumor evolved, the need for tailored therapeutic approaches becomes increasingly apparent. This understanding informed the decision to shift treatment modalities over time, particularly toward the combination of radiotherapy and capecitabine (RT-capecitabine) in the P2T3 time point, an approach employed for locally advanced tumors as part of a neoadjuvant therapy strategy.

Additionally, ANOVA analysis (p-value < 0,01) of common proteins across P2T1, P2T2, and P2T3 identified 1754 significative proteins over 6050, forming distinct clusters of overexpressed proteins in each tissue (**Figure 44D**). Further pathway analysis of these clusters using FUNRICH confirmed previous observations (**Figure 44E**). For instance, the integrin cell surface signaling pathway was associated with interactions between cancer cells and the surrounding stromal cells in the TME, promoting tumor growth and metastasis by activating EMT. This pathway, particularly the involvement of β integrin 1, is closely linked to the EMT process, which drives the tumor toward a more aggressive phenotype (S. Li et al., 2023). Moreover, the presence of the Tumor Necrosis Factor-Related Apoptosis-Inducing Ligand (TRAIL) signaling pathway further highlights the tumor's aggressiveness (Kundu et al., 2022). TRAIL, part of the TNF family, induces apoptosis selectively in cancer cells and can be stimulated by chemotherapies like paclitaxel (Taxol), which was administered to P2T2, EGFR targeted therapies, or immune checkpoint inhibitors. A decrease in TRAIL was observed in P2T2, possibly reflecting this treatment's effect. However, despite being a promising therapeutic target, resistance to TRAIL-induced apoptosis like Taxol (Rahman et al., 2009) is still a possibility to keep in mind.

Potential protein markers associated with taxol resistance were identified in the proteomic analysis of patient 2's dataset. Notably, several proteins previously linked to drug resistance were observed. Specifically, the overexpression of the TUBB3 gene in samples P2T1 and P2T2 is significant, as TUBB3 is a well-known marker of multidrug resistance. High levels of TUBB3 can alter microtubule dynamics, potentially decreasing the binding efficacy of taxol, thereby promoting resistance (Stengel et al., 2009; Tame et al., 2017). In addition, P2T1 exhibited exclusive expression of the CA12 gene, particularly in clones A and C. The presence of CA12 is noteworthy because it has been implicated in drug resistance mechanisms, especially in hypoxic tumor environments, where it contributes to pH regulation (Tonissen & Poulsen, 2021). This is particularly concerning when CA12 coexists with overexpression of P-glycoprotein (PGP), a well-established efflux transporter responsible for pumping drugs like taxol out of cells, reducing intracellular drug concentration (Tonissen & Poulsen, 2021). PGP was found to be overexpressed in both P2T2 and P2T3, further suggesting a strong association with taxol resistance in these samples. In this way, the combination of TUBB3 overexpression with elevated levels of CA12 and PGP in patient 2's samples indicates a robust multidrug resistance profile, highlighting the potential challenge of overcoming taxol resistance in this patient’s tumor.

Thus, the proteomic analysis revealed significant changes in tumor heterogeneity over time in patient 2, with distinct proteome profiles and biological pathways emerging at each time point. The increased involvement of glypican and VEGFR pathways, alongside changes in MET/EMT processes, suggests a progression towards a more aggressive tumor phenotype. These findings, combined with observed shifts in TRAIL signaling and integrin pathways, provide insights into the tumor's evolution and potential therapeutic targets, highlighting the dynamic nature of breast cancer progression and treatment resistance response.

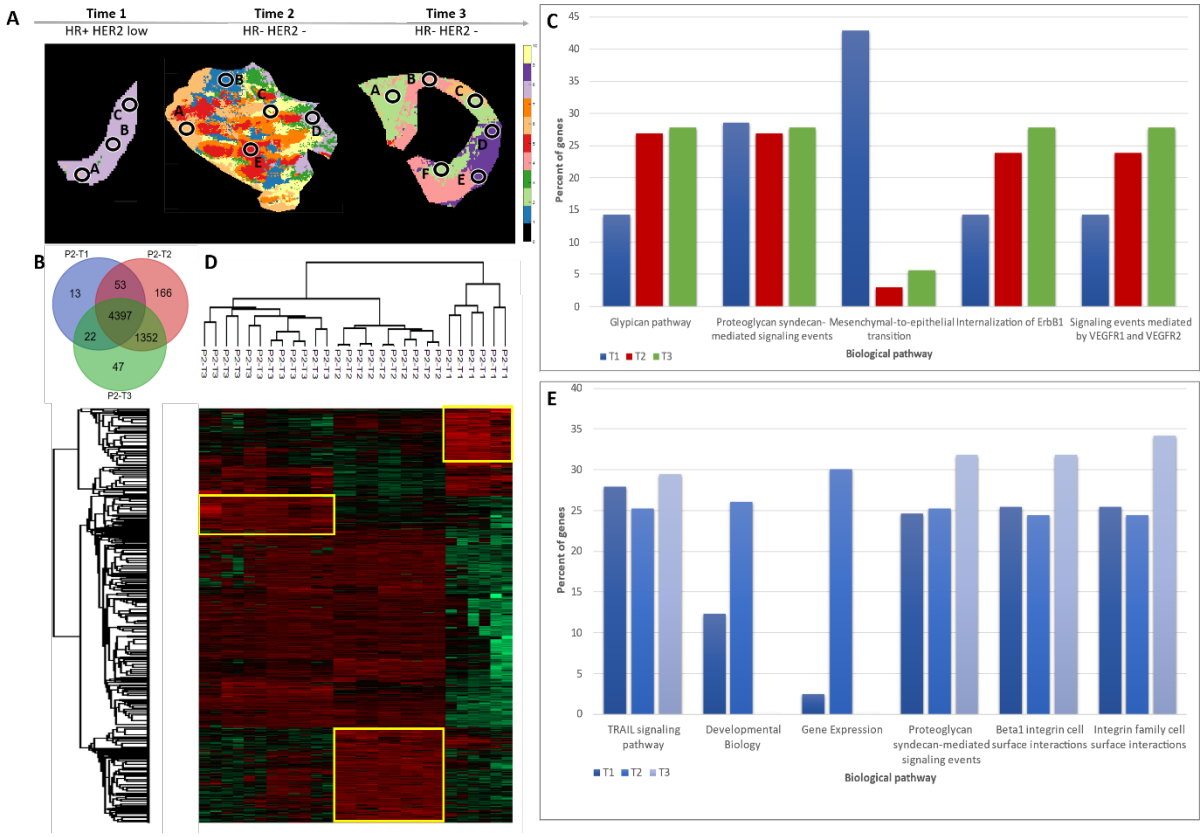


Figure 44: Patient 2 over time tumoral heterogeneity evolution analysis. A) Time point tissue co-segmentation. B) Venn diagram of time point tissues proteome comparison with C) FUNRICH biological pathway analysis involving exclusive proteins of each tissue. D) Heatmap of over-expressed proteins according to time point tissues after ANOVA (p-value<0,01), and E) FUNRICH biological pathway analysis involving over-express protein cluster specific to each tissue.

Patient 11 Tumoral Heterogeneity Over Time Analysis

The results for patient 11 were more encouraging for recovery. Three tissue samples were analyzed (P11T1, P11T2, and P11T3), all from BC metastases treated with trastuzumab. P11T1 and P11T2 exhibited an HR-/HER2+ subtype, while P11T3 shifted to an HR+/HER2low subtype (**Table 9**). The co-segmentation analysis revealed significant differences in tumor heterogeneity between these time points, identifying 11 distinct clusters based on the Silhouette criterion (**Figure 45A**). In P11T1, clusters 5 and 10 were predominant, suggesting a specific molecular landscape early in the tumor's progression. However, by the time of P11T2 and P11T3, other clusters became more prominent, with cluster 5 disappearing completely and cluster 10 persisting across all three time points. This

persistence of cluster 10, coupled with the disappearance of cluster 5, highlights a significant shift in the tumor's molecular composition following the first time point. These observations suggest that trastuzumab treatment may have altered the tumor's internal composition, reducing the dominance of certain molecular profiles (e.g., cluster 5) while allowing others to persist (e.g., cluster 10). Trastuzumab, which targets the HER2 receptor, likely influenced the HER2-driven tumor cells' survival, selectively reducing HER2+ cells while shifting the tumor's molecular profile. Though P11T2 and P11T3 shared some clusters and presented a more stable heterogeneity, further differences became evident between these later time points. Notably, clusters 4, 7, and 8 emerged and expanded in P11T3, reflecting continued evolution in the tumor's molecular landscape. These new clusters may represent a shift in the tumor's biological behavior, correlating with its transition to the HR+/HER2low subtype. This transformation could indicate the tumor's response to previous treatments, including trastuzumab, which likely contributed to changes in cell signaling pathways and tumor receptor status over time. The shift from HR-/HER2+ to HR+/HER2low in Patient 11 is particularly significant. HR-/HER2+ tumors are known for their aggressive nature, driven by HER2 overexpression and the absence of hormone receptors, which leads to rapid cell proliferation, increased invasiveness, and a higher risk of metastasis. These tumors typically respond well to targeted HER2 therapies, such as trastuzumab. In contrast, the emergence of the HR+/HER2low subtype in P11T3 represents a shift toward a less aggressive form of breast cancer. HR+/HER2low tumors express hormone receptors, which makes them more responsive to hormone therapies. However, HER2low expression adds complexity to treatment, it is typically not as strong a driver of aggressive behavior as full HER2 positivity.

The results observed through imaging data were corroborated with proteomic data. Indeed, the Venn diagram revealed an important number of exclusive proteins for each time point tumor biopsies (**Figure 45B**). As previously, biological pathway involved by the later were highlighted thanks to FUNRICH analysis (**Figure 45C**). Notably, a decline was observed in the percentage of genes involved in several key pathways, including TRAIL signaling, integrin family cell surface interactions, IFN-γ signaling, and PAR1 (protease-activated receptor 1)-mediated thrombin signaling pathways in the P11 tissues over time. The integrin-mediated cell surface interactions are vital for regulating various cellular processes, including tumor cell death, adhesion, migration, and metastasis. These interactions are particularly important in the context of the TME, where they facilitate communication between tumor cells and surrounding stromal cells. The observed reduction in genes associated with these pathways over time suggests a potential decrease in tumor aggressiveness, indicating that the tumor may be becoming less invasive. This hypothesis was further supported by the notable decrease in IFN-γ and PAR1-mediated thrombin signaling pathways. The PAR1 pathway is well-documented as a marker of poor prognostic outcomes, primarily due to its involvement in

promoting tumor aggressiveness, angiogenesis, and metastasis through thrombin activation in the TME (Boire et al., 2005). The reduction in PAR1 signaling aligns with previous findings that indicate a regression in tumor aggressiveness. Moreover, the decline in TRAIL and IFN-γ signaling pathways may reflect the effective action of the treatment, since both of which are crucial for mediating tumor immunity and regulating the TME. IFN-γ is a cytokine produced by activated T cells, NK cells, and macrophages, playing a crucial role in mounting immune responses against tumors (Ding et al., 2022). The observed decrease in these pathways suggests a favorable immunological environment, which could enhance the effectiveness of therapeutic interventions.

Additionally, an ANOVA analysis (p-value < 0,01) of the common proteins across P11T1, P11T2, and P11T3 identified 2391 significative proteins over 6044, among which distinct clusters of overexpressed proteins were observed in each tissue (**Figure 45D**). Further analysis of these protein clusters using FUNRICH (**Figure 45E**) confirmed the persistence of TRAIL and integrin signaling pathways across the different clusters. This suggests that, despite the overall decrease in tumor aggressiveness, these pathways remain integral to the tumor's biology, particularly due to the metastatic nature of the tissues. Interestingly, there was a noted decrease in pathways associated with developmental biology and metabolism. This finding implies a reduction in the energy demands of the tumor, which could indicate a state of dormancy or a halt in tumor growth. Such changes are consistent with a regression in tumor aggressiveness, highlighting the potential for a more favorable clinical outcome. Indeed, TRAIL and integrin signaling pathways were still involved in the different clusters of over-expressed proteins. This is due to the metastatic nature of the tissues. What was interesting was the decrease of developmental biology and metabolism pathways, suggesting a decrease of energy demands in tumor and consequently a dormancy or a stop in tumor growth, which was also in line with tumor aggressiveness regression.

The protein data from patient 11 revealed several significant and deeper findings that may have important implications for treatment and disease progression. One of the most noteworthy observations was the overexpression of ERBB2 in the tumor samples from stages P11T1 and P11T2. This overexpression is indicative of a HR-/HER2+ phenotype, which is characterized by the presence of HER2 protein on the surface of cancer cells (Dean & Kane, 2021). This observation is also in line with the transition of the tumor to an HR+/HER2low subtype in P11T3 tissue. Importantly, the sustained overexpression of ERBB2 is likely linked to an improved response to targeted therapies, particularly trastuzumab, which specifically targets HER2-positive breast cancer cells, inhibiting their growth and promoting apoptosis (Dean & Kane, 2021). Therefore, the initial HER2 positivity may provide a therapeutic window during which effective targeted treatment can be applied. In addition to ERBB2, the protein TOP2A was also found to be overexpressed in the P11T1 stage. This finding is

particularly relevant as it suggests that patients receiving anti-HER2 therapies may experience significant tumor regression. Indeed, TOP2A is a critical enzyme involved in DNA replication and repair, and its overexpression has been correlated with enhanced sensitivity to certain chemotherapy agents, especially anthracyclines (Fountzilas et al., 2012). This indicates that early intervention with anti-HER2 therapies may lead to a more favorable treatment outcome in this patient. However, the data also revealed concerning trends regarding treatment resistance. The overexpression of SRC in the P11T1 and P11T2 stages, along with CCND1 in P11T3, suggests a potential development of treatment resistance. SRC is a non-receptor tyrosine kinase implicated in various signaling pathways that promote cell proliferation, survival, and metastasis. Its overexpression may contribute to tumor aggressiveness and resistance to therapies (Peiró et al., 2014).Similarly, the increased levels of CCND1 in the P11T3 stage indicate a transition to a more proliferative and possibly resistant tumor subtype. Cyclin D1 plays a crucial role in cell cycle regulation, and its overexpression can drive the cell cycle forward, leading to uncontrolled cell proliferation (Tanioka et al., 2014). Finally, while the initial overexpression of ERBB2 and TOP2A suggests opportunities for effective targeted therapy, the subsequent increases in SRC and CCND1 raise concerns about the potential for treatment resistance as the disease progresses.

To resume, the analysis of patient 11's tumor progression indicated a significant shift from an aggressive HR-/HER2+ subtype to a less aggressive HR+/HER2low subtype, suggesting a favorable response to treatment. The corroboration of imaging and proteomic data revealed distinct changes in tumor biology, including a decrease in pathways associated with aggressiveness, such as PAR1 and IFN-γ signaling. The observed reduction in energy-demanding pathways further supported the notion of diminished tumor activity, potentially reflecting a state of dormancy or slowed growth. Overall, these findings provided a promising outlook for patient 11, highlighting the importance of continuous monitoring and tailored therapeutic strategies in breast cancer management.

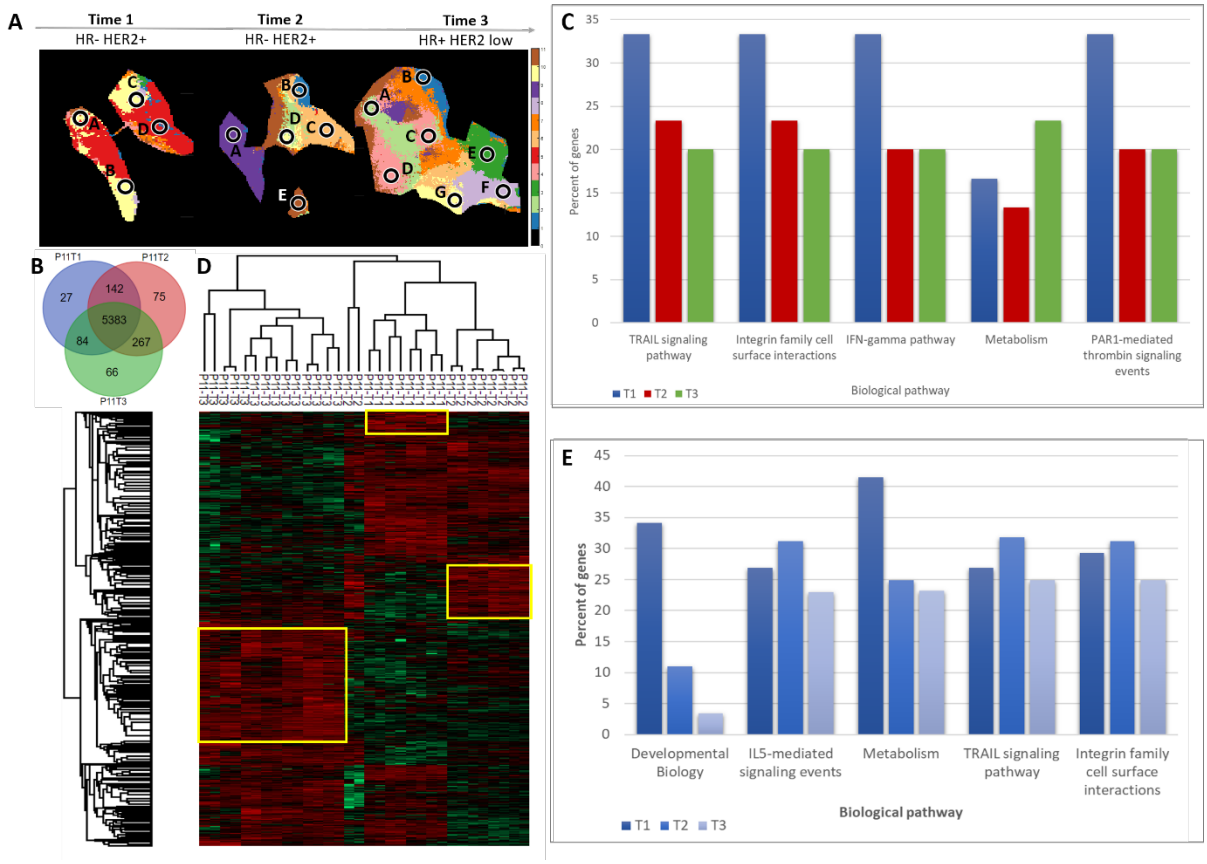


Figure 45: Patient 11 over time tumoral heterogeneity evolution analysis. A) Time point tissue co-segmentation. B) Venn diagram of time point tissues proteome comparison with C) FUNRICH biological pathway analysis involving exclusive proteins of each tissue. D) Heatmap of over-expressed proteins according to time point tissues after ANOVA (p-value<0,01), and E) FUNRICH biological pathway analysis involving over-express protein cluster specific to each tissue.

Patient 18 Tumoral Heterogeneity Over Time Analysis

Patient 18 was a different case from previous ones. Four tissue samples were analyzed over time. The first two samples, P18T1 and P18T2, were BC metastatic tissues treated with a combination of fulvestrant and palbociclib, which is a targeted therapy approach for treating HR+ and HER2-negative breast cancer. Notably, P18T1 exhibited a HR-/HER2- subtype, while P18T2 showed a HR-/HER2low subtype. The subsequent samples, P18T3 and P18T4, were also metastatic and derived from liver tissue, both classified as HR-/HER2low subtypes. However, P18T4 received a different treatment approach involving atezolizumab and bdb001, which are designed to enhance immune system activity against tumors (**Table 9**).

Subtype switch was also observed in patient 18 over time through tissue co-segmentation analysis (**Figure 46A**) across four samples (P18T1, P18T2, P18T3, and P18T4). The co-segmentation image was composed of 10 clusters according to Silhouette criterion. In the first sample, P18T1, the tissue was predominantly clustered with cluster 5 (represented in red), indicating a specific subtype or tumor environment at this time point. However, in the second sample, P18T2, a significant shift was detected, with the tissue spreading across eight distinct clusters, reflecting considerable heterogeneity in the tumor's cellular composition or microenvironment. The comparison between

the first two time points (P18T1 and P18T2) and the later samples (P18T3 and P18T4) revealed notable differences in tissue characteristics. Both P18T3 and P18T4 showed a strong association with cluster 9 (in dark purple), which was absent in the earlier samples. This persistent presence of cluster 9 in the later samples suggests that this cluster may be specific to the development of liver metastasis, indicating a shift in the disease's progression toward metastatic behavior. Moreover, the appearance of cluster 2 (light green), cluster 3 (in green) and 6 (light orange), first identified in P18T2 and later reappearing in P18T4, is of particular interest. This shared cluster between P18T2 (an earlier time point) and P18T4 (a later metastatic sample) implies a potential link between the primary tumor and the metastatic site in the liver. The presence of common clusters suggests that the liver metastasis may have originated from the breast cancer tumor, likely involving shared molecular characteristics or cellular pathways driving both the primary tumor and the metastatic spread.

After processing the protein data, the Venn diagram revealed that each tissue sample at different time points exhibited a unique proteome characterized by distinct exclusive proteins (**Figure 46B**). This finding underscores the complexity of the proteomic landscape associated with each time point in the study. Notably, the FUNRICH analysis (**Figure 46C**) highlighted significant differences between the early time points (P18T1 and P18T2) and the later samples (P18T3 and P18T4). These differences are primarily attributed to the inherent variations in tissue types rather than tumor progression. Specifically, P18T1 and P18T2 demonstrated a higher percentage of genes involved in various biological pathways compared to the later samples. This observation suggests that the exclusive proteins identified in these early time points may stem from the unique molecular characteristics of each tissue type, rather than reflecting changes related to tumor development. Thus, the distinct proteomic profiles observed may be indicative of the physiological roles of these tissues at different stages rather than being directly linked to cancer progression.

In this context, analyzing the common proteins shared among the different tissues provided more significant and insightful information regarding the biological processes involved. The heatmap generated through ANOVA (p-value < 0,01) testing of these common proteins identified 2754 significative proteins out of 6413, facilitating the identification of various clusters of overexpressed proteins (**Figure 46D**). Interestingly, unlike findings from previous patient studies, which often reported specific clusters of overexpressed proteins linked to individual tissue types, this analysis revealed less specificity in the overexpressed protein clusters among the samples. Distinct clusters were identified for P18T1, P18T2, P18T3, P18T2T3T4, and P18T4, suggesting that while certain proteins were consistently overexpressed across different tissues, their expression patterns were not restricted to a single tissue type. To further explore the implications of these findings, a FUNRICH analysis were conducted (**Figure 46E**), which highlighted the involvement of the overexpressed proteins in several biological pathways previously implicated in tumor progression and metastasis.

Notable pathways included glypican signaling, integrin cell surface signaling, and the VEGF signaling network. These pathways are known to play critical roles in regulating cell growth, migration, and the formation of blood vessels, all of which are vital processes in tumor development.

Importantly, our analysis also revealed a trend toward decreased involvement of associated genes in these pathways as tumor progression advanced. This observation aligns with established patterns in metastatic tumors, where the functional contribution of specific genes may diminish as cancer cells acquire the ability to invade surrounding tissues and spread to distant sites.

In summary, the analysis of patient 18 revealed significant insights into the evolving nature of breast cancer and its metastatic behavior. The examination of four tissue samples over time highlighted a subtype switch from HR-/HER2- to HR-/HER2low, emphasizing the heterogeneity of the tumor environment and the dynamic nature of cancer progression. The emergence of new clusters in the later samples suggested a shift in the tumor's characteristics, particularly concerning liver metastasis, potentially indicating shared molecular pathways between the breast metastasis and the metastatic liver tissue. The distinct proteomic profiles across samples underscored the complexity of the disease, with early samples exhibiting unique proteins linked to their specific tissue types rather than to tumor development. Furthermore, the analysis of common proteins revealed critical biological pathways involved in tumor progression and metastasis.

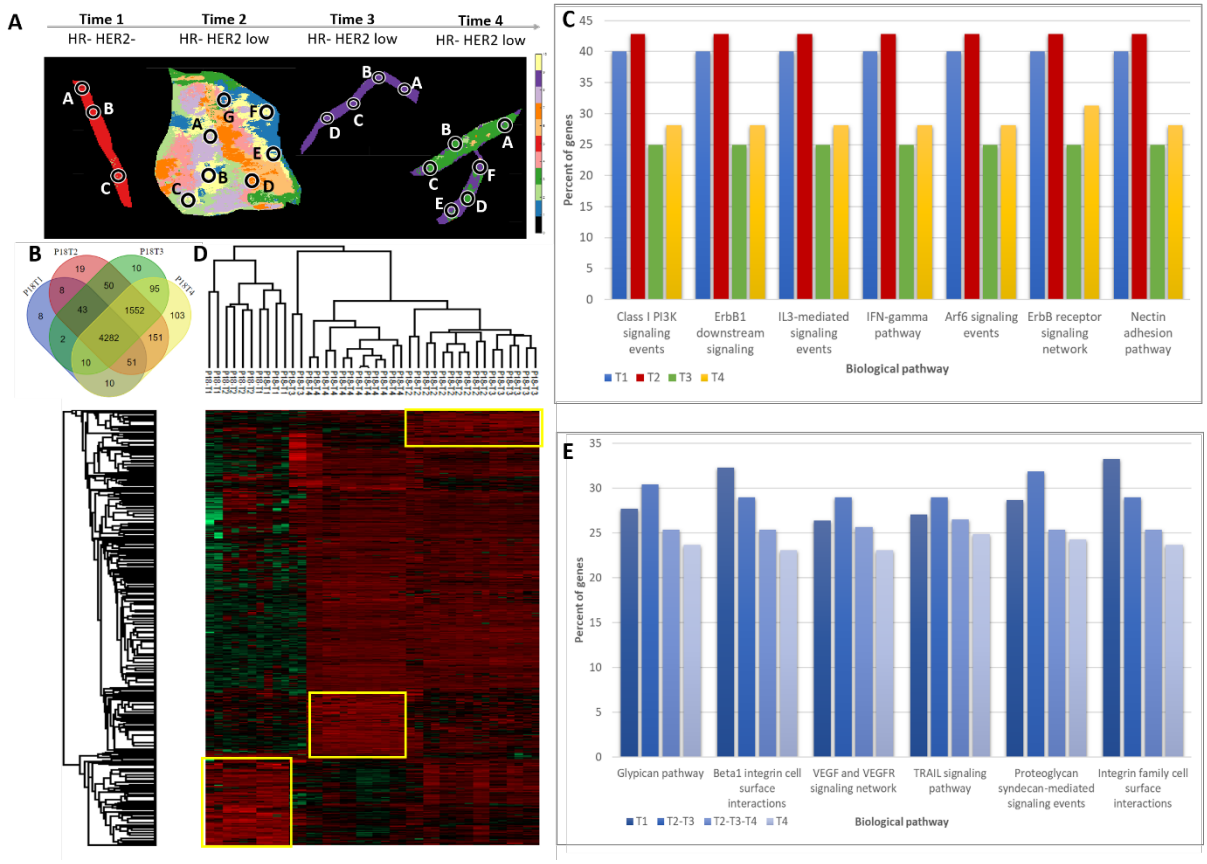


Figure 46: Patient 18 over time tumoral heterogeneity evolution analysis. A) Time point tissue co-segmentation. B) Venn diagram of time point tissues proteome comparison with C) FUNRICH biological pathway analysis involving exclusive proteins of each tissue. D) Heatmap of over-expressed proteins according to time point tissues after ANOVA (p-value<0,01), and E) FUNRICH biological pathway analysis involving over-express protein cluster specific to each tissue.

The analysis of tumor heterogeneity evolution over time in patients 2, 11, and 18 has significantly highlighted the role of the tumor microenvironment. Notably, the results indicated that heterogeneous clusters within the tumor are interacting, playing crucial roles in tumor development, recovery, and drug resistance.

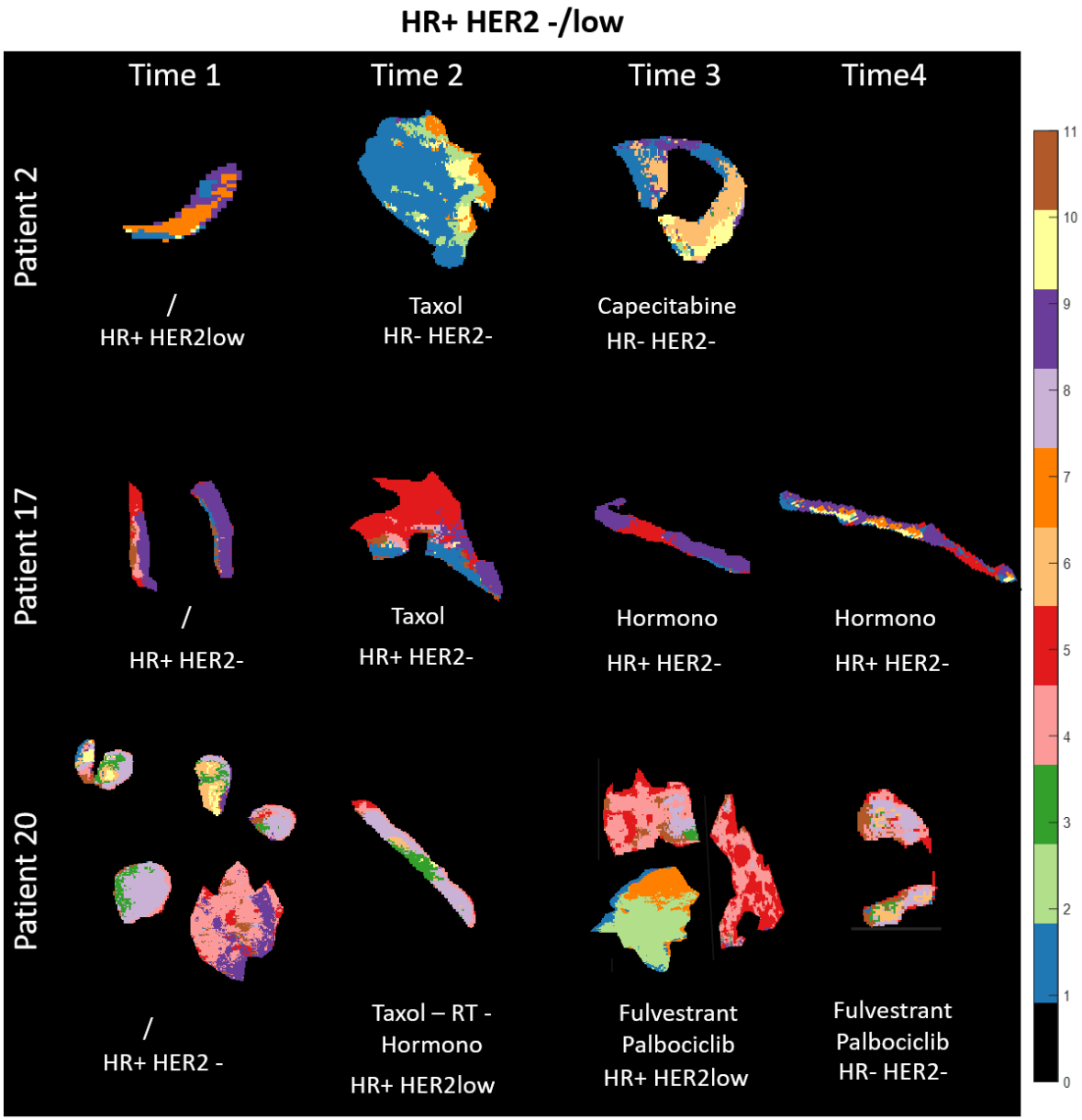
Each patient's tissue co-segmentation analysis revealed that while some clusters were consistently present across all time points, others appeared or disappeared at specific intervals. This observation suggests a dynamic communication between clusters, indicating that certain cellular subpopulations may influence each other's behavior as the disease progresses. Such interactions within the TME could facilitate adaptation mechanisms, contributing to the tumor's ability to evolve and resist treatment. These findings underscore the importance of understanding the interplay between tumor heterogeneity and the TME, as this relationship is pivotal in shaping the tumor's response to therapies and its overall progression. Here is introduced the notion of breast cancer heterogeneity community.

Breast Cancer Heterogeneity Community Study, Towards Psychohistory

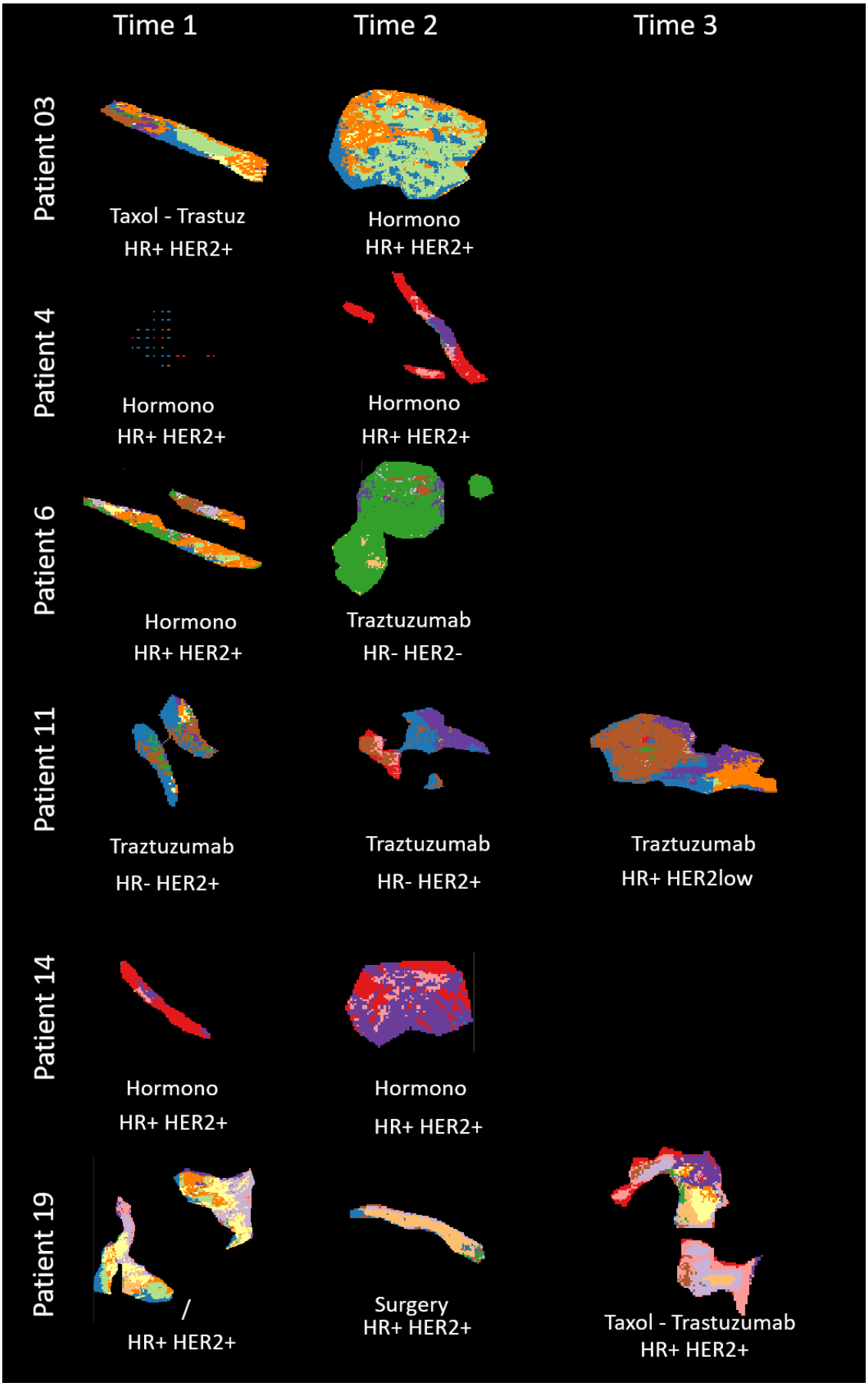
To gain a comprehensive view of BC heterogeneity in relation to treatment, tumor subtype, and progression over time, a co-segmentation analysis was conducted on the entire cohort of tumor tissues (**Figure 47**). The segmentation identified 11 clusters based on the Silhouette criterion. The results were unexpected, as no specific cluster could be distinctly associated with a particular BC subtype or treatment regimen.

For example, despite the differences in subtypes, Patient 2 (HR+ HER2-/low), Patient 3 (HER2+), and Patient 8 (HR- HER2-/low) all displayed similar clusters, such as cluster 1 (blue), cluster 2 (light green), cluster 7 (orange), and cluster 10 (yellow) (**Figure 47**). This indicates that while these tumors are classified into different molecular subtypes based on receptor status, they exhibit overlapping patterns of tissue heterogeneity. This observation challenges the conventional understanding that breast cancer subtypes are always reflected in distinct tissue architectures. Instead, it suggests that factors beyond subtype classification, such as the tumor microenvironment and possibly the evolutionary history of the tumor, might contribute to these shared heterogeneity profiles.

Similarly, the co-segmentation analysis showed no clear pattern based on treatment regimens. For instance, even though Patients 17, 3, and 8 all received hormone therapy, their tumor profiles differed significantly. Patient 17 exhibited clusters 5 and 9 (red and dark purple), whereas Patients 3 and 8 displayed clusters 1, 2, 7, and 10 (**Figure 47**). This suggests that treatment alone may not fully determine the tumor's heterogeneity profile, reinforcing the idea that breast cancer's response to therapy is highly individualized, potentially influenced by a variety of biological and microenvironmental factors.


HR+ HER2 -/low
Time 1
Time 2
Time 3
Time4
Patient 2
/
HR+ HER2low
Taxol
HR- HER2-
Capecitabine
HR- HER2-
Patient 17
/
HR+ HER2-
Taxol
HR+ HER2-
Hormono
HR+ HER2-
Hormono
HR+ HER2-
Patient 20
/
HR+ HER2 -
Taxol – RT -
Hormono
HR+ HER2low
Fulvestrant
Palbociclib
HR+ HER2low
Fulvestrant
Palbociclib
HR- HER2-
11
10
9
8
7
6
5
4
3
2
1
0

HER2 +


Time 1
Time 2
Time 3
Patient 03
Taxol - Trastuz
HR+ HER2+
Hormono
HR+ HER2+
Patient 4
Hormono
HR+ HER2+
Hormono
HR+ HER2+
Patient 6
Hormono
HR+ HER2+
Traztuzumab
HR- HER2-
Patient 11
Traztuzumab
HR- HER2+
Traztuzumab
HR- HER2+
Traztuzumab
HR+ HER2low
Patient 14
Hormono
HR+ HER2+
Hormono
HR+ HER2+
Patient 19
/
HR+ HER2+
Surgery
HR+ HER2+
Taxol - Trastuzumab
HR+ HER2+

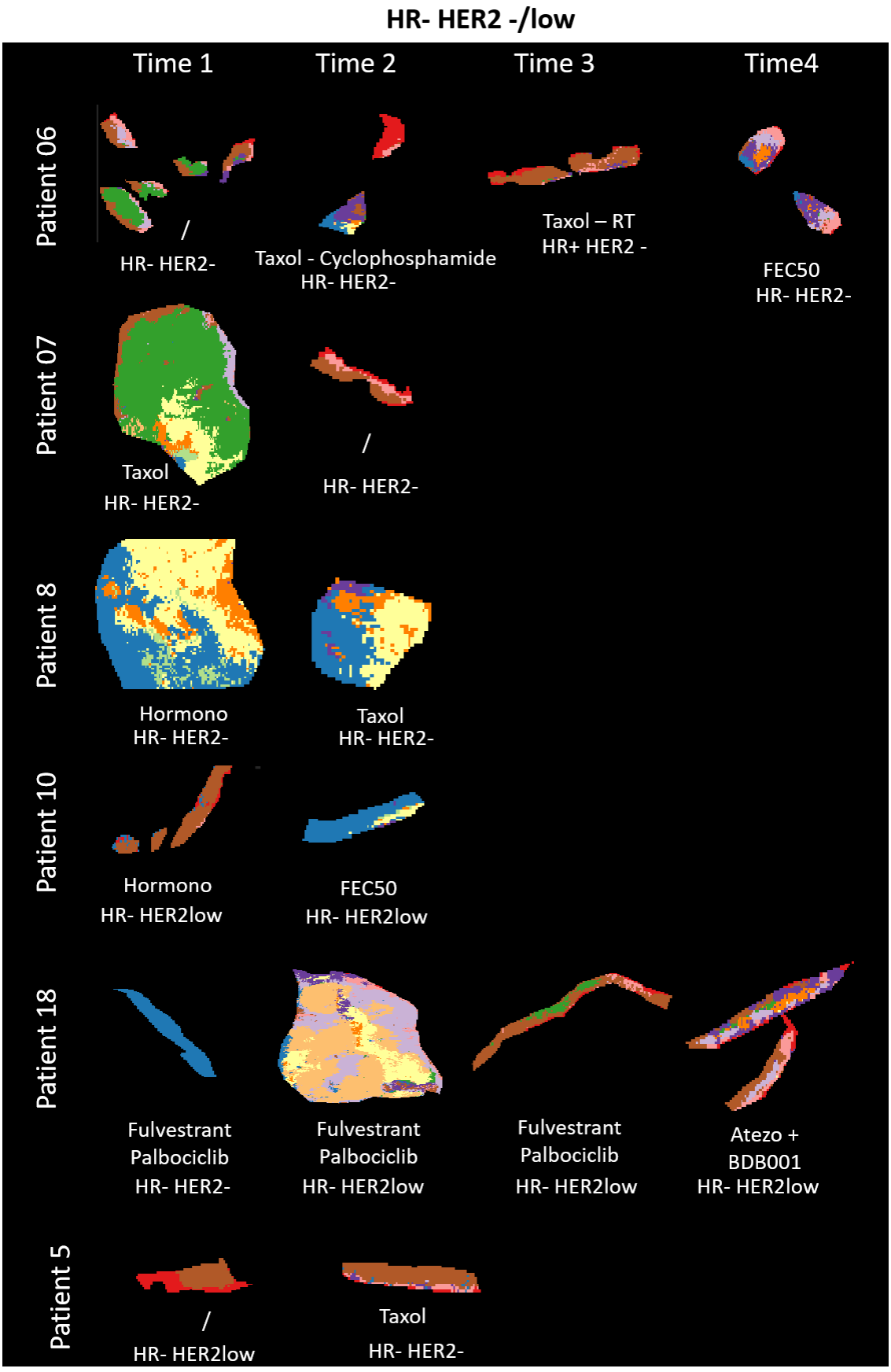


Figure 47: Co-segmentation over whole breast cancer cohort, segmented with 11 clusters according to Silhouette criterion.

Despite this variability, tumor progression over time was still evident in the data. For patients 2, 11, and 18 (**Figure 48**), comparing the over-time tissue co-segmentation (**Figure 48A**) with the whole cohort segmentation (**Figure 48B**) highlighted consistent patterns of heterogeneity and tumor evolution. Both individual and cohort-level segmentations aligned with the overall trends seen in previous analyses, confirming tumor progression according to response to the treatment. This further suggests that as tumors evolve, their microenvironment becomes increasingly diverse, possibly contributing to resistance mechanisms and therapeutic failures. The patterns of progression observed were in line with previously identified molecular markers of tumor advancement, reinforcing the importance of continuously monitoring tumor heterogeneity over time to better understand its implications for treatment resistance and disease progression.

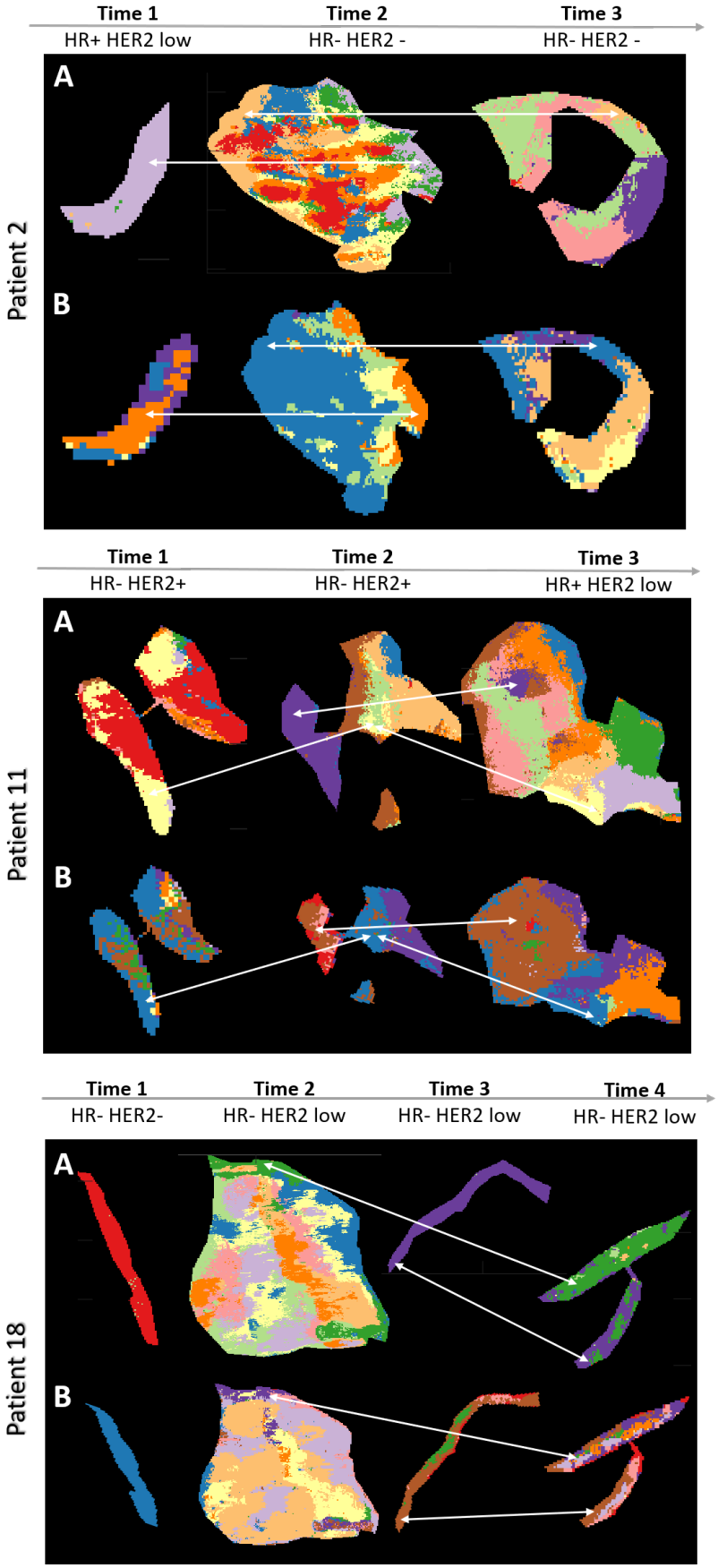


Figure 48: Comparison of A) over-time tissue co-segmentation and B) whole cohort segmentation focusing on patient 2, 11 and 18.

These findings underscored the complex, dynamic nature of breast cancer. The fact that tumor heterogeneity evolves over time, regardless of treatment and subtype, highlights the challenges of treating BC. This study highlighted the pivotal role of the BC tumor microenvironment in shaping tumor progression and treatment outcomes. Despite the different molecular subtypes and treatment regimens across the cohort, tumor evolution largely followed its initial molecular patterns, independent of both subtype and therapy. It was shown that the TME exerted a significant influence on the tumor's adaptive mechanisms, allowing it to evolve and diversify in response to the microenvironment, rather than being solely determined by subtype classification. The consistency in tumor heterogeneity patterns over time, observed even across different treatments, reinforced the idea that tumors possessed an inherent "psychohistory", like a biological narrative shaped by their original molecular characteristics and the evolving dynamics of the TME. This psychohistory appeared to govern how tumors adapted to treatments, developed resistance, and continued progressing despite therapeutic interventions. Therefore, understanding this evolutionary trajectory and the role of the TME was crucial for developing more effective, personalized treatment strategies that addressed the tumor's broader biological context, including its capacity for adaptation. Given these findings, it becomes clear that patient heterogeneity must be analyzed on a case-by-case basis. Tumor evolution cannot be generalized solely by subtype or treatment regimen, as individual tumor progression is deeply influenced by unique molecular characteristics and interactions with the TME. Personalized approaches, focusing on the specific heterogeneity of each patient's tumor, are essential for improving therapeutic efficacy and patient outcomes.

Conclusion and Perspectives

This chapter offered an in-depth examination of intra-tumor heterogeneity in breast cancer, utilizing peptide MALDI MSI and spatial proteomic data from a cohort of 16 patients with varying subtypes, including luminal, triple-negative, and HER2-low breast cancer, each contributing multiple biopsies over time. By concentrating on three specific patients (Patient 2 (HR+/HER2- or low), Patient 11 (HER2+), and Patient 18 (HR-/HER2- or low)) the study illuminated the intricate biology of tumors and the differences in their behavior, which could significantly affect prognosis and treatment responses.

The discovery of molecular clusters and distinct proteomic profiles underscored the dynamic evolution of tumors over time and treatment. For example, Patient 2 exhibited notable shifts in tumor heterogeneity over time, with different proteomic profiles and biological pathways emerging at each stage. The heightened involvement of glypican and VEGFR pathways, along with changes in MET/EMT processes, indicated a transition toward a more aggressive tumor phenotype. These findings, combined with alterations in TRAIL signaling and integrin pathways, demonstrated the

complex nature of breast cancer progression and the development of resistance to treatments. Patient 11's analysis revealed a significant shift from a highly aggressive HR-/HER2+ subtype to a less aggressive HR+/HER2low subtype, suggesting a positive response to therapy. The integration of imaging and proteomic data highlighted a reduction in pathways linked to aggressive behavior, emphasizing the necessity for ongoing monitoring and customized treatment strategies in managing breast cancer. In the same way, the analysis of patient 18 presented critical insights into the evolving characteristics of breast cancer and its metastatic behavior. The review of four metastatic tissue samples over time revealed a shift in subtype from HR-/HER2- to HR-/HER2low, reflecting the tumor's heterogeneity and the evolving nature of cancer progression. The emergence of new clusters in later samples indicated changes in the tumor's characteristics, especially concerning liver metastasis, hinting at potential shared molecular pathways between the primary tumor and its metastatic sites. The distinct proteomic profiles across samples further emphasized the complexity of the disease, with earlier samples exhibiting unique proteins associated with their specific tissue types.

Importantly, this analysis revealed the essential role of the tumor microenvironment and the interactions among heterogeneous cellular clusters. These interactions promoted adaptive mechanisms that contributed to tumor evolution and resistance to treatment. The idea of a "breast cancer heterogeneity community" emerged, illustrating how different cellular subpopulations communicated and influenced each other's behavior throughout the progression of the disease.

To gain an understanding of breast cancer heterogeneity concerning treatment, tumor subtype, TME and progression over time, a co-segmentation analysis was conducted on the entire cohort of tumor tissues. This analysis identified 11 clusters using the Silhouette criterion, demonstrating that no specific cluster could be directly linked to a particular breast cancer subtype or treatment approach. Nevertheless, tumor progression over time was clear, with consistent patterns of heterogeneity and evolution seen across patients 2, 11, and 18. The study highlighted the multifaceted and evolving nature of breast cancer, revealing that tumor heterogeneity changes over time, irrespective of treatment or subtype. These results illustrate the substantial challenges associated with treating breast cancer and stress the crucial role of the TME in guiding tumor progression and impacting treatment outcomes. Despite differences in molecular subtypes and therapeutic regimens, tumor evolution primarily adhered to its initial molecular characteristics, indicating a degree of independence from both subtype classification and therapy.

Finally, the study demonstrated that the TME plays a vital role in shaping the tumor's adaptive strategies, enabling it to diversify and evolve in response to its environment. This adaptability reinforces the notion of a tumor's "psychohistory," a biological narrative formed by its

foundational molecular traits and the shifting dynamics of the TME, which governs how tumors respond to therapies, develop resistance, and continue to progress despite treatments.

Consequently, understanding the evolutionary path of tumors and the influence of the TME is essential for creating more effective and personalized treatment strategies. Given these insights, it is evident that patient heterogeneity should be evaluated on an individual basis. Tumor evolution cannot be broadly generalized based solely on subtype or treatment regimen; rather, individual tumor progression is deeply affected by unique molecular features and interactions with the TME. Therefore, personalized approaches that prioritize the specific heterogeneity of each patient's tumor are crucial for enhancing therapeutic effectiveness and improving patient outcomes.

Future directions for this research will focus on conducting a thorough analysis of the patient cohort by integrating advanced machine learning techniques. This will enable us to track the evolution of tumor cluster communities over time and in response to various treatments using MALDI MSI and spatial proteomic data. By employing machine learning, we can simplify the complexity of the data generated, leading to clearer insights into the relationships among different tumor populations and their microenvironments. This will enhance our understanding of tumor biology and help us identify distinct molecular signatures associated with various breast cancer subtypes.

We plan to develop a sophisticated machine learning model based on the dry proteomic concept, which will automate the identification of heterogeneous breast cancer clusters from MALDI MSI data. This model will analyze large datasets to detect patterns that characterize different tumor microenvironments and their proteomic profiles. By linking protein expression data with clinical outcomes, the model will provide vital information for therapeutic decision-making, enabling more precise diagnoses tailored to each patient’s unique tumor characteristics. Additionally, its predictive capabilities will allow us to anticipate drug resistance and tumor evolution, empowering clinicians to adjust treatment plans proactively.

Integrating this research with organoid technology offers an exciting opportunity to further investigate tumor heterogeneity in a controlled laboratory setting. By creating organoid models that mimic the diversity of breast cancer subtypes, we can test various treatment strategies and monitor how tumors respond over time. This experimental framework will help us understand how different therapies impact distinct tumor clusters and uncover the mechanisms behind treatment resistance. Observing these dynamics in real time will allow us to identify key signaling pathways and molecular changes that drive tumor adaptation.

We also plan to expand the cohort to include a broader range of breast cancer clinical characteristics. This expansion will involve incorporating tissue samples from patients whose subtype diagnoses

could not be determined through conventional techniques. By including these challenging cases, we aim to gain insights into less common or atypical breast cancer subtypes that may not be well represented in current datasets. Additionally, we also could generate computational data that will complement our tissue samples.

Ultimately, these combined efforts aim to deepen our understanding of breast cancer heterogeneity and its implications for treatment. By merging advanced computational techniques with experimental models, we hope to create a comprehensive platform that clarifies the complexities of tumor biology and translates these insights into effective therapeutic strategies. This holistic approach will enhance our ability to predict how patients will respond to treatments, refine interventions, and improve outcomes for individuals facing breast cancer, marking a significant advancement in the field of precision oncology.

CHAPTER 5

General Conclusion and Perspectives

Chapter 5 : General Conclusion and Perspectives

Comprehensive Insights and Future Directions in Breast Cancer Heterogeneity Analysis

Advances in technology have fundamentally reshaped medical practice and clinical research, which once relied heavily on physical examinations and patient histories. The introduction of DNA sequencing marked a turning point, giving rise to genomics and the era of personalized and precision medicine. The premise was that sequencing a patient's genome would allow treatments to be tailored specifically to their genetic profile. Genomics has greatly enhanced our understanding of the genetic underpinnings of both normal physiology and disease. However, while genetic mutations can help refine treatments for some, relying on genomics alone has proven insufficient for fully addressing the complexity of many diseases. One of the major challenges is tumor heterogeneity, where different cells within the same tumor exhibit distinct genetic and molecular characteristics. This diversity leads to varied responses to treatment and contributes to drug resistance. To confront this issue, researches have been expanded beyond genomics to include large-scale analysis of gene products, such as metabolites, lipids and proteins. Precision medicine is increasingly moving towards integrating data from multiple "omics" layers: genomics, proteomics and metabolomics. This approach offers a more detailed understanding of disease biology, particularly in complex cancer conditions, and holds the potential to overcome the challenges posed by tumor heterogeneity and improve treatment outcomes.

During this thesis, many projects were engaged to try to tackle the tumor heterogeneity to improve diagnosis and treatments and patient prognosis. To do so the study was focused on breast cancer, which the primary cause on cancer death within woman. Technics of MSI and spatial proteomics used in this study, allowed to better understand the tumor heterogeneity.

Firstly, as detailed in **CHAPTER 2: Organoids for Luminal Breast Cancer Therapy Guidance Including Molecular Heterogeneity**, a comprehensive analysis of luminal breast cancer tissues using MSI and spatial proteomics revealed the intra-tumoral complexity in four different breast cancer samples. By combining MSI with image clustering techniques, several clusters of distinct submolecular populations were observed. These findings were further supported by protein data, which identified unique proteins involved in various biological pathways within these clusters. This confirmed that each cluster had a distinct proteome and biological network.

The analysis of over-expressed proteins within these common clusters of proteins reinforced these observations, as different clusters exhibited specific sets of over-expressed proteins linked to distinct biological pathways. Thus, a tumor is composed of heterogeneous molecular clusters, each with its

own proteome and pathway involvement. This supports the hypothesis that different molecular clusters, due to their specific biology, may respond differently to tumor progression, depending on their sensitivity to treatment, tumor microenvironment, and patient-specific factors. Notably, the protein data also highlighted potential markers of treatment resistance, as well as potential therapeutic or druggable targets that were either cluster-specific or shared across clusters.

These insights enabled the proposal of treatments tailored to the tumor's protein data, specifically targeting the heterogeneous clusters. Protein analysis of the organoids was conducted to confirm the molecular correlation between the primary tumor and the organoids. This validation was primarily achieved by verifying the presence of protein markers that were initially identified in the tumor's protein data. Organoids treated with drugs based on the protein analysis showed a higher rate of tumor cell death, a promising result that demonstrated the effectiveness of MSI combined with spatial proteomics in addressing tumor heterogeneity. For example in tumor 1, key pathways such as telomerase activity in clone A and integrin-related functions in clones B and C were identified as critical drivers of cancer cell behavior, highlighting potential drug targets like telomerase inhibitors and integrin modulators. The proteomic data enabled the identification of specific drug targets, such as FASN in clone A and PSMB1 in clone B, leading to tailored drug combinations like cerulenin and sunitinib. This personalized strategy has the potential to enhance treatment efficacy by directly addressing the specific vulnerabilities of each clone, thereby overcoming the limitations of traditional therapies. Validation of these proteomic-based treatment regimens in organoid models demonstrated their effectiveness, as indicated by lower IC50 values compared to conventional drugs like paclitaxel, which showed potential resistance associated with proteins like EDIL3 and CA12 in resistant clones.

Thus, this approach proved to be an effective method for personalizing treatment by addressing tumor heterogeneity, fully aligning with the principles of personalized medicine. Additionally, a co-segmentation analysis of the four tumors identified shared clusters and biological pathways across certain tumors, suggesting that specific treatments could potentially target similar clusters in different patients, enhancing treatment efficacy across diverse cases.

However, it is important to note that this process requires significant expertise and time, making it difficult to implement in routine clinical practice.

This is where the "dry proteomic" concept, presented in **CHAPTER 3: Dry Proteomic Concept Based on Lipid MALDI MSI**, becomes relevant. The dry proteomic approach is based on the idea that if clusters with the same spatial localization are consistently observed in MSI data, regardless of the type of omics-MSI used, it suggests that these clusters possess unique and distinct molecular signatures associated with specific biological pathways. Essentially, each cluster can be thought of as

having a unique "barcode" that distinguishes it from others. The advantage of this concept is that if a specific cluster is identified through lipid MSI (the most rapid and straightforward MSI technique, which can be easily integrated into routine clinical practice), its corresponding proteins and biological pathways can be automatically inferred without the need for additional experiments. During this study, machine learning was integrated into the MSI and spatial proteomics workflow to develop this predictive approach.

The method, developed through lipid MALDI MSI, utilizes a machine learning model trained to identify distinct clusters based on their lipid signatures. Once the clusters are identified, the model can automatically link them to their specific protein profiles and associated biological pathways. This includes providing valuable therapeutic insights, such as potential drug targets, markers of drug resistance, and prognostic indicators, as discussed earlier in **CHAPTER 2**.

CHAPTER 2 further elaborates on the development of this dry proteomic machine learning method, initially tested on rat brain tissue and then applied to a glioblastoma study, leveraging data from Duhamel et al., 2022 research.

In summary, the first two studies demonstrated the intra-tumoral complexity of BC using MSI and spatial proteomics, revealing distinct molecular clusters with unique proteomes and biological networks that may respond differently to tumor progression and treatment. A detailed analysis of the protein data from these heterogeneous clusters enabled the proposal of personalized treatments tailored to target them. Validation using organoid models showed a higher rate of tumor cell death when treated with drugs selected based on this protein analysis, highlighting the promise of this approach for personalized medicine. However, the complexity and time required for such in-depth analysis present challenges for routine clinical use. To overcome this, the "dry proteomic" concept was introduced as a streamlined tool. This approach identifies molecular clusters based on lipid MSI signatures and automatically associates them with specific protein profiles and biological pathways, providing a predictive and efficient method for clinical application. It allows the rapid identification of therapeutic targets, drug resistance markers, and prognostic indicators without the need for additional experiments. As a result, all the workflows, knowledge, and machine learning tools developed are now ready to be applied to a larger cohort of BC cases.

CHAPTER 4: 4D Longitudinal Proteomics Tracking of Breast Cancer Heterogeneity Community Response to Therapeutics presented preliminary results from a study on a complex breast cancer cohort, which included FFPE tissue samples from 16 patients with different subtypes: HR-/HER2- or low, HER2+, and HR+/HER2- or low. For each patient, multiple tissue samples were available, representing different time points due to relapse or metastasis, resulting in a total of 48 tissues analyzed. Using peptide MALDI MSI and spatial proteomics, nearly 200 molecular clusters

were identified and analyzed for their proteomic profiles. As the intra-tumoral heterogeneity analysis was already developed in previous chapters, this study was focused on the tumoral heterogeneity evolution according to time and treatments response. The study focused on three specific patients (Patients 2, 11, and 18), uncovering the complex and evolving nature of tumors, highlighting how distinct molecular clusters, each with their unique proteomic profiles, influence tumor progression and treatment response.

Important findings included the dynamic evolution of tumors, with distinct changes in proteomic profiles and biological pathways as cancer progressed or metastasized. For example, in Patient 2, shifts in glypican, VEGFR, MET/EMT, and TRAIL signaling pathways reflected an increasingly aggressive tumor phenotype over time. Patient 11 showed a positive response to therapy, with a shift from a highly aggressive subtype to a less aggressive one, underscoring the need for tailored treatment strategies. Patient 18's analysis revealed evolving tumor characteristics, particularly in relation to liver metastasis, demonstrating the tumor's adaptability and heterogeneity.

Interestingly, the study highlighted the critical role of the tumor microenvironment in driving tumor progression and resistance to treatment. A co-segmentation analysis of the entire cohort identified 11 molecular clusters, though none could be directly tied to a specific breast cancer subtype or treatment. This indicated that tumor heterogeneity is influenced by factors beyond molecular subtype, with progression largely governed by the tumor's initial molecular traits and the surrounding TME.

In conclusion first results provide an overview on the tumor's evolutionary path and its interaction with the TME is crucial for developing effective, personalized treatment strategies. The complexity of breast cancer requires patient-specific approaches, as tumor evolution is highly individualized and cannot be generalized based on subtype or treatment alone.

Looking ahead, the research will take a major step forward by integrating advanced machine learning techniques to further analyze and track tumor cluster evolution. These tools will allow for the identification of specific and common molecular signatures across various breast cancer subtypes, offering a deeper understanding of how tumors evolve in response to different treatments and over time. By continuously refining these machine learning algorithms, the research aims to reveal patterns within the data that may not be immediately apparent, enhancing the ability to predict how individual tumors will behave, adapt, and potentially resist treatment. An important part of this next phase will be the expansion of the patient cohort to include more diverse and challenging breast cancer cases. This will allow the study to capture a wider range of tumor behaviors and subtypes, including rarer and more aggressive forms of breast cancer that may not have been well represented in the initial analysis. By incorporating these diverse cases, the research aims to build a

more robust dataset that better reflects the full spectrum of breast cancer, enabling the discovery of molecular patterns that are universally applicable or unique to specific subtypes. The comprehensive approach will need computational models with experimental research, aiming to provide clearer insights into the dynamic nature of breast cancer. A significant innovation in this process will be the development of a machine learning model based on the "dry proteomic" concept.

In parallel, the study will utilize organoid models to investigate tumor responses to various treatments in real-time. These organoid models offer a controlled environment that allows researchers to closely observe how tumors adapt, develop resistance, and progress under different therapeutic conditions. By examining the effects of various therapies on these models, the research aims to generate valuable insights into the mechanisms underlying treatment resistance. This understanding will be crucial for refining future interventions and enhancing patient outcomes. The data generated from these studies will be integrated into the machine learning model, further improving treatment recommendations tailored to specific tumors. By incorporating insights regarding potential resistance mechanisms and effective responses, the model will be able to predict how individual tumors may respond to different therapies. This approach will ensure that treatment strategies are not only personalized but also optimized for the unique characteristics of each patient's tumor, ultimately leading to more effective and targeted therapeutic interventions.

Additionally, the major objective of the next phase is to compare the molecular clusters identified in breast cancer with those found in other cancer types. By doing so, the research aims to determine whether certain clusters are "pan-cancer," meaning they are common across multiple cancer types, or whether they are specific to breast cancer or other malignancies. This comparative analysis could help identify cancer-specific clusters and characteristics, providing insights into shared mechanisms of cancer development and resistance that transcend individual cancer types. Such findings would not only deepen our understanding of breast cancer but also have broader implications for oncology, potentially leading to the identification of new therapeutic targets or universal biomarkers that could be applied across multiple forms of cancer.

In conclusion, this comprehensive, multi-faceted approach aims to transform the way we understand and treat breast cancer. By combining cutting-edge machine learning with experimental techniques like MALDI MSI and organoid modeling, the research will provide a platform for more personalized, precise, and effective treatment strategies. These advances will improve the ability to predict how each patient's tumor will respond to therapy, enabling clinicians to make more informed decisions and adjust treatments proactively to counteract tumor evolution and resistance. In doing so, this work will contribute to the ongoing evolution of precision oncology, aiming to significantly improve outcomes for patients facing breast cancer.

Perspectives

At the clonal level, consider how many cells are different from one clone to another. What are the characteristics of the cells that are different from the rest of the population? If EMT is shaping this cell during the development of the tumor, it is understandable that these cells already have differences in their nature compared to the others. This is like what we see in human populations. For example, during war, some people become resistant while most of the population does not. Why do these people choose to be resistant? What elements have made them different? Social environment? In utero environment? Family pressure? Such an analogy can be surprising, but if we think of the clone as a population of cells, in which whatever their difference, they will develop in the same way compared to some that go to resistance. It can be tangible to make this kind of comparison. It may be understandable that the development of a BC tumor at an early stage could be dependent on its environment. Tumor cells growing among immune cells, if they survive, can more easily switch to a resistant profile, whereas those growing in a protective environment do not face the same selection pressure and can be destroyed more easily. Similarly, if a tumor starts in an area with a high density of breast cells compared to an area with a much lower cell density, the ability of the tumor cells to escape and switch to metastasis is understandable. In this context, from a population of cells, the switch can be made by a few cells rather than the whole population. So it would be a challenge to find the characteristics of the cells that share some characteristics that then confer the ability to switch to resistance. So, we can speculate that there is such a determinism in BC tumors. Some of the cells will resist any treatment. Finding such cells is key to understanding how the cells switch and how we can change the fate of these deterministic cells. This requires single cell analyses for each clone. To follow these presumed future resistant cells. Integrating these data with those from spatial proteomics will allow us to evaluate how many of these cells are found per clone and how they evolve over time with the different treatments used. It would be interesting to localize these cells in the tumor. Using machine learning, we will know how many of them are present in different cones of a complete cohort of several patients and specify the EMT of these cells. Deep learning can then be applied to this population of cells that share specific characteristics, and following these cells over time could be a great advantage in determining how these cells will evolve in the tumor over the course of treatment. Developing an AI that can predict how these cells will evolve can change the trajectory of such resistant cells. Recently, researchers at the National Institutes of Health (NIH) developed an AI tool that uses data from individual cells within tumors to predict whether a person's cancer will respond to a particular drug (S. Sinha et al., 2024). They use transfer learning to train an AI model to predict drug response using widely available bulk RNA sequencing data, but then fine-tune this model using single-cell RNA sequencing data. Using this approach on published cell line data from large-scale drug screens, the researchers-built AI models

for 44 Food and Drug Administration-approved cancer drugs. The AI models accurately predicted how individual cells would respond to both single drugs and drug combinations. The researchers then tested their approach on published data from 41 multiple myeloma patients treated with a combination of four drugs and 33 breast cancer patients treated with a combination of two drugs. The researchers discovered that if only one clone was resistant to a particular drug, the patient would not respond to that drug, even if all the other clones responded. In addition, the AI model successfully predicted the development of resistance in published data from 24 patients treated with targeted therapies for non-small cell lung cancer.

However, the predictive power of single-cell transcriptomic approaches is a far cry from what we can find at the proteomic level. The proteome provides unique insights into disease biology beyond the genome and transcriptome. In fact, there is a large bias between transcriptomic and proteomic data, as has already been shown (Nagaraj et al., 2011; Wisztorski et al., 2023; Y. Wu et al., 2023). Machine learning developments include Transpro, a deep learning model that predicts chemical proteomic profiles for uncharacterized cell lines using transcriptomic data, explicitly modelling information transfer from RNAs to proteins. Recently, a pan-cancer proteomic study was conducted on 949 cancer cell lines from 28 tissue types (Gonçalves et al., 2022). Integrating multi-omics, drug response and CRISPR-Cas9 gene essentiality screens with a deep learning-based pipeline reveals thousands of protein biomarkers of cancer susceptibility that are not significant at the transcript level. Randomly down sampling to just 1500 proteins has a limited impact on predictive power, consistent with highly interconnected and co-regulated protein networks. However, no specific work on such robust prediction cells is currently under investigation.

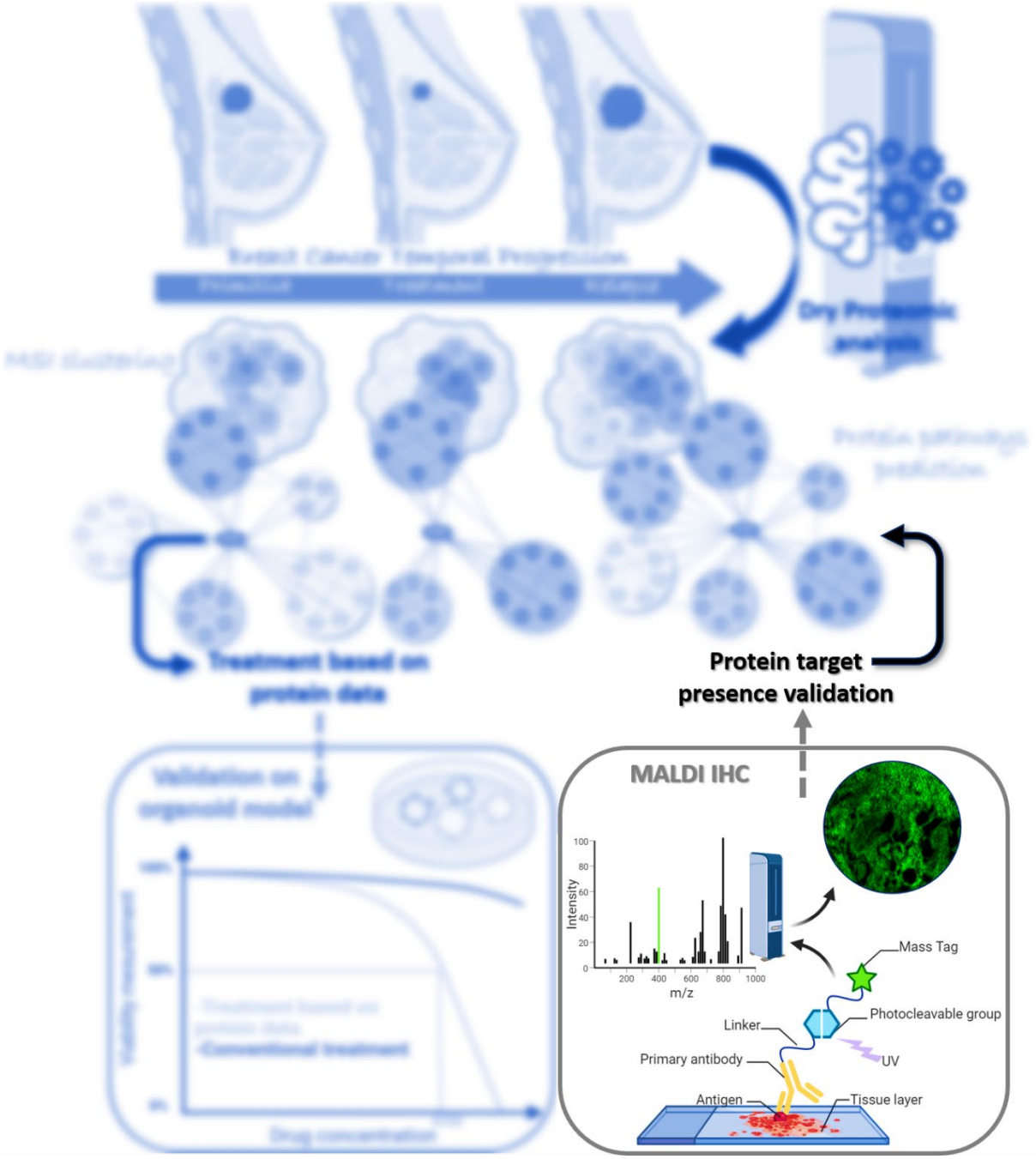
Another important point to consider when treating the tumor is the determinism of some tumor cells. As I mentioned earlier, tumor evolution is linked to the tissue and cell environment that conditions its witch to metastasize. Changing the trajectory of such resistant cells is a key to treatment by remodeling EMT. Several possibilities can be considered. In fact, tumor-associated tumors (TAM) are epigenetically modified to adopt a pro-tumor profile. Since TREM2 is a key membrane receptor in myeloid cells to switch these cells to an anti-inflammatory profile, the use of TREM2 inhibitors as ADC antibodies in conjunction with cytokines that re-activate NK cells, such as interleukin 15, which can switch NK cells to an activated phenotype, may be one direction. IL15 is also involved in switching tumor-infiltrating lymphocytes to a pro-cytotoxic profile. Using TAM as a Trojan horse could be a good angle, knowing that tumor development is based on the Red King theory of VAN VALEN LEIGH (1973), the paradox of evolution. In fact, a balance is needed between the tumor, the host and the future resistant cells. Without this balance, the host will soon stain and the tumor will stain. Thus, the future resistant cells are few and do not grow differently from the others, except when the therapeutic is used and selects between the two populations, making the

selection in favor of the resistant one. Identifying the future resistant cells and the resistant cells from the EMT by selecting the specific drug will help to modify the fate of the tumor.

We need more data at the single cell level, more spatial data and new AI that considers the psychohistory of the tumor, the red king theory and the evolutionary theory to playchess mate with the tumor cells.

CHAPTER 6

Annex Contributions Involving Tag Mass Technology for MALDI IHC Applications



CHAPTER 6: Annex Contributions Involving Tag Mass Technology for MALDI IHC Applications

Introduction

Immunohistochemistry is a laboratory technique used in fields such as biology, pathology, and medical research to investigate the presence and distribution of specific proteins within cells of tissue samples. By employing antibodies that bind to target proteins, IHC enables the visualization of these proteins' locations, thereby providing a detailed map of the molecular architecture of tissues. This ability to identify and localize specific biomarkers is essential in differentiating between various cellular components and understanding the complex spatial organization of biological tissues.

In clinical settings, IHC is indispensable for diagnostic pathology, particularly in oncology, where it helps detect markers that indicate cell proliferation, differentiation, apoptosis, hormone receptors, enzymes, and other molecular targets, which can aid in tailoring specific treatments for patients based on their unique biomarker profile. Beyond diagnostics, IHC also plays a role in research, where it is used to study the distribution and localization of various biomarkers across different tissue types. IHC technics allows to gain insights into cellular processes, disease mechanisms, and the effects of therapeutic interventions. In case of pharmacology, IHC helps evaluate the effectiveness and safety of drugs by analyzing their distribution and interaction with target tissues, thereby offering crucial data on drug mechanisms and potential side effects.

The methodology behind IHC involves the use of antibodies that are chemically linked to substances capable of producing a visible signal, such as enzymes that catalyze chromogenic reactions or fluorescent dyes that emit light when exposed to specific wavelengths. This technique allows for the detection of single or multiple proteins within a single tissue section. Multiplex IHC, which involves labeling several biomarkers at once, is particularly useful in analyzing complex and heterogeneous tissues like tumors. However, conventional multiplexing using fluorescence microscopy is often limited to detecting 3-5 biomarkers simultaneously due to spectral overlap, where the excitation and emission spectra of different fluorescent dyes interfere with one another. Advanced techniques like hyperspectral or multispectral imaging can extend the detection range to up to 8 biomarkers by better distinguishing overlapping spectral signals. Nonetheless, fluorescence-based multiplexing remains limited in the number of targets it can analyze concurrently.

To overcome these limitations, mass spectrometry has emerged as a promising alternative for high-throughput, multiplexed analysis of tissues. MS has been integrated with IHC techniques to enhance its capabilities, particularly through the use of photocleavable mass tags in mass

spectrometry imaging, such as MALDI MSI. In this approach, antibodies are modified with a mass tag that can be cleaved by UV light and detected by MS, enabling the simultaneous imaging of multiple biomarkers within a single tissue sample. MS-based methods provide a significant advantage over traditional fluorescence approaches because they can distinguish a vast array of molecules with a resolution finer than 1 Da, allowing for the precise quantification and identification of numerous proteins and other biomolecules. This technology was introduced in 2005 by PRISM (US20050687848P) and published in 2007 (Lemaire et al., 2007) based on polypeptides mass tags for MALDI or DESI or metals for SIMS and ICP-MS (US8221972B2). Recently, the technique has been commercialized by companies like Ambergen or Fluidigm. This method further enhances the ability to analyze multiple biomarkers simultaneously, overcoming the traditional limitations of multiplexing and providing a more comprehensive view of the molecular landscape of tissues.

Results

The emergence of MALDI IHC technology enabled the project to reach several conclusions for future applications:

- Validation of Biomarkers: The technology confirmed the presence of potential or predicted biomarkers in patient tissue sections, providing robust evidence for their relevance.
- Role of Immune Cells: It underscored the critical role that immune cells play in cancer development and progression, highlighting their importance in the tumor microenvironment.

During annex projects, a collaboration with Ambergen provided access to specific probes targeting cancer cells, connective tissues, and immune cells for MALDI IHC experiments. This collaboration enabled the validation of the predictive methodologies developed with SpiderMass technology across various contexts, two of which are presented in this manuscript.

The first study, titled '**SpiderMass and Machine Learning-Based Lipids Immunoscoring for Ovarian Cancer Diagnosis and Prognosis**', aimed to develop a robust classification model based on convolutional neural networks (CNN). This model can discriminate between different ovarian cancer subtypes in real time using ex vivo morphological and molecular data acquired through mass spectrometry. Since patient survival is closely linked to immune cell infiltration, we also developed a novel mass spectrometry imaging model. This model enables direct visualization of immune cell presence and localization within tissue sections. Using this data, we established an immune score based on the proportion of immune cells in the tissue to aid in diagnosis and prognosis. The immune cells detected by SpiderMass were validated through MALDI-IHC experiments, providing not only accurate diagnoses of ovarian cancer subtypes but also prognostic information to guide clinicians in selecting the most appropriate therapy for each patient.

In the second study '**Development of Molecular Digital Twins Based on Ambient Ionization Mass Spectrometry Imaging for Application in Cancer Surgery**', we introduced the concept of digital twins using precise, high-throughput molecular data from mass spectrometry imaging. The integration of digital twin (DT) technology has recently ushered in a new era of precision and efficiency in cancer surgery, building on its success in the industrial sector. We developed a machine-learning-based pipeline capable of depicting cancer cell infiltration into normal tissue, offering precise delineation of tumor margins with the help of SpiderMass. The immune cells prediction through by SpiderMass MSI were validated through MALDI-IHC experiments. This approach also facilitates the prediction of bacterial strain presence in both tumoral and healthy mammary gland tissues.

SpiderMass and Machine Learning-Based Lipids Immunoscoring for Ovarian Cancer Diagnosis and Prognosis

Introduction

Ovarian cancer is recognized as one of the deadliest cancers, with 198,000 deaths reported out of 294,000 cases worldwide (Sung et al., 2021). Advanced stages are diagnosed in 75% of patients, resulting in a 5-year survival rate of only 46% overall and 26% for stage IV (Lheureux et al., 2019). The majority of OC cases (90%) are epithelial, encompassing serous, endometrioid, mucinous, and clear cell subtypes (Köbel & Kang, 2022; Lisio et al., 2019). Among serous tumors, classifications include high-grade serous carcinoma (HGSC), low-grade serous carcinoma (LGSC), and serous borderline tumors (SBL).

Regarding the patient management, the therapeutic sequence is determined by a multidisciplinary evaluation based on an estimation of the tumor burden, staging according to FIGO classification and the histological diagnosis by laparotomy(Berek et al., 2021). For advanced stages, both American and European recommendations are to combine surgery and chemotherapy as first-line therapy(Colombo et al., 2019). The reference systemic chemotherapy is based on 6 cycles of chemotherapy with carboplatin AUC 6 and paclitaxel 175 mg/m2 every 3 weeks. PARP inhibitors can also be used as maintenance treatment for patients tested positive for homologous recombination deficiency (Alvarez Secord et al., 2021). Cytoreduction surgery (CRS) is the cornerstone of OC treatment, aiming for complete resection without residual disease to enhance survival (Barakat et al., 2022). Complete CRS, specifically achieving a CC-0 classification (no residual disease), is a significant prognostic factor for improved overall survival. CC-0 is associated with a disease-free survival of 22.3 months, in contrast to 12.3 months for CC-1 (residual disease < 2.5 mm) and 6.3 months for CC-2 (residual disease between 2.5 mm and 25 mm) (Bois et al., 2009). Median overall survival is notably impacted, with 106 months for no gross residual, 66 months for residual disease ≤ 5 mm, and 34 months for > 2 cm ($p < 0.01$) (Chi et al., 2006). If the goal is to achieve a complete resection through

radical surgery, it is preferable to limit the extent of resection to only what is strictly necessary, as it can lead to significant morbidity. The surgical complexity score, designed to summarize the extent of procedures in OC surgery, indicates that heightened complexity correlates with increased surgical morbidity (Aletti et al., 2007). Similarly, as the operating time increases, the risk of postoperative complications rises (Gerestein et al., 2010).

However, achieving optimal surgery in OC remains challenging due to a lack of tools for tissue assessment. While intraoperative frozen section analysis offers high sensitivity and specificity for suspicious pelvic masses (Ratnavelu et al., 2016), it is time-consuming and limited to 2-3 times per patient. Given the difficult context of OC, late diagnoses, and potential cancer spread in the abdominal cavities, there's a clear need for real-time tumor tracking technology during surgery to enhance patient survival and reduce post-operative issues, especially for later stages of CRS. In this context, we investigate the impact of the SpiderMass technology associated with machine learning to perform real-time diagnosis and prognosis of patient survival with OC. We introduce as well, associated with SpiderMass technology, the creation of a real-time immunoscoring from patients excised tissues (**Figure 49**).

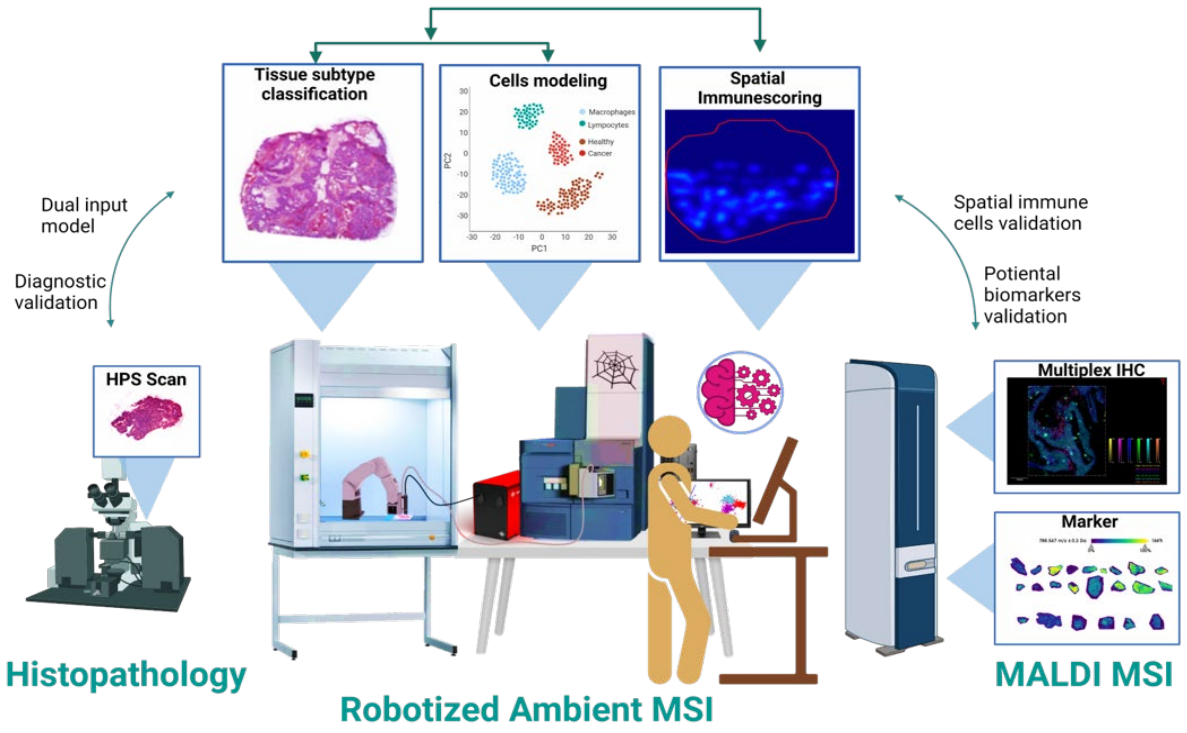


Figure 49: Overall workflow developed in the study. A robust classification model based on the combination of H&E staining and MS spectra obtained with SpiderMass was created for diagnosis of ovarian cancer. A second model was also built to predict the presence of different cell types (including immune cells) in the TME and create an immunoscore for diagnosis and prognosis.

Material and Method

Experimental Model Details

Ovarian cancer cohort

This is a single cohort study conducted at the Oscar Lambret Center in Lille, France. This is an observational study. The Institutional Review Board of the Oscar Lambret Center has confirmed that no ethical approval is required (number 2022-08). The study complies with the "reference methodology" MR004 adopted by the French Data Protection Authority (CNIL), and we checked that patients did not object to the use of their anonymized clinical data for the research purpose. Data was obtained through electronic medical record review. Oscar Lambret is a European Society of Gynecologic Oncology (ESGO)-accredited center for advanced ovarian cancer surgery.

Patient samples

As a reference center for the management of rare ovarian tumors, the Centre Oscar Lambret (COL) is once again certified by the European Society of Gynecologic Oncology (ESGO) for surgery on advanced ovarian cancer. Thus, tissue samples were obtained from patients by COL in Lille, France as a prospective cohort. Patients provided written informed consent before participating in the trial. To protect patient privacy, no personal information was used in these experiments, and a random number was assigned to each sample. A cohort of 78 fresh samples were processed between September 2020 and January 2021 including 41 normal ovaries versus 37 OC including (10 serous borderline ovaries, 13 serous high-grade ovaries, 8 mucinous ovaries, 6 endometrioids). Moreover, 79 frozen samples of endometrium issued by the same center were processed in September 2021 including 42 normal, 31 endometrioid carcinoma, and 6 serous high-grade carcinomas. A validation cohort of 24 samples has been used to validate the model. Finally, a FFPE retrospective cohort of 47 patients has been analyzed. A retrospective cohort of 83 FFPE specimens was issued by Oscar-Lambret center tissue bank including 33 HGSOC, 5 CCC, 12 SBL.

Pathology Review and Histology Control

For histology, one gynecologist pathologist (QP) read and annotated HPS stained tissues, and two human Ovarian cancer board-certified pathologists (Dr. Anne-Sophie Lemaire and Dr. Camille Pasquesonne, Oscar Lambret Center) commented and validated these annotations. The pathologists were blind to any information about the acquisition from MS studies. The annotations were done on Hemalum/Phloxine/Saffran (HPS) stained tissues. For HPS staining, a 10 µm thick tissue slice consecutive to the SpiderMass-analyzed tissue was treated with hemalum solution for 1 minute and rinsed with tap water. Then the tissue section was stained in phloxine 0.1% solution for 10 sec and rinsed with tap water before dehydration in 70% and 100% ethanol baths. Finally, the sections were dipped in saffron for 5 sec, rinsed twice in alcohol, cleaned in xylene, and mounted with cover slips

and the EUKITT® slide mounting medium. After HPS staining, the nuclei were colored dark blue, the cytoplasm pink, and the conjunctive tissue orange. The stained slide was scanned for digital image acquisition using the Pannoramic MIDI slide scanner (3DHISTECH LTD. Budapest, Hungary) and the images were viewed and exported using Pannoramic viewer 1.15.

Experimental Design

The data were collected and processed in blind after anonymization of the patients using an internal laboratory labelling. All samples of the cohort were analyzed and since ovarian cancer is a heterogeneous class of cancer originating from various types of tissues, we took great care that features discriminating the different types of tissues contributed less weight to the classification model than the other features of interest such as grading or typing. No patients or generated data were excluded from the study and outliers were defined per se when their value was more than 5 time the mean's SD. All cross-validation results are shown with and without the outliers considered. The tissues were analyzed in 5 different locations to replicate the measure and in both mode of analysis.

Tissue Preparation

Frozen samples had been snap-frozen and stored at −80°C. Samples were cut in slice of 7µm, stained with HPS and sent to pathologists for annotations using CryoStats (Leica Microsystems, Nanterre, France). The mirror of this tissues was also cut at 20µm for SpiderMass analysis and cross-validated with another cut of 12µm for MALDI-MSI analysis on ITO-coated glass slides (LaserBioLabs, Valbonne, France) and stored at -80°C. Annotated slide were used to identify the different areas to shot with SpiderMass.

Cell Lines

One healthy cell line and four ovarian cancer (PA-1, SK-OV-3, PEO4, THP-1) cell lines were cultured. Immortalized healthy epithelial ovarian cell were cultured in Prigrow I medium. PA1 cells were cultured in DMEM. SKOV3 cells were cultured in McCoy medium. PEO4 and THP1 cells were cultured with RPMI medium. All the medium were supplemented with 10% fetal bovine serum and 100 U/mL penicillin-streptomycin in a humidified air incubator at 37 °C under an atmosphere of 5% CO2. After 70% confluence, cells were washed two times with DPBS, dried under PSM during 10min at RT than analyzed directly into cell plate.

Primary Macrophages Isolation

Blood was diluted two times into PBS-EDTA. Leucocytes were isolated with 25min centrifugation 2200rpm with a Ficoll gradient. Leucocytes were than washed three times with PBS-EDTA. Leucocytes were than resuspended into RPMI medium and incubated into a cell plate 1h30 at

37°C. Cell plate were than washed three times with PBS. Macrophages were than grown 7 days with RPMI medium with 10% fetal bovine serum, 100U/mL penicillin-streptomycin and MCSF.

Macrophages Stimulation

TPH1 cell line were stimulated with 10ng/mL PMA for the macrophage differentiation. TPH1 as well as primary macrophages were stimulated into two different conditions. M1-like macrophages were stimulated with 0.5mg/mL of LPS and 20ng/mL of IFN-γ during 48h. M2-like macrophages were stimulated with 20ng/mL of Il4 during 48h.

Primary lymphocyte isolation

Peripheral blood mononuclear cells (PBMCs) were isolated from whole blood samples using density gradient centrifugation (Ficoll Paque Plus (GE Healthcare)). Then, cells were labeled with mix antibodies (Sony) : anti CD3 PE (clone SK7), anti CD4 FITC (clone A161A1), anti CD8 APC (clone HIT8a) and anti CD7 PE-Cy5 (clone CD7-6B7) for 20 minutes at 4°C in dark. After washing, CD3+ CD4+ cells, D3+CD8+ cells and CD3-CD7+ cells were sorted using the BD FACS ARIA II SORP. One million of each population was transferred onto glass slides using a Cytospin™ centrifuge (Thermo Shandon Cytospin) and conserved at -80°C.

SpiderMass Analysis

The global design of the instrument setup is described in a previous study (Saudemont et al, 2018). In brief, the system is made up of three parts: the mass spectrometer itself, a laser system for remote micro-sampling of tissues and a transfer line allowing for the transfer of the micro-sampled material. The first component consists of a pulsed Nd:YAG laser (pulse duration: 5 ns, = 1064 nm, Quantel, Les Ulis, France) pumping a tunable wavelength OPO (Radiant version 1.0.1, OPOTEK Inc., Carlsbad, CA, USA). A handpiece with a 4 cm focusing lens is attached to the end of the biocompatible laser fiber, which is connected to the laser system output and has an inner diameter of 450 microns and a length of 1 m. In these studies, the laser intensity was set to 4 mJ/pulse for a fixed irradiation time of 10 s, resulting in a laser fluence of approximately 3 J/cm2. The second component of the system is a 2 m transfer line made of Tygon ND 100-65 tubing (Akron, Ohio, USA, 2.4 mm inner diameter, 4 mm outer diameter). The transfer line is directly connected to the mass spectrometer (Xevo G2-XS, Waters, Manchester, United Kingdom) from which the conventional electrospray source was removed and replaced by a REIMS interface on one side and is attached to the laser handpiece on the other. A 200 µL/min infusion of isopropanol was administered before each acquisition. 200 µg/mL of Leucine enkephalin was added to the infusion to play the role of a lockmass. The sampling position was determined based on the histopathological annotations. The acquisition was composed of a burst of 10 laser shots resulting in an individual spectrum. Spectral

acquisition was performed both in positive and negative ion mode in sensitivity mode with a scan time of 1 s. The mass range was set to *m/z* 50–2000.

MALDI Mass Spectrometry Imaging

The SpiderMass setup was described in the previous section (Ogrinc, Saudemont, et al., 2021a). To perform imaging analysis, the Spider-Mass microprobe was coupled to a stiff robotic arm described in a previous work. The spatial step size was set to 250 µm for fresh frozen tissue to achieve oversampling. The mass-range was fixed between *m/z* 100–1500. The acquisition sequence was composed of 3 consecutive laser shots and 3 s between each step. The laser bursts and the spectrometer acquisition were automatically triggered through a MATLAB in-house user interface developed for the robotic WALDI-MSI (Ogrinc, Saudemont, et al., 2021a). The data was acquired in negative and sensitivity ion mode.

MALDI Immunohistochemistry (MALDI-IHC)

Multiplex imaging was conducted on two fresh-frozen GBM tissue samples that were previously analyzed using SpiderMass-MSI. One tissue sample had a survival rate of less than 10 months, while the other had an OS of more than 36 months. The MALDI-IHC analysis was made on an adjacent tissue section from the same tumor analyzed by SpiderMass. The tissue preparation and imaging protocol utilized was the recommended one for AmberGen (Billerica, MA) Miralys probes. Initially, the tissues were vacuum-dried for 10 min and then fixed with 1% PFA for 30 min. Subsequently, the tissues underwent a series of baths: one bath in PBS for 10 min, two baths in acetone for 3 min each, and one bath in Carnoy solution for 3 min. Following this, the tissues were rehydrated through two baths in 100% ethanol for 2 min each, succeeded by three consecutive baths in 95% EtOH, 70% EtOH, and 50% EtOH, each for 3 min. A 10-min TBS bath (50 mM Tris, pH 7.5, 200 mM NaCl) preceded antigen retrieval, which occurred in 20 mM Tris buffer at pH 9 for 30 min at 95°C. After a 10-min TBS wash, the tissues were treated with a tissue blocking buffer (2% each of normal mouse and rabbit serum and 5% BSA in TBS-OG [TBS with 0.05% w/v Octyl β-D-glucopyranoside]) for 1 h. Following this, the tissues were incubated at 4°C overnight in the same blocking buffer, which contained CD68, CD8α, Ki67, Vimentin, and collagen probes at a concentration of 2.5 µg/mL. Each slide was individually washed with three 5-min TBS baths and three 2-min baths in 50 mM NH4HCO3, all conducted in darkness. The tissues were then vacuum-dried for 1 h and 30 min at room temperature before subjecting them to a 365 nm UV exposure (Miralys Light Box from AmberGen, Inc., Billerica, MA) for 10 min to cleave the probes. DHB matrix at a concentration of 20 mg/mL in MeOH:TFA 0.1% (70:30, v/v) was sprayed onto the tissues using the HTX sprayer M5 from HTX technologies, LLC (Chappel Hill, NC). The two slides were subjected to MALDI-MSI analysis using a rapifleX MALDI-TOF-MS instrument (Bruker Daltonics, Germany) in reflector mode, positive ion

mode, with a laser spot size of 20 µm and continuous raster scanning of 20 µm. Each pixel underwent 500 laser shots, and a TIC normalization was employed for multiplex image processing. The resulting images were analyzed using flexImaging (Bruker Daltonics, Billerica, MA).

Lipid Identification

m/z intervals corresponding to loadings with the largest contribution to the explained variance observed in the different groups were selected for MS/MS-based identification. For these experiments, the settings were the same as described in the SpiderMass section. Full scans were acquired in the Xevo G2-XS. The identifications were performed directly on the tissue by doing a full scan first to verify the presence of the targeted masses. Then after switching to MS/MS mode, the ions were selected "on the fly" for with collision induced dissociation (CID) with an isolation window of 0.1 *m/z*. MS/MS spectra were acquired for a continuous irradiation time of 5 s. The lipid annotations were performed manually through Metfrag, LipidMaps and Alex 123 data-bases.

Data Processing

AMX classification

For data analysis, all raw data files produced with the SpiderMass instrument were imported into the Offline Model Builder software. After importation, spectra were first pre-processed. The pre-processing steps include background subtraction, total ion count normalization, and re-binning to a 0.1 Da window. All the processed MS spectra obtained from the 78 histologically validated samples were then used to build a principal component analysis and linear discriminant analysis (PCA-LDA) classification model. The first step consisted of PCA to reduce data multidimensionality by generating features that explain most of the variance observed. These features were then subjected to supervised analysis using LDA by setting the classes that the model will be based upon. LDA attempts to classify the sample spectra and assess the model by cross validation. Cross validation was carried out by the "leave one patient out" methods. In this method, the spectra are grouped by patient and left out one by one; at each step the model without the patient is interrogated against this model.

Statistical analysis from classification

m/z intervals corresponding to loading scores with the largest contribution to the first principal components (i.e., where ideally 80% of the variance is explained) were obtained and their normalized intensities across different classes were plotted using GraphPad Prism v9.5.1. To consider the imbalance in the numbers of sample per class , non-parametrical two-sided ANOVA (Kruskal-Wallis) was used, followed by Dunn's test and adjusted P-value to account for the multiple comparisons with a family-wise significance and confidence level of 0.05.

Optimal classification model, cross-validation and blind prediction

The Lazy Predict library (https://lazypredict.readthedocs.io/en/latest/) was used to build multiple models from the scikit-learn library by training and testing a range of 24 classifiers. The random state was always kept at 1. Subsequently, the optimal model was reconstructed individually using the scikit-learn library, which enabled tuning of its parameters for optimization and evaluation of its accuracy. To further evaluate the model's performance, 20-fold cross-validation was performed using KFold and cross_val_score functions, and the classification report was generated using the classification_report function. Additionally, the ConfusionMatrixDisplay function from the matplotlib library was used to display the confusion matrix. The optimal model was then saved and loaded for blind prediction using the joblib library, with the joblib.dump and joblib.load functions. The Local Interpretable Model-agnostic Explanations (LIME) algorithm was employed to explain decision making of the classification model. This algorithm calculates feature contributions that can be positive or negative. The ELI5 library was utilized to generate a LIME table containing the weight of feature contributions using the explain_prediction function (https://eli5.readthedocs.io/en/latest/overview.html). Next, a non- parametric statistical test Kruskal-Wallis with Bonferroni correction, employing the stats.kruskal function from the Scipy library, was used to evaluate the significance of each high feature contribution. Only significant features with a p-value of equal or less than 0.05 were retained. Finally, a step of filtering was added to only keep the mono-isotopic peaks corresponding to molecules in the final list. The corresponding box plots were then generated from the seaborn library.

Multi input neuronal network

To build and train a neuronal network model with multiple input spectra and HPS images (by using QuPath), the TensorFlow and Keras libraries were used. Spectra in csv format and HPS images in png format were combined by converting them into NumPy arrays. The model architecture is defined with separate branches for image (2D-CNN) and spectra (MLP) data inputs. Branches are concatenated and additional layers are added for classification. The model is compiled with the appropriate parameters. The script divides the data into training, validation, and testing sets (train: 0.6, validation: 0.2 and test: 0.2).

Immunoscore classification model

The LightGBM Python library was employed to train a LightGBM model, which is a gradient boosting framework developed by Microsoft. The immunoscore models were built based on cell spectra in the mass range *m/z* 600-1100 in negative ion mode including the following categories and corresponding spectra counts : macrophages (M1 & M2) : 216 spectra, healthy cell line : 163 spectra, cancer cell lines : 146 spectra, lymphocytes (NK, CD8, CD4) : 114 spectra. To retrieve the prediction scores for each cell type, the predict_proba function from the LightGBM model was used instead of

the predict function. This approach is favored because it provides probability estimates for each class by obtaining the scores for each cell type being present.

SpiderMass MSI immunoscoring

A MATLAB code was developed to extract the MS scans in each pixel from the Waters Raw file. The raw SpiderMass files were converted to mzML using MSConvert (Proteowizard). The imzML converter was used to reconstruct the imaging files. A Python code was developed for further analysis and utilizes the ImzMLParser class from the pyimzml library. It imports and parses the imzML file within the Python environment. The script iterates over each coordinate (x, y, z) in the imzML file and retrieves the corresponding mass spectra using the getspectrum method. It applies a desired mass range filter to select specific mass values and stores the spectra along with their respective coordinates. Several preprocessing steps are performed in the Python code. This includes identifying non-zero pixels, normalization, and applying Gaussian smoothing using the gaussian_filter function from the scipy.ndimage.filters module. The spatial resolution is also increased by a factor 2 using interpolation with the zoom function from scipy.ndimage. For predictions, a pre-trained LightGBM model is loaded from a joblib file using the joblib.load function. The script iterates over each spectrum in the image and prepares them for prediction. The predict_proba method of the model is used to obtain the predicted scores and labels for each spectrum. After obtaining the predicted labels, the script visualizes the label maps for each class predicted by the model. It creates a colormap using the 'jet' color scheme and iterates over each class to display the label maps using the imshow function from the matplotlib.pyplot module. A color bar of scores is also added for reference.

Results

Ovarian Cancer Histological Subtyping Based on Molecular SpiderMass Data.

The initial objective was to determine whether SpiderMass technology could provide real-time molecular typing of various cancer histotypes (HGSC, SBL, MC, and EC) in comparison to normal ovaries. Molecular data from histological sections of the prospective OC cohort (**Appendix C, Table 10**) were obtained by SpiderMass and show differences according to cancer type in both negative (**Figure 50A and B**) and positive ion mode (**Appendix C, Figure 71A and B**). Different machine learning algorithms were evaluated to train a classification model that could distinguish between them. First, multivariate analysis models based on PCA-LDA were constructed using AMX software in both ion modes using the 78 samples from the prospective cohort (**Figure 50C**). This resulted in correct classifications of 93.8% and 96.1% in negative and positive ion modes, respectively, after "20% out" validation (**Figure 50C and Appendix C, Figure 71C and D**).

To improve the classification, data extension was performed by combining the already fresh frozen analyzed tissue with 83 additional samples from a retrospective cohort of FFPE tissues (**Appendix C, Table 11**). The extended data set did not improve the accuracy of the classification. This may imply that the low accuracy is due to the use of the AMX software, which has limitations such as using only LDA classification, lack of cross-validation, and lack of explanation for predictions.

To address this limitation, 24 classification algorithms were tested with 5-fold cross-validation on the same FF dataset. The RidgeClassifier achieved the highest accuracy of 100% in training and 97% and 92% after 5-fold cross-validation in negative and positive ion mode on the FF cohort. The Ridge classifier (Dijkstra, 2014), suitable for high-dimensional datasets such as ours with 5000 *m/z* dimensions, is a linear classifier that uses L2 regularization to avoid overfitting and has shown strong performance on MS tissue data (Cortes et al., 2012). The FFPE tissue cohort was also used to build another model, resulting in 100% accuracy in training and 94% after 5-fold cross-validation. When FFPE and FF tissues were combined, the classification model yielded an excellent accuracy of 100% in training and 97% after 5-fold cross-validation (**Figure 50E and F**). The different models were then evaluated by blindly analyzing a validation cohort of 24 independent samples, further annotated by a pathologist (**Appendix C, Table 12**). For each of the 24 tissues, 3 different sites were analyzed for a total of 72 blinded analyses. Comparing the results obtained for the FF AMX, mixed AMX and Ridge Classifier mixed models, the latter appears to have the lowest error rate, with only 12 poorly predicted regions (**Appendix C, Table 13**), mostly endometrioid tissue. This can be explained by the lack of endometrioid tissue in the training cohort. When endometrioid tissue is removed, this model finishes with a correct prediction rate of 95.2%.

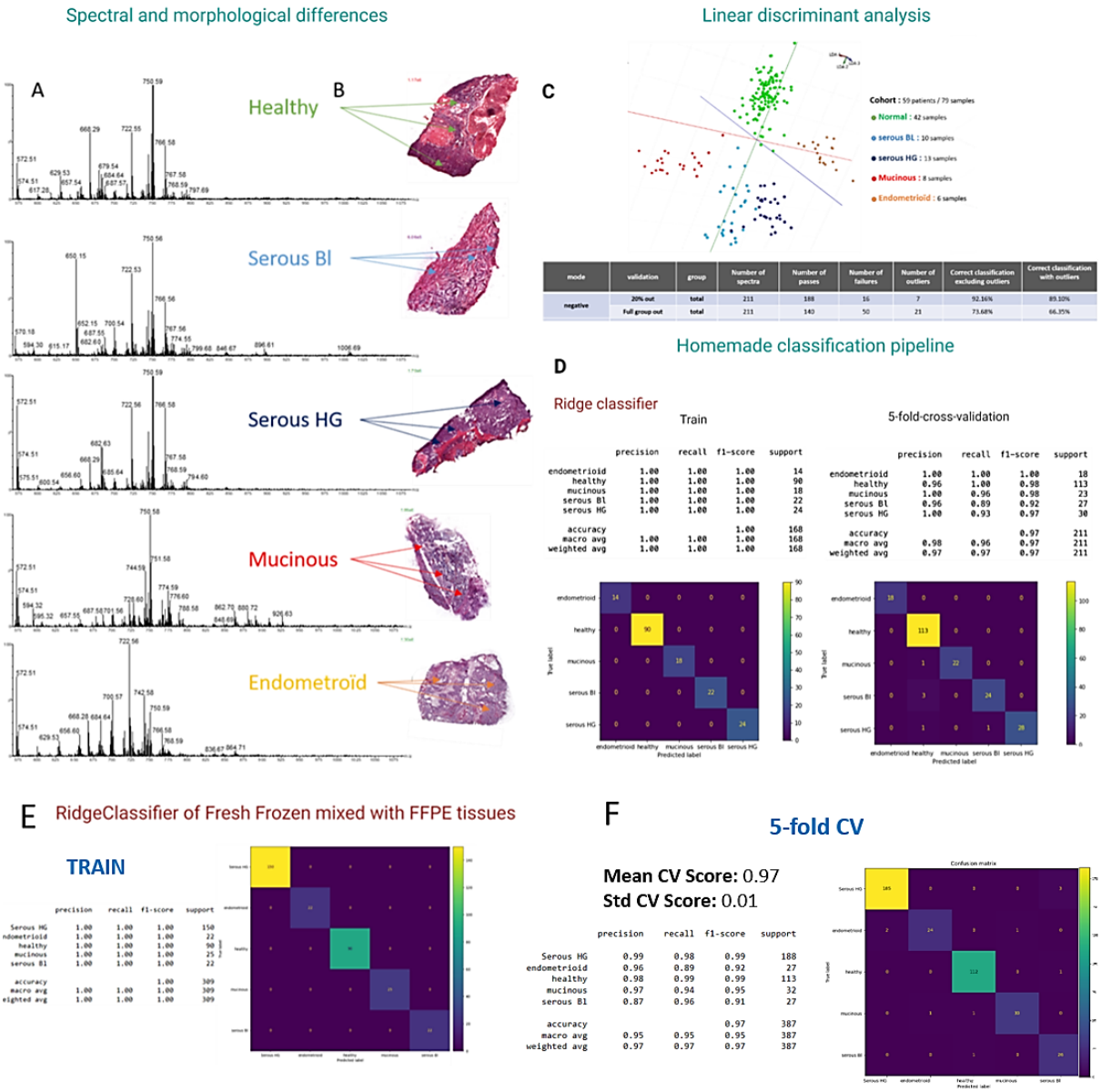


Figure 50: Multivariate statistical analysis-based models for ovarian cancer in negative ion mode. (A- B) Morphological and spectral fingerprint of different OC subtypes. (C-D) LDA and Ridge classification models based on the MS spectra from FF tissues. (E-F) Classification report and matrix confusion obtained with Ridge classification model on both FF and FFPE tissues for train and 5-fold cross-validation sets.

To further improve the performance of the model, the scans of stained tissue were integrated with the MS spectra to create a dual-input deep learning model. Indeed, a multi-input neural network model combining a multi-layer perceptron (MLP) (for molecular data) and a 2D convolutional neural network (2D-CNN) (for morphological data) was trained (**Figure 51A**). This approach proves to be highly robust, achieving 99% accuracy in negative ion mode through 5-fold cross-validation (**Figure 51B**). It achieves 100% accuracy in blind diagnosis when tested on 40 spectra images (**Figure 51C**). Notably, the model shows 100% sensitivity and specificity in both validation and test sets, surpassing the 88% accuracy of the sole molecular branch (Perceptron algorithm). The integration of histologic information with molecular data significantly improves the ability to

diagnose various OCs. This model is particularly compelling because it incorporates molecular information that is lacking in other models that focus on morphological features (Allaume et al., 2023).

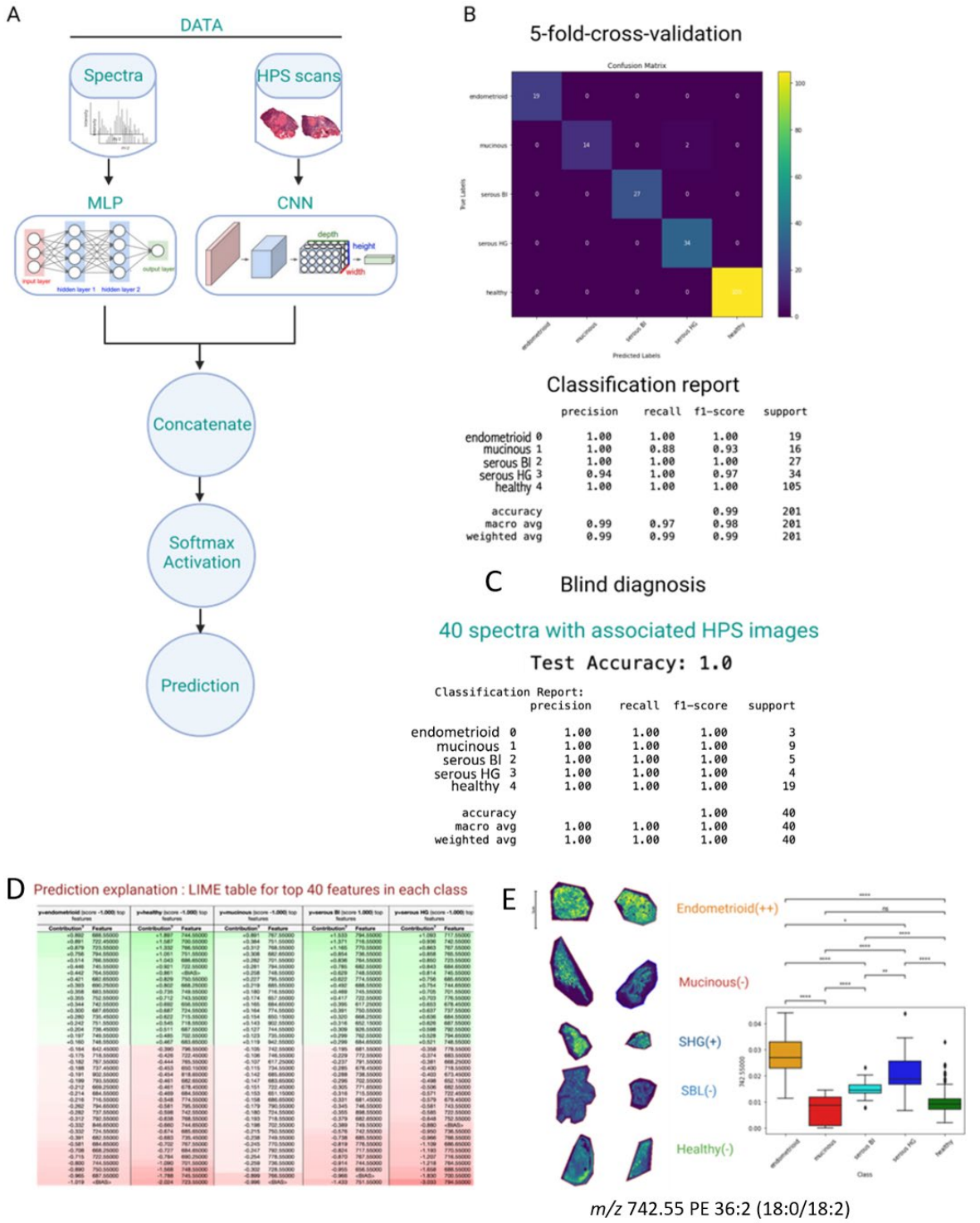


Figure 51: Multi-input model and the discovered lipid biomarkers. (A) Overview pipeline of the multi-input model. (B-C) Corresponding performance results after 5-fold cross-validation and for prediction on 40 blind image-spectra. (D) Top 40 *m/z* positive and negative contributions to differentiate each tissue type. (E) Example of a lipid marker, PE 18:2/18:0 (*m/z* 742.55), differentially expressed between the different OC histotypes.

Lipids Biomarkers Associated to the Different OC Subtypes

To provide a biological justification by finding lipid biomarkers, the LIME algorithm was used to determine the positive or negative contribution of each *m/z* feature to the classification of each cancer type (**Figure 51D**). A statistical test was used to validate the significance of each potential biomarker. Furthermore, each identified biomarker was cross-validated by MALDI-MSI to ensure its localization in the histological tissue region. In fact, ion *m/z* 742.5 (PE 18:0_18:2) was found to be significantly present in HGSC and EC, which was also confirmed by the high distribution in these two histotypes by MALDI-MSI (**Figure 51E**). Twenty-six lipids corresponding to potential biomarkers were identified in both ion modes (**Appendix C, Table 14 and Figure 69**). High relative abundance of PAs and PEs such as *m/z* 747.5 and *m/z* 718.5 were observed in HGSC. In positive ion mode, PC species including PC 36:1 and PC 30:4 are highly abundant in MC, whereas PS (as *m/z* 776.55) and PE (as *m/z* 748.55) were more representative of HGSC and SBL. In addition, several lipids segregated between HGSC and SBL, such as PI (22:1_18:0) and PE (18:0_22:4) in SBL versus PE (18:0_16:0), PA (18:1_20:2) in HGSC.

These results were in line with previous DESI-MS (Sans et al., 2017), MasspecPen (Sans et al., 2019) and MALDI-MSI (Meriaux et al., 2010) results. The lipid composition in different OC subtypes differs from normal tissues, with increased levels of PS, PA and PE, consistent with findings in other cancers (Koundouros & Poulogiannis, 2020). The emerging hypothesis is that elevated PS content affects the mechanical properties of PS/PC bilayers, affecting interfacial tension and contributing to cancer cell migration (Pöyry & Vattulainen, 2016). Dysregulation of PS in cancer is associated with surface exposure on tumor cells, leading to immunosuppression (Stoica et al., 2022). In addition, PS exposure shields cancer cells from NK activity and other cytotoxic immune cells (Lankry et al., 2013). Fatty acid remodeling of phospholipids is an adaptive response to the acidic tumor microenvironment (Urbanelli et al., 2020), supporting rapid cell proliferation and increased phospholipid metabolism. Changes in phospholipid levels can also affect cellular signaling pathways, influencing cell proliferation and survival and promoting tumorigenesis (Stoica et al., 2022). Notably, PA is more abundant in various cancer types compared to normal tissues and plays a role in the activation of kinases such as MAPK in stress signaling pathways (Putta et al., 2016). Overall, real-time molecular diagnostics reflect the dynamic changes in lipid metabolism in different cancer subtypes. The differential abundance of phospholipids is closely related to the progression of the tumor microenvironment (TME). Therefore, further exploration of the tumor microenvironment will be undertaken to improve understanding and diagnostic approaches.

The TME is known to play a key role in cancer progression, survival, and migration (Quail & Joyce, 2013). To discriminate the most abundant cell types within the TME, several cancer cell lines (SKOV3, POE4, PA1, SW626), a healthy ovarian cell line, dedifferentiated macrophage cell lines, sorted primary macrophages and lymphocytes were directly analyzed on cell plates/slides using SpiderMass. Based on their specific molecular signature, a PCA/LDA model (**Figure 52A and B**) was obtained with a correct classification rate of 97%. The few misclassifications were caused by the discrimination of the macrophage M1 and M2 subpopulations. By mixing the two phenotypes in the "macrophage" class, this model increased to 99.7%.

In addition, it was possible to discriminate the macrophage and lymphocyte phenotypes with 100% good classification (**Figure 52C and D**). Lipid markers were found to be highly abundant in macrophages, such as the ions *m/z* 818.65 and *m/z* 846.65, both identified as glucosylceramides (**Appendix C, Figure 70**). Indeed, glucosylceramides are known to be involved in inflammation as a regulator of the immune system and to reduce the inflammatory response during bacterial infection (Yeom et al., 2015). In addition, other lipid markers were found to be differentially expressed in macrophages according to their phenotypes. For M2-like, a significant increase in relative abundance was found for *m/z* 819.55 (PG 18:1_22:6) and *m/z* 867.55 (PI O-16:0_22:6) (**Figure 52E**). On the other hand, M1-like were highly discriminated by *m/z* 748.55 (PE 16:0_22:5), which was already abundant in normal ovarian tissue. Thus, a relative increase in polyunsaturated fatty acid PGs was observed for M2-like macrophages. Previously, Zhang et al. (C. Zhang et al., 2017). showed a lower concentration of smaller fatty acid chain PGs on M2-like macrophages in mice and human cell line. In addition, other studies demonstrated the colocalization of polyunsaturated PGs with inflammatory cells in cancer (King et al., 2023). For lymphocytes, several lipids were found to be highly expressed, such as *m/z* 768.55 (PE 38:3) and *m/z* 738.55 (PE 20:4_16:0). Some ions were also discriminative of lymphocyte phenotype such as *m/z* 794.55 (PE 18:0_22:4) for NK cells, in contrast to *m/z* 722.55 (PE P-16:0_20:4) for CD4 cells (**Appendix C, Figure 71**). The abundance of macrophage-specific signals (*m/z* 818.65 and *m/z* 819.55) is significantly higher in the HGSC subtype, in contrast to minimal expression in normal tissues (**Figure 52F**). Analysis of an endometrial cohort (79 patients) revealed resident macrophages in normal endometrium, with higher abundance in EC compared to normal ovarian tissue (**Appendix C, Figure 72**). While macrophages play a critical role in ovarian homeostasis (Z. Zhang et al., 2021), A total of 36 lipid markers were identified as being capable of discriminating between the various OC subtypes. The endometrium is more susceptible to inflammation. The cyclical changes in the endometrium involving proliferation and desquamation create a favorable environment for inflammation, which may be promoted by macrophages (Hogg et al., 2021) EC,

which is often derived from endometriosis, shows an increased macrophage presence compared to types such as HGSC, which is derived from the fallopian tube and is less exposed to inflammation.

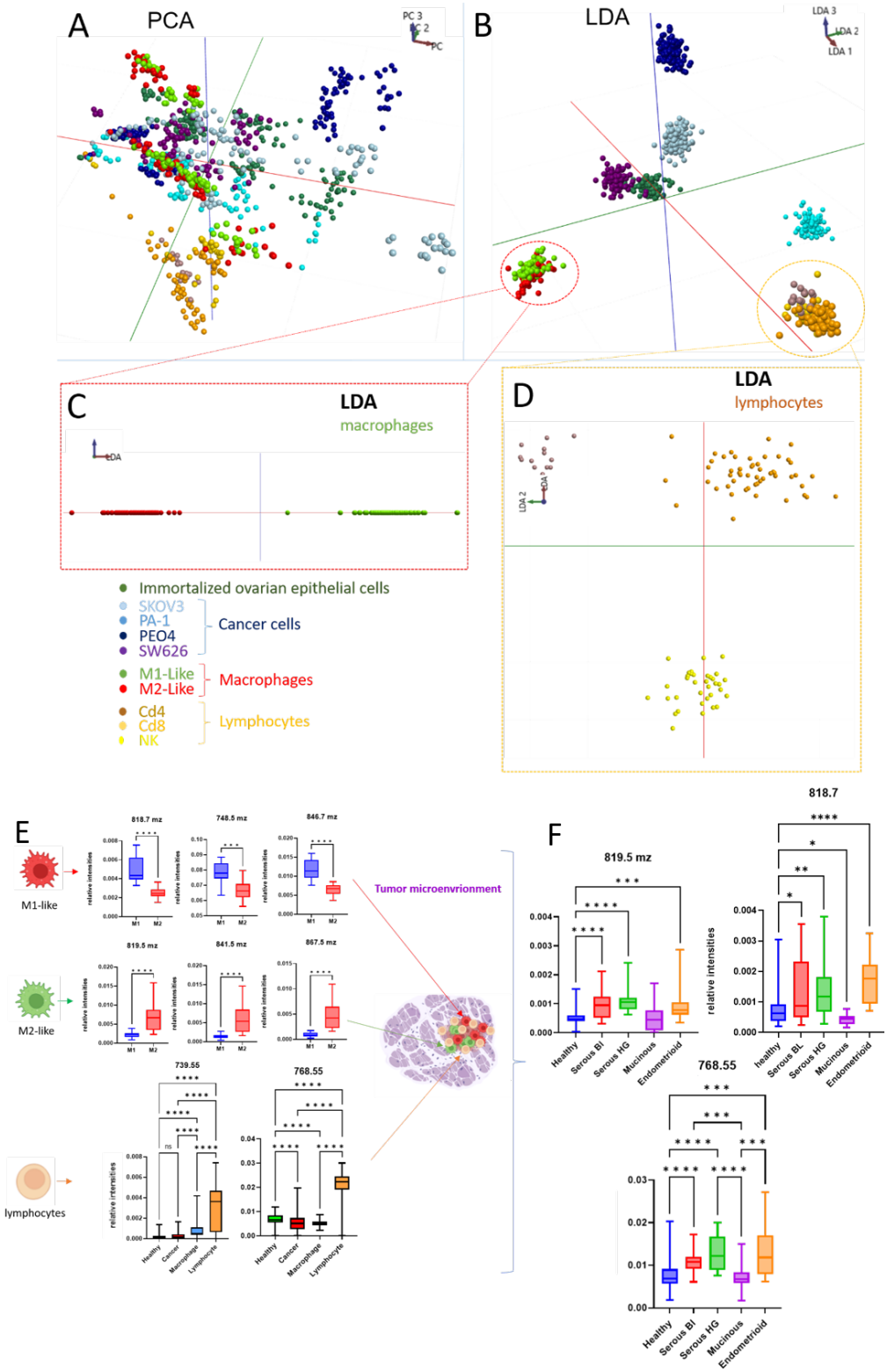


Figure 52: Classification of the immune cell types and their associated specific markers. (A-B) PCA and LDA models of ovarian cancer cell lines and immune cells. (C) LDA model of M1-like versus M2-like macrophages. (D) LDA model of lymphocytes (NK, CD4 and CD8). (E) Boxplots of 8 immune cells biomarkers. (F) Boxplots of 3 robust biomarkers within the OC cohort.

SpiderMass MS analysis provides a fast and direct way to obtain molecular information from immune cells. A method was developed to learn from the molecular fingerprint of different cell types and predict their probability of presence in tissues to analyze TME composition. This immunoscore model was trained using the Light Gradient Boosting Machine (LGBM) (Kanber et al., 2024) Python library by using cell line and immune cell spectra in the *m/z* 600-1100 range in negative ion mode, resulting in a successful classification rate of 100% (**Figure 53A and B**). This highlights different cellular signatures according to cancer subtype, such as MC, SBL and EC showing a higher presence of cancer cells, lymphocytes/macrophages and normal cells respectively (**Figure 53B**). It was also possible to obtain a probability of presence for each cell type in the pixels of each tissue, allowing tissues to be distinguished according to their immunoscore (**Figure 53C**). For this purpose, SpiderMass-MSI was performed on 3 samples of each OC subtype to visualize the different immune cell distributions. To validate the distributions based on LGBM predictions, multiplex MALDI immunohistochemistry (MALDI-IHC) (Lemaire et al., 2007; Yagnik et al., 2021) was performed using CD68, CD8, and Ki67 markers, which are specific targets of macrophages, lymphocytes, and cancer cell proliferation, respectively. The distribution of the cell population obtained by LGBM was confirmed by MALDI-IHC, although a higher spatial resolution (20 µm) was used for MALDI-IHC. As expected, a difference in the probability of the presence of macrophages and lymphocytes is observed according to the OC histotype. In fact, HGSC and EC show a higher abundance of macrophages. SBL and HGSC show a rather homogeneous spatial distribution of macrophages throughout the tissue, whereas EC and MC show a more heterogeneous distribution of macrophages with islands of highly intense signal (**Figure 54A and B**). Interestingly, SpiderMass MSI based on immunoscore allows the distinction of different immune cell subpopulations present in the TME without the need of any probe technique (IHC).

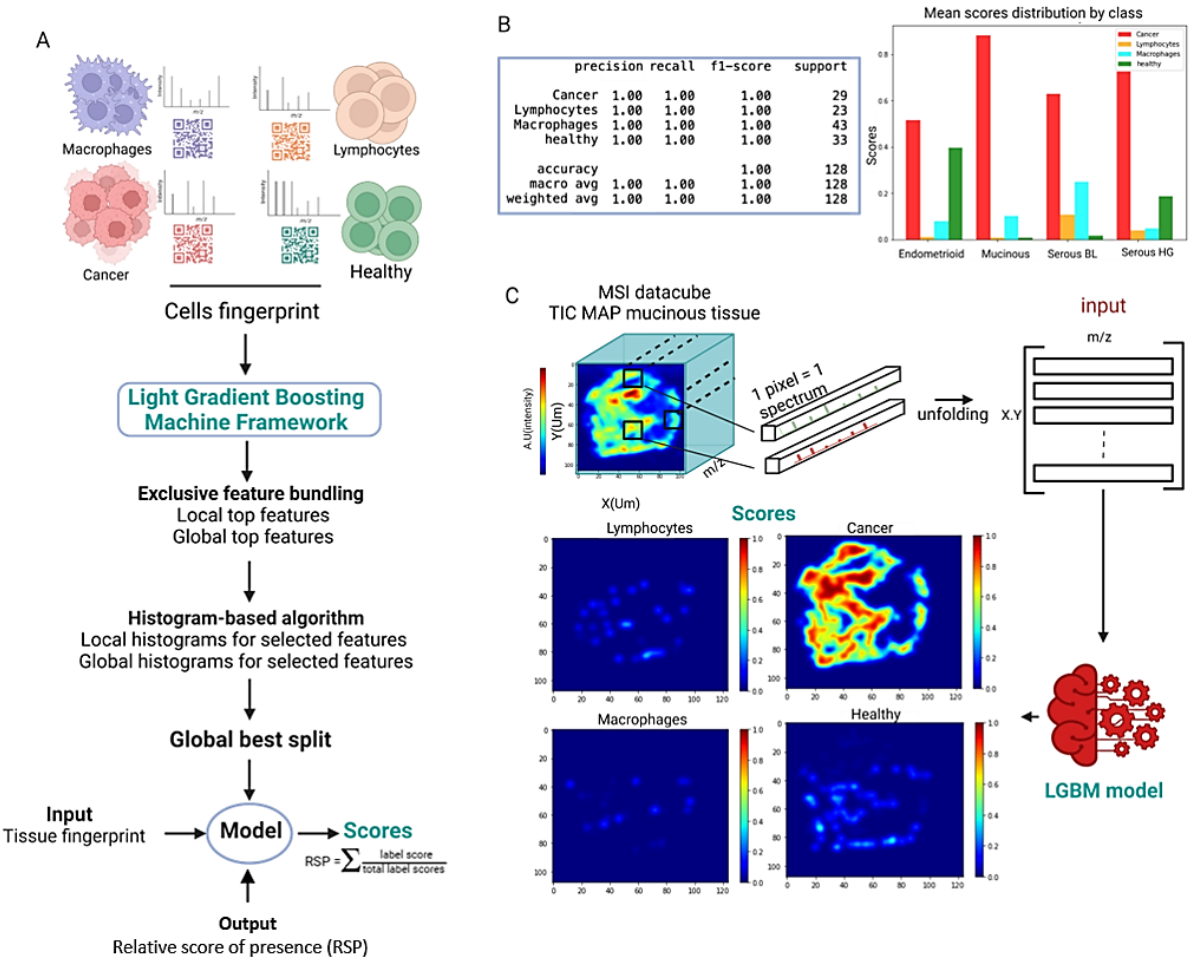


Figure 53: Workflow to create an immunoscore based on SpiderMass MSI. (A) Overview pipeline to train an LGBM model based on cells fingerprint. (B) Performances of the immunoscoring model and the mean scores distribution of each cell in OC subtypes. (C) The overall pipeline for seeing the distribution of cancer and immune cells in OC tissue analyzed by SpiderMass-MSI.

Since the immunoscore allowed to decipher the different OC histotypes, correlation with patient survival using HGSC tissues was instigated. The SpiderMass immunoscore was obtained from 8 different HGSC patients divided into 2 groups. Two patient groups were distinguished: one with no recurrence and OS > 50 months, and the other with recurrence and OS < 42 months (**Appendix C, Table 15**). To eliminate therapy-related bias, the analysis was restricted to surgical specimens obtained prior to chemotherapy. Tissues from patients with longer survival show a high abundance of both lymphocytes (increase of 56%) and M1-like macrophages (increase of 74%) compared to patients with shorter survival with a higher score of cancer cells (increase of 25%) and M2 macrophages (**Figure 54C, E and Appendix C, Table 16 and Table 17**). It's correlated with the expected results, knowing that M1- and M2-like macrophages are pro-inflammatory and pro-tumoral, respectively (Jayasingam et al., 2020; Mantovani et al., 2006). Moreover, the immunoscore of patients who received neo-adjuvant chemotherapy showed that the number of macrophages and lymphocytes detected was drastically reduced. The M1/M2 ratio shows that a better prognosis is associated with a high infiltration of pro-inflammatory macrophages, demonstrating the potential of

this ratio as a patient prognostic indicator that can help clinicians adjust treatment and surgery. This immunoscoring could be performed in real time to achieve precision surgery, as the SpiderMass is capable of in vivo analysis. This model can be further improved by increasing the number of immune cell subpopulations studied (such as monocytes, TAMs, Treg and TILs), but as presented represents the first immune cell-based approach to patient prognosis during surgery.

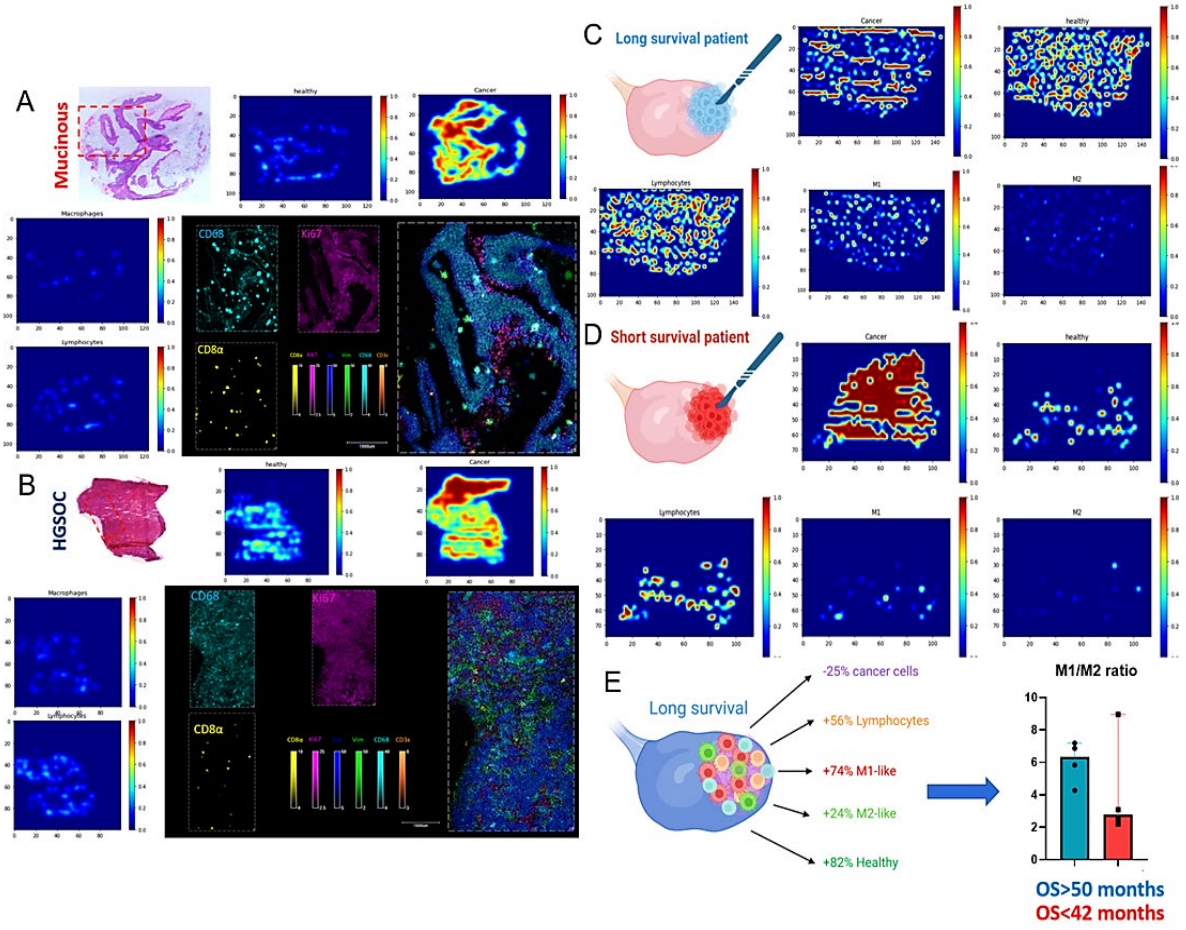


Figure 54: SpiderMass-based immunoscore for the diagnosis and prognosis of OC. (A-B) Predicted presence of cell populations using SpiderMas immunoscore model and MALDI-IHC for a mucinous and a HGSC carcinoma tissue respectively. (C-D) Immunoscore in a patient with long-term and short-term survival respectively. (E) Comparison of the M1/M2 ratios between patients with OS <42 months and those with OS>50 months.

Discussion

Accurate diagnostic and prognostic methods are considered essential for the management of ovarian cancer, particularly given the typically late detection of this disease. The acquisition of real-time data during surgery could significantly assist surgeons in customizing surgical approaches and managing patients. In this study, the potential of SpiderMass to achieve this goal was explored by integrating molecular data collection with advanced machine learning processing. The study was conducted on both prospective and retrospective cohorts of patients with ovarian cancer, and 100 percent correct classification was demonstrated after training, with 97 percent accuracy achieved

following cross-validation. The use of a dual-input model that combined both molecular and morphological information further enhanced the accuracy to 100 percent.

To ensure the biological validity of the findings, lipid markers that are differentially regulated across various subtypes of ovarian cancer were identified, specifically phosphatidylserine and phosphatidic acid. Phosphatidylserine, located in the inner membrane, is associated with immunosuppressive responses, while phosphatidic acid activates the mitogen-activated protein kinase pathway, which plays a critical role in stress signaling. To achieve a more comprehensive characterization of the tumor microenvironment a workflow was developed using the Light Gradient Boosting Machine algorithm to predict the relative abundance of different cell populations at each analytical pixel. This innovative approach enabled the differentiation of tissue regions containing cancer cells based on immune cell distribution, derived from SpiderMass mass spectrometry imaging data.

To validate these findings, MALDI IHC was employed, which proved to be crucial in confirming the accuracy of the SpiderMass immunoscore. The model was further tested for prognostic purposes in patients with high-grade serous carcinoma, and an increased infiltration of lymphocytes and M1 macrophages was observed in patients with longer overall survival. Notably, matrix-assisted laser desorption/ionization immunohistochemistry provided critical confirmation of these results, reinforcing its importance as a gold standard in the validation of immunoscoring. Given that some patients with ovarian cancer experience shorter survival times despite the absence of recurrence, the use of immunoscoring could open new possibilities for more personalized treatment strategies.

While the study has so far relied on post-operative tissue sections, the aim is to extend this approach to real-time data collection during surgery, enabling dynamic immunoscoring. The use of MALDI-IHC remain essential for validating these real-time assessments, ensuring both their reliability and clinical applicability.

Development of Molecular Digital Twins Based on Ambient Ionization Mass Spectrometry Imaging for Application in Cancer Surgery

Introduction

Cancer surgery most often remains the first pillar of therapy in oncology. Surgery quality is of utmost importance because of the huge impact it has on cancer recurrence and patient survival. Besides, the therapeutic management choice will very often depend on the outcome of the surgery and the post-surgical assessment of excised tissues. The aim of any cancer surgery is to remove the cancer with an adequate margin of normal tissue but with minimal morbidity. During the procedure,

surgeons cannot access in real-time to data that will help them discriminate normal tissues from tissues infiltrated with cancer cells. They thus must rely on their experience and training to make their decision. This lack of objective data lead, in application of the precautionary principle, to take wider surgical margins increasing morbidity and decreasing patient life quality. Conversely, surgical reintervention must have to be carried out for positive margins detected at the post-surgery diagnosis. Yet the gold standard procedure is based on the histopathological examination of the excised specimens. Intraoperative examination is possible and beneficial to improve the surgery, but due to the constraints of the process, it's limited to a few parts of the tissues not well recapitulating the specimen globality. For logistic reasons and lack of pathologists, it is also very difficult to implement for all surgeries. The final diagnosis is therefore only made during the post-treatment examination and may lead to the discovery of positive margins or, conversely, the removal of tissue that was unnecessary. Besides, only excised tissues are examined and the status of the bedside or adjacent tissues remain unknown. Further improving the surgical oncology procedures implies moving forward more personalized surgery. However, tailoring the surgery is tightly linked to the ability to collect robust molecular data already by the time point of the procedure. Collecting accurate molecular information will help the surgeon by offering in real-time the knowledge on the tissue status that is central for the margin delineation or to find the loco-regional extension of the cancer (e.g. status of the sentinel node). Additionally, it could also be used to get a real-time diagnosis and prognosis which according to the type of solid tumor could be taken into consideration to adapt the surgery (e.g. detection of an aggressive cancer subtype).

Digital Twin (DT) represents a paradigm shift for precision medicine (Björnsson et al., 2019) and paves the way in oncology to enter the era of precision cancer care (Hernandez-Boussard et al., 2021). If the concept of precision oncology is most often used for adapting the treatment of patients based on the administration of medicines, it is however an emerging concept in surgery. Introduced by (Shafto et al., 2010), the concept of DT initially found its roots in manufacturing and engineering, primarily serving purposes in product design, service management, and the real-time monitoring of industrial equipment (Jiang et al., 2021). Building on the successful applications within the industrial sector, the DT concept has now become one of the most promising advancements in healthcare with the creation of patients DT. When discussing digital twins in healthcare, the focus often revolves around the integration of connected devices like smart sensors to enhance data collection, like blood glucose levels, MRI, temperature, CT scans, cardiac electrophysiology or even clinical data (C. Wu et al., 2022). This integration facilitates simulations with real-world scenarios, thereby reducing medical risks and costs, and enhancing the quality of diagnosis, treatment, and disease prediction. DT of patients help in developing a more personalized and precise medicine by gathering real-time

monitoring and adaptation of treatments as well as improved management by predictions through machine learning. Cancer patients digital twins (CPDT) have proven to be of significant value in the field of oncology (Meijer et al., 2023). Their potential applications are therefore diverse, including monitoring, diagnosis, cancer surgery and development of therapeutics strategies tailored to individual patients (Croatti et al., 2020). Notably, there is a growing interest in exploring CPDT for the discovery of novel treatment targets and the prediction of drugs effects. Aggregating data from different modalities into a single representation provides indeed a more accurate overall view of patient characteristics, which can be used to predict response to treatment and improve care. On the other side, DT concept also find important application in surgery to better prepare the surgery, improve the post-surgical management using DT recorded during the surgery or using DT during the surgery as surgical (Rouhollahi et al., 2023; Servin et al., 2023; Shu et al., 2023). For example, the anatomical structure of the patient can be augmented by molecular data to help the surgeon. In cancer surgery, this involves the creation of virtual replicas of a patient's tumor and surrounding anatomical structures, using real-time data and advanced simulations. These digital counterparts enable oncologic surgeons to meticulously plan and execute procedures with an unprecedented level of precision, ultimately leading to enhanced patient outcomes (Hernandez-Boussard et al., 2021; Tardini et al., 2022). The fusion of this digital duplicate with cutting-edge technologies like cloud computing, artificial intelligence (AI) and machine learning will facilitate the comprehensive understanding, analysis, manipulation of data and informed decision-making. For example, Angulo et al., 2020 have introduced a digital twin for monitoring lung cancer behavior in patients, tailoring healthcare interventions based on the disease's specific impact on individuals. (Bagaria et al., 2020) have proposed a DT for monitoring heart rate and galvanic response to prevent heart diseases. For instance, Meraghni et al., 2021 have suggested a DT for breast cancer using temperature sensor information collected by portable intelligent devices. In health treatments, DT can simulate new experimental decisions within a virtual "real" environment to assess treatment outcomes (Bruynseels et al., 2018). Using a digital twin combined with AI offers the chance to create personalized recommendations tailored to the specific circumstances of the patient (Corral-Acero et al., 2020; Kaul et al., 2023). Clinicians can then utilize these recommendations to make decisions that are not only more accurate but also highly personalized and effective.

Over the past decade mass spectrometry has emerged as an interesting technology to assist surgery. MS can be performed in-situ and provide non-targeted molecular data. As MS separates molecules according to their molecular weight, the MS spectra provide a profile of the molecules detected from the tissues based on hundreds of different compounds with various intensities. These molecular profiles have been shown to be very specific of the cell phenotypes. Hence the presence of cancer

cells, or existence of different cancer subtypes driven by different pathophysiological mechanisms are traduced in the MS molecular profiles. Using Mass Spectrometry Imaging, these molecular profiles can be recorded in a systematic manner by scanning a region of interest to get the spatial distribution of the molecules within this ROI. MSI has demonstrated to be an essential tool in cancer research, revolutionizing our ability to explore the molecular intricacies of tumor tissues (Duhamel et al., 2022; Vaysse et al., 2017). By combining the precision of MS with spatial information, MSI allows researchers to create detailed maps of various biomolecules such as metabolites, lipids, and proteins within cancerous tissues (Buchberger et al., 2017; Neumann et al., 2020). This spatially resolved molecular data unveils the complex heterogeneity of tumors, shedding light on their metabolism, biomarker profiles, and microenvironment (Calligaris et al., 2013). Importantly, ambient ionization mass spectrometry (AIMS) based technologies have emerged to enable in-man MS analysis during surgery (Ifa & Eberlin, 2016; Saudemont et al., 2018; Tzafetas et al., 2020). Among AIMS technologies, only few were shown to be operable in vivo for surgical purposes (Ogrinc, Saudemont, et al., 2021b). Advantageously, SpiderMass technology (Ogrinc et al., 2019), which is promoting a micro-sampling of the tissue through resonant excitation of water molecules endogenous to the tissues, was demonstrated to be a micro-invasive, painless and contactless technology, for intraoperative analysis (Fatou et al., 2016; Ogrinc et al., 2019). Thanks to the limited damage to the tissues (a few µm depth) and the contactless analysis, SpiderMass stands out as the only MS solution permitting in man MSI (Ledoux et al., 2023; Ogrinc et al., 2021). MSI with the SpiderMass technology (Ogrinc et al., 2021) is obtained thanks to an integration of the laser probe of the system onto a 6-axis robotic arm which can be moved to scan tissues according to a defined raster. Interestingly, a distance sensor is also added into the head of the robotic arm to measure the sample topography as well. Thus, during a MSI acquisition, the system capture at the same time both the molecular data and the topography, providing access to the reconstruction of 3D molecular imaging maps.

Cancer surgery aims at removing the cancer but uniquely the diseased tissues while preserving normal ones. However, practitioners are often at a loss as they are not always able to determine the status of the tissue and make the best decision for the patient. Highly invasive surgery is associated with limited recurrence but significant surgical complications and high morbidity rates. Conversely, leaving cancerous tissues/cells after surgery is associated with much lower 5-year survival rates. Hence, the accurate delineation of the resection margins and the loco-regional extension of the disease are paramount to the quality of the surgery. The development of new technologies, such as AIMS, offers an interesting prospect for the surgeon in guiding his decisions. Nevertheless, due to its format and the large amount of information it contains, the MS data collected intraoperatively cannot be used directly and require an interpretation and have to be

represented in such a way that the end-users can obtain the information they need in the most readable format.

This study thus focuses on a new concept of molecular digital twins for solid tumor surgery with the generation of digital twins from non-targeted molecular data based on MSI collected thanks to the SpiderMass technology. To create these CPDT, we combine the molecular data from MSI interpreted through a machine learning based interpretation pipeline and their 3D virtual counterparts to offer a better delineation of the margins, and their infiltration by cancer cells, to the practician leading to a more tailored surgery. Classifying cells of homogeneous (or very closed) phenotypes has already been demonstrated for various intraoperative technologies including MS and can be achieved thanks to conventional machine learning algorithms (Balog et al., 2010; Eberlin et al., 2013; Gredell et al., 2019). However, interpreting mixed molecular profiles issued from mixed cell populations is above the capacity of simple classification and generally lead to false positive or negative results. More specifically, predicting the ratio of the different cells subpopulation in an image pixel require a more complex processing (Zirem et al., 2024). Hence, we are presenting the development of a robust and novel processing pipeline to generate these MS-based CPDT. With this one, we can predict the ratio of different cell population in each pixel of the molecular image, and for example the ratio of cancerous cells versus normal ones to get the margin delineation. Additionally, this pipeline allows also to predict the ratio of different bacterial strain in different region, and according to cell phenotypes, of the tissue. This can be applied to highlight the presence of specific bacterial populations according to cancer subtypes or in different areas of the tissues for diagnosis and prognosis purposes. To proof-of-concept these novel CPDT, we have worked with transgenic mice models, which spontaneously develop mammary tumors, as a mimic of triple negative breast cancer, and generated DT after scanning tumors from different mice. We are now investigating generating these scores in real-time and unrolling them on the topographic maps recorded during the MSI to reconstruct DT based on the MS Imaging scoring which can then be used by the physicians to personalize their surgery. We illuminate how molecular DT can reshape the landscape of cancer surgery, offering new avenues for enhanced positive patient outcomes and medical education.

Material and Method

Experimental Model and Subject Details

TgC(1)3 mice model

Twelve FVB-Tg(C3-1-Tag) cJeg/JegJ mice were obtained from Charles River Laboratories (L'Arbresle, France). They were housed in the Pasteur Institute animal facility (PLETHA) and bred under a protocol approved by the Animal Protocol Review Committees of the Pasteur Institute (Lille, France) following European regulation (#25871-20200522117321730). For genotyping, ear clips were

digested in lysis buffer (KAPA Biosystems). Samples were employed for PCR reactions to amplify T antigen cDNA using the primers TA1: 5'-GACCTGTGGCTGAGTTTGCTCA-3' and TA2: 5'-GCTTTATTTGTAACCATTATAAG-3'. Products of the amplification were analyzed by agarose gel electrophoresis. As previously described, female mice expressing T antigen developed multiple mammary tumors by 4–5 months of age42. One healthy female mouse was obtained from the Animal facility of University of Lille. They were euthanized by exposure to a rising concentration of CO2.

Bacterial strain

Several bacterial strains (Streptococcus infantis, Staphylococcus lugdunensis, Methylobacterium radiotolerans) were obtained from the American Type Culture Collection (ATCC).

TgC(1)3 Mice Model Analysis

All of the mice were used for post-mortem imaging right after being sacrificed. Indeed, the mice were attached to the polystyrene foam using needles. The skin was notched with scissors without opening the peritoneum. To guide the scissors, a Brodie fistula director grooved with probe end was introduced between the skin and the peritoneum. After opening the shin, it was attached to the polystyrene foam using needles, thus exposing the mammary glands for 3D imaging analysis.

SpiderMass MS Imaging

The overall layout of the instrument setup has already been covered elsewhere43. In addition, here, the laser system used was an Opolette 2940 laser with a reinforced jacketed fiber. To perform imaging analysis, the Spider-Mass microprobe was coupled to a commercially available stiff 6D-axis precision MECA robotic arm (MECADEMIC, Montreal, Canada) described in a previous work30. The spatial step size was set either to 500 or 250 µm by achieving oversampling. The mass-range was fixed between *m/z* 100-1500. The acquisition sequence was composed of 3 consecutive laser shots and 3 seconds between each step. The laser bursts and the spectrometer acquisition were automatically triggered through a MATLAB in-house user interface developed for the robotic WALDI-MSI30. The data was acquired in positive and negative, sensitivity ion mode on a Xevo (G2-S, Q-TOF, Waters, Manchester, UK) mass analyzer through the REIMS prototype interface with biocompatible 2 meters Tygon tubing.

MS/MS Analysis

SpiderMass technology facilitated the MS/MS investigation using the Xevo G2-S instrument from Waters. MS/MS spectra were recorded following the isolation of the precursor ion, which were then subjected to collision-induced dissociation (CID) in the transfer cell. The collision energy used for this ranged from 30 to 40 eV, depending on the specific precursor ion chosen. To identify lipids, a manual annotation process was employed, guided by fragmentation spectra guidelines, and the

results were compared against databases such as LipidMaps, Alex123, MetFrag database (Ruttkies et al., 2016), and relevant literature.

Bacterial Strain Analysis

The bacterial strain used in this study are Streptococcus infantis (ATCC 700779), Staphylococcus lugdunensis (ATCC 49576), Methylobacterium radiotolerans (ATCC 27329). S. infantis and S. lugdunensis were cultured on Tryptic soy (BD 236950) agar plates when M. radiotolerans was cultured on Nutrient (BD 213000) agar plates, at 37°C for 24 hours. The SpiderMass analysis was made directly in the Petri dish in which the bacteria were grown. Several colonies were analyzed to get a good representativeness. To enable a good match of the molecular profiles, the SpiderMass MS spectra of the grown bacteria were obtained from the same spot size as the tissues.

Data Processing and Analysis

Image processing

The raw data files generated by the Spidermass instrument in both ionization modes were initially converted into the imzml format19 using MSConvert (Proteowizard ToolKit) (Chambers et al., 2012). To conduct data analysis, all imzml data files were imported into Python using Pyimzl library and pyimzmachine learning .Imzml Parser module. Once the data was imported, several pre-processing steps were applied to the spectra. These preprocessing steps involved normalization of total ion count (TIC) and binning to a window of 0.1 *m/z*. All final data sets contain 5000 *m/z* data points with the corresponding surface using x and y pixels coordinates. Indeed, for every raw file, the relevant map file was retrieved to extract topographic details (z-values) from the 3D image in CSV format. Ultimately, a 3D map was created using the provided z-values, accompanied by an interactive surface plot generated using the Plotly library. The resulting 3D plot facilitated the visualization of MSI data, including the exploration of spatial distributions of ions, TIC map, segmentation map, prediction label map as well as bacterioscoring map.

Segmentation

Spatial segmentation of SpiderMass imaging data was performed individually using the *k*-means++ algorithm from the scikit-learn library. This comprehensive spatial segmentation enabled the identification of regions of interest. To estimate the right number of clusters, not subjectively, the Silhouette criterion was employed. This criterion served multiple purposes, aiding in finding the optimal number of clusters and assessing their stability and compactness. Spectra associated with clusters, representing healthy, tumoral, or peritumoral regions, were labeled and grouped in a CSV file, forming the basis for building the classification model in both ionization modes.

Classification model and blind prediction34

The Lazy Predict library (https://lazypredict.readthedocs.io/en/latest/) was used to build multiple models from the scikit-learn library, encompassing 24 classifiers including linear models, ensembles, gradient boosting and neuronal network. The LGBMClassifier (Light Gradient Boosting Machine) (Ke et al., 2017) emerged as the optimal model. Model assessment involved a 20% out, a 5-fold cross-validation and a 20-fold cross-validation using the KFold and cross_val_score functions. The classification report was generated, and the ConfusionMatrixDisplay function from the matplotlib library was used to visualized the confusion matrix. The optimal model was stored and blind predictions were executed using the joblib library's dump and load functions.

Margin delineation

The constructed models were employed to predict the class of each pixel in an image using the "predict" function, which returns the class label with an argmax greater than 0.5. Consequently, pixels belonging to the same class are assigned the same color. To delineate the margin between classes, such as between a tumor and its surrounding peri-tumor region, the probability scores returned by the predict_proba function are utilized. The LGBM model effectively handles the predict_proba function. For models that do not inherently support probabilities, the CalibratedClassifierCV method could be used to make them probabilistic. The predict_proba method return continuous values (or predicted probabilities) that represent the likelihood of each new input belonging to each class (normal vs cancer). This is based on the features (ions) more or less present in the new spectrum compared to the characteristic spectrum of each class (tumor and healthy).

Bacterioscore

The bacterioscoring model underwent training using either the SGD Python library in negative ion mode or Ridge Classifier (sigmoid method) in positive ion mode. Both model used cell spectra within the *m/z* range of 600 to 1100 in both ion modes. The spectra were categorized into distinct bacterial strains, including Streptococcus, Staphylococcus lugdunensis, and Methylobacterium radiotolerans, each consisting of 70 spectra. To predict the bacterial strain in SpiderMass images, the model's predict_proba function was employed, offering probability estimates for bacterial strains and facilitating a nuanced understanding of the likelihood of each strain's presence in the local environment. Moreover, ratio scores were calculated to estimate the relative presence of each bacterial strain across the entire image. These scores were determined by summing the scores for each strain and dividing it by the sum of the total scores across all labels. The ratios provided insights into the distribution of the trained bacterial strains throughout the image, offering a comprehensive assessment of the bacterial landscape in the analyzed sample.

Statistical tests

The discriminative *m/z* of the different segmentation's clusters were found using a non-parametric statistical test Kruskal-Wallis with Bonferroni correction, employing the stats.kruskal function from the scipy library. First, a peak picking algorithm, the find_peaks_cwt function, was applied to remove instrument noise using a signal/noise ratio > 10. The corresponding box plots were then generated using GraphPad software where the legend correspond to *p value ≤ 0.05, **p value ≤ 0.01, ***p value ≤ 0.001, ****p value ≤ 0.0001.

Quantification and Statistical Analysis

Three datasets were employed for training, cross-validation, and testing of all classification models, including those for tissue type, and bacterioscore. Evaluation involved a 20% validation split and 20-fold cross-validation, with a classification report providing metrics such as accuracy, recall, precision, and F1 score. To assess the statistical significance of biomarkers, a non-parametric Kruskal-Wallis test was employed. Bonferroni corrections were applied to adjust p-values for multiple comparisons. Values are presented as medians and visualized through box plots.

Results

SpiderMass MS Imaging and Molecular DT Creation Workflow

The CPDT based on MS Imaging data was proof-of-concept on transgenic TgC(1)3 mice showing spontaneous development of mammary gland tumors. The overall design and workflow of the generation of molecular DT is shown in **Figure 55**. The MS Imaging SpiderMass system used for the DT is presented in **Figure 55A and B** and corresponds to an upgrade from our first lab prototype (Ogrinc et al., 2021). Briefly, the handheld of a fibered OPO laser equipped with a 1m length optical fiber and emitting at 2.94 µm (to promote resonant excitation of water) is connected to the head of a robotic arm which also includes a distance sensor. A tygon tubing of 1.5m is also connected to the head of the robotic arm to aspire the aerosol generated during the desorption-ionization process upon laser irradiation and transport it back the to the MS instrument thanks to high vacuum inside the mass spectrometer. The MS instrument without its conventional ion source is equipped with an interface which process the aerosol and avoid extensive fouling of the instrument while helping with desolvating the aggregates contained within the aerosol. This enables MS spectra to be generated in real-time (ms scale). The robotic arm is programed to move above the sample surface, to perform point-to-point MSI acquisition while adjusting the focal distance. In the present version, we have improved the scanning system and data acquisition for simultaneously collecting both the topographical and the molecular data (**Figure 55C**). Additionally, a high-precision laser distance sensor with 5 µm accuracy in all dimensions was incorporated and the SpiderMass laser microprobe is placed at an angle of 30° to ensure the convergence of the two laser beams (distance sensor and MS laser) in the focal plane. The system is powered by a novel version of OPO midIR fibered laser

based on an Opolette 2940 mid-IR pulsed laser (Opotek, USA, CA) which is designed for a plug & play connection through high energy SMA connectors of a metal jacketed fiber. The injection in the optical fiber is made in a N2 flowed chamber to prevent the presence of dust and hance extensive burning of the fiber entrance. On the other side, at the fiber outlet, a handheld is mounted which includes a focusing CaF2 lens with a 4 cm focal distance. For each experiment the post-mortem mice were place underneath the sensor and microprobe as shown in **Figure 55B**. Using this SpiderMass MSI prototype, we have created a complete workflow ranging from the data acquisition to the DT creation (**Figure 55D-H**). The data are automatically acquired using a homemade acquisition interface programed under MATLAB in 20 to 30 min runs depending on the size (around one or more centimeters) of the region to be imaged (tumor and peritumoral area) (**Figure 55D**). Data can be collected either in the positive or negative ion mode, or both (one after each other) depending the cancer type and which mode provides the best molecular discrimination. These result in a topographical image and molecular data recorded in every pixel. Optical images of the samples are also collected prior to the acquisition. An example of collected optical and topographical images of tumors M15-T2 and M2-T1 are show in **Figure 55E**. The optical image shows two tumors from two different mice imaged with their corresponding topographies. Based on segmentation of molecular images plot in their corresponding topographical images, the molecular profiles of each class were used to train machine learning models (**Figure 55F-G**). **Figure 55F** shows an example of the segmentation obtained for the M3-T2 tumor, using Silhouette plot to find the number of clusters which best match the molecular data and in **Figure 55G** of a machine learning model, here shown in the for positive ion mode. The models can then be validated in blind from different samples and the molecular DT created by transforming the physical information into the virtual space using the topographical data. All the registered data are, therefore, combined and can be displayed according to various representations including Total Ion Current map, molecular segmentation map, margin delineation map or even bacterioscoring maps, as exemplified for tumor M9-T1 in **Figure 55H**. While the training of the machine learning models from the initial sample cohort and their validation can take sometimes (typically a week weeks depending the size of the cohort and its availability), the creation of DTs by querying the models in real-time during an application of the technology is fast, with instant feedback to the user.

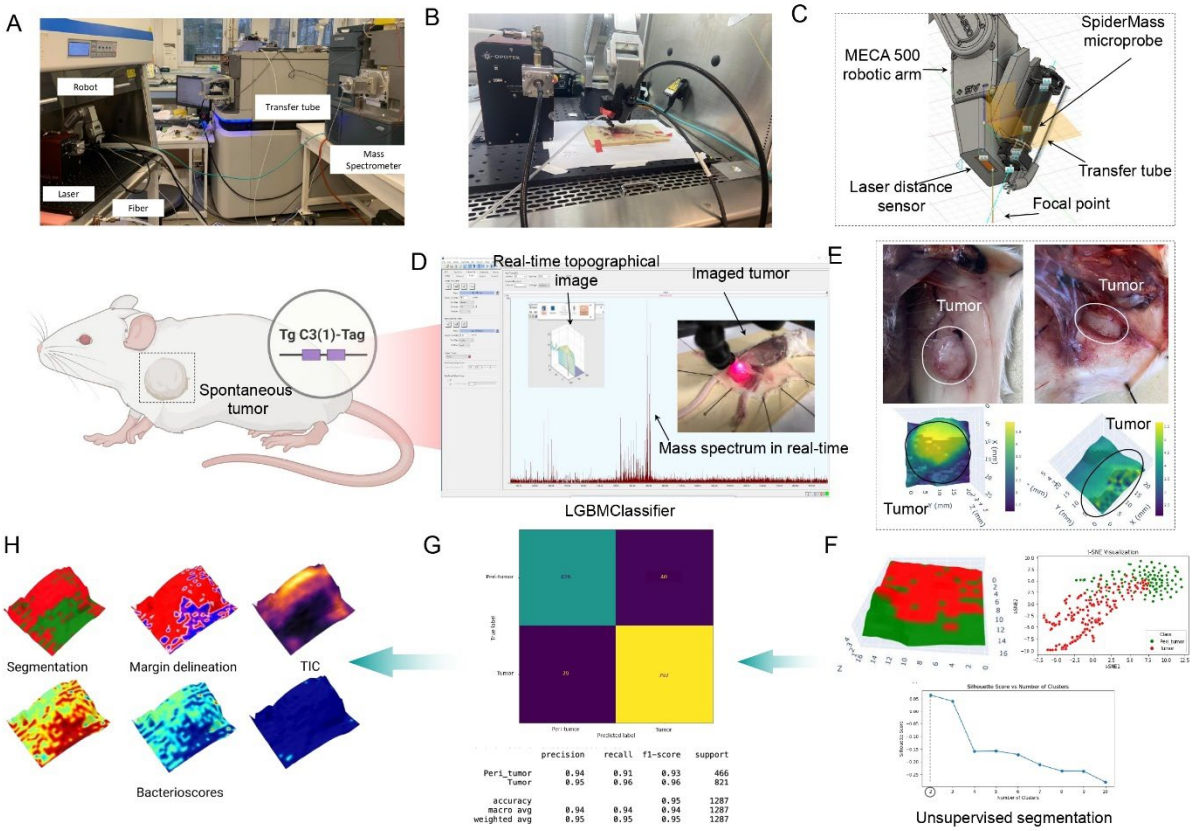


Figure 55: Workflow for the generation of the MS-based molecular digital twins. (A-B) Photo of the imaging setup including the Opolette 2940 laser with a reinforced jacketed fiber and an example of a mouse imaging experiment. The post-mortem mouse is exposed to reveal the tumor region and placed underneath the scanning system. (C) Schematic representation of the improved laser scanning system linked to the SpiderMass laser microprobe and transfer tubing on the robotic arm. (D) An example of 3D imaging acquisition. The image includes a real-time display of a real-time topography acquisition, mass spectrum and the photo of an imaged tumor. (E) Optical images of two mice with exposed tumor areas and corresponding topographical images obtained of the selected region. (F) Unsupervised segmentation to distinguish between tumoral and peritumoral areas. These clusters are then used to create the classification model. (G) Confusion matrix and classification report of the LGBMClassifier classification model built in positive ion mode from molecular profiles of tumoral and peri-tumoral regions. (H) Generation of different molecular digital twins based on various MS data.

SpiderMass MS Imaging Based DT Training

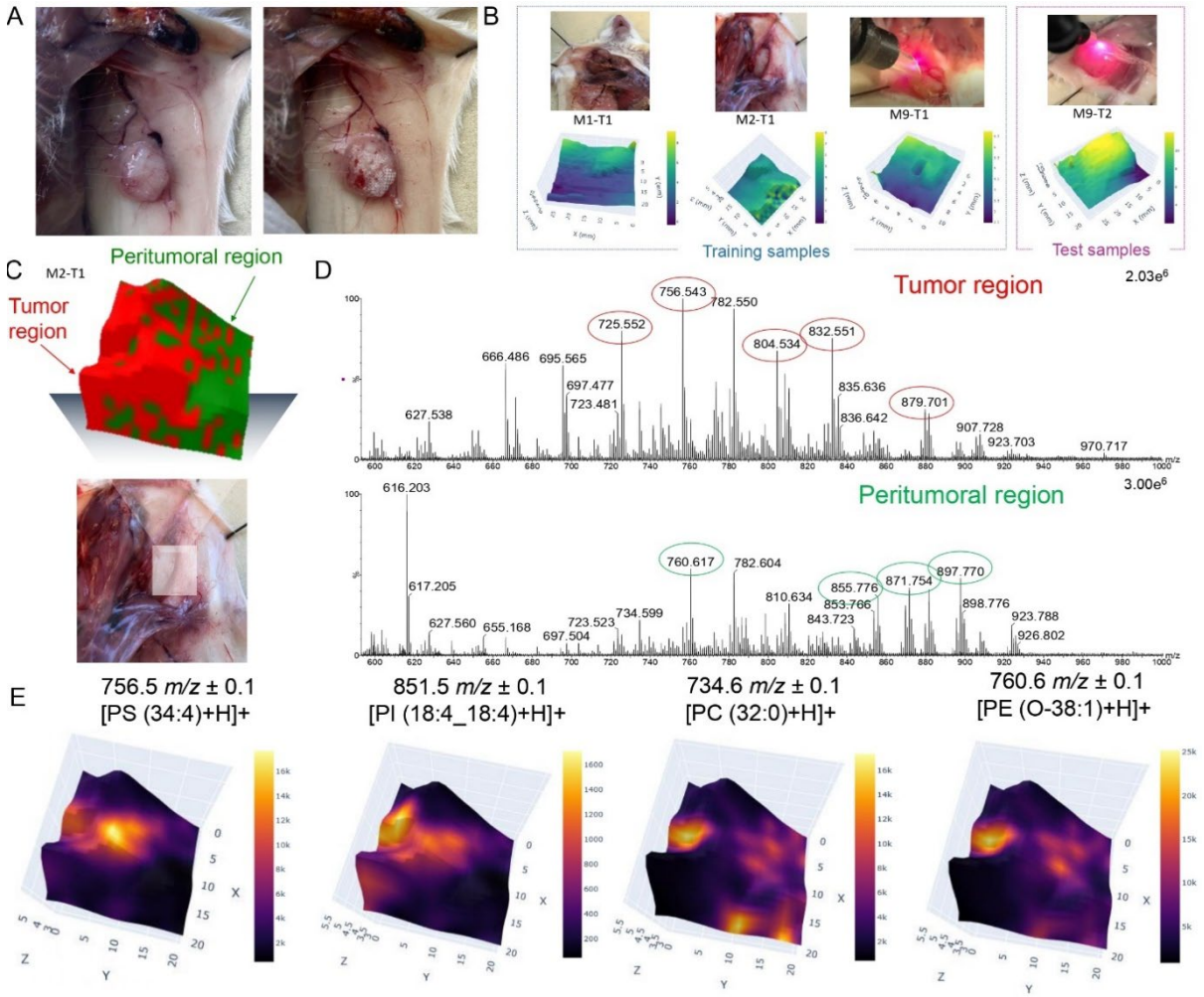
A total of 12 transgenic mice were imaged post-mortem with SpiderMass after dissection. MS offers the possibility to work in both positive and negative ion modes, and some family of molecules can be better detected in one or the other mode. Most of experiments were conducted in negative ion mode involving 15 mammary glands from 9 animals while only 4 tumors from 3 mice were used for the positive ion mode. In addition, one healthy mammary gland from a separate healthy mouse was also subjected to MSI with SpiderMass, in negative ion mode only, as a negative control. All data were collected at 500 µm spatial resolution. Photographs of one mouse tumor before and after an imaging experiment is shown in **Figure 56A**. Example of some optical images of mammary glands, with their corresponding topography images are presented **Figure 56B** and **Figure 57A** (framed in blue tumors used for training while in purple those for testing in blind). Three tumors were used for the training and one for the blind prediction in positive ion mode while in negative ion mode, fourteen were used for training purposes and two tumors were assigned to the validation

blind set. After mapping back the molecular data onto the 3D surfaces, image segmentation was performed by *k*-means ++ algorithm on all images and it highlights the presence of 2 clusters knowing that the choice of the optimal number of clusters was obtained using the Silhouette criterion. These 2 clusters correspond to 2 molecularly distinct regions which correspond to tumor (red) and peritumoral region (green) (**Figure 56C** and **Figure 57B**). The assessment of clusters into peritumoral or tumoral tissue was visually conducted, leveraging the tumor's topography, typically positioned higher contrasted with the peritumoral tissue located on lower sides. While this instance exemplified a straightforward class assignment, on intricate samples, multiple pathologists will assess each cluster prior creating the classification model. Furthermore, the distribution of the dataset for each tumor were depicted employing t-SNE dimension reduction. Two distinct point clouds are clearly observed, aligning with the presence of two clusters, one for the tumor and another for the peritumoral region (**Figure 57B**).

Molecular differences between the two clusters are well observed in the extracted average mass spectra from the tumor and peritumoral region as shown for the *m/z* 600-1000 range in positive ion mode (**Figure 56D**). In a visual manner, some signals seem to be exclusive to one of the clusters, as the *m/z* 756.5 [PS (34:4)+H]+ and *m/z* 832.5 [PS (40:8)+H]+ which seems only present in the tumor regions in contrary to ions *m/z* 760.6 [PE (O-38:1)+H]+ and *m/z* 871.7 which seems specific of the peritumoral regions. More interestingly, some ions were found to be zone-exclusive using a Kruskall-Wallis significance test in both ion mode. In fact, in positive ion mode, as the **Figure 56E** show, the selected ion 3D image *m/z* 756.5 ± 0.1 and *m/z* 851.5 ± 0.1 [PI (18:4_18:4)+H]+ overlap with the tumor region while the ion image of *m/z* 734.6 ± 0.1 [PC (32:0)+H]+ and *m/z* 760.6 ± 0.1 overlay with the peritumoral region. In addition, in negative ion mode, the selected ion image of the *m/z* 913.6 ± 0.1 [PI (40:4)-H]- and *m/z* 885.5 ± 0.1 [PI (20:4_18:0)-H]- overlap with the tumor region, whereas the ion images of *m/z* 762.5 ± 0.1 [PE (16:0_22:6)-H]- and *m/z* 747.5 ± 0.1 [PA (18:0_22:6)-H]- overlay with the peritumoral region (**Figure 57C**). Furthermore, this region specificity is confirmed by the corresponding boxplot representations (**Figure 57D**). To be noted that the healthy mammary gland is defined by the absence of the ions which are found to be very abundant in either the tumor or peritumoral regions rather than being discriminated by the presence of specific ions.

Two classification models, in each ion mode, were then constructed using all corresponding MS spectra acquired from the training samples. LGBMClassifier was found to be the optimal algorithm for both ion mode datasets, according to our new developed pipeline34. Remarkably, the classification models obtained 95% and 91% of correct classification rate after 20-fold cross-validation, in positive and negative ion mode respectively. Moreover, the sensitivity of these classification models, based on unsupervised segmentation of molecular data, are 91% and 95%

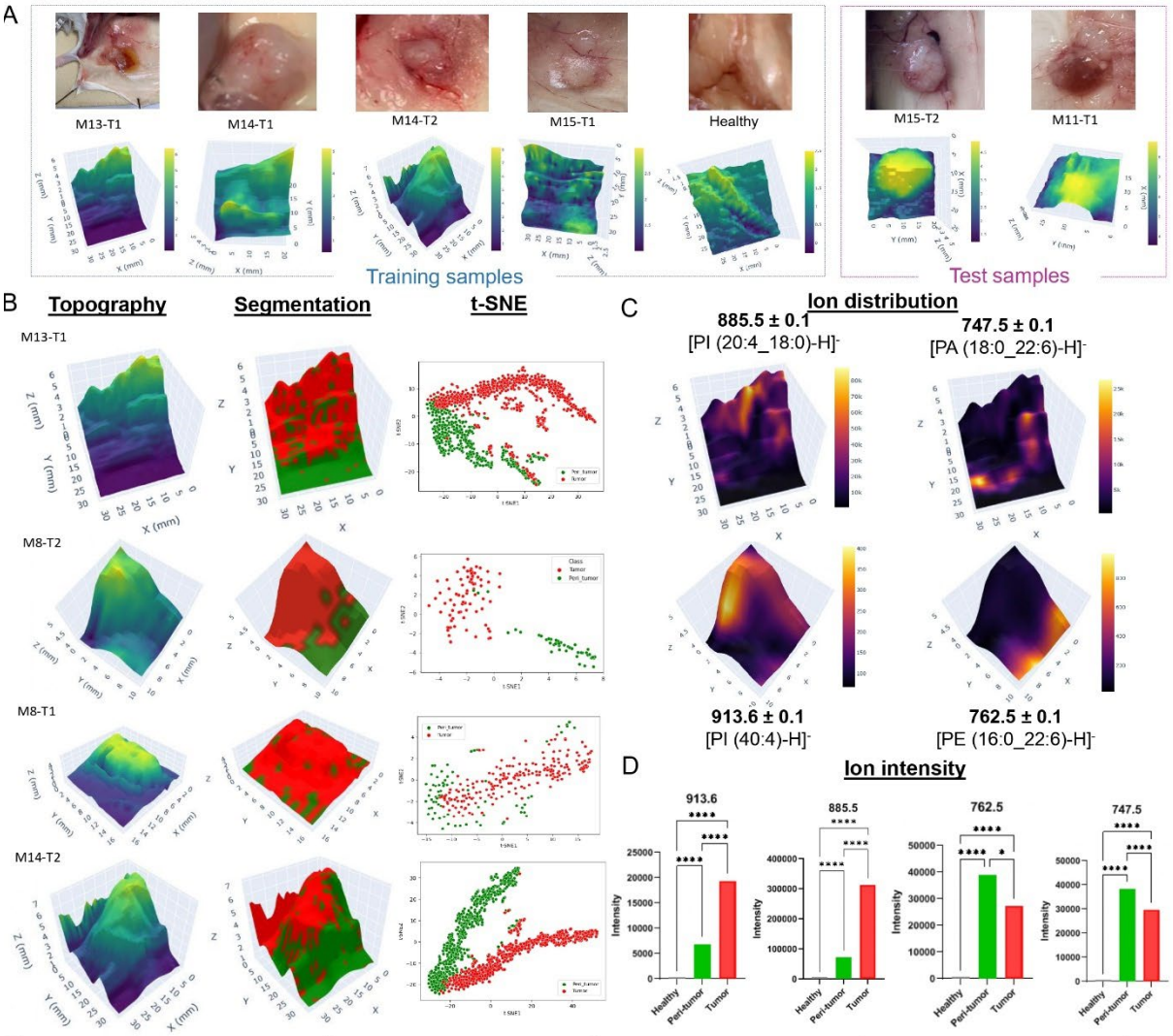
respectively in positive and negative ion mode while their specificities are 94% for both (**Figure 56F** and **Figure 57E**).



F

	Accuracy	Sensitivity	Specificity
Tumor regions	95%	96%	96.4%
Peritumoral regions	94%	91%	91.4%
Average	94.6%	95%	94%

Figure 56: SpiderMass MS Imaging of TgC(1)3 mice mammary tumors in positive MS ion mode. (A) Photograph of the mouse tumor before and after the MSI experiment. The experiment leaves white dots of dehydration indicating the imaged area (tumor and subsequent peritumoral). (B) Several tumors from different mice were used as a training and validation cohort in positive ion mode. The optical images of tumors with the corresponding topography highlighted in the blue box served as training samples. The optical images and corresponding topography highlighted in the magenta box served as the validation cohort. (C) Imaged region of the M2-T1 tumor with the corresponding *k*-means ++ segmentation. This one depicts 2 clusters corresponding to tumor (red) and peritumoral region (green). (D) Extracted mass spectra from the tumor and peritumoral regions at 600-1000 *m/z*. The distinct peaks for each area are circled in green or red, respectively. (E) Single ion 3D reconstruction, on M2-T1 tumor, for *m/z* 756.5 ± 0.1, *m/z* 851.5 ± 0.1, *m/z* 734.6 ± 0.1 and *m/z* 760.6 ± 0.1. (F) Table with accuracies, sensitivities and specificities for tumor regions, peritumoral regions and in average after 20-fold cross-validation ending to 94.6% correct class prediction.



E

	Accuracy	Sensitivity	Specificity
Tumor regions	90%	94%	86%
Peritumoral regions	88%	81%	95%
Healthy regions	98%	97%	100%
Average	91%	91%	94%

Figure 57: SpiderMass-MSI analysis on TgC(1)3 mice tumor regions in negative ion mode. (A) Several tumors from different mice were used as a training and validation cohort in negative ion mode. The optical images of tumors and a healthy mammary gland with the corresponding topography highlighted in the blue box served as training samples. The optical images and corresponding topography highlighted in the magenta box served as the validation cohort. (B) The 3D topographical image, the *k*-means ++ segmentation and the t-SNE visualization of 4 tumor examples (M13-T1, M8-T1, M8-T2 and M14-T2). Each individual segmentation reveals 2 distinct clusters mainly tumor (red) and peritumoral region (green). (C) The 3D selected ion images *m/z* 913.6 ± 0.1, *m/z* 885.5 ± 0.1, *m/z* 762.5 ± 0.1 and *m/z* 747.5 ± 0.1 on M13-T1 and M8-T2 respectively. (D) Corresponding boxplot representations of specific *m/z* values for each cluster. *p ≤ 0.05, **p ≤ 0.01, ***p ≤ 0.001, ****p ≤ 0.0001. (E) Table with accuracies, sensitivities and specificities for tumor regions, peritumoral regions, healthy regions and in average after 20-fold cross-validation ending to a 91% correct class prediction.

SpiderMass Based 3D DT from Blind Prediction

The built classification models were tested on additional tissues to the training cohort for the identification of tumoral versus peritumoral versus healthy regions in blind. The prediction scores resulting from the blind interrogation were plotted back onto the topographic maps to generate 3D prediction DT (**Figure 58**). In the positive MS mode (**Figure 58A**), blind prediction shows that the

imaged mammary gland, M9-T1, is composed by 75.6% of tumor and 24.4% of peritumoral areas. These blind predictions from supervised machine learning approach were compared to unsupervised *k*-means ++ segmentation. The unsupervised approach gives 37.4% for peritumoral and 62.6 % for tumor area which is different from the supervised approach. The prediction similarity between supervised and non-supervised is then 64.2%. Similarly, in the negative MS ion mode (**Figure 58B**), two tumors were submitted to blind prediction. No normal tissues, as expected, are found in both mammary gland mice tumors. The first tumor, M11-T1, blind prediction reveals 71.9% tumoral and 27.9% peritumoral regions. A comparison with segmentation demonstrated a highly promising 90.03% similarity between the result for supervised versus unsupervised approaches. For the second tumor, M15-T2, the prediction indicated 78.6% tumoral and 21.3% peritumoral regions but 63.8% and 36.2% for the segmentation. This tumor was thus exhibiting only 58% similarity between supervised and non-supervised. This highlights that molecular DT based on the prediction of cell populations can be built either using supervised or non-supervised processing. The non-supervised approach is attractive because it does not require training, but on the other hand it doesn't give any clue about the class correspondence. This demonstrates that employing molecular DT has a significant potential for oncological surgery, through a novel approach using real-time ambient mass spectrometry imaging.

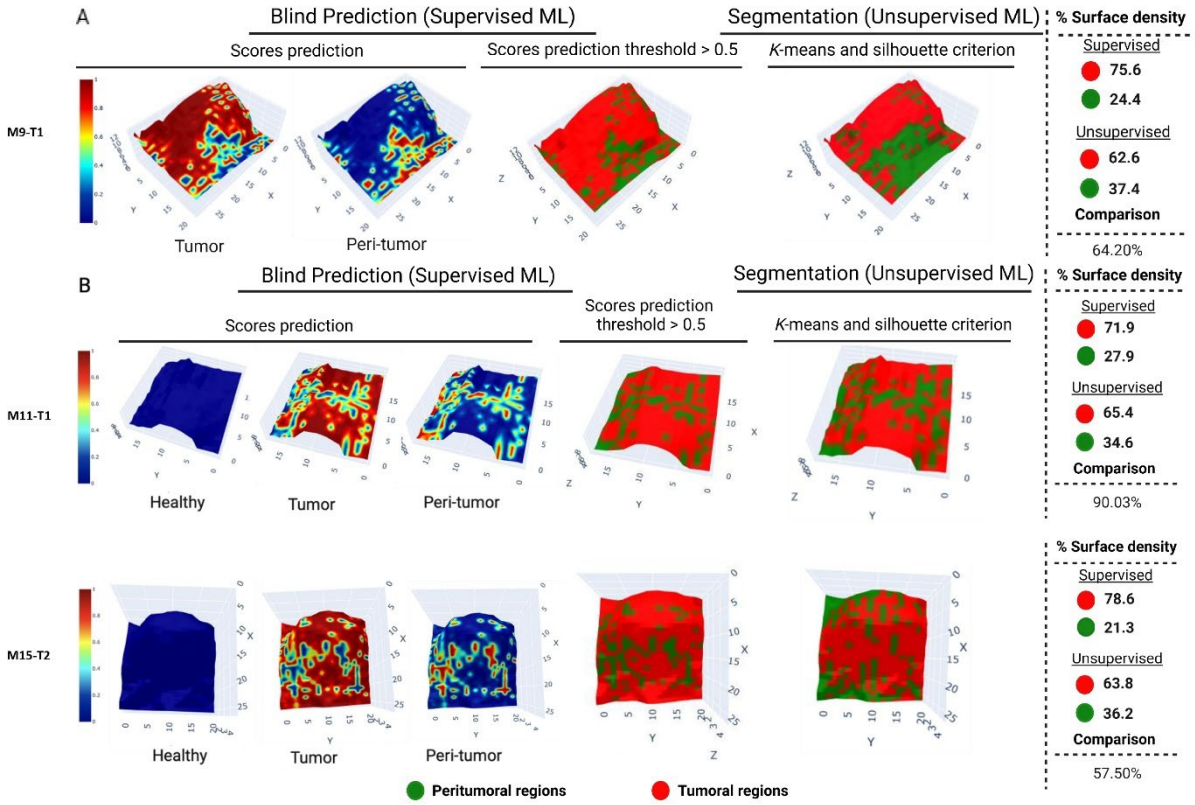


Figure 58: 3D reconstruction of SpiderMass blind prediction. The prediction scores for tumor, peri-tumor, and healthy regions are reconstructed in 3D and the 3D map is also obtained with scores exceeding a threshold of 0.5, all this achieved through supervised machine learning. The corresponding segmentation obtained by unsupervised machine learning is also

displayed. Furthermore, the surface density for each area is calculated and compared between the supervised and unsupervised approaches. (A) positive ion mode and (B) negative ion mode.

Creating Molecular DT for Margin Delineation

Accurate margin delineation is the main objective of surgeon while removing the cancer tissues. Because of the delicate balance between the removal of cancerous tissue and the preservation of healthy tissue, the composition of normal and cancerous cells in the tissue must be precisely determined. Yet, SpiderMass technology analyzes at a spatial resolution of 500 µm which appears to be the best compromise between excision accuracy and time. However, with such a spatial resolution each analytical spot contains about 1000-2000 different cells which can be of different nature. This means that non-supervised segmentation will not be able to cope with predicting the cell population ratios within in a pixel of the SpiderMass image. To address this aspect, we have predicted the ratio of the different cell populations using the LGBMClassifier to construct the classification models. For margin delineation purposes, we only searched for the prediction of 2 cell populations, according to our model, which are cancer versus peritumoral cells. We exemplified the margin delineation DT by plotting these predicted ratios with different colors for 4 to 5 different ratio levels on the topographic maps, as illustrated in **Figure 59**. The four levels depict an intermediate region with tumor / peritumor ratios between 0.33 and 0.66, unlike the second model that portrays two in-between regions, each encompassing between 25% and 50% of either cancer or peritumoral cells. Each level ratio includes a range between 0 and 0.1, designated to emphasize regions deemed truly healthy, free from tumor infiltration. This processing was tested on a sample (M8-T2) from the training cohort (**Figure 59A**), showing the accuracy gained, compared to unsupervised segmentation only, for accurately finding the boundaries between peritumoral, tumoral and mixed cell area. More interestingly, the pipeline was also then applied on data issued from the three tumors employed for blind prediction (**Figure 59B**) in both positive and negative ion mode. In these three tissues (M9-T1, M11-T1 and M15-T2), locating the boundary between tumor and peritumoral tissue proves to be a delicate task without adequate molecular information. The DT based on predicted cell ratios showed that non-cancerous cells are infiltrating the tumor tissue. Only the M8-T2 and M9-T1 tumors contains areas with less than 10% tumor cells. This highlights the potential of DT based on predicted cell ratios to enhance precision enabling surgeons to make informed decisions regarding the optimal size for tumor resection, and ultimately enhancing patient prognosis and post-operative quality of life.

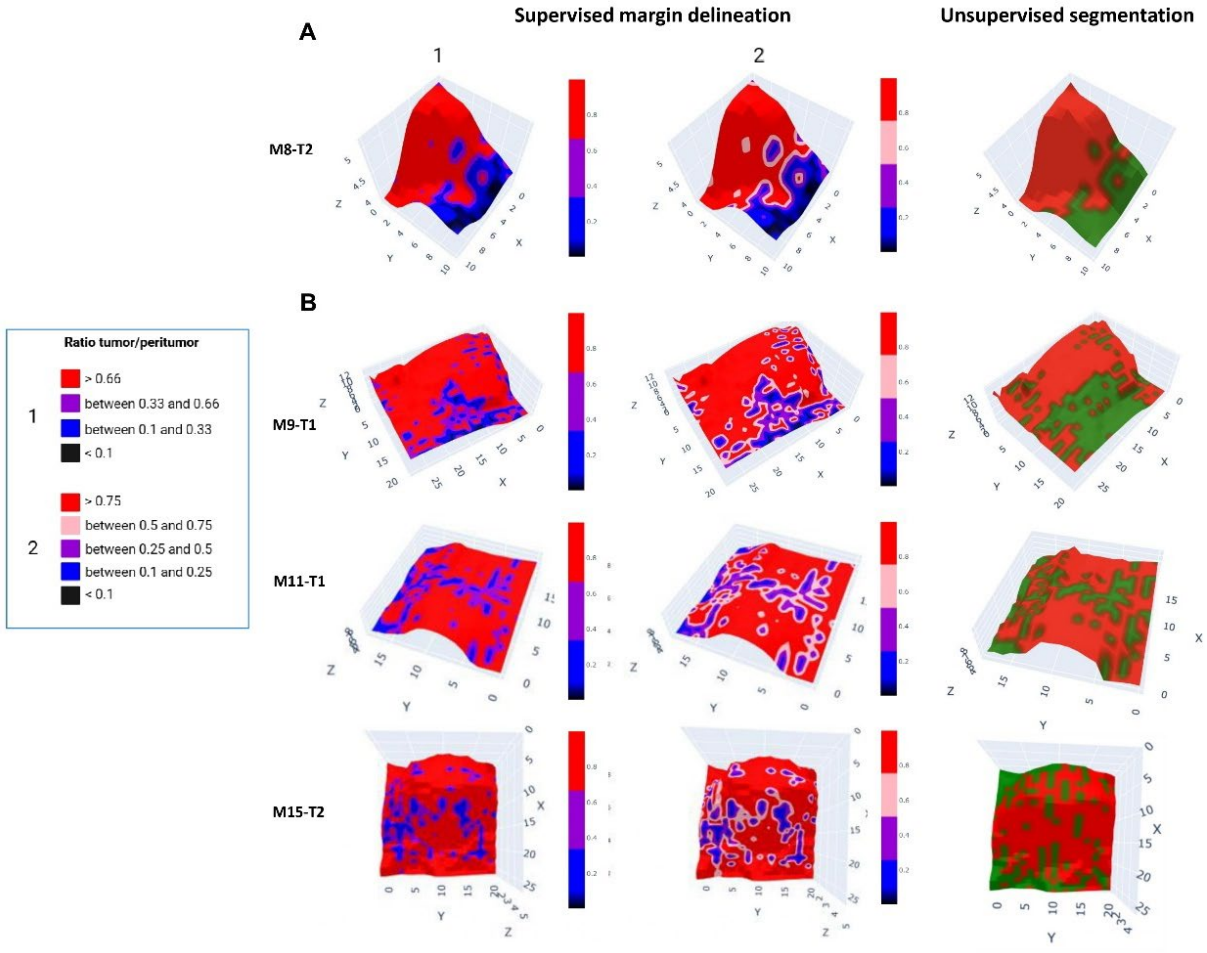
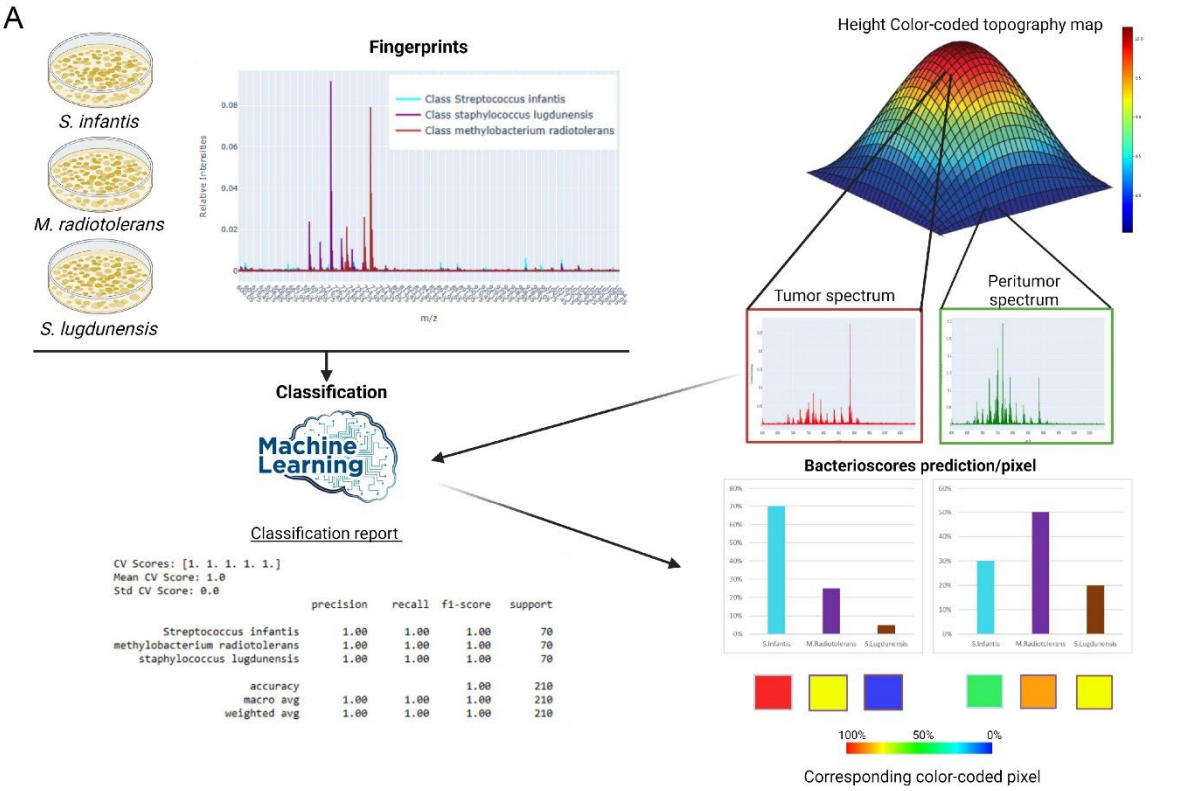


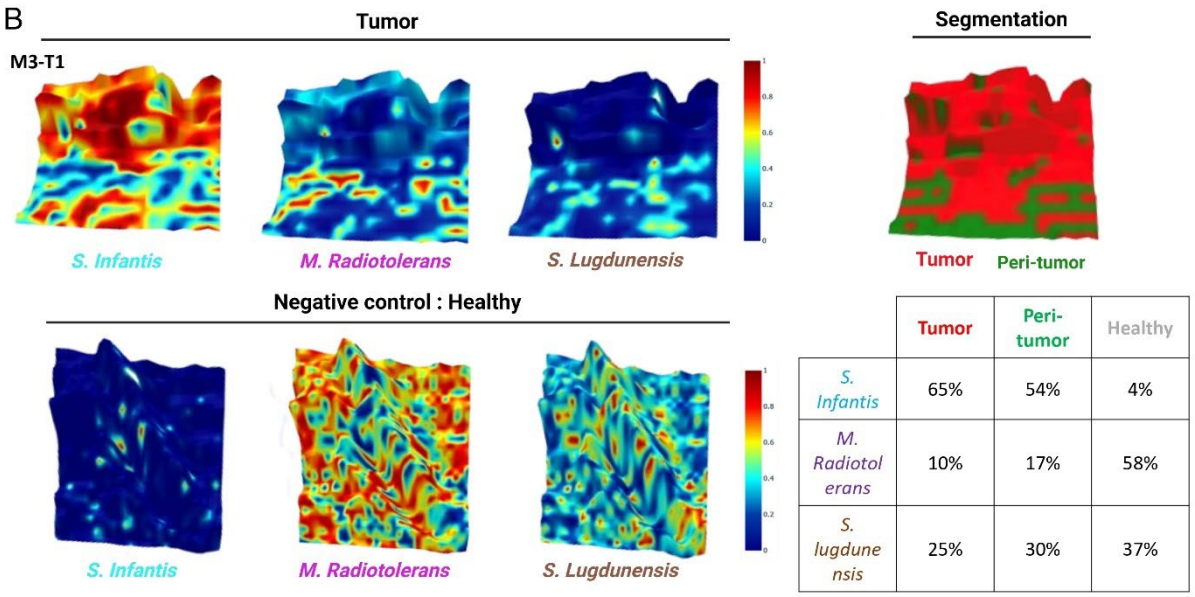
Figure 59: Margin delineation. The boundaries of tumor and peritumoral regions were delineated based on a ratio between tumor and peritumoral for (A) a mice tumor (M8-T2) used in the training of the classification model and for (B) three mice tumors (M9-T1, M11-T2 and M15-T2), not used in the classification model training, all using supervised blind prediction. For the color bar, the two options involve either a margin based on four ratio levels or on five ratio levels. Either illustrating an intermediate zone for ratios ranging from 0.33 to 0.66, or showing two intermediary regions, each covering between 25% and 50% of either peritumor or tumor areas.

Bacterioscore-based DT

There are, over the past 5-years, growing evidence of the importance of the microbiota in cancer. Different bacterial populations are indeed found according to the cancer type and subtype (Nejman et al., 2020). It was also recently demonstrated that there is causal link between the presence of bacteria and the tumor behavior. Tissue microenvironment microbiota was shown to alter the biology of cell compartments and thus the immunity response as well as the migration of the cancer cells (Galeano Niño et al., 2022; Long et al., 2023). Microbiotic niches represent an interesting new avenue for therapy and for understanding treatment resistance. Therefore, we have been interested to evaluate the possible creation of DT based on the specific detection of bacterial strains within the tissues of the mouse mammary gland to provide a comprehensive assessment of the bacterial landscape using SpiderMass-MSI and develop a bacterioscore-based DT (**Figure 60**). To create the bacteriocore, we first analyzed by SpiderMass directly on agar plates three different bacterial strains (S. infantis, S. lugdunensis and M. radiotolerans) which were previously shown to be

present in different breast cancers. Using machine learning algorithms, we trained a model in both positive and negative MS ion mode using 70 MS spectra per bacterial strain, we achieved a correct classification rate of 100% in both training and cross-validation sets (**Figure 60A**). The training was then unrolled on the topographic molecular images of the tumor mammary glands. For each pixel of the image, it is possible to predict the ratio of the different bacteria within the pixel (Zirem et al., 2024). The obtained results furnish estimated scores for each bacterial strain, and ratio scores are calculated to determine the relative score presence of each bacterial type across the entire image and plot this result on the topographic map to create the bacterioscore-based DT. This innovative approach allowed the reconstruction of a 3D image for all tumors, providing the probability of presence of each bacterial strain expressed in each pixel. Indeed, more the pixel is red more the bacterium is present in the pixel (**Figure 60A**). Interestingly, the expression levels of bacteria were found to be associated with specific regions of the tissue and consistent results were obtained in both MS ion modes. All the predicted percentage in the different peritumoral, tumor and healthy regions for the different mammary glands imaged by SpiderMass can be found in **Appendix C, Table 18**. Specifically, the S. infantis strain exhibited a higher predicted abundance in tumor tissues (80%) compared to peritumoral (66%) and healthy tissues (4.3%) which is in line with previous results obtained by sequencing (Nejman et al., 2020). In contrast, S. lugdunensis and M. radiotolerans demonstrated higher expression in healthy tissues than in tumor tissues (**Figure 60B**). There is a significant contrast in the expression levels of M. radiotolerans, as it shows a relative presence of 58% in a normal mammary gland, whereas its presence decreases to 10% in tissues adjacent to and within tumors. This correlates with finding from a study indicating that the bacterial strain M. radiotolerans is more commonly present in hormone-dependent (estrogen-positive) rather than in triple-negative breast cancer tissue, the transgenic mice models used in our study mimicking triple-negative subtype (Alpuim Costa et al., 2021).





	Tumor	Peri-tumor	Healthy
S. Infantis	65%	54%	4%
M. Radiotolerans	10%	17%	58%
S. lugdunensis	25%	30%	37%

Figure 60: Automated bacterial strain recognition on 3D tumor. (A) Methodology employed to obtain the different bacterioscores prediction/pixel. (B) 3D reconstruction of bacterioscores for 3 bacterial strains in one mice tumor (M3-T1) and in one healthy mammary gland.

Discussion

DT research is highly promising for customized patient care, with the potential to significantly impact the future landscape of healthcare innovation. A DT corresponds to a virtual embodiment of real-word objects or systems. In industry, DT were shown to be able to predict the behavior and response of a real-world object. In the case of cancers, evolution and response are linked to a

complex interplay of different factors including the patient's biology (i.e. genetics metabolism, etc.), lifestyle and external exposure factors (e.g. environment, pollution, etc.). Ideally, the creation of DT should consider all of these parameters to provide accurate prediction of the patient's evolution with cancer and their response to treatment. It is extremely difficult though to obtain all this data on patients including their lifestyle and the environmental pressures to which they were subjected over the decades preceding the development of the cancer. However, these external factors influence the human biology and are reflected in the signaling pathways and levels of different biomolecules and xenobiotics which are expressed in patients' tissues and fluids. The impact of external factors on the internal biology of patients demonstrates the value of using molecular data (derived from omics) to predict the evolution of the organism in relation to pathophysiological mechanisms. The development of CPDTs based on omics data is therefore of particular interest in cancer management by accurate forecasting. In the field of precision medicine, this is generally foreseen for adapting patient treatment without including the initial surgery within the therapy pipeline. Nonetheless, in solid tumors, the quality of the surgery is known to greatly impact patient prognosis and accurate diagnosis and prognosis by the time of surgery, opens the door a better tailored surgery as well as improved management. Thus, we have been investigating the development of CPDT based on lipidomic data issued from MS as potential tool for advanced diagnosis and prognosis which could be presented to guide surgeons in optimizing their procedure and patient management. Thanks to a recently developed ambient-ionization MS device the SpiderMass, lipidomic data can be monitored in real-time in-man. These data are obtained through the coupling to a robotic arm, in the imaging mode by scanning the laser probe of the SpiderMass above the ROI with the arm. From these molecular imaging data, we can create CPDT thanks to machine learning and present them as 3D map which are interpreted for the clinician. In the present study we have demonstrated the creation of these CPDT for surgery by analyzing TgC(1)3 mice which spontaneously develop mammary gland tumors and mimic triple negative BC. The mammary glands with cancer present two main regions which are the tumor area versus the peritumoral tissues. We show that a clear discrimination between the tumoral and peritumoral region is possible thanks to segmentation and involve ions which are different between the 2 regions and distinct from those found in the mammary gland of a normal mouse which is not a transgenic model. The silhouette scores and tests confirm that only 2 clusters are expected from the image data as awaited for these model tissues. Similar results are obtained in both positive and negative MS ion mode; discriminative lipids being different in the 2 modes though. Thus, these data can be used to train a classification model which enable to classify the tumoral and the peritumoral area versus the normal tissues, reaching up to respectively 95% and 91% in positive and negative MS ion modes. Interestingly, the model reaches a sensitivity of 95% and a specificity of 94% in positive ion mode while 91% sensitivity and 94% specificity are obtained in

negative ion mode. These models are also validated in blind with class recognition and >90% similarity is obtained between the blind prediction and the non-supervised segmentation. Ex vivo part of the study can take several weeks to be done (to obtain the optimal classification model), yet querying the model in real time in vivo requires only a few milliseconds. The deployment of the prediction map in 3D leads to an easy-to-read representation that could be project in the surgery room for surgery guidance. However, affecting each image pixel to a single class is not sufficiently accurate for margin delineation which is one of the main objectives of the cancer surgeon. Indeed, each pixel which corresponds to a laser spot size of 500 µm, which appears to be the best compromise between excision accuracy and time for the end-user, includes 1000-2000 cells which can differ in their nature and phenotypes. To address this problem, we have developed an approach to predict the cell ratio of each pixel, here applied to 2 cell type situations with the tumor versus peritumoral cells. Thanks to this approach, we can offer CPDT presenting the area with different cell ratios highlighting area with only tumor cells versus area with only peritumoral cells as well as intermediate margin areas with different ratios of cell populations. The DT demonstrates that the peritumoral regions still contain cancer cells and that this tissue, despite presenting a different molecular profile and being segmented separately, cannot be consider as normal and should better be removed during the surgical process. It demonstrates that these CPDTs based on the ratio of cells map in 3D provide the surgeons with the right information to define the excision margins. Since images during surgery need to be acquire in a limited time (5-10 min maximum) instead of acquiring high spatial resolution (e.g. single cell) molecular images which would take far too much time (several hours), we have opted for a different approach where we keep lower spatial resolution (a few hundred micrometers) but we predict the ratio of the different cell populations within these pixels which contains a few thousands of cells including cancer and normal cells as well as immune cell infiltrated populations.

To validate this approach, we have completed our study by some experiments on ovarian cancer tissues performed in 2D from which we have cross-validated the results from the prediction based on the SpiderMass MSI by a targeted approach using MALDI-MSI IHC (**Figure 61**). With the MALDI-IHC we can image in multiplex the distribution of several specific markers which target the different cell populations (cancer cells, conjunctive tissues and immune cells) and validate the predicted distribution for these cells from the SpiderMass molecular MSI data.

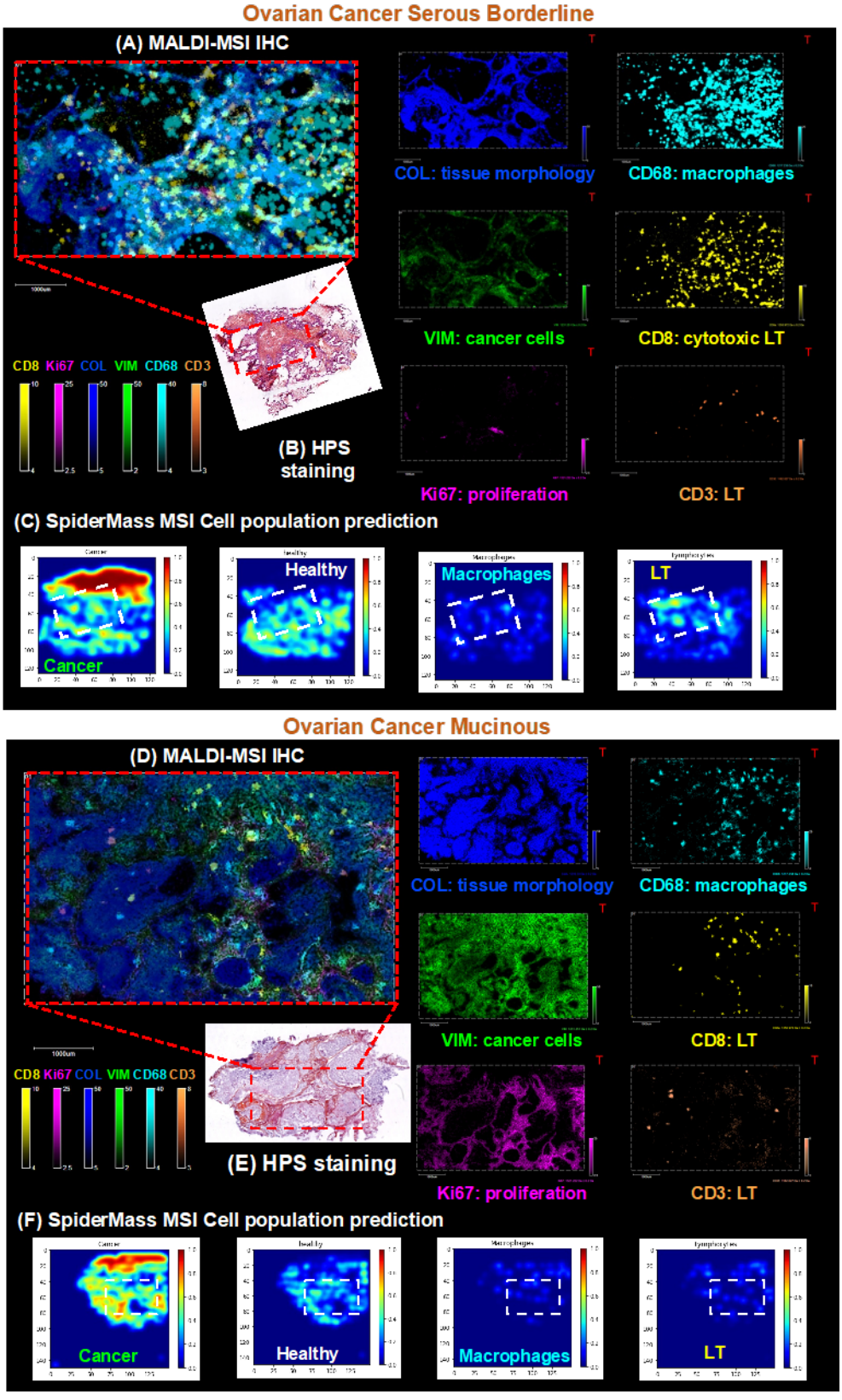


Figure 61: Cross-validation of the different cell population predictions from the SpiderMass MSI data by MALDI-MSI IHC against markers of normal, cancer and immune cells. MALDI MSI-IHC (A, C) in 6-plex against CD8 (Lymphocytes T cytotoxic), Ki67 (proliferation), collagen (tissue morphology), vimentin (cancer cells), CD68 (macrophages), CD3 (Lymphocyte T); HPS staining (B, E) and different cell populations prediction (cancer, normal cells, macrophages, LT) based on SpiderMass MSI data (C, F). Related to Figure 5.

Lastly, there are increased evidence of the role of bacteria in the onset of BC and its evolution. We wanted to assess the sensitivity of the SpiderMass technology to assess the presence and distribution of bacterial population in cancer tissues. By analyzing grown population of bacteria known to be present in BC, we were able to predict their presence in the molecular profiles recorded from the mammary gland tissues of transgenic and normal mice. By scoring their presence in each pixel, it was possible to create CPDT based on 3D maps bacterioscoring. Interestingly, these maps reveal that the S. infantis bacterial strain, which is known from previous work to be one of the most abundant bacteria found in BC (Nejman et al., 2020), is highly present in the cancer region with a limited presence in peritumoral tissues and almost completely absent from mammary glands of normal non-transgenic mice. Conversely, M. radiotolerans and S. lugdunensis are found in peritumoral regions with limited presence even for S. lugdunensis, while there are highly abundant in normal mammary glands. These 2 last bacteria were also described in cancer in human tissues; however here we used mice for the study which might explain the difference of the bacteria expression. In addition, other study indicating that M. radiotolerans was more prevalent in estrogen-positive BC tissue compared to triple-negative BC tissue (Alpuim Costa et al., 2021). Certain bacteria are shown to be associated with BC subtypes and immune cell niches in the tissue and could relate to the aggressivity of the cancer. In this context, knowing the cancer aggressivity would be an important information for the surgeon to consider tailoring the surgery within mind the future evolution of the patient. We have made yet the demonstration from transgenic mice model and we need in the next step to look forward to translate these developments to in-man application at the surgery room. For these, we will need to take further steps to enable the DT resulting from the model interpretation in real-time. Additionally, we need to address the speed of image acquisition to increase from the current 2.6 pixel/s screened. This will be made possible by continuous rather than spot to spot acquisition. Lastly, the bacterioscoring is extremely interesting and promising though we need to dig more into the microbiota of human BC by screening for more bacterial strains to better determine the location of different bacteria population and better understand their role in the cancer onset and progression. Finally, we demonstrate the interest of developing CPDT based on MS imaging data processed by machine learning and unrolled as 3D maps. Several CPDT representations will be accessible to the surgeon who will only have to pick the one that best match the needs (margin delineation, diagnosis, prognosis,…) according to the cancer type, subtype and the nature of the surgery. In the future, we are looking forward to decreasing the acquisition time using continuous acquisition mode to enable molecular topographic MSI data to be recorded in no more than 5 min followed by instant visualization of the DT. Hence, we do believe that these MSI-based CPDT will be very useful in the future for cancer surgery guidance giving an accurate hand to the surgeon and promoting the evolution towards precision surgery.

Conclusion and Perspectives for the Annexed Publications

These projects underscored the potential of MALDI-IHC in cancer diagnostics and treatment. This emerging technology confirmed the presence of critical biomarkers within cancer tissues, which enabled the development of more accurate and personalized diagnostic approaches.

During the two presented project, MALDI-IHC allowed to provide invaluable insights into immune cell infiltration. This capability links immune profiles directly to cancer progression and patient prognosis, thus enhancing our understanding of how the immune system interacts with tumors. By illustrating these complex interactions within the tumor microenvironment, MALDI-IHC not only identifies biomarkers but also reveals potential therapeutic targets. This understanding is crucial for developing therapies that can enhance the body's immune response against cancer, ultimately leading to better patient outcomes.

Furthermore, MALDI-IHC played a pivotal role in validating innovative concepts that integrated SpiderMass technology and machine learning. These advancements enabled real-time classification of ovarian cancer subtypes and facilitated the visualization of immune cells. Such capabilities are particularly valuable during surgical procedures, as they provide surgeons with immediate, actionable data to inform their decisions, thereby improving surgical precision and overall outcomes.

The development of Molecular Digital Twins and Cancer Patient Digital Twins also significantly benefited from MALDI-IHC experiments. These digital models offered precise tumor margin mapping and provided detailed insights into tumor composition, further enhancing surgical precision. The ability to visualize the intricate details of tumor architecture in real-time empowers surgeons to make more informed decisions, ultimately leading to more effective and personalized surgical interventions.

Moreover, the integration of MALDI-IHC with advanced imaging techniques and machine learning approaches enhances the overall capability of cancer diagnostics. This combination allows for comprehensive immune profiling, offering a deeper understanding of tumor biology that can directly inform treatment strategies. By validating biomarker presence and immune cell interactions, MALDI-IHC supports the advancement of personalized therapy approaches, which are essential in tackling the complexities of cancer treatment.

In this chapter, we developed the integration of SpiderMass technology and machine learning with MALDI-IHC in the first study. This combination enabled the development of sophisticated models capable of accurately classifying ovarian cancer subtypes based on real-time tissue analysis. In addition to providing clear visualizations of immune cells within tissue sections,

these models held tremendous potential for guiding surgeons during cancer surgeries. The ability to analyze and visualize molecular and cellular data in real time offered surgeons invaluable insights during operations, allowing them to make more precise decisions about tumor excision and improving surgical outcomes.

Another aspect of one project was the development of Molecular Digital Twins using mass spectrometry data. This concept allowed the creation of virtual tumor representations that could predict cancer behavior and offer highly accurate tumor margin mapping. During surgeries, these digital twins served as guides, helping surgeons distinguish between cancerous and healthy tissue with greater precision. By enhancing the accuracy of tumor removal, digital twin technology helped minimize the risk of leaving behind cancerous cells, reducing recurrence rates and improving patient survival. Cancer Patient Digital Twins concept was also introduced, which used 3D maps to offer critical insights into tumor composition. These models highlighted key features like cell population ratios and bacterial presence within tumors, helping surgeons identify areas requiring removal. This level of detail in real time proved especially beneficial in complex surgeries, where understanding the precise nature of the tissue being excised could determine the success of the procedure and reduce the risk of cancer recurrence.

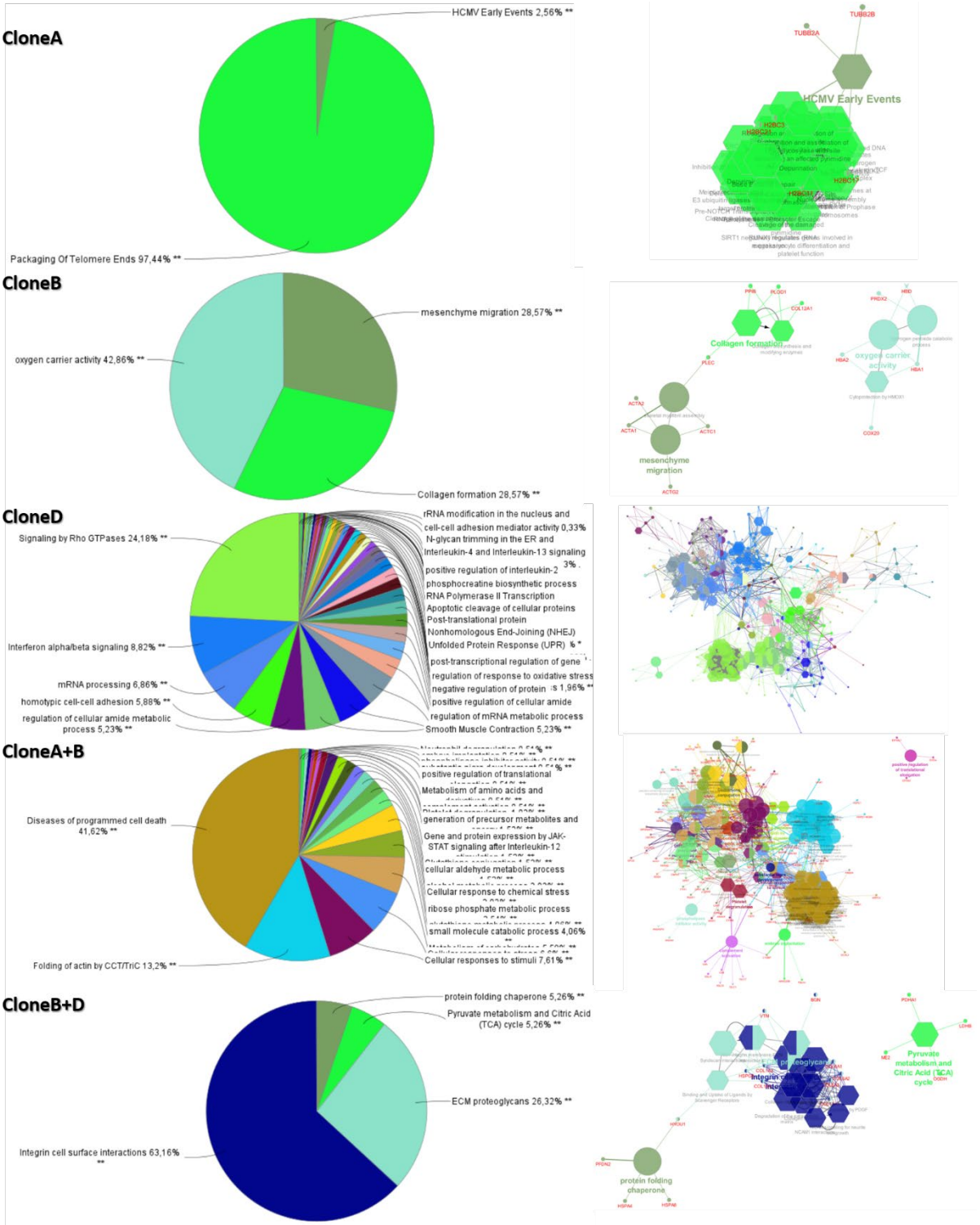
Additionally, presented studies applied immunoscoring and bacterioscoring for personalized treatment strategies. By using MALDI-IHC to assess immune cell activity and bacterial presence within tumors, clinicians gained a deeper understanding of the tumor's biological and microbial environment. This allowed for the development of treatments tailored to the specific characteristics of the tumor and its microenvironment. Such personalized approaches were especially important for aggressive or advanced cancers, where tailored treatment strategies significantly improved patient survival rates.

In conclusion, MALDI-IHC and SpiderMass technologies demonstrated remarkable potential for enhancing cancer diagnosis, prognosis, and treatment. By validating biomarkers, enabling real-time visualization of immune cells, and integrating predictive models through technologies like SpiderMass and Molecular Digital Twins, these advancements drove the future of personalized cancer care and precision surgery. The ability to guide surgeries with real-time data and develop individualized treatment plans based on tumor biology offered a promising pathway to improving cancer outcomes and reducing recurrence rates.

Appendices

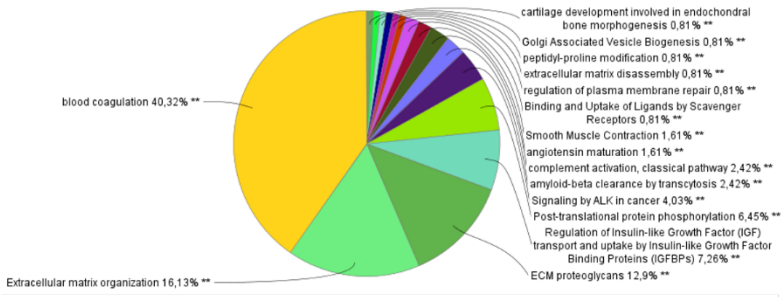
Appendix A

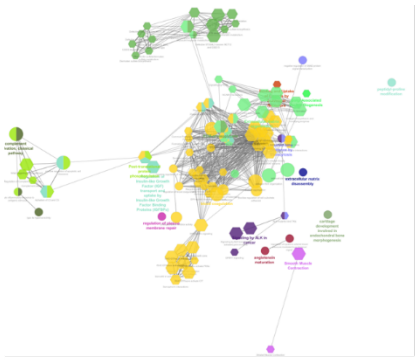
A



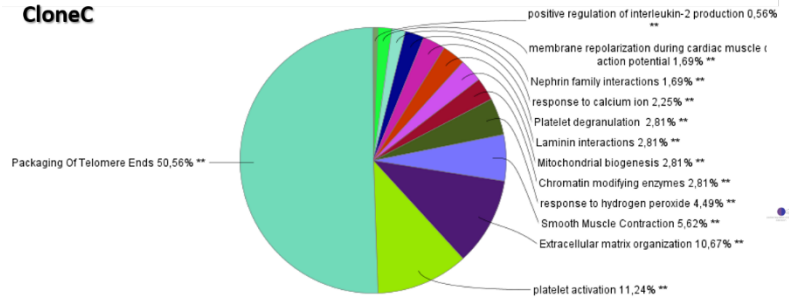
B

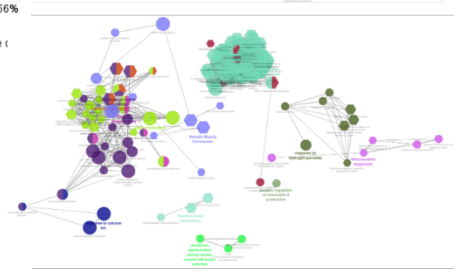
CloneA



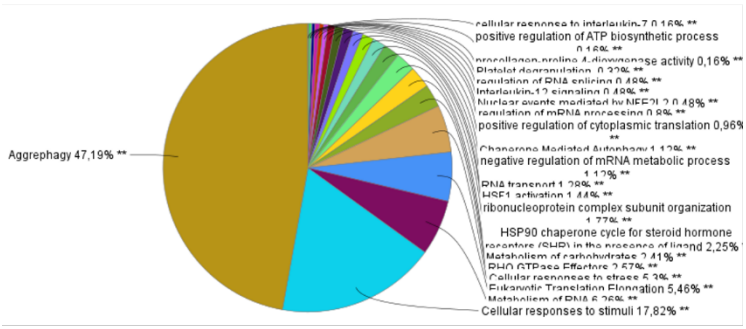


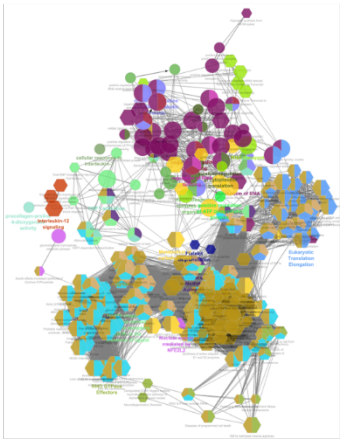
CloneC



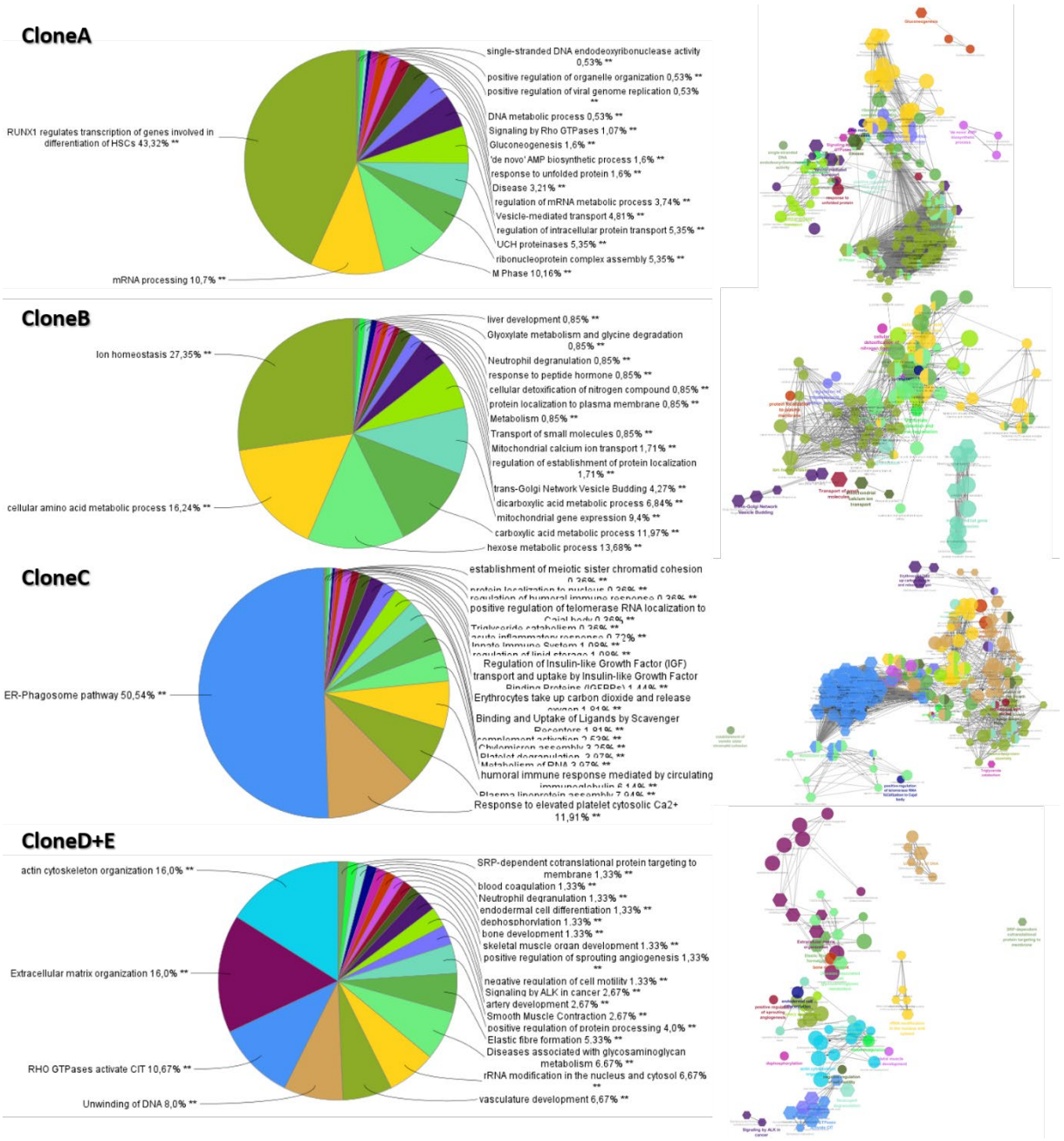


CloneB+D





C


CloneA
single-stranded DNA endodeoxyribonuclease activity 0,53% **
positive regulation of organelle organization 0,53% **
positive regulation of viral genome replication 0,53% **
DNA metabolic process 0,53% **
Signaling by Rho GTPases 1,07% **
Gluconeogenesis 1,6% **
'de novo' AMP biosynthetic process 1,6% **
response to unfolded protein 1,6% **
Disease 3,21% **
regulation of mRNA metabolic process 3,74% **
Vesicle-mediated transport 4,81% **
regulation of intracellular protein transport 5,35% **
UCH proteinases 5,35% **
ribonucleoprotein complex assembly 5,35% **
M Phase 10,16% **
mRNA processing 10,7% **
RUNX1 regulates transcription of genes involved in differentiation of HSCs 43,32% **
CloneB
liver development 0,85% **
Glyoxylate metabolism and glycine degradation 0,85% **
Neutrophil degranulation 0,85% **
response to peptide hormone 0,85% **
cellular detoxification of nitrogen compound 0,85% **
protein localization to plasma membrane 0,85% **
Metabolism 0,85% **
Transport of small molecules 0,85% **
Mitochondrial calcium ion transport 1,71% **
regulation of establishment of protein localization 1,71% **
trans-Golgi Network Vesicle Budding 4,27% **
dicarboxylic acid metabolic process 6,84% **
mitochondrial gene expression 9,4% **
carboxylic acid metabolic process 11,97% **
hexose metabolic process 13,68% **
Ion homeostasis 27,35% **
cellular amino acid metabolic process 16,24% **
CloneC
establishment of meiotic sister chromatid cohesion
positive regulation of telomerase RNA localization to
Regulation of Insulin-like Growth Factor (IGF)
transport and uptake by Insulin-like Growth Factor
Erythrocytes take up carbon dioxide and release
Binding and Uptake of Ligands by Scavenger
humoral immune response mediated by circulating
Response to elevated platelet cytosolic Ca2+
11,91% **
ER-Phagosome pathway 50,54% **
CloneD+E
SRP-dependent cotranslational protein targeting to membrane 1,33% **
blood coagulation 1,33% **
Neutrophil degranulation 1,33% **
endodermal cell differentiation 1,33% **
dephosphorylation 1,33% **
bone development 1,33% **
skeletal muscle organ development 1,33% **
positive regulation of sprouting angiogenesis 1,33% **
negative regulation of cell motility 1,33% **
Signaling by ALK in cancer 2,67% **
artery development 2,67% **
Smooth Muscle Contraction 2,67% **
positive regulation of protein processing 4,0% **
Elastic fibre formation 5,33% **
Diseases associated with glycosaminoglycan metabolism 6,67% **
rRNA modification in the nucleus and cytosol 6,67% **
vasculature development 6,67% **
actin cytoskeleton organization 16,0% **
Extracellular matrix organization 16,0% **
RHO GTPases activate CIT 10,67% **
Unwinding of DNA 8,0% **

D

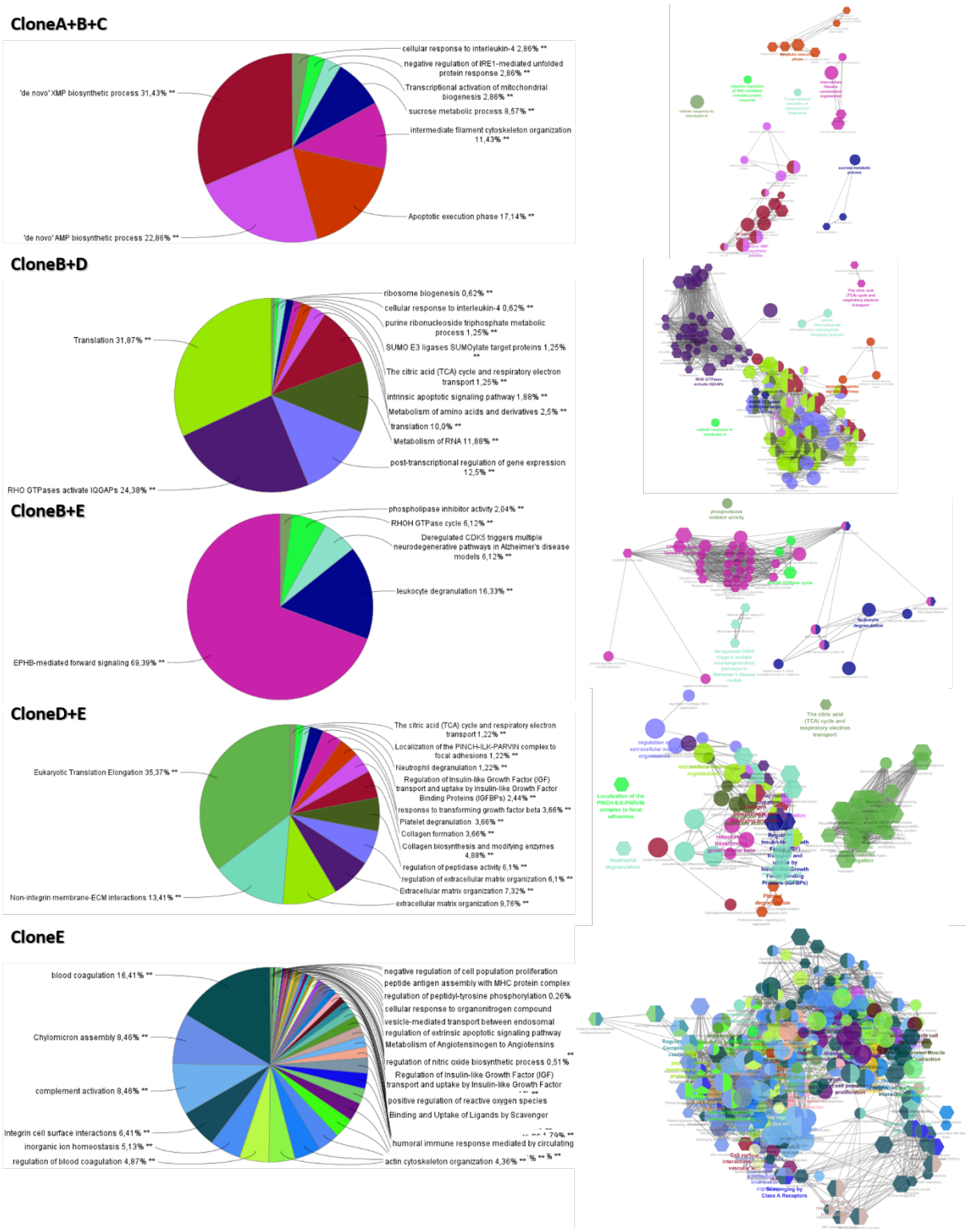


Figure 62: ClueGo biological pathways associated to over-expressed proteins involved in the different clones from A) tumor 1, B) tumor 2, C) tumor 3, or D) tumor 4.

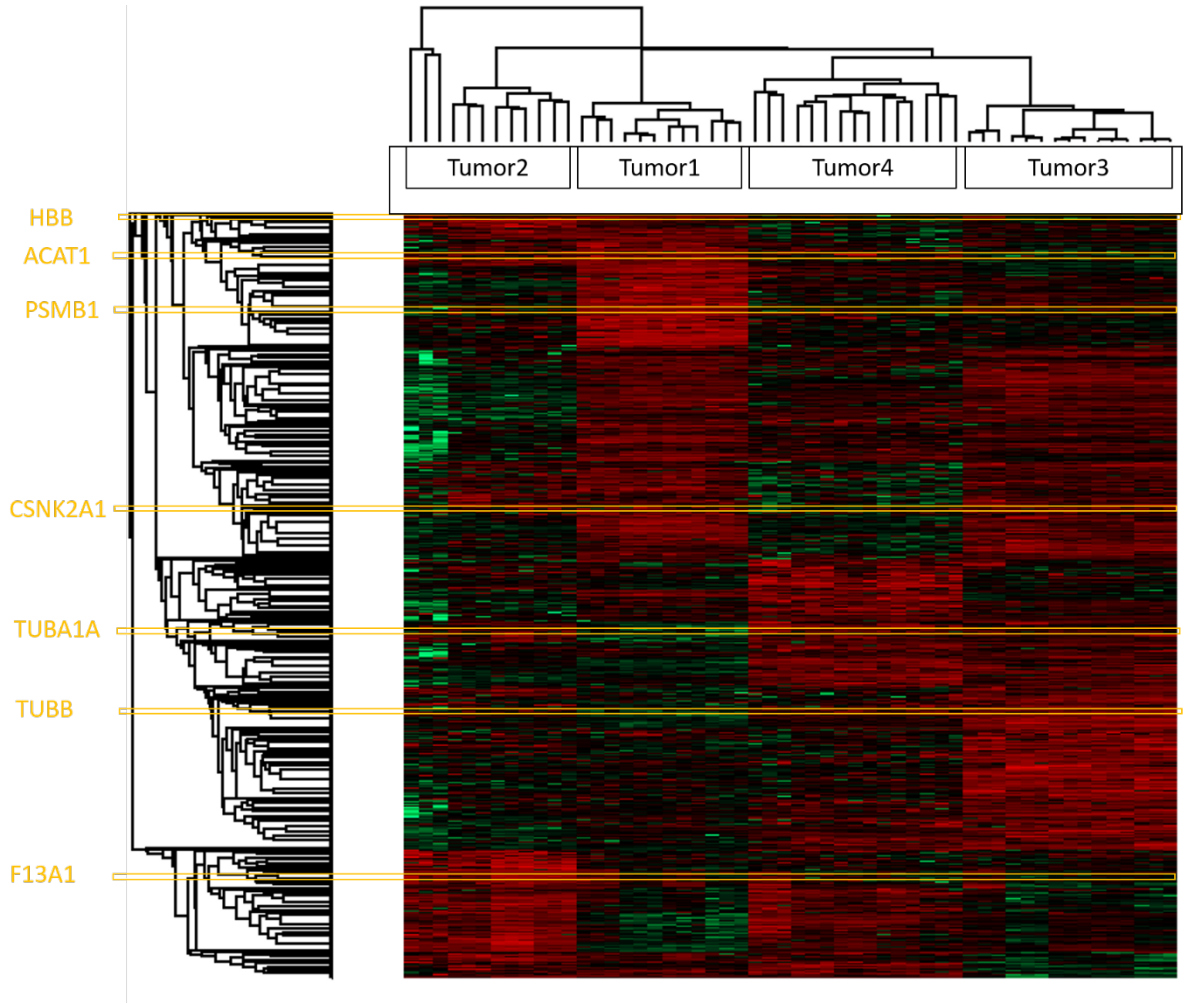


Figure 63: Inter-tumor heterogeneity analysis between tumor 1, 2, 3 and 4 highlighting therapeutic target over-expression.

Appendix B

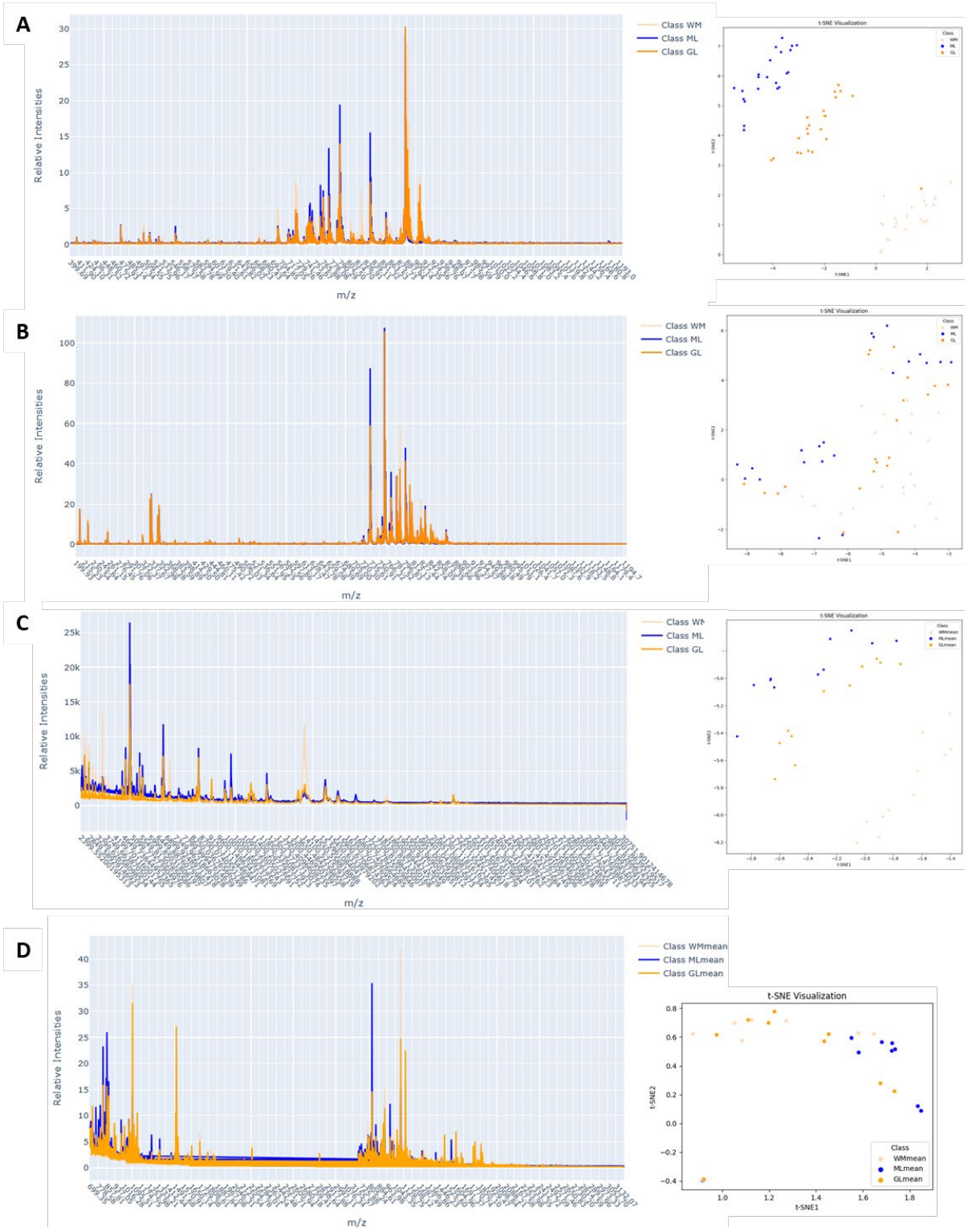


Figure 64: Cerebellum rat brain WM, ML and GL mean spectra and t-SNE separation for A) Lipid (-), B) Lipid (+), C) Protein, and D) Peptide MSI.

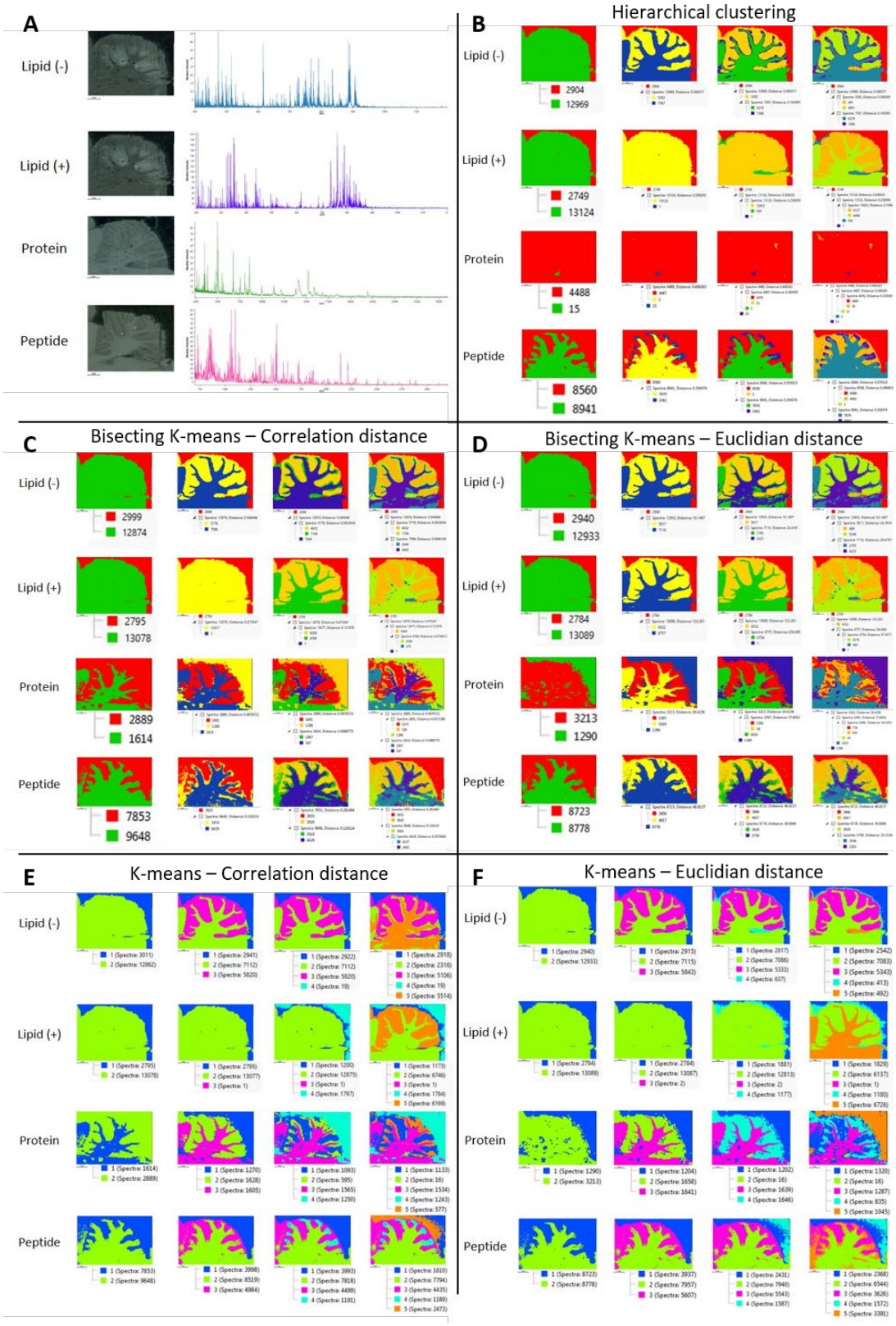
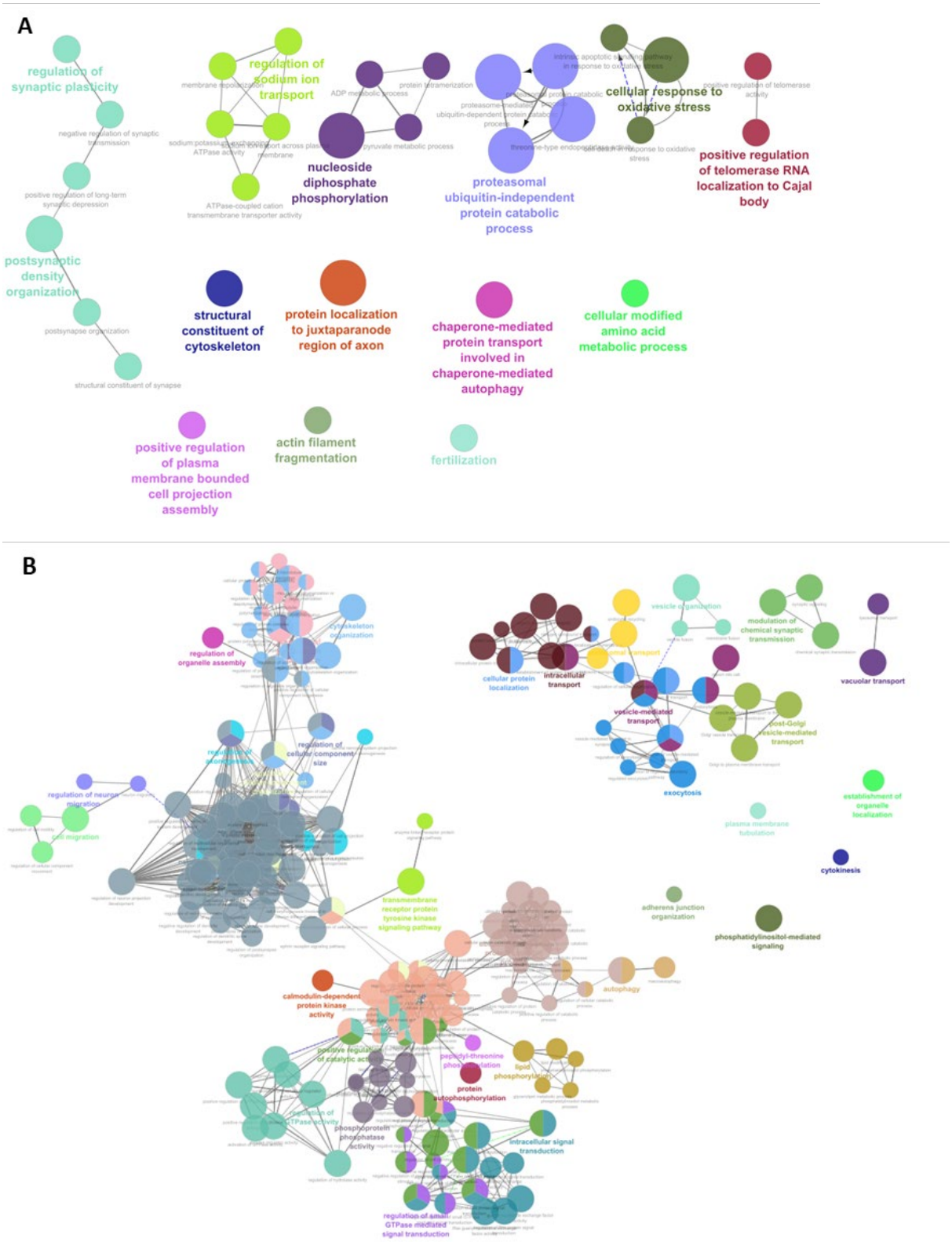
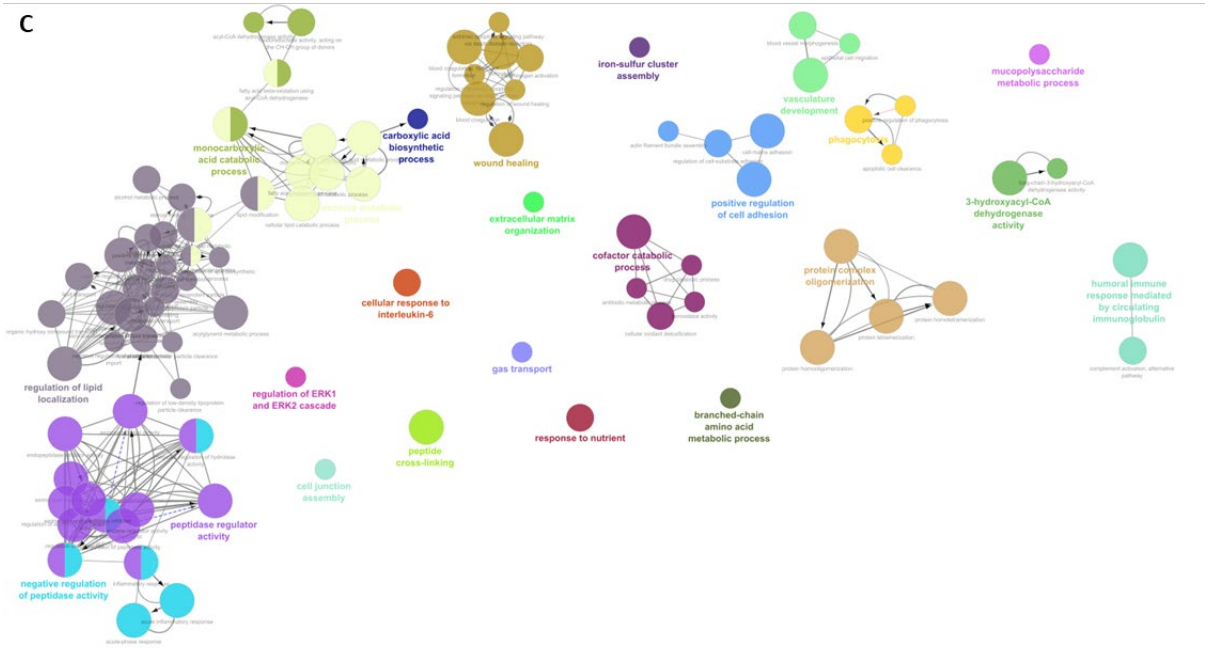


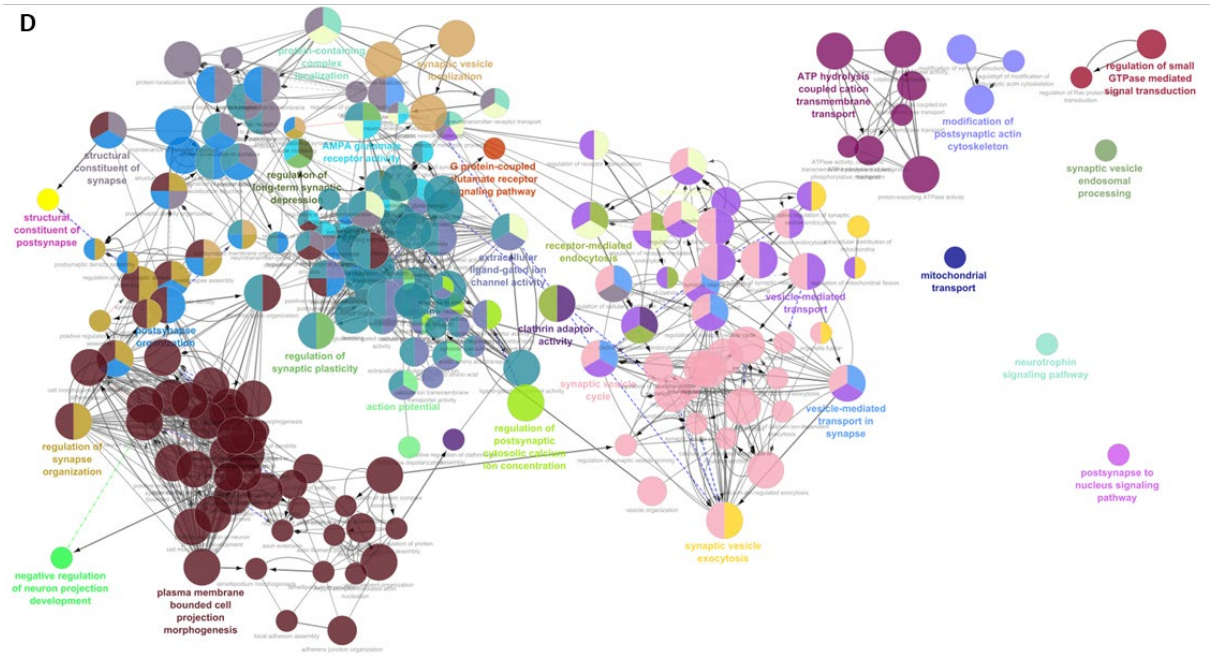
Figure 65: Comparison of different SCiLS clustering methods applied to lipid negative mode, lipid positive mode, protein and peptide imaging. A) Scan of rat brain cerebellum analyzed tissues and mean MSI spectra. Segmented images for each omics MSI analysis processed with B) Hierarchical clustering, C) Bisecting *k*-means with correlation distance, D) Bisecting *k*-means with Euclidean distance, E) *k*-means with correlation distance for 2 to 5 clusters, or E) Bisecting *k*-means with Euclidian distance for 2 to 5 clusters.


A
regulation of synaptic plasticity
negative regulation of synaptic transmission
positive regulation of long-term synaptic depression
postsynaptic density organization
postsynapse organization
structural constituent of synapse
regulation of sodium ion transport
membrane repolarization
sodium:potassium-exchanging ATPase activity
sodium ion transport across plasma membrane
ATPase-coupled cation transmembrane transporter activity
nucleoside diphosphate phosphorylation
ADP metabolic process
protein tetramerization
pyruvate metabolic process
proteasomal ubiquitin-independent protein catabolic process
proteasome-mediated ubiquitin-dependent protein catabolic process
threonine-type endopeptidase activity
cellular response to oxidative stress
positive regulation of telomerase activity
positive regulation of telomerase RNA localization to Cajal body
structural constituent of cytoskeleton
protein localization to juxtaparanode region of axon
chaperone-mediated protein transport involved in chaperone-mediated autophagy
cellular modified amino acid metabolic process
positive regulation of plasma membrane bounded cell projection assembly
actin filament fragmentation
fertilization
B
cytoskeleton organization
regulation of organelle assembly
regulation of cellular component size
regulation of neuron migration
cell migration
transmembrane receptor protein tyrosine kinase signaling pathway
calmodulin-dependent protein kinase activity
peptidyl-threonine phosphorylation
lipid phosphorylation
protein autophosphorylation
phosphoprotein phosphatase activity
intracellular signal transduction
autophagy
vesicle organization
cellular protein localization
intracellular transport
vesicle-mediated transport
modulation of chemical synaptic transmission
vacuolar transport
post-Golgi vesicle-mediated transport
exocytosis
establishment of organelle localization
plasma membrane tubulation
cytokinesis
adherens junction organization
phosphatidylinositol-mediated signaling

C


monocarboxylic acid catabolic process
carboxylic acid biosynthetic process
wound healing
iron-sulfur cluster assembly
vasculature development
phagocytosis
mucopolysaccharide metabolic process
positive regulation of cell adhesion
3-hydroxyacyl-CoA dehydrogenase activity
extracellular matrix organization
cofactor catabolic process
protein complex oligomerization
humoral immune response mediated by circulating immunoglobulin
cellular response to interleukin-6
gas transport
regulation of lipid localization
regulation of ERK1 and ERK2 cascade
response to nutrient
branched-chain amino acid metabolic process
peptide cross-linking
cell junction assembly
peptidase regulator activity
negative regulation of peptidase activity

D


protein-containing complex localization
synaptic vesicle localization
ATP hydrolysis coupled cation transmembrane transport
regulation of small GTPase mediated signal transduction
modification of postsynaptic actin cytoskeleton
structural constituent of synapse
AMPA glutamate receptor activity
G protein-coupled glutamate receptor signaling pathway
regulation of long-term synaptic depression
synaptic vesicle endosomal processing
structural constituent of postsynapse
receptor-mediated endocytosis
extracellular ligand-gated ion channel activity
mitochondrial transport
vesicle-mediated transport
postsynapse organization
clathrin adaptor activity
regulation of synaptic plasticity
neurotrophin signaling pathway
synaptic vesicle cycle
action potential
vesicle-mediated transport in synapse
regulation of postsynaptic cytosolic calcium ion concentration
regulation of synapse organization
postsynapse to nucleus signaling pathway
synaptic vesicle exocytosis
negative regulation of neuron projection development
plasma membrane bounded cell projection morphogenesis

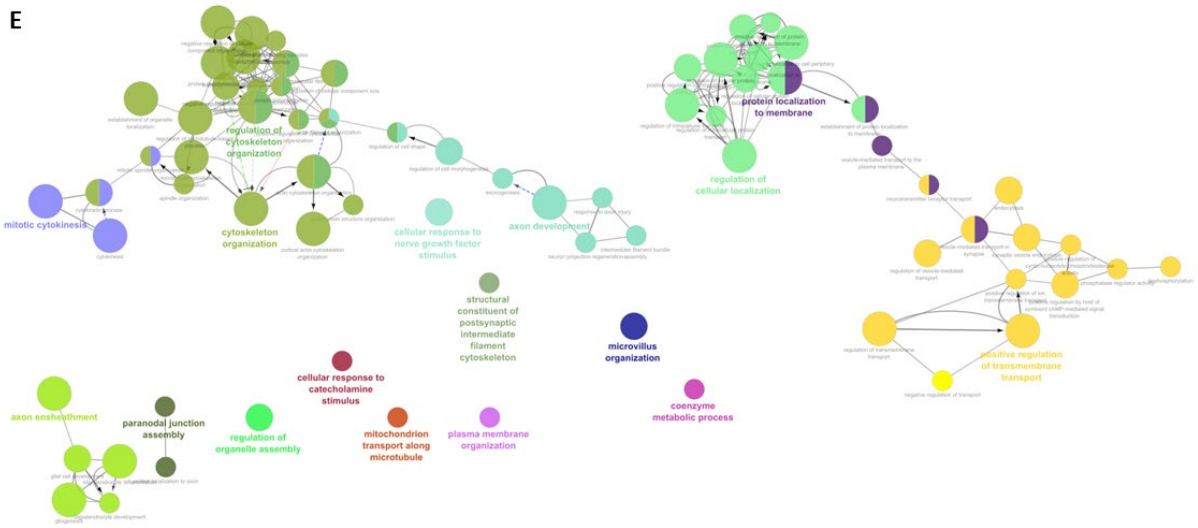


Figure 66: ClueGO biological pathways involving the significant proteins found in A) cerebellum, B) all clusters excluding cerebellum, C) ventricular system, D) cerebral cortex and E) corpus callosum.

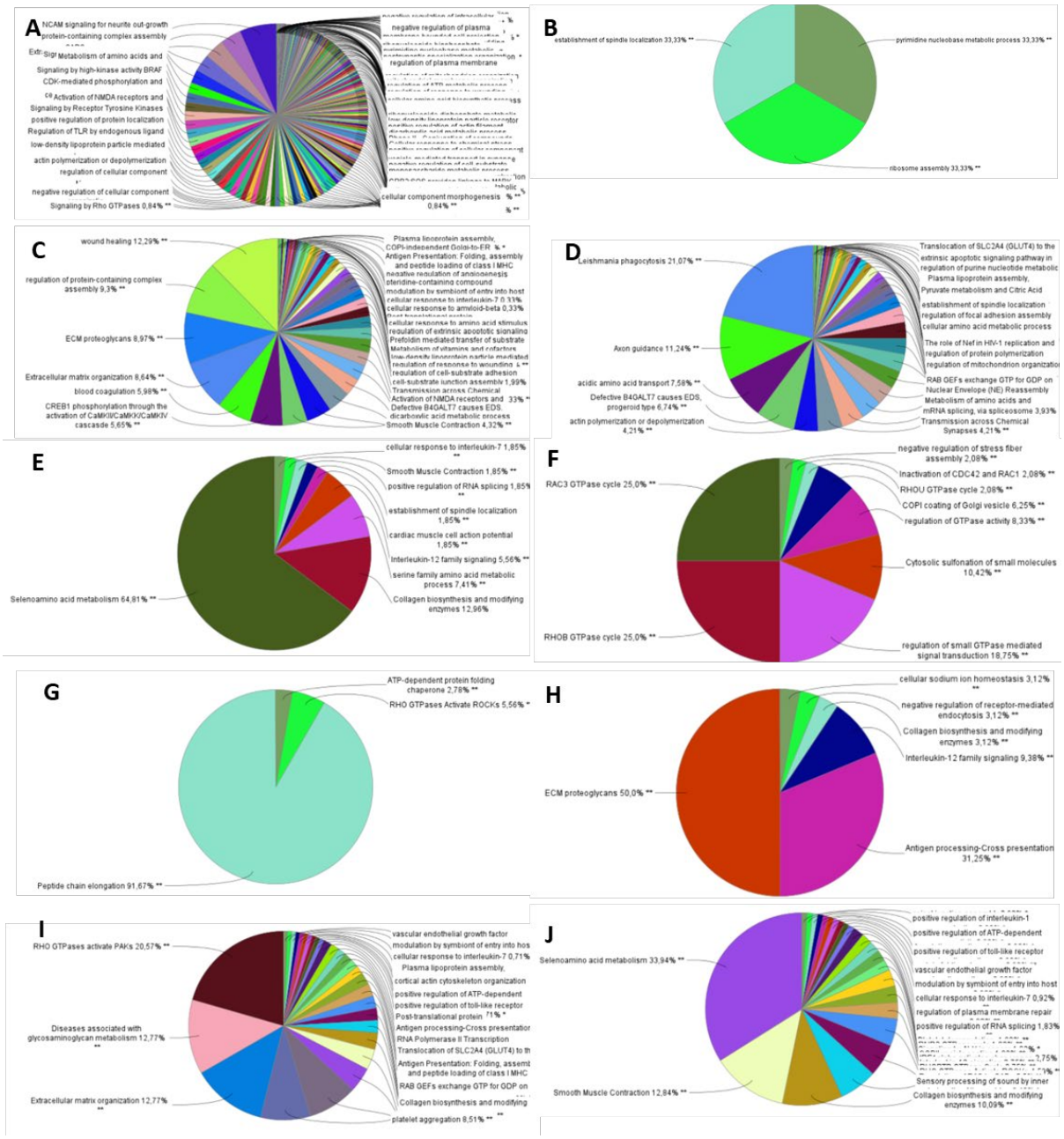


Figure 67: ClueGO biological process and reactome pathways analysis for GBM lipid clusters. A) lipid cluster 1, B) lipid cluster 2, C) lipid cluster 4, D) lipid cluster 5, E) lipid cluster 6 and 10, F) lipid cluster 7, G) lipid cluster 8, H) lipid cluster 9, I) lipid cluster 12, and J) lipid cluster 13 according to K) lipid co-segmentatioin of 9 GBM tissue patient images.

Appendix C

Table 10: Clinical data of patients included in the FF OC cohort.

Patient	Age	Recurrence (months)	Survival (months)	Death	First Treatment
			Serous High Grade		
47	37	5	28	yes	chemotherapy
48	61	33	86	no	chemotherapy
44	58	52	52	no	surgery
62	60	7	42	yes	surgery
63	41	20	43	yes	chemotherapy
86	59	17	63	yes	surgery
89	69	8	42	yes	chemotherapy
95	65	8	60	yes	surgery
114	64	25	40	yes	surgery
42	50	84	84	no	surgery
187	75	12	24	yes	chemotherapy
183	41	24	54	no	chemotherapy
			Serous Borderline		
13	38	17	63	yes	surgery
22	52	92	92	no	surgery
64	51	70	70	no	surgery
65	52	63	63	no	surgery
113	26	6	66	no	surgery
174	67	6	48	no	surgery
			Mucinous		
6	70	94	94	no	surgery
8	54	7	7	no	surgery
23	43	17	31	no	chemotherapy
55	86	21	21	no	surgery
68	52	4	4	no	surgery
84	54	51	51	no	surgery
103	19	64	64	no	surgery
135	53	59	59	no	surgery
			Endometrioid		
108	58	18	57	non	surgery
179	67	16	38	yes	surgery
181	49	43	43	non	surgery
			Healthy		
5	24	no	no	no	surgery
20	53	6	22	yes	chemotherapy
22	52	no	no	no	surgery
48	61	33	no	no	surgery
51	65	no	72	yes	surgery
52	66	39	68	yes	surgery
53	57	no	no	no	surgery
54	38	25	no	no	chemotherapy
56	81	no	no	no	chemotherapy
59	53	no	no	no	surgery
61	56	12	24	yes	chemotherapy
63	41	20	43	yes	chemotherapy

Table 11: Clinical data of patients included in the FFPE cohort.

Patient	*Age*	*Recurrence (yes/no)*	*Recurrence (months)*	*Survival*	*First treatment*
		Carcinosarcome			
4	57	yes	1	15	surgery
		CCC			
26	NA	no		36	surgery
32	NA	no		31	surgery
40	NA	yes	13	15	surgery
42	68	yes	1	2	surgery
74	74	no		5	surgery
		Endometrioid			
51	51	no		48	surgery
55	55	yes	16	32	surgery
13	NA	no		48	surgery
37	59	no		22	surgery
		HGSC			
1	66	yes	39	50	surgery
2	67	no		27	surgery
3	NA	no		14	surgery
5	NA	no		25	surgery
6	48	no		36	surgery
7	NA	no		3	surgery
8	66	yes	13	56	surgery
9	67	yes	28	45	surgery
10	NA	no		55	surgery
11	61	yes	15	48	surgery
12	NA	no		36	surgery
14	48	yes	22	46	surgery
15	74	no		47	surgery
16	58	yes	39	45	surgery
19	NA	no		32	surgery
20	74	no		41	surgery
22	73	yes	14	41	surgery
23	NA	no		38	surgery
24	NA	no		39	surgery
27	78	no		37	surgery
28	58	yes	10	30	surgery
29	NA	no		38	surgery
30	71	yes	8	38	surgery
33	65	yes	24	31	surgery
34	NA	no		31	surgery
35	62	no		27	surgery
38	61	no		20	surgery
39	62	no		18	surgery
41	77	no		12	surgery
43	69	no		5	surgery
44	77	no		3	surgery
45	65	no		3	surgery

Table 12: List of samples used for the validation test in blind.

Sample	*Age*	*Class*	*Recurrence*	*OS*	*Death*	*FIGO*	*First treatment*
102.1	64	HGSOC	52	52	no	IIIC	Surgery
102.2	64	Normal			no		Surgery
104.1	44	HGSOC	24	60	no	IIIC	Chemotherapy
104.2	44	Normal			no		Chemotherapy
106.1	38	HGSOC	48	48	no	IIIC	Chemotherapy
106.2	38	Normal			no		Chemotherapy
108.1	46	HGSOC	50	50	no	IIIC	Surgery
108.2	46	Normal			no		Surgery
109.1	48	Endometroid	20	20	no	IA	Surgery
109.2	48	Normal			no		Surgery
110.1	61	HGSOC	33	33	no	IIIC	Chemotherapy
110.2	61	Normal			no		Chemotherapy
220.1	62	Endometrioid			no		Chemotherapy
220.2	62	Normal			no		Chemotherapy
221.1	72	HGSOC	6	6	no	IIIC	Surgery
222.1	55	Endometrioid			no		Surgery
222.2	55	Normal			no		Surgery
223.1	65	SBL	7	7	no	IIIC	Surgery
223.2	65	Normal			no		Surgery
224.1	68	Mucinous	8	8	no	IIIC	Chemotherapy
224.2	68	Normal			no		Chemotherapy
225.1	60	HGSOC	6	6	no	IIIC	Surgery
226.1	64	Normal			no		Surgery
227.1	56	endometrioid	11	11	no	IIIC	Chemotherapy
228.1	67	HGSOC	11	11	no	IVB	Chemotherapy
229.1	65	Mucinous	2	2	no	IA	Surgery

Table 13: Comparison of the pathologist annotation and the prediction obtained for the LDA FF, LDA mixed and Ridge Classifier mixed models on 72 blinded analyses.

Sample	*Pathologist Annotation*	*LDA FF prediction*			*LDA Mixed prediction*			*Ridge Classifier mixed prediction*		
102.1	HGSC	HGSC	HGSC	HGSC	HGSC	HGSC	HGSC	HGSC	HGSC	HGSC
104.1	HGSC and Heal	HGSC	Heal	Heal	SBL	SBL	Heal	Heal	Heal	Heal
104.2	Heal	Heal	Heal	Heal	Heal	Heal	SBL	Heal	Heal	Heal
106.1	SBL	SBL	SBL	SBL	SBL	SBL	SBL	SBL	SBL	SBL
106.2	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal
108.1	HGSC	HGSC	HGSC	HGSC	HGSC	HGSC	HGSC	HGSC	HGSC	EC
108.2	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal
109.1	EC	MC	MC	SBL	MC	MC	SBL	MC	MC	HGSC
109.2	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal
110.1	HGSC	HGSC	HGSC	SBL	HGSC	HGSC	HGSC	HGSC	HGSC	HGSC
220.1	EC	MC	MC	SBL	MC	SBL	SBL	Heal	Heal	HGSC
220.2	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal
221.1	HGSC	MC	MC	HGSC	MC	MC	HGSC	HGSC	HGSC	HGSC
222.1	Heal	Heal	Heal	HGSC	SBL	SBL	SBL	Heal	Heal	Heal
222.2	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal
223.1	SBL	SBL	SBL	Healthy	SBL	SBL	SBL	SBL	Heal	Heal
223.2	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal
224.1	Heal with mucinous	MC	MC	Heal	MC	MC	SBL	MC	MC	Heal
224.2	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal
225.1	HGSC	Heal	Heal	SBL	HGSC	HGSC	Heal	HGSC	HGSC	HGSC
226.1	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal	Heal
227.1	EC	SBL	SBL	Heal	SBL	SBL	Heal	HGSC	HGSC	Heal
228.1	Heal	Heal	Heal	Heal	Heal	Heal	SBL	Heal	Heal	Heal
229.1	MC	MC	MC	Heal	MC	SBL	SBL	MC	MC	MC

Table 14: Discriminative lipids annotated by MS/MS for the different OC subtypes.

Ionisation mode	*Sample*	*Potential identification*	*Peak m/z*
Negative $[M-H]^-$	Normal	PE (16:0_22:5)	748.55
	Normal	PS (18:0_18:1)	788.55
	Normal	PI (20:4_18:0)	885.55
	Normal	PA (18:1_18:0)	701.55
	SBL	PI (22:1_18:0)	919.79
	SBL	TG (56 :12)	893.63
	SBL	PE (18:0_22:4)	794.55
	HGSC	PE (18:0_16:0)	718.55
	HGSC	PE(18:0_18:2)	742.55
	HGSC	PA (18:1_20:2)	725.55
	HGSC	PE (18:0_18:1)	744.55
	HGSC	PA (18:0_22:6)	747.55
	MC	PA (18:0_20:2)	727.55
	MC	PS (22:2_20:4)	862.65
	EC	PS(22:0_22:2) or PS (22:1_22:1)	898.55
Positive $[M+H]^+$	SBL/HGSC	PE (16:0_20:0)	748.55
	SBL/HGSC	PE (18:2_22:4)	792.55
	SBL	PS (P-16:0_18:2)	744.55
	SBL	PS (18:4_18:1)	782.55
	HGSC	PS (O-16:0_20:1)	776.55
	HGSC	PE (O-18:0_18:2)	730.55
	MC	PS (O-18:0_18:1)	776.55
	MC	PC (20:1_16:1)	788.65
	MC	PE (O-18:0_20:4)	752.55
	MC	PC (18:4_12:0)	698.55
	EC	PE (16:0_20:5)	724.55

Table 15: Samples used to calculated immunoscore based on SpiderMass-MSI.

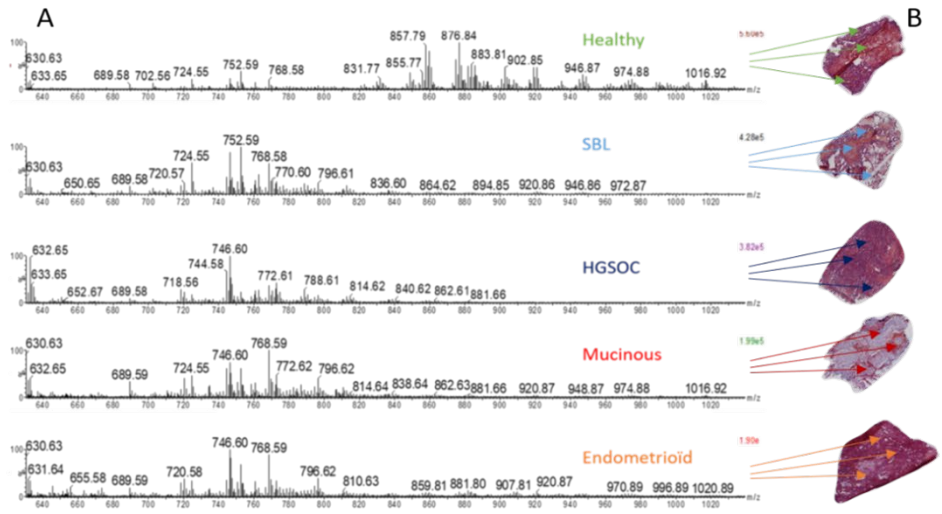
Survival	*Sample*	*Age*	*Recurrence*	*OS*	*Death*	*First treatment*
Long survival <50 months	89-3	59	17	42	yes	surgery
	66-4	74	8	42	yes	surgery
	114.1	64	25	40	yes	surgery
	62.1	60	7	42	yes	surgery
Short survival >50 months	102.1	64	52	52	no	surgery
	108.1	46	50	50	no	surgery
	44-2	50	51	51	no	surgery
	42-1	50	84	84	no	surgery

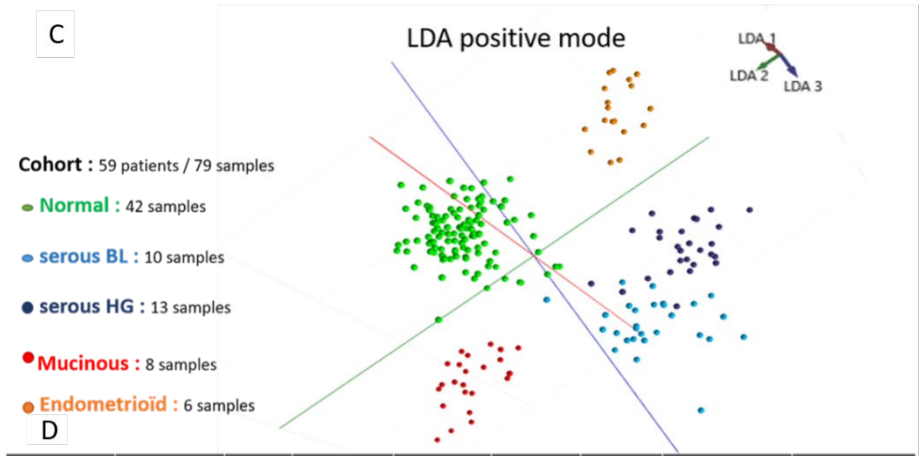
Table 16: M1- and M2-like macrophages immunoscores and their M1/M2 ratio obtained for each SpiderMass-MSI analyzed tissue.

	Long survival			*Short survival*		
Phenotype	**M1-like**	**M2-like**	**Ratio**	**M1-like**	**M2-like**	**Ratio**
Immunoscore	0.029	0.007	4.270	0.062	0.007	8.954
	0.087	0.012	7.164	0.007	0.002	2.513
	0.033	0.004	6.864	0.014	0.005	3.077
	0.029	0.005	5.841	0.019	0.009	2.194
Mean	**0.045**	**0.007**	**6.035**	**0.025**	**0.005**	**4.184**

Table 17: Quantification of macrophages (Ki67 marker) in the different OC subtypes calculated from MALDI-IHC.

	SBL	*HGSC*	*MC*	*EC*	*Healthy*
	19.59	47.25	17.64	8.08	0.02
Quantification	42.01	22.04	42.64	15.20	6.00
	1.10	15.05	10.38	16.31	26.18
Mean	**20.90**	**28.11**	**23.55**	**13.19**	**10.3**
Median	**19.59**	**22.04**	**17.64**	**15.20**	**6.00**





D

mode	validation	group	Number of spectra	Number of passes	Number of failures	Number of outliers	Correct classification excluding outliers	Correct classification with outliers
positive	20% out	total	207	197	8	2	96.10%	95.17%
	Full group out	total	207	138	48	21	74.19%	66.67%

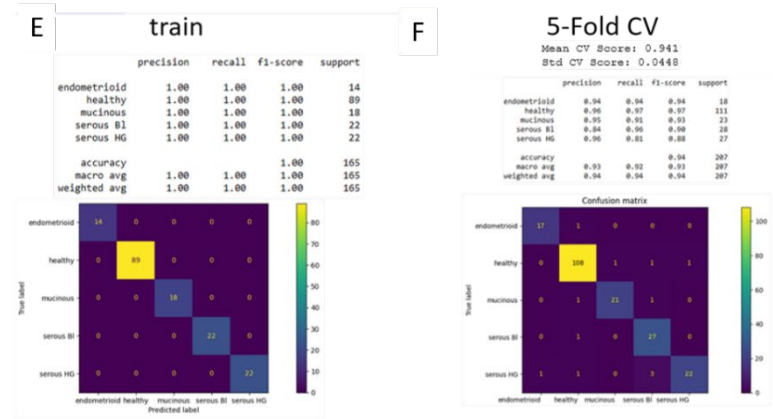


Figure 68: Morphological and spectral fingerprint of ovarian cancer subtypes in positive ion mode. **(A)** Mean spectra for each ovarian cancer subtype (4mJ/shot, burst mode, 10 shots, 1s/spectrum). (B) HPS staining of the corresponding histological sections. The arrows indicate the location where the laser was fired. Classification models and their cross-validation for OC subtypes in the positive ion mode. (C) Linear discriminant analysis model for the different OC subtypes. (D) Cross-validations of the LDA model by "20out" and "full group out" methods with and without outliers. (E) Training of the model based on the RIDGE classifier. (F) Cross-validation report of the RIDGE model by 5-fold method.

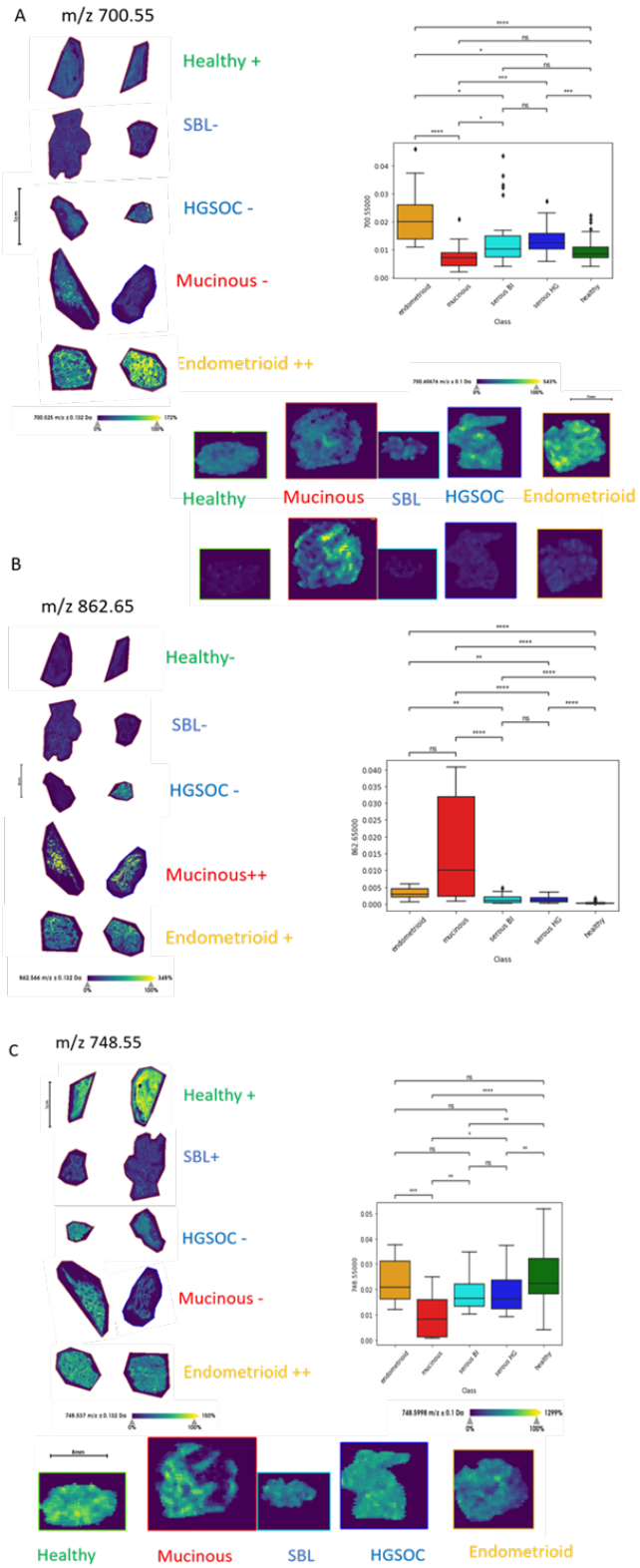


Figure 69: Cross-validation of the OC subtypes lipid markers by MALDI-MSI and SpiderMass. (A) Example of ion *m/z* 700.55 which is specific to endometrioid. (B) Example of ion *m/z* 862.65 which is specific to mucinous carcinoma. (C) Example of ion *m/z* 748.55 which is specific to endometrioid and normal tissues. The contribution of each ion in each tissue calculated by LIME is represented by a (+) if positive and by a (-) if negative. A Kruskal-Wallis test was performed on the SpiderMass data for this ion and the relative intensities are represented as a boxplot. (**** $p < 0.0001$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$) NS for $p > 0.05$).

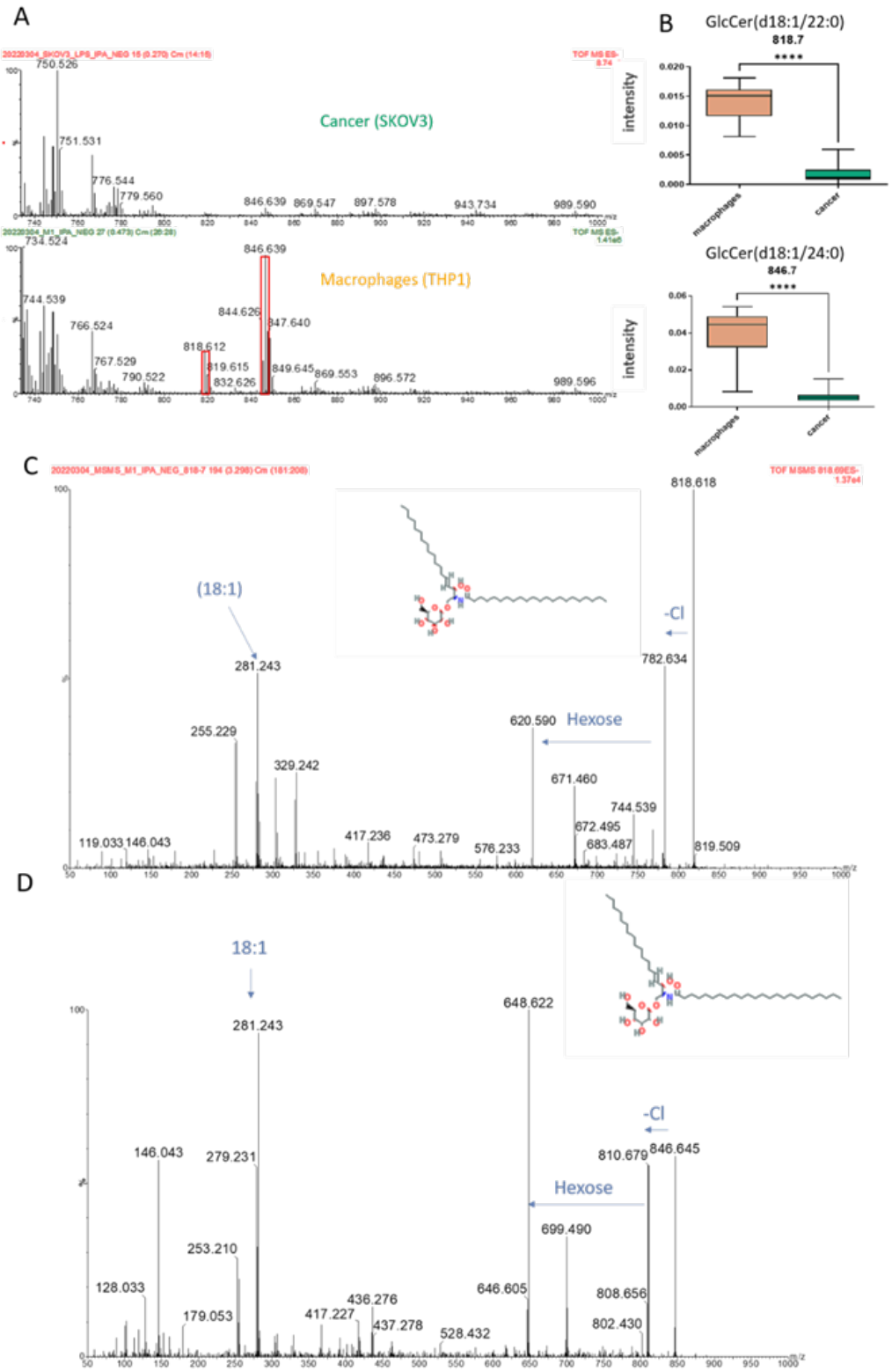


Figure 70: MS/MS identification and variation of abundance of two hexosylceramides between cancer cells and macrophages. (A) SpiderMass mean spectra obtained from the THP1 macrophage versus SKOV3 cancer cell lines. The red frames indicate the ions *m/z* 818.65 and *m/z* 846.65 specific to the macrophages. (B) Boxplot based on the relative intensities of these two ions for SKOV3 vs. THP1 (****=pvalue<0.0001) showing the higher abundance of these two markers in the macrophages. (C) MS/MS spectrum of the ion *m/z* 818.65 identified as GlcCer d40:1 (d18:1_22:0). (D) MS/MS spectrum of the ion *m/z* 846.65 identified as GlcCer d42:1 (d18:1_24:0).

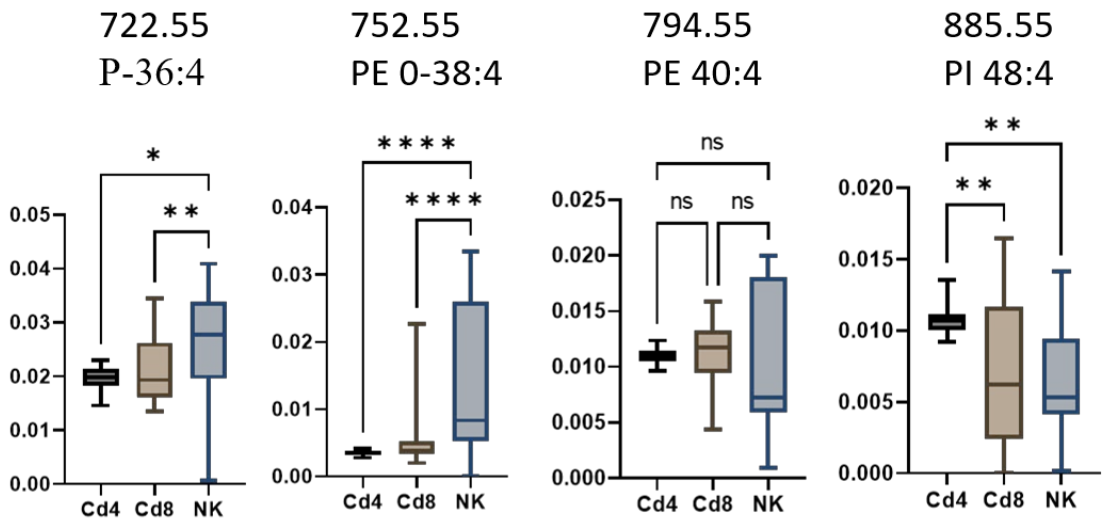


Figure 71: Discriminative lipid markers of lymphocyte cells. Boxplots based on the relative intensities of ions *m/z* 722.55, 752.55, 794.55 and 885.55 plotted for the different subpopulation of lymphocytes (NK, CD8 or CD4) (**** $p < 0.0001$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$) NS for $p > 0.05$).

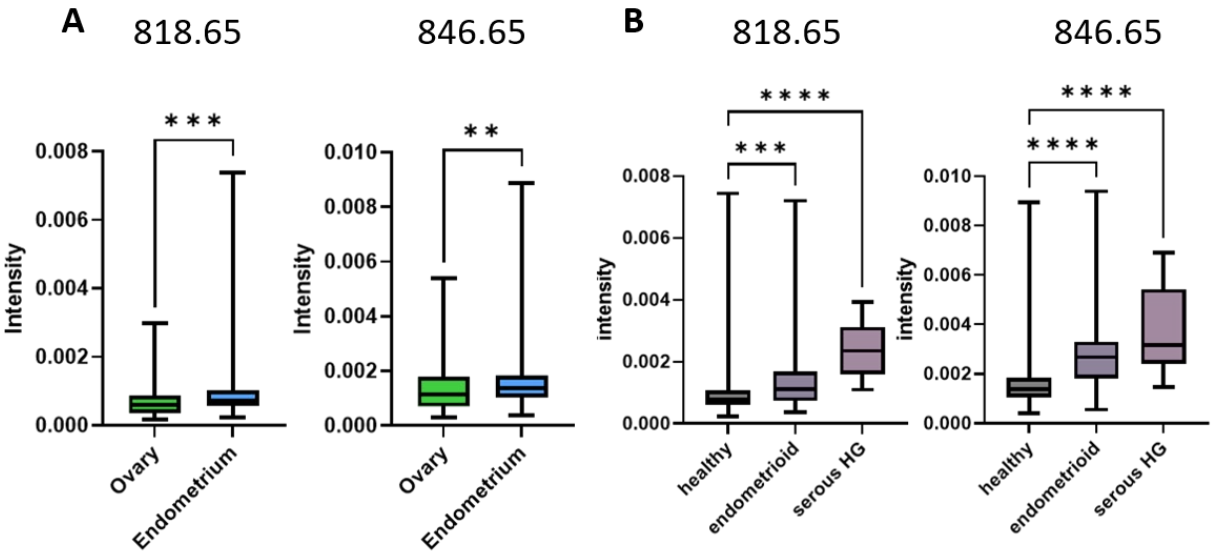


Figure 72: Abundance of lipid markers specific to macrophages in normal ovary and endometrium tissues. (A) In normal endometrium and ovary tissues. (B) In the different endometrium cancer subtypes (Healthy, HGSC or endometrioid). (**** $p < 0.0001$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$) NS for $p > 0.05$).

Table 18: Percentages of S. infantis, S. lugdunensis, M. radiotolerans bacteria in the tumor and peritumoral region of the different mice tumors. Percentages obtained for the three bacterial strains in each mammary gland tissues imaged by SpiderMass in both MS ion modes.

Mice	Tissue area	*S. infantis* (%)	*S. lugdunensis* (%)	*M. radiotolerans* (%)
Negative ion mode				
M3-T1	Tumor	65	25	10
	Peritumor	54	30	17
M3-T2	Tumor	91	9	0.4
	Peritumor	67	3	30
M3-T3	Tumor	56	35	19
	Peritumor	46	29	16
M4-T1	Tumor	5	17	78
	Peritumor	23	34	43
M7-T2	Tumor	81	20	0
	Peritumor	75	18	8
M7-T3	Tumor	80	20	0
	Peritumor	66	29	6
M8-T1	Tumor	72	7	21
	Peritumor	52	43	4
M8-T2	Tumor	83	15	2
	Peritumor	68	22	10
M12-T1	Tumor	50	40	11
	Peritumor	25	34	42
M13-T1	Tumor	80	16	3
	Peritumor	77	16	7
M14-T1	Tumor	83	8	9
	Peritumor	66	23	11
M14-T2	Tumor	80	15	5
	Peritumor	82	17	2
M15-T1	Tumor	91	7	2
	Peritumor	63	37	0.7
Healthy		**4.3**	**37**	**58**
Median in peritumoral regions		**66**	**29**	**10.5**
Median in tumor regions		**80**	**16**	**10**
Positive ion mode				
M1-T2	Tumor	60	20	36
	Peritumor	13	17	70
M2-T2	Tumor	75	0	25
	Peritumor	56	5	40
M9-T2	Tumor	55	0	45
	Peritumor	53	2	46
Median in peritumoral regions		**53**	**5**	**46**
Median in tumor regions		**60**	**0**	**36**

Bibliography

Abdelmoula, W. M., Pezzotti, N., Hölt, T., Dijkstra, J., Vilanova, A., McDonnell, L. A., & Lelieveldt, B. P. F. (2018). Interactive Visual Exploration of 3D Mass Spectrometry Imaging Data Using Hierarchical Stochastic Neighbor Embedding Reveals Spatiomolecular Structures at Full Data Resolution. *Journal of Proteome Research*, *17*(3), 1054–1064. https://doi.org/10.1021/ACS.JPROTEOME.7B00725

Aletti, G. D., Dowdy, S. C., Podratz, K. C., & Cliby, W. A. (2007). Relationship among surgical complexity, short-term morbidity, and overall survival in primary surgery for advanced ovarian cancer. *American Journal of Obstetrics and Gynecology*, *197*(6), 676.e1-676.e7. https://doi.org/10.1016/J.AJOG.2007.10.495

Alexandrov, T. (2012). MALDI imaging mass spectrometry: statistical data analysis and current computational challenges. *BMC Bioinformatics*, *13*(Suppl 16), S11. https://doi.org/10.1186/1471-2105-13-S16-S11

Allaume, P., Rabilloud, N., Turlin, B., Bardou-Jacquet, E., Loréal, O., Calderaro, J., Khene, Z. E., Acosta, O., De Crevoisier, R., Rioux-Leclercq, N., Pecot, T., & Kammerer-Jacquet, S. F. (2023). Artificial Intelligence-Based Opportunities in Liver Pathology—A Systematic Review. *Diagnostics*, *13*(10), 1799. https://doi.org/10.3390/DIAGNOSTICS13101799/S1

Alpuim Costa, D., Nobre, J. G., Batista, M. V., Ribeiro, C., Calle, C., Cortes, A., Marhold, M., Negreiros, I., Borralho, P., Brito, M., Cortes, J., Braga, S. A., & Costa, L. (2021). Human Microbiota and Breast Cancer-Is There Any Relevant Link?-A Literature Review and New Horizons Toward Personalised Medicine. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/FMICB.2021.584332

Alshammari, F. O. F. O., Al-Saraireh, Y. M., Youssef, A. M. M., Al-Sarayra, Y. M., & Alrawashdeh, H. M. (2021). Glypican-1 Overexpression in Different Types of Breast Cancers. *OncoTargets and Therapy*, *14*, 4309. https://doi.org/10.2147/OTT.S315200

Alvarez Secord, A., O'Malley, D. M., Sood, A. K., Westin, S. N., & Liu, J. F. (2021). Rationale for combination PARP inhibitor and antiangiogenic treatment in advanced epithelial ovarian cancer: A review. *Gynecologic Oncology*, *162*(2), 482–495. https://doi.org/10.1016/J.YGYNO.2021.05.018

Angulo, C., Gonzalez-Abril, L., Raya, C., & Ortega, J. A. (2020). A Proposal to Evolving Towards Digital Twins in Healthcare. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12108 LNBI*, 418–426. https://doi.org/10.1007/978-3-030-45385-5_37

Arthur, D., & Vassilvitskii, S. (n.d.-a). *k-means++: The Advantages of Careful Seeding*.

Arthur, D., & Vassilvitskii, S. (n.d.-b). *k-means++: The Advantages of Careful Seeding*.

Ayasoufi, K., Wolf, D. M., Namen, S. L., Jin, F., Tritz, Z. P., Pfaller, C. K., Zheng, J., Goddery, E. N., Fain, C. E., Gulbicki, L. R., Borchers, A. L., Reesman, R. A., Yokanovich, L. T., Maynes, M. A., Bamkole, M. A., Khadka, R. H., Hansen, M. J., Wu, L. J., & Johnson, A. J. (2023). Brain resident memory T cells rapidly expand and initiate neuroinflammatory responses following CNS viral infection. *Brain, Behavior, and Immunity*, *112*, 51–76. https://doi.org/10.1016/J.BBI.2023.05.009

Bagaria, N., Laamarti, F., Badawi, H. F., Albraikan, A., Velazquez, R. A. M., & El Saddik, A. (2020). Health 4.0: Digital Twins for Health and Well-Being. *Connected Health in Smart Cities*, 143–152. https://doi.org/10.1007/978-3-030-27844-1_7

Balog, J., Szaniszlo, T., Schaefer, K. C., Denes, J., Lopata, A., Godorhazy, L., Szalay, D., Balogh, L., Sasi-Szabo, L., Toth, M., & Takats, Z. (2010). Identification of biological tissues by rapid evaporative ionization mass spectrometry. *Analytical Chemistry*, *82*(17), 7343–7350. https://doi.org/10.1021/AC101283X

Banks, W. A., Farr, S. A., Salameh, T. S., Niehoff, M. L., Rhea, E. M., Morley, J. E., Hanson, A. J., Hansen, K. M., & Craft, S. (2018). Triglycerides cross the blood-brain barrier and induce central leptin and insulin receptor resistance. *International Journal of Obesity (2005)*, *42*(3), 391–397. https://doi.org/10.1038/IJO.2017.231

Barakat, C., Escalante, G., Stevenson, S. W., Bradshaw, J. T., Barsuhn, A., Tinsley, G. M., & Walters, J. (2022). Can Bodybuilding Peak Week Manipulations Favorably Affect Muscle Size, Subcutaneous Thickness, and Related Body Composition Variables? A Case Study. *Sports (Basel, Switzerland)*, *10*(7). https://doi.org/10.3390/SPORTS10070106

Batterman, K. V., Cabrera, P. E., Moore, T. L., & Rosene, D. L. (2021). T Cells Actively Infiltrate the White Matter of the Aging Monkey Brain in Relation to Increased Microglial Reactivity and Cognitive Decline. *Frontiers in Immunology*, *12*. https://doi.org/10.3389/FIMMU.2021.607691

Bazira, P. J., Ellis, H., & Mahadevan, V. (2022). Anatomy and physiology of the breast. *Surgery (United Kingdom)*, *40*(2), 79–83. https://doi.org/10.1016/j.mpsur.2021.11.015

Berek, J. S., Renz, M., Kehoe, S., Kumar, L., & Friedlander, M. (2021). Cancer of the ovary, fallopian tube, and peritoneum: 2021 update. *International Journal of Gynaecology and Obstetrics: The Official Organ of the International Federation of Gynaecology and Obstetrics*, *155 Suppl 1*(Suppl 1), 61–85. https://doi.org/10.1002/IJGO.13878

Bikfalvi, A., da Costa, C. A., Avril, T., Barnier, J. V., Bauchet, L., Brisson, L., Cartron, P. F., Castel, H., Chevet, E., Chneiweiss, H., Clavreul, A., Constantin, B., Coronas, V., Daubon, T., Dontenwill, M., Ducray, F., Enz-Werle, N., Figarella-Branger, D., Fournier, I., … Virolle, T. (2023). Challenges in glioblastoma research: focus on the tumor microenvironment. *Trends in Cancer*, *9*(1), 9–27. https://doi.org/10.1016/J.TRECAN.2022.09.005

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W. H., Pagès, F., Trajanoski, Z., & Galon, J. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics (Oxford, England)*, *25*(8), 1091–1093. https://doi.org/10.1093/BIOINFORMATICS/BTP101

Björnsson, B., Borrebaeck, C., Elander, N., Gasslander, T., Gawel, D. R., Gustafsson, M., Jörnsten, R., Lee, E. J., Li, X., Lilja, S., Martínez-Enguita, D., Matussek, A., Sandström, P., Schäfer, S., Stenmarker, M., Sun, X. F., Sysoev, O., Zhang, H., & Benson, M. (2019). Digital twins to personalize medicine. *Genome Medicine*, *12*(1). https://doi.org/10.1186/S13073-019-0701-3

Bloom, H. J., & Richardson, W. W. (1957). Histological Grading and Prognosis in Breast Cancer. *British Journal of Cancer 1957 11:3*, *11*(3), 359–377. https://doi.org/10.1038/bjc.1957.43

Blusztajn, J. K., Liscovitch, M., & Richardson, U. I. (1987). Synthesis of acetylcholine from choline derived from phosphatidylcholine in a human neuronal cell line. *Proceedings of the National*

Academy of Sciences of the United States of America, *84*(15), 5474–5477. https://doi.org/10.1073/PNAS.84.15.5474

Boire, A., Covic, L., Agarwal, A., Jacques, S., Sherifi, S., & Kuliopulos, A. (2005). PAR1 Is a Matrix Metalloprotease-1 Receptor that Promotes Invasion and Tumorigenesis of Breast Cancer Cells. *Cell*, *120*(3), 303–313. https://doi.org/10.1016/J.CELL.2004.12.018

Bois, A. Du, Reuss, A., Pujade-Lauraine, E., Harter, P., Ray-Coquard, I., & Pfisterer, J. (2009). Role of surgical outcome as prognostic factor in advanced epithelial ovarian cancer: a combined exploratory analysis of 3 prospectively randomized phase 3 multicenter trials: by the Arbeitsgemeinschaft Gynaekologische Onkologie Studiengruppe Ovarialkarzinom (AGO-OVAR) and the Groupe d'Investigateurs Nationaux Pour les Etudes des Cancers de l'Ovaire (GINECO). *Cancer*, *115*(6), 1234–1244. https://doi.org/10.1002/CNCR.24149

Bonnel, D., Franck, J., Mériaux, C., Salzet, M., & Fournier, I. (2013). Ionic matrices pre-spotted matrix-assisted laser desorption/ionization plates for patient maker following in course of treatment, drug titration, and MALDI mass spectrometry imaging. *Analytical Biochemistry*, *434*(1), 187–198. https://doi.org/10.1016/J.AB.2012.10.035

Bonnel, D., Longuespee, R., Franck, J., Roudbaraki, M., Gosset, P., Day, R., Salzet, M., & Fournier, I. (2011). Multivariate analyses for biomarkers hunting and validation through on-tissue bottom-up or in-source decay in MALDI-MSI: application to prostate cancer. *Analytical and Bioanalytical Chemistry*, *401*(1), 149–165. https://doi.org/10.1007/S00216-011-5020-5

Brunelle, A., & Laprévote, O. (2012). MALDI Imaging Mass Spectrometry. *Electrospray and MALDI Mass Spectrometry: Fundamentals, Instrumentation, Practicalities, and Biological Applications: Second Edition*, 245–261. https://doi.org/10.1002/9780470588901.CH8

Bruynseels, K., de Sio, F. S., & van den Hoven, J. (2018). Digital Twins in health care: Ethical implications of an emerging engineering paradigm. *Frontiers in Genetics*, *9*(FEB), 320848. https://doi.org/10.3389/FGENE.2018.00031/BIBTEX

Buchberger, A. R., DeLaney, K., Johnson, J., & Li, L. (2017). Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights. *Analytical Chemistry*, *90*(1), 240. https://doi.org/10.1021/ACS.ANALCHEM.7B04733

Burrell, R. A., McGranahan, N., Bartek, J., & Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, *501*(7467), 338–345. https://doi.org/10.1038/nature12625

Caiado, F., Silva-Santos, B., & Norell, H. (2016). Intra-tumour heterogeneity – going beyond genetics. *FEBS Journal*, *283*, 2245–2258. https://doi.org/10.1111/febs.13705

Calligaris, D., Norton, I., Feldman, D. R., Ide, J. L., Dunn, I. F., Eberlin, L. S., Graham Cooks, R., Jolesz, F. A., Golby, A. J., Santagata, S., & Agar, N. Y. (2013). Mass Spectrometry Imaging as a Tool for Surgical Decision-Making. *Journal of Mass Spectrometry : JMS*, *48*(11), 1178. https://doi.org/10.1002/JMS.3295

Campaner, E., Zannini, A., Santorsola, M., Bonazza, D., Bottin, C., Cancila, V., Tripodo, C., Bortul, M., Zanconati, F., Schoeftner, S., & Del Sal, G. (2020). Breast Cancer Organoids Model Patient-Specific Response to Drug Treatment. *Cancers*, *12*(12), 1–19. https://doi.org/10.3390/CANCERS12123869

Caprioli, R. M., Farmer, T. B., & Gile, J. (1997). Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS. *Analytical Chemistry*, *69*(23), 4751–4760. https://doi.org/10.1021/ac970888i

Ceci, C., Atzori, M. G., Lacal, P. M., & Graziani, G. (2020). Role of VEGFs/VEGFR-1 Signaling and Its Inhibition in Modulating Tumor Invasion: Experimental Evidence in Different Metastatic Cancer Models. *International Journal of Molecular Sciences*, *21*(4). https://doi.org/10.3390/IJMS21041388

Chambers, M. C., MacLean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T. A., Brusniak, M. Y., Paulse, C., Creasy, D., … Mallick, P. (2012). A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology*, *30*(10), 918–920. https://doi.org/10.1038/NBT.2377

Chédotal, A., Kerjan, G., & Moreau-Fauvarque, C. (2005). The brain within the tumor: new roles for axon guidance molecules in cancers. *Cell Death & Differentiation 2005 12:8*, *12*(8), 1044–1056. https://doi.org/10.1038/sj.cdd.4401707

Chi, D. S., Eisenhauer, E. L., Lang, J., Huh, J., Haddad, L., Abu-Rustum, N. R., Sonoda, Y., Levine, D. A., Hensley, M., & Barakat, R. R. (2006). What is the optimal goal of primary cytoreductive surgery for bulky stage IIIC epithelial ovarian carcinoma (EOC)? *Gynecologic Oncology*, *103*(2), 559–564. https://doi.org/10.1016/J.YGYNO.2006.03.051

Cirella, A., Luri-Rey, C., Di Trani, C. A., Teijeira, A., Olivera, I., Bolaños, E., Castañón, E., Palencia, B., Brocco, D., Fernández-Sendin, M., Aranda, F., Berraondo, P., & Melero, I. (2022). Novel strategies exploiting interleukin-12 in cancer immunotherapy. *Pharmacology & Therapeutics*, *239*, 108189. https://doi.org/10.1016/J.PHARMTHERA.2022.108189

Colombo, N., Sessa, C., Du Bois, A., Ledermann, J., McCluggage, W. G., McNeish, I., Morice, P., Pignata, S., Ray-Coquard, I., Vergote, I., Baert, T., Belaroussi, I., Dashora, A., Olbrecht, S., Planchamp, F., & Querleu, D. (2019). ESMO-ESGO consensus conference recommendations on ovarian cancer: pathology and molecular biology, early and advanced stages, borderline tumours and recurrent disease†. *Annals of Oncology : Official Journal of the European Society for Medical Oncology*, *30*(5), 672–705. https://doi.org/10.1093/ANNONC/MDZ062

Cong, L., Sugden, S. M., Leclair, P., Lim, C. J., Pham, T. N. Q., & Cohena, E. A. (2021). HIV-1 Vpu Promotes Phagocytosis of Infected CD4+ T Cells by Macrophages through Downregulation of CD47. *MBio*, *12*(4). https://doi.org/10.1128/MBIO.01920-21

Corral-Acero, J., Margara, F., Marciniak, M., Rodero, C., Loncaric, F., Feng, Y., Gilbert, A., Fernandes, J. F., Bukhari, H. A., Wajdan, A., Martinez, M. V., Santos, M. S., Shamohammdi, M., Luo, H., Westphal, P., Leeson, P., DiAchille, P., Gurev, V., Mayr, M., … Lamata, P. (2020). The "Digital Twin" to enable the vision of precision cardiology. *European Heart Journal*, *41*(48), 4556-4564B. https://doi.org/10.1093/EURHEARTJ/EHAA159

Cortes, C., Mohri, M., & Rostamizadeh, A. (2012). L2 Regularization for Learning Kernels. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009*, 109–116. https://arxiv.org/abs/1205.2653v1

Cowppli-Bony, A., Uhry, Z., Remontet, L., Voirin, N., Guizard, A. V., Trétarre, B., Bouvier, A. M., Colonna, M., Bossard, N., Woronoff, A. S., & Grosclaude, P. (2017). Survival of solid cancer

patients in France, 1989-2013: A population-based study. *European Journal of Cancer Prevention*, *26*(6), 461–468. https://doi.org/10.1097/CEJ.0000000000000372

Croatti, A., Gabellini, M., Montagna, S., & Ricci, A. (2020). On the Integration of Agents and Digital Twins in Healthcare. *Journal of Medical Systems*, *44*(9), 1–8. https://doi.org/10.1007/S10916-020-01623-5/FIGURES/2

Cronin, K. A., Lake, A. J., Scott, S., Sherman, R. L., Noone, A.-M., Howlader, N., Henley, ; S Jane, Anderson, R. N., Firth, A. U., Ma, J., Kohler, B. A., & Jemal, A. (2018). The Authors. Cancer; 6 Surveillance and Health Services Research. *Cancer*, *124*, 2801–2815. https://doi.org/10.1002/cncr.31551

Czajka, M. L., & Pfeifer, C. (2022). Breast Cancer Surgery. *British Journal of Medical Practitioners*, *13*(1), 1. https://doi.org/10.1093/med/9780198839248.003.0013

Dagogo-Jack, I., & Shaw, A. T. (2017). Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology 2017 15:2*, *15*(2), 81–94. https://doi.org/10.1038/nrclinonc.2017.166

de Curtis, I. (2019). The Rac3 GTPase in Neuronal Development, Neurodevelopmental Disorders, and Cancer. *Cells*, *8*(9). https://doi.org/10.3390/CELLS8091063

Dean, L., & Kane, M. (2021). Trastuzumab Therapy and ERBB2 Genotype. *Medical Genetics Summaries*. https://www.ncbi.nlm.nih.gov/books/NBK310376/

Debien, V., De Caluwé, A., Wang, X., Piccart-Gebhart, M., Tuohy, V. K., Romano, E., & Buisseret, L. (2023). Immunotherapy in breast cancer: an overview of current strategies and perspectives. *Npj Breast Cancer 2023 9:1*, *9*(1), 1–10. https://doi.org/10.1038/s41523-023-00508-3

Deininger, S. O., Cornett, D. S., Paape, R., Becker, M., Pineau, C., Rauser, S., Walch, A., & Wolski, E. (2011). Normalization in MALDI-TOF imaging datasets of proteins: practical considerations. *Analytical and Bioanalytical Chemistry*, *401*(1), 167. https://doi.org/10.1007/S00216-011-4929-Z

Delcourt, V., Franck, J., Quanico, J., Gimeno, J. P., Wisztorski, M., Raffo-Romero, A., Kobeissy, F., Roucou, X., Salzet, M., & Fournier, I. (2018). Spatially-Resolved Top-down Proteomics Bridged to MALDI MS Imaging Reveals the Molecular Physiome of Brain Regions. *Molecular & Cellular Proteomics : MCP*, *17*(2), 357–372. https://doi.org/10.1074/MCP.M116.065755

Dewez, F., Oejten, J., Henkel, C., Hebeler, R., Neuweger, H., De Pauw, E., Heeren, R. M. A., & Balluff, B. (2020). MS Imaging-Guided Microproteomics for Spatial Omics on a Single Instrument. *Proteomics*, *20*(23). https://doi.org/10.1002/PMIC.201900369

Dey, A., Ghosh, S., Jha, S., Hazra, S., Srivastava, N., Chakraborty, U., & Roy, A. G. (2023). Recent advancement in breast cancer treatment using CAR T cell therapy:- A review. *Advances in Cancer Biology - Metastasis*, *7*, 100090. https://doi.org/10.1016/J.ADCANC.2023.100090

Dijkstra, T. K. (2014). Ridge regression and its degrees of freedom. *Quality and Quantity*, *48*(6), 3185–3193. https://doi.org/10.1007/S11135-013-9949-7/METRICS

Ding, H., Wang, G., Yu, Z., Sun, H., & Wang, L. (2022). Role of interferon-gamma (IFN-γ) and IFN-γ receptor 1/2 (IFNγR1/2) in regulation of immunity, infection, and cancer development: IFN-γ-dependent or independent pathway. *Biomedicine & Pharmacotherapy*, *155*, 113683. https://doi.org/10.1016/J.BIOPHA.2022.113683

Donnarumma, F., & Murray, K. K. (2016). Laser ablation sample transfer for localized LC-MS/MS proteomic analysis of tissue. *Journal of Mass Spectrometry : JMS*, *51*(4), 261–268. https://doi.org/10.1002/JMS.3744

Drost, J., & Clevers, H. (2018). Organoids in cancer research. *Nature Reviews Cancer 2018 18:7*, *18*(7), 407–418. https://doi.org/10.1038/s41568-018-0007-6

Drucker, H. (1997). Improving Regressors using Boosting Techniques. *International Conference on Machine Learning*. https://www.academia.edu/28038947/Improving_Regressors_using_Boosting_Techniques

Duda, R. O., & Peter, E. (2012). *Pattern Classi fi cation , 2nd Edition*. https://www.librairie-ledivan.com/ebook/9781118586006-pattern-classification-richard-o-duda-peter-e-hart-david-g-stork/

Duhamel, M., Drelich, L., Wisztorski, M., Aboulouard, S., Gimeno, J. P., Ogrinc, N., Devos, P., Cardon, T., Weller, M., Escande, F., Zairi, F., Maurage, C. A., Le Rhun, É., Fournier, I., & Salzet, M. (2022). Spatial analysis of the glioblastoma proteome reveals specific molecular signatures and markers of survival. *Nature Communications*, *13*(1). https://doi.org/10.1038/S41467-022-34208-6

Eberlin, L. S., Norton, I., Orringer, D., Dunn, I. F., Liu, X., Ide, J. L., Jarmusch, A. K., Ligon, K. L., Jolesz, F. A., Golby, A. J., Santagata, S., Agar, N. Y. R., & Cooks, R. G. (2013). Ambient mass spectrometry for the intraoperative molecular diagnosis of human brain tumors. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(5), 1611–1616. https://doi.org/10.1073/PNAS.1215687110

Elsberger, B. (2014). Translational evidence on the role of Src kinase and activated Src kinase in invasive breast cancer. *Critical Reviews in Oncology/Hematology*, *89*(3), 343–351. https://doi.org/10.1016/J.CRITREVONC.2013.12.009

Evangelisti, C., Rusciano, I., Mongiorgi, S., Ramazzotti, G., Lattanzi, G., Manzoli, L., Cocco, L., & Ratti, S. (2022). The wide and growing range of lamin B-related diseases: from laminopathies to cancer. *Cellular and Molecular Life Sciences : CMLS*, *79*(2). https://doi.org/10.1007/S00018-021-04084-2

Famta, P., Shah, S., Dey, B., Kumar, K. C., Bagasariya, D., Vambhurkar, G., Pandey, G., Sharma, A., Srinivasarao, D. A., Kumar, R., Guru, S. K., Raghuvanshi, R. S., & Srivastava, S. (2024). Despicable role of epithelial–mesenchymal transition in breast cancer metastasis: Exhibiting de novo restorative regimens. *Cancer Pathogenesis and Therapy*. https://doi.org/10.1016/J.CPT.2024.01.001

Fatou, B., Saudemont, P., Leblanc, E., Vinatier, D., Mesdag, V., Wisztorski, M., Focsa, C., Salzet, M., Ziskind, M., & Fournier, I. (2016). In vivo Real-Time Mass Spectrometry for Guided Surgery Application. *Scientific Reports*, *6*. https://doi.org/10.1038/srep25919

Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., & Whitehouse, C. M. (1989). Electrospray ionization for mass spectrometry of large biomolecules. *Science (New York, N.Y.)*, *246*(4926), 64–71. https://doi.org/10.1126/SCIENCE.2675315

Firdaus, I. (2020). de Lahunta's Veterinary Neuroanatomy and Clinical Neurology. *Alexander de Lahunta, Eric Glass, Marc Kent*. https://www.academia.edu/49074255/de_Lahunta_s_Veterinary_Neuroanatomy_and_Clinical_Neurology

Fonville, J. M., Carter, C., Cloarec, O., Nicholson, J. K., Lindon, J. C., Bunch, J., & Holmes, E. (2012). Robust data processing and normalization strategy for MALDI mass spectrometric imaging. *Analytical Chemistry*, *84*(3), 1310–1319. https://doi.org/10.1021/AC201767G

Fountzilas, G., Christodoulou, C., Bobos, M., Kotoula, V., Eleftheraki, A. G., Xanthakis, I., Batistatou, A., Pentheroudakis, G., Xiros, N., Papaspirou, I., Koumarianou, A., Papakostas, P., Bafaloukos, D., Skarlos, D. V., & Kalogeras, K. T. (2012). Topoisomerase II alpha gene amplification is a favorable prognostic factor in patients with HER2-positive metastatic breast cancer treated with trastuzumab. *Journal of Translational Medicine*, *10*(1). https://doi.org/10.1186/1479-5876-10-212

Franck, J., Arafah, K., Barnes, A., Wisztorski, M., Salzet, M., & Fournier, I. (2009). Improving tissue preparation for matrix-assisted laser desorption ionization mass spectrometry imaging. Part 1: using microspotting. *Analytical Chemistry*, *81*(19), 8193–8202. https://doi.org/10.1021/AC901328P

Füzéry, A. K., Levin, J., Chan, M. M., & Chan, D. W. (2013). Translation of proteomic biomarkers into FDA approved cancer diagnostics: Issues and challenges. *Clinical Proteomics*, *10*(1), 1–14. https://doi.org/10.1186/1559-0275-10-13/FIGURES/2

Galeano Niño, J. L., Wu, H., LaCourse, K. D., Kempchinsky, A. G., Baryiames, A., Barber, B., Futran, N., Houlton, J., Sather, C., Sicinska, E., Taylor, A., Minot, S. S., Johnston, C. D., & Bullman, S. (2022). Effect of the intratumoral microbiota on spatial and cellular heterogeneity in cancer. *Nature*, *611*(7937), 810–817. https://doi.org/10.1038/S41586-022-05435-0

Gaud, C., Sousa, B. C., Nguyen, A., Fedorova, M., Ni, Z., O'Donnell, V. B., Wakelam, M. J. O., Andrews, S., & Lopez-Clavijo, A. F. (2021). BioPAN: a web-based tool to explore mammalian lipidome metabolic pathways on LIPID MAPS. *F1000Research*, *10*, 1–18. https://doi.org/10.12688/F1000RESEARCH.28022.2/DOI

Gerestein, C. G., Nieuwenhuyzen-de Boer, G. M., Eijkemans, M. J., Kooi, G. S., & Burger, C. W. (2010). Prediction of 30-day morbidity after primary cytoreductive surgery for advanced stage ovarian cancer. *European Journal of Cancer (Oxford, England : 1990)*, *46*(1), 102–109. https://doi.org/10.1016/J.EJCA.2009.10.017

Gianola, S., Savio, T., Schwab, M. E., & Rossi, F. (2003). Cell-autonomous mechanisms and myelin-associated factors contribute to the development of Purkinje axon intracortical plexus in the rat cerebellum. *Journal of Neuroscience*, *23*(11), 4613–4624. https://doi.org/10.1523/JNEUROSCI.23-11-04613.2003

Glaser, P. E., & Gross, R. W. (1995). Rapid Plasmenylethanolamine-Selective Fusion of Membrane Bilayers Catalyzed by an Isoform of Glyceraldehyde-3-Phosphate Dehydrogenase: Discrimination between Glycolytic and Fusogenic Roles of Individual Isoforms. *Biochemistry*, *34*(38), 12193–12203. https://doi.org/10.1021/BI00038A013/ASSET/BI00038A013.FP.PNG_V03

Gonçalves, E., Poulos, R. C., Cai, Z., Barthorpe, S., Manda, S. S., Lucas, N., Beck, A., Bucio-Noble, D., Dausmann, M., Hall, C., Hecker, M., Koh, J., Lightfoot, H., Mahboob, S., Mali, I., Morris, J., Richardson, L., Seneviratne, A. J., Shepherd, R., … Reddel, R. R. (2022). Pan-cancer proteomic map of 949 human cell lines. *Cancer Cell*, *40*(8), 835-849.e8. https://doi.org/10.1016/J.CCELL.2022.06.010

Gredell, D. A., Schroeder, A. R., Belk, K. E., Broeckling, C. D., Heuberger, A. L., Kim, S. Y., King, D. A., Shackelford, S. D., Sharp, J. L., Wheeler, T. L., Woerner, D. R., & Prenni, J. E. (2019). Comparison of Machine Learning Algorithms for Predictive Modeling of Beef Attributes Using Rapid Evaporative Ionization Mass Spectrometry (REIMS) Data. *Scientific Reports*, *9*(1), 5721. https://doi.org/10.1038/S41598-019-40927-6

Grillo, P. K., Győrffy, B., & Götte, M. (2021). Prognostic impact of the glypican family of heparan sulfate proteoglycans on the survival of breast cancer patients. *Journal of Cancer Research and Clinical Oncology*, *147*(7), 1937. https://doi.org/10.1007/S00432-021-03597-4

Gu, Y., Fang, Y., Wu, X., Xu, T., Hu, T., Xu, Y., Ma, P., Wang, Q., & Shu, Y. (2023). The emerging roles of SUMOylation in the tumor microenvironment and therapeutic implications. *Experimental Hematology & Oncology 2023 12:1*, *12*(1), 1–22. https://doi.org/10.1186/S40164-023-00420-3

Hajjaji, N., Aboulouard, S., Cardon, T., Bertin, D., Robin, Y. M., Fournier, I., & Salzet, M. (2022a). Path to Clonal Theranostics in Luminal Breast Cancers. *Frontiers in Oncology*, *11*, 1. https://doi.org/10.3389/FONC.2021.802177/FULL

Hajjaji, N., Aboulouard, S., Cardon, T., Bertin, D., Robin, Y. M., Fournier, I., & Salzet, M. (2022b). Path to Clonal Theranostics in Luminal Breast Cancers. *Frontiers in Oncology*, *11*, 1. https://doi.org/10.3389/FONC.2021.802177/FULL

Hajjaji, N., Aboulouard, S., Cardon, T., Bertin, D., Robin, YM., Fournier, I., & Salzet, M. (2021). Path to drugging functional clones of luminal breast cancers using in-depth proteomics with spatially resolved mass spectrometry guided by MALDI imaging. *MedRxiv*, 2021.02.16.21251694. https://doi.org/10.1101/2021.02.16.21251694

Harvey, R. J. (1980). Cerebellar regulation in movement control. *Trends in Neurosciences*, *3*(11 C), 281–284. https://doi.org/10.1016/0166-2236(80)90102-2

Hastie, T., Rosset, S., Zhu, J., & Zou, H. (2009). Multi-class AdaBoost. *Statistics and Its Interface*, *2*(3), 349–360. https://doi.org/10.4310/SII.2009.V2.N3.A8

Heem Wong, C., Wei Siah, K., & Lo, A. W. (2019). Estimation of clinical trial success rates and related parameters. *Biostatistics*, *20*, 273–286. https://doi.org/10.1093/biostatistics/kxx069

Hernandez-Boussard, T., Macklin, P., Greenspan, E. J., Gryshuk, A. L., Stahlberg, E., Syeda-Mahmood, T., & Shmulevich, I. (2021). Digital twins for predictive oncology will be a paradigm shift for precision cancer care. *Nature Medicine*, *27*(12), 2065–2066. https://doi.org/10.1038/S41591-021-01558-5

Hillenkamp, F., Karas, M., Beavis, R. C., & Chait, B. T. (1991). Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Analytical Chemistry*, *63*(24), 1193 A-1202 A. https://doi.org/10.1021/AC00024A002

Hogg, C., Panir, K., Dhami, P., Rosser, M., Mack, M., Soong, D., Pollard, J. W., Jenkins, S. J., Horne, A. W., & Greaves, E. (2021). Macrophages inhibit and enhance endometriosis depending on their origin. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(6). https://doi.org/10.1073/PNAS.2013776118

Hu, H., & Laskin, J. (2022). Emerging Computational Methods in Mass Spectrometry Imaging. *Advanced Science (Weinheim, Baden-Wurttemberg, Germany)*, *9*(34). https://doi.org/10.1002/ADVS.202203339

Hulet, S. W., Menzies, S., & Connor, J. R. (2002). Ferritin Binding in the Developing Mouse Brain Follows a Pattern Similar to Myelination and Is Unaffected by the Jimpy Mutation. *Developmental Neuroscience*, *24*(2–3), 208–213. https://doi.org/10.1159/000065704

Hulet, S. W., Powers, S., & Connor, J. R. (1999). Distribution of transferrin and ferritin binding in normal and multiple sclerotic human brains. *Journal of the Neurological Sciences*, *165*(1), 48–55. https://doi.org/10.1016/S0022-510X(99)00077-5

Ifa, D. R., & Eberlin, L. S. (2016). Ambient Ionization Mass Spectrometry for Cancer Diagnosis and Surgical Margin Evaluation. *Clinical Chemistry*, *62*(1), 111–123. https://doi.org/10.1373/CLINCHEM.2014.237172

Jardin-Mathé, O., Bonnel, D., Franck, J., Wisztorski, M., Macagno, E., Fournier, I., & Salzet, M. (2008). MITICS (MALDI Imaging Team Imaging Computing System): a new open source mass spectrometry imaging software. *Journal of Proteomics*, *71*(3), 332–345. https://doi.org/10.1016/J.JPROT.2008.07.004

Javed, K., Reddy, V., & Lui, F. (2023). Neuroanatomy, Cerebral Cortex. *StatPearls*. https://www.ncbi.nlm.nih.gov/books/NBK537247/

Jayasingam, S. D., Citartan, M., Thang, T. H., Mat Zin, A. A., Ang, K. C., & Ch'ng, E. S. (2020). Evaluating the Polarization of Tumor-Associated Macrophages Into M1 and M2 Phenotypes in Human Cancer Tissue: Technicalities and Challenges in Routine Clinical Practice. *Frontiers in Oncology*, *9*. https://doi.org/10.3389/FONC.2019.01512

Jiang, Y., Yin, S., Li, K., Luo, H., & Kaynak, O. (2021). Industrial applications of digital twins. *Philosophical Transactions of the Royal Society A*, *379*(2207). https://doi.org/10.1098/RSTA.2020.0360

Journé, F., Body, J. J., Leclercq, G., & Laurent, G. (2008). Hormone therapy for breast cancer, with an emphasis on the pure antiestrogen fulvestrant: Mode of action, antitumor efficacy and effects on bone health. *Expert Opinion on Drug Safety*, *7*(3), 241–258. https://doi.org/10.1517/14740338.7.3.241

Judasz, E., Lisiak, N., Kopczyński, P., Taube, M., & Rubiś, B. (2022). The Role of Telomerase in Breast Cancer's Response to Therapy. *International Journal of Molecular Sciences 2022, Vol. 23, Page 12844*, *23*(21), 12844. https://doi.org/10.3390/IJMS232112844

Kanber, B. M., Smadi, A. Al, Noaman, N. F., Liu, B., Gou, S., & Alsmadi, M. K. (2024). LightGBM: A Leading Force in Breast Cancer Diagnosis Through Machine Learning and Image Processing. *IEEE Access*, *12*, 39811–39832. https://doi.org/10.1109/ACCESS.2024.3375755

Katikireddy, K. R., & O'Sullivan, F. (2011). Immunohistochemical and Immunofluorescence Procedures for Protein Analysis. *Methods in Molecular Biology*, *784*, 155–167. https://doi.org/10.1007/978-1-61779-289-2_11

Kaul, R., Ossai, C., Forkan, A. R. M., Jayaraman, P. P., Zelcer, J., Vaughan, S., & Wickramasinghe, N. (2023). The role of AI for developing digital twins in healthcare: The case of cancer care. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *13*(1), e1480. https://doi.org/10.1002/WIDM.1480

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (n.d.). *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*. Retrieved July 18, 2024, from https://github.com/Microsoft/LightGBM.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, *30*. https://github.com/Microsoft/LightGBM.

Kellogg, M. K., Tikhonova, E. B., & Karamyshev, A. L. (2022). Signal Recognition Particle in Human Diseases. *Frontiers in Genetics*, *13*. https://doi.org/10.3389/FGENE.2022.898083

Kenner, B. J., Go, V. L. W., Chari, S. T., Goldberg, A. E., & Rothschild, L. J. (2017). Early Detection of Pancreatic Cancer: The Role of Industry in the Development of Biomarkers. *Pancreas*, *46*(10), 1238–1241. https://doi.org/10.1097/MPA.0000000000000939

Kertesz, V., Weiskittel, T. M., & Van Berkel, G. J. (2015). An enhanced droplet-based liquid microjunction surface sampling system coupled with HPLC-ESI-MS/MS for spatially resolved analysis. *Analytical and Bioanalytical Chemistry*, *407*(8), 2117–2125. https://doi.org/10.1007/S00216-014-8287-5

Ketkar, N. (2017). Stochastic Gradient Descent. *Deep Learning with Python*, 113–132. https://doi.org/10.1007/978-1-4842-2766-4_8

King, M. E., Yuan, R., Chen, J., Pradhan, K., Sariol, I., Li, S., Chakraborty, A., Ekpenyong, O., Yearley, J. H., Wong, J. C., Zúñiga, L., Tomazela, D., Beaumont, M., Han, J. H., & Eberlin, L. S. (2023). Long-chain polyunsaturated lipids associated with responsiveness to anti-PD-1 therapy are colocalized with immune infiltrates in the tumor microenvironment. *The Journal of Biological Chemistry*, *299*(3). https://doi.org/10.1016/J.JBC.2023.102902

Kirilina, E., Helbling, S., Morawski, M., Pine, K., Reimann, K., Jankuhn, S., Dinse, J., Deistung, A., Reichenbach, J. R., Trampel, R., Geyer, S., Müller, L., Jakubowski, N., Arendt, T., Bazin, P. L., & Weiskopf, N. (2020). Superficial white matter imaging: Contrast mechanisms and whole-brain in vivo mapping. *Science Advances*, *6*(41). https://doi.org/10.1126/SCIADV.AAZ9281

Köbel, M., & Kang, E. Y. (2022). The Evolution of Ovarian Carcinoma Subclassification. *Cancers*, *14*(2). https://doi.org/10.3390/CANCERS14020416

Kohale, I. N., Yu, J., Zhuang, Y., Fan, X., Reddy, R. J., Sinnwell, J., Kalari, K. R., Boughey, J. C., Carter, J. M., Goetz, M. P., Wang, L., & White, F. M. (2022). Identification of Src Family Kinases as Potential Therapeutic Targets for Chemotherapy-Resistant Triple Negative Breast Cancer. *Cancers*, *14*(17). https://doi.org/10.3390/CANCERS14174220

Kohavi, R. (n.d.). *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*. Retrieved November 8, 2023, from http//roboticsStanfordedu/"ronnyk

Koundouros, N., & Poulogiannis, G. (2020). Reprogramming of fatty acid metabolism in cancer. *British Journal of Cancer*, *122*(1), 4–22. https://doi.org/10.1038/S41416-019-0650-Z

Kundu, M., Greer, Y. E., Dine, J. L., & Lipkowitz, S. (2022). Targeting TRAIL Death Receptors in Triple-Negative Breast Cancers: Challenges and Strategies for Cancer Therapy. *Cells*, *11*(23). https://doi.org/10.3390/CELLS11233717

Lambert, A. W., Pattabiraman, D. R., & Weinberg, R. A. (2017). Emerging Biological Principles of Metastasis. *Cell*, *168*(4), 670–691. https://doi.org/10.1016/J.CELL.2016.11.037

Lamont, L., Baumert, M., Ogrinc Potočnik, N., Allen, M., Vreeken, R., Heeren, R. M. A., & Porta, T. (2017). Integration of Ion Mobility MSE after Fully Automated, Online, High-Resolution Liquid Extraction Surface Analysis Micro-Liquid Chromatography. *Analytical Chemistry*, *89*(20), 11143–11150. https://doi.org/10.1021/ACS.ANALCHEM.7B03512

Lankry, D., Rovis, T. L., Jonjic, S., & Mandelboim, O. (2013). The interaction between CD300a and phosphatidylserine inhibits tumor cell killing by NK cells. *European Journal of Immunology*, *43*(8), 2151–2161. https://doi.org/10.1002/EJI.201343433

Le Doussal, V., Tubiana-Hulin, M., Friedman, S., Hacene, K., Spyratos, F., & Brunet, M. (1989). Prognostic value of histologic grade nuclear components of Scarff-Bloom-Richardson (SBR). An improved score modification based on a multivariate analysis of 1262 invasive ductal breast carcinomas. *Cancer*, *64*(9), 1914–1921. https://doi.org/10.1002/1097-0142(19891101)64:9

Ledoux, L., Zirem, Y., Renaud, F., Duponchel, L., Salzet, M., Ogrinc, N., & Fournier, I. (2023). Comparing MS imaging of lipids by WALDI and MALDI: two technologies for evaluating a common ground truth in MS imaging. *The Analyst*, *148*(20), 4982–4986. https://doi.org/10.1039/D3AN01096A

Lee, J., Musyimi, H. K., Soper, S. A., & Murray, K. K. (2008). Development of an automated digestion and droplet deposition microfluidic chip for MALDI-TOF MS. *Journal of the American Society for Mass Spectrometry*, *19*(7), 964–972. https://doi.org/10.1016/J.JASMS.2008.03.015

Lemaire, R., Stauber, J., Wisztorski, M., Van Camp, C., Desmons, A., Deschamps, M., Proess, G., Rudlof, I., Woods, A. S., Day, R., Salzet, M., & Fournier, I. (2007). Tag-mass: specific molecular imaging of transcriptome and proteome by mass spectrometry based on photocleavable tag. *Journal of Proteome Research*, *6*(6), 2057–2067. https://doi.org/10.1021/PR0700044

Leuschner, J., Schmidt, M., Fernsel, P., Lachmund, D., Boskamp, T., & Maass, P. (2019). Supervised non-negative matrix factorization methods for MALDI imaging applications. *Bioinformatics (Oxford, England)*, *35*(11), 1940–1947. https://doi.org/10.1093/BIOINFORMATICS/BTY909

Lheureux, S., Braunstein, M., & Oza, A. M. (2019). Epithelial ovarian cancer: Evolution of management in the era of precision medicine. *CA: A Cancer Journal for Clinicians*, *69*(4), 280–304. https://doi.org/10.3322/CAAC.21559

Li, S., Sampson, C., Liu, C., Piao, H. long, & Liu, H. X. (2023). Integrin signaling in cancer: bidirectional mechanisms and therapeutic opportunities. *Cell Communication and Signaling 2023 21:1*, *21*(1), 1–19. https://doi.org/10.1186/S12964-023-01264-4

Li, Y., Tang, P., Cai, S., Peng, J., & Hua, G. (2020). Organoid based personalized medicine: from bench to bedside. *Cell Regeneration 2020 9:1*, *9*(1), 1–33. https://doi.org/10.1186/S13619-020-00059-Z

Li, Z. H., Li, B., Zhang, X. Y., & Zhu, J. N. (2024). Neuropeptides and Their Roles in the Cerebellum. *International Journal of Molecular Sciences*, *25*(4). https://doi.org/10.3390/IJMS25042332

Lisio, M. A., Fu, L., Goyeneche, A., Gao, Z. H., & Telleria, C. (2019). High-Grade Serous Ovarian Cancer: Basic Sciences, Clinical and Therapeutic Standpoints. *International Journal of Molecular Sciences*, *20*(4). https://doi.org/10.3390/IJMS20040952

Liu, F., Wu, Q., Dong, Z., & Liu, K. (2023). Integrins in cancer: Emerging mechanisms and therapeutic opportunities. *Pharmacology & Therapeutics*, *247*, 108458. https://doi.org/10.1016/J.PHARMTHERA.2023.108458

Lohner, K. (1996). Is the high propensity of ethanolamine plasmalogens to form non-lamellar lipid structures manifested in the properties of biomembranes? *Chemistry and Physics of Lipids*, *81*(2), 167–184. https://doi.org/10.1016/0009-3084(96)02580-7

Long, Y., Tang, L., Zhou, Y., Zhao, S., & Zhu, H. (2023). Causal relationship between gut microbiota and cancers: a two-sample Mendelian randomisation study. *BMC Medicine*, *21*(1), 1–14. https://doi.org/10.1186/S12916-023-02761-6/FIGURES/6

Luo, J., Zou, H., Guo, Y., Tong, T., Ye, L., Zhu, C., Deng, L., Wang, B., Pan, Y., & Li, P. (2022). SRC kinase-mediated signaling pathways and targeted therapies in breast cancer. *Breast Cancer Research : BCR*, *24*(1). https://doi.org/10.1186/S13058-022-01596-Y

Ma, M., Futia, G. L., de Souza, F. M. S., Ozbay, B. N., Llano, I., Gibson, E. A., & Restrepo, D. (2020). Molecular layer interneurons in the cerebellum encode for valence in associative learning. *Nature Communications 2020 11:1*, *11*(1), 1–16. https://doi.org/10.1038/s41467-020-18034-2

Mallah, K., Quanico, J., Raffo-Romero, A., Cardon, T., Aboulouard, S., Devos, D., Kobeissy, F., Zibara, K., Salzet, M., & Fournier, I. (2019). Mapping Spatiotemporal Microproteomics Landscape in Experimental Model of Traumatic Brain Injury Unveils a link to Parkinson's Disease. *Molecular & Cellular Proteomics : MCP*, *18*(8), 1669–1682. https://doi.org/10.1074/MCP.RA119.001604

Mallah, K., Zibara, K., Kerbaj, C., Eid, A., Khoshman, N., Ousseily, Z., Kobeissy, A., Cardon, T., Cizkova, D., Kobeissy, F., Fournier, I., & Salzet, M. (2023). Neurotrauma investigation through spatial omics guided by mass spectrometry imaging: Target identification and clinical applications. *Mass Spectrometry Reviews*, *42*(1), 189–205. https://doi.org/10.1002/MAS.21719

Mantovani, A., Schioppa, T., Porta, C., Allavena, P., & Sica, A. (2006). Role of tumor-associated macrophages in tumor progression and invasion. *Cancer and Metastasis Reviews*, *25*(3), 315–322. https://doi.org/10.1007/S10555-006-9001-7/METRICS

Marcos, P., Hernández-Pérez, C., Weruaga, E., & Díaz, D. (2023). Lobe X of the Cerebellum: A Natural Neuro-Resistant Region. *Anatomia 2023, Vol. 2, Pages 43-62*, *2*(1), 43–62. https://doi.org/10.3390/ANATOMIA2010005

Marine, J. C., Dawson, S. J., & Dawson, M. A. (2020). Non-genetic mechanisms of therapeutic resistance in cancer. *Nature Reviews Cancer 2020 20:12*, *20*(12), 743–756. https://doi.org/10.1038/s41568-020-00302-4

Marshall, L. J., Triunfol, M., & Seidle, T. (2014). *Organoids: A Preliminary Analysis of Cancer Research Output, Funding and Human Health Impact in*. *10*. https://doi.org/10.3390/ani10101923

Meier, F., Brunner, A. D., Frank, M., Ha, A., Bludau, I., Voytik, E., Kaspar-Schoenefeld, S., Lubeck, M., Raether, O., Bache, N., Aebersold, R., Collins, B. C., Röst, H. L., & Mann, M. (2020). diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nature Methods 2020 17:12*, *17*(12), 1229–1236. https://doi.org/10.1038/s41592-020-00998-0

Meijer, C., Uh, H. W., & el Bouhaddani, S. (2023). Digital Twins in Healthcare: Methodological Challenges and Opportunities. *Journal of Personalized Medicine*, *13*(10). https://doi.org/10.3390/JPM13101522

Meraghni, S., Benaggoune, K., Al Masry, Z., Terrissa, L. S., Devalland, C., & Zerhouni, N. (2021). Towards Digital Twins Driven Breast Cancer Detection. *Lecture Notes in Networks and Systems*, *285*, 87–99. https://doi.org/10.1007/978-3-030-80129-8_7/FIGURES/6

Meriaux, C., Franck, J., Wisztorski, M., Salzet, M., & Fournier, I. (2010). Liquid ionic matrixes for MALDI mass spectrometry imaging of lipids. *Journal of Proteomics*, *73*(6), 1204–1218. https://doi.org/10.1016/J.JPROT.2010.02.010

Mezger, S. T. P., Mingels, A. M. A., Bekers, O., Heeren, R. M. A., & Cillero-Pastor, B. (2021). Mass Spectrometry Spatial-Omics on a Single Conductive Slide. *Analytical Chemistry*, *93*(4), 2527–2533. https://doi.org/10.1021/ACS.ANALCHEM.0C04572

Mordente, A., Meucci, E., Martorana, G. E., & Silvestrini, A. (2015). Cancer biomarkers discovery and validation: State of the art, problems and future perspectives. *Advances in Experimental Medicine and Biology*, *867*, 9–26. https://doi.org/10.1007/978-94-017-7215-0_2

Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S., & Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular Systems Biology*, *7*, 548. https://doi.org/10.1038/MSB.2011.81/SUPPL_FILE/MSB201181.REVIEWER_COMMENTS.PDF

Nardecchia, A., Fabre, C., Cauzid, J., Pelascini, F., Motto-Ros, V., & Duponchel, L. (2020). Detection of minor compounds in complex mineral samples from millions of spectra: A new data analysis strategy in LIBS imaging. *Analytica Chimica Acta*, *1114*, 66–73. https://doi.org/10.1016/J.ACA.2020.04.005

Nejman, D., Livyatan, I., Fuks, G., Gavert, N., Zwang, Y., Geller, L. T., Rotter-Maskowitz, A., Weiser, R., Mallel, G., Gigi, E., Meltser, A., Douglas, G. M., Kamer, I., Gopalakrishnan, V., Dadosh, T., Levin-Zaidman, S., Avnet, S., Atlan, T., Cooper, Z. A., … Straussman, R. (2020). The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science (New York, N.Y.)*, *368*(6494), 973–980. https://doi.org/10.1126/SCIENCE.AAY9189

Neumann, E. K., Djambazova, K. V., Caprioli, R. M., & Spraggins, J. M. (2020). Multimodal Imaging Mass Spectrometry: Next Generation Molecular Mapping in Biology and Medicine. *Journal of the American Society for Mass Spectrometry*, *31*(12), 2401. https://doi.org/10.1021/JASMS.0C00232

Nguyen, T. N. Q., Jeannesson, P., Groh, A., Guenot, D., & Gobinet, C. (2015). Development of a hierarchical double application of crisp cluster validity indices: A proof-of-concept study for automated FTIR spectral histology. *Analyst*, *140*(7), 2439–2448. https://doi.org/10.1039/C4AN01937G

Nijs, M., Smets, T., Waelkens, E., & De Moor, B. (2021). A mathematical comparison of non-negative matrix factorization related methods with practical implications for the analysis of mass spectrometry imaging data. *Rapid Communications in Mass Spectrometry : RCM*, *35*(21). https://doi.org/10.1002/RCM.9181

O'brien, J. S., Sampson, E. L., Brien, O. ', Fillerup, D. L., Mead, J. F., & Lz, J. (1965). *Lipid composition of the normal human brain: gray matter, white matter, and myelin"*. *5*, 329. https://doi.org/10.1016/S0022-2275(20)39619-X

Ogrinc, N., Kruszewski, A., Chaillou, P., Saudemont, P., Lagadec, C., Salzet, M., Duriez, C., & Fournier, I. (2021). Robot-Assisted SpiderMass forIn VivoReal-Time Topography Mass Spectrometry

Imaging. *Analytical Chemistry*, *93*(43), 14383–14391. https://doi.org/10.1021/ACS.ANALCHEM.1C01692

Ogrinc, N., Saudemont, P., Balog, J., Robin, Y. M., Gimeno, J. P., Pascal, Q., Tierny, D., Takats, Z., Salzet, M., & Fournier, I. (2019). Water-assisted laser desorption/ionization mass spectrometry for minimally invasive in vivo and real-time surface analysis using SpiderMass. *Nature Protocols*, *14*(11), 3162–3182. https://doi.org/10.1038/S41596-019-0217-8

Ogrinc, N., Saudemont, P., Takats, Z., Salzet, M., & Fournier, I. (2021a). Cancer Surgery 2.0: Guidance by Real-Time Molecular Technologies. *Trends in Molecular Medicine*, *27*, 602–615. https://doi.org/10.1016/j.molmed.2021.04.001

Ogrinc, N., Saudemont, P., Takats, Z., Salzet, M., & Fournier, I. (2021b). Cancer Surgery 2.0: Guidance by Real-Time Molecular Technologies. *Trends in Molecular Medicine*, *27*(6), 602–615. https://doi.org/10.1016/J.MOLMED.2021.04.001

Pandrangi, S. L., Chittineedi, P., Chikati, R., Mosquera, J. A. N., Llaguno, S. N. S., Mohiddin, G. J., Lanka, S., Chalumuri, S. S., & Maddu, N. (2022). Role of Lipoproteins in the Pathophysiology of Breast Cancer. *Membranes*, *12*(5). https://doi.org/10.3390/MEMBRANES12050532

Parra, E. R., Uraoka, N., Jiang, M., Cook, P., Gibbons, D., Forget, M. A., Bernatchez, C., Haymaker, C., Wistuba, I. I., & Rodriguez-Canales, J. (2017). Validation of multiplex immunofluorescence panels using multispectral microscopy for immune-profiling of formalin-fixed and paraffin-embedded human tumor tissues. *Scientific Reports 2017 7:1*, *7*(1), 1–11. https://doi.org/10.1038/s41598-017-13942-8

Peiró, G., Ortiz-Martínez, F., Gallardo, A., Pérez-Balaguer, A., Sánchez-Payá, J., Ponce, J. J., Tibau, A., López-Vilaro, L., Escuin, D., Adrover, E., Barnadas, A., & Lerma, E. (2014). Src, a potential target for overcoming trastuzumab resistance in HER2-positive breast carcinoma. *British Journal of Cancer*, *111*(4), 689. https://doi.org/10.1038/BJC.2014.327

Pogrebniak, K. L., & Curtis, C. (2018). Harnessing tumor evolution to circumvent resistance. *Trends in Genetics : TIG*, *34*(8), 639. https://doi.org/10.1016/J.TIG.2018.05.007

Polyak, K. (2007). Breast cancer: origins and evolution. *The Journal of Clinical Investigation*, *117*(11), 3155. https://doi.org/10.1172/JCI33295

Pöyry, S., & Vattulainen, I. (2016). Role of charged lipids in membrane structures - Insight given by simulations. *Biochimica et Biophysica Acta*, *1858*(10), 2322–2333. https://doi.org/10.1016/J.BBAMEM.2016.03.016

Psilopatis, I., Souferi-Chronopoulou, E., Vrettou, K., Troungos, C., & Theocharis, S. (2022). EPH/Ephrin-Targeting Treatment in Breast Cancer: A New Chapter in Breast Cancer Therapy. *International Journal of Molecular Sciences*, *23*(23). https://doi.org/10.3390/IJMS232315275

Putta, P., Rankenberg, J., Korver, R. A., van Wijk, R., Munnik, T., Testerink, C., & Kooijman, E. E. (2016). Phosphatidic acid binding proteins display differential binding as a function of membrane curvature stress and chemical properties. *Biochimica et Biophysica Acta*, *1858*(11), 2709–2716. https://doi.org/10.1016/J.BBAMEM.2016.07.014

Quail, D. F., & Joyce, J. A. (2013). Microenvironmental regulation of tumor progression and metastasis. *Nature Medicine*, *19*(11), 1423–1437. https://doi.org/10.1038/NM.3394

Quanico, J., Franck, J., Cardon, T., Leblanc, E., Wisztorski, M., Salzet, M., & Fournier, I. (2017). NanoLC-MS coupling of liquid microjunction microextraction for on-tissue proteomic analysis. *Biochimica et Biophysica Acta. Proteins and Proteomics*, *1865*(7), 891–900. https://doi.org/10.1016/J.BBAPAP.2016.11.002

Quanico, J., Franck, J., Dauly, C., Strupat, K., Dupuy, J., Day, R., Salzet, M., Fournier, I., & Wisztorski, M. (2013). Development of liquid microjunction extraction strategy for improving protein identification from tissue sections. *Journal of Proteomics*, *79*, 200–218. https://doi.org/10.1016/J.JPROT.2012.11.025

Quanico, J., Franck, J., Wisztorski, M., Salzet, M., & Fournier, I. (2017). Integrated mass spectrometry imaging and omics workflows on the same tissue section using grid-aided, parafilm-assisted microdissection. *Biochimica et Biophysica Acta. General Subjects*, *1861*(7), 1702–1714. https://doi.org/10.1016/J.BBAGEN.2017.03.006

Raffo-Romero, A., Aboulouard, S., Bouchaert, E., Rybicka, A., Tierny, D., Hajjaji, N., Fournier, I., Salzet, M., & Duhamel, M. (2023). Establishment and characterization of canine mammary tumoroids for translational research. *BMC Biology*, *21*(1). https://doi.org/10.1186/S12915-023-01516-2

Rahman, M., Pumphrey, J. G., & Lipkowitz, S. (2009). The TRAIL to targeted therapy of breast cancer. *Advances in Cancer Research*, *103*(C), 43. https://doi.org/10.1016/S0065-230X(09)03003-6

Ratnavelu, N. D., Brown, A. P., Mallett, S., Scholten, R. J., Patel, A., Founta, C., Galaal, K., Cross, P., & Naik, R. (2016). Intraoperative frozen section analysis for the diagnosis of early stage ovarian cancer in suspicious pelvic masses. *The Cochrane Database of Systematic Reviews*, *3*(3). https://doi.org/10.1002/14651858.CD010360.PUB2

Römpp, A., Schramm, T., Hester, A., Klinkert, I., Both, J. P., Heeren, R. M. A., Stöckli, M., & Spengler, B. (2011a). imzML: Imaging Mass Spectrometry Markup Language: A common data format for mass spectrometry imaging. *Methods in Molecular Biology (Clifton, N.J.)*, *696*, 205–224. https://doi.org/10.1007/978-1-60761-987-1_12

Römpp, A., Schramm, T., Hester, A., Klinkert, I., Both, J. P., Heeren, R. M. A., Stöckli, M., & Spengler, B. (2011b). imzML: Imaging Mass Spectrometry Markup Language: A Common Data Format for Mass Spectrometry Imaging. *Methods in Molecular Biology*, *696*, 205–224. https://doi.org/10.1007/978-1-60761-987-1_12/COVER

Rouhollahi, A., Willi, J. N., Haltmeier, S., Mehrtash, A., Straughan, R., Javadikasgari, H., Brown, J., Itoh, A., de la Cruz, K. I., Aikawa, E., Edelman, E. R., & Nezami, F. R. (2023). CardioVision: A fully automated deep learning package for medical image segmentation and reconstruction generating digital twins for patients with aortic stenosis. *Computerized Medical Imaging and Graphics : The Official Journal of the Computerized Medical Imaging Society*, *109*. https://doi.org/10.1016/J.COMPMEDIMAG.2023.102289

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J., & Neumann, S. (2016). MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *Journal of Cheminformatics*, *8*(1), 3. https://doi.org/10.1186/S13321-016-0115-9

Sans, M., Gharpure, K., Tibshirani, R., Zhang, J., Liang, L., Liu, J., Young, J. H., Dood, R. L., Sood, A. K., & Eberlin, L. S. (2017). Metabolic Markers and Statistical Prediction of Serous Ovarian Cancer

Aggressiveness by Ambient Ionization Mass Spectrometry Imaging. *Cancer Research*, *77*(11), 2903–2913. https://doi.org/10.1158/0008-5472.CAN-16-3044

Sans, M., Zhang, J., Lin, J. Q., Feider, C. L., Giese, N., Breen, M. T., Sebastian, K., Liu, J., Sood, A. K., & Eberlin, L. S. (2019). Performance of the MasSpec Pen for Rapid Diagnosis of Ovarian Cancer. *Clinical Chemistry*, *65*(5), 674–683. https://doi.org/10.1373/CLINCHEM.2018.299289

Saudemont, P., Quanico, J., Robin, Y. M., Baud, A., Balog, J., Fatou, B., Tierny, D., Pascal, Q., Minier, K., Pottier, M., Focsa, C., Ziskind, M., Takats, Z., Salzet, M., & Fournier, I. (2018). Real-Time Molecular Diagnosis of Tumors Using Water-Assisted Laser Desorption/Ionization Mass Spectrometry Technology. *Cancer Cell*, *34*(5), 840-851.e4. https://doi.org/10.1016/J.CCELL.2018.09.009

Schönthal, A. H., Marino, S., Menna, G., Di Bonaventura, R., Lisi, L., Mattogno, P., Figà, F., Bilgin, L., Giorgio D'alessandris, Q., Olivi, A., & Della Pepa, G. M. (2023). *The Extracellular Matrix in Glioblastomas: A Glance at Its Structural Modifications in Shaping the Tumoral Microenvironment-A Systematic Review*. https://doi.org/10.3390/cancers15061879

Servin, F., Collins, J. A., Heiselman, J. S., Frederick-Dyer, K. C., Planz, V. B., Geevarghese, S. K., Brown, D. B., Jarnagin, W. R., & Miga, M. I. (2023). Simulation of Image-Guided Microwave Ablation Therapy Using a Digital Twin Computational Model. *IEEE Open Journal of Engineering in Medicine and Biology*, *5*, 107. https://doi.org/10.1109/OJEMB.2023.3345733

Shafto, M., Rich, M. C., Glaessgen, D. E., Kemp, C., Lemoigne, J., & Wang, L. (2010). *DRAFT MoDeling, SiMulATion, inFoRMATion Technology & PRoceSSing RoADMAP Technology Area 11*.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, *13*(11), 2498. https://doi.org/10.1101/GR.1239303

Shehata, M., Grimm, L., Ballantyne, N., Lourenco, A., Demello, L. R., Kilgore, M. R., & Rahbar, H. (2019). Ductal Carcinoma in Situ: Current Concepts in Biology, Imaging, and Treatment. *Journal of Breast Imaging*, *1*(3), 166–176. https://doi.org/10.1093/JBI/WBZ039

Sheils, T., Mathias, S. L., Siramshetty, V. B., Bocci, G., Bologa, C. G., Yang, J. J., Waller, A., Southall, N., Nguyen, D. T., & Oprea, T. I. (2020). How to Illuminate the Druggable Genome using Pharos. *Current Protocols in Bioinformatics*, *69*(1), e92. https://doi.org/10.1002/CPBI.92

Shu, H., Liang, R., Li, Z., Goodridge, A., Zhang, X., Ding, H., Nagururu, N., Sahu, M., Creighton, F. X., Taylor, R. H., Munawar, A., & Unberath, M. (2023). Twin-S: a digital twin for skull base surgery. *International Journal of Computer Assisted Radiology and Surgery*, *18*(6), 1077–1084. https://doi.org/10.1007/S11548-023-02863-9

Shulgin, A. A., Lebedev, T. D., Prassolov, V. S., & Spirin, P. V. (2021). Plasmolipin and Its Role in Cell Processes. *Molecular Biology*, *55*(6), 773–785. https://doi.org/10.1134/S0026893321050113

Sinha, I., Fogle, R. L., Gulfidan, G., Stanley, A. E., Walter, V., Hollenbeak, C. S., Arga, K. Y., & Sinha, R. (2023). Potential Early Markers for Breast Cancer: A Proteomic Approach Comparing Saliva and Serum Samples in a Pilot Study. *International Journal of Molecular Sciences*, *24*(4), 4164. https://doi.org/10.3390/IJMS24044164

Sinha, S., Vegesna, R., Mukherjee, S., Kammula, A. V., Dhruba, S. R., Wu, W., Kerr, D. L., Nair, N. U., Jones, M. G., Yosef, N., Stroganov, O. V., Grishagin, I., Aldape, K. D., Blakely, C. M., Jiang, P.,

Thomas, C. J., Benes, C. H., Bivona, T. G., Schäffer, A. A., & Ruppin, E. (2024). PERCEPTION predicts patient response and resistance to treatment using single-cell transcriptomics of their tumors. *Nature Cancer*, *5*(6), 938–952. https://doi.org/10.1038/S43018-024-00756-7

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lønning, P. E., & Børresen-Dale, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(19), 10869–10874. https://doi.org/10.1073/PNAS.191367098

Stack, E. C., Wang, C., Roman, K. A., & Hoyt, C. C. (2014). Multiplexed immunohistochemistry, imaging, and quantitation: A review, with an assessment of Tyramide signal amplification, multispectral imaging and multiplex analysis. *Methods*, *70*(1), 46–58. https://doi.org/10.1016/J.YMETH.2014.08.016

Stauber, J., Ayed, M. El, Wisztorski, M., Salzet, M., & Fournier, I. (2010). Specific MALDI-MSI: Tag-Mass. *Methods in Molecular Biology (Clifton, N.J.)*, *656*, 339–361. https://doi.org/10.1007/978-1-60761-746-4_20

Stengel, C., Newman, S. P., Leese, M. P., Potter, B. V. L., Reed, M. J., & Purohit, A. (2009). Class III β-tubulin expression and in vitro resistance to microtubule targeting agents. *British Journal of Cancer 2010 102:2*, *102*(2), 316–324. https://doi.org/10.1038/sj.bjc.6605489

Stoeckli, M., Farmer, T. B., & Caprioli, R. M. (1999). Automated mass spectrometry imaging with a matrix-assisted laser desorption ionization time-of-flight instrument. *Journal of the American Society for Mass Spectrometry*, *10*(1), 67–71. https://doi.org/10.1016/S1044-0305(98)00126-3

Stoica, C., Ferreira, A. K., Hannan, K., & Bakovic, M. (2022). Bilayer Forming Phospholipids as Targets for Cancer Therapy. *International Journal of Molecular Sciences*, *23*(9). https://doi.org/10.3390/IJMS23095266

Sun, C., Wang, A., Zhou, Y., Chen, P., Wang, X., Huang, J., Gao, J., Wang, X., Shu, L., Lu, J., Dai, W., Bu, Z., Ji, J., & He, J. (2023). Spatially resolved multi-omics highlights cell-specific metabolic remodeling and interactions in gastric cancer. *Nature Communications*, *14*(1). https://doi.org/10.1038/S41467-023-38360-5

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209–249. https://doi.org/10.3322/CAAC.21660

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J., & Von Mering, C. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, *43*(Database issue), D447–D452. https://doi.org/10.1093/NAR/GKU1003

Tame, M. A., Manjón, A. G., Belokhvostova, D., Raaijmakers, J. A., & Medema, R. H. (2017). TUBB3 overexpression has a negligible effect on the sensitivity to taxol in cultured cell lines. *Oncotarget*, *8*(42), 71536–71547. https://doi.org/10.18632/ONCOTARGET.17740

Tanioka, M., Sakai, K., Sudo, T., Sakuma, T., Kajimoto, K., Hirokaga, K., Takao, S., Negoro, S., Minami, H., Nakagawa, K., & Nishio, K. (2014). Transcriptional CCND1 expression as a predictor of poor response to neoadjuvant chemotherapy with trastuzumab in HER2-positive/ER-positive breast

cancer. *Breast Cancer Research and Treatment*, *147*(3), 513–525. https://doi.org/10.1007/S10549-014-3121-5

Tardini, E., Zhang, X., Canahuate, G., Wentzel, A., Mohamed, A. S. R., Van Dijk, L., Fuller, C. D., & Marai, G. E. (2022). Optimal Treatment Selection in Sequential Systemic and Locoregional Therapy of Oropharyngeal Squamous Carcinomas: Deep Q-Learning With a Patient-Physician Digital Twin Dyad. *Journal of Medical Internet Research*, *24*(4). https://doi.org/10.2196/29455

Tonissen, K. F., & Poulsen, S. A. (2021). Carbonic anhydrase XII inhibition overcomes P-glycoprotein-mediated drug resistance: a potential new combination therapy in cancer. *Cancer Drug Resistance*, *4*(2), 343. https://doi.org/10.20517/CDR.2020.110

Tracey, T. J., Kirk, S. E., Steyn, F. J., & Ngo, S. T. (2021). The role of lipids in the central nervous system and their pathological implications in amyotrophic lateral sclerosis. *Seminars in Cell and Developmental Biology*, *112*(June), 69–81. https://doi.org/10.1016/j.semcdb.2020.08.012

Trim, P. J., Atkinson, S. J., Princivalle, A. P., Marshall, P. S., West, A., & Clench, M. R. (2008). Matrix-assisted laser desorption/ionisation mass spectrometry imaging of lipids in rat brain tissue with integrated unsupervised and supervised multivariant statistical analysis. *Rapid Communications in Mass Spectrometry : RCM*, *22*(10), 1503–1509. https://doi.org/10.1002/RCM.3498

Tsurui, H., Nishimura, H., Hattori, S., Hirose, S., Okumura, K., & Shirai, T. (2000). Seven-color fluorescence imaging of tissue samples based on fourier spectroscopy and singular value decomposition. *Journal of Histochemistry and Cytochemistry*, *48*(5), 653–662. https://doi.org/10.1177/002215540004800509/ASSET/IMAGES/LARGE/10.1177_002215540004800509-FIG6.JPEG

Tzafetas, M., Mitra, A., Paraskevaidi, M., Bodai, Z., Kalliala, I., Bowden, S., Lathouras, K., Rosini, F., Szasz, M., Savage, A., Balog, J., McKenzie, J., Lyons, D., Bennett, P., MacIntyre, D., Ghaem-Maghami, S., Takats, Z., & Kyrgiou, M. (2020). The intelligent knife (iKnife) and its intraoperative diagnostic advantage for the treatment of cervical disease. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(13), 7338–7346. https://doi.org/10.1073/PNAS.1916960117/-/DCSUPPLEMENTAL

Urbanelli, L., Buratta, S., Logozzi, M., Mitro, N., Sagini, K., Raimo, R. Di, Caruso, D., Fais, S., & Emiliani, C. (2020). Lipidomic analysis of cancer cells cultivated at acidic pH reveals phospholipid fatty acids remodelling associated with transcriptional reprogramming. *Journal of Enzyme Inhibition and Medicinal Chemistry*, *35*(1), 963. https://doi.org/10.1080/14756366.2020.1748025

Vaysse, P. M., Heeren, R. M. A., Porta, T., & Balluff, B. (2017). Mass spectrometry imaging for clinical research - latest developments, applications, and current limitations. *The Analyst*, *142*(15), 2690–2712. https://doi.org/10.1039/C7AN00565B

Wang, Z., Zhang, Y., Tian, R., Luo, Z., Zhang, R., Li, X., & Abliz, Z. (2022). Data-Driven Deciphering of Latent Lesions in Heterogeneous Tissue Using Function-Directed t-SNE of Mass Spectrometry Imaging Data. *Analytical Chemistry*, *94*(40), 13927–13935. https://doi.org/10.1021/ACS.ANALCHEM.2C02990

Wind, N. S., & Holen, I. (2011). Multidrug resistance in breast cancer: from in vitro models to clinical studies. *International Journal of Breast Cancer*, *2011*, 1–12. https://doi.org/10.4061/2011/967419

Wisztorski, M., Aboulouard, S., Roussel, L., Duhamel, M., Saudemont, P., Cardon, T., Narducci, F., Robin, Y. M., Lemaire, A. S., Bertin, D., Hajjaji, N., Kobeissy, F., Leblanc, E., Fournier, I., & Salzet, M. (2023). Fallopian tube lesions as potential precursors of early ovarian cancer: a comprehensive proteomic analysis. *Cell Death & Disease*, *14*(9). https://doi.org/10.1038/S41419-023-06165-5

Wisztorski, M., Desmons, A., Quanico, J., Fatou, B., Gimeno, J.-P., Franck, J., Salzet, M., & Fournier, I. (2016). Spatially-resolved protein surface microsampling from tissue sections using liquid extraction surface analysis. *PROTEOMICS*, *16*(11–12), 1622–1632. https://doi.org/10.1002/pmic.201500508

Wisztorski, M., Fatou, B., Franck, J., Desmons, A., Farré, I., Leblanc, E., Fournier, I., & Salzet, M. (2013). Microproteomics by liquid extraction surface analysis: application to FFPE tissue to study the fimbria region of tubo-ovarian cancer. *Proteomics. Clinical Applications*, *7*(3–4), 234–240. https://doi.org/10.1002/PRCA.201200070

Wu, C., Lorenzo, G., Hormuth, D. A., Lima, E. A. B. F., Slavkova, K. P., DiCarlo, J. C., Virostko, J., Phillips, C. M., Patt, D., Chung, C., & Yankeelov, T. E. (2022). Integrating mechanism-based modeling with biomedical imaging to build practical digital twins for clinical oncology. *Biophysics Reviews*, *3*(2). https://doi.org/10.1063/5.0086789

Wu, Y., Liu, Q., & Xie, L. (2023). Hierarchical multi-omics data integration and modeling predict cell-specific chemical proteomics and drug responses. *Cell Reports Methods*, *3*(4), 100452. https://doi.org/10.1016/j.crmeth.2023.100452

Xu, Y., & Goldkorn, A. (2016). Telomere and Telomerase Therapeutics in Cancer. *Genes*, *7*(6). https://doi.org/10.3390/GENES7060022

Yagnik, G., Liu, Z., Rothschild, K. J., & Lim, M. J. (2021). *Highly Multiplexed Immunohistochemical MALDI-MS Imaging of Biomarkers in Tissues*. https://doi.org/10.1021/jasms.0c00473

Yamazaki, Y., Hikishima, K., Saiki, M., Inada, M., Sasaki, E., Lemon, R. N., Price, C. J., Okano, H., & Iriki, A. (2016). Neural changes in the primate brain correlated with the evolution of complex motor skills. *Scientific Reports*, *6*. https://doi.org/10.1038/SREP31084

Yates, L. R. (2017). Intratumoral heterogeneity and subclonal diversification of early breast cancer. *Breast*, *34*, S36–S42. https://doi.org/10.1016/j.breast.2017.06.025

Yeom, M., Park, J., Lim, C., Sur, B., Lee, B., Han, J. J., Choi, H. D., Lee, H., & Hahm, D. H. (2015). Glucosylceramide attenuates the inflammatory mediator expression in lipopolysaccharide-stimulated RAW264.7 cells. *Nutrition Research (New York, N.Y.)*, *35*(3), 241–250. https://doi.org/10.1016/J.NUTRES.2015.01.001

Zhang, C., Wang, Y., Wang, F., Wang, Z., Lu, Y., Xu, Y., Wang, K., Shen, H., Yang, P., Li, S., Qin, X., & Yu, H. (2017). Quantitative profiling of glycerophospholipids during mouse and human macrophage differentiation using targeted mass spectrometry. *Scientific Reports*, *7*(1). https://doi.org/10.1038/S41598-017-00341-2

Zhang, Z., Huang, L., & Brayboy, L. (2021). Macrophages: an indispensable piece of ovarian health. *Biology of Reproduction*, *104*(3), 527–538. https://doi.org/10.1093/BIOLRE/IOAA219

Zirem, Y., Ledoux, L., Roussel, L., Maurage, C. A., Tirilly, P., Le Rhun, É., Meresse, B., Yagnik, G., Lim, M. J., Rothschild, K. J., Duhamel, M., Salzet, M., & Fournier, I. (2024). Real-time glioblastoma

tumor microenvironment assessment by SpiderMass for improved patient management. *Cell Reports. Medicine*, 101482. https://doi.org/10.1016/J.XCRM.2024.101482

Journal Pre-proof

Heterogeneity Assessment and Protein Pathway Prediction via Spatial Lipidomic and Proteomic Correlation: Advancing Dry Proteomics concept for Human Glioblastoma Prognosis

Laurine Lagache, Yanis Zirem, Émilie Le Rhun, Isabelle Fournier, Michel Salzet

PII: S1535-9476(24)00181-6

DOI: https://doi.org/10.1016/j.mcpro.2024.100891

Reference: MCPRO 100891

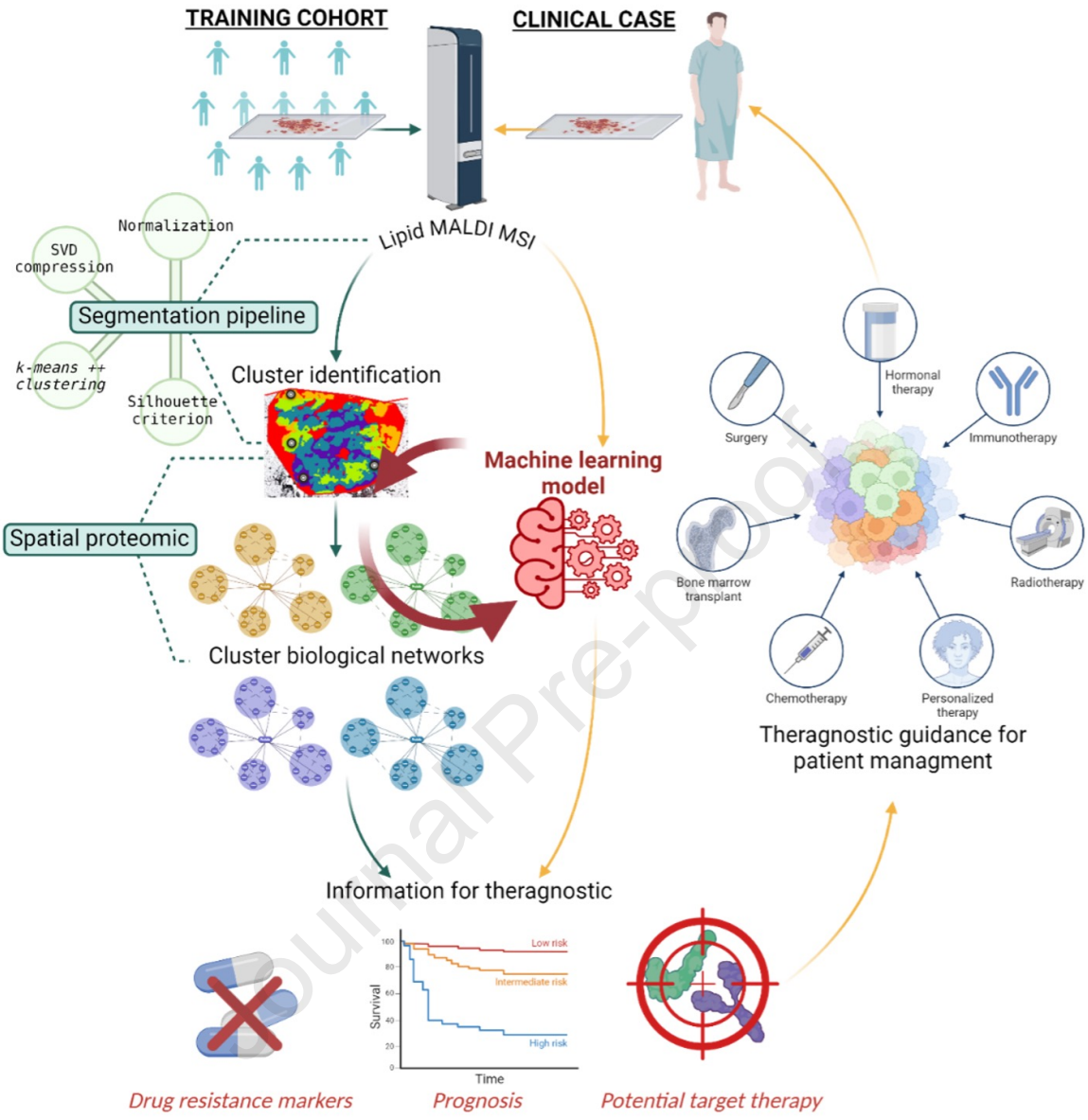
To appear in: *Molecular & Cellular Proteomics*

Received Date: 9 May 2024

Revised Date: 20 November 2024

Accepted Date: 4 December 2024

Please cite this article as: Lagache L, Zirem Y, Le Rhun É, Fournier I, Salzet M, Heterogeneity Assessment and Protein Pathway Prediction via Spatial Lipidomic and Proteomic Correlation: Advancing Dry Proteomics concept for Human Glioblastoma Prognosis, *Molecular & Cellular Proteomics* (2025), doi: https://doi.org/10.1016/j.mcpro.2024.100891.



Heterogeneity Assessment and Protein Pathway Prediction via Spatial Lipidomic and Proteomic Correlation: Advancing Dry Proteomics concept for Human Glioblastoma Prognosis

Laurine Lagache[1†], Yanis Zirem[1†], Émilie Le Rhun[1,2],

Isabelle Fournier[1,3*†] and Michel Salzet[1,3*†]

[1]Univ.Lille, Inserm, CHU Lille, U1192 – Proteomics Inflammatory Response Mass Spectrometry- PRISM, F-59000 Lille, France

[2]Department of Neurosurgery and Neurology, Clinical Neuroscience Center, University Hospital Zurich and University of Zurich, Zurich, Switzerland

[3]Institut Universitaire de France, 75000 Paris

†Equal contribution

*Co-corresponding

RUNNING TITLE: Heterogeneity Assessment and Protein Pathway Prediction via Spatial Lipidomic and Proteomic Correlation: Advancing Dry Proteomics concept for Human Glioblastoma Prognosis

CORRESPONDING AUTHORS

Prof. Michel Salzet and Prof. Isabelle Fournier. Laboratoire Protéomique, Réponse Inflammatoire et Spectrométrie de Masse (PRISM) - Inserm U1192, Bât SN3, 1er étage, Campus Cité Scientifique, Université de Lille, F-59655 Villeneuve d'Ascq Cedex. Phone: +33 (0)320 434 194; email: michel.salzet@univ-lille.fr; email: isabelle.fournier@univ-lille.fr.

ABBREVIATIONS

CID - Collision Induced Dissociation

DG - Diglyceride

FF - Fresh Frozen

GBM - Glioblastoma

GL - Granular Layer

GM - Grey Matter

GO - Gene Ontology

LIME – Local Interpretable Model-agnostic Explanations

LGBM - Light Gradient Boosting Machine

MALDI – Matrix Assisted Laser Desorption/Ionization

ML - Molecular Layer

MSI – Mass Spectrometry Imaging

NNMF - Non-Negative Matrix Factorization

PC - Phosphatidylcholine

PCA - Principal Component Analysis

PE – Phosphatidylethanolamine

RB - Rat Brain

RMS – Root Mean Square

ROIs - Regions of Interest

SGD - Stochastic Gradient Descent

SVD - Singular Value Decomposition

TG - Triglyceride

t-SNE - t-distributed Stochastic Neighbor Embedding

WM - White Matter

ABSTRACT

Prediction of proteins and associated biological pathways from lipid analyses via MALDI MSI is a pressing challenge. We introduced "dry proteomics," using MALDI MSI to validate spatial localization of identified optimal clusters in lipid imaging. Consistent cluster appearance across omics images suggests association with specific lipid and protein in distinct biological pathways, forming the basis of dry proteomics. The methodology was refined using rat brain tissue as a model, then applied to human glioblastoma, a highly heterogeneous cancer. Sequential tissue sections underwent omics MALDI MSI and unsupervised clustering. Spatial omics analysis facilitated lipid and protein characterization, leading to a predictive model identifying clusters in any tissue based on unique lipid signatures and predicting associated protein pathways. Application to rat brain slices revealed diverse tissue subpopulations, including successfully predicted cerebellum areas. Similarly, the methodology was applied to a dataset from a cohort of 50 glioblastoma patients, reused from a previous study. However, among the 50 patients, only 13 lipid signatures from MALDI MSI data were available, allowing for the identification of lipid-protein associations that correlated with patient prognosis. For cases lacking lipid imaging data, a classification model based on protein data was developed from dry proteomic results to effectively categorize the remaining cohort.

Keywords: Dry Proteomics, MALDI MSI multi-omics, Heterogeneity, Cluster estimation, Spatial Proteomics, Machine Learning, Glioblastoma prognosis

INTRODUCTION

Since the gap between mass spectrometry imaging (MSI) and proteomics has been bridged by the development of spatially resolved proteomics guided by MALDI MSI[1–5], the next challenge was to perform multi-omics analyses at the spatial level[6–11]. Nevertheless, there are still developments to be performed to correlate from lipid MSI data, proteins and lipid networks to retrieved functions. Multi-omics MSI is particularly valuable for the analysis of heterogeneous biological samples, such as brain or tumours, which consist of different cell types and regions with distinct molecular composition and function[3]. Indeed, tumour heterogeneity is a significant and growing area in cancer research. An overview on tumoral heterogeneous proteome is subsequently linked to therapeutic, allowing drug resistance analysis and optimized treatment guideline proposal, tending to personalize medicine strategy. However, the complex nature of protein annotation and the lack of standardized methodologies pose challenges to the effectiveness of MALDI-MSI data analysis, especially in multi-omics clinical research. The interpretation and integration of the vast amount of data generated by these technologies remains a significant limitation[12]. Extracting meaningful insights from complex datasets therefore requires sophisticated computational approaches and bioinformatic analysis[13]. MALDI MSI data analysis involves pre-processing and processing stages, preparing them for subsequent statistical analysis. Reduction techniques, like PCA (Principal Component Analysis)[14,15], t-SNE (t-distributed Stochastic Neighbour Embedding)[16,17], or NNMF (Non-negative matrix factorisation)[18,19], are particularly useful for exploring the spatial distribution of molecular features in MALDI MSI data[20,21]. In addition, the combination of MSI and machine learning methods is widely used in the processing step to effectively extract the essential information contained in complex MSI data. The emergence of segmentation methods, such as bisecting *k*-means, hierarchical clustering and *k*-means clustering[22,23], provides valuable insights from complex data like meaningful regions corresponding to biological features in heterogeneous sample. However, choosing the right number of *k*-clusters is not straightforward, limiting biological conclusions[22,23]. The common method involves performing *k*-means clustering for different k values ($2<k<k_{max}$) and calculate the distances between clusters. The aim is to find the optimal *k* that minimizes intra-class distances while maximizing inter-class distances. Several statistical indices, called criterions, have been developed for this purpose[24,25] .

Here, we introduce the concept of dry proteomics, an automated procedure capable of identifying heterogeneous clusters of biological samples according to their lipid signature, thought lipid MALDI MSI, and automatically providing their associated protein data without any proteomic experiments. The development of this machine learning method required overcoming several challenges (see, Graphical Abstract).The central hypothesis was that if a cluster appeared identical in both lipid and protein images, it should possess lipids and paired proteins related to a specific biological pathway, like a unique barcode that allows one cluster to be distinguished from others. Thus, the

correlation between lipids and proteins in a biological network, within different clusters, forms the basis of dry proteomics. The data processing workflow was first developed on lipid, protein, and peptide MSI datasets performed on rat brain (RB) tissue. We succeeded in building a segmentation pipeline, consisting of Singular Value Decomposition (SVD) data compression pre-processing and *k*-means++ segmentation processing steps. The integration of the silhouette criterion allowed to optimize and automate the optimal number of clusters finding for MSI analysis, corresponding to the sample heterogeneity. The next step was to develop a prediction model that could blindly identify the different RB clusters from a lipid MS image according to their spectral fingerprint. The prediction model was complemented by discriminative lipid and protein identifications for each cluster, forming a dry proteomic reference dataset for RB tissue section.

Finally, the dry proteomics concept is a simple and rapid procedure, as the user only needs to perform lipid MALDI MSI to automatically identify the heterogeneous clusters present in a sample and obtain their specific proteome. The development of this tool is aimed at clinical application for patient therapeutic guidance. Indeed, the protein information provided by the dry proteomics process can be related to drug resistance, potential therapeutic target or patient survival, which could help the oncologist to propose a therapeutic guideline adapted to the patient's tumour. In this way, the ultimate phase of presented research involved the application of this innovative concept to intricate and heterogeneous pathology samples, particularly human Glioblastoma [26–28]. In addition, by applying the dry proteomics workflow, correlation between predicted protein and patient survival outcome information allowed to establish a robust model for glioblastoma patient survival prediction. This crucial validation step not only enhances confidence in the reliability of this approach but also holds significant promise for advancing personalized medicine strategies in the management of this challenging disease. Indeed, the assessment of heterogeneity, whether intra or interpatient, is pivotal in personalized medicine, as it allows for the identification of unique molecular profiles that can inform tailored treatment strategies for individual patients.

EXPERIMENTAL PROCEDURES

Experimental Design and Statistical Rationale

For MALDI imaging and spatial omics development studies n = 3 male Wistar rats were sacrificed. All the experiments were performed in biological triplicate to ensure data reproducibility. For the proteomic statistical analysis of conditioned media, as a criterion of significance, we applied an ANOVA significance threshold of p-value ≤ 0.01, and heat maps were generated. Normalization was achieved using a Z-score with matrix access by rows. To assess the statistical significance of biomarkers for lipids MSI biomarkers, a non-parametric Kruskal-Wallis test was employed. Bonferroni corrections were applied to adjust p-values for multiple comparisons. Values are presented as medians and visualized through scatter boxplots.

A retrospective cohort of 50 fresh frozen (FF) glioblastoma tissues was obtained from the Pathology department of Lille Hospital, France. A prospective cohort of 50 FF glioblastoma tissues were also included in this study. 50 patients with newly diagnosed glioblastoma were prospectively enrolled between September 2014 and November 2018 at Lille University Hospital, France (NCT02473484). This research complies with all relevant ethical regulations. Approval of the study protocol was obtained from the Lille Hospital research ethics committee (ID-RCB 2014-A00185-42) before the initiation of the study. The study adhered to the principles of the Declaration of Helsinki and the Guidelines for Good Clinical Practice and is registered at NCT02473484. Informed consent was obtained from patients. Participants did not receive any compensation. According to the French Public Health Code and in application of the General Data Protection Regulations, all patients had been informed at the time of care that their standard clinical and biological data could be used for research purposes regarding the retrospective analysis of FF samples, and none had expressed his opposition. Regarding the prospective collection of samples, each patient's informed consent for the collection and publication of clinical and biological data was obtained at the time of hospitalization prior to surgical intervention[27,28]. Tissue sections were subject to H&E coloration for histopathological analysis. The regions annotations were made by an anatomopathologist.

Chemical products and Material

Water (H_2O), ethanol (EtOH), acetic acid, acetonitrile (ACN) and methanol (MeOH) were obtained from Thermo Fisher Scientific (Courtaboeuf, France). 99% pure trifluoroacetic acid (TFA), α-cyano-4-hydroxycinnamic acid (HCCA), sinapinic acid (SA), 2,5-dihydroxybenzoic acid (2,5-DHB), aniline, formic acid (FA) and ammonium bicarbonate (NH_4HCO_3) were purchased from Sigma-Aldrich (Saint-Quentin Fallavier, France). The chloroform ($CHCl_3$) was obtained from Carlo Erba Reagents (Val-de-Reuil, France). Porcine Trypsin Sequencing Grade was from Promega (Charbonnières, France).

Tissues were cut on a cryostat (Leica Microsystems, Nanterre, France). Indium Tin Oxide slides were purchased from LaserBio Labs (Valbonne, France), whereas the poly-lysine coated slides were from Epredia™ (Braunschweig, Germany). The MALDI matrices and the trypsin were deposited on the tissue sections using the HTX M5-Sprayer™ (HTX Technologies, Carboro, NC, USA). Mass spectrometry imaging analyses were performed using the MALDI-TOF Rapiflex Tissuetyper (Bruker Daltonics, Bremen, Germany) equipped with the Smart Beam 3D laser. Spatial proteomic analysis were carried out through the utilization of chemical printer (CHIP-1000, Shimadzu, Kyoto, Japan) and the TriVersa Nanomate device (Advion Biosciences Inc, Ithaca, NY, USA). Samples were dried in a SpeedVac (SPD13DPA, Thermo Fisher Scientific, Waltham, Massachusetts, USA). nLC-MS/MS analysis were performed with TimsTOF Flex (Bruker) coupled to an EVOSEP One (EVOSEP).

Sample preparation

Rat brains were obtained from our collaborator Dr. Dasa Cizkova (Institute of Neuroimmunology, Slovak Academy of Science, Bratislava). Male Wistar rats of adult age were sacrificed by CO_2 asphyxiation and dissected. Brain tissues were frozen in isopentane at -50 °C and stored at -80 °C until use. Experiments on animals were carried out according to institutional animal care guidelines conforming to international standards and were approved by the State Veterinary and Food Committee of Slovak Republic (Ro-4081/17-221), and by the Ethics Committee of the Institute of Neuroimmunology, Slovak Academy of Science, Bratislava. For this study, FF rat brain tissues were cut using a cryostat at -20° C. All sections were obtained at the same time and stored at -80° C until their use. Rat brain sagittal 12 µm sections were prepared, to finally reach 22 batch of 4 consecutive sections. Tissues were fixed on ITO slides and respectively intended to: back-up, lipid in negative and positive mode imaging, protein imaging and peptide imaging in positive mode[29,30].

10 others consecutive rat brain sagittal sections of 12 µm were mounted on poly-lysine coated slide for lipid analysis carried out by SpiderMass technology. Three consecutive another 20 µm sections were fixed on poly-lysine coated slide for spatial proteomic analysis.

Finally, 3 different rat brain sagittal 12 µm section were fixed onto ITO coated slide as a validation cohort for the lipid predictive model.

For the analysis of horizontal rat brain tissues, 4 consecutives sections were prepared for multi-omics MSI analysis as describe bellow, followed by another consecutive sections for spatial proteomic analysis. This schema was repeated on 4 different rat brains.

Lipid MALDI MS imaging

Tissues were dried in a desiccator before a matrix deposition. Norharman was used as MALDI matrix for positive and negative lipid imaging. The matrix was deposited at 7 mg/mL in $CHCl_3$: MeOH (2:1, *v/v*). The HTX parameters for norharman spray were: spray at 30° C with 10 psi pressure, a pattern CC, a flow rate of 0.1 mL/min, a velocity of 1200 mm/min, for 12 passages with 2 mm track spacing. Lipid images were performed on the MALDI-TOF Rapiflex Tissuetyper mass spectrometer. The spectra were acquired within the *m/z* 200-1200 range in positive ion mode and the *m/z* 400-1500 range in negative ion mode. All data were performed in the delayed extraction reflectron mode with an average of 300 laser shots per pixel for a spatial resolution of 50 µm. The laser energy was set around 60 % and the voltages of the ion source were 20 kV and 11 kV for the lens. Same protocol was applied for 10 µm lipid imaging.

Other images were performed with DHB matrix in positive ion mode. The matrix was deposited at 10 mg/mL in MeOH: TFA 0.1% (7:3, *v/v*). The HTX parameters for DHB spray were: spray at 75° C, tray at 55° C, with 10 psi pressure, a pattern CC, a flow rate of 0.1 mL/min, a velocity of 1200 mm/min, for 8 passages with 2 mm track spacing. Lipid images were performed on the MALDI-TOF Rapiflex Tissuetyper mass spectrometer. The spectra were acquired within the *m/z* 200-1200 range in positive ion mode. All data were performed in the delayed extraction reflectron mode with an average of 300 laser shots per pixel for a spatial resolution of 50 µm. The laser energy was set around 85 % and the voltages of the ion source were 20 kV and 11 kV for the lens.

Protein MALDI MS imaging

Tissues were vacuum dried before being subjected to delipidation using sequential baths of EtOH: H_2O (70:30, *v/v*) for 30 s, EtOH 100% for 30 s, Carnoy solution (EtOH/Chloroform/Acetic acid, 3:6:1, *v/v/v*) for 2 min, EtOH 100% for 30 s, TFA 0.1%/H_2O for 30 s and EtOH 100% for 30 s. After drying the sections, SA-Aniline (SA-ANI) MALDI matrix was deposited on tissue. SA-Aniline was prepared by dissolving sinapinic acid matrix at 10 mg/mL in ACN/TFA 0.1% (50:50, *v/v*) and adding 24.3 µL of aniline. The HTX parameters included a temperature of spray at 75° C with 10 psi pressure, a pattern CC, a flow rate of 0.1 mL/min, a velocity of 1100 mm/min, a temperature of tray at 55° C, for 8 passages with 2 mm track spacing. The slides were analyzed on the MALDI-TOF Rapiflex Tissuetyper mass spectrometer. MS spectra were acquired in the positive linear delayed extraction mode, on the *m/z* 2400-30,000 range with an average of 700 laser shots per pixel and at a spatial resolution of 50 µm. The laser energy was set around 90 %. The voltages of the ion source were 20 kV and 9 kV for the lens.

Peptide MALDI MS imaging

For peptide imaging, the slides were dried and delipidated using a similar protocol as for protein MS Imaging. The tissue sections were then submitted to trypsin digestion. The tryptic digestion was performed by applying trypsin (40 µg/mL in NH_4HCO_3 50 mM). The HTX parameters included a temperature of spray at 65° C with 10 psi pressure, a pattern CC, a flow rate of 0.1 mL/min, a velocity of 1100 mm/min, for 12 passages with 2 mm track spacing. Once the trypsin was deposited the slides were incubated overnight at 56° C in a humidified box containing MeOH/H_2O. The slides were then dried under vacuum over the next day. An HCCA-aniline matrix was deposited by the HTX M5-Sprayer. Briefly, 43.2 µL of aniline were added to 5 mL of a solution of 10 mg/mL HCCA dissolved in ACN/TFA 0.1% (7:3, *v/v*). Slides were analyzed on a MALDI-TOF Rapiflex. Spectras were obtained in the positive delayed extraction reflector mode analysis, with a mass range of 700-3200 *m/z*, and averaged from 500 laser shots per pixel for a spatial resolution of 50 µm. The laser energy was set around 40 %. The voltages of the ion source were 20 kV and 11 kV for the lens.

Multi-Omics MSI segmentation

The raw MALDI MSI data for lipids in both ionization modes, peptide and protein data were initially converted into the imzML format[31] using SCiLS lab software. Subsequently, the imzML converter, version 1.3.3, was employed to import these datasets into MATLAB R2019a. It's worth noting that MSI data is characterized by high dimensionality, often reaching sizes of up to 100 GB per image. This magnitude makes it infeasible to analyze such data. To address this issue and prevent data loss using peak list generation, data compression was implemented as a preprocessing step before segmentation. Several data reduction (compression) algorithms were explored, including t-SNE (t-distributed Stochastic Neighbor Embedding), NNMF (Non-Negative Matrix Factorization) and SVD (Singular Value Decomposition). For the segmentation process, the *k*-means++ algorithm was utilized, implemented as the '*k*-means' function in the MATLAB Statistics Toolbox. *K*-means++ offers an improved initialization of centroids, enhancing the quality of clustering[32]. The cosine distance metric was employed to calculate the cosine angle between two spectra for quantifying the similarity. For visualization, each cluster's pixels are uniformly assigned a specific color, facilitating the creation of a segmentation map. This map delineates the cluster or region of interest to which each pixel (spectrum) belongs. To estimate the right numbers of clusters, the Silhouette criterion was used. After predefining the number of clusters, the silhouette plot method was used to assess the stability of the clusters. The silhouette plot displays a measure of the proximity of each point in a cluster. This measure has a range (-1, 1). A value close to 1 indicates that the cluster is distant from neighboring clusters (the spectra are very compact within the cluster to which it belongs and distant from other clusters). A value of 0 indicates that the sample is very close to the decision boundary between two neighboring clusters (overlapping clusters). Negative values indicate that these samples may have been assigned to the wrong cluster[33]. Silhouette plot was calculated using the function silhouette in Matlab. Subsequently, each centroid within these clusters is thoughtfully exported in CSV format, ready for further in-depth analysis and exploration.

Differential analysis between clusters

The centroids generated from the image segmentation were imported into Python using the panda’s library. All centroid data was structured into a data frame. A custom script was developed to automate the execution of a statistical test. This script iterates over all *m/z* variables, identifying ions that exhibited statistical significance between the regions of interest (ROIs). To enhance data quality, a peak picking algorithm was employed. Specifically, the find_peaks_cwt function from the Scipy library was utilized to effectively remove instrument noise. A non-parametric statistical test, the Kruskal-Wallis test with Bonferroni correction, was conducted. Only features deemed

statistically significant, with a p-value equal to or less than 0.05, were retained. A manual step is added to isolate and retain only the mono-isotopic peaks. The seaborn library was utilized to generate corresponding box plots.

Prediction model based on lipid MALDI imaging and associated proteins pathways

The previously developed pipeline[27] served as the foundation for constructing the optimal model adapted to the dataset based on highest accuracy and F1-score. These predictive models are designed to classify new MSI-lipid samples pixel by pixel, or the centroid of clusters after segmentation. While models cannot directly predict protein pathways, clusters previously associated with detected proteins using spatially resolved proteomics can indicate these pathways. Therefore, a logical algorithm was integrated into the prediction process. When a model predicts clusters, it also highlights the associated pathways and the corresponding list of proteins.

The three selected models for both rat brain optimization and glioblastoma applications were Stochastic Gradient Descent (SGD)[34], RidgeClassifier[35] and Light Gradient Boosting Machine (LGBM)[36]. The **Table 1** summarize the performance of each model in both rat brain and glioblastoma analysis. In addition, LIME (Local Interpretable Model-agnostic Explanations) was used for each model to understand the decision-making process of the models and thus identify the molecules that contribute most to predicting each cluster. The highest-contributing molecules are considered potential biomarkers.

Lipid annotation by SpiderMass technology

The basic design of the instrument setup has been described in detail elsewhere[37]. In addition, here, the laser system used was an Opolette 2940 laser (OPOTEK Inc., Carlsbad, California, USA). The infrared laser microprobe was turned at 2.94 µm to excite the most vibrational band of water (O-H). The laser beam was injected into a 1 m reinforced jacketed fiber of 450 µm inner core diameter equipped at its extremity with a handheld including a focusing lens with 4 cm focal distance to get a 500 µm spot on the tissue. To aspirate and analyze the ablated material, a Tygon® tubing (Akron, OH, USA) is directly connected to Q-TOF mass spectrometer (Xevo, Waters, UK) through a REIMS interface. Each rat brain cerebellum clusters, observed by MSI, were analyzed by SpiderMass with four independent biological repetitions. Briefly, the laser was directly placed above the region of interest at the 4 cm focal distance. The laser energy was fixed to 4 mJ/pulse[38]. On each spot, three acquisitions of 10 repetitive laser shots (10 Hz) were performed which resulted in 3 individual MS spectra. The data were acquired in both negative and positive polarities, in the sensitivity mode over a *m/z* 100-2000 range. The previously identified discriminative ions were selected for MS/MS analysis with 0.1 *m/z* isolation window. MS/MS was performed using collision induced dissociation (CID) with argon as collision gas and a collision energy of 25 eV.

Spatially resolved proteomics extraction

The different clusters identified by the segmentation process were submitted to spatially resolved proteomics. Each cluster was analyzed in triplicate from the same tissue section as describe bellow. A localized digestion was carried out by deposing a trypsin solution (40 µg/mL in NH_4HCO_3 50mM), on a region of 500 μm^2 of tissue (4 x 4 droplets of 200 µm in diameter), using CHIP-1000. The deposition method comprises approximately 1205 cycles per digestion spot, i.e., 3 hours of deposition, with a drop volume of 150 pL. Finally, each spot was digested with 0.112 µg of trypsin. Following the micro-digestion, each spot was extracted by liquid micro-junction using the TriVersa Nanomate device, with LESA (Liquid Extraction and Surface Analysis) parameters[1]. The tryptic peptides were extracted by performing 2 consecutive extraction cycles for three different solvents mixtures (TFA 0.1%; ACN/0.1% TFA (8:2, *v/v*); and MeOH/0.1% TFA (7:3, *v/v*)) for a total of 6 extractions. For each cycle, 2 µL of solvent was drawn into the tip of the pipette, of which 0.8 pL was brought into contact with the surface. 15 back and forth movements were performed to extract the peptides before collecting the solution in a recovery tube. All extracts were pulled in one tube and 50 µL of ACN were finally added before drying the samples in a SpeedVac. The samples were then stored at -20° C prior to nLC-MS/MS analysis.

nLC-MS/MS bottom-up analysis

All sample analysis was performed on a timsTOF fleX mass spectrometer online coupled to an Evosep One nano-flow liquid chromatography system. Peptides were separated using an 8 cm x 150 µm C18 column with 1.5 µm beads and the 60 samples per day method from Evosep One. The mobile phases comprised 0.1% FA in water as solution A and 0.1% FA in ACN as solution B. To perform DIA analysis in PASEF mode[39], one MS1 scan was followed by 10 dia-PASEF scans from *m/z* 100 to 1700. The ion mobility range was set to 1.42 and 0.65 $V.s/cm^{-2}$. The accumulation and ramp times were specified as 100 ms. As a result, each MS1 scan and each MS2/dia-PASEF scan last 100 ms plus additional transfer time, and a dia-PASEF method with 22 dia-PASEF scans has a cycle time of 1.06s. The mass spectrometer was operated in high sensitivity mode, with a collision energy ramped linearly as a function of the ion mobility from 59 eV at $1/K_0$=1.6$Vs.cm^{-2}$ to 20 eV at $1/K_0$=0.6$Vs.cm^{-2}$. The ion mobility was calibrated with three Agilent ESI Tuning Mix ions (*m/z*, $1/K_0$: 622.02, 0.98 $V.cm^{-2}$, 922.01, 1.19 $V.cm^{-2}$, 1221.99, and 1.38 $V.cm^{-2}$).

Proteomic data analysis

DIA-NN version 1.8.1 was used to search DIA raw files and dia-PASEF files. A Rattus library was generated with the software parameters set as following: complete proteome of Rattus norvegicus from UniProt database (Release January 2024, 92958 entries), Trypsin protease with 2 missed cleavages and a maximum number of variable modification at 3, methionine oxidation as variable, peptide length range from 7 to 30, precursor charge

range from 1 to 4, precursor m/z range comprised between 100 and 1700, fragment ion m/z range between 200 and 1700, 0.1% precursor FDR, protein inference set on 'genes', neural network classifier on single-pass mode, quantification strategy set on robust LC (high accuracy), RT-dependent cross-run normalization, and library generation fixed on the 'IDs, RT & IM profiling' ruban. Samples were interrogated according the resulting Rattus library with the same options. Data are available via ProteomeXchange with identifier PXD054488**.** Statistical analyses were carried out using Perseus software v2.0.5.0. ANOVA tests were performed with p-value ≤ 0.01 to be statistically significant and generate heat maps of differentially expresses proteins across sample. Gene Ontology (GO) analysis were performed using ClueGO[40] with GO term database, on Cytoscape v3.10.2[41].

RESULTS

The main goal of this study was to develop a machine learning pipeline capable of automatically identifying tissue heterogeneity clusters from lipid MSI data and providing associated protein networks without requiring additional protein experiments. To this end, the first challenge was to demonstrate that identified clusters are specifically spatially localized by MSI, regardless of whether lipid or protein imaging is used. Following this idea, if a cluster is identical on these omics images, it should possess specific lipid and protein pathways, tending to the basis of the dry proteomics concept.

Segmentation workflow development on RB cerebellum omics MSI.

Clustering multi-omics MALDI MSI workflow optimization

The machine learning clustering processing was the first development to adapt a workflow for multi-omics MALDI MSI. This step was focused on RB cerebral lump, a model whose anatomical and molecular characteristics are already well referenced. For the latter, four main clusters are described (**Fig 1B**): the white matter (WM) and the grey matter (GM), composed of the molecular layer (ML), Purkinje cells and the granular layer (GL) (**Fig. S1**). The first aim was to demonstrate that these clusters could be observed with the same spatial localization in each omics image, using an adapted segmentation process script.

For that, 22 RB sagittal sections were analyzed for lipid in negative (-) and positive (+) ion mode, while 12 slides were analyzed for protein and peptide, focusing on the RB cerebellum area. First, the MS spectra revealed different molecular fingerprints regarding WM, GL, and ML clusters for each molecular MSI (**Fig. S2**), confirmed for lipid (-) and protein data by t-SNE revealing clear separations of the different ROIs. On the contrary, the t-SNE obtained for lipid (+) and peptides did not show a clear separation of the different ROIs, which could predict difficulties for data processing of the latter.

To generate the most relevant segmented images, the image data was first analyzed on SCiLS lab software using RMS (Root Mean Square) normalization. The SCiLS software allows to play with different clustering parameters. Several segmentation methods were tested, including bisecting *k*-means, hierarchical clustering and *k*-means segmentation using correlation or Euclidean distance metrics. As shown in **Fig. S3**, the use of bisecting *k*-means and hierarchical clustering were ruled out due to the difficulty of interpreting the results for several reasons. First, the complexity of manually determining the desired number of clusters, which can be difficult in the case of a complex and unknown image. In addition, the spatial connectivity limitations of bisecting *k*-means do not adequately account for the connectivity between pixels in an image. This oversight can lead to segmentation discontinuities that undermine the overall accuracy and coherence of the segmentation process. *k*-means segmentation appeared to be more user-friendly, with multiple clusters defined subjectively. Unfortunately, it seems that poor centroid initialization led to insufficient clustering performance, rendering the segmentations of lipid, protein, and peptide images incomparable despite using the same number of clusters

To find a more transparent and robust strategy, data from SCiLS was imported into MATLAB software. To improve the previous clustering performance, the *k*-means++ segmentation algorithm with cosine distance metric was used. This algorithm ensures more intelligent centroid initialization, thereby improving the overall quality of the clustering. Beyond the initialization step, the rest of the algorithm remains consistent. To overcome the high dimensionality MSI data problem and to avoid data loss due to peak list generation, a pre-processing step involving data compression was introduced prior to segmentation.

For this purpose, several data reduction algorithms were investigated, including t-SNE, NNMF and SVD. As for PCA, t-SNE and NNMF are common preprocessing methods used for MSI data processing, compared to SVD. PCA and SVD are known to be suitable for linear dimensionality reduction and preserving global structure, NNMF is useful for non-negative data and part-based representation, while t-SNE excels in visualizing high-dimensional data. As shown in **Fig. 1A**, SVD compression was found to be optimal. Indeed, even if t-SNE presented good segmentations for lipid (-) and peptide images, it was difficult to distinguish the GL from the ML and WM in lipid (+) and protein cases. The results using NNMF and SVD were correct, observing the three RB areas in each omics image (**Fig. 1A**). It can be added that the images generated by the latter have a better resolution and are more looking alike. Therefore, the SVD compression was kept for the future to obtain the best possible segmentations.

It is noteworthy that within the context of this investigation, only three out of the four primary cerebellum clusters were discernible using a 50 µm MSI spatial resolution: the ML, WM, and GL in conjunction with Purkinje cells. Additionally, lipid cerebellum images were captured at a finer resolution of 10 µm (**Fig. 1C**) and subsequently processed, thereby confirming the distinct visualization of all four cerebellum clusters. This underscores the crucial

role that spatial resolution plays in the generation and differentiation of clusters, yet the spatial resolution was fixed at 50 µm since proteins imaging needs a higher resolution to get enough signals. Despite the potential for finer resolution to improve cellular component discrimination within the cerebellum, it was pragmatically determined that the 50 µm resolution sufficed for the objectives of this study, given the constraints and goals at hand.

Unsupervised cluster number estimation

We have shown that the three regions of the cerebellum can be observed in a similar way on lipid or protein image constructed with five clusters. However, the choice of the number of clusters was made in a semi-supervised manner. To automate the process of lipid-based proteomics, it was necessary to implement a tool capable of estimating the optimal number of clusters. To estimate the correct number of clusters in a non-subjective way, the silhouette criterion was used. The advantage is that it can be used multiple times, both to find the optimal number of clusters and to assess their stability and compactness.

As shown in **Fig. 1D and S4**, Silhouette estimated the optimal number of clusters at 5 for the lipid images, which was a coherent result with respect to the previously selected semi-supervised number. Furthermore, the fact that the same results were obtained for the lipids in negative or positive mode was expected due to their identical nature and metabolism. The 5 estimated clusters included 4 corresponding to the ML, GL, WM and brainstem regions of the rat brain, while 1 cluster represented a tissue-free area containing only matrix. These clusters were also observable for protein and peptide images with 5 clusters.

When analyzing protein and peptide data, predictions yielded a slightly higher number of clusters, typically between 9 and 10, reflecting the greater heterogeneity of proteins compared to lipids. Proteins are made up of a combination of 20 different amino acids, which may explain the presence of more protein clusters in the depth of the tissue compared to what is observed by lipid imaging or immunohistochemistry**.** Moreover, artefacts in tissue-free regions, likely due to inhomogeneous crystallization of the matrix, may have contributed to this variability. While we expected a single cluster to represent the matrix, as seen in the lipid data, we instead observed three distinct clusters, likely due to inhomogeneous crystallization (due to the nature of the matrix i.e. Norharman for lipids and HCCA-aniline vs SA-aniline for proteins). HCCA-aniline and SA-aniline are ionic matrices based on two component which explain the fact that we have 3 clusters instead to get only one (corresponding to HCCA, HCCA-aniline and aniline clusters or for SA, SA-aniline and aniline clusters). Taking account that in proteins and peptides due to the nature of the ionic matrix giving 3 additional clusters we can remove them and at the end we only have 7 clusters related to the tissue. Subdivisions were also observed in two clusters for molecular layer (possibly linked to the presence of Purkinje cells in some pixels) and brainstem, which were also found with lipids images with 10 clusters. Thus, in total we have 7 clusters for lipids and 7 clusters for peptides and proteins, as it can be seen in the **Figure 1D** for

the 10 clusters images, still suggesting a degree of concordance between lipid and protein clustering images. Thus, by considering previous explanations, this observed consistency reinforces the validity of dry proteomics for imaging, regardless of whether 5 or 10 clusters are used.

Finally, dry proteomics was based on lipid images which does not require additional sample preparation steps, protects the tissue from artifacts and potential degradation, and is less time consuming for routine analysis. Consequently, the clusters identified in lipid images are more representative of the RB cerebellum anatomy. In this study, we adopted the principle of dry proteomics through lipid imaging and selected the 5-cluster omics images for further analysis on RB cerebellum, as this segmentation was determined to be optimal based on the silhouette criterion for lipid images as explained above.

Finally, the optimal segmentation workflow developed (**Fig. 1E**) was a MATLAB script, integrating a SVD compression of data with 10 principal components, combined with a *k*-means++ segmentation using a cosine distance with a silhouette criterion. This approach allowed the visualization of the three main clusters of the RB cerebellum (ML in blue, GL in orange, WM in light orange), in an identical and specific spatial localization, from the 5 cluster images respectively generated for: lipid (-) and (+) MSI with Norharman matrix (**Fig. S5 and S6**), lipid (+) MSI with DHB matrix (**Fig. S7**), protein MSI (**Fig. S8**) and peptide MSI (**Fig. S9**), with semi-supervised observation.

Prediction model on lipid MALDI imaging

To automatically identify each cluster present in a tissue from a lipid image, a machine learning algorithm was trained on the 22 positive and negative lipid imaging datasets previously obtained. The ML, GL and WM centroids were extracted from the 5-cluster segmented lipid images and imported into Python. The datasets were subjected to peak picking and a non-parametric Kruskal-Wallis test to compare the significance of each ion between each ROI. Only features with a p-value equal to or less than 0.05 were retained as discriminant ions (**Fig. S10 and S11**). After isotope filtering, a final list of 36 lipid (-) and 19 lipid (+) discriminant ML, GL and WM ions were identified (**Fig. 2A-B**). The spatial distribution of each ion also confirms its specificity to its assigned cluster (**Fig. S12 and S13**). The annotation of the discriminant lipid ions was performed by SpiderMass MS/MS experiments, as its highest lipidomic similarities with the MALDI[38] (**Fig. S14 and S15**). All the specific ions of the lipids in a region are listed in an internal database, which is used to predict regions on MALDI lipid images. Various prediction models were evaluated, respectively for each lipid mode analysis, taking in account the discriminant ions previously set out. In case of lipid (-) datasets, the SGD[42,43] algorithm was selected as optimal model (**Fig. S16A**) and was validated using a 5-fold cross-validation[44] with an accuracy of 94% (**Fig. S16B**). The Ridge classifier model was the one adapted to the lipid (+) datasets (**Fig. S17A**), with 98% accuracy after 5-fold cross-validation (**Fig. S17B**). The robustness of the developed models were subsequently evaluated by blind cohort validation, which included three

different datasets of cerebellum RB lipid images for both polarity modes. Notably, in each instance, the model achieved 100% accuracy in its classifications (**Fig. S16C and S17C**).

The same data processing was performed on the protein imaging datasets to provide the corresponding discriminant protein ions for each RB cerebellum clusters (**Fig. 2C and S18**). The list of discriminant protein ions was then added to discriminant lipid ions for each cluster in order to create discriminant protein and lipid ions dataset reference for each RB cerebellum area.

Lipids biological network analysis

Based on the compilation of annotated lipids, a greater number of lipids have been specifically identified in WM compared to GM. This observation is in line, considering that the WM is predominantly comprised of myelin, a substance containing a higher lipid content (78-81%) than both white (49-66%) and grey matter (36-40%)[45]. In the same way, myelin is composed of a high percentage of galactoceroboside and cholesterol compared to GM, which is why more diglycerides (DG), triglycerides (TG) and fatty compounds are identified in the latter. On the other hand, GM presents a higher percentage of phosphatidylethanolamines (PE) and phosphatidylcholines (PC), which correlate with presented annotations[45].

To highlight the biological process involved by lipid data, a comparison between WM and GM discriminant lipid was performed on BioPAN[46]. The results, shown in **Fig. 2D** and **Fig. S19A**, revealed PC biosynthesis as the most active pathway in WM (with the involvement of PEMT predicted gene), whereas PE biosynthesis was observed as the most active pathway in GM (with the involvement of PISD predicted gene). These results clearly indicate discriminants lipids involved in specific biological pathways associated to distinct cerebellum regions.

Biologically, phosphatidylcholine is an essential choline reservoir for brain function[47]. In fact, choline is an important molecule for neurotransmission in neurons, which may explain the high activation of PC biosynthesis in WM. Phosphatidylethanolamine's biological function is more due to its small chemical structure, which allows fluidity of the neuronal membrane[48]. The hypothesis is that this facilitates vesicle budding and membrane fusion[49], a key step in synaptic transmission in GM. Finally, biological pathway based on lipids analysis showed that PC may be involved in the neurotransmission process in WM, whereas PE is more involved in synaptic transmission in GM[50]. These conclusions were further corroborated by Reactome analysis of the lipid dataset (**Fig. S19B**), which demonstrated their involvement in the neural system, signal transduction, small molecule transport or metabolism of protein and vesicle-mediated transport pathways.

Consolidation method by protein pathway analysis

As discriminant biological pathways were defined for different regions of the cerebellum RB based on lipid species, the proteomes of these regions were analyzed to validate the hypothesis of a correlation between lipids and proteins within the same biological network. This analysis aimed to consolidate the dry proteomics processes. The ML, GL and WM regions observed in the multi-omics MALDI MSI were therefore subjected to spatial proteomics using the micro-proteomics workflow on three different RB sections[51,52]. By regrouping the triplicates for each cluster, a total of 5270 proteins were identified for WM, 5390 for GL and 5354 for ML (**Supplemental Spreadsheet S1**). This study confirmed the spatial heterogeneity of proteins previously observed in imaging. The results showed that discriminant lipid species for each ROIs are consistently linked to specific proteins in the same ROI, thereby forming region-specific pathways and functions.

Indeed, the Venn diagram, shown in **Fig. 3A**, considers the protein diversity between each region by the presence of proteins exclusive to each of them. In total, 85 proteins were exclusive, of which 7 were specific for WM, 11 for ML and 67 for GL (**Supplemental Spreadsheet S2**). It must be noted it was found among the 11 specific proteins in ML, two important enzymes involved in lipids metabolism *e.g.* Phosphoinositide phospholipase C and Inositol monophosphatase 1 whereas in GL, the Gamma-butyrobetaine dioxygenase know to catalyze the formation of L-carnitine and the Plasmolipin in WM, a main component of the myelin sheath involved in intracellular transport, lipid raft formation, and Notch signaling were identified[53]. The GL contain several neuropeptides or neuropeptide hormone activity such as Corticotropin-like intermediary peptide, Somatostatin-14; Pro-thyrotropin-releasing hormone, cholecystokinin-12; Neurokinin-B, Cocaine- and amphetamine-regulated transcript protein; Pituitary adenylate cyclase-activating polypeptide 27 or Ephexin-1[54]. In ML, among the identified protein the lamin B-binding protein (BAF: Barrier-to-autointegration factor) and Myogenin are of particular interest. In fact, BAF is required during brain development as a regulator of nuclear migration during neurogenesis of the CNS[55]. Myogenin is also detected in Allan brain atlas and is linked to motor neurons[56]. Similarly, in WM among the specific proteins identified, the Lymphocyte specific 1 is recently known to be correlated with tissue resident memory T cells[56] and T cell infiltration[57]. Interestingly, Phosphatidylserine decarboxylase proenzyme (PISD) was also found in both WM and ML regions and was a predicted gene previously reported in BioPAN GM lipid pathway (**Fig. S19A**). The presence of PISD protein may explain the amount of PE identified in the ML region. In this context, PE may contribute to the integrity and function of neuronal membranes, influence synaptic transmission, and participate in signaling events. This again demonstrated the relevance of the different clusters by MSI, which predicted their own lipid/protein pathway and therefore biological heterogeneity.

To go further, the common proteins were subjected to an ANOVA test (p-value < 0.01) and showed that 2204 out of 5465 proteins have a significant variability of expression (**Supplemental Spreadsheet S3**). According to Allan

brain Atlas, based on transcriptomic analyses, 196 genes are Cerebellum enriched gene and 59 out of those genes show highest expression levels in cerebellum. 90% of their corresponding proteins have been identified such as CBLN1 and CBLN3. Among them, some are known to be specifically located to the Purkinje layer which was re-grouped with the GL after clustering**.** We were able to identify specific proteins from the Purkinje cells (MYH10, HOMER3, KIT, QKI, MX1, PCP-2, PP1R17, ARGEF33). For example, QKI protein expressed by radial astrocytes (Bergmann glia) with processes through the molecular layer all the way to the pial surface of the cerebellar cortex has been identified. MX1 is known to be in the dendritic processes of Purkinje cells. Moreover, other specific proteins of granular layer, GABRB2, TMEM6 and KCNIP4, markers of synaptic glomeruli from granular cells are also detected.

This was reflected by the presence of different clusters of over- or under-expressed proteins between each RB cerebellum area (**Fig. 3B**). The gene lists corresponding to over-expressed protein clusters were analyzed using ClueGO software to identify the biological pathways associated with the significant proteins identified in each distinct cluster. It turns out that the overexpressed proteins in the WM are mainly involved in myelination, glucose and neurofilament metabolism (**Fig. 3C and S20A**), which is a consistent result according to the bibliography[58]. In fact, WM consists of myelinated axons, so it's involved in the transmission of nerve impulses by axons. The presence of glucose metabolism is also interesting when correlated with the galactoceroboside myelin composition previously suggested by lipid WM analysis. Furthermore, iron metabolism is another important biological process in the white matter, e.g. for myelin formation, redox reactions or neuronal development and synaptic plasticity[59–61]. This information can be linked to biological pathways previously found by lipids analysis, which also highlighted the neurotransmission pathway in WM. Regarding the GL, the neuropeptide hormone activity pathway was found to play a role in the processing and regulation of peptides that influence synaptic transmission, neural signaling and modulation of neuronal activity (**Fig. 3D and S20B**). Purine metabolism also plays a crucial role thereby influencing various physiological processes such as neurotransmission, synaptic plasticity, and energy metabolism. Dysregulation of purine metabolism in the brain has been implicated in several neurological disorders, including epilepsy, Parkinson's disease, and neurodegenerative diseases. Similarly, the relevance of synaptic organization and sodium ion transport pathways involved in the molecular layer (**Fig. 3E and S20C**) were expected results given their role in neurotransmission and synaptic signaling between these cell types.[62]. It's interesting to remember that the biological processes of synaptic transmission, vesicle transport and signaling were also predominant pathways in the previous lipid study. Thus, it has been shown that the ML, WM, and GL have their own specific proteome that can be correlated with specific lipid associated to distinct biological pathways.

Dry proteomics based on RB horizontal lipid imaging application

Multi-omics RB horizontal sections generation

To validate the dry proteomics workflow to more complex tissue, the analysis was widened to total horizontal RB sections. As previously, multi-omics MALDI MSI were performed on 4 different sets of consecutive horizontal RB sections, and resulting data were submitted to the imaging data processing workflow, excluding matrix signal. The Silhouette criterion was around 11 for each lipid replicate, leading to multi-omics images composed of 11 clusters (**Fig. 4A and Fig. S21**). A similar spatial clustering shape was observed for each lipid image, including the well-known areas of the cerebellum, as well as other specific areas such as: the corpus callosum subdivided into clusters white, green and yellow, the cerebral cortex and thalamus in purple, red and pink, and the ventricular system in brown. These specific regions were also observed on the protein and peptide images built with 11 clusters, again confirming the lipid/protein pathway cluster appurtenance.

RB cerebellum lipid classification model: prediction on horizontal sections

Four replicate lipid (-) horizontal RB images were blindly analyzed using the pre-built classification model trained on 22 RB cerebellum lipid (-) MSI datasets (**Fig. S22**). The model returned a confidence score for predicting each ROI. Since the model was trained on three ROIs, the default confidence score to predict an ROI was >33%. The model successfully predicted the ML area with a mean confidence score of 100%, WM with a confidence score of 52%, and GL with 89% (**Fig. 4B**). For WM, although 52% is significantly higher than 33%, the lower confidence score may be due to the discrepancy in surface area between the sagittal and horizontal brain slices of the rats, with the former showing a significantly greater extent of WM. Other clusters were also analyzed using the predictive model (**Fig. 4B**) with interesting results. The light green and yellow clusters (corpus callosum region) were predicted as WM with confidence scores of 75% and 61%, respectively. Similarly, the green cluster (colliculus regions) was predicted as GL with a confidence score of 71%. A Pearson's correlation of the discriminant lipid negative ions, shown in **Fig. 4C**, further validated these predictions. Two main clustering branches were identified: one leading to correlated cluster 1 associated with ML, and another leading to two separate clusters, correlated cluster 2 associated to GL and correlated cluster 3 associated to WM. In correlated cluster 1, dark purple and brown ROIs were grouped with ML, sharing the 774.6 and 790.6 lipid (-) ions (**Fig. 4D**). In correlated cluster 2, WM was grouped with the yellow and light green ROIs, as predicted by the model, with the main involvement of the 888.7 and 906.7 lipid (-) ions (**Fig. 4D**). Biologically, these results were expected. The corpus callosum (light green and yellow clusters) forms the largest commissural WM bundle in the brain, which has a distinct molecular composition due to its significant size and role, explaining the observed clustering[63]. Similar observations were valuable for the colliculus (green cluster), which also contains a superficial grey layer[63]. This explains the presence of orange color in both granular

layer and colliculus clusters, corresponding to GM, and accounts for the 71% confidence score prediction explaining similarity[65].

With the aim to justify the images segmentation, discriminant lipid (-) ions were identified for different cluster observed on the horizontal RB section lipid (-) image (**Fig. S23**). Many peaks were spatially distributed regrouping multiple clusters. For example, common ions were spatially distributed in ML, cerebral cortex and hypothalamus regions, like *m/z* 790.4, 834.4 and 886.5. The ion *m/z* 599.4 was collocated in GL and green cluster. Same observations for both the WM and the corpus collosum, for which ions as *m/z* 701.6, 889.6 and 904.7 were also spatially present. These ions could explain the correlation clusters highlighted by the Pearson's correlation graph in **Fig. 4D**. However, discriminant ions were also extracted for specific regions, explaining their segmentation as single cluster during MSI data processing. The ventricular region (brown cluster) possessed various discriminant ions such as *m/z* 473.2 and 615.1. Specific ions were also discriminant for the red cluster, regrouping cerebral cortex and hypothalamus regions (*m/z* 746.6, 766.6, and 834.5), as well as for the corpus callosum like *m/z* 806.6. This ions list was added to the prediction model, to refine predictions.

Proteome horizontal RB section cluster comparison

To have a look at the proteome specificity of the RB horizontal section clusters, spatial proteomic analysis was also performed on the 7 clusters observed from the rat brain 11-cluster segmentation image (excluding the cerebellum cluster, already analyzed) (**Fig. 5A**). Proteins from red cluster were extracted from hypothalamus region, while proteins from purple and pink clusters were extracted from cerebral cortex. The green cluster was extracted from colliculus area, brown cluster from ventricular system and yellow and light green clusters from corpus callosum. Experiments were performed in biological triplicate. Data were processed with ML, WM and GL previous data in DIA-NN software for protein identification, quantification and correlation. By regrouping the triplicates for each cluster, more than 17243 proteins were identified, among them 5498 were proteotypics (**Fig. 5B**) (**Supplemental Spreadsheet S4**). Common proteins were subjected to an ANOVA test (p-value < 0.0001) and showed that 4481 out of 7223 proteins had a significant variability in expression. This was represented by the presence of different clusters of over- or under-expressed proteins between each extracted region (**Fig. 5C**). The resulting heatmap highlighted different clusters of overexpressed proteins (**Supplemental Spreadsheet S5**). First, cluster i consisted of proteins overexpressed in the cerebellum regions (ML, GL, and WM), while cluster v consisted of proteins overexpressed in the other regions. Specific overexpressed protein clusters were also highlighted for the ventricular system in cluster ii and for the corpus callosum in cluster iii. It was also observed that cluster iii was involved in WM, confirming their correlation in the previous lipid Pearson's analysis (**Fig. 4C**). The overexpressed protein

cluster iv was involved in the cerebral cortex and hypothalamus brain regions, explaining their similar image segmentation in the red cluster (**Fig. 5A**), Pearson's correlation and prediction model using lipid (-) data (**Fig. 4C**).

Biological pathway analysis of these later over-expressed protein clusters also confirms this observation (**Fig. S24**). Indeed, biological pathways involved in cerebellum (cluster i) were mainly concentered around synapse metabolism, with myelination, paranodal metabolism, neurofilament assembly, and calcium/sodium transport (**Fig. S24A**)**.** We could notice that this biological process also resumed the one's independently found for ML, GL and WM (**Fig. 3**). At the opposite, biological process involved in the cerebral cortex (cluster v) contributed to at least NMDA selective glutamate receptor signaling, regulation of neurotransmitter receptor transport (endosome to postsynaptic membrane) (**Fig. S24B**). This distinction of biological process well defined and distinguished the cerebellum and cerebral cortex regions of the brain. Indeed, the cerebellum is primarily involved in coordinating motor movements, maintaining posture and balance, and motor learning[64], whereas the cerebral cortex is responsible for higher cognitive functions including perception, memory, attention, language, and consciousness[65].

Same conclusions were observable analyzing biological pathways specifically involved in cerebral cortex and hypothalamus, in cluster iv, where mains pathways regrouped vocal and auditory learning, memory and feeling process with serotonin metabolism (**Fig. S24D**). Likewise, myelination and neurofilament pathways were involved in cluster iii, for corpus callosum RB area (**Fig. S24E**), which was linked to WM biological pathways. The biological pathways for cluster ii, specific of ventricular system RB region, was also analyzed. It turned out that cholesterol, triglyceride, and blood coagulation regulation were the most relevant pathways (**Fig. S24C**). These results fit with the neuroanatomy of ventricular system, where cerebrospinal fluid flows in the regions thanks to blood pulsations in surrounding blood vessels[66]. Furthermore, triglycerides cross the blood-brain barrier and are found in cerebrospinal fluid helping in satiety and cognition mechanisms[67].

In this way, we were able to show from a protein pathway point of view that cerebellum regions are distinct from the cerebral cortex regions, which itself consists of several specific areas. Their proteomes were also integrated into the model with their paired lipid clusters. In addition, proteomic data of this study were in line with previous analysis already performed on RB regions from published studies. This allowed to add more information to the RB dry proteomics model. First, we compared proteins identified here in bottom-up, with proteins identified by top-down in the hippocampus and corpus callosum RB areas, presented in a previous study[3]. According to Delcourt, V. *et al.,* 2018[3], 16 over 22 proteins identified in top-down for the corpus callosum were also identified and over-expressed in this area according to presented protein dataset. Same observations for 15 proteins over the 20 identified in top-down for the hippocampus (**Supplemental Spreadsheet S6**).

Workflow robustness

The robustness of the dry proteomics workflow was thoroughly assessed by examining the redundancy of spectral lipidome and proteome identifications within each cluster across independent triplicates. To ascertain clustering repeatability, the spectral lipid (-) dataset from each cluster was compared among triplicates, as illustrated in **Fig. S25A**. Impressively, an average of 99% of common lipid (-) ions was consistently identified across replicates within clusters (refer to **Fig. S25C**). Similarly, an in-depth analysis of the spatial proteomic dataset, with a specific focus on distinct clusters, revealed a remarkable consistency, with 93% of the proteins consistently identified across each replicate extraction point within a cluster. This robustness is highlighted in **Fig. S25C**, which succinctly summarizes the percentage of common protein identifications in replicates for each cluster (**Fig. S25B**). Notably, it's worth mentioning that proteins involved in cluster-specific pathways, as previously depicted in **Fig. 4A** were fully recovered at a 100% rate in subsequent analyses. This underscores the reliability and reproducibility of the methodology employed in capturing proteomic signatures associated with distinct cellular clusters. This reproducibility is the essence of dry proteomics. For future analyses, there's no need to redo spatially resolved proteomics. Simply start with a lipid image and query the dry proteomics model to reliably determine the cluster type, associated proteins, and relevant biological pathways.

Glioblastoma tumoral heterogeneity analysis

Lipid and peptide MSI segmentation correlation

Finally, we performed the dry proteomics workflow on a prospective and retrospective cohort of glioblastoma (GBM), re-using collected data from Duhamel, M. *et al*., 2022 study[27,28]. The previous study performed patient's stratification based on spatial proteomic and spatial lipidomic guided by MALDI MSI associated to patient survival[27,28]. The cohort consisted of 50 GBM patient tissues, referenced to P1 to P53 (**Fig. S26 and S27**). Peptide MALDI MSI was performed for all samples, and lipid MSI was conducted for 13 of these tissues. Thus, peptide and paired lipid images were collected for these 13 patients and were processed through developed data imaging workflow. Initially, each tissue was analyzed individually to assess its heterogeneity using Silhouette criterion and generate segmented images. Subsequently, peptide and lipid images were created with 8 to 13 clusters each. The findings of this study revealed an intriguing correlation between lipid and peptide distributions in samples labeled P1 to P14, as evidenced by the generation of highly similar numbers of clusters in both types of images. This correlation underscores the inherent link between the spatial heterogeneity of peptides and lipids within the tissue microenvironment (refer to **Fig. 6A and Fig. S26**). Furthermore, segmentation analysis effectively mirrored histological annotations, enabling the delineation of distinct regions of tumoral proliferation from necrotic or inflammatory areas (as depicted in **Fig. 6A**), as it was also evocated by Duhamel, M. *et al*., 2022 in previous studies[27,28]. Prior investigations have primarily relied on lipid and protein to differentiate between these three main tissue types based

on specific molecular signatures. In contrast, dry proteomics segmentation workflow offers a more detailed representation of the intricate composition of biological tissues. This enhanced segmentation, not only facilitates the precise identification of pathological features but also reveals previously undetected levels of heterogeneity within tumor, necrotic, and inflammatory regions. This not only achieved improved delineation between annotated areas but also unveiled a greater-than-expected level of heterogeneity within these regions. This heightened resolution enhances understanding of tissue composition and offers valuable insights into the underlying biological processes driving tumor progression and response to treatment (**Fig. 6A**). For instance, in the case of P12, a nuanced examination of the proteomic extraction points unveiled intriguing insights. Points 12.1, 12.3, and 12.4, which were annotated as tumoral in the histopathological scan, exhibited a complex molecular landscape. Notably, point 12.2 was identified as bearing both tumor and inflammation characteristics. However, upon closer inspection using lipid and peptide MSI, it became evident that point 12.4 shared similar molecular profiles with point 12.2, in stark contrast to points 12.1 and 12.3. This striking observation was further corroborated by protein extraction analysis, which revealed distinct correlations among the points. Specifically, points 12.1 and 12.3 exhibited a notable correlation, indicating shared molecular features, while points 12.2 and 12.4 formed a separate correlated cluster (**Fig. 6D**). This delineation underscored the intricate heterogeneity within the tumor microenvironment, where discrete molecular signatures delineated different regions, potentially indicative of diverse biological processes or cellular compositions.

Lipid-MSI clusters classification and proteomic correlation

To have a large view on the general heterogeneity on the whole cohort, a co-segmentation was performed on 9 lipid images dataset. It turned out that 13 different clusters were shared between these 9 patients' tissues (**Fig. 6B**). Some clusters were correlated to biological specific tissues regions according to histopathological annotations. In this way, clusters 4 (light pink) and 9 (dark purple) were identified as necrosis tissues, clusters 1 (blue), 2 (light green) and 7 (orange) seemed to be specific tumors, whereas clusters 3 (green) and 5 (red) were tumoral areas near to inflammation and clusters 6 (light orange), 8 (light purple), 10 (yellow), 12 (light blue) and 13 (pink) were tumoral areas with necrosis. Clusters were predominantly identified within specific tissues, such as cluster 9 primarily present in P9, or shared across multiple tissues, as observed with cluster 3 in P1, P2, and P13. Once more, the segmentation underscored the molecular diversity within necrotic and tumoral regions, revealing a mosaic of numerous clusters.

A t-SNE representation of tissue lipid imaging clusters allowed to distinguish two mains groups of clusters based on lipid MSI (**Fig. 6C**): group A was regrouping clusters 6, 8, 9, 10, 12 and 13, while group B regrouped clusters 1,

2, 3, 4, 5, and 7. Cluster 11 was shared between the two groups. A correlation heatmap, presented in **Fig. 6D**, also highlighted the correlation between lipid clusters regrouped in group A and B.

The proteomic data obtained from nine distinct tissue samples were leveraged to conduct a comparative analysis of the various clusters identified through lipid imaging segmentation. Notably, specific extraction points analyzed in this study correlated with clusters identified in lipid imaging (**Fig. 6B**). Through statistical analysis, employing an ANOVA test with a significance threshold set at $p < 0.01$, we identified 373 out of 3616 proteins exhibiting significant variability in expression levels (**Supplemental Spreadsheet S7**). First, biological pathways were identified for each cluster through ClueGo analysis (**Fig S28**), based on the overexpressed proteins present in each. Interestingly, some pathways were specific to particular lipid clusters. For example, the RAC3 GTPase cycle pathway was unique to cluster 7 (**Fig S28F**), playing an important role in neuronal development and tumor progression[68]. L1CAM expression was particularly found in cluster 1 (**Fig S28A**), underscoring the tumor aggressiveness of this cluster. This pathway is a focal point of active investigation in GBM due to its profound implications for tumor aggressiveness, invasion, therapeutic resistance, and poor prognosis. Similarly, overexpressed proteins in cluster 5 were specifically involved in the axon guidance pathway[69], which is currently a therapeutic area of research for the treatment of malignancy. On the other hand, some biological pathways were common across multiple clusters. Notably, the interleukin-12 family signaling pathway[70], a current therapeutic target in cancer immunotherapy, was identified in clusters 6, 9, 10, and 13 (**Fig S28E, H and J).** Similarly, the ECM proteoglycans pathway[71] associated with tumor development in GBM was found in clusters 4 and 9 (**Fig S28C and H**). Finally, the biological pathway analysis of each cluster revealed distinct characteristics: some clusters exhibited a more aggressive GBM pattern, whereas others showed a less aggressive pattern and identified potential therapeutic targets.

Further investigation on protein data allowed to compare the proteome of each cluster and identify correlations between them. It revealed the presence of 2 distinct clusters of over-expressed proteins, namely protein cluster A and B (**Fig. 6D**)**.** Of particular interest, protein cluster A was found to correspond to regions of necrotic tissue, encompassing the imaging clusters 9 and 4 previously described. To gain deeper insights into the biological processes associated with these necrotic regions, we performed ClueGO analysis on protein cluster A, utilizing GOterms and Reactome databases (**Fig. S29A**). This analysis unveiled a multitude of signaling pathways implicated in necrosis processes. Notably, pathways such as platelet degranulation, blood coagulation, MyD88 deficiency, and IRE1 chaperone activation emerged as significant contributors in modulating cell death processes, including necrosis and can influence tissue damage and disease progression in various pathological conditions such as GBM. In the same way, an intriguing correlation in protein cluster A was observed among protein extracted

from lipid imaging clusters 6, 8, 10, 12, and 13, as depicted in **Fig. 6D**. The later result confirmed the lipid image cluster classification in group A, proposed previously according lipid MSI co-segmentation analysis (**Fig. 6C**). This cluster notably encompassed tumoral clones characterized by the presence of necrotic regions. Through ClueGO analysis, the significant implication of selenoamino acid metabolism within this cluster was unveiled, shedding light on its pivotal role in the pathogenesis of glioblastoma. This pathway was also individually identified previously in **Fig. S28E and J** in lipid cluster 6, 10 and 13. Selenoamino acids, such as selenocysteine and selenomethionine, are fundamental constituents of selenoproteins, where selenium, an essential trace element, is incorporated. These selenoproteins orchestrate a myriad of cellular processes, including antioxidant defense, redox regulation, and DNA synthesis and repair. The dysregulation of selenoamino acid metabolism has been implicated in the intricate progression of GBM through various mechanisms, contributing to disease aggressiveness and resistance to therapy. Similarly, the over-expressed proteins identified within protein cluster B, primarily comprising lipid imaging clusters 3, 4, 5, and 7, yielded significant insights, particularly regarding the involvement of L1CAM interactions (**Fig S29B**) from cluster 1 **(Fig. S28A).** Protein cluster B suggested a more aggressive tumor phenotype compared to those within protein cluster a, with implications for poor prognosis or short survival prediction. The intricate interplay between selenoamino acid metabolism and L1CAM interactions underscored the multifaceted nature of GBM pathogenesis, highlighting potential avenues for targeted therapeutic interventions and personalized treatment strategies aimed at mitigating tumor progression and improving patient outcomes.

Finally, two distinct classification groups, labeled group A and group B, were highlighted and cross-validated between lipid MSI and proteomic analysis. Proteins from the over-express protein cluster A were associated to lipid cluster A, resulting in group A. Thus, group A was associated to the lipid clusters 6, 8, 9, 10, 12, 13, and protein, involving specific protein pathways with a pivotal role in GBM, such as selenoamino acid metabolism. In another hand, group B regrouped lipid clusters 1, 2, 3, 4, 5, 7, and the over-expressed protein cluster B, which possessed more aggressive protein pathways with the implication L1CAM interactions.

Patient proteome blind prediction based on lipid cluster classification

To predict patient proteome through group A and B, two distinct classifications models were developed. Firstly, a model was trained on the lipid-MSI data from the 13 clusters comprising groups A and B. The aim of this classification model was to classify patient tissue according to lipid images, and associate their paired protein pathway. The resulting model was built with LGBM algorithm with an accuracy of 97% after 5-fold cross validation with an individual accuracy up to 95% for each cluster (**Table 1**, **Fig. S30B-C**). Specific lipid ions involved in the model were extracted and identified in specifics clusters using LIME algorithm. The top lipid biomarkers implicated to classify each cluster with 82.3% of contribution were summarized in **Fig. S30E and S31**. For example, lipid with

m/z 770.35 was specific to group A, with a significative presence in lipid cluster 8 with highest contribution weight at 70% (**Fig. S30C**). Likewise, the *m/z* 798.64 was specially distributed in lipid MSI cluster 5 (**Fig. S32**) with a contribution weight at 61% (**Fig. S30C**), associating it with group B specific marker.

Thus, the classification of all 9 patients was carefully reviewed according to the patient group A or B classification model, based on the presence of specific lipid clusters in tumoral tissue. In scenarios where tissue samples exhibited clusters overlapping both group A and B, they were unequivocally classified into group B, prioritizing the presence of markers indicative of unfavorable outcomes. This approach ensured a rigorous and systematic evaluation, wherein each case was subjected to thorough examination, with particular emphasis placed on identifying and prioritizing markers associated with poorer prognostic indicators. By adhering to the following patient classification method, the prognostic assessment process maintained an exemplary level of precision and consistency, empowering clinicians to render well-informed decisions regarding patient management and treatment strategies. As depicted in **Fig. 6E**, 4 patients were classified in group B and 5 patients in group A. It's noteworthy that previous investigations have emphasized the importance of assessing patient classification based on the expression levels of key proteins[28]. In 9 patient's cohort (outlined in **Fig. 6E**), prior studies classified 2 patients in group B and 7 patients in group A using this protein panel[28]. However, resulting analysis unveiled a nuanced disparity in group classification for patients P10 and P12. This discrepancy can be attributed to the incorporation of molecular heterogeneity into analysis, offering additional insights into survival prediction. Furthermore, upon scrutinizing the co-segmentation analysis illustrated in **Fig. 6B**, it became evident that P12 and P1 shared significant cluster composition. Given P1's association with group B, it was reasonable to surmise that P12, sharing similar cluster characteristics, would also be classified within group B. This observation underscores the importance of integrating molecular heterogeneity and comprehensive data analysis techniques to refine classification assessments and enhance clinical decision-making processes.

The 4 last patient tissues, for which lipid-MSI and protein data were available (P3, P5, P6 and P11), were blindly interrogated in classification model based on lipid-MSI clusters. P3 and P11 presented the IDH1 mutation and were not considered in the studies of [27,28]. Upon blind interrogation of the lipid cluster images, patients 3 and 6 harbored a non-negligible percentage of lipid clusters 4 and 2, leading to the prediction to proteins associated to wound healing, or ECM proteoglycans biological pathways for example (**Fig. 6F**). The presence of the latter lipid cluster and biological pathways in patient 3 and 6 were thus indicative of the group B classification. Conversely, patients 5 and 11 mainly predicted with high percentage of lipid cluster 6 and 8, allowed the prediction of proteins associated to biological pathways such as Interleukin-12 family signaling, peptide chain elongation or RHO GTPase active ROCKS (**Fig. 6F**). In this way, patient 5 and 11 were classified in group A. This result also correlated with a lipid

MSI co-segmentation performed on the 13 tissues (**Fig. S33**). The resulting image was composed of 14 clusters according Silhouette criterion. Interestingly, P6 and P3 were segmented apart of the rest of the cohort, suggesting a possible new lipid class. P5 and P11 tissue associated to group A were sharing specific clusters with P8 and P14, already previously classified in group A.

Complementary, a second classification model was constructed using RidgeClassifier with group A and B protein data (**Fig. S34**). The objective was first to intricately cross-validate the lipid MSI-based classification model. The resulting model had an accuracy of 96%, with a 5-fold cross validation (**Table 1** and **Fig. S34F**). Specific proteins involved in the model decision-making were identified in specifics clusters (**Fig. S35**). Among them, group A and group B biomarkers were distinct, referring to selenoamino acid metabolism or L1CAM interaction pathway for instance. This sophisticated approach underscored the synergy between lipidomic and proteomic analyses in refining group A and B classification for glioblastoma patients, thus paving the way for personalized therapeutic interventions tailored to individual risk profiles.

Thus, previous finding was further reinforced by the protein classification model, which concurred in its classification assessment, designating patients P3 and P6 to group B, whereas P11 and P5 were classified to group A. Hence, both the lipid-MSI clusters and protein models converged in classifying these patients within group A, or B. This alignment serves to authenticate the reliability and validity of the classification model, as well as enhancing the dry proteomics concept on clinical study as GBM[28]

Groups classification and patient outcome correlation

The dry proteomics developed pipeline, in both using lipid-MSI data and proteins classification models, led to the discernment of two distinct classes in GBM study, labeled as classification group A and B, illustrated in **Fig. 6-D**. Leveraging patient survival data, prognostic outcomes were correlated with specific lipid-MSI clusters. The clinical characteristics of the patient, evocated in studies[27,28], revealed that patients involved in group A, through lipid-MSI clusters 6, 8, 9, 10, 12, and 13 implication, were upper the survival interquartile range with a survival outcome surpassing 32 months. In the same logic, patients associated to group B, with the presence of lipid clusters 1, 2, 3, 4, 5, and 7, had a poorer survival prognosis of less than 30 months.

Indeed, some of lipid biomarkers involved in lipid-MSI classification model were already recognized as prognostic markers in previous research[27]. For instance, lipid ions with *m/z* of 864.7, 866.7, and 881.7 were identified in both studies as markers for survival outcomes exceeding 36 months, primarily present in clusters 8 and 9 from group A. Conversely, lipid ions such as *m/z* 760.6, 788.6, and 810.6 were associated with shorter survival durations, less than 30 months, and were distinctly present in clusters 2 and 5 from group B. Moreover, these significant findings were consistent with prior investigations, reinforcing the notion that protein group B typically correlates with a poorer

prognosis compared to group A. Particularly notable was the identification of over-expressed proteins ANXA6 and GPHN within group B, both previously implicated as unfavorable prognostic indicators [28]. Conversely, group A exhibited elevated expressions of proteins RPS14 and MTDH, associated with more favorable prognostic outcomes[28]. Thus, the identification of group A and B lipid features by MALDI MSI, would automatically provide the paired protein pathways (**Supplemental Spreadsheet S8**), associated to short or long survival patient outcome.

As the left 37 patients were only analyzed through peptide MALDI MSI and spatial proteomics due to the data reuse, the later were interrogated through classification model with proteomics data, to predict their appurtenance to group A and group B, and thus their protein networks and prognosis (**Fig. S36**). Finally, among the cohort of 50 patients, 11 patients were classified in group A with a prognosis survival outcome >32 months, whereas 39 patients were classified in group B with a survival outcome <30 months. The latter results correlated with the clinical characteristics of the patient evocated in study [28]. Indeed, 4 patients with IDH mutation were excluded, 12 patients were upper the survival interquartile range (IQR) set at 13.5 and 32 months, 23 patients were intermediate IQR, and 11 patients were lower IQR.

Dry proteomics limitations

Although the dry proteomics model is robust, fast, and simplifies the analysis of complex heterogeneous tissues, it has some experimental and predictive limitations.

Technically, it is impossible to obtain identical consecutive tissue sections due to the z-dimensional factor related to tissue depth during cryostat sectioning. For example, we observed less structural changes between consecutive sections of the cerebellum. However, in horizontal sections, where the anatomy is more complex and variable, differences between consecutive sections are noticeable. These differences affect the imaging of lipids, proteins and peptides due to anatomical changes with depth. To address this issue, we performed spatially resolved proteomics on the same section used for lipid imaging. Once the model is trained, dry proteomics becomes a useful tool because only one lipid image is needed to assess the heterogeneity, identify the clusters, and associate the proteome, avoiding issues related to anatomical changes in consecutive sections. The second limitation concerns the predictive ability of the model, which is based on experimental data of clusters obtained by segmentation of lipid images. A reliable and accurate model requires a large cohort with representative replicates of the studied population. Building a generalizable model is challenging because some tissue-specific clusters may not be represented in our analyses. When the model encounters an unknown cluster that it hasn't been trained on, it will likely misclassify it by approximating a known cluster. There are two ways to address this problem. First, by checking the approximation of an unknown cluster by the unsupervised k-means++ and t-SNE models. This involves plotting the

matrix of this cluster on the k-means++ and t-SNE axes to see which known cluster it is close to, thereby confirming or disproving the model's predicted approximation. Second, consider the use of self-training algorithms in the future [72]. This involves retraining our model with known labeled clusters and new unknown and unlabeled clusters to improve and update the model specifically for clinical routine use. In this case, it will also be necessary to update the proteomic data for the new unknown clusters.

To extrapolate the strategy of dry proteomic to other tissue types or diseases, different learning model approaches are possible. The first one consists in a specific model for a specific tissue type or disease. In this case, the model would be trained on clusters specific to a particular tissue type or disease. While this approach is limited to the heterogeneity of that single tissue or disease, it offers greater accuracy by focusing on fewer clusters, which reduces the risk of false positive predictions (fewer classes in a multi-class classification task). This results in a more targeted and precise model. The second possibility is to improve the model in an agnostic model. This is a global model designed to work across multiple tissue types or diseases. To improve its performance, the model would need to be trained on clusters from various diseases and tissue types. Such a model would be capable of predicting and identifying clusters specific to particular tissues or diseases, while also recognizing common clusters across different tissue types. This approach could be especially useful for large-scale studies, such as PAN-cancer research. However, agnostic models are typically less accurate and require sophisticated feature engineering to enhance their performance. Another strategy to improve agnostic models is to use a transfer learning approach, where specific models are trained on individual diseases and then adapted for broader applications. Once refined, this type of agnostic model could also be applied to study metastasis and help trace the origin of cancers.

DISCUSSION

We presented an automated dry proteomics approach based on lipid MALDI-MSI, addressing several challenges to establish cluster-specific lipid and protein correlations in terms of imaging and pathways. Optimizations in the segmentation pipeline using SVD data compression, k-means++ segmentation and the Silhouette criterion enabled the correlation of multi-omics MALDI-MSI data. The integration of the Silhouette criterion proved useful for determining tissue heterogeneity, identifying the most optimized number of clusters in a fully automated and unsupervised manner.

Using the RB cerebellum tissue model, we demonstrated the workflow's suitability for lipid, protein, and peptide imaging, outperforming other segmentation algorithms. The robustness of our MS image processing model was confirmed through numerous experimental replicates. Multi-omics segmented images revealed the presence of RB cerebellum clusters ML, GL, and WM, each with specific spatial localizations, distinct lipid and protein compositions,

and associated biological pathways. A predictive model was developed based on these specific lipid fingerprints, complemented by the unique protein compositions and paired biological pathways of each cluster. Pathway analysis validated the dry proteomics approach for GL, ML, and WM.

To extend the prediction model beyond cerebellum regions to more complex tissues, RB horizontal slices were analyzed using multi-omics MALDI MSI. This analysis successfully identified several clusters with unique spatial localizations, including cerebellum clusters. The model, trained with cerebellum lipid datasets, effectively annotated these areas and provided insights into their specific lipids, proteins, and associated biological pathways. Further analysis refined the model for more accurate predictions, improving our understanding of complex tissue composition and highlighting the potential of dry proteomics to elucidate intricate biological processes.

Applying the dry proteomics workflow to glioblastoma patient cohorts provided profound insights into the spatial heterogeneity of peptides and lipids within the tumor microenvironment. By combining previous research data with cutting-edge imaging techniques, we uncovered previously unexplored complexity within GBM tissues. The segmentation process accurately delineated pathological features and revealed nuanced variations within tumor, necrotic, and inflammatory regions, providing a detailed representation of tissue composition.

The observed correlation between lipid and peptide distributions underscores their potential as robust biomarkers for tumor characterization. Our analysis revealed distinct molecular signatures within different tumor regions, indicating distinct biological processes and cellular compositions. Co-segmentation identified 13 discrete clusters among patients that corresponded to specific biological tissue regions. Proteomic data integration enriched our understanding of the molecular landscape within these clusters. Statistical analyses revealed significant protein expression variability across clusters, identifying distinct biological pathways. Indeed, the ClueGO analysis highlighted the involvement of different pathways, such as selenoamino acid metabolism or L1CAM interactions, which have a remarkable impact on GBM pathogenesis.

By integrating dry proteomics with prognosis, this study culminated in the development of a sophisticated classification model for GBM patients to identify the type of clusters and corresponding proteomic data with region-specific pathways and functions to stratify different prognostic categories. Two resulting GBM patient groups, A and B, were predicted using a model based on GBM lipid MSI that incorporated molecular heterogeneity within tumor tissues. The model was also translated into a proteomic model capable of distinguishing groups A and B based on protein data. It validated the model based on lipid MSI data and classified patients using only proteomic data. The protein networks of group A correlated with survival of more than 32 months, while those of group B correlated with

survival of less than 30 months. By classifying patients' tumors into groups A or B, we were able to predict tumor protein networks that correlated with classification group membership and survival prognosis.

In essence, this project highlights the importance of integrating multi-omics approaches for comprehensive prognostic assessment in GBM. By unraveling the interplay between molecular features and clinical outcomes, the developed model provides invaluable insights to inform personalized treatment strategies and improve patient management in the complex GBM landscape.

Finally, the concept of dry proteomics, which first identifies tissue heterogeneity and distinct clusters by lipid imaging and then automatically associates specific proteins and biological pathways involved in each cluster, has proven essential for clinical applications. These insights facilitate the identification of potential therapeutic targets or prognostic markers, as demonstrated in the glioblastoma study, paving the way for improved patient outcomes and personalized treatment strategies.

DATA AVAILABILITY

The data from this study, including MS DIA raw files, DIA-NN files, and annotated MS/MS datasets, have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD054488 (**Username:** reviewer_pxd054488@ebi.ac.uk - **Password:** vg74fZoVimg3)

Lipid MS/MS spectra are available at: https://doi.org/10.7910/DVN/MFKI8I

Code developed can be retrieved at: https://github.com/yanisZirem/Spatial_multi-omics_guided_by_SVD_kmeans_clustering_and_statistical_estimation_of_heterogeneity.git

The data from Duhamel, M, *& al,.* for glioblastoma study, including MS raw files, MaxQuant files, and annotated MS/MS datasets, have been deposited to the ProteomeXchange Consortium Via PRIDE partner repository with the accession code PXD016165.

SUPPLEMENTAL DATA

Supplementary information's represents the 36 supplementary figures.

CONFLICT OF INTEREST

The authors declare no competing interests.

ACKNOLEDGMENTS

This research was supported by grants from Ministère de l'Enseignement Supérieur et de la Recherche (MESR), Inserm specific funding for SpiderMass project (I.F.)**, Regional Council Hauts de France, The Metropole**

Européenne de Lille (MEL) and I-site ULNE (grant MutiOmics), Inserm and Institut Universitaire de France (I.F. and M.S.). L.R. PhD was funded by University of Lille. Y.Z. is supported by ANR Click-Detect AAP CE29 (designed and built the machine learning workflow). L.L., Y.Z., I.F., M.S. and E.LR. corrected the manuscript. I.F. and M.S. supervised the project and provided the funding.

REFERENCES

1. Quanico, J. *et al.* Development of liquid microjunction extraction strategy for improving protein identification from tissue sections. *J Proteomics* **79**, 200–218 (2013).

2. Wisztorski, M. *et al.* Microproteomics by liquid extraction surface analysis: application to FFPE tissue to study the fimbria region of tubo-ovarian cancer. *Proteomics Clin Appl* **7**, 234–240 (2013).

3. Delcourt, V. *et al.* Spatially-Resolved Top-down Proteomics Bridged to MALDI MS Imaging Reveals the Molecular Physiome of Brain Regions. *Mol Cell Proteomics* **17**, 357–372 (2018).

4. Lee, J., Musyimi, H. K., Soper, S. A. & Murray, K. K. Development of an automated digestion and droplet deposition microfluidic chip for MALDI-TOF MS. *J Am Soc Mass Spectrom* **19**, 964–972 (2008).

5. Kertesz, V., Weiskittel, T. M. & Van Berkel, G. J. An enhanced droplet-based liquid microjunction surface sampling system coupled with HPLC-ESI-MS/MS for spatially resolved analysis. *Anal Bioanal Chem* **407**, 2117–2125 (2015).

6. Quanico, J., Franck, J., Wisztorski, M., Salzet, M. & Fournier, I. Integrated mass spectrometry imaging and omics workflows on the same tissue section using grid-aided, parafilm-assisted microdissection. *Biochim Biophys Acta Gen Subj* **1861**, 1702–1714 (2017).

7. Sun, C. *et al.* Spatially resolved multi-omics highlights cell-specific metabolic remodeling and interactions in gastric cancer. *Nat Commun* **14**, (2023).

8. Dewez, F. *et al.* MS Imaging-Guided Microproteomics for Spatial Omics on a Single Instrument. *Proteomics* **20**, (2020).

9. Mezger, S. T. P., Mingels, A. M. A., Bekers, O., Heeren, R. M. A. & Cillero-Pastor, B. Mass Spectrometry Spatial-Omics on a Single Conductive Slide. *Anal Chem* **93**, 2527–2533 (2021).

10. Donnarumma, F. & Murray, K. K. Laser ablation sample transfer for localized LC-MS/MS proteomic analysis of tissue. *J Mass Spectrom* **51**, 261–268 (2016).

11. Lamont, L. *et al.* Integration of Ion Mobility MSE after Fully Automated, Online, High-Resolution Liquid Extraction Surface Analysis Micro-Liquid Chromatography. *Anal Chem* **89**, 11143–11150 (2017).

12. Quanico, J. *et al.* NanoLC-MS coupling of liquid microjunction microextraction for on-tissue proteomic analysis. *Biochim Biophys Acta Proteins Proteom* **1865**, 891–900 (2017).

13. Alexandrov, T. MALDI imaging mass spectrometry: statistical data analysis and current computational challenges. *BMC Bioinformatics* **13**, S11 (2012).

14. Fonville, J. M. *et al.* Robust data processing and normalization strategy for MALDI mass spectrometric imaging. *Anal Chem* **84**, 1310–1319 (2012).

15. Trim, P. J. *et al.* Matrix-assisted laser desorption/ionisation mass spectrometry imaging of lipids in rat brain tissue with integrated unsupervised and supervised multivariant statistical analysis. *Rapid Commun Mass Spectrom* **22**, 1503–1509 (2008).

16. Wang, Z. *et al.* Data-Driven Deciphering of Latent Lesions in Heterogeneous Tissue Using Function-Directed t-SNE of Mass Spectrometry Imaging Data. *Anal Chem* **94**, 13927–13935 (2022).

17. Abdelmoula, W. M. *et al.* Interactive Visual Exploration of 3D Mass Spectrometry Imaging Data Using Hierarchical Stochastic Neighbor Embedding Reveals Spatiomolecular Structures at Full Data Resolution. *J Proteome Res* **17**, 1054–1064 (2018).

18. Nijs, M., Smets, T., Waelkens, E. & De Moor, B. A mathematical comparison of non-negative matrix factorization related methods with practical implications for the analysis of mass spectrometry imaging data. *Rapid Commun Mass Spectrom* **35**, (2021).

19. Leuschner, J. *et al.* Supervised non-negative matrix factorization methods for MALDI imaging applications. *Bioinformatics* **35**, 1940–1947 (2019).

20. Deininger, S. O. *et al.* Normalization in MALDI-TOF imaging datasets of proteins: practical considerations. *Anal Bioanal Chem* **401**, 167 (2011).

21. Brunelle, A. & Laprévote, O. MALDI Imaging Mass Spectrometry. *Electrospray and MALDI Mass Spectrometry: Fundamentals, Instrumentation, Practicalities, and Biological Applications: Second Edition* 245–261 (2012) doi:10.1002/9780470588901.CH8.

22. Duda, R. O. & Peter, E. *Pattern Classi Fi Cation , 2nd Edition*. (2012).

23. Arthur, D. & Vassilvitskii, S. k-means++: The Advantages of Careful Seeding.

24. Nardecchia, A. *et al.* Detection of minor compounds in complex mineral samples from millions of spectra: A new data analysis strategy in LIBS imaging. *Anal Chim Acta* **1114**, 66–73 (2020).

25. Nguyen, T. N. Q., Jeannesson, P., Groh, A., Guenot, D. & Gobinet, C. Development of a hierarchical double application of crisp cluster validity indices: A proof-of-concept study for automated FTIR spectral histology. *Analyst* **140**, 2439–2448 (2015).

26. Bikfalvi, A. *et al.* Challenges in glioblastoma research: focus on the tumor microenvironment. *Trends Cancer* **9**, 9–27 (2023).

27. Zirem, Y. *et al.* Real-time glioblastoma tumor microenvironment assessment by SpiderMass for improved patient management. *Cell Rep Med* 101482 (2024) doi:10.1016/J.XCRM.2024.101482.

28. Duhamel, M. *et al.* Spatial analysis of the glioblastoma proteome reveals specific molecular signatures and markers of survival. *Nat Commun* **13**, (2022).

29. Caprioli, R. M., Farmer, T. B. & Gile, J. Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS. *Anal Chem* **69**, 4751–4760 (1997).

30. Hajjaji, N. *et al.* Path to Clonal Theranostics in Luminal Breast Cancers. *Front Oncol* **11**, 1 (2022).

31. Römpp, A. *et al.* imzML: Imaging Mass Spectrometry Markup Language: A Common Data Format for Mass Spectrometry Imaging. *Methods in Molecular Biology* **696**, 205–224 (2011).

32. Arthur, D. & Vassilvitskii, S. k-means++: The Advantages of Careful Seeding.

33. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* **20**, 53–65 (1987).

34. Ketkar, N. Stochastic Gradient Descent. *Deep Learning with Python* 113–132 (2017) doi:10.1007/978-1-4842-2766-4_8.

35. Dijkstra, T. K. Ridge regression and its degrees of freedom. *Qual Quant* **48**, 3185–3193 (2014).

36. Ke, G. *et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree.

37. Saudemont, P. *et al.* Real-Time Molecular Diagnosis of Tumors Using Water-Assisted Laser Desorption/Ionization Mass Spectrometry Technology. *Cancer Cell* **34**, 840-851.e4 (2018).

38. Ledoux, L. *et al.* Comparing MS imaging of lipids by WALDI and MALDI: two technologies for evaluating a common ground truth in MS imaging. *Analyst* **148**, 4982–4986 (2023).

39. Meier, F. *et al.* diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nature Methods 2020 17:12* **17**, 1229–1236 (2020).

40. Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091–1093 (2009).

41. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res* **13**, 2498 (2003).

42. Hastie, T., Rosset, S., Zhu, J. & Zou, H. Multi-class AdaBoost. *Stat Interface* **2**, 349–360 (2009).

43. Drucker, H. Improving Regressors using Boosting Techniques. *International Conference on Machine Learning* (1997).

44. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.

45. O'brien, J. S. *et al.* Lipid composition of the normal human brain: gray matter, white matter, and myelin". **5**, 329 (1965).

46. Gaud, C. *et al.* BioPAN: a web-based tool to explore mammalian lipidome metabolic pathways on LIPID MAPS. *F1000Res* **10**, 1–18 (2021).

47. Blusztajn, J. K., Liscovitch, M. & Richardson, U. I. Synthesis of acetylcholine from choline derived from phosphatidylcholine in a human neuronal cell line. *Proc Natl Acad Sci U S A* **84**, 5474–5477 (1987).

48. Lohner, K. Is the high propensity of ethanolamine plasmalogens to form non-lamellar lipid structures manifested in the properties of biomembranes? *Chem Phys Lipids* **81**, 167–184 (1996).

49. Glaser, P. E. & Gross, R. W. Rapid Plasmenylethanolamine-Selective Fusion of Membrane Bilayers Catalyzed by an Isoform of Glyceraldehyde-3-Phosphate Dehydrogenase: Discrimination between Glycolytic and Fusogenic Roles of Individual Isoforms. *Biochemistry* **34**, 12193–12203 (1995).

50. Tracey, T. J., Kirk, S. E., Steyn, F. J. & Ngo, S. T. The role of lipids in the central nervous system and their pathological implications in amyotrophic lateral sclerosis. *Semin Cell Dev Biol* **112**, 69–81 (2021).

51. Mallah, K. *et al.* Neurotrauma investigation through spatial omics guided by mass spectrometry imaging: Target identification and clinical applications. *Mass Spectrom Rev* **42**, 189–205 (2023).

52. Mallah, K. *et al.* Mapping Spatiotemporal Microproteomics Landscape in Experimental Model of Traumatic Brain Injury Unveils a link to Parkinson's Disease. *Mol Cell Proteomics* **18**, 1669–1682 (2019).

53. Shulgin, A. A., Lebedev, T. D., Prassolov, V. S. & Spirin, P. V. Plasmolipin and Its Role in Cell Processes. *Mol Biol* **55**, 773–785 (2021).

54. Li, Z. H., Li, B., Zhang, X. Y. & Zhu, J. N. Neuropeptides and Their Roles in the Cerebellum. *Int J Mol Sci* **25**, (2024).

55. Evangelisti, C. *et al.* The wide and growing range of lamin B-related diseases: from laminopathies to cancer. *Cell Mol Life Sci* **79**, (2022).

56. Ayasoufi, K. *et al.* Brain resident memory T cells rapidly expand and initiate neuroinflammatory responses following CNS viral infection. *Brain Behav Immun* **112**, 51–76 (2023).

57. Batterman, K. V., Cabrera, P. E., Moore, T. L. & Rosene, D. L. T Cells Actively Infiltrate the White Matter of the Aging Monkey Brain in Relation to Increased Microglial Reactivity and Cognitive Decline. *Front Immunol* **12**, (2021).

58. Gianola, S., Savio, T., Schwab, M. E. & Rossi, F. Cell-autonomous mechanisms and myelin-associated factors contribute to the development of Purkinje axon intracortical plexus in the rat cerebellum. *Journal of Neuroscience* **23**, 4613–4624 (2003).

59. Hulet, S. W., Powers, S. & Connor, J. R. Distribution of transferrin and ferritin binding in normal and multiple sclerotic human brains. *J Neurol Sci* **165**, 48–55 (1999).

60. Hulet, S. W., Menzies, S. & Connor, J. R. Ferritin Binding in the Developing Mouse Brain Follows a Pattern Similar to Myelination and Is Unaffected by the Jimpy Mutation. *Dev Neurosci* **24**, 208–213 (2002).

61. Kirilina, E. *et al.* Superficial white matter imaging: Contrast mechanisms and whole-brain in vivo mapping. *Sci Adv* **6**, (2020).

62. Ma, M. *et al.* Molecular layer interneurons in the cerebellum encode for valence in associative learning. *Nature Communications 2020 11:1* **11**, 1–16 (2020).

63. Yamazaki, Y. *et al.* Neural changes in the primate brain correlated with the evolution of complex motor skills. *Sci Rep* **6**, (2016).

64. Harvey, R. J. Cerebellar regulation in movement control. *Trends Neurosci* **3**, 281–284 (1980).

65. Javed, K., Reddy, V. & Lui, F. Neuroanatomy, Cerebral Cortex. *StatPearls* (2023).

66. Firdaus, I. de Lahunta's Veterinary Neuroanatomy and Clinical Neurology. *Alexander de Lahunta, Eric Glass, Marc Kent* (2020).

67. Banks, W. A. *et al.* Triglycerides cross the blood-brain barrier and induce central leptin and insulin receptor resistance. *Int J Obes (Lond)* **42**, 391–397 (2018).

68. de Curtis, I. The Rac3 GTPase in Neuronal Development, Neurodevelopmental Disorders, and Cancer. *Cells* **8**, (2019).

69. Chédotal, A., Kerjan, G. & Moreau-Fauvarque, C. The brain within the tumor: new roles for axon guidance molecules in cancers. *Cell Death & Differentiation 2005 12:8* **12**, 1044–1056 (2005).

70. Cirella, A. *et al.* Novel strategies exploiting interleukin-12 in cancer immunotherapy. *Pharmacol Ther* **239**, 108189 (2022).

71. Schonthal, A. H. *et al.* The Extracellular Matrix in Glioblastomas: A Glance at Its Structural Modifications in Shaping the Tumoral Microenvironment-A Systematic Review. (2023) doi:10.3390/cancers15061879.

72. Hu, H. & Laskin, J. Emerging Computational Methods in Mass Spectrometry Imaging. *Adv Sci (Weinh)* **9**, (2022).

FIGURE LEGENDS

Fig. 1: Omics MALDI MSI clustering procedure optimization on rat brain cerebellum. A) Comparison of t-SNE, NNMF and SVD data compression followed by *k*-means++ segmentation for 2 to 5 clusters applied to lipid negative mode, lipid positive mode, protein, and peptide MSI. **B)** Rat brain sagittal section HPS coloration and cerebellum annotations. **C)** Lipid MALDI MSI in negative and positive mode with 10 µm spatial resolution with image segmentation composed by 5 clusters, and ion spatial distribution specific to Purkinje cells, ML, GL and WM. **D)** Use of Silhouette criterion for the number of cluster estimation and each cluster value determination applied to lipid negative mode, lipid positive mode, protein, and peptide imaging. **E)** Optimal segmentation workflow developed on MATLAB integrating SVD compression data with 10 principal components, combined with a *k*-means++ segmentation using a cosine score with a Silhouette criterion

Fig. 2: Discriminant lipid and protein ions present in RB cerebellum with BioPAN lipid pathways. Exhaustive list of **A)** 36 lipid (-), **B)** 19 lipid (+) and **C)** protein discriminant ML, GL, and WM cerebellum ions. **D)** BioPAN biological lipid pathways involved in white matter represented according to lipid species and lipid classes, with nodes legend.

Fig. 3: Rat brain cerebellum regions spatial proteomic analysis. A) Venn diagram of the specific proteins per layer. **B)** Heatmap after ANOVA (p-value <0.01) analysis demonstrated the presence of different of overexpressed proteins. ClueGO biological pathways involving the significant proteins found in **C)** granular layer**, D)** white matter, and **E)** molecular layer of the cerebellum.

Fig. 4: Horizontal rat brain section omics MALDI MSI analysis. A) Lipid (-), lipid (+), protein and peptide MSI segmentation images with 11 clusters and Silhouette criterion. **B)** Clusters mean scores prediction based on rat brain cerebellum lipid (-) model. **C)** Clusters Pearson's correlation. **D)** Prediction lipid (-) model peaks involvement.

Fig. 5: Spatial proteomic analysis of rat brain horizontal clusters. A) 10 different clusters identified thanks to lipid (-) lipid MSI and spatial proteomic extraction points. **B)** Protein Venn diagram. **C)** Heatmap after ANOVA (p-value 0.0001) analysis demonstrated the presence of different of overexpressed proteins.

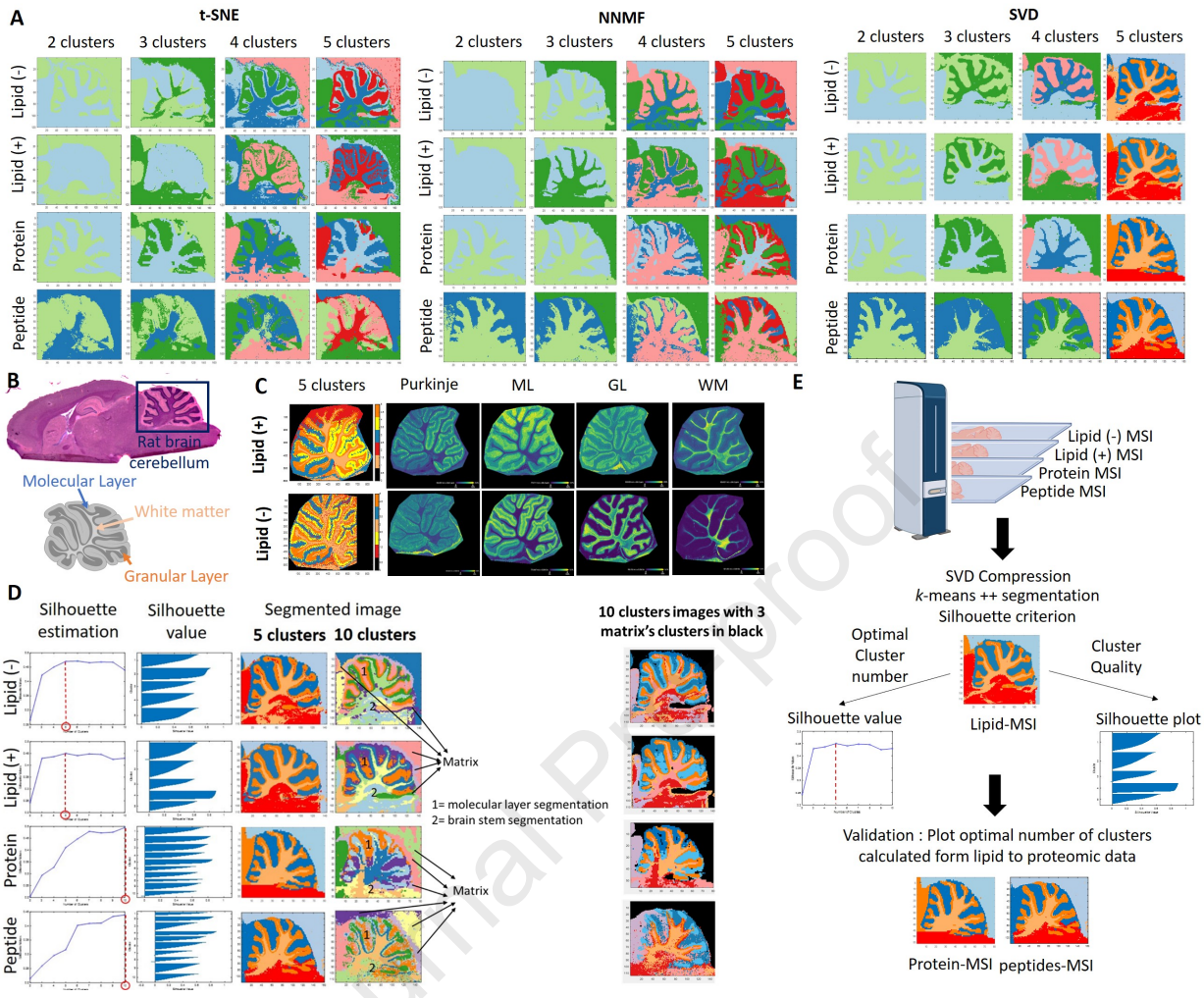
Fig. 6: Glioblastoma patient lipid and protein heterogeneity analysis. A) P9 and P12 lipid and peptide MSI with histopathological annotations. **B)** Co-segmentation of 9 tissues previously analyzed by lipid MALDI MSI. **C)** t-SNE representation of each cluster identified through lipid co-segmentation. **D)** Protein heat map after ANOVA (p-

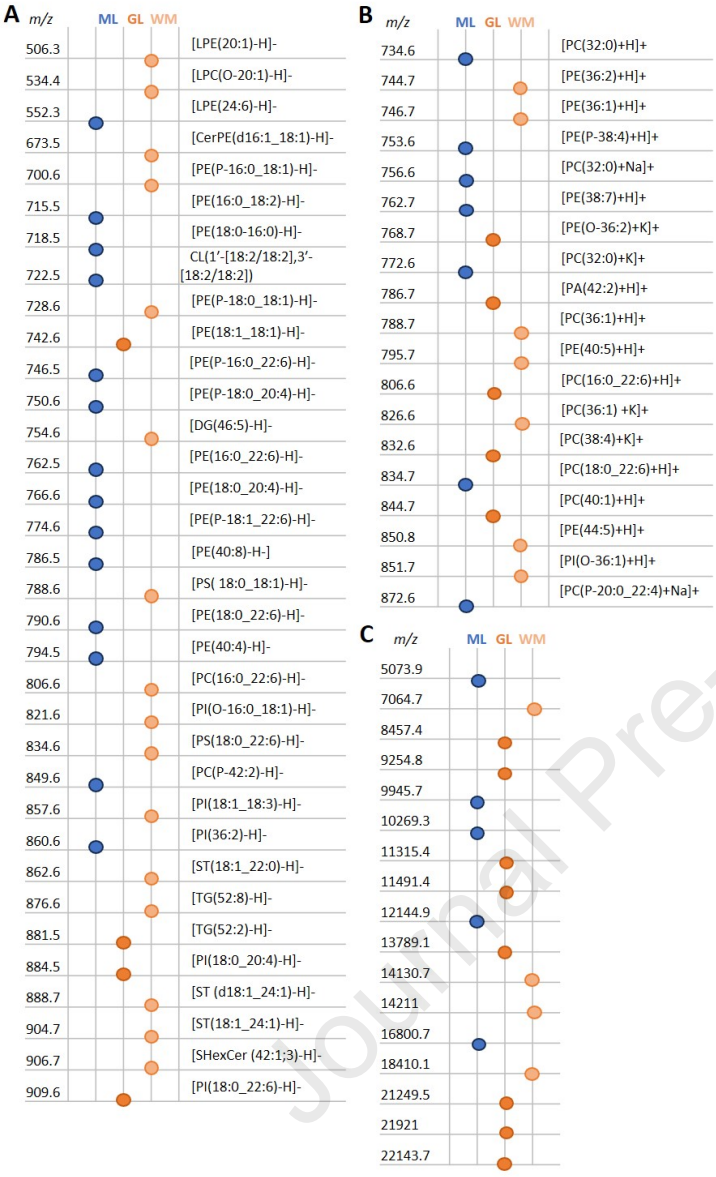
value 0.01) analysis demonstrating the presence of different of over-expressed proteins according lipid clusters. **E)** Patient classification group A and B prediction according lipid and protein model. **F)** Lipid cluster and associated protein blind prediction on patient P3, P5, P6 and P11.

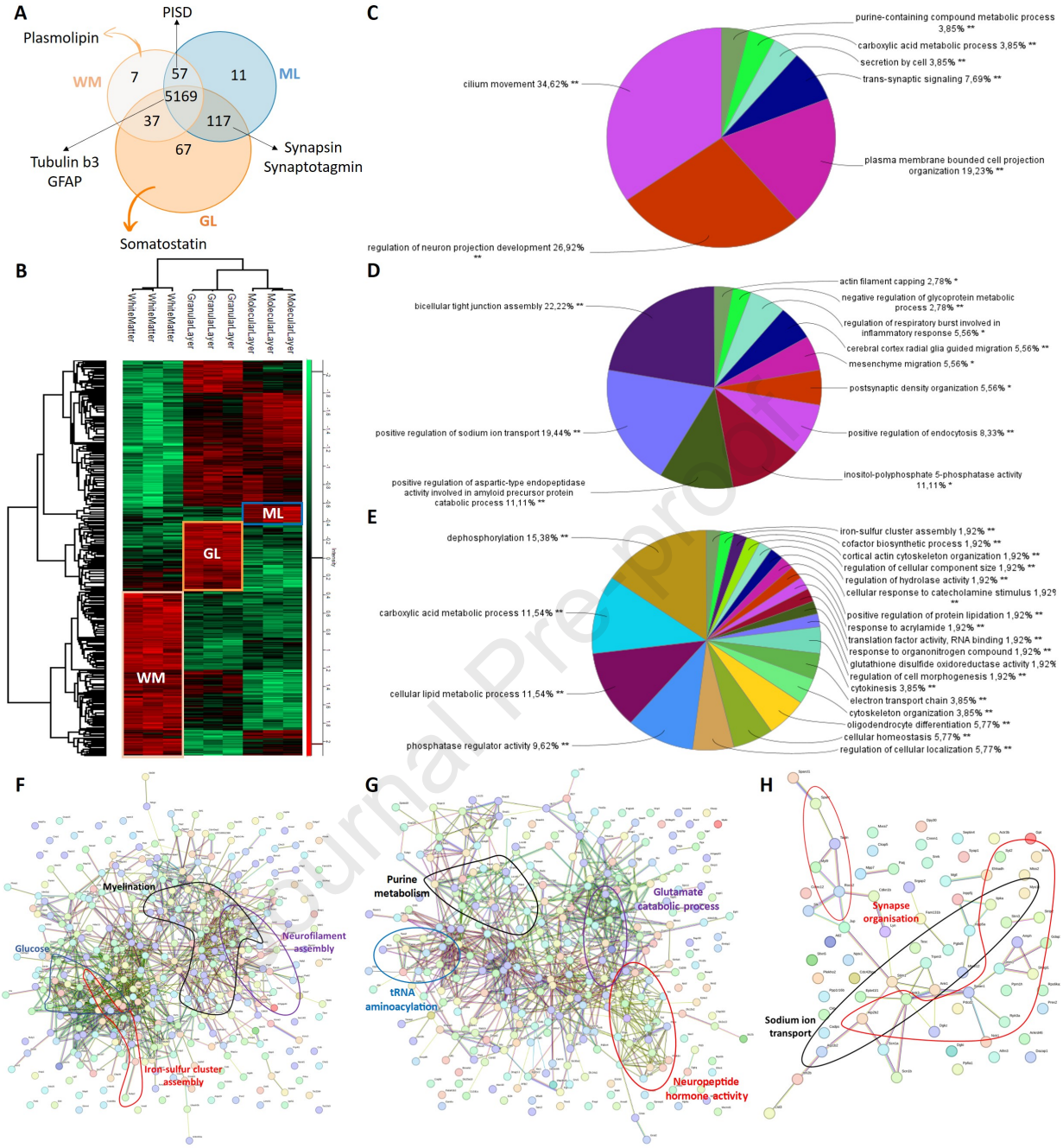
TABLE

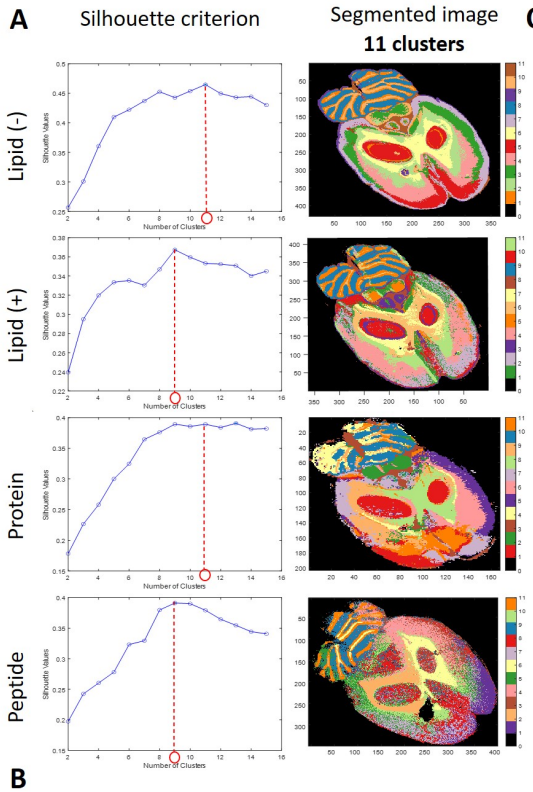
Table 1: Model algorithms implication

Model	*Algorithm*	*F1 score*
RB cerebellum clusters lipid (-)	SGD	94%
RB cerebellum clusters lipid (+)	RidgeClassifier	98%
GBM lipid classification	LGBM	97%
GBM protein classification	RidgeClassifier	96%



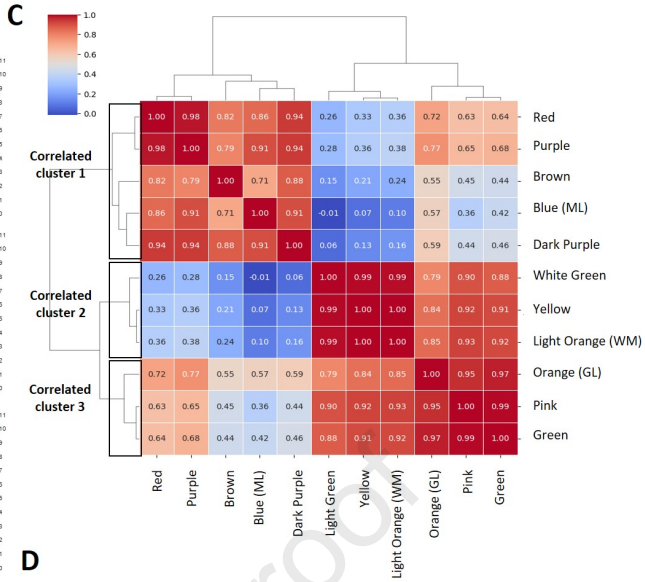


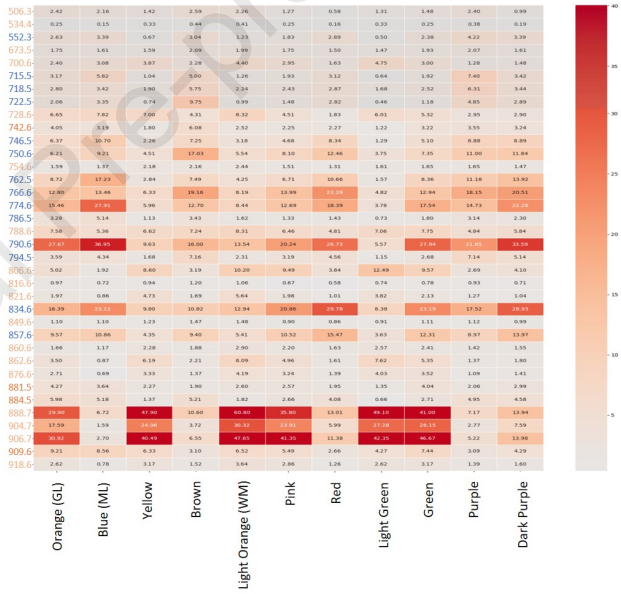




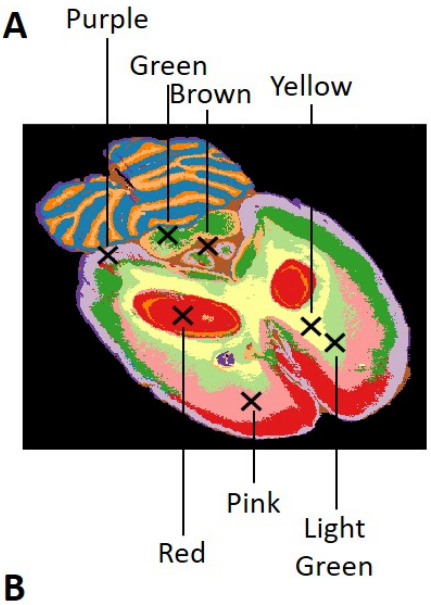
B

Index	GL	ML	WM
Orange (GL)	**89%**	9%	2%
Light Green	25%	0%	75%
Green	71%	13%	16%
Pink	61%	16%	21%
Red	5%	90%	5%
Yellow	39%	0%	61%
Purple	11%	81%	8%
Blue (ML)	0%	**100%**	0%
Dark Purple	26%	74%	0%
Light Orange (WM)	43%	5%	**52%**
Brown	33%	65%	2%





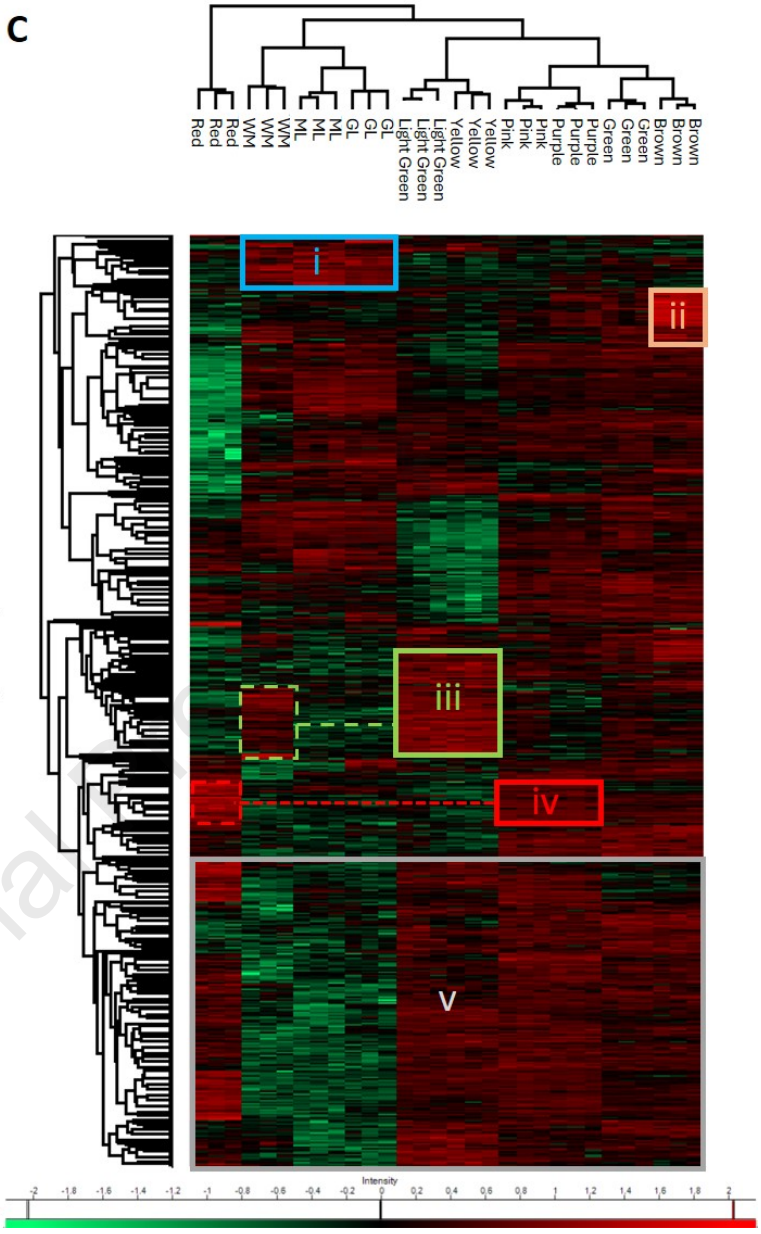
A

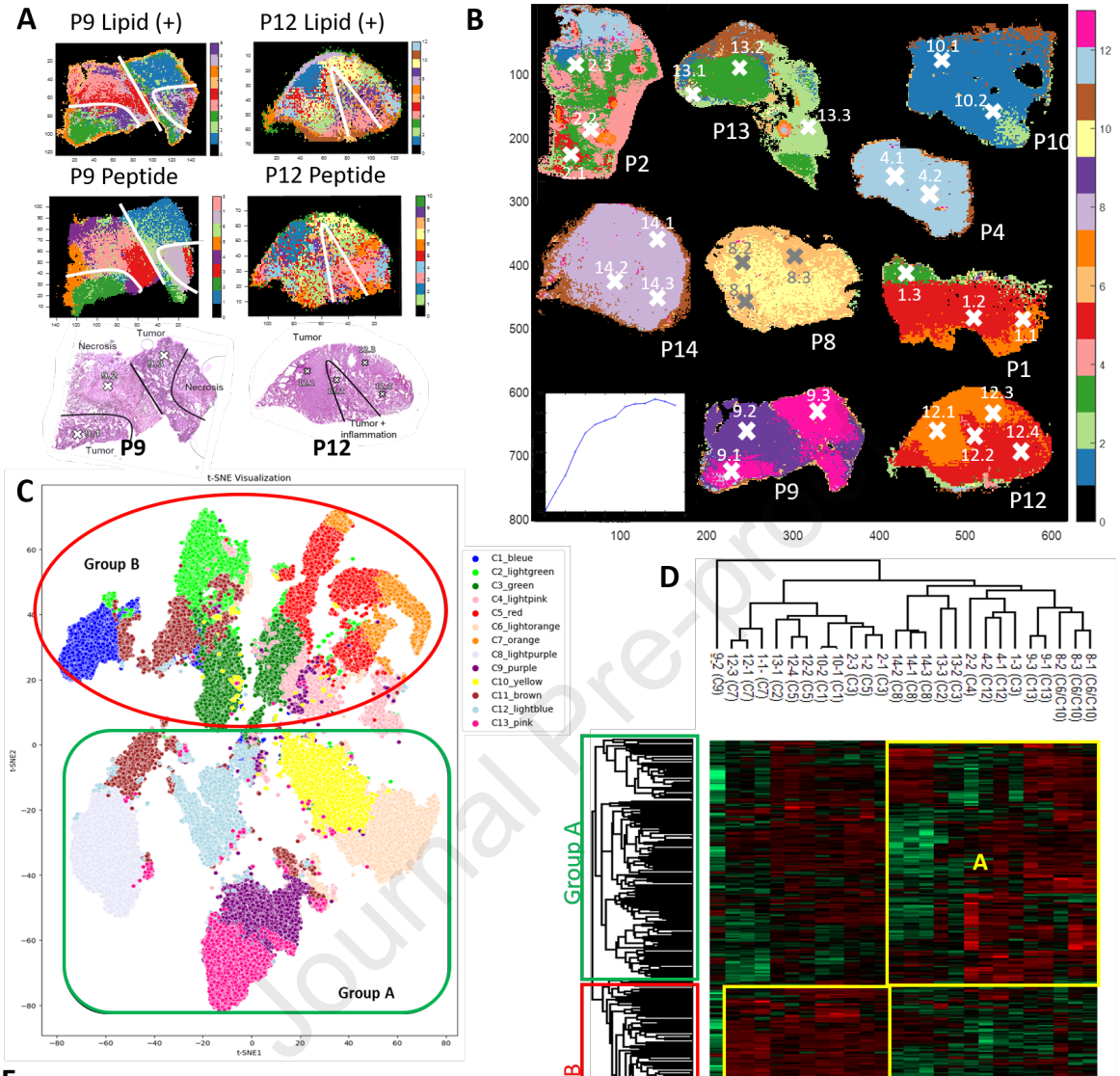


B

Cluster	Total proteins	Exclusive proteins
WM	6133	0
Light Green	6667	0
Yellow	6524	1
GL	6463	29
Pink	6918	0
Green	7045	3
ML	6457	1
Red	6606	4
Brown	6983	5
Purple	6946	0

C





F

Image lipid clusters		Cluster prediction according lipid (-) model and associated proteins pathways			
		Patient 3	Patient 5	Patient 6	Patient 11
C1 Blue	B	/	/	/	/
C2 Light green	B	/	/	**22%**	3%
C3 Green	B	/	/	/	4%
C4 Light pink	B	**39%**	5.4%	**25%**	1.3%
C5 Red	B	/	/	/	2%
C6 Light orange	A	34%	**75%**	50%	/
C7 Orange	B	/	/	/	/
C8 Light purple	A	0.1%	0.2%	/	**51%**
C9 Purple	A	/	/	/	0.3%
C10 Yellow	A	/	/	/	/
C11 Brown	NA	6.7%	19%	3%	36%
C12 Light blue	A	/	0.3%	/	1.5%
C13 Pink	A	21%	/	/	1.1%
Main biological associated pathways		-Wound healing -Regulation of protein-containing complex assembly -ECM proteoglycans -Selenoamino acid metabolism	-Serine family amino acid metabolic -Interleukin-12 family signaling	-Establishment of spindle localization -Wound healing -ECM proteoglycans -Interleukin-12 family signaling	-Peptide chain elongation -RHO GTPase active ROCKS - ATP dependent protein folding chaperone
Patient classification		B	A	B	A

E

Patient	*Classification group according heterogeneity*	*Classification group according original study*
P1	Group B	Group B
P2	Group B	Group B
P4	Group A	Group A
P8	Group A	Group A
P9	Group A	Group A
P10	Group B	Group A
P12	Group B	Group A
P13	Group A	Group A
P14	Group A	Group A

HIGHLIGHTS

- Tissue heterogeneity assessment pipeline was developed based on lipid MALDI MSI.
- Artificial Intelligence model predicts lipid clusters with associated proteomes and biological pathways.
- Developed strategy applied to glioblastoma deciphered heterogeneity for prognosis.
- Dry proteomics: rapid, robust cancer tissues analysis for theragnostic management.

IN BRIEF

This study introduces a robust "dry proteomics" workflow based on lipid MALDI MSI, validated on rat brain tissues and applied to human glioblastoma. The innovative multi-omics MSI data analysis addresses a crucial gap in spatial data processing by integrating tissue heterogeneity assessment. Notably, SVD data compression, ***k*-**means++ segmentation and silhouette criterion yielded optimal results. This workflow offers novel insights into glioblastoma biology and patient survival, presenting promising tool for clinical studies and patient theragnostic management, marking a significant advancement in cancer research.

Heterogeneity Assessment and Protein Pathway Prediction via Spatial Lipidomic and Proteomic Correlation: Advancing Dry Proteomics concept for Human Glioblastoma Prognosis

Laurine Lagache[1†], Yanis Zirem[1†], Émilie Le Rhun[1,2], Isabelle Fournier[1,3*†] and Michel Salzet[1,3*†]

[1]Univ.Lille, Inserm, CHU Lille, U1192 – Proteomics Inflammatory Response Mass Spectrometry-PRISM, F-59000 Lille, France

[2]Department of Neurosurgery and Neurology, Clinical Neuroscience Center, University Hospital Zurich and University of Zurich, Zurich, Switzerland

[3]Institut Universitaire de France, 75000 Paris

†Equal contribution

*Co-corresponding

CONFLICT OF INTEREST

The authors declare no competing interests.

Résumé

Le cancer demeure un défi majeur pour la santé mondiale, avec près de 10 millions de décès et 19 millions de nouveaux cas signalés en 2020. Le cancer du sein est le type le plus fréquent chez les femmes, avec 2,26 millions de cas et environ 685 000 décès la même année. Le diagnostic repose souvent sur des échantillons de biopsie analysés par histopathologie pour identifier les caractéristiques moléculaires. Le cancer du sein est classé en sous-types selon divers biomarqueurs, ce qui permet d'orienter les traitements. Cependant, environ 30 % des patientes subissent encore des récidives ou des métastases en raison de l'hétérogénéité tumorale, causée par des sous-populations génétiquement diverses au sein des tumeurs, entraînant des réponses thérapeutiques variées. Cette étude vise à aborder cette complexité en améliorant la caractérisation des tumeurs grâce à l'analyse protéomique, en se concentrant sur la diversité moléculaire du cancer du sein. En utilisant la spectrométrie de masse, en particulier la technique MALDI MSI, la recherche examine les sous-populations tumorales pour identifier des cibles thérapeutiques potentielles et améliorer les approches de traitement personnalisé. Les premiers résultats obtenus à partir d'échantillons tumoraux de patientes montrent que les traitements guidés par la protéomique sont plus efficaces que les thérapies conventionnelles et permettent d'identifier des biomarqueurs de résistance aux médicaments. Cependant, des défis tels que la nécessité de grandes quantités de matériel biologique et le temps requis pour cette méthode limitent sa mise en œuvre à grande échelle en clinique. Pour simplifier le processus, l'étude explore également un modèle d'apprentissage automatique permettant de prédire les informations protéiques à partir de l'analyse lipidique via MALDI MSI, réduisant ainsi le besoin d'expériences protéomiques distinctes. Cette approche, validée sur des tissus cérébraux de rat et sur le glioblastome, pourrait accélérer l'analyse tumorale et rendre la technique plus accessible en pratique clinique. De plus, l'étude applique cette approche de 'dry proteomic' pour mieux comprendre l'hétérogénéité spatiale et temporelle du cancer du sein, dans le but d'identifier des cibles thérapeutiques à différents stades de la maladie. Enfin, une nouvelle technique de multiplexage MALDI IHC est en cours de développement pour détecter rapidement les principales cibles protéiques et cartographier leurs interactions spatiales, notamment dans le contexte de l'immunothérapie. Dans l'ensemble, cette étude vise à améliorer le traitement du cancer du sein en intégrant des techniques de protéomique, d'apprentissage automatique et d'imagerie, pour aborder l'hétérogénéité tumorale et améliorer les stratégies thérapeutiques personnalisées.

Summary

Cancer remains a global health challenge, with nearly 10 million deaths and 19 million new cases reported in 2020. Breast cancer is the most common type among women, responsible for 2.26 million cases and about 685,000 deaths in the same year. Diagnosis often involves biopsy samples analyzed through histopathology to identify molecular characteristics. Breast cancer is classified into subtypes based on various biomarkers, helping to guide treatment. However, around 30% of patients still face recurrence or metastasis due to tumor heterogeneity, which stems from genetically diverse subpopulations within tumors, leading to varied treatment responses. This study aims to address this complexity by enhancing tumor characterization through proteomic analysis, focusing on breast cancer's molecular diversity. Using mass spectrometry, specifically MALDI MSI, the research examines tumor subpopulations to identify potential therapeutic targets and improve personalized treatment approaches. Early results from patient tumor samples show that proteomics-guided treatments are more effective than conventional therapies and help identify drug resistance biomarkers. However, challenges such as the need for large biological samples and the time-intensive nature of this method limit its clinical scalability. To streamline the process, the study also explores a machine learning model to predict protein information from lipid analysis using MALDI MSI, reducing the need for separate proteomics experiments. This approach, validated in rat brain tissues and glioblastoma, could speed up tumor analysis and make the technique more accessible for clinical use. Further, the study applies this "dry proteomics" approach to better understand breast cancer's spatial and temporal heterogeneity, aiming to identify treatment targets at different disease stages. Lastly, a new MALDI IHC multiplex technique is being developed to rapidly detect key protein targets and map their spatial interactions, particularly in the context of immunotherapy. Overall, this research seeks to improve breast cancer treatment by integrating proteomics, machine learning, and imaging techniques, addressing tumor heterogeneity, and enhancing personalized therapeutic strategies.