Université de Strasbourg

UNIVERSITÉ DE STRASBOURG

EDSC École Doctorale des Sciences Chimiques

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

IPHC – UMR 7178

THÈSE présentée par :

Marie GEBELIN

soutenue le : **29 septembre 2023**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Chimie Analytique

Développement de méthodes d'analyse protéomique et phosphoprotéomique à haut débit et leur application pour la recherche de biomarqueurs de pathologies sur de larges cohortes

THÈSE dirigée par :

Dr. CARAPITO Christine	Directrice de recherche, CNRS, Strasbourg
Dr. SCHAEFFER Christine	Ingénieure de recherche, CNRS, Strasbourg

RAPPORTEURS :

[Dr. PINEAU Charles]	Directeur de recherche, INSERM, Université de Rennes
[Pr. HIRTZ Christophe]	Professeur, Université de Montpellier

AUTRES MEMBRES DU JURY :

[Dr. PFLIEGER Delphine]	Chargée de recherche, CNRS, Grenoble

Université de Strasbourg

UNIVERSITÉ DE STRASBOURG

EDSC École Doctorale des Sciences Chimiques

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

IPHC – UMR 7178

THÈSE présentée par :

Marie GEBELIN

soutenue le : **29 septembre 2023**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Chimie Analytique

Development of high throughput proteomic and phosphoproteomic analytical methods and their application for pathologies' biomarker discovery on large cohorts

THÈSE dirigée par :

Dr. CARAPITO Christine	Directrice de recherche, CNRS, Strasbourg
Dr. SCHAEFFER Christine	Ingénieure de recherche, CNRS, Strasbourg

RAPPORTEURS :

Dr. PINEAU Charles	Directeur de recherche, INSERM, Université de Rennes
Pr. HIRTZ Christophe	Professeur, Université de Montpellier

AUTRES MEMBRES DU JURY :

Dr. PFLIEGER Delphine	Chargée de recherche, CNRS, Grenoble

À mes parents,

« Le pire n'est pas d'échouer mais de n'avoir jamais pris la peine d'essayer. Les erreurs sont une preuve que tu n'as jamais abandonné. »

La Passeuse de Mots, Alric & Jennifer Twice

Acknowledgments

Tout d'abord, je souhaite remercier Sarah Cianferani, de m'avoir offert l'opportunité de réaliser ma thèse au sein du laboratoire de Spectrométrie de Masse BioOrganique de l'Institut Pluridisciplinaire Hubert Curien (IPHC, UMR7178). C'est un cadre de travail de qualité pour se former à la spectrométrie de masse et au travail de recherche, et propice à l'entraide entre chercheurs.

Je tiens tout particulièrement à remercier mes directrices de thèse, le duo de Christines, Christine Carapito et Christine Schaeffer, de m'avoir encadrée pendant ces trois années de thèse. Christine S., merci de m'avoir accueillie, moi et mon stress légendaire, pour me faire découvrir le monde de la spectrométrie de masse et des phosphorylations pendant mon stage de master. Merci pour ta bienveillance, tes conseils avisés, et ton écoute. Christine C., merci de m'avoir embarquée en cours de route dans ce projet MAXOMOD, qui en plus de m'avoir formée en tant que protéomiste, m'a permis de faire de superbes rencontres scientifiques. Merci pour ton encadrement, ta confiance et ta gentillesse. Je ressors grandie de cette expérience à vos côtés, et je vous en suis grandement reconnaissante.

Je remercie chaleureusement Delphine Pflieger, Charles Pineau, et Christophe Hirtz, d'avoir eu la gentillesse de lire mon manuscrit et d'évaluer mon travail de thèse.

Je souhaite également remercier l'ensemble des collaborateurs avec lesquels j'ai pu travailler pendant ces trois années. Notamment l'ensemble du consortium MAXOMOD, mais plus particulièrement Paul Lingor, Lucas Caldi-Gomes, Laura Tzeplaeff, Sergio Oller, Sonja Hänzelmann. Je remercie également l'ensemble de l'équipe du laboratoire CompOmics, mais tout particulièrement : Lennart Martens, Nina Demeulemeester, Tine Claeys, Arthur Declercq, et Pathmanaban Ramasamy. Enfin, je souhaite également remercier Hubert Schaller, Marc Graille, Catarina Paiva et Maria Zeniou.

Le plus grand merci à l'ensemble du LSMBO, pour l'entraide, pour votre accueil, et pour tous les bons moments partagés. Merci Laurence, par tes cours à l'ECPM, de m'avoir fait découvrir le monde de la spectrométrie de masse et de la protéomique. Je pense à Agnès, Hélène, Véronique (merci de m'avoir acceptée comme colocataire pendant quelques mois), Fabrice. Merci Delphine pour ton aide précieuse sur les phosphos. Merci Martine pour ton aide pour toutes les démarches administratives. Magali, François, merci d'avoir toujours pris le temps de répondre à mes (nombreuses) questions sur Proline, MaxQuant, et PRIDE. Thanks Sarahi for your help for manipulations, and thanks for all the delicious sweets you brought. Valériane, je te confie le bébé Bravo, je suis sûre que tu sauras très bien t'en occuper.

Merci Alex, de reprendre le flambeau du HF-X, et pour ta bonne humeur malgré ses caprices souvent le soir à 18h. C'était un plaisir de te former dessus, je n'ai aucun doute que tu vas être un parfait membre de la confrérie des responsables HF-X. Je ne peux pas parler du HF-X sans remercier Jean-Marc, alias l'homme qui chuchote à l'oreille des chromatos et spectros. Merci pour ta disponibilité et ton aide sur les machines, et surtout avec ces fameuses aiguilles.

Acknowledgements

Merci à la team bioinfo, pour votre aide précieuse. Fabrice V., Alex B., merci pour votre disponibilité et réactivité lors de mes (très) nombreux soucis de session. Adrien, merci pour ton « iiincroyable » aide avec les codes R et Python.

Un immense merci au bureau des goûters, mon bureau pendant ces trois années. Merci Aurélie, Jeewan, Noélia, Bastien (si si je te compte ici même si ce n'était qu'un court séjour), pour votre bonne humeur, pour ces nombreux goûters remplis de rire. Bastien, tu nous as rejoint au Club Jeunes avant même de nous rejoindre au labo, et maintenant tu peux être fier d'être à la fois le président du CJ et le champion invétéré du ping-pong. Jeewan, thanks for always being in a good mood, even when I was asking you for the hundredth time the same question about the Tims.

Merci à la team des supramoelleux, Hugo, Oscar, Jérôme, Rania. Jérôme, ce fut un plaisir de m'engager avec toi pendant quelques temps le bureau des doctorants. Ne t'inquiète pas, si tu as besoin d'un compagnon pour une coupe de crémant, tu pourras toujours compter sur moi. Ma chère Rania, merci d'avoir toujours été à l'écoute et d'avoir toujours été là pour me remonter le moral en chantant à tue-tête sur Taylor Swift. Merci pour les rires, et surtout pour ton cœur en or. Merci de t'être occupée d'Aria de la meilleure façon qui soit : en la faisant danser sur Taylor Swift.

Je n'oublie pas ceux qui ont quitté le laboratoire, et notamment, Marie C., Marie L., Marziyeh, Joanna, Kévin, Nathan. Nathan merci pour ta bonne humeur (bien cachée sous ta bougonnerie) et tes blagues. Marie L. merci pour ton écoute et tes conseils.

Marie C., on a partagé tellement de choses que je ne saurais pas par où commencer. Merci d'avoir toujours été là pour me faire une place sous le plaid du déni. Merci pour tous ces moments à parler de tout et rien : de nos connaissances médicales acquises grâce à Grey's Anatomy (on est presque médecins en fait), de photos de chats mignons, sans oublier de toutes les infos et théories possibles et imaginables sur notre chère TayTay. On se revoit vite pour de superbes moments en 2024 à Lyon et à Londres (et à Paris … ?). Pour résumer l'essentiel : « *All I can say is, I was enchanted to meet you ».*

Merci au bureau du fond ou bureau du fun (même si clairement moins fun que le bureau des goûters). Charline, je suis ravie que tu ais rejoins le labo pour les trois prochaines années et te souhaite bonne chance pour la suite. Je suis sûre que tu seras très bien entourée par cette fine équipe.

Pauline, c'est un plaisir de travailler avec toi et ta bonne humeur communicative. Merci d'avoir été ma confidente et ma *partner in crime* dans la folie Taylor Swift. Je compte sur toi pour continuer de mettre du rose et des paillettes partout où tu peux, et continuer à sauver des petites grenouilles en détresse. Je suis plus qu'heureuse de t'avoir rencontrée, et d'avoir partagé ces moments, pro et perso, avec toi. Saches que tu pourras toujours compter sur moi pour qu'on continuer à partager d'autres moments ensemble, que ce soit aller acheter nos légumes au marché ensemble, aller boire un verre (parce qu'on aime bien ça quand même), ou allez voir TayTay en concert (see you in London).

J'ai gardé le trio gagnant pour la fin : Charlotte, Aurélie, et Corentin. On aura peut-être toujours pas réussi à se faire ce week-end tous les 4 mais je ne perds pas espoir car j'espère qu'on partagera encore des moments tous ensemble. Ces trois années sont remplies de riches souvenirs à vos côtés, que je ne suis pas prête d'oublier. Merci pour tous ces bonnes soirées autour d'une table et de verres bien remplis, merci pour tous ces fous rire, merci pour ces petits voyages partagés à l'autre bout de la France ou au fin fond de la contrée allemande, et merci d'avoir supporté tous mes « on est d'accords que ? ». Aurélie, merci, dès mes premiers jours au labo, de m'avoir accueillie à bras ouverts. Merci de m'avoir

transmis toutes tes connaissances en salle bio. Merci d'être la maman du groupe, à toujours avoir de quoi nous nourrir (ou nourrir un régiment), à nous border après une soirée, à toujours être à l'écoute de nos petits (et gros) soucis. On se souviendra encore longtemps de ton sommeil de plomb qui a failli nous coûter une nuit sur le balcon de l'hôtel, mais surtout du fou rire qui s'en est suivi. Bon allez maintenant que j'ai fini la thèse on reprend le sport : tu préfères quoi, on se (re)met une énième fois à la piscine ou on tente enfin l'ultimate ? Corentin, on t'embête parfois mais on sait que tu nous aimes bien quand même dans le fond. Merci d'avoir été là pour répondre à toutes mes questions sur R et de toujours être là pour nous rappeler les paroles des classiques français en fin de soirée. Charlotte, merci de m'avoir pris sous ton aile à mon arrivée au labo, que ce soit pour me former sur la phospho, le bébé Bravo, ou le HF-X. C'était un plaisir d'avoir travaillé pendant 1 an et demi sur la phospho avec toi, à avoir des idées émerger un samedi soir à 22h autour d'un verre. Nos MacDo réconfortants, pendant une longue soirée sur le Bravo, ou pendant une journée difficile, m'ont beaucoup manqué cette dernière année. Merci d'avoir été là pour moi, pour partager ces moments tant professionnels que personnels. Pour résumer, je suis très heureuse de vous avoir rencontrés et d'avoir fait partie de notre quatuor de folie (#Papi, maman et les deux sales gosses).

Un énorme merci à ma famille, mes amis et mes proches qui m'ont soutenue pendant ces trois années et bien plus.

Merci à la bande des Trouffois, pour tous ces moments de galères partagés (quelle idée de tous faire une thèse), pour ces soirées déguisées inoubliables, et cette super escapade au pays de la Sangria.

Madelaine, Marie, Tamara, merci pour votre présence et votre soutien depuis toutes ces années, surtout les trois dernières, malgré la distance. Merci d'être là pour les Skypes interminables, pour les fous rires à se plier par terre, pour les séjours déjantés que ce soit Au Petit Goût De ou bien dans des lieux douteux au milieu de la forêt de Twilight, et les milliers d'autres moments qu'on a pu partager ces 3 dernières années et qu'on continuera à partager dans les années à venir.

Mes plus tendres remerciements à toi, Théo. Merci de m'avoir soutenue durant les moments difficiles, merci de me redonner le courage et le sourire quand j'en ai eu besoin, malgré les kilomètres qui nous séparent. Je suis chanceuse de t'avoir à mes côtés, et j'ai hâte de continuer cette aventure avec toi.

Je tiens enfin à remercier du fond du cœur mes parents, pour leur soutien incommensurable ces 26 dernières années mais surtout ces 3 dernières années. Je ne serai pas où j'en suis aujourd'hui sans vous, pour votre écoute, votre soutien, votre confiance et vos encouragements, merci.

Table of contents

Abbreviations

AA	Amino acid
ABC	Ammonium bicarbonate
ACN	Acetonitrile
AD	Alzheimer's disease
AGC	Automatic gain control
ALS	Amyotrophic lateral sclerosis
ATP	Adenosine triphosphatase
C18	Carbon 18
CCS	Collision cross section
CHAPS	3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate
CID	Collision induced dissociation
CSF	Cerebrospinal fluid
CTRL	Control
CV	Coefficient of variation
Da	Dalton
DDA	Data dependent acquisition
DEP	Differentially expressed protein
DEpS	Differentially expressed phosphosite
DIA	Data independent acquisition
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol
EBI	European informatics institute
EDTA	Ethylenediaminetetraacetic acid
ERLIC	Electrostatic interaction liquid chromatography
ESI	Electro-spray ionization
ETD	Electron-transfer dissociation
EthCD	Electron-transfer/higher energy collision dissociation
EtOH	Ethanol
eV	Electron volt
FA	Formic acid
FAIMS	Field asymmetric ion mobility spectrometry
FAIR	Findability, accessibility, interoperability and reusability
FASP	Filter-aided sample preparation
FDR	False discovery rate
FFPE	Formalin-fixed paraffin-embedded
FLR	False localization rate
GPF	Gas phase fractionation
HCD	Higher energy collision dissociation
HILIC	Hydrophilic interaction chromatography
IAM	Iodoacetamide
IDA	Iminodiacetic acid
IGD	In-gel digestion
IM	Ion mobility

IMAC	Immobilized metal affinity chromatography
IP	Immuno-precipitation
iRT	Indexed retention time
ISD	In-solution digestion
iST	In-stage tip
IT	Ion-trap
LC	Liquid chromatography
LFQ	Label-free quantitation
LC-MS/MS	Liquid chromatography coupled to tandem mass spectrometry
M	Molar
MAXOMOD	Multi-omic analysis of axono-synaptic degeneration in motoneuron disease
MC	Missed-cleavage
meOH	Methanol
MOAC	Metal oxide affinity chromatography
MOFA	Multi-omics factor analysis
MRM	Multiple reaction monitoring
MS	Mass spectrometry
MV	Missing value
NCBI	National center of biological information
NTA	Nitrilotriacetic acid
OMS	Open modification search
OT	Orbitrap
PASEF	Parallel accumulation serial fragmentation
PD	Proteome discoverer
PFC	Pre-frontal cortex
PIR	Protein information resource
PSM	Peptide spectrum match
PTM	Post translational modification
Q	Quadrupole
QC	Quality control
RNA	Ribonucleic acid
RP HPLC	Reversed-phase high-pressure liquid chromatography
RT	Retention time
S.pombe	Schizosaccharomyces pombe
SCX	Strong cation exchange
SDC	Sodium deoxycholate
SDS	Sodium dodecyl sulphate
SDS-PAGE	Sodium dodecyl sulphate – polyacrylamide gel electrophoresis
Ser	Serine
SIB	Swiss institute of bioinformatics
SILAC	Stable isotope labeling by amino acids in cell culture
SNF	Similarity network fusion
SP	Spectronaut
Sp3	Single-pot solid-phase-enhanced sample preparation
SPE	Solid phase extraction
Strap	Suspension trapping filter

SWATH	Sequential windowed acquisition of all theoritical fragment ion
TCEP	Tris(2-carboxyéthyl)phosphine
TD	Target decoy
TFA	Tirifluoroacetic acid
TG	Transgenic
Thr	Threonine
TIC	Total ion chromatogram
TIMS	Trapped ion mobility spectrometry
TMT	Tandem mass tag
TOF	Time of flight
Tyr	Tyrosine
UHPLC	Ultra high performance liquid chromatography
UVPD	Ultraviolet photodissociation
WT	Wild-type
XIC	Extracted ion chromatogram

List of communications

PUBLICATIONS

Demeulemeester N., Gebelin M., Caldi-Gomes L., Lingor P., Carapito C., Martens L., Clement L., **msqrob2PTM: differential abundance and differential usage analysis of MS-based proteomics data at the post-translational modification and peptidoform level**, *Molecular & Cellular Proteomics*, in review.

Caldi Gomes L.*, Hänzelmann S.*, Oller S., Parvaz M., Hausmann F., Khatri R., Ebbing M., Holzapfel C., Pasetto L., Columbro S.F., Scozzari S., Gebelin M., Knöferle J., Cordts I., Demleitner A.F., Tzeplaeff L., Deschauer M., Dufke C., Sturm M., Zhou Q., Zelina P., Sudria-Lopez E., Haack T.B., Streb S., Kuzma-Kozakiewicz M., Edbauer D., Pasterkamp R.J., Laczko E.,Rehrauer H., Schlapbach R., Carapito C., Bonetto V., Bonn S.[§], Lingor P[§]., **Multiomic ALS signatures highlight sex differences, molecular subclusters and the MAPK pathway as therapeutic target**, *Nat med*, in publication. *,[§]Equal contribution

ORAL COMMUNICATIONS

Selected oral presentation for Analytics congress, co-organised by the French Proteomic Society (FPS), the French Society of Mass Spectrometry (SFSM), the French Association of Separative Sciences (AFSEP) and the French Network of Metabolomics and Fluxomics (RFMF), 5th to 8th of September 2022, Nantes (France),

Gebelin M., Schaeffer-Reiss C., Caldi-Gomes L., Sonja Hänzelmann, Sergio Oller, Mojan Parvaz, Fabian Hausmann, Robin Khatri, Melanie Ebbing, Constantin Holzapfel, Laura Pasetto, Stefano Columbro, Serena Scozzari, Johanna Knöferle, Isabell Cordts, Antonia F. Demleitner, Laura Tzeplaeff, Marcus Deschauer, Claudia Dufke, Marc Sturm, Qihui Zhou, Pavol Zelina, Emma Sudria-Lopez, Tobias B. Haack, Sebastian Streb, Magdalena Kuzma-Kozakiewicz, Dieter Edbauer, R. Jeroen Pasterkamp, Endre Laczko, Hubert Rehrauer, Ralph Schlapbach, Valentina Bonetto, Stefan Bonn, Lingor P., Carapito C., "**Multi-omic approach for discovery of amyotrophic lateral sclerosis' biomarker: the input of mass spectrometry proteomics & phosphoproteomics**"

Selected oral presentation for PhD student day of ED222, 30th of November 2021, online,

Gebelin M., Brun C., Carapito C., Schaeffer-Reiss C., "**Elucidation of a phosphoproteomic workflow: from sample preparation to data treatment**".

Co-speaker for a French Proteomics Society (FPS) webinar, 30th of September, online,

Gebelin M., Brun C., Bertile F., Carapito C., Schaeffer-Reiss C., "**Elucidation of a phosphoproteomic workflow: from sample preparation to data treatment**".

POSTERS

Selected poster for the American Society of Mass Spectrometry (ASMS) congress, 4th to 8th of June 2023, Houston (USA),

Gebelin M., Schaeffer-Reiss C., Babu Rijal J., Carapito C., "**Optimization of a dia-PASEF method for phosphoproteomics analysis of mouse brain tissues**".

Selected poster for European Proteomics Association (EuPa) congress, 3rd to 7th of April 2021, Leipzig (Germany),

Gebelin M., Brun C., Carapito C., Schaeffer-Reiss C., "**Optimization of a high-throughput phosphoproteomics workflow on complex proteomes**"

PART I

Résumé en français

Chapitre 1 : Etat de l'art de la protéomique quantitative

La spectrométrie de masse (MS) est devenue un outil de choix pour l'analyse des protéines grâce à de nombreuses avancées technologiques et génère des données qualitatives, quantitatives et/ou structurales sur des protéines issues de divers échantillons (cellules, tissus, fluides…), permettant ainsi d'appréhender certains processus biologiques. L'analyse protéomique comporte trois étapes clefs, représentées en **Figure 1** : la préparation de l'échantillon, l'analyse par chromatographie liquide en phase inverse (RP-LC) couplée à la spectrométrie de masse et le traitement de données.

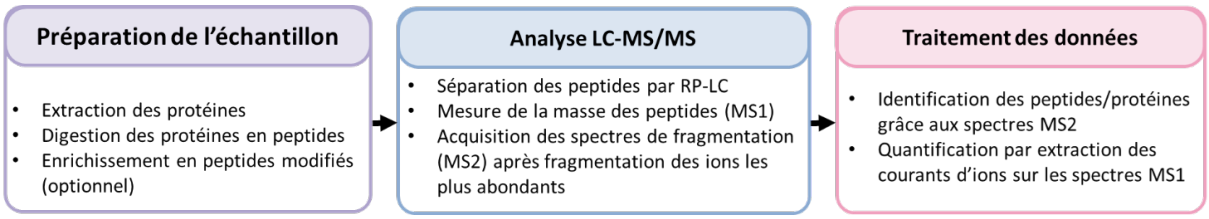


Figure 1: Principales étapes de l'analyse protéomique sans marquage ou *label-free.*

Chacune de ces étapes a bénéficié et bénéficie toujours de progrès technologiques majeurs depuis 30 ans. Parmi eux on peut notamment citer : (i) les avancées dans l'automatisation des étapes de préparation d'échantillon[1] (ii) le développement de nouvelles méthodes d'acquisition pour l'analyse MS/MS (Data Independent Acquisition (DIA)[2] par exemple) (iii) ou encore l'émergence de logiciels de traitement de données basés sur le *deep learning* et l'intelligence artificielle[3].

La spectrométrie de masse est également une méthode communément utilisée pour l'étude des modifications post-traductionnelles (PTMs) des protéines, qui contrôlent notamment l'activité des protéines, leur demi-vie, leur conformation[4],… L'une des PTMs la plus répandue et la plus étudiée est notamment la phosphorylation. En effet, la dérégulation de la réaction de phosphorylation est impliquée dans le processus de nombreuses maladies[5,6]. Cependant, malgré les avancées dans ce domaine, l'étude des phosphorylations à large échelle reste toujours un défi analytique. Que ce soit au niveau de la préparation des échantillons (abondance faible des peptides phosphorylés, conservation de la modification très labile, ...), en passant par l'analyse LC-MS/MS (difficultés d'ionisation et/ou de fragmentation des phosphopeptides) et jusqu'au traitement des données de phosphoprotéomique (outils classiques non adaptés, difficultés à localiser le site modifié, …), l'analyse des phosphopeptides rencontre de nombreuses difficultés[7].

Des difficultés supplémentaires s'ajoutent à la mise en place d'un protocole lorsque l'analyse (phospho)protéomique doit être réalisée dans le cadre d'une étude multi-omique dans un contexte d'application clinique. La multi-omique est définie comme une analyse regroupant des données issues de différentes sciences -omiques, telles que la génomique, la transcriptomique ou la protéomique entre autres. Ce type d'étude s'est révélée indispensable pour l'étude de pathologies complexes, pour lesquelles l'utilisation des différentes sciences -omiques séparément s'est avérée insuffisante pour expliquer les différents processus en cause dans ces maladies[8]. Cependant, dans un projet multi-omique, la quantité de matériel de départ à disposition est souvent réduite, puisque les échantillons biologiques utilisés sont la plupart du temps issus de patients et l'échantillon précieux récupéré doit être utilisé pour les différentes analyses omiques. De plus, ces études regroupent généralement de grandes cohortes d'échantillons, afin d'être le plus représentatif possible de la population étudiée. Les protocoles mis en place doivent donc pouvoir être appliqués à de petites quantités d'échantillons de

départ, tout en permettant une analyse haut débit aussi reproductible que possible sur des centaines d'échantillons.

L'objectif de ma thèse contenait ainsi deux axes principaux :

- Le développement d'un protocole de phosphoprotéomique à haut débit, de la préparation d'échantillon, en passant par l'analyse LC-MS/MS et le traitement des données générées. Ainsi, j'ai développé un protocole automatisé de préparation d'échantillons adapté à des petites quantités de tissus cérébraux, comprenant notamment un enrichissement des phosphopeptides par Immobilized Metal Affinity Chromatography (IMAC). J'ai également évalué les paramètres chromatographiques, différentes énergies de collision mais aussi différentes méthodes d'acquisition (Data Dependent Acquisition (DDA) et Data Independent Acquisition (DIA)) sur une plateforme LC-MS/MS de dernière génération, afin de développer une méthode optimisée pour l'analyse de ces phosphopeptides. Enfin, différents logiciels de traitement des données ont été évalués pour l'identification, la quantification et la localisation des sites de phosphorylation, cette information de localisation étant cruciale pour l'interprétation biologique des résultats.
- Ce protocole optimisé a ensuite été appliqué dans le cadre d'un projet multi-omique (Multi-omic analysis of axono-synaptic degeneration in motoneuron disease, MAXOMOD) pour lequel j'ai réalisé l'analyse protéomique et phosphoprotéomique de centaines d'échantillons d'origines biologiques différentes (tissus cérébraux murins et humains, liquides céphalo-rachidiens humains). Pour l'analyse des échantillons de liquide céphalo-rachidien (CSF), une méthode de préparation d'échantillon a également été développée, compatible à la fois avec l'analyse protéomique et phosphoprotéomique. Des contrôles qualités, transférables à d'autres applications cliniques, ont été mis en place pour s'assurer de la qualité et de la reproductibilité de l'ensemble des analyses (phospho)protéomiques réalisées sur cette large cohorte d'échantillons.

Chapitre 2 : Développement d'un protocole de phosphoprotéomique automatisé et à haut débit

1. Développement d'un protocole de préparation d'échantillons automatisé et haut débit pour l'analyse phosphoprotéomique

De nombreuses optimisations ont été réalisées pour la mise en place d'un protocole complet de phosphoprotéomique, automatisé et applicable à des analyses à haut débit**.** L'extraction des protéines a tout d'abord été optimisée en comparant 5 protocoles différents (**Figure 2 – (A)**).

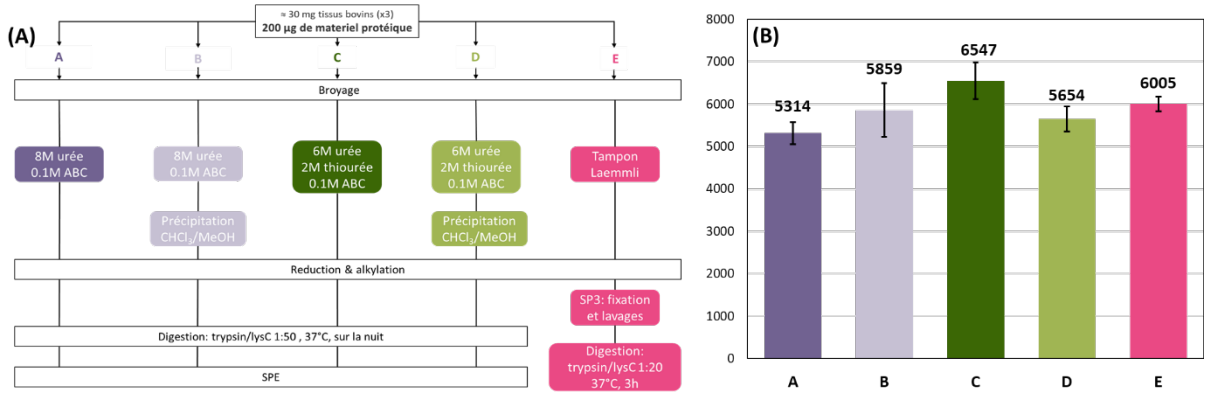


Figure 2: (A) Différents protocoles évalués pour la préparation d'échantillons et l'extraction des protéines (B) Nombre moyen de peptides identifiés et validés pour chaque protocole d'extraction.

En **Figure 2 – (B)**, nous concluons que les deux protocoles d'extraction donnant les meilleurs résultats sont les protocoles C (6M urée, 2M thio-urée sans précipitation) et E (Laemmli). En effet, après digestion, ils permettent d'identifier entre 6000 et 6500 peptides contre moins de 6000 peptides pour les autres protocoles. Ces deux protocoles ont ainsi été sélectionnés pour la suite des optimisations.

Afin de palier à la faible stœchiométrie des phosphopeptides comparés aux peptides non phosphorylés dans les échantillons[9], une étape d'enrichissement des phosphopeptides par chromatographie d'affinité est nécessaire. Ici, nous avons utilisé une étape d'enrichissement automatisée sur cartouches IMAC à l'aide d'un robot Bravo AssayMAP (Agilent Technologies). Les extraits peptidiques des deux conditions retenues ont ainsi été soumis à cet enrichissement, et ont conduit aux résultats présentés en **Figure 3**.

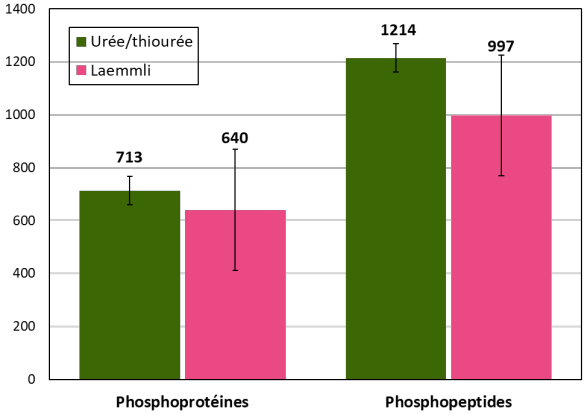


Figure 3: Nombre moyen de phosphoprotéines et phosphopeptides identifiés à l'aide des deux protocoles après enrichissement IMAC.

Le protocole urée/thiourée présente les meilleurs résultats en terme d'identification de phosphoprotéines et phosphopeptides. Cependant, le protocole avec un tampon « Laemmli » ayant été retenu pour le projet multi-omique décrit en **Chapitre 2 - 1.Analyse protéomique haut débit de large cohortes de tissus cérébraux murins et humains**, seuls les résultats obtenus avec ces conditions seront présentés dans la suite de ce travail.

2. Optimisation d'une méthode LC-MS/MS pour l'analyse des phosphopeptides

Les échantillons enrichis sur cartouches IMAC ont ensuite été utilisés afin d'optimiser une méthode LC-MS/MS pour l'analyse des phosphopeptides. En effet, l'analyse par spectrométrie de masse des phosphopeptides est loin d'être triviale, de part notamment la labilité de la modification. A partir d'une méthode sur une Q-Exactive HF-X, non optimisée pour la phosphoprotéomique, nous avons ainsi profité des performances d'un instrument de dernière génération, le TimsTOF Pro, pour développer une méthode adaptée à l'analyse des phosphopeptides. Sur ce couplage LC-MS/MS, deux modes d'acquisition différents ont été testés : le mode DDA et le mode DIA (**Figure 4**).

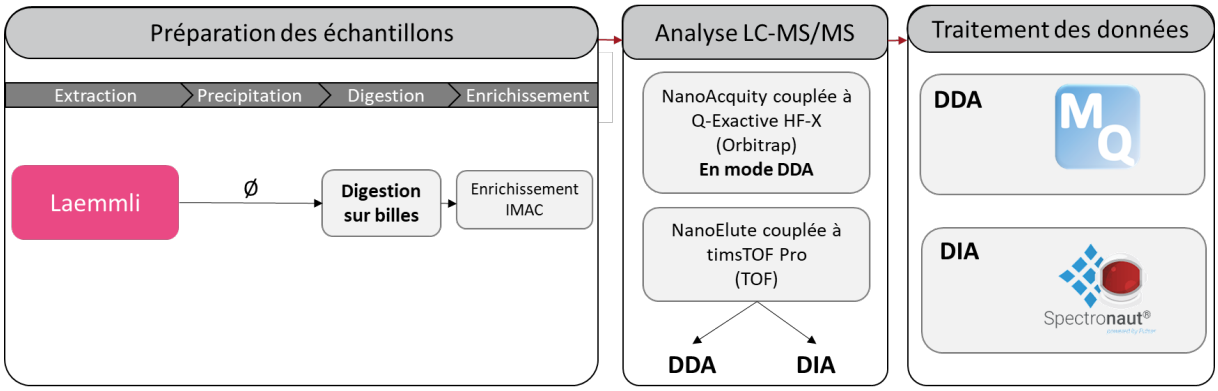


Figure 4: Schéma analytique pour l'optimisation d'une méthode LC-MS/MS pour l'analyse des phosphopeptides.

Dans un premier temps, les paramètres communs aux deux modes d'acquisition ont été évalués en mode DDA : optimisation du temps d'accumulation, de la rampe d'énergie de collision et de la fenêtre de mobilité ionique. Ces optimisations ont permis d'augmenter le nombre de phosphosites de façon significative (plus de 35%) comparé à la méthode non optimisée (**Figure 5**, méthodes 3 et 4).

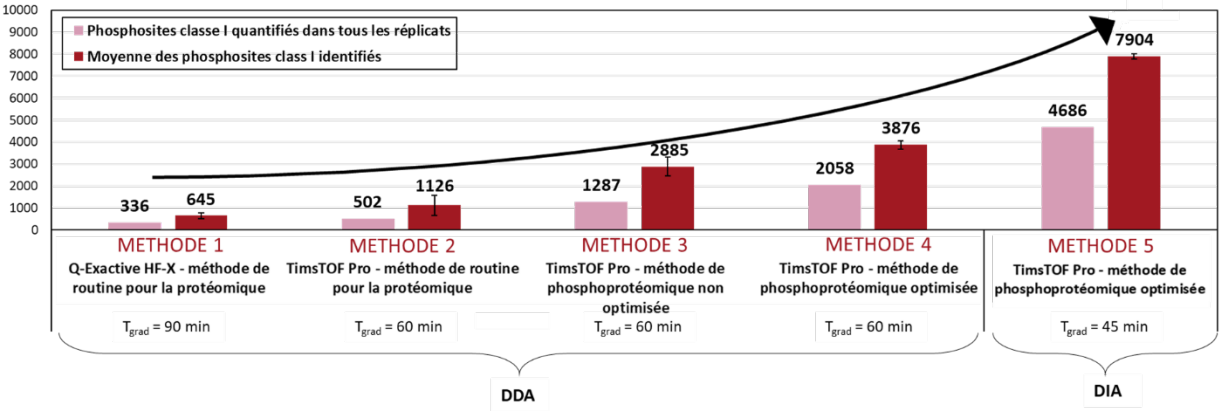


Figure 5: Résultats d'identification et de quantification des phosphosites classe I à travers les différentes optimisations LC-MS/MS.

Une des limitations majeures du mode d'acquisition DDA est sa nature semi-stochastique. En effet, seuls les N ions précurseurs les plus intenses en MS1 vont être sélectionnés pour la fragmentation, limitant la gamme dynamique et la reproductibilité de la méthode. La DIA offre une alternative prometteuse[10], puisqu'elle permet de fragmenter tous les ions compris dans une fenêtre d'isolation de masse définie. Cette méthode d'acquisition permet ainsi d'atteindre une meilleure justesse et sensibilité d'analyse. De plus, les premières études récentes décrites dans la littérature qui combinent une approche DIA avec la technologie de *Parallel Accumulation Serial Fragmentation* (PASEF) ont montré une augmentation considérable de la couverture du protéome et phosphoprotéome[10,11], révélant ainsi les capacités plus que prometteuses de ces nouvelles générations. C'est pour cela que j'ai mis en place une méthode DIA pour l'analyse des phosphopeptides sur le TimsTOF Pro. Plusieurs paramètres ont ainsi été évalués (largeur de fenêtre d'isolation, temps d'accumulation, temps de cycle, plage de mobilité ionique) afin d'augmenter le nombre de sites de phosphorylation identifiés et quantifiés. Ces optimisations ont conduit à une méthode DIA adaptée pour la phosphoprotéomique, permettant ainsi d'atteindre de presque 8000 phosphosites class I identifiés (**Figure 5**, méthode 5).

3. Evaluation de différentes méthodes de traitement de données pour l'identification, la quantification et la localisation de phosphorylation

L'analyse des données en phosphoprotéomique est une étape difficile car, outre l'identification des peptides, les sites de phosphorylation doivent être localisés. L'identification et la quantification des sites de phosphorylation peuvent être effectuées par différents algorithmes. Ces derniers génèrent également différents scores pour évaluer la fiabilité de la localisation de la phosphorylation. Cependant, les données de la littérature ne permettent pas toujours de comparer directement les résultats obtenus par les différents algorithmes. C'est pourquoi durant ma thèse, j'ai comparé plusieurs logiciels de traitement des données de phosphoprotéomique, à la fois pour l'analyse DDA et DIA (**Figure 6**).

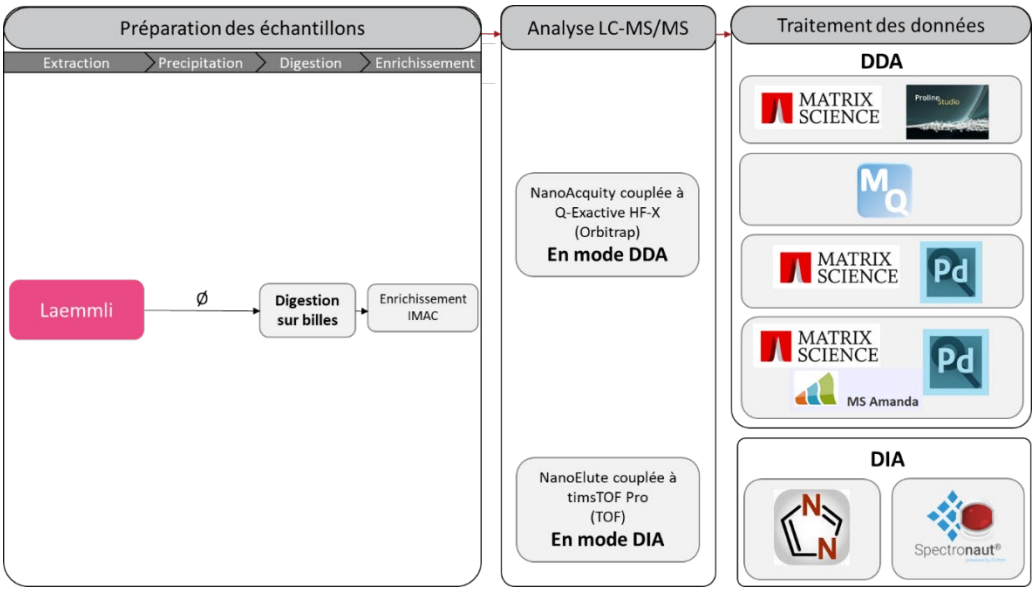


Figure 6: Schéma analytique pour l'optimisation du traitement de données pour l'analyse DDA et DIA des phosphopeptides.

Dans un premier temps, les résultats de comparaison des méthodes de traitements de données phosphoprotéomique issues de la DDA sont illustrés en **Figure 7 – (A)**. Des nombres plus élevés

d'identification sont obtenus en utilisant la combinaison des deux moteurs de recherche, Mascot et de MS Amanda, avec plus de 1300 phosphosites, soit 8% de plus qu'en utilisant Mascot seul. Cependant, en regardant plus en détail dans les données et notamment les spectres de certains phosphopeptides identifiés uniquement par MS Amanda, il semble que ces spectres soient moins informatifs que ceux obtenus par Mascot. Pour l'évaluation des logiciels de DIA, DIA-NN permet une nette augmentation du nombre de phosphopeptides identifiés avec en moyenne 20% supplementaires sur l'ensemble des méthodes (**Figure 7 – (B)**).

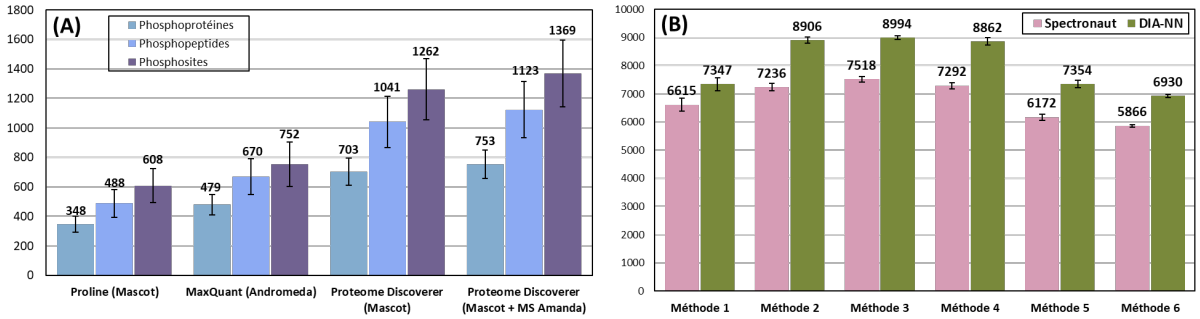


Figure 7: Comparaisons des logiciels de traitement des données phosphoprotéomique (A) DDA : Nombres moyens de phosphoprotéines, -peptides, et -sites identifiés (B) DIA : Nombres moyens de phosphopeptides identifiés.

La définition d'un site de phosphorylation est dépendante du logiciel. En général, un site de phosphorylation fait référence à la localisation d'un acide aminé dans la séquence peptidique portant une phosphorylation. La notion de site donne plus d'informations que le phosphopeptide car un peptide peut porter plusieurs phosphorylations et plusieurs peptides peuvent porter le même site de phosphorylation. Ainsi, si la quantification est effectuée au niveau du peptide plutôt qu'au niveau du site, un biais est introduit. En mode DDA, si Proline et Proteome Discoverer permettent tous deux d'identifier les sites de phosphorylation, ils ne donnent accès à aucune information de quantification au niveau du site. MaxQuant est, à ce jour, le seul logiciel DDA à permettre la quantification des phosphosites en additionnant les intensités de tous les phosphopeptides impliqués dans un site, contenant ainsi les peptides avec des clivages manqués et ceux avec des modifications additionnelles. Similairement en DIA, Spectronaut est le seul logiciel à permettre la quantification des phosphosites. Pour cette raison, MaxQuant pour les données DDA et Spectronaut pour la DIA ont été les deux logiciels utilisés pour le traitement des données de phosphoprotéomique durant le reste de ma thèse.

Chapitre 3 : Application des développements méthodologiques à une étude multi-omique pour la recherche de biomarqueurs de la sclérose latérale amyotrophique - projet MAXOMOD

La sclérose latérale amyotrophique[12] (SLA) est une maladie neurodégénérative qui provoque des faiblesses musculaires progressives suivies du décès de la personne malade dans les 3 à 5 ans suivants les premiers symptômes. Son diagnostic est basé sur des critères cliniques et survient relativement tardivement. En raison de son mauvais diagnostic et des options thérapeutiques limitées, une meilleure caractérisation des événements déclencheurs du développement de la SLA est nécessaire. Le but du projet européen MAXOMOD, qui a financé une partie de mes travaux de thèse, était d'adopter une approche multi-omique pour identifier de nouvelles voies et biomarqueurs liés à la SLA.

1. Analyse protéomique haut débit de large cohortes de tissus cérébraux murins et humains

L'agrégation de différentes protéines, telles que SOD1, TDP-43, C9ORF72 ou FUS dans le cerveau, et notamment dans le cortex frontal, est bien connue comme l'une des principales caractéristiques de la SLA[13,14]. Par conséquent, l'étude (phospho)protéomique des tissus post-mortem du cortex préfrontal humain et de souris transgéniques peut permettre d'identifier des protéines spécifiques à la maladie. Ces protéines, participant à des processus pathologiques clés, peuvent être utilisées comme biomarqueurs potentiels de la SLA. Dans ce contexte, je me suis concentrée sur les développements à mener pour obtenir des protocoles complets pour l'analyse par LC-MS/MS à haut débit du protéome et du phosphoprotéome de grandes cohortes de tissus cérébraux. Pour chaque modèle étudié, la moitié des échantillons provenait d'échantillons contrôle et l'autre moitié d'échantillons ALS ou de souris transgéniques.

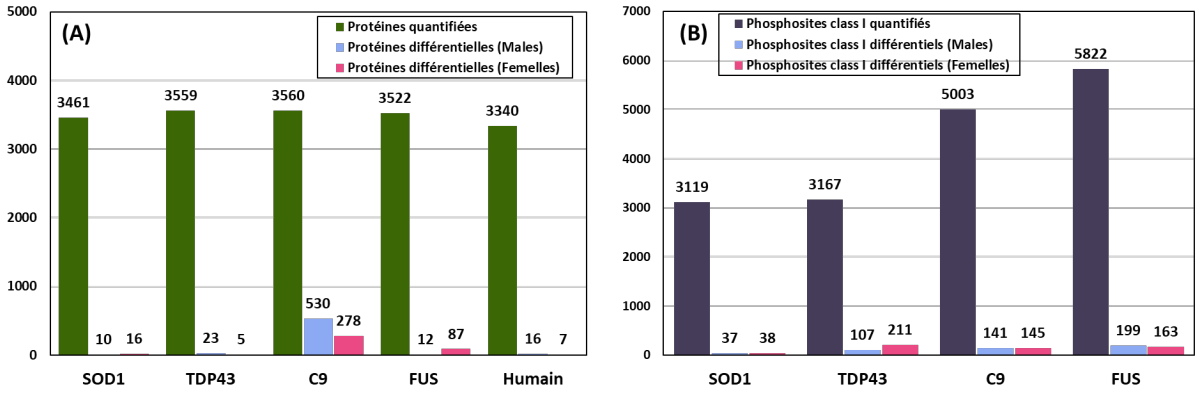


Figure 8: Pour les différents modèles étudiés, résultats globaux de quantification et issus de l'analyse différentielle entre les conditions transgénique et contrôle, chez les mâles et femelles (A) Protéomique globale (B) Phosphoprotéomique.

L'analyse des protéomes totaux a ainsi permis la quantification de plus de 3000 protéines (**Figure 8 – (A)**). Parmi elles, seul un faible pourcentage de protéines a été quantifié différentiellement entre les deux conditions, que ce soit chez les mâles ou les femelles. Pour l'analyse phosphoprotéomique des tissus (murins uniquement), le protocole développé précédemment (détaillé en **Chapitre 1 - 1.**

Développement d'un protocole de préparation d'échantillons automatisé et haut débit pour l'analyse phosphoprotéomique) a été appliqué sur les tissus. Ce protocole optimisé a permis la quantification de 3000 à 5000 phosphosites class I (**Figure 8 – (B)**). Comme pour la protéomique, l'analyse différentielle n'a révélé que peu de phosphosites différentiels. Ces résultats mettent en avant le peu de différence entre le protéome d'une personne saine et d'une personne atteinte de la SLA, ainsi que la complexité de la maladie. Malgré les faibles changements observés, ces résultats, mis en commun avec les résultats des autres sciences –omiques, ont permis l'identification de potentiels biomarqueurs de la SLA.

2. Développement d'un protocole commun pour l'analyse protéomique et phosphoprotéomique de liquide céphalo-rachidien

Pour l'analyse protéomique et phosphoprotéomique de liquide céphalo-rachidien (LCR), différents protocoles de préparation d'échantillons ont été évalués. L'enjeu était de mettre en place un protocole commun à l'analyse protéomique, phosphoprotéomique et métabolomique, à partir du même échantillon.

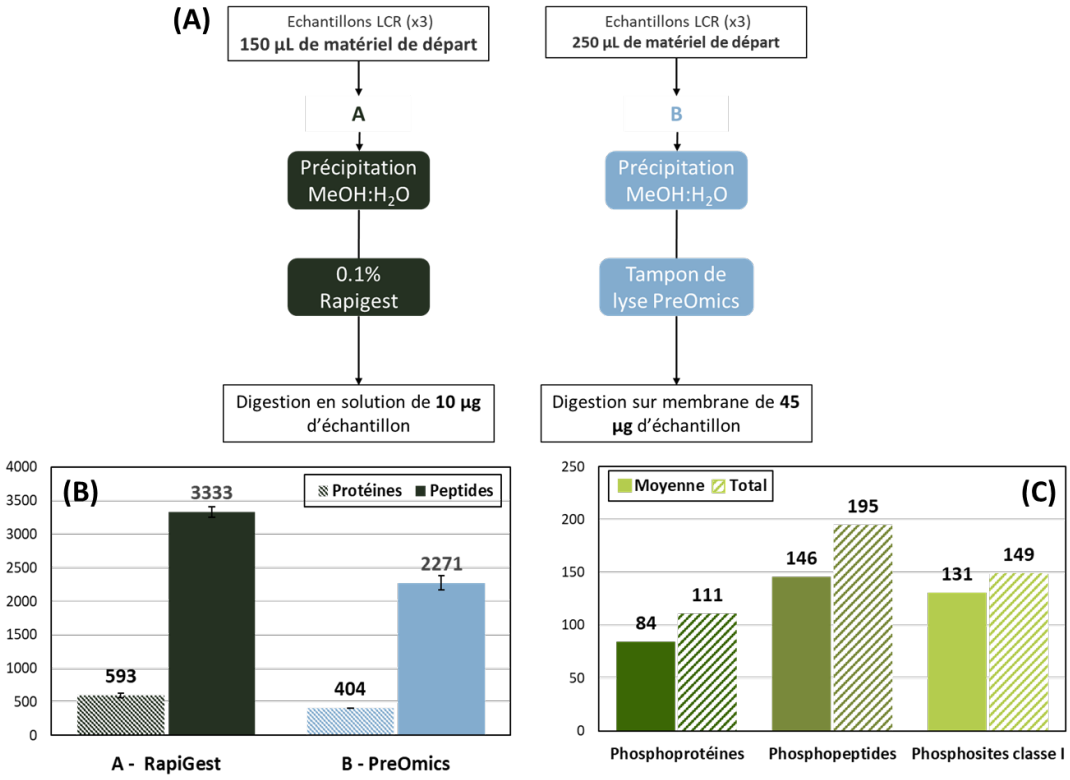


Figure 9: (A) Schéma analytique pour la préparation des échantillons LCR (B) Résultats de la comparaison des deux protocoles en termes de protéines et peptides identifiés (C) Résultats de l'analyse phosphoprotéomique de LCR à partir du protocole de préparation d'échantillon au RapiGest.

Après une étape de précipitation au méthanol pour extraire les métabolites des échantillons et concentrer les échantillons, deux protocoles de digestion ont été comparés (**Figure 9 – (A)**). Le premier correspond à l'utilisation du kit commercial de préparation d'échantillon iST PreOmics consistant en la solubilisation des protéines dans leur tampon commercial, suivie du protocole iST de digestion sur membrane. Le second correspond à la reprise des protéines dans 0.1% de surfactant RapiGest, suivi

d'un protocole de digestion liquide. Les résultats de protéomique globale (**Figure 9 – (B)**) montrent que le protocole « RapiGest » permet d'obtenir un plus grand nombre d'identifications (plus de 3300 peptides contre moins de 2300 peptides). Les performances du protocole RapiGest ont ensuite été évaluées pour l'analyse phosphoprotéomique. Pour cela, une étape d'enrichissement des phosphopeptides a été ajoutée. Grâce à ce protocole optimisé, des résultats prometteurs pour l'analyse des phosphorylations du LCR ont été obtenus, représentés en **Figure 9 – (C)**.

Le protocole de préparation des échantillons de LCR avec du RapiGest a ensuite été appliqué pour l'analyse protéomique et phosphoprotéomique des plus de 100 échantillons de LCR cliniques de patients contrôles et de patients atteints de la SLA. L'analyse a permis la quantification de 669 protéines, dont 59 exprimées différentiellement chez les mâles et seulement 12 chez les femelles. Pour la phosphoprotéomique, plus de 360 phosphosites classe I ont été quantifiés. Parmi eux, 23 phosphosites classe I différentiels chez les mâles et 28 chez les femelles. Ces résultats, mis en commun par la suite avec les autres équipes des différentes omiques du projet, permettent l'identification de potentielles cibles biomarqueurs de la SLA.

3. Contrôles qualité pour l'analyse protéomique et phosphoprotéomique de larges cohortes d'échantillons

Afin de s'assurer de la robustesse et répétabilité de notre préparation d'échantillons et de nos analyses sur de si grands nombres d'échantillons, différents contrôles qualités (QC) ont été mis en place.

Pour les analyses protéomiques, des peptides synthétiques standards (iRT, Biognosys) ont été ajoutés à tous les échantillons avant leur injection en LC-MS/MS. Ces derniers nous ont permis de vérifier l'alignement des temps de rétention au cours de nos longues séquences d'injections. Ainsi, pour les 9 peptides synthétiques détectés sur l'ensemble des cohortes, le coefficient de variation (CV) moyen sur les temps de rétention est inférieur à 4%, soulignant la stabilité du système chromatographique tout au long des plus de 1000 injections (**Figure 10 – (A)**). En plus de ce QC interne, un mélange de tous les échantillons a été assemblé pour chaque cohorte avant les étapes de réduction et alkylation des protéines. Ce pool a ensuite suivi exactement les mêmes étapes de préparation que les autres échantillons et a été injecté à intervalles très réguliers durant nos analyses. Cela nous a permis de nous assurer de la stabilité du signal de MS. En effet, comme illustré en **Figure 10 – (B)**, l'abondances des protéines identifiées par les pools restent stable au cours des pools injectés *ie* au cours du temps.

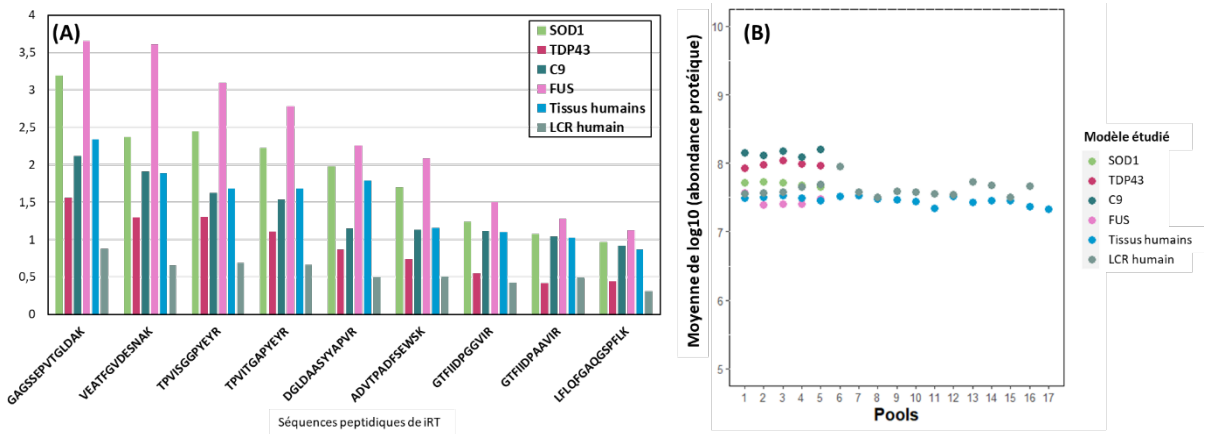


Figure 10 : (A) Coefficients de variation (en %) des temps de rétention des différentes peptides iRT synthétiques dans les échantillons des différentes cohortes de protéomique (B) Nuages de points représentant la variation de l'abondance protéique moyenne des différents pools d'échantillons.

Pour les analyses de phosphoprotéomique, un mélange de phosphopeptides synthétiques standards (Phosphomix, Sigma Aldrich) a été utilisé. Ces derniers ont été ajoutés sous leur forme non marquée « *light* » avant l'étape d'enrichissement sur cartouches IMAC, et sous leur forme isotopiquement marquée « *heavy* » après l'enrichissement. En calculant ensuite les ratios de l'intensité du peptide *light* sur l'intensité du peptide *heavy*, cela nous permet d'estimer un ratio d'enrichissement pour chaque phosphopeptide synthétique de chaque cohorte. Ainsi, en moyenne, entre 20% et 50% du matériel peptidique a été enrichi (**Figure 11 – (A)**). Afin d'évaluer la stabilité du système chromatographique, nous avons représenté les CV sur les temps de rétention des différents phosphopeptides synthétiques dans chaque cohorte (**Figure 11 – (B)**). Pour tous les modèles de souris, presque tous les CV sont inférieurs à 1 %, ce qui souligne la grande stabilité du système LC. De plus, même pour la grande cohorte de LCR, les CV sont tous inférieurs à 2 %. Cela permet de conclure que le système chromatographique est très stable sur des centaines d'injections.

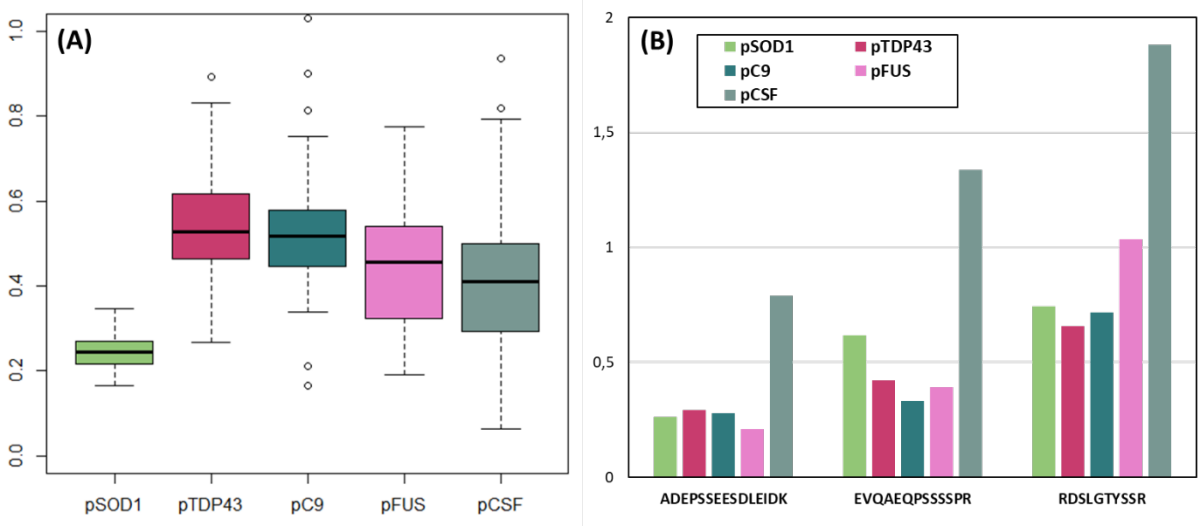


Figure 11: (A) Diagrammes en boîte représentant le ratio de phosphopeptides enrichis pour chaque cohorte (B) Coefficients de variation (en %) des temps de rétention des différents phosphopeptides synthétiques Phosphomix dans les échantillons des différentes cohortes.

4. Evaluation des performances d'un logiciel d'*Open Modification Search* pour l'identification de phosphorylations

Lors d'une analyse protéomique, un grand nombre de spectres ne sont généralement pas attribués à des peptides. L'une des raisons pour cela est l'espace de recherche restreint des moteurs de recherche actuels, qui ne peuvent donc pas identifier les peptides présentant des modifications inattendues. En effet, si une modification particulière n'a pas été spécifiée dans les paramètres de recherche, les spectres correspondant aux peptides portant cette modification peuvent se faire attribuer une séquence d'acides aminés incorrecte[15]. Dans cette optique, des outils d'*Open Modifications Search* (OMS) ont été développés pour identifier les spectres modifiés. Les logiciels OMS sont donc particulièrement prometteurs pour l'étude des PTMs.

Pour mes travaux de thèse, j'ai donc évalué les performances d'un outil OMS, IonBot[16], sur nos différents ensembles de données de tissus cérébraux de souris enrichis en phosphopeptides. Les

performances de IonBot (v.0.10.0) ont été comparées à celles de MaxQuant (v.1.6.14), pour l'identification des phosphorylations.

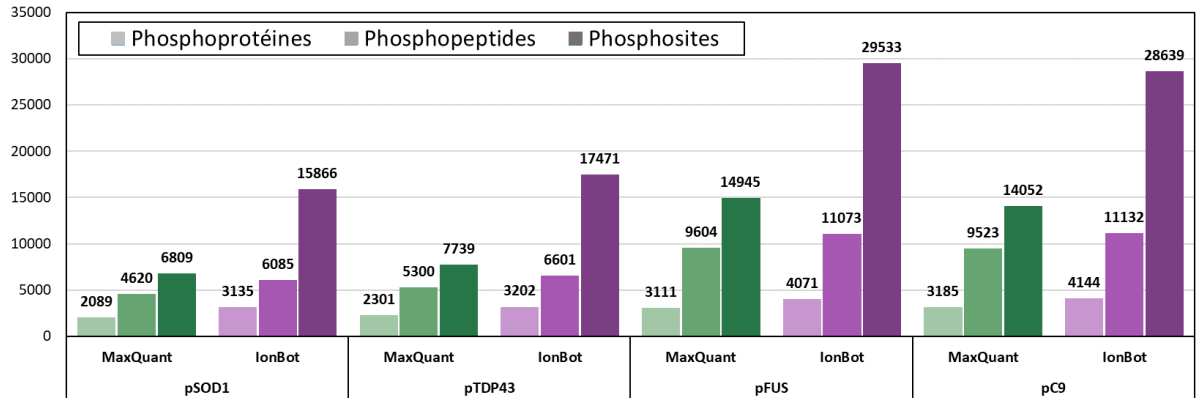


Figure 12: Résultats de l'identification des phosphoprotéines, phosphopeptides et phosphosites pour les différents modèles de souris en utilisant Andromeda (MaxQuant) ou IonBot pour la recherche.

IonBot permet ainsi d'augmenter les identifications de 15% (phosphopeptides) à 50% (phosphoprotéines). L'augmentation est encore plus impressionnante au niveau des sites avec en moyenne 10 000 phosphosites supplémentaires identifiés grâce à IonBot. En combinant les identifications des différentes cohortes, j'ai ensuite étudiés les différentes populations identifiées par les deux logiciels. IonBot permet ainsi l'identification de plus de 6700 phosphopeptides uniques comme représenté en **Figure 13 – (A)**. En se concentrant sur les phosphopeptides identifiés uniquement par IonBot, j'ai pu mettre entre avant que la plupart d'entre eux sont également identifiés par IonBot mais avec des caractéristiques différentes (portant une autre modification, non modifié…). De plus, un filtre limitant la longueur des peptides à maximum 30 amino-acides afin de limiter l'espace de recherche est appliqué automatiquement lors de la recherche IonBot. Ainsi, plus de 800 phosphopeptides sont uniquement identifiés par MaxQuant avec une longueur de peptide supérieure à 30 amino-acides puisque MaxQuant ne fixe pas de filtre à ce niveau.

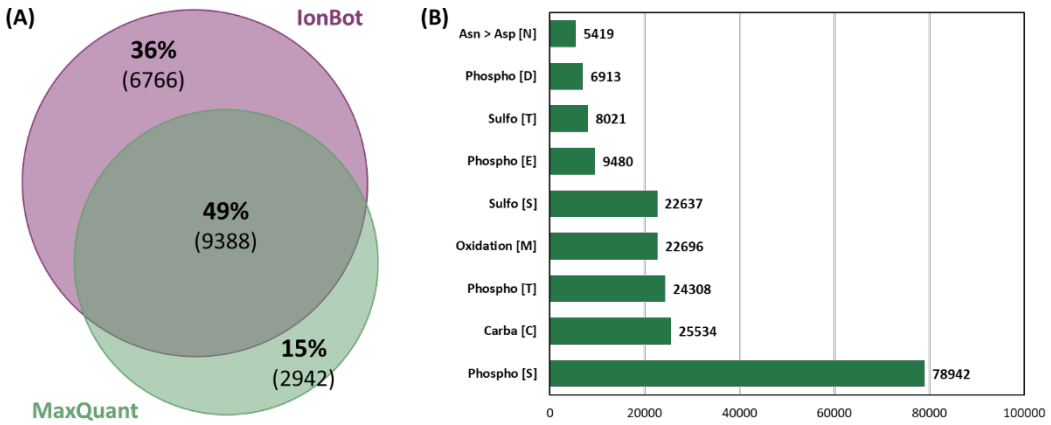


Figure 13: (A) Recouvrement des phosphopeptides identifiés entre IonBot et Andromeda (MaxQuant) (B) Modifications les plus abondantes avec IonBot et le nombre correspondant de peptides modifiés identifiés**.**

Chapitre 4 : Conclusion générale

Grâce à ces travaux de thèse, j'ai acquis une expertise dans le domaine de la protéomique et phosphoprotéomique basée sur la spectrométrie de masse.

La première partie de ma thèse s'est concentrée sur les nombreux défis de l'analyse phosphoprotéomique. J'ai ainsi développé des protocoles complets et automatisés pour l'analyse des phosphopeptides, adaptés à la nature des échantillons du projet. J'ai commencé par l'optimisation de la préparation des échantillons et notamment celle de l'extraction des protéines et de l'étape d'enrichissement des phosphopeptides automatisée. Je me suis ensuite concentrée sur le développement de méthodes LC-MS/MS pour l'analyse des phosphopeptides sur un TimsTOF Pro. J'ai ainsi évalué les performances et l'influence de différents paramètres, à la fois en mode DDA et DIA, sur la qualité des résultats d'analyse. Ce travail a donné lieu à une méthode optimisée pour la phosphoprotéomique sur le TimsTOF Pro. Enfin, j'ai approfondi ces travaux jusqu'au choix du logiciel le plus adapté pour l'identification, la quantification et la localisation des sites de phosphorylation. J'ai ainsi évalué plusieurs logiciels proposés pour l'analyse – délicate et difficile – des données de phosphoprotéomique, qu'elles soient générées en DDA ou en DIA.

La seconde partie de mes travaux de thèse est axée sur l'application d'une partie de ce travail dans le cadre d'un projet multi-omique sur la SLA. Dans ce contexte, j'ai réalisé des analyses protéomiques et phosphoprotéomiques sur un grand nombre et variété d'échantillons cliniques. Pour cela, j'ai notamment optimisé un protocole de préparation d'échantillon pour le LCR, commun à la fois à l'analyse protéomique, phosphoprotéomique et métabolomique. L'analyse, protéomique et phosphoprotéomique, des tissus et des échantillons LCR, a permis d'identifier de potentielles cibles pour la SLA qui seront par la suite validées en tant que biomarqueurs. La qualité de l'ensemble de ces analyses a pu être évaluée grâce à de nombreux contrôles qualités attestant de la reproductibilité des protocoles développés sur de larges cohortes d'échantillons. Enfin, j'ai évalué les performances d'un logiciel d'OMS pour l'identification de phosphorylations. Les résultats de cette expérience, bien que préliminaires, sont très prometteurs avec une augmentation des identifications de plus de 50%.

General introduction

General introduction

Proteins are large and complex molecules made of hundreds to thousands of amino acids, linked by peptidic bonds. The variety of proteins is extreme and collectively, proteins catalyze and control most tasks within biological systems. Proteomics is defined as the large-scale characterization of the entire protein complement of a cell, tissue or organism, at a specific time and location, and under given physiologic/pathologic conditions[17]. The proteome is thus highly dynamic and extremely complex. Indeed, in response to internal or external stimuli, proteins can be synthetized, modified by post-translational modifications (PTMs), undergo translocations within the cell or be degraded. The transcription into RNA of the approximately 20,000 protein-coding human genes produces approximately 100,000 transcripts, which will be then translated into more than half a million proteins. It is estimated that at least half of these proteins contain modification sites that may carry a PTM, creating a total of more than a million different proteic forms, called proteoforms[18]. To date, more than 300 types of PTMs[19] such as acetylation, glycosylation, phosphorylation, etc..., are described to occur physiologically and the implication of some of them has already been clearly recognized during the transformation of normal cells into tumor cells[20].

Strong from these specificities, proteomics represents a great opportunity to investigate biological processes and their modulating mechanisms upon genetic variability, environmental or physiological perturbations. Additionally, thanks to its complementarity with other –omics sciences (genomics, transcriptomics, metabolomics …), proteomics is a major counterpart in the context of multi-omics studies. These studies, while facing a combination of challenges both intrinsic to each individual omics and more generally for multi-level data integration, are still the most promising route to better understand cellular processes involved in many diseases[21,22].

The increasing interest for proteomics has followed the development of new technologies adapted for peptides/proteins separation and analysis. Since the 80's during which proteomics analysis was performed on bi-dimensional electrophoresis gels, mass spectrometry has become a tool of choice for proteomic analysis thanks to numerous technological advances, together with continuous developments in sample preparation and data analysis[17,23,24]. Nowadays, high throughput label-free proteomics allows the identification and quantification of thousands of proteins in a few hours[25]. This label free approach can also be applied to study PTMs. However, they add an additional level of complexity to the analysis and despite recent advances, remains an analytical challenge. Difficulties emerge from sample preparation (labile PTMs, incomplete digestion, site localization, …), from the LC-MS/MS analysis (ionization and fragmentation issues, low stoichiometry…) and from data treatment (localization of the modified site, quantification at the site level, …).

My PhD work is in line with this context. Indeed, it was focused on the development of high-throughput proteomics and phosphoproteomics methods, and their application for amyotrophic lateral sclerosis (ALS) biomarker discovery on large cohorts.

This manuscript is structured in five parts that are here briefly presented:

- Part I of the manuscript is a brief summary in French of my PhD work.

- Part II corresponds to an overview of the state of the art in quantitative proteomics. It includes the description of the three main steps of the label free bottom-up proteomics workflow: sample preparation, LC-MS/MS analysis, and data treatment. It includes also the specificities and difficulties of protein phosphorylation analysis. A specific focus is also put on Data Independent Acquisition (DIA) methods, its different strategies, challenges, and its application to phosphoproteomics. Finally, a short review on multi-omics approaches to disease sums up the different challenges faced with these kind of studies.

- Part III focuses on the analytical developments conducted to set up a fully automated high throughput phosphoproteomics workflow:
    - Chapter 1 describes the optimization of phosphoproteomics sample preparation on bovine brain tissues. It includes protein extraction and digestion optimizations, as well as phosphopeptide enrichment efficiency and reproducibility evaluation.
    - Chapter 2 details the developments performed for the optimization of a nanoLC-MS/MS method for phosphopeptides analysis. In this chapter, different nanoLC-MS/MS platforms and fragmentation methods were compared for phosphoproteomic analysis. NanoLC-MS/MS methods were then optimized using different modes of acquisition to study phosphopeptides.
    - In Chapter 3, different pipelines for phosphoproteomics data analysis were compared. Their performances were evaluated for the identification, quantification and localization of phosphorylation events.

- Part IV is dedicated to the developments conducted and results obtained in the context of the Multi-omic analysis of AXOno-synaptic degeneration in MOtoneuron Disease (MAXOMOD) project (European consortium that has partially funded this PhD work).
    - Chapter 1 focuses on the application of proteomics and phosphoproteomics workflows for high throughput analysis of large cohorts of brain tissues. It sums up the protocols used as well as the global identification and quantification results achieved. The differential expression of some proteins/phosphosites is also investigated, as well as the biological relevance of those proteins/phosphosites.
    - In Chapter 2 the development of a protocol for proteomics and phosphoproteomics analysis of cerebrospinal fluid (CSF) samples is described. The evaluation of different sample preparation workflows and their performances for both proteomics and phosphoproteomics are evaluated. The results of the analysis of more than 100 samples of clinical CSF for the MAXOMOD project are also detailed. The biological relevance of potential biomarker target was discussed.
    - Chapter 3 addresses the need of quality controls for proteomics and phosphoproteomics analysis of large cohorts of samples.
    - In Chapter 4, the use of an open modification search (OMS) software is evaluated to further increase phosphoproteome coverage in the future.

- Part V details all experimental procedures and protocols used for the presented PhD work.

PART II

State of the art in quantitative proteomics and phosphoproteomics

Chapter 1: Mass spectrometry analysis of proteins

Proteins are functional molecules, coded from genetic information, that catalyze biochemical reaction in the cells of an organism. The first notion of proteome was introduced in 1994 by Marc Wilkins[26] by analogy with the genome term. Unlike the genome, the proteome is a highly dynamic entity and protein expression varies with time, depending on their localization and in response to diverse stimuli, making its study challenging[23]. Proteomics were introduced a few years later by Peter James[27]. Proteomic is the science that studies the proteome and is defined as the large-scale characterization of the entire protein complement of a cell, tissue or organism, at a specific time and location, and under given physiologic/pathologic conditions[17]. Proteome adaptability to environmental stimuli, while challenging for proteomics analysis, is the reason proteomics studies are so popular in various areas such as clinical applications[28,29], food industries[30,31], plant interaction systems[32,33], paleontology[34,35] or art conservation[36,37].

Mass spectrometry has become a tool of choice for proteomic analysis thanks to numerous technological advances[17,23,24], such as the development of new soft ionization sources (ESI[38] or MALDI[39]) and of new acquisition methods. The constant instrumental developments of mass spectrometers and liquid-chromatography have also allowed for great improvement in sensitivity, specificity and resolution of proteomics analysis. Moreover, the continuous developments in automation of the workflow as well as relentless improvements of bioinformatics tools for data treatments have greatly contributed to the popularity of mass spectrometry proteomics.

1. The different strategies for mass spectrometry-based proteomics

Mass spectrometry-based proteomics can be performed either by keeping protein intact, or after enzymatic digestion, as illustrated on **Figure 14**.

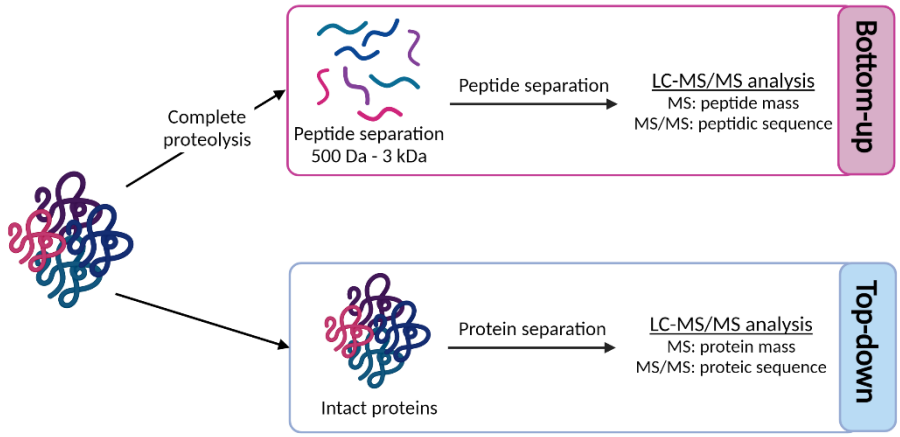


Figure 14: Schematic representation of the bottom-up and top-down approaches (Figure created with BioRender.com).

The bottom-up approach consists in characterizing protein through the analysis of their peptides, obtained after enzymatic digestion. Generated peptides, with a mass lower than 3 kDa, are separated by liquid chromatography, ionized and analyzed by MS to obtain their mass. Those ions are then selected and fragmented in a collision cell, and masses of the fragment ions are measured by MS/MS.

Peptide identification is achieved by comparing those measured masses to a theoretical mass list obtained by *in-silico* digestion of proteic sequences from the chosen database. Protein identification is then performed by inference[23]. As peptides are not necessary uniquely assigned to a single protein but can also be shared by more than one protein, the identified proteins may be grouped. However, this inference process remains complex and the results of the bottom-up analysis will be the smallest list of protein satisfying the principle of parsimony[40].

On the other hand, top-down approach allows the analysis of intact proteins by MS (protein mass) and MS/MS (protein sequence) without enzymatic digestion. It aims at providing high coverage and complete characterization of a targeted protein. It is especially used to differentiate proteoforms and to study PTMs[41,42]. It has however some limitations[43–45], particularly regarding sample preparation which requires proper extraction and solubilization of native proteins. Moreover, ionization and fragmentation of intact proteins require fine tuning, and are limiting factors of the approach's sensitivity. Mass spectrometers with high resolution, mass accuracy and scan speed are thus required in top-down analysis to finely resolve the isotopic envelopes of the multiple charge states proteins analyzed. Progress are also still ongoing regarding the interpretation of the complex spectra generated and their statistical evaluation.

While progress are ongoing to make top-down approach compatible with the study of complex protein mixture, bottom-up is still established as the gold standard for large scale and high throughput proteomics[46]. Moreover, MS is currently more widely used to study protein phosphorylation as the identification of phosphorylated residues within proteins is mostly done by MS these days, and bottom-up technique involves analyzing peptides derived from protein digests[47]. The work presented in this manuscript is only based on this approach and therefore is described step-by-step in the following section.

2. Bottom-up sample preparation

The bottom-up approach includes three key steps that are described in **Figure 15:** sample preparation, LC-MS/MS analysis, and data analysis.

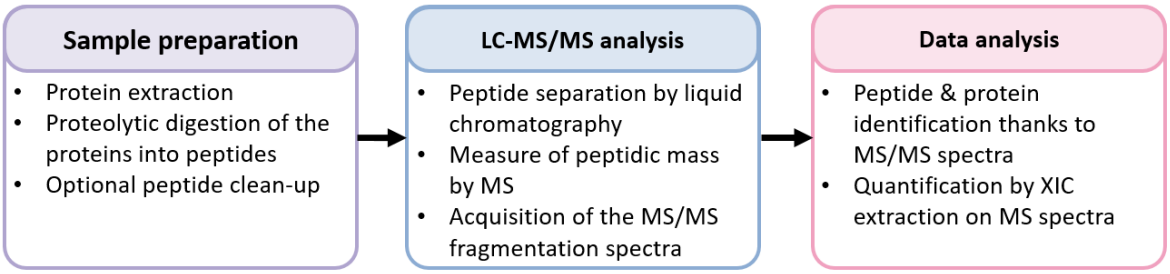


Figure 15: Description of the main steps of a bottom-up proteomics workflow**.**

The quality and the repeatability of bottom-up proteomics analysis greatly relies on sample preparation. Therefore, each step, from protein extraction to the injection on the LC-MS/MS system needs to be finely tuned for the analysis and the type of sample.

i. Protein extraction

One of the main strength of proteomic analysis is its versatility of applications. Indeed, it can be applied on various type of samples ranging from bio fluids such as CSF [48], plasma [49] or urine[50], but also tissues[51], micro-organisms[52], plant cells[32], or single cells[53]. The extraction of proteins from those samples is a major step, as proteins need to be available to protease for the enzymatic digestion. The goal is to

solubilize as many proteins as possible without degrading, modifying or introducing any bias. Cell lysis and protein extraction can be performed either mechanically or chemically.

Mechanical approaches, such as manual mortar and pestle or beads grinding under liquid nitrogen, are most used for tissues and cells. Samples can also be sonicated or ultra-sonicated to help lyse and/or denaturize the proteins [54]. Sonication is also helpful to desorb the proteins that may be adsorbed on tube walls. Many developments are being made in order to improve the efficiency, reproducibility and reliability of this step, for example: the Bioruptor (Diagenode)[55,56] , the CryoPrep extraction system (Covaris), or the BeatBox (PreOmics)[57].

The chemical approach is based on the use of denaturing detergents to maximize protein extraction and solubilization via micelle formation. Those detergents can be either ionic (SDS), non-ionic (Triton), zwitter-ionic (3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate [CHAPS]), or salts of bile acids that are less denaturant (sodium deoxycholate, SDC) [54,58]. All of those are not compatible with the enzymatic digestion and therefore need to be removed before the digestion. Other ionic detergents are however compatible with the enzymatic digestion RapiGest (Waters)[59,60] for example. The addition of chaotropic agents such as urea or thio-urea allows protein denaturation and unfolding of their structure. Organic solvents like ACN or MeOH also facilitate protein denaturation. Other conditions such as the concentration of the lysis buffer, its pH or the temperature, need to be adapted to the extraction process and the type of sample[54,58,61]. A combination of both chemical lysis and mechanical stimuli is however the most common approach[61].

Using detergents may not always be compatible with the LC-MS/MS analysis as they may contains impurities (salts, Triton…). Other impurities such as lipids or nucleic acids may be presents in the samples and interfere with the analysis by co-eluting with peptides of interest. A precipitation step may be added to remove get rid of those unwanted components. Most common protein precipitation are performed using trichloroacetic acid, ice cold ethanol, methanol or acetone, or a combination of two of those organic solvents together[54,62,63]. Other techniques might be used such as Solid-Phase Extraction (SPE), or filter based methods to remove contaminants. However, they have been shown to be more time consuming and prone to sample loss[64].

ii. Enzymatic digestion

Enzymatic digestion is one of the key step of bottom-up proteomics sample preparation. Indeed, an incomplete digestion might result in a high rate of missed cleavages, generating peptides not selected for MS fragmentation or leading to lower quality spectra[65,66]. Before performing the digestion step, there is usually a step to reduce and alkylate the disulfide bridges of the proteins in order to ensure to proteases the access to cleavage sites.

a. Reduction and alkylation

Disulfide bonds are the results of a covalent bond between two cysteine's' thiol group on a protein. They contribute to the stability of its structure[67]. Two of the most used reducing agents used in proteomics to reduce disulfide bridges are dithiothreitol (DTT)[68] and tris-2(-carboxyethyl)-phosphine (TCEP)[69]. They are both an alternative to beta-mercaptoethanol which is more toxic and less stable[70]. Sulfide groups are then alkylated to prevent the bridges to form again. Several alkylating agents may be used, the most common of them is iodoacetamide (IAM), but alkylation can also be performed using acrylamide or chloroacetamide[70]. The performances of those different reagents were evaluated on HeLa cells[70], highlighting the great impact of the choice of an adapted reductant and alkylant. Indeed, they have a non-negligible impact on the identified peptides as well as the reproducibility of the experiment.

b. *Used proteases*

After reduction and alkylation, proteins are digested, most of the time using trypsin. Trypsin owes its popularity to its high efficiency, cleavage-site specificity and relatively easy accessibility[66,71,72]. Trypsin cleaves specifically at the C-terminal end of lysine and arginine, except when they are followed by a proline, due to steric hindrance[70,71]. Thanks to the frequency of lysine and arginine residues, it generates peptides with a mass usually between 500 to 3000 Da, therefore suitable for the LC-MS/MS analysis. Moreover, a positive charge is present after cleavage at the basic C-terminal end of the peptides, which will favor their ionization and fragmentation, according to the mobile proton model, thus making tryptic peptides even more adapted to LC-MS/MS analysis. Tryptic digestion is usually performed at 37°C and pH 7, overnight (from 12 to 17 h). However, if an excess of enzyme is added or if the digestion is performed longer than recommended, trypsin might autolyze, leading to unwanted peptides that will compete with peptides of interest during the LC-MS/MS analysis[73].

For the past years, other proteases have been presented as alternatives. Due to the overall short lengths of tryptic peptides (56% of all generated tryptic peptides are smaller than 6 residues), they may be not identified by, therefore allowing only a limited segment of the proteome to be covered, especially when PTMs or proteoforms are involved[72]. To access a wider coverage, alternatives proteases have been investigated: chymotrypsin, pepsin, LysN, AspN, GluC, LysC or ArgC [66,74–76]. The combination of trypsin with another protease seems particularly efficient as shown on *S.Pombe* cells where a mix of trypsin and AspN allowed a fine improvement in terms of peptide identification[74]. In particular, the association of trypsin to LysC displays higher sequence coverage and its use has spread across the scientific community[77–79]. LysC has the advantage of cleaving on the C-terminal of lysine residues, complementing the action of trypsin. Trypsin digestion (or a combination of trypsin and lysC) remains however the golden standard in proteomics. The digestion can be performed through different protocols, presented in the next paragraphs.

c. *In-solution digestion*

With in-solution digestion (ISD), the enzyme is added directly into the proteic medium. Its main assets is the simplicity of the process, as described in **Figure 16**, and its low cost[55,58].

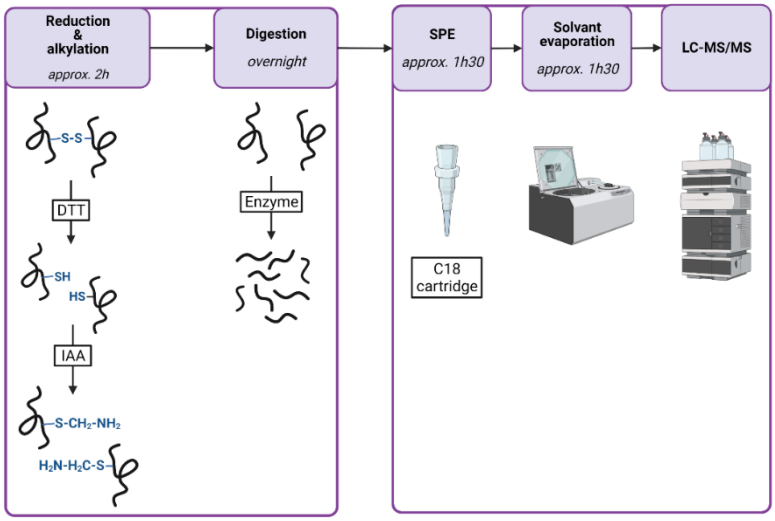


Figure 16: General scheme of in-solution digestion (Figure created with BioRender.com).

However, it requires the use of a buffer compatible with enzymatic digestion and LC-MS/MS analysis. Therefore, urea-based buffers are largely used as they are compatible with enzymatic digestion once

diluted. A clean-up step is then needed to remove urea and other salts that are not compatible with the LC-MS/MS analysis. Using C18 cartridges mostly, Solid Phase Extraction (SPE) is most of the time performed to clean-up peptides. They are retained based on their hydrophobicity, while salts and certain other contaminants are discarded. However, this SPE step needs to be performed with caution, as very hydrophobic peptides can remain attached to the cartridges. Moreover, adding an additional step to the workflow might lead to sample loss, especially if performed manually and when working on very low starting material amounts[64]. An alternative to classical SPE, called SP2, was recently developed[80]. It is based on the use of carboxylate-modified paramagnetic beads to retain peptides and therefore purify them from detergents and other contaminants. This protocol clean-up may be applied to phosphopeptides and glycopeptides, and might also be automated[80,81].

d. In-gel digestion

In-gel preparation is typically employed when SDS was used during extraction step. Indeed, Sodium Dodecyl Sulphate – PolyAcrylamide Gel Electrophoresis (SDS-PAGE) protocol, illustrated in **Figure 17**, was developed as a method compatible with SDS use.

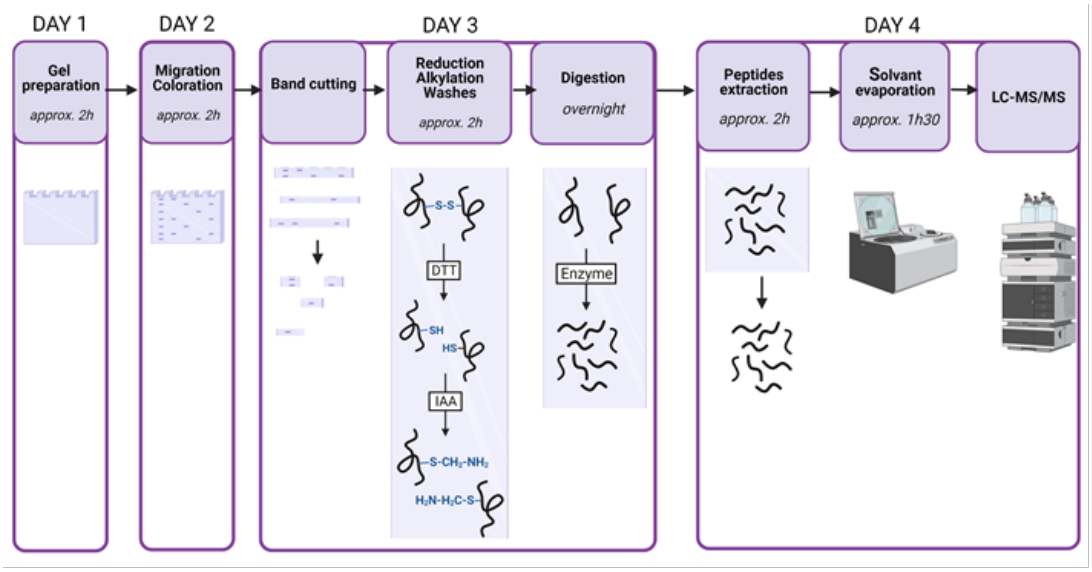


Figure 17: General scheme of SDS-PAGE in-gel digestion (Figure created with BioRender.com).

Extracted protein are first linearized and negatively charged due to SDS. They are then loaded on the gel and protein migration is performed with the application of an electric field. Different migration scenarios are possible:

- 1D SDS-PAGE: proteins are first concentrated in the stacking gel (composed of 4-5% of acrylamide/bis-acrylamide). Then, thanks to SDS that uniforms proteic charges, they are separated according to their molecular weight by migrating in the separation gel (composed of 8-15% of acrylamide/bis-acrylamide). This method allows fractionating samples in several bands.
- The stacking gel: proteins migration is performed only a few centimeters in the stacking, and stopped before entering the separation gel.

Protein presence is then revealed thanks to Coomassie blue. Gel bands are cut, proteins trapped in the gel are reduced, alkylated and gel pieces washed to remove SDS and other contaminants. Proteins are digested directly in the gel as trypsin is small enough to enter the gel, thanks to dehydration/hydration processes. Finally, peptides are extracted from the gel, with again dehydration/hydration processes[58,82]. In-gel protocol was for long the golden standard of proteomic sample preparation but

it is time consuming and hardly automatable[52,58].Therefore, for the past years, new sample preparation protocols, more adapted to high throughput and compatible with SDS have emerged.

e. *Membrane based digestion*

The In-Stage Tip (iST) protocol is filter-aided sample preparation-based method, in which all steps are performed in a single, enclosed, volume[83]. It is commercialized by PreOmics (Planegg-Martinsried, Germany) and the general protocol is described in **Figure 18.**

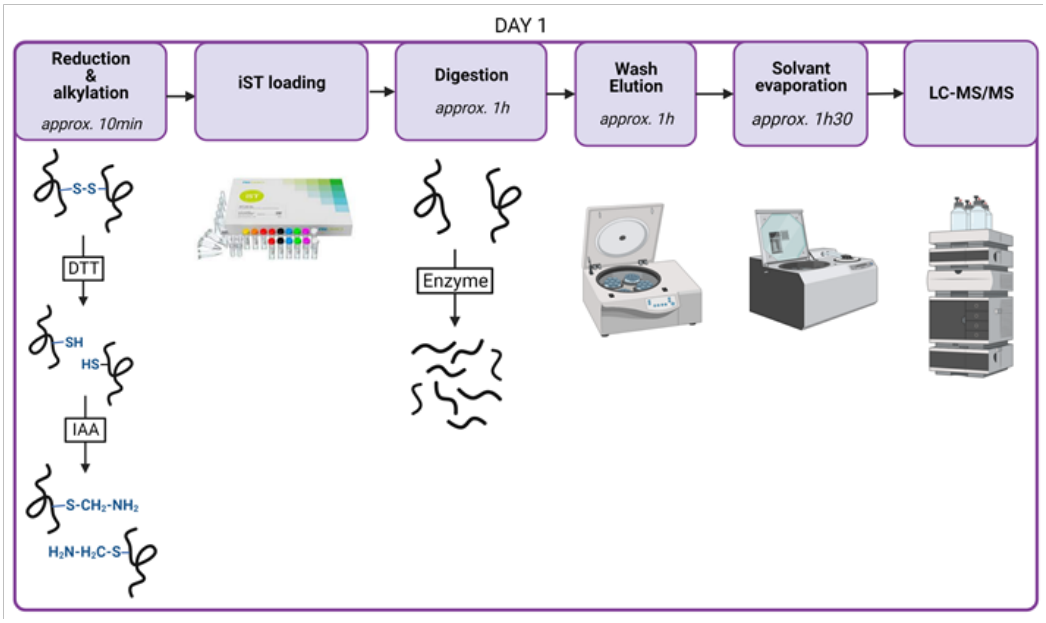


Figure 18: General scheme of on filter digestion, iST protocol (Figure created with BioRender.com).

First, cell are lysed, proteins are solubilized, reduced and alkylated using provided lysis buffer. All of these steps are performed under only 10 minutes. Proteins are then loaded on the iST cartridges to be digested at 37°C for 1 h. Peptides are washed thanks to centrifugation cycles and finally eluted from the cartridges and recovered. As the whole sample preparation is performed in a single reactor, it highly simplifies the workflow and thus minimize sample loss. Moreover, iST protocol is very fast and has shown robust results on various type of samples such as urine samples[50], plasma[84], HeLa cells[55] or cerebrospinal fluid[85]. IST protocol displayed high level of reproducibility as well as the best digestion efficiency when compared to other in-solution and S-Trap protocols[50,55]. However, as a commercial kit, the exact composition of the different buffers is unknown, and the cost of using the kit for sample preparation remains significantly higher than for other techniques such as in-solution or in-gel protocols. An automated version of the protocol is available on the PreOn robot also commercialized by PreOmics.

Other filter-based techniques, commercialized or not, exist, such as Filter Aided Sample Preparation (FASP)[86], Suspension Trap (S-Trap)[87,88], MStern[89], Sample Preparation Kit (Biognosys) or Pierce™ Mass Spec Sample Pre Kit (ThermoFisher Scientific). They are all based on a similar principle: proteins are trapped on a filter, then washed and digested on the filter before peptide recovery by centrifugation.

f. Digestion on beads

The Single-Pot, Solid-Phase-Enhanced, Sample Preparation (SP3) protocol was first described by Hughes *et al*. in 2014[90], then later improved[91,92] and finally automated on a BravoAssay Map platform (Agilent) in 2020[93]. SP3 protocol, described in **Figure 19**, uses paramagnetic beads functionalized with carboxylate groups. Proteins are bound to those beads in specific conditions, usually at physiological pH and with a 50% concentration in ACN. Mechanisms advanced to explain this binding are similar to those in play in Hydrophilic Interaction Chromatography (HILIC) and Electrostatic interaction Liquid Chromatography (ERLIC)[80,90]. Extended investigation showed that protein immobilization is also driven by protein aggregation induced by the high concentration of organic solvent such as ACN[94]. Beads carrying the beads are then retained thanks to a magnetic rack to go through a series of washing step to get rid of contaminants. Proteins are then enzymatically digested, and peptides are recovered from the beads again using a magnetic rack.

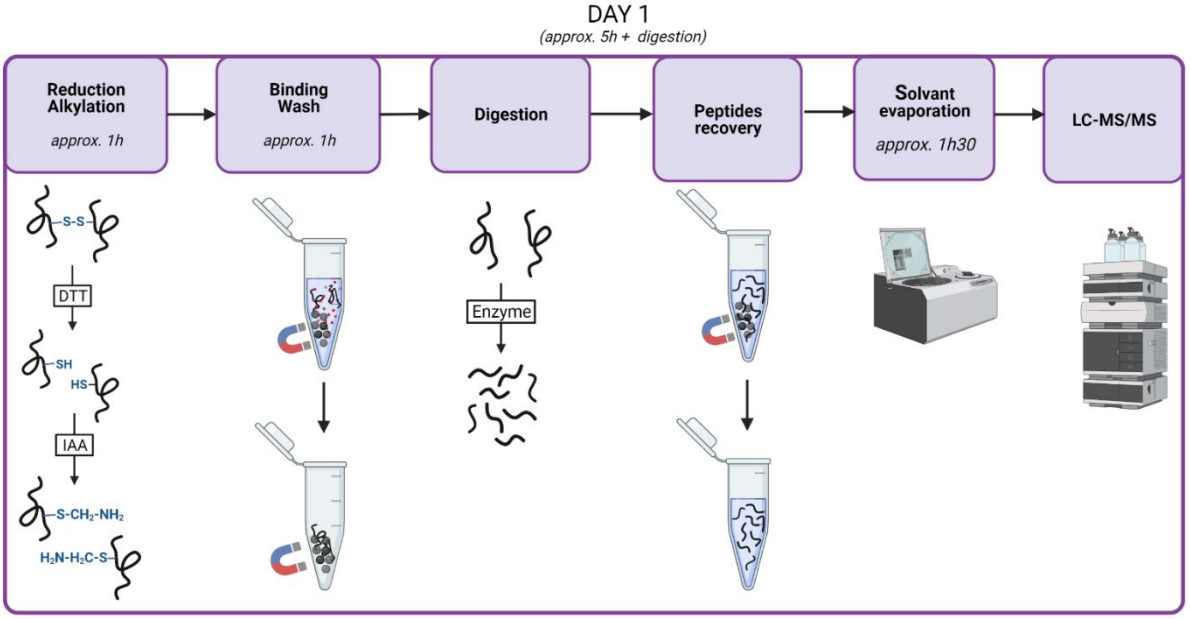


Figure 19: General scheme of bead-based digestion, SP3 protocol (Figure created with Biorender.com)**.**

The SP3 protocol has been applied to a wide variety of samples such as plants[95], aquaculture species[96], immunoprecipitation (IP)[97], paleoproteomics human bone samples [34], yeast[98], FFPE tissues[99]. SP3 protocol has also been combined to labeling method such as TMT[98,99] or SILAC[100] labeling. Comparing SP3 to both in-gel and S-Trap digestion protocols, on micro-organisms, SP3 outperformed the two other methods in terms of speed, and allowed more peptides and proteins identification than in-gel protocol[52]. On different quantities of HeLa cells (from 1 µg to 20 µg), SP3 was compared to iST and FASP methods[101]. All three methods showed comparable performances for 20 µg of starting material. However, for low amount of material (1 µg to 5 µg), SP3 and iST allowed higher proteome coverage and reproducibility compared to FASP. On plasma samples, SP3 was tested against in-solution digestion, and outperformed the latest in terms of quantified proteins[102]. Finally, on plant pathogen extract, S-Trap protocol increased protein identification and reproducibility compared to SP3[103]. Many other publications and reviews compare sample preparation methods together, highlighting that there is no universal protocol suitable for every single kind of sample[60,96,104–107].

SP3 protocol main advantage is its versatility, as it is applicable to almost any kind of bottom-up studies, but was also applied on top-down analysis[108,109]. It Is compatible with SDS up to 10%[91] and can

be easily applied to low protein amounts by adjusting the bead quantity, and was even applied on single cells[110]. The SP3 protocol has also already shown its aptitude to be automatized, in particular on an AssayMAP Bravo platform (Agilent), as shown by different groups[93,111]. SP3 protocol was also adapted as SP2 to purify peptides[80] , as solvent precipitation SP3 (SP4)[112] using non-magnetic inert beads and as universal SP3 (USP3)[109] for both bottom-up and top-down sample preparation.

iii. Automation of proteomic sample preparation

For the last years, proteomics have evolved to become a method of choice to perform high throughput quantitative analysis on large cohorts of samples in the hope of finding new biomarkers for various diseases[113,114]. Sample preparation is a tedious, time consuming step, at the root of many unwanted variations. Therefore, there is a real need for automation, not only to increase the throughput, but most of all to improve the reproducibility and repeatability of the analysis. Automation also allows to reduce even more the amounts of starting material[115]

Many steps of proteomics sample preparation, even complete protocols have been automatized. Liquid digestion, as well as S-Trap, SP3[93], or TMT labeling[116] are compatible with automated platforms. Reduction and alkylation steps, SPE, reversed-phase fractionation may also be performed in an automated way[1]. For the analysis of PTMs such as phosphorylation or others, the enrichment step might also be automatized[117–119]. Many liquid handling platforms have been developed. Two of the most widespread are the AssayMAP Bravo (Agilent) and the KingFisher Flex (Thermo), both equipped with a 96 tip automatized head. Other platforms might be mentioned such as the MicroLab Star (Hamilton), the Resolvex A200 (Tecan), the Biomek workstation series (Beckman Coulter life science) or the PreOn platform (PreOmics)[115]. Recently, interfaces between sample preparation steps and LC-MS/MS system have emerged, such as the cellenONE robot (CELLENION)[120] to perform single cell analysis, or the ADE-OPI-MS developed by AB Sciex[121]. All these techniques allow to reduce drastically potential loss of material as well as analysis time.

3. Liquid chromatography coupled to tandem mass spectrometry

i. Peptidic separation by liquid chromatography

While essential to classical bottom-up proteomics approach, enzymatic digestion greatly increases sample complexity with tens of thousands of peptides to analyze. Therefore, liquid chromatography is most of the time used to decomplexify the peptidic mixture before MS analysis. It thus allows to improve ionization efficiency, as well as sensitivity, specificity, and proteome coverage[122]. For proteomics analysis, the most common system is reversed-phase high performance liquid chromatography (RP-HPLC). This type of chromatography elutes peptides according to their hydrophobicity properties by decreasing the polarity of the mobile phase using a mixture of water and ACN. For the results presented in this manuscript, two different LC systems were used and are described in the following **Table 1**.

LC system	UHPLC NanoAcquity	UHPLC NanoElute
System vendor	Waters	Bruker Daltonics GmbH
Column vendor	Waters	Ion Opticks
Stationary phase	C18	C18
Column length (mm)	250	250
Internal diameter (µm)	75	75
Particle size (µm)	1.7	1.6
Pore size (Å)	130	120
Flow rate (µL/min)	350 or 400	300

Table 1: Description of the LC systems used in my PhD work.

Multiple factors are involved in the quality of the peptide's separation by LC. They are of different types: (i) linked to the system itself (composition of the solvents, flow rate, gradient…) or (ii) inherent to the column (internal diameter, particle size, pore size, length of the column)[123–125]. Ultra High Pressure Liquid Chromatography (UHPLC) refers to chromatographic systems that are working with nano-columns at flow rates lower than µL per minute (nanoLC) and at high pressure (>400 bars). They require much lower amount of loaded material (between 100 to 350 ng of peptides), while increasing the resolution, sensitivity and peak capacity, rendering them well suited for proteomic analysis of low abundant samples. However, they face common technical problems (spray instability, hardly detectable leaks, and dead volumes) and thus require expert handling.

ii. Tandem mass spectrometry

Once eluted, peptides enter the mass spectrometer and are ionized via Electro Spray Ionization (ESI). Data Dependent Acquisition (DDA) is the most popular mode in bottom up analysis. Using this mode of acquisition, the mass over charge ratio (m/z) and the intensity of every ion are measured to generate a MS spectrum, as described in **Figure 20**. The top N most intense precursors are then selected, fragmented and analyzed, generating MS/MS spectra.

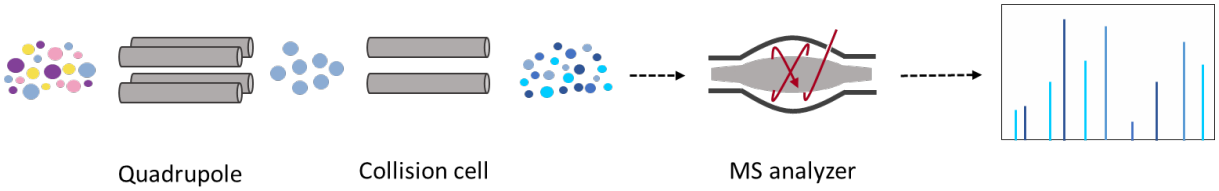


Figure 20: Schematic representation of a DDA analysis. An MS spectrum is acquired then the N most intense ions are isolated one by one by decreasing intensities. Once isolated, they are fragmented in the collision cell and fragments ions are analyzed generating MS/MS spectrum**.**

DDA mode enables the identification of thousands of proteins, offering a deep proteome coverage. However, as the ion selection for fragmentation is solely based on intensities, DDA remains a stochastic approach with a great lack of reproducibility especially for low abundant peptides. To increase the proteome coverage, some experimental parameters can be tuned such as dynamic exclusion to reduce spectral redundancy. Lists of exclusion or inclusion can also be set up for the same purpose[126].

Tandem mass spectrometry, defined by this successive acquisition of MS and MS/MS, is used on hybrid instruments that combine different mass analyzers (quadrupole (Q), Orbitrap (OT), time of flight (TOF), ion trap (IT) and Fourier transform ion cyclotron resonance (FT-ICR)). For the results presented in this manuscript, three mass spectrometers with two different configurations were used and are described in the following **Table 2**. One of them, the TimsTOF Pro, is equipped with an additional dimension of separation that is detailed in the next part, ion mobility.

MS system	Q-Exactive Plus	Q-Exactive HF-X	TimsTOF Pro
System vendor	Thermo Fisher Scientific	Thermo Fisher Scientific	Bruker Daltonics GmbH
Analyzer	Q-Orbitrap	Q-Orbitrap	Q-TOF
Resolution	17 500 to 140 000 (at 200 m/z)	7 500 to 240 000 (at 200 m/z)	40 000 (at 622 m/z)
Mass accuracy	5 ppm	5 ppm	10 ppm
Acquisition speed	Up to 12 Hz	Up to 40 Hz	> 100 Hz
Fragmentation type	HCD	HCD	CID
Ion mobility	/	/	TIMS
Year of installation	2014	2017	2019

Table 2: Description of the MS systems used in my PhD work.

iii. Ion mobility spectrometry

Ion mobility (IM) spectrometry has been used for a long time combined to MS for different applications (isomers separation, signal filtering, characterization of intact proteins,…)[127,128]. It was only recently extended to bottom-up proteomics with the development of the Field Asymmetric Ion Mobility Spectrometry (FAIMS, ThermoFisher Scientific)[129] device and the Trapped Ion Mobility Spectrometry (TIMS, Bruker)[130]. Implemented in TimsTOF Pro mass spectrometer, the dual TIMS cell just before the quadrupole allows to separate ions by their charge and size in gas phase and to have an additional level of information: the ion mobility value (K0). Most of the time, we are talking about the reduced ion mobility (1/K0) or the Collision Cross Section (CCS) which is the 2D projection of the 3D structure of an ion. The CCS can be calculated from the ion mobility, with the Mason-Schamp equation. To fully take advantage of this additional separation dimension, Parallel Accumulation Serial Fragmentation (PASEF)[130] has been developed and is composed of two main steps:

- Ion separation in the dual TIMS cell: the first part of the TIMS cell is used to accumulate ions. Ions are then driven in the TIMS cell by an inert gas flow and retained by the application of a static electrical magnetic field. When the set accumulation time is reached, accumulated ions are transferred into the second part of the TIMS cell where they are separated according to their shape and charge in gas phase. For a same charge state, larger ions will by dragged by the gas flow further in the funnel and therefore be closer to its exit. Ions are then sequentially eluted by decreasing CCS into the quadruple by decreasing the electrical field. At the same time, the first part of the dual cell has accumulated other ions, and the cycle is repeated allowing a duty cycle of nearly 100%, meaning 100% of the ions incoming in the source are transmitted into the analyzer.

- Targeted isolation of the eluted ions by the quadrupole: the elution of the precursors from the second part of the TIMS cell is synchronized with their selection by the quadrupole using real time MS data treatment. It is possible to select precursors with a specific m/z and CCS using the PASEF scan mode. The focusing effect of the TIMS cell improve the ion utilization rate while increasing the duty cycle. Every MS/MS spectra corresponds to a given elution voltage and a given ion mobility value.

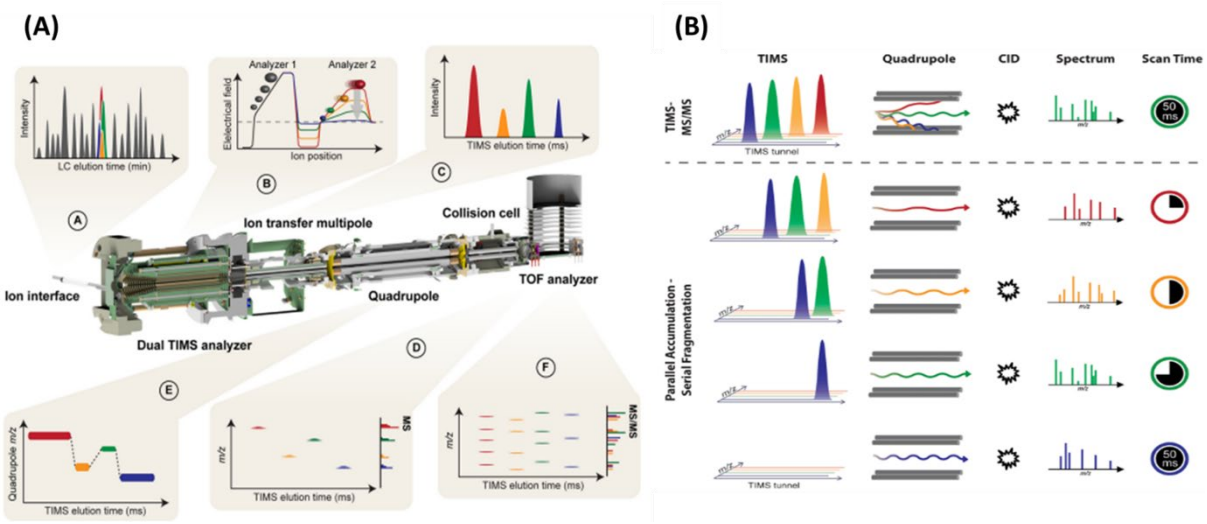


Figure 21: (A) A full PASEF cycle: (a) peptides elution and ionization (b) accumulation of the ions in the first cell and ion separation in the second cell (c)/(d) simultaneous sequential elution from the second cell and TOF analysis to produce MS spectra (e) targeted isolation of the eluted ions by the quadrupole (f) multiple ion mobility resolved mass spectra obtained[131] ; (B) PASEF mode principle: depending on the ion mobility value, m/z is selected in a targeted manner by the quadrupole[130].

This additional separation allows to isolate co-eluted peptides, for example isobaric phosphorylated peptides[132,133] (bearing the same modification located on different amino acids). Moreover, the PASEF acquisition mode offers to the TimsTOF Pro an extremely fast acquisition speed and sensitivity with more than 100 MS/MS spectra acquired in 1 second. As an additional separation dimension is added, we are talking about a four-dimension separation: retention time, peptide mass measured by MS; fragment mass measured by MS/MS and ion mobility value.

iv. Peptide fragmentation

The goal of peptides' fragmentation is to break the peptidic bond between amino acids, generating fragments that will generate MS/MS spectra. Various fragmentation methods have been implemented each fragmented at different locations on peptides. The different types of fragments generated are named according to the Biemman classification as illustrated on **Figure 22**.

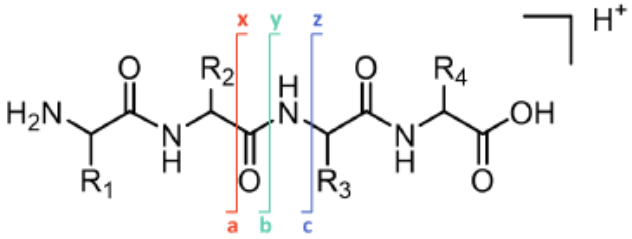


Figure 22: Biemann nomenclature for peptide fragmentation. a-, b-, c- ions carry the positive charge at the N-terminal part while x-, y-, z- ions carry it on the C-terminal part**.**

In bottom-up proteomics, Collision Induced Dissociation (CID) and Higher Collision Dissociation (HCD) are the most common fragmentation techniques. In both cases, ions are accelerated to increase their kinetic energy and then collide with neutral gas atoms (argon, helium, nitrogen) in the collision cell. The produced kinetic energy is converted into internal energy, inducing the breaking of the peptidic bond according to the mobile proton model[134]. With HCD, ions are accumulated in the C-Trap prior to fragmentation then sent to the collision cell, fragmented, sent back to C-Trap that will finally sent the fragments to the analyzer. Therefore, HCD fragmentation is specific of Orbitrap instruments. Both CID and HCD methods generate b- and y- ions, as described in **Figure 23**.

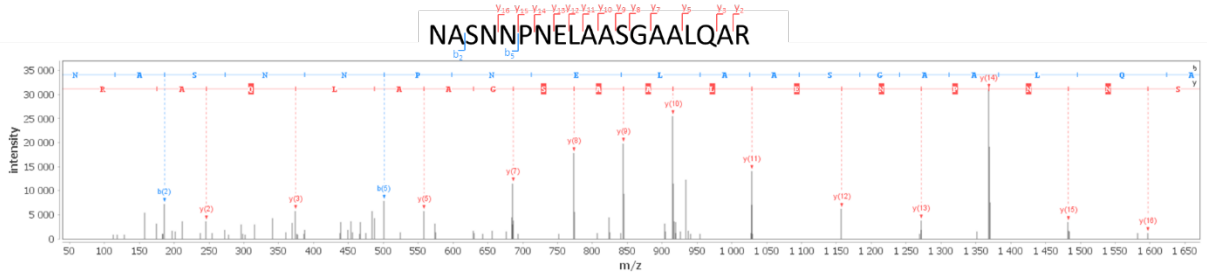


Figure 23: MS/MS spectrum from NASNNPNELAASGAALQAR peptide obtained on a Q-Exactive Plus with a HCD fragmentation cell. Mostly b and y ions are generated and allow the identification of the amino acids of the peptide.

Electron Transfer Dissociation (ETD)[33] can also be used, and generates mostly c- and z-ions. The reaction of the peptide with an anion (most of the times fluoranthene obtained by chemical ionization) via an electron transfer leading to the fragmentation along the peptidic chain. ETD fragmentation is considered softer than other techniques, thus being useful for labile PTMs analysis[135,136]. Other fragmentation techniques exist such as Electron Capture Dissociation (ECD)[137], UltraViolet Photo-Dissociation (UVPD)[138], Electron-Transfer/Higher-Energy Collision Dissociation (EThcD, which combines ETD and HCD)[139], but they are not used in the presented work and therefore will not be detailed here.

4. Data analysis and interpretation

i. Protein identification and validation

a. Search engines

In a typical computational pipeline, data acquired by mass spectrometer, namely m/z ratios from MS/MS spectra, ion measured intensities and retention times, are compiled into a peaklist. Those experimental data are then searched against theoretical spectra corresponding to peptides from *in-silico* digested proteins, generating Peptide Spectrum Matches (PSMs)[140]. Identification of peptides then lead to protein identification through inference process. Peptides might be either unique to a protein or shared between several proteins. In the latest case, proteins are then grouped in a protein group, while respecting the principle of parsimony so that the protein group is the smallest list possible. Protein inference remains a tricky process, as the same peptide sequence can be present in multiple proteins, therefore leading the identification of such shared peptides to ambiguities as to the correct identity of the sample proteins[40]. Those steps are performed automatically by search engines such as Andromeda[141], Mascot[142] (Matrix Science), MS Amanda[143], Pulsar (Spectronaut, Biognosys), Sequest (Thermo) and others.

Various information is needed by those software to perform the search, regarding both experimental and instrumental parameters listed below:

- The sequence database defining the search environment
- The enzyme used for proteic digestion and the maximal number of authorized missed cleavages
- The tolerance on the m/z ratios at both precursor level (MS spectra) and fragment level (MS/MS spectra)
- The allowed charge for both precursors and fragment ions
- The modifications of some amino acids that need to be searched, either fixed or variable
- The type of fragmentation

For the work presented in this manuscript, only Mascot, Andromeda, MS Amanda and Pulsar were used and therefore will be detailed here.

Mascot is a proprietary search engine, commercialized by Matrix Science (London, UK). Its probability-based algorithm calculates from each MS/MS spectrum an ion score. The ion score evaluates the probability that the match between an experimental MS/MS spectrum and the theoretical one occurs by chance, *i.e.* is a false positive. The higher the score, the more robust the peptide identification. The score is given at the PSM level which corresponds to all the identifications associated to a MS spectrum.

Andromeda is a search engine developed by Cox *et al.*[141] and integrated to the MaxQuant software. It works similarly to Mascot (a score is attributed to identifications through a notation system based on probabilities) and displays comparable results[141].

MS Amanda is a free right but not open source algorithm developed by Karl Mechtler's group in the Institute of Molecular Pathology (Vienna)[143]. It can be either used on its own or through Proteome Discoverer (Thermo) software. Its principle is similar to both Mascot and Andromeda, but was especially designed for high mass accuracy data. It shows increased peptide identification performances compared to Mascot on both HCD and ETD data[143].

Pulsar search engine was developed by Biognosys (Schlieren, Switzerland) and has been implemented in Spectronaut software. It is dedicated to the generation of spectral libraries, used for DIA data extraction in the Spectronaut workflow. As it is a commercial solution, its algorithm is not available.

Despite all the software developments, in average around 60% to 75% of MS/MS spectra remains unidentified[144,145]. This can be explained in a non-exhaustive manner by:

- Errors during data processing such as incorrect peak extraction, incorrect monoisotopic peak attribution, or attribution to a wrong charge state.
- A poor or insufficient MS/MS spectra quality, if for example the global intensity is too low or if the amount of fragments in insufficient.
- The quality of the database is also of foremost importance (see **4.i.b.Protein databases**), it needs to be adequate to the sample studied and as complete as possible.
- Chimeric spectra might be generated if multiple precursors are co-isolated and co-fragmented despite the small DDA isolation window (between 1-3 m/z). To overcome this issue, Mascot (since version 2.5) allows to identify all peptides possible from chimeric spectra. Also, Andromeda offers a second peptide search on MS/MS spectra, if parent ions with close masses have been detected, performed after retrieving fragments that were used for the first peptide identification. Since 2018, Charmer is implemented in MS Amanda to specifically deal with chimeric spectra with a similar principle than MaxQuant second peptide search[146].
- The presence of various modifications on peptides can induce mass variations compared to the non-modified peptide sequence. If the modification was no specified in the search parameters, modified peptides will not be identified. A large fraction of unidentified spectra is suspected to be due to this phenomenon[145,147,148]. Indeed, PTMs identification trough "classical" searches is restricted as (i) it assumes prior knowledge of the modifications present

in the sample (ii) the consideration of multiple modifications lead to an explosion of needed informatics resources and calculation time. Therefore, new search engines are being developed, faster and specifically adapted to enzyme free searches or multiple PTMs searches. Those open searches software such as IonBot[16], MSFragger[149] or SpecOMS[150] will be detailed in the appropriate later in this state of the art.

b. *Protein databases*

The choice of the appropriate database is crucial for the correct identification of peptides and proteins as it defines the search environment. A proteic database with redundant or incorrect sequences can greatly impair the analysis, thus a high quality database (achieved by data curation and sequence annotation) is necessary. Several databases are available, differentiated by their quality, completeness, and degree of sequence redundancy, among which:

- UniprotKB[151,152] , created through the collaboration of the European Informatics Institute (EBI), the Protein Information Resource (PIR) and the Swiss Institute of Bioinformatics (SIB) in 2002. It contains two different banks:
    - SwissProt is the reference proteic database used in proteomics when there are enough entries for the studied organism. All entries have been manually annotated and curated to keep only high quality information on protein sequences. Additionally to the proteic sequence, a wealth of information is available to the user such as its function, sub-cellular localization, known PTMs, sequence variant, interactions with other proteins, or structure.
    - TrEMBL contains automatically annotated and classified translating coding sequences from GenBank, as well as sequences from literature. Those sequences are waiting to be manually curated before validation to integrate SwissProt. The quality of those sequences is relatively poor as they contain many redundancies and errors.

  Both are regularly updated: on the 11th of April 2023, SwissProt counted 569 213 reviewed entries while TrEMBL counted 245 871 724 unreviewed entries.
- Refseq[153] was created by the National Center of Biological Information (NCBI). Its data are coming from automatized annotation of genomic data. A portion of these data is validated manually, leading to proteins with a NP prefix, while others have a XP prefix. For all proteins, the link between gene, transcript, and protein is available. On the 11th of April 2023, it counted 254 500 694 entries. However, this database may be subject to sequence errors due to issues while translating nucleotide sequences into peptides sequences, which can negatively impact MS data interpretation[154].

Databases are in constant evolution thanks to day-to-day curation, annotation and discovery of alternative/variant sequences for instance. Therefore, re-analyzing data with updated databases allows to extract more information and to decrease the amount of non-exploited MS/MS spectra[155,156], as previously discussed.

c. *Validation of protein identifications*

Search engines are not perfect and might introduce false peptide/protein assignations in the results. The score associated to each identification is not sufficient to assess the veracity of identification and thus it is necessary to validate the peptide and protein identifications. Automated validation methods have been developed, the most widespread being the target-decoy (TD) strategy[157] used to evaluate the false discovery rate in a dataset. To do so, decoy proteins are added to the database. Decoy

sequences are usually created from reversing the proteic sequences from the original database, thus obtaining the same number of decoy as of real proteins, which have a very negligible probability to exist in experimental data. Then every decoy identification is by definition a wrong identification, allowing to evaluate a false discovery rate (FDR) for the dataset. The FDR is calculated as followed:

$$FDR\ (in\ \%) = \frac{Number\ of\ assigned\ decoy\ sequences}{Number\ of assigned\ decoy\ sequences + Number\ of\ assigned\ target\ sequences} * 100$$

Equation 1: FDR calculation by target/decoy approach.

FDR can be determined at PSM, peptide and/or protein levels. Assignations are sorted by score and filtered until obtaining a FDR lower that the defined threshold. In most proteomics articles and journals, a 1% FDR is the common guideline but is not a global consensus. Different software allow to filter thanks to the TD strategy among which Proline[158], Andromeda or Proteome Discoverer. The latest however also enables the use of another validation method based on machine learning algorithm, namely Percolator[159].

ii. Open modification searches

As detailed previously, a large amount of spectra are usually unmatched in a classical proteomics analysis. One of the reason for this is the restricted search space of current engines that can therefore not identify peptides with unexpected modifications. Indeed, if a particular modification has not been specified in the search settings, then spectra corresponding to peptides bearing this modification will be assigned an incorrect amino acid sequence[15]. These missed modifications have a damaging impact on identification performance, as such false hits tends to obtain higher scores than other false-positives matches[160]. In this optic, Open Modifications Searching (OMS) tools have been developed to identify modified spectra.

During OMS, a very wide precursor mass window exceeding the delta mass induced by the PTM is used, making possible the comparison of modified query spectrum to its unmodified counterpart. Compared with the number of candidate peptides in a classical search, it is approximately three orders of magnitude higher using OMS[161]. This increased in search space, while allowing to identify a wide range of modified spectra, comes at the cost of a drastic increase in computational cost[15]. One of the approach to overcome this is the development of OMS software using spectral libraries for the identification of unanticipated modifications. Among those tools can be named QuickMod[162], pMatch[163] , the hybrid method[164] or the Approximate Nearest Neighbor Spectral Library (ANN-SoLo)[15] search tool.

Another way to speed up OMS are based on the use of a sequence database instead of a spectral library. MSFragger[149] uses an index of theoretical fragments, created from *in silico* digestion of the protein database, to quickly compute the number of shared fragments ions between a query spectrum and theoretical spectra. MSFragger allowed for a 100-fold improvement in speed over existing proteome database search tools[149]. SpecOMS[150] uses an algorithm to compare within minutes a whole set of experimental spectra to a whole set of theoretical spectra deduced from a protein database. Other similar OMS tools using sequence database can be named such as PTM-Invariant Peptide Identification (PIPI)[165], MODa[166], TagGraph[167] or Open-pFind[168]. PIPI, MODa and TagGraph engines work using a similar strategy: sequence tags are matched to measured spectra, then multiple matching tags are aligned and delta masses between the tags or between the sum of the tags and the precursor mass are reporter as mass(es) of the modification(s)[169]. Recently, Schulze *et al.*, based on the results of combining results from multiple classical search engines, demonstrated an increase in 8-18% in PSM by combining MSFragger, PIPI, MODa and TagGraph open search results[169].

Machine learning, as for classical search engines, is getting more and more popular with OMS. Indeed, OMS tools implemented with deep-learning and machine-learning based algorithms are emerging such as yHydra[170], IonBot[16] or the last version of the ANN-SoLo tool[171]. ANN-SoLo and IonBot both use different rescoring algorithms, with the difference that IonBot rescoring is not limited only to first-ranked PSMs and takes into account retention times (DeepLC[172]) and peak intensity (MS2PIP[173]) predictions, allowing to increase even more identifications compared to other OMS tools[16].

iii. Strategies for label free quantification

There are various strategies for protein quantification. They can be either label-based, meaning samples will be labeled before LC-MS/MS analysis (chemical, enzymatic and metabolic labeling), or label-free. Even if labeling strategies are often considered more accurate for protein quantitation compared to label-free, they are much more expensive (expensive isotope labels), they need specific software for data analysis, and specific expertise to analyze generated data. Moreover, label-free approaches can be applied to every type of samples and are not limited in number of samples compared to label-based where samples are multiplexed[174]. For those reasons, label-free approaches are most of the time preferred for proteomics analysis. However, every step of the workflow needs to be perfectly reproducible as every sample are processed and analyzed on their own. In the work presented in this manuscript, peptides' and proteins' quantification was performed exclusively using a label-free extracted ion chromatogram (XIC) approach, which is the most widespread technique used in proteomics.

In the XIC approach, the intensity of the chromatographic signal obtained in MS (if using DDA) or in MS/MS (if using DIA) is considered proportional to the peptide abundance. The ion current corresponding to a peptide is extracted, then the area under the curve measured to obtain the quantitative values. XIC-based quantification is performed at the MS level, but peptide identification is enabled by data collected at the MS/MS level. Therefore, the acquisition parameters of the method need to be optimized on one hand to obtain enough MS spectra to define the chromatographic peak, but also on the other hand to generate enough high quality MS/MS spectra to reach a satisfying depth of coverage[175,176].

This technic however requires complex data processing, as retention time alignment and intensities normalization might be necessary to compensate for potential variations induced throughout the workflow (sample preparation, chromatographic fluctuations, and signal instability). Different software packages offer solutions such as in MaxQuant (Max Plank Institut), Proline (Profi) or Proteome Discoverer (Thermo).

In order to partially overcome the stochasticity of this method in DDA, an algorithm can be used to search for the identity of a peptide detected in MS in an analysis but not identified due to the poor quality of the MS/MS spectrum or the absence of an MS/MS spectrum. The information of the unidentified MS signal, *i.e.* its m/z ratio and retention time, will be searched in other analyses processed in parallel. If an MS signal corresponding to an identified peptide is detected in another run, the identification will be transferred to the first assay whose signal did not trigger identification in the first place. This option is called Match Between Runs (MBR) in MaxQuant[177] or Cross Alignment in Proline[158].

To ensure quality and robustness of the label-free workflow, quality controls must be set up to control the stability and reproducibility of analysis. Indeed, as label-free approaches are not limited in the number of samples, injections might sometimes last for weeks. It is therefore necessary to ensure the stability of both the chromatographic system and the mass spectrometer.

In the context of the label-free analysis performed in this manuscript, two different techniques have been used, either separately or together:

- Internal standards were added to all samples: indexed Retention Time (iRT) peptides[178] are a mixture of 11 synthetic peptides commercialized by Biognosys (Schlieren, Switzerland). They are added in equivalent quantity in each sample, and their retention times and intensities are then followed throughout the analyses.
- External standards were analyzed regularly throughout the cohorts of analyses. The goal of those standards is to be representative of the biological samples, therefore they were constituted from an equivalent quantity of each analyzed sample. Those standards were made up before enzymatic digestion. Then this reference sample (called pool) follows the same sample preparation steps than any other sample. It is injected at regular intervals throughout the course of injections. The variations of the proteins' intensities are then measured during the sequence of analysis.

Once the intensities of the proteins are calculated for both internal and external standards, statistical tests are carried out to check if the protein abundance varies significantly across samples.

Chapter 2: Challenges in phosphoproteomics

1. The biological importance of protein phosphorylation

Post-translational modification (PTM) refers to the modification of a protein through the attachment of functional groups to its amino acid chain, or the side chain of a particular amino acid as well as a proteolytic cleavage. PTMs are usually implemented by dedicated and highly specific enzymatic systems (enzymatic PTM). Identification, characterization and mapping of these modifications to specific amino acids residues on proteins are critical to decipher their functional significance. Understanding these modifications is also important to develop novel targeted therapies for many pathologies such as cancers and neurodegenerative diseases[179]. To date, more than 300 types of PTMs[19] such as acetylation, carbonylation, phosphorylation, etc... , are known to occur physiologically and the implication of some of them has already been clearly recognized during the transformation of normal cells into tumor cells[20].

Among them, protein phosphorylation is one of the most abundant and important. Phosphorylation is a reversible reaction tightly controlled by balanced activities of two enzyme groups: kinases and phosphatases. Kinases substitute a neutral hydroxyl group (OH) by a tetrahedral phosphate group (PO_4^{2-}) coming from Adenosine TriPhosphatase (ATP) molecule while phosphatases remove it[180,181]. This highly dynamic and fast (of the order of a second[182]) reaction mechanism is described in **Figure 24**.

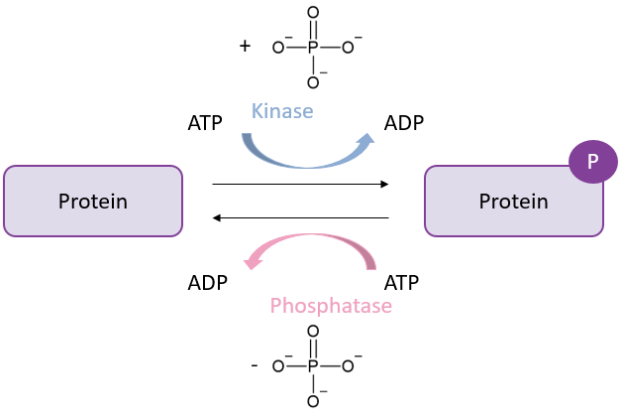


Figure 24: Schematic representation of the phosphorylation process. ATP = Adenosine TriPhosphatase ADP = Adenosine DiPhosphatase P=Phosphorylation.

Nearly 75% of all human proteins may be phosphorylated, and these phosphorylation events mainly occur on serine (86.4%), threonine (11.8%) and tyrosine (1.8%) residues[181]. Other amino acids might be phosphorylated like cysteine, lysine, arginine, histidine, aspartic and glutamic acids[181,183]; they however represent an extreme minority. Phosphorylation can happen on one (mono-) or multiple (multi-) sites and can co-exist with other PTMs, generating different proteoforms[184]. In the human proteome, it is estimated that there are approximately 13 000 phosphoproteins and around 230 000 phosphorylation sites[185], and in the human genome, 568 kinases and 156 phosphatases that regulated phosphorylation events[181]. An estimated 30-75% of proteins might be phosphorylated, although their precise function is still unknown[182,186]. This modification of charge state and steric environment modifies chemical and electrostatic properties and creates opportunities for new interactions. For all those reasons, protein phosphorylation is one of the most studied and important modifications. The balance between phosphorylation and de-phosphorylation is tightly regulated and controls many cellular processes, as described in **Figure 25**.

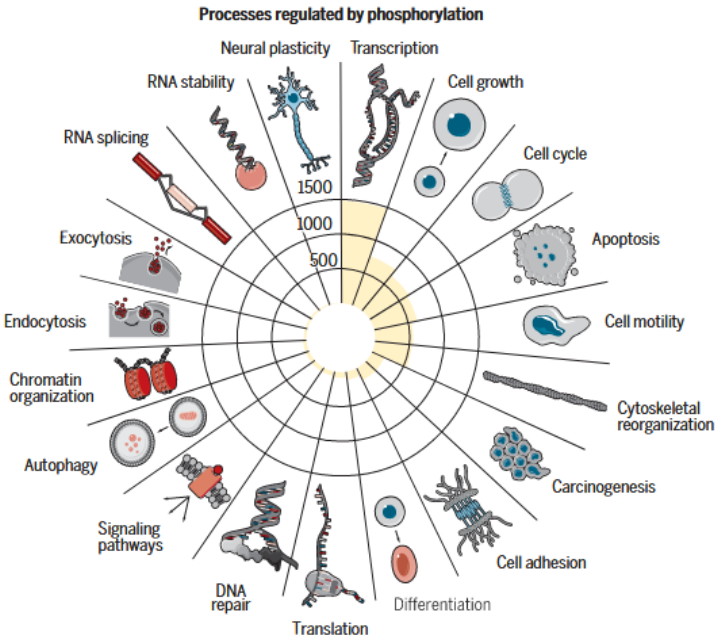


Figure 25: Processes regulated by phosphorylation[4].

Dysregulation of phosphorylation plays a key role in the development of various diseases such as numerous cancers[187,188], neurodegenerative pathologies like Alzheimer's disease (AD)[6,189] or Amyotrophic Lateral Sclerosis (ALS)[190–192], cardiovascular diseases[193], or diabetes[194]. Currently, there are 72 Food and Drug Administration (FDA)-approved therapeutic agents that exist and target about two dozen different protein kinases. Three of these drugs were approved in 2022[195].

The study of phosphorylation is promising for the discovery of new treatments and it is therefore necessary to determine the localization sites, abundance and role of phosphorylations to better comprehend cellular signaling and dysregulations. Mass spectrometry has step up as one of the gold standard to study protein phosphorylation, thanks to its untargeted nature and its high throughput capacity[196,197].

2. Analytical challenges of the study of protein phosphorylation by mass spectrometry

To study protein phosphorylation, MS-based methods have become popular in recent years. Indeed, mass spectrometry enables accurate identification and quantification of expressed proteins as well as identification and localization of PTMs[17]. Yet, the analysis of protein phosphorylation poses many technical hurdles at every step of the phosphoproteomic workflow[7,182], as detailed in **Figure 26**.

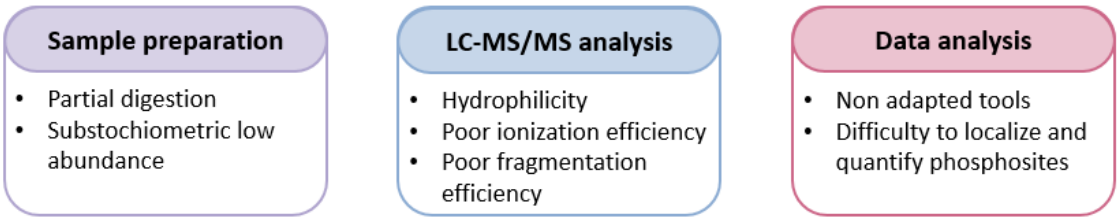


Figure 26: Challenges encountered in the study of phosphorylation during the 3 main steps of a phosphoproteomic workflow: sample preparation, LC-MS/MS analysis and data analysis.

i. Sample preparation

First of all, as phosphorylation are labile and thermo-sensitive modifications[182,183], sample preparation steps must be, when possible, realized ice-cold. To keep the original phosphorylation status and stop the reversible reaction, protease and phosphatase inhibitors might be added during cell lysis[183]. For proteolytic digestion, even if trypsin is usually the enzyme of choice in proteomics, its efficiency is strongly impaired on phosphorylated proteins. Indeed, the formation of hydrogen bonds and salt bridges between phosphate group and basic amino acids residues, such as arginine or lysine, might lead to missed cleavages and complicate phosphoproteome analysis[182,197]. This can be partially compensated by using optimized digestion, such as increasing the amount of protease to a 1:20 ratio (enzyme:protein) instead of classical 1:50 or 1:100 ratios in global proteomics[7]. A multi-enzymatic digestion, using for example consecutively LysC and trypsin, allowed to increase by 40% more phosphorylation sites than a one-step tryptic digestion[197]. Another technique has been developed, adding a digallium complex that exhibits a high binding affinity to the phosphate group and therefore mitigating the salt bridge formation. It has been shown to improve phosphopeptides abundance by approximately 17%[198]. Phosphorylated peptides are present in really low abundance in samples (< 2-3%) and therefore, their LC-MS/MS detection is hindered by the much more abundant unmodified peptides[186]. We might face signal suppression if peptides are co-eluted but also phosphopeptides may not being fragmented as in DDA, only the top N most abundant ions are isolated for fragmentation. Therefore, an additional step of phosphoprotein or phosphopeptide enrichment is commonly added to the classical proteomics workflow.

With phosphoprotein enrichment, proteins extracted from lysed cells are enriched, then digested into peptides and analyzed by LC-MS/MS. One of the advantages of phosphoprotein enrichment approach is that intact proteins are separated. Therefore, the peptide spectrum obtained is mostly derived from one protein. Protein identification is more likely since it has been achieved on the basis of several peptides (including non-phosphorylated ones) and not according to only one single peptide. However, one main resulting disadvantage is that this technic is not very specific and therefore not suitable for the study of less abundant proteins. Moreover, as this approach mainly generate non-phosphorylated peptides, phosphopeptides are very difficult to detect as their signal is hindered by the high abundant peptides.

For peptide enrichment, proteins extracts are first digested, then subjected to an enrichment step. Finally, enriched peptidic digest is analyzed by LC-MS/MS. With this approach, phosphopeptides identification should increase as we removed the unmodified peptides. Phosphopeptides enrichment approach is almost always preferred as it can led to precise phosphorylation sites localization and it is more easily automated than phosphoprotein enrichment[199]. Different techniques were developed to enrich phosphopeptides; the three most commonly used o are represented in the following **Figure 27**.

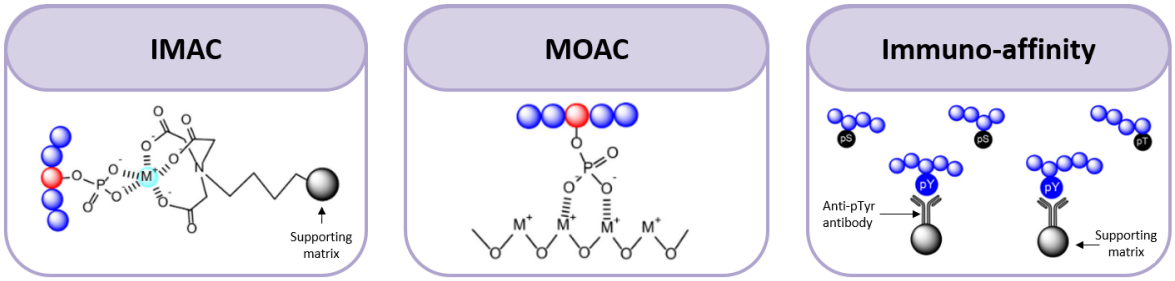


Figure 27: Principle of the three main strategies for phosphoproteomics enrichment: Immobilized Metal Ion Affinity Chromatography (IMAC), Metal Oxide Affinity Chromatography (MOAC) and immuno-affinity enrichment using anti-pTyr antibodies. In blue are represented amino acid residues and in red serine, threonine and tyrosine residues. (M+)= Metal oxide (from Arrington *et al.*, 2017[200]).

Immobilized Metal Affinity Chromatography (IMAC) exploits the affinity of phosphate groups for transition metal ions. The interaction is based on metal chelation and electronic attraction. There are three main components to an IMAC column: the metallic cations (usually Fe^{3+}, Ga^{3+}, Al^{3+}, Zr^{4+} or Ti^{4+}), immobilized by chelation using a coated ligand (mostly with iminodiacetic acid (IDA) or nitrilotriacetic acid (NTA)) on the stationary phase (magnetic beads or silica-based resins)[183,197]. Phosphopeptides are retained by the stationary phase in acidic conditions and then eluted in basic conditions[184].

Metal Oxide Affinity Chromatography (MOAC) is based on the formation of a stable bond between a metal cation in a metal oxide (TiO_2, ZrO_2 or Fe_3O_4) and an oxygen anion of the phosphate group[182,197]. Phosphopeptides are loaded onto the metal oxide in acidic conditions, usually with specific additives (lactic acid, glycerol, citric acid,..) to improve MOAC efficiency[186,201]. Phosphopeptides are then eluted at pH > 10.

Both IMAC and MOAC techniques are widely spread in laboratories are they are available at rather low cost if performed manually. Enrichment recovery and overall sensitivity of the two methods are generally comparable[182]. It was recently shown in a targeted phosphorylation quantification workflow that, while TiO_2 MOAC enrichment displayed higher recovery rate than Fe^{3+} IMAC (respectively 70% and 40%), IMAC provided a nearly perfect specificity (99% compared to 70% for MOAC)[202]. When comparing the two approaches, IMAC usually results in higher identification number of multiple-phosphopeptides while MOAC displays higher identification number of mono-phosphopeptides. It has been also shown on HeLa cells that the two techniques seem complementary as a phosphopeptides library recovery of only 42% was observed between MOAC and IMAC[203]. As of this date, no clear consensus exists on the optimum technique for phosphorylation enrichment.

Different strategies have been developed to increase the efficiency of those enrichment techniques. In the Sequential elution from IMAC (SIMAC) method, three fractions are eluted from IMAC phase: the non-retained fraction from sample loading, one fraction eluted at acid pH and one eluted at basic pH. The non-retained and acidic fractions are then mixed and undergo an additional MOAC enrichment step. The basic fraction contains multi-phosphorylated peptides whereas MOAC enriched fractions contain more mono-phosphorylated peptides[204]. Similarly, Sequential enrichment from MOAC (SMOAC), which uses serial enrichment with TiO_2 and Fe-NTA, has shown promising results[186,205]. Another strategy is also to decomplexify samples before or after phosphopeptide enrichment by fractionation using Hydrophilic Interaction Liquid Chromatography (HILIC)[206], Strong Cation Exchange (SCX)[183,207] or basic Reversed Phase (RP) chromatography[208].

An interest has especially been taken in the study of phosphotyrosine (pTyr) peptides. Tyrosine kinases, which account for only 0.3% of the genome yet, contribute to a disproportionally large percent (about 30%) of the known 100 dominant oncogenes. The regulation of kinase and phosphatase activates is crucial, and highlights the importance of an unbiased study of pTyr peptides to identify novel targeted therapies for patients[209]. For this reason, immune affinity based enrichment, using anti-pTyr antibodies, is specifically utilized to enrich and study those particular phosphorylations. However, this type of enrichment requires large amount of starting material as pTyr are present in really low abundance and as a preliminary IMAC step might sometimes be added[182,210]. This technique is also more expensive as the cost of the antibody needs to be considered.

Those different enrichment technics (IMAC, MOAC, anti-pTyr antibodies) can be used under different formats: HPLC columns[211], cartridges[212], magnetic beads[213]. Phosphopeptides enrichment requires high amount of starting material (usually > 1 mg of peptides) as only a fraction of the original sample is kept. This might be especially crippling when studying phosphorylation on primary cell cultures, micro dissected cells or tissues samples[182]. To increase throughput and repeatability of phosphopeptides enrichment but also limit sample loss through manual handling, automatized platforms offer solutions. The EasyPhos protocol, in which digestion and enrichment are performed on

functionalized magnetic beads in a single container, was developed in this optic. It can be easily automatized as the steps are realized in 96 well plates. Results show that the protocol has high reproducibility and small sample size requirement (<200 µg of proteic starting material)[214,215]. Similarly, µPhos was developed as a phosphoproteomics platform allowing phosphopeptide enrichment on 24- to 96-well plates in a single-pot manner, minimizing sample loss and processing time while increasing sensitivity, specificity and reproducibility[216]. The Rapid-Robotic Phosphoproteomics (R2-P2) platform has also been developed for automation of streamlined phosphopeptide enrichment. The R2-P2 platform enables the identification of more than 4000 phosphopeptides from 2.5 µg of yeast protein[217]. Phosphopeptide enrichment protocols using commercialized TiO_2 or NTA- Fe^{3+} cartridges can be also used on AssayMAP Bravo platform (Agilent Technologies)[118,212,218]. Automatized IMAC protocol on AssayMAP Bravo enables also to greatly decrease the amount of sample starting material as it displayed promising results even with only 1 µg of HeLa cells[219]. This automatized protocol offers promising results on various other samples such as rat neuron lysate[219], malignant melanoma tissues[212], bacteria[118], human cancer cell line[220] or Formalin-Fixed Paraffin-Embedded (FFPE) tissues[221].

ii. Mass spectrometry analysis of phosphorylated peptides

Two main issues are encountered when analyzing phosphorylated peptides by mass spectrometry:

- The impact of a phosphorylation on a peptide upon ionization is widely discussed. It has been shown to affect the ionization efficiency of phosphorylated peptides compared to their non-phosphorylated counterpart[222]. However, other studies on synthetic phosphopeptides suggest that the issue of ionization efficiency of phosphopeptides versus their unmodified counterparts is not as straightforward as suspected as no clear evidence for decreased ionization was found[223].
- The labile character of the phosphoester bond leads to the potential loss of phosphate group during fragmentation. Indeed, as the phosphate bond is very labile, it tends to break first, generating a 98 Da loss of phosphoric acid (also called neutral loss). It is especially the case when using CID fragmentation and can therefore drastically impair the fragmentation behavior when compared to non-modified peptides. The main risk is obtaining MS/MS spectra dominated by this neutral loss signal, resulting in poor quality for identification[7,224]. Two different mechanisms are possible for this neutral loss reaction; they are both described in **Figure 28**.

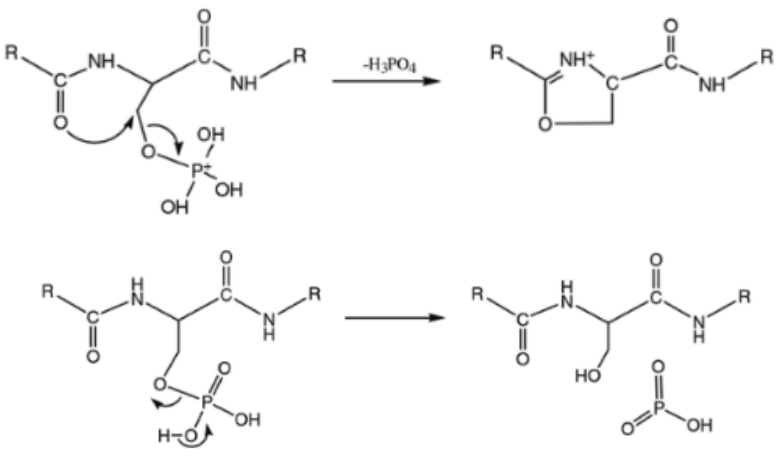


Figure 28: Mechanisms of H_3PO_4 neutral loss (A) via formation of a five-membered oxazoline (B) under mobile proton conditions, via SN2 mechanism[224].

As HCD generates more energy during fragmentation than CID, more informative spectra with less neutral loss are observed[7,186]. Other fragmentation methods were tested for phosphopeptides study such as ETD[136], EThCD[139] or UVPD[225]. As described in previously, ETD generates mostly c- and z- ions. ETD allows to preserve the labile phosphate group and therefore facilitate phosphopeptides identification. However, the technic suffers from a very long scanning speed due to the long reaction time between the anion that bear the electron and the peptide cation. Moreover, its efficiency of fragmentation is strongly dependent of the ion charge state, thus doubly charged peptides generally suffers from a lowest fragmentation efficiency with ETD than CID/HCD[186,224]. For these reasons, even if ETD spectra are more informative than CID's (see Figure 29), they are generally used as complementary information to CID/HCD, which remains usually the method of choice to analyze phosphopeptides[200].

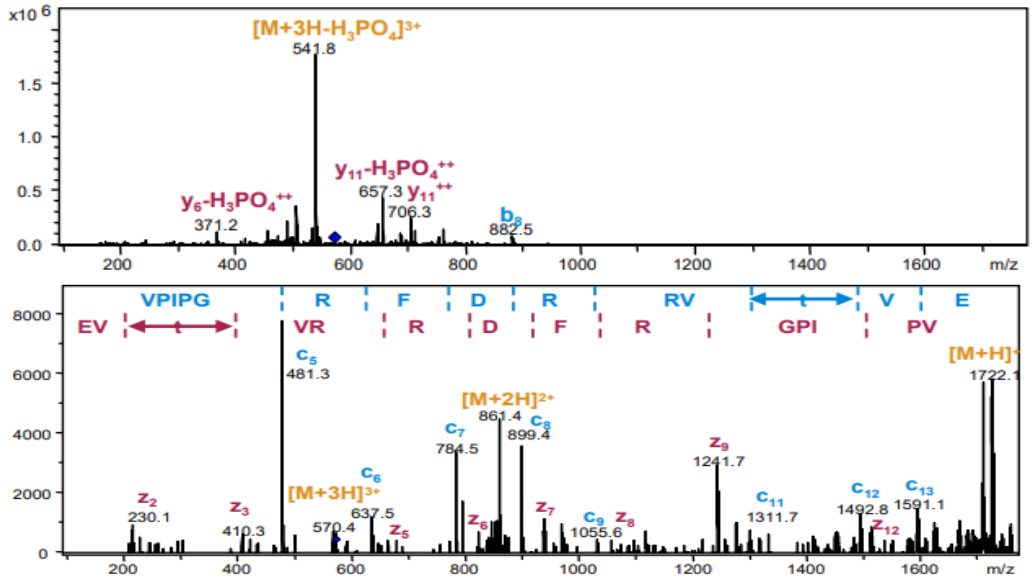


Figure 29: MS/MS spectra of VPIPGRFDRRVtVE phosphopeptide by (A) CID and (B) ETD. The most intense peak in CID spectrum corresponds to - 98 Da neutral loss of phosphate group. On ETD spectrum, mass differences between c11 and c12, z2 and z3 is equal to 181 Da corresponding to a threonine phosphorylation[226].

The combination of ETD and HCD into EThcD allows the combination of the fragmentation speed of HCD and the fragmentation efficiency of ETD. Peptides are thus fragmented twice, generating both y- and b- ions from HCD and z- and c- ions form ETD, as shown in **Figure 30**. This hybrid mode of fragmentation has shown promising results on phosphorylated peptides, as it allowed the identification of 3942 phosphosites (with a 99% localization probability at least) compared to 2002 and 4291 respectively for ETD and HCD alone. However, 95% of all EThcD phosphosites were identified with a localization probability of at least 99% whereas only 89% for HCD. Therefore, even less phosphosites are identified with EThcD, the sequence coverage is higher than with HCD[139].

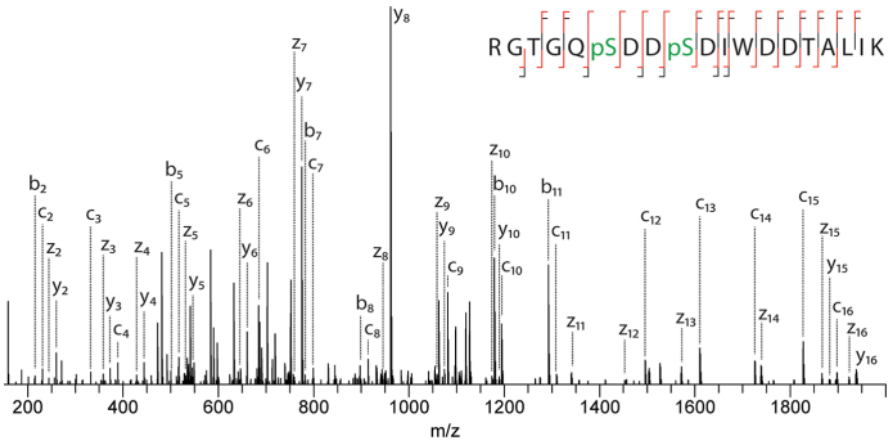


Figure 30: EThcD MS/MS spectrum of a doubly phosphorylated peptide. RGTGQsDDsDIWDDTALIK is doubly phosphorylated and contains in total four potential phosphorylation sites. EThcD generates dual ion series that enable phosphorylation site localization with very high confidence (phosphoRS site probabilities: T(3), 0.0%; S(6), 100.0%; S(9), 100.0%; T(15), 0.0%)[139].

In UVPD fragmentation, a high flow of protons is sent through pulsed laser beam during a short amount of time (< 1 µs)[138]. This process generates complex spectra with a-, b-, c-, x-, y- and z- ions, as shown in **Figure 31**. Important parameters affecting the fragmentation efficiency and the fragment distribution are the number of impulsions, the energy of the laser, and the wavelength[138,225]. All mechanisms involved in this fragmentation method have not been yet fully comprehended. While UVPD has shown to identify less phosphopeptides than HCD on HeLa cell lysate, the unique identifications of UVPD make this technique complementary to HCD for phosphorylation analysis[225].

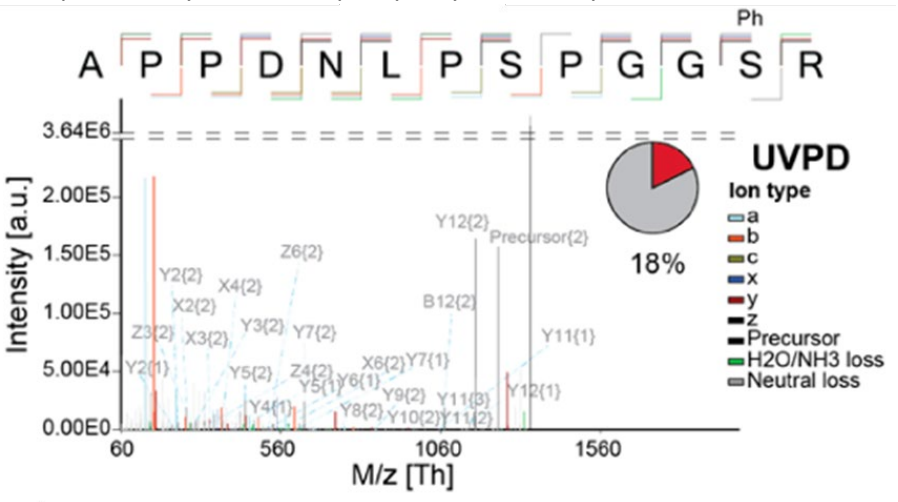


Figure 31: UVPD-generated fragmentation spectrum of APPDNLPSPGGsR phosphopeptide. The modification losses are encoded as {1} H_3PO_4 +H_2O, {2} H_3PO_4, and {3} full modification or HPO_3[138].

iii. Quantification strategies of phosphorylated peptides

Quantitative phosphoproteomics analysis can be performed using two different approaches: either with a label-free strategy[7] or by labeling, mostly using Tandem Mass Tag (TMT)[227,228]. Labeling techniques such as isobaric Tag for Relative and Absolute Quantification (iTRAQ)[229] or Stable Isotopic Labeling by Amino acids in Cell culture (SILAC)[230] have also been applied to phosphoproteomics but are not predominant.

In classical proteomics label free approach, a protein is usually identified with multiple peptides. However, in phosphoproteomics, quantification is performed at the peptide level, meaning only one

peptide is used for the quantification of a phosphorylation event. For one protein, quantitative values of different phosphosites of this protein might be different. It is thus necessary to detect the modified peptide in each analysis in order to identify and quantify the phosphorylation site. Because of the stochasticity of DDA label free method, the lability of the modification and the limited reproducibility of the phosphopeptide enrichment, the depth of the analysis can be relatively limited[197].

To overcome these issues, TMT labeling is more and more used for phosphoproteomics analysis[207,228,231,232]. TMT labeling decreases analysis time on instrument as samples are multiplexed while limiting missing values that might be numerous in label-free approach. However, it is much more expensive than label-free and limited in terms of sample number. Indeed, to date, several commercial TMT labeling kits are available, allowing to multiplex up to 18 samples[233]. Some in-house protocols have been developed to go further in the number of samples, allowing to analyze 21[234] or even 27 samples[235]. TMT stable isotopic labeling strategy is illustrated in **Figure 32**.

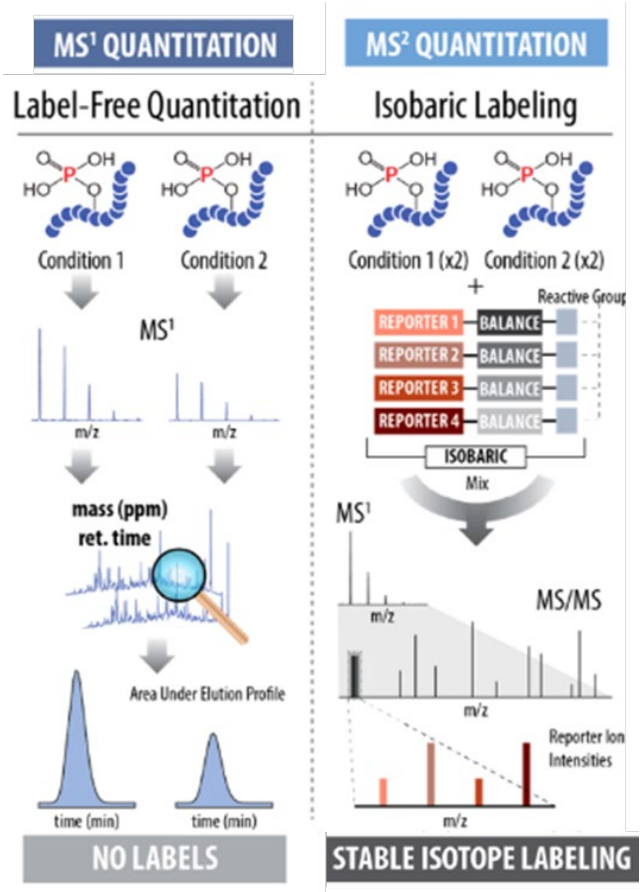


Figure 32: Quantitative strategies for global phosphoproteomics, adapted from Riley et al[197]. Label-free quantitation requires no additional steps in the phosphoproteomic workflow, and samples are analyzed individually. Quantitation is then performed across separate LC−MS/MS analyses using accurate mass and retention time windows to compare phosphopeptides from different samples. In contrast, stable isotope labeling methods permit multiplexing, where multiple samples can be mixed after labeling and then analyzed in the same LC−MS/MS analysis. Isobaric labeling uses a reactive tag that labels peptide functional groups, but quantitation is achieved at the MS2 level. The intact mass of each label is the same based on the coupling of reporter and balance regions that have an equivalent number of total heavy isotopes. Upon phosphopeptide dissociation, the reporter ions fragment off, allowing comparison of relative reporter ion intensities for quantitative measurements between samples, all within the same scan that provides phosphopeptide identification.

In a TMT labeling phosphoproteomics workflow, phosphopeptides can either be labeled before[236] or after[232] phosphopeptide enrichment. Labeling before phosphopeptide enrichment is highly expensive

as all the non-phosphorylated peptides that will mostly get lost during enrichment will also be labeled but shows reduced variability in sample preparation compared to labeling after enrichment[186]. However, Ogata *et al.*[237], when comparing TiO_2 enrichment of phosphopeptides either before or after TMT-labeling, showed that TMT-labeled phosphopeptides appear to have a tendency to flow through TiO_2 columns. They thus suggest that the labeling step should preferably be performed after phosphopeptide enrichment. Once labeled, samples are pooled and analyzed by LC-MS/MS. As the labeled peptides have identical chemical features and only differ by their mass, they will co-elute in the LC and enter the mass spectrometer together. However, during fragmentation, each marker will generate a unique reporter ion that will be then detected by MS/MS. Peak heights provide relative quantification of all the different samples. Peptides identification is performed on fragmentation spectra and quantification on reporter ions in MS/MS spectra also[238].

Various studies comparing label-approaches to TMT labeling phosphoproteomics analysis have been performed. On ovarian cancer tissues, TMT method displayed the highest precision and robustness in phosphosites quantification while label-free quantification offered the highest number of identifications[236]. Similar conclusions were obtained on another study conducted on MOAC enriched DiFi cells lysate[239].

To improve identification and quantification of phosphopeptides, and especially co-eluted phosphopeptides, an additional separation dimension can be added, namely ion mobility. The two most used technologies are TIMS[133,240,241] and FAIMS[242–244]. On FFPE mantle cell lymphoma samples, enriched by Fe^{3+} IMAC, more than 7000 class I phosphosites (meaning with a localization probability greater than 0.75) were quantified using a TimsTOF Pro[245]. Using FAIMS technology, around 15-20% additional phosphorylation sites were identified compared to same experiment performed without it[244]. As phosphorylation has shown to have a great impact on the Collision Cross Section (CCS) of peptides ions, an algorithm called TIMScore was developed by Brukerto predict CCS values of tryptic and phosphorylated peptides. It allows to increase phosphopeptides identification by 10 to 25% compared to analysis realized without TIMScore[246]. Another label-free approach to increase identification and quantification results in phosphoproteomics is Data Independent Acquisition (DIA)[247]. Detailed principle of this acquisition mode and its application to phosphoproteomics are detailed in **Chapter 3: Data Independent Acquisition – 4.DIA for phosphoproteomics**.

3. Bioinformatics tools for phosphoproteomics

For phosphoproteomics, or more generally PTMs study, a step is added in the proteomics pipeline, after data search identification and protein/peptide validation. This additional step consists of assigning the modification site localization and is followed by biological interpretation (**Figure 33**).

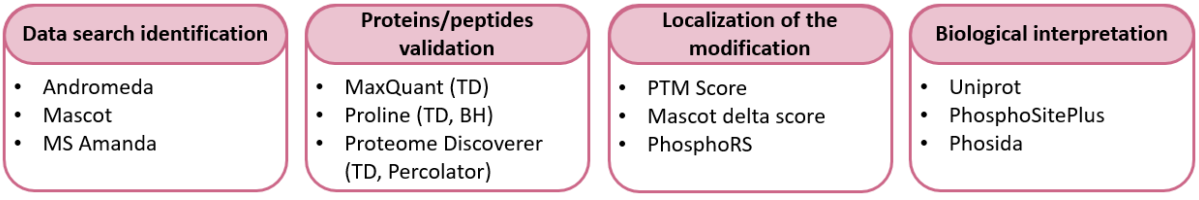


Figure 33: Representation of the main data treatment steps for phosphoproteomics analysis**.**

i. Identification and localization of phosphorylation

For identification of modified peptides, most search algorithms such as Andromeda[141], Mascot[142], or MS Amanda[143], allow variable modification search by adding a mass delta. However, if too many

modifications are searched at the same time, it exponentially increases search space and time. OMS software such as MSFragger[149] or IonBot[16] offer promising results to partially overcome this issue.

Correct phosphorylation site assignment is a crucial aspect of phosphoproteomics analysis, as its determination allows for functional characterization of the modifications observed[197,248]. However, in a typical phosphoproteomics experiment, from 20 to 40% of identified phosphopeptides are lost due to ambiguous modification localization[197]. Accurate phosphosite localization might be hindered by neutral losses, as fragments generated from a neutral loss and unmodified fragment have the same mass and are thus undifferentiable. Moreover, phosphorylation can be bared by various residues (Ser, Thr, and Tyr) that can be present in the same peptide, and thus phosphosites localization might need complete peptide sequence coverage to be accurate[224]. Once the phosphorylation is localized and attributed to its corresponding peptide, a score to evaluate the probability of presence of the modification needs to be calculated. To do so, several algorithms have been developed, based on: (i) computing the probability of an incorrect match for each phosphopeptide isoforms (phosphorylations placed on the different amino acids potentially phosphorylated in the peptide) or (ii) computing the difference between score of the different peptide isoforms[224]. PTM Score (implemented in Andromeda search engine when using MaxQuant)[141] or PhosphoRS (also called ptm-RS, implemented in MS Amanda, Mascot or Sequest search engines when using Proteome Discoverer)[249] algorithms are based on the first method whereas Mascot delta score[250] is based the second one[224]. The two most commonly used of them are PTM Score and PhosphoRS[224], and their overall operating workflow are presented in **Figure 34**.

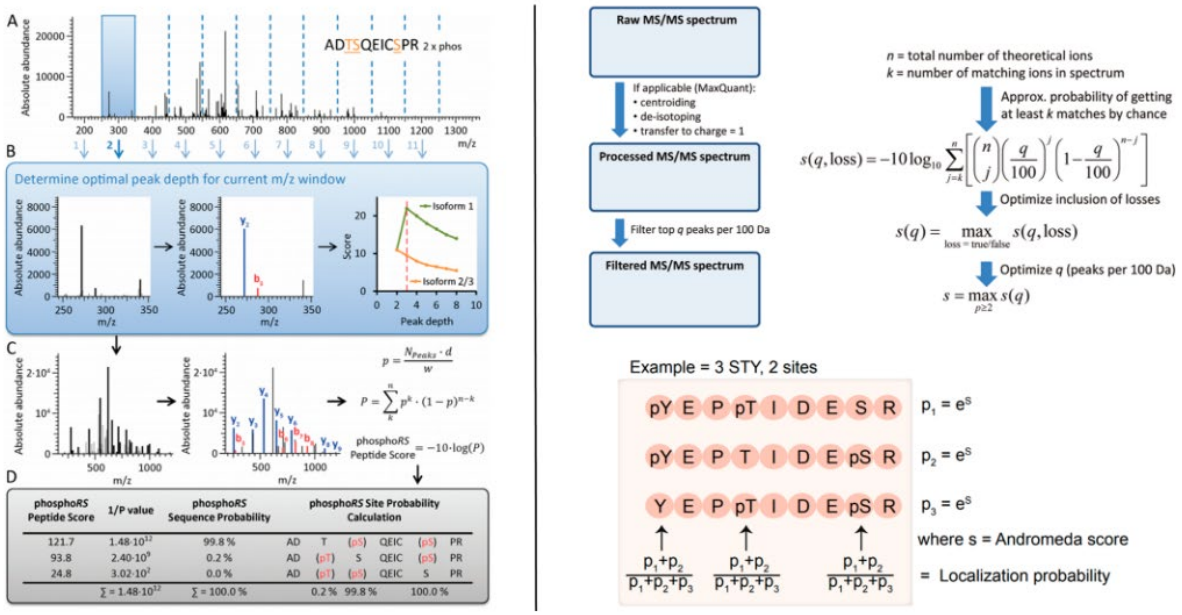


Figure 34: Workflows for the calculation of phosphosite localization probabilities by phosphoRS (left), and MaxQuant/Andromeda (PTM-score, right)[141,224,249].

Many studies have compared the results obtained by the different algorithms[224,251,252] and few explanations have been proposed to explain the observed differences : (i) fragment selection, (ii) noise thresholding, (iii) neutral loss consideration, (iv) the peak depth (how many of the most intense peaks per m/z are considered for score calculation) and (v) whether the algorithm was developed for low or high mass resolution and accuracy measurments[224,251]. Taking neutral loss into account is non editable in PTM Score but can be changed in PhosphoRS[224]. Comparing 22 phosphoproteomics pipelines for peptide identification and site localization, Locard *et al*. showed that only 50% of identified peptides were common between the Andromeda-PTM Score and Mascot-PhosphoRS. The main source of differences came from peptides identification but scoring played also an important role[251].

ii. Validation of phosphosites and False Localization Rate (FLR) estimation

In most phosphoproteomics studies, phosphosite validation is realized by applying an arbitrary filter of 75% on the localization probability: we are talking about class I phosphosites. Most of the algorithms score the confidence of the localization but do not estimated the False Localization Rate (FLR), an estimation of the number of wrong localization, which is similar but less straightforward than the FDR. Indeed, for an identified phosphopeptide with a given sequence, an incorrect localization phosphosite is not a random match as many fragments will match both correct and incorrect localizations, so decoy sequences do not provide an error estimation[253]. There are however a few attempts at finding an alternative such as the LuciPHor algorithm, a modified target-decoy approach, where decoy phosphopeptides are generated by adding an artificial phosphorylation on each non-candidate residues of identified phosphopeptide[254]. As the number of decoy sites is therefore usually higher than the number of target sites, this approach is considered more conservative[224]. The Site Localization In Peptide (SLIP) method performs the search with decoy phosphorylation on glutamic acid (E) and proline (P) residues allowed[255]. As the combined score of both E and P residues matched approximately the one of threonine (around 16%), and that E and P residues are present in some of the most prevailing phosphorylation motifs, they ensure the presence of decoy sites in the close environment of the real phosphosites[224]. A method described by Ramsbottom *et al*. has explored the possibility of using amino-acids as decoy, with alanine and leucine displaying the best results for reliable FLR estimation[256]. More recently, Zong *et al.* [245] proposed a deep learning based approach to control FLR in phosphoproteomics. DeepFLR combines deep learning-based phosphopeptides MS/MS spectra prediction to a target-decoy approach for FLR control. On both synthetic and biological phosphopeptides datasets from various LC-MS/MS platforms, DeepFLR demonstrated accurate FLR estimation and led to additional phosphopeptides identifications.

To date, all those alternatives are not unanimously approved among the scientific community and the only FLR evaluation accepted for now is on synthetic phosphopeptide libraries, as the correct phosphorylation sites are known, but do not reflect the complexity of true biological samples[224].

iii. Biological interpretation

To extract the functional and biological role of phosphorylation from the data, various bioinformatics solutions have been developed. PaDua uses Python language[257] while PhosR is based on R language[258]. Both packages are open source but only source code of PaDua is freely available. Scop3P is a freely accessible resource which allows to visualize the tridimensional structure of the protein but also give access to a variety of other information[259].

Phosphoproteomic data biological interpretation relies on databases that list all known phosphorylation sites, information on associated kinases and phosphatases, and the function affected by the phosphorylation of a given residue. PhosphositePlus[260] is the golden standard in this area, with more than 300 000 described phosphorylation sites, that are well curated and regularly updated. Other resources exists such as Phosida[261] , HRPD, Swiss-Prot or PhosphoELM. Both HRDP and Swiss-Port are general databases for all proteins while PhosphoSitePlus and PhosphoELM contains specific phosphorylation site information. Unfortunately, both HRPD and PhosphoELM were not updated since 2010. Many other phosphorylation sites exist, and their review can be found in literature[262]. As every database, they have their shortcomings, and are missing information as more than 95% of reported human phosphosites have no known kinase that regulates them[4]. For comprehensive information on kinase and phosphatase signaling, different knowledge bases were developed[262].

Chapter 3: Data Independent Acquisition

Ideally, proteomics would be able to quantify the vast majority of proteins in large sample cohort. Realistically, DDA-based approached successfully quantity thousands of proteins but in a restricted number of samples, and display limited sensitivity and restricted dynamic range. On the other hand, targeted approaches allows the quantification on only a limited number of proteins but in a multitude of samples and with increased sensitivity, specificity and dynamic range. Data-Independent Acquisition (DIA) strategies offers a promising combination of those two approaches, with the quantification of a large number of proteins, high sensitivity, specificity and accuracy, and a wide dynamic range[2,263]. Although the first DIA-based concept was reported in the early 2000 by Venable *et al.*[2,264,265], instrument improvements over the past decade, especially their resolution and speed of acquisition, have been the main reason for the overgrowing interest for DIA proteomics analysis. Moreover, the rapid development of DIA data-dedicated algorithms has also largely contributed to this increased enthusiasm for DIA[2,265] (**Figure 35**).

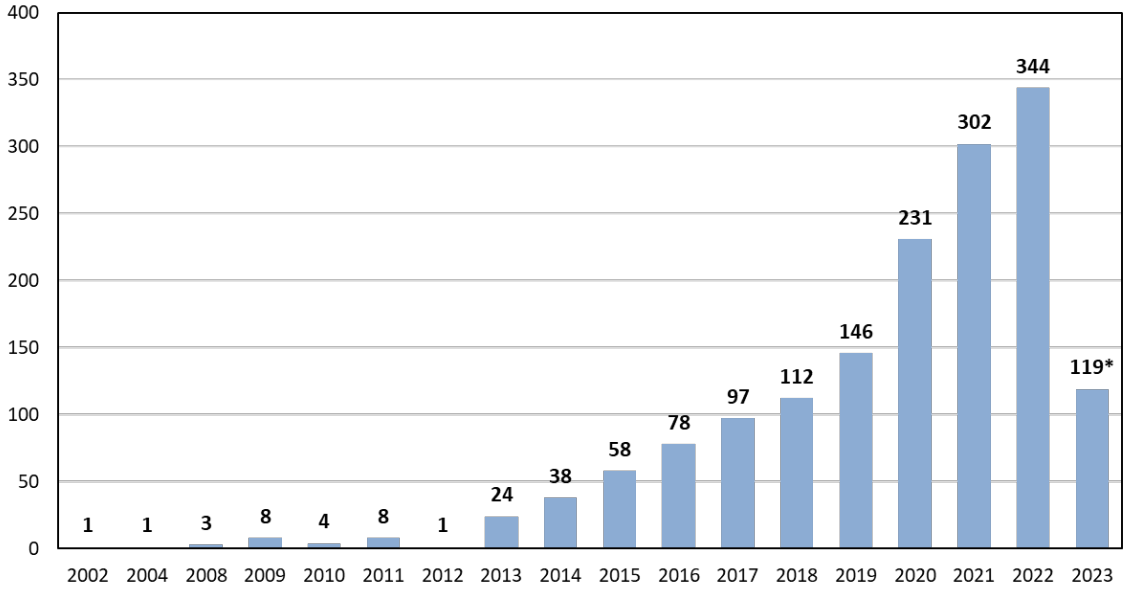


Figure 35: Number of publications whose abstract contains the term "data-independent acquisition" in PubMed. *For 2023, the number of publications was collected on the 11th of April 2023.

1. Principle and assay development of data independent acquisition

Compared to DDA mode, it is not the N most intense ions that fragmented but rather all precursor ions on a defined mass range (m/z isolation window). The acquisition of the fragment from all co-isolated and co-fragmented peptides results in multiplexed MS/MS spectra and chromatographic peak are extracted for all detected fragment ions to perform their quantification (**Figure 36**). DIA mode allows the collection of MS/MS spectra of peptides along the chromatographic gradient in an untargeted and unbiased manner, getting rid of the stochasticity and under-sampling issues of DDA[265].

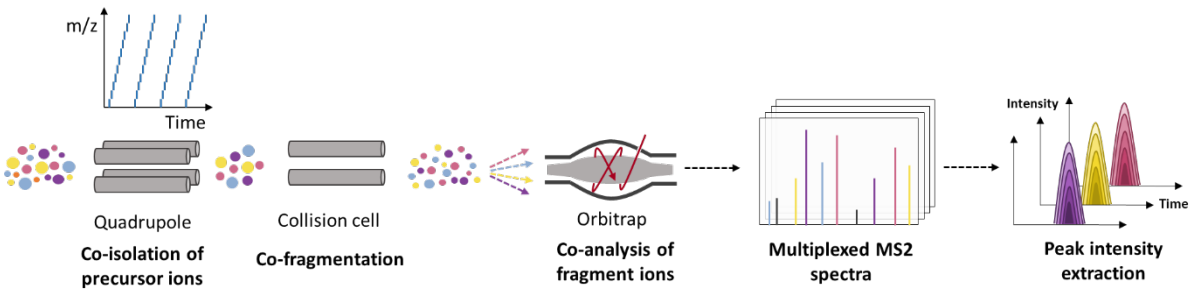


Figure 36: Schematic representation of Data-Independent Acquisition mode based on isolation windows on an Orbitrap instrument.

2. Evolution of DIA based strategies

The first proof-of-principle of DIA appeared in 2000 by Masselon *et al.*, in an experiment in which several polypeptides were characterized from multiplexed MS/MS spectra generated on a Fourier transform ion cyclotron resonance mass spectrometer (FT-ICR). They pointed out that the gain in throughput and sensibility can only be achieved thanks to highly accurate mass measurements[288].

DIA Method	Year	Isolation windows	Reference
Shotgun CID	2003	Full range	Purvine *et al.*[266]
DIA	2004	10 m/z	Venable *et al.*[264]
MS^E	2006	Full range	Silva *et al.*[267]
PAcIFIC	2009	2.5 m/z	Panchaud *et al.*[268]
AIF	2010	Full range	Geiger *et al.*[269]
XDIA	2010	20 m/z	Carvalho *et al.*[270]
FT-ARM	2012	100 m/z	Weisbrod *et al.*[271]
SWATH	2012	25 m/z	Gillet *et al.*[272]
HDMS^E	2012	Full range	Geromanos *et al.*[273]
MSX	2013	4 m/z	Egertson *et al.*[274]
WiSIM	2014	200 m/z and 20 m/z	Zabrouskov *et al.* [275]
pSMART	2014	5 – 20 m/z	Prakash *et al.*[276]
UDMS^E	2014	Full range	Distler *et al.*[277]
SWATH (variable windows)	2015	8-85 m/z variable	Zhang *et al.*[278]
HRM	2015	24 – 220 m/z variable	Bruderer *et al.*[279]
SONAR	2018	24 m/z	Moseley et *al.*[280]
BoxCar DIA	2018	-	Meier *et al.*[281]
DIA-FAIMS	2020	13.7 m/z	Bekker-Jensen *et al.*[282]
dia-PASEF	2020	25 m/z	Meier *et al.*[11]
DDIA	2020	12 m/z	Guan *et al.*[283]
Scanning SWATH	2021	5 m/z	Messner *et al.*[284]
PulseDIA	2021	variable	Cai *et al.*[285]
BoxCarMax	2021	22 m/z and 2.5 m/z	Salovoska *et al.*[286]
DaDIA	2022	25 m/z and 100 m/z	Guo *et al.*[287]

Table 3: Non exhaustive list of the evolution of DIA modes and associated parameters, adapted from Kitata *and al.*[2].

In 2003, a first evolution of the concept was proposed by Purvine *et al.*, under the name of shotgun CID. Here, a first analysis was performed at low source voltage to limit fragmentation and obtain MS spectra of precursor ions; and then a second one at higher source voltage to induce fragmentation of peptide and generate MS spectra of fragment ions. The term data-independent acquisition was introduced in 2004 by Venable *et al.*, when they proposed a new alternative strategy to sequentially isolate and fragment precursor ions in 10 m/z isolation windows using a linear ion trap mass spectrometer[264]. Since then, variations and improvements of the method were continuously developed, as shown in **Table 3.**

They can be divided in two categories: the approaches working on the complete mass range and the ones using isolation windows.

In the work presented in this manuscript, only methods based on isolation windows were used and therefore detailed here. However, a general overview of the principle of all DIA based strategies is presented in **Figure 37**.

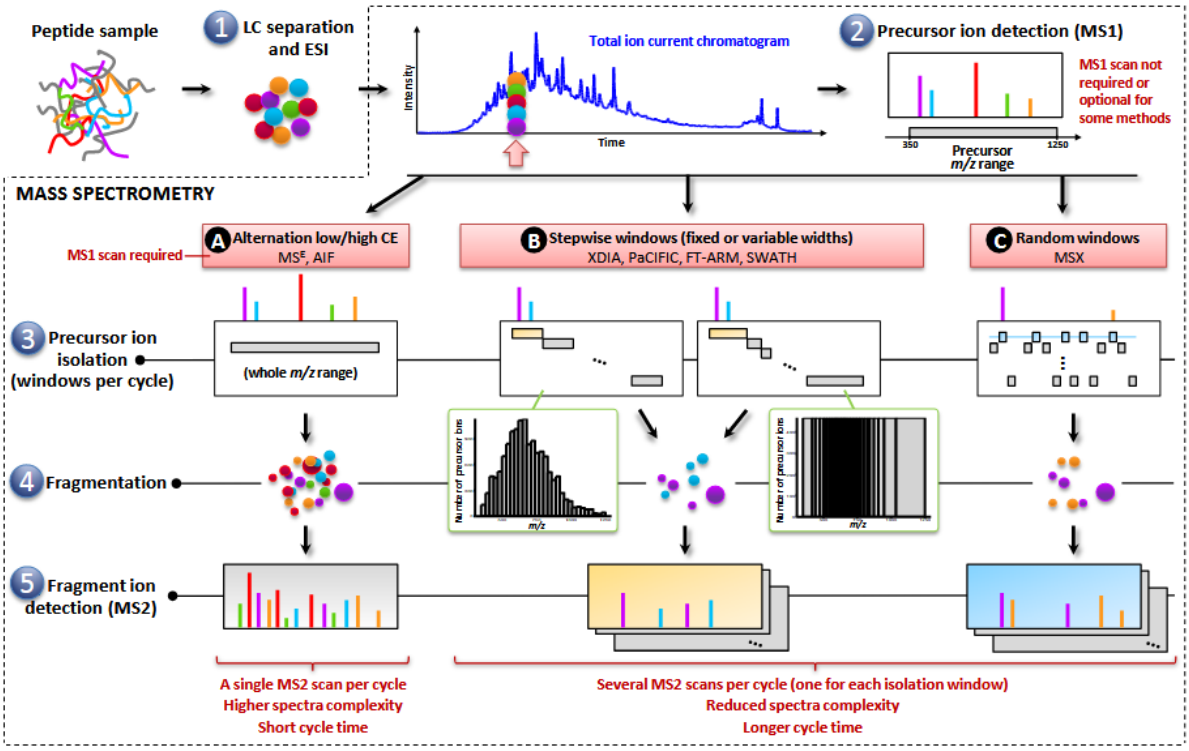


Figure 37: Data-Independent Acquisition schemes in bottom-up proteomics, from Bilbao *et al.*[289]. (A) Full MS range-based DIA strategies (B) Isolation window-based DIA based strategies, with either fixed or variable windows (C) Multiplexed DIA strategy with randomly chosen isolation windows.

i. DIA based strategies based on isolation windows

a. Consecutive fixed windows

Venable *et al.*[264] were the first to propose the use of isolation windows for fragmentation, many have followed afterwards. The Precursor Acquisition Independent From Ion Count (PAcIFIC) approach, proposed by Panchaud *et al.*, is based on the use of narrow isolation windows (2.5 m/z) to reduce MS/MS spectra complexity. Unfortunately, it requires multiple injection of samples over several days to cover the entire mass range[268]. They improved this method a few years later thanks to instrumental developments[290]. A similar method was developed more recently by Cai *et al*. called PulseDIA, which

aims to reduce the number of isolation windows and thus the number of injections by using window size adapted to the ion density[285].

A variation of this approach was introduced in 2010 by Carvalho *et al.*, called eXtended Data-Independent Acquisition (XDIA). It acquires an additional high resolution MS spectrum at the beginning of each cycle, followed by a combined CID and ETD fragmentation[270].

The Fourier Transform-All Reaction Monitoring (FT-ARM) strategy was presented by Weisbrod *et al*. in 2012, and uses fixed windows of 12 m/z or 100 m/z on LTQ-Ft or LTQ-Orbitrap instruments[271]. The same year a similar strategy was developed by Gillet *et al.* on a Q-TOF instrument and using 25 m/z windows[272]. This method, namely Sequential Windowed Acquisition of All THeoritical fragment ion (SWATH), is now commercialized by Sciex.

Two similar approaches relies on the parallelization capacity of Q-Exactive instruments by acquiring high resolution MS spectra independently from MS/MS spectra that are obtained after isolation of precursors in restricted mass windows. MS spectra are used for quantification while MS/MS spectra for identification. These approaches, called pSMART and Wide Selected-Ion Monitoring (WiSIM), were developed respectively by Prakash *et al.*[276] and Zabrouskov *et al.*[275] .

Additionally, the BoxCar method has emerged and was first implemented in 2018 by Meier *et al*.[281]. It relies on MS acquisition of narrow mass windows to increase the dynamic range of MS1 signals. This approach combined to DIA is promising as it results in much better precursor information. Recently, an optimized version of this method, BoxCarMax, was developed for the analysis of complex samples obtained after SILAC and pulseSILAC labeling[286].

Recently, new DIA strategies have been developed thanks to ion mobility. DIA-FAIMS[282] has been implemented on latest generation Q-Orbitrap instruments while dia-PASEF[11] was applied to TimsTOF platforms. The dia-PASEF approach will be furthered described later in the manuscript.

A hybrid method combining DDA and DIA was developed by Guan *et al*. in 2020[283]. The cycle is divided in three phases. First, precursor ions are analyzed generating MS spectra. Then, as in a DDA method, the top N most intense ions are selected and sequentially fragmented. Finally, a multiplexed MS/MS spectrum is acquired, resulting from the co-fragmentation of ions in isolation windows covering the entire mass range. The method showed promising results as, compared to a classical DIA method, it was able to identify a similar number of peptide but almost twice as many protein groups, while requiring much less sophisticated data treatment. Another hybrid approach combining both DDA and DIA has been proposed recently by Guo *et al.* called DaDIA[287], in which individual biological samples are analyzed in DIA, whereas the pooled quality control (QC) samples are analyzed by DDA. DDA and DIA data are then integrated altogether using a dedicated algorithm.

b. Consecutive variable width windows

The peptide distribution across the m/z range, time dimension and ion mobility range, if present, is not homogeneous. A possibility would be to reduce the size of the windows to reduce the complexity of the MS/MS spectra but this will inherently increase the cycle time needed to cover the whole mass range and thus reduce proteome coverage. For this reason, using different size of isolation windows based on ion density would allow the ion population to be more evenly distributed. Large windows are used in sparse regions while smaller windows are used in in dense regions to reduce spectra complexity while limiting the impact on the cycle time. In this optic, a SWATH approach with variable windows was proposed by Zhang *et al*. in 2015, an evolution of the original method with fixed size windows[278]. Corresponding software were developed, such as swathTUNER, to optimize the isolation window size to the precursor ion density in an automated way. This strategy is implemented as SWATH 2.0 by Sciex.

A similar approach, Hyper Reaction Monitoring (HRM) was implemented in 2015 by Bruderer *et al*. on Q-Orbitrap instruments[279] and is now owned by Biognosys company. In this method, the variable windows need to be generated manually.

c. *Multiplexed strategies*

A strategy based on the use of sequentially co-isolating the precursor ions contained in randomly selected isolation windows was proposed by Egertson *et al*. in 2013 called MSX[274]. The mass range 500-900 m/z is divided into 100 windows of 4 m/z each. MS/MS spectra are then computationally demultiplexed to increase selectivity and signal-to-noise ratio. This increase is particularly adapted for the analysis of modified peptides, as they might be difficult to distinguish if they are isolated in the same window due to their similar fragmentation. This approach is based on the multiplexing capabilities of Q-Orbitrap instruments but it may suffer from a loss of sensitivity due to the limited time for ion trapping.

d. *Overlapping windows*

Another approach relies on overlapping isolation windows: the isolation scheme is composed of windows that cover half the masse range covered by the previous window, resulting in increased selectivity. Information from overlapping MS/MS spectra are used to generate demultiplexed MS/MS spectra. In this optic, the SONAR approach was developed on a Q-TOF instrument, were MS/MS spectra are continuously acquired over the 400-900 m/z range with 24 m/z windows[280]. Similarly and more recently, the scanning SWATH strategy was proposed by Messner *et al.* for restricted chromatographic gradients[284]. Cycle time is decreased compared to classical DIA methods as the successive window acquisition is replaced by a continuous scanning with the first quadrupole.

e. *Dia-PASEF approach*

The dia-PASEF approach is a more recent DIA acquisition method which is specific to the TimsTOF Pro and uses to its advantage the PASEF technology described earlier[11]. PASEF is beneficial to DIA as it allows for increased speed of acquisition, noise reduction, improvement of the signal thanks to the ion accumulation, and better separation of co-eluting peptides thanks to the ion mobility dimension. It also allows the use of up to 100% of the ions entering the mass spectrometer and the elimination of the interferences linked to mono-charged ions. The principle of dia-PASEF on a TimsTOF Pro is detailed in **Figure 38**.

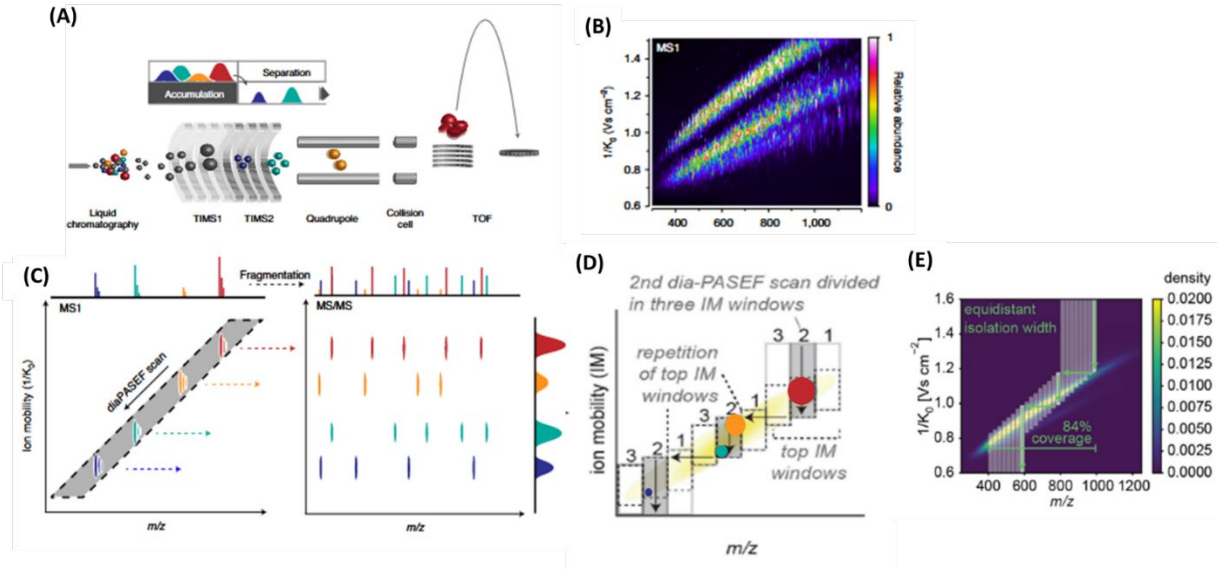


Figure 38: The dia-PASEF acquisition method, adapted from Meier *et al*.[11] and Skowronek *et al.*[241]. (A) Schematic ion path of the TimsTOF Pro mass spectrometer (B) Correlation of IM and m/z in a tryptic digest of HeLa cell lysate (C) In diaPASEF, the quadrupole isolation window (gray) is dynamically positioned as a function of ion mobility (arrow). In a single TIMS scan, ions from the selected mass ranges are fragmented to record ion mobility–resolved MS/MS spectra of all precursors (D) Representation of the dia-PASEF acquisition scheme depicting three dia-PASEF scans divided into three IM windows (E) Original dia-PASEF acquisition scheme

Briefly, ions enter the mass spectrometer and are accumulated in the first part of the IM cell. They are then separated according to their IM and sequentially eluted from the second part IM cell. As with dda-PASEF, the accumulation and separation/elution of the ions is performed simultaneously so that no ions are lost (100 % duty cycle), as shown in **Figure 38-(A)**. As for peptide ions of a given charge state, ion mobilities and masses are correlated (**Figure 38-(B)**), this feature can be used to isolate precursor mass windows for DIA without losing the ions outside the respective windows, unlike in other DIA acquisition scheme. When separating ions according to their ion mobility, those with the lower mobility (usually high m/z) are released first, the m/z window of the quadrupole start at high m/z values. As higher mobility ions (usually low m/z) are then released from the TIMS, the quadrupole mass isolation should slide down to lower m/z values in synchronization with the elution of ions (**Figure 38-(C)**), to fully transmit the ion cloud. This movement happens in distinct steps and thus divides one PASEF scan into multiple IM windows (**Figure 38-(D) and (E)**). The isolation windows for dia-PASEF are designed to cover the most part of the doubly and triply charged ions, as they fragment more easily and thus are more informative, and thus constitute a two dimension acquisition scheme (IM and m/z dimensions). The isolation windows are also located in the most precursors-dense regions, as illustrated in **Figure 38-(E)**. Once a m/z window is selected by the quadrupole, precursor ions are sent into the collision cell where collision energy is applied according to the 1/K0 range. This means that higher energies are applied for lower IM coefficient, *ie* for ions that are harder to fragment. The applied collision energy will slide toward lower energies in synchronization with the ions' elution from the IM cell. Finally, ions are sent to the TOF to measure their m/z and intensity .

The dia-PASEF strategy is growing popularity amongst the scientific community and various applications of it are emerging[291–295]. Among the most recent ones are Synchro-PASEF[295], Slice-PASEF[296] or midiaPASEF[297].

3. Different approaches to overcome the challenge of DIA data processing

In DIA, multiple peptides that are co-isolated and co-fragmented in the same precursors window generate highly convoluted fragmentation spectra, which can lead to the loss of a direct relationship between the precursors and its fragment ions, making the distinction of those multiple peptides complicated. DIA thus requires much more sophisticated data analysis post acquisition than DDA, and dedicated strategies have been developed as those generated multiplexed spectra cannot be analyzed by conventional DDA tools[263,265,289]. DIA data processing relies on two different approaches, represented in **Figure 39**: the peptide-centric approach and the spectrum-centric approach.

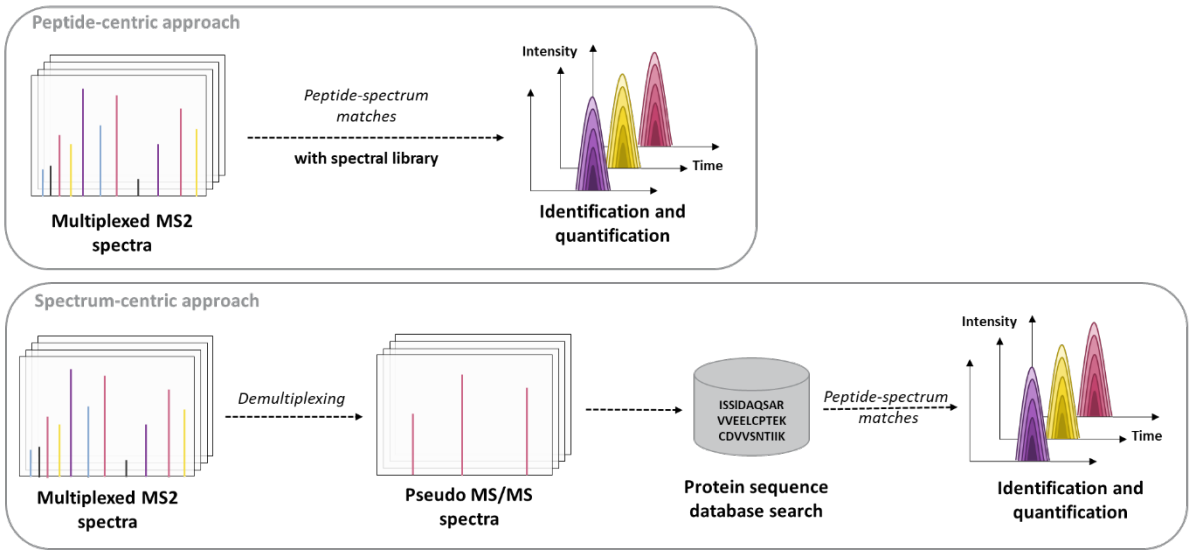


Figure 39: Schematic representation of the peptide-centric and spectrum-centric analysis, adapted from Ting *et al.*[298]. The peptide-centric approach uses previously generated spectral library to query for the presence of peptides in the DIA MS/MS spectra. The spectrum-centric approach identify peptides directly from DIA spectra. To do so, pseudo MS/MS DDA spectra are generated from DIA MS/MS spectra, and a classical database search is performed.

i. Peptide centric approach

The peptide-centric approach is based on the use of a previously generated spectral library, either performing targeted data extraction or by directly matching spectra.

a. Spectral libraries generation

A spectral library is defined as a condensation of MS/MS spectra confidently assigned to a specific peptide sequence[299]. A spectral library is usually composed of different features such as: precursor m/z, fragment ion m/z and relative intensity, standardized retention time (RT) for each peptide precursor, and sometimes ion mobility values. Peptides present in the library are usually in their dominant charge states, unique to a protein sequence, easily detectable *ie* with enough abundant ions[265].

Spectral libraries are usually generated from multiple fractionated DDA acquisitions of the same sample type of sample used for the DIA analysis. A search is then performed against a protein sequence database to identify peptides in the shotgun proteomics spectra, and the search results of multiple observations of the same peptides are collapsed together into a single entry to build a consensus

spectral library[265,300]. As only the peptides present in the generated library can be detected in the DIA analysis, it is utterly important to use a library as complete and qualitative as possible. Indeed, it assumes that precursors ions contained in the spectra are correctly identified, meaning false positives become true positives. Therefore, the FDR of the spectral library generation is generally controlled tightly at 1%[300]. This approach, which increases the coverage of the proteome and thus the search space thanks to fractionation, is however tedious, time and material consuming[300].

Another approach uses Gas Phase Fractionation (GPF) acquisition scheme for DIA library generation. GPF performs multiple analysis of a sample by mass spectrometry over multiple small m/z ranges[289]. Recently, a GPF scheme was implemented in combination with ion mobility on TimsTOF Pro. The study showed that the diaPASEF GPF generated library exceeded the performance of libraries generated directly from diaPASEF data[301].

Various software has been designed to build libraries from peptide spectrum matches results of DDA analysis. In SpectraST[302], one of the most used, a library can be built from replicate spectra identified to the same peptide ion and combined with quality filters to remove low-quality spectra from the library. Skyline builds a library using a set of tools from BiblioSpec, to assemble a redundant library which is then filtered to create a non-redundant library. The best spectrum within a group with the best score and highest total ion chromatograph (TIC) is chosen[303]. Spectronaut Pulsar might also be used to generate spectral libraries[265,304].

To avoid the generation of spectral libraries, which is both time- and sample-consuming, various platforms provide public spectral libraries for the extraction of DIA-SWATH data. These include Peptide Atlas[305,306], MassIVE[307], Proteomics IDEntification database (PRIDE)[308], or SWATHAtlas[309]. Unfortunately, these public libraries cover only a limited variety of organism's proteomes. For example, as of April 2023, only 16 organisms are represented by PRIDE's spectral libraries, and 18 in SWATHAtlas, six of which are human.

The use of a spectral library generated from data acquired on the same LC-MS coupling and conditions is the optimal method for DIA analysis. However, studies have shown that MS/MS spectra acquired on different instruments both using CID fragmentation can be compared and thus used for cross-instrument library generation, if the peptides' elution order is the same[265,310]. Recently, Multiple Characteristic Intensity Pattern (MCIP) approach was introduced to better take into account spectral variability when building spectral libraries[311].

b. Artificial intelligence-predicted libraries

Recently, deep learning and artificial intelligence growth has been reflected on proteomics data processing and especially with DIA data processing. Different tools have emerged to predict library assay with deep learning computational fragment ion prediction, reaching the same level of accuracy than with empirical methods. Prosit is a deep neural network, trained on ProteomeTools synthetic peptide library, that can learn and predict both chromatographic RT and fragmentation ion intensity of any peptide with extremely high quality[312]. It is implemented in ProteomicsDB[313], allowing custom *in-silico* spectral library generation. DeepMass:Prism[314] predicts peptide fragmentation pattern using a deep-learning based algorithm trained on millions of MS/MS spectra and generates experimentally equivalent mass spectra, used to create *in-silico* spectral library[265]. pDeep[315] is also a deep-learning based method, with extensive PTM support, predicting the intensity distribution of product ions of a peptide. It work well to predict not only HCD spectra but also ETD and EThcD spectra[265]. An important point to be raised, is that the intensities of the computationally predicted fragment are instrument dependent[265]. An other deep-learning based approach, DeepDIA[316] , aims at training instrument-

specific models for accurate MS/MS spectrum and RT prediction. This approach is shown to outperform both Prosit and pDeep approaches for in-silico spectral library generation[316,317].

DIA data processing software such as Spectronaut (Biognosys) or DIA-NN[3] also use artificial intelligence and machine learning to improve their data processing. Benchmarking studies comparing the different software and approaches are widely developing[318–320]. One of the current challenges in the field of deep-learning predicted libraries is the prediction of spectra of modified peptides[321], even if some methods are emerging, such as DeepPhospho[322], a deep learning network specifically designed to generate phosphoproteomics *in-silico* libraries.

c. *Targeted data extraction*

Targeted data extraction approach was first proposed in 2012 by Gillet *et al.*[272] to process SWATH data. All information in the spectral library are used to extract XICs from MS/MS spectra and interrogate the presence of peptides in the data. Potential elution peaks of each peptide from these fragment ion chromatograms are evaluated according to different criteria[263]:

- The shape of the chromatographic peak
- The peptide sequence and normalized RT
- The precursors ions' m/z and charge
- The correlation of the fragment ions over elution time and how well their relative intensities match their corresponding library spectrum

Similarly to DDA, statistical validation of peptide identifications is performed using a targeted strategy to assess the FDR. Several software are based on a targeted data extraction approach such as OpenSWATH[323], DIA-NN[3], PeakView (AbSciex), Skyline[303] and Spectronaut (Biognosys).

An incorrect peptide detection can be due to multiple factors: (i) the queried peptide is not in the sample (ii) it is in the sample but at an amount below the instrumental limit of detection (iii) the correct peak group results in a lower score than the wrong one. To overcome at least partially these issues, algorithms have been developed, such as Transfer of Identification Confidence (TRIC)[324], DIAlignR[325] or recently DeepRTAlign[326], to normalize retention time between analyses and limit the proportion of false identifications. Other software have been also developed to support the DIA data including ion mobility dimension namely Mobi-DIK[11], Spectronaut (Biognosys), or more recently DIA-NN.

d. *Direct spectrum matches*

Another way of processing DIA data is much similar to DDA search algorithms as it consists of directly scoring the match between the MS/MS spectrum and a theoretical assigned MS/MS spectrum contained in the spectral library[263]. It is thus an untargeted detection of peptide, as opposed to the previously described approach. These methods introduce additional heuristic filtering such as considering only observed spectra that match a threshold number of peaks in a theoretical spectrum.

The first developed software tool was ProbIDtree, which includes an algorithm that identifies, for each multiplexed MS/MS spectrum, all potential precursor ions contained in an isolation window and above a user-defined intensity threshold, using the corresponding MS spectrum[327]. From the list of potential precursor ions, it then calculates a probability score for the identification of each peptide and a peptide probability is constructed. At each iteration, a new DIA MS/MS spectrum is generated by removing the already matched fragments. Another software, Mixture-Spectrum Partitioning using Libraries of Identified Tandem mass spectra (MSPLIT-DIA)[328], deconvolutes the DIA MS/MS spectra by evaluating the similarities between them and MS/MS spectra from the spectral library. Spectra that are too similar are removed from the targeted extraction and the quality of the results is evaluated through RT score

and FDR statistical validation. On the other hand, the EncyclopeDIA[329] software uses the PEptide-Centric ANalysis (PECAN)[330] algorithm, which is based on spectral library generation from multiple injections of the sample in DIA using narrow isolation windows (4 m/z).

ii. Spectrum-centric approach

Spectrum-centric (also called "library-free") approaches rely on the generation by algorithms of multiple pseudo-MS/MS spectra from DIA MS/MS spectra, each containing the fragments of only a single peptide in the mixture. The intensities of different fragment of a same peptide should correlate over elution time, and it is this correlation that is used to assign fragment ions from MS/MS scans to their intact peptide specie in MS scans. The pseudo MS/MS spectra are then submitted to a classical DDA database search[263]. This approach was first reported by Purvine *et al*.[266] in 2003, where they constructed pseudo DDA spectra, from analyses performed at low and high voltages, and based on the similar chromatographic characteristics of precursor and fragment ion to identify them manually.

Since then, many algorithms have emerged to perform this task automatically. Among them, DIA-Umpire[331] software propose the untargeted identification and quantification of peptides. Moreover, it allows to generate a spectral library from the obtained identification results, further used to perform targeted extraction of peptides in the initial DIA data. More recently, thanks to the implementation of Pulsar search engine in Spectronaut software, a directDIA algorithm[10] has been implemented, allowing peptide-centric extraction of data. Similarly, the discovery mode of MaxDIA also allows the use of a library-free approach[332]. Finally, DIA-NN software allows performing DIA data extraction with both peptide- and spectra-centric approaches[3].

4. DIA for phosphoproteomics

In classical DDA approach, isobaric phosphopeptides are difficult to sample and assign, as they share the same mass, similar retention times, and many fragment ions. DIA offers promising solution to study these co-eluting phosphopeptides, as multiple precursor ions are fragmented in parallel to trace peptide fragment ions along their chromatographic gradient and uses fragment ions to perform quantification[247,333] (**Figure 40**).

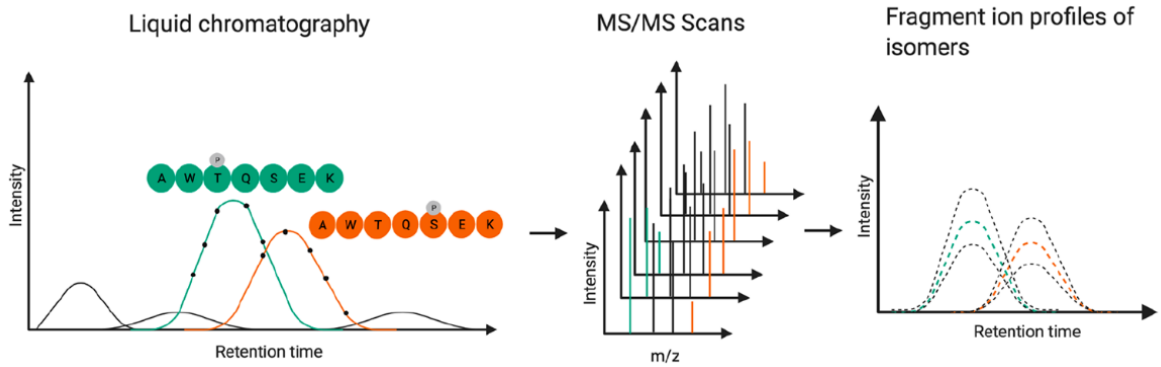


Figure 40: Data-independent acquisition mass spectrometry deterministically samples chromatographic peaks at multiple time points. Due to the reconstruction of fragment ion elution profiles enabled by this method, chromatographically co-eluting phosphopeptide isomers may be distinguished from each other by their site-specific fragment ions, from Srinivasan *et al*.[247].

Even if DIA method is not yet unanimously adopted to study phosphopeptides, some studies comparing DIA to DDA for phosphoproteomics are emerging. They mostly demonstrates that DIA allows better quantitative reproducibility and superior quantitation over a larger dynamic range for

phosphopeptides[10,247,334–336]. On fungus samples, the number of quantifiable peptides was 35% higher using DIA than DDA while the data completeness was also increased with DIA[334]. On cell lysate, using a DIA method allowed an almost 2 fold increase in the number of quantified phosphosites, while also highly increasing the percentage of phosphosites quantified with a coefficient of variation (CV) lower than 20%[336]. Comparing the two approaches on a dataset from human cell line doped with synthetic phosphopeptides, Srinivasan *et al*. have shown that DIA allows a more robust phosphopeptide identification, with 66% of phosphopeptides identified in 5 out of 10 replicates for only 32% in DDA[247]. Bekker-Jensen *et al*. used an optimized DIA phosphoproteomics protocol on HeLa cell lysate and identified approximately twice as many phosphopeptides over a larger intensity range compared to DDA, with superior quantitative reproducibility between technical replicates[10].

Taking advantage of the instrument developments, and especially ion mobility separation using dia-PASEF on a TimsTOF Pro, Skowronek *et al*. identified more than 35 000 class I phosphosites on stimulated HeLa cells, 20 000 of which were quantified in all replicates of at least one experimental condition[241]. Using also dia-PASEF technology, Oliinyk *et al*. quantified on 20 µg of human cancer cell line over 13 000 phosphosites without spectral library[133].

Chapter 4: Multi-omic approaches to disease

The addition of 'omic' to a molecular term implies a comprehensive or global assessment of a set of molecules. The application of the different individual omic has successfully allowed to better understand multiple cellular processes involved in many diseases[21,22,337,338]. These omics data are useful as marker for the disease progress and to give insight into biological pathways or processes are involved in it. However, each of these omic taken separately have not been yet able to fully comprehend the cause of disease but are rather reflecting of the reactive process arising from it. It is in this context that multi-omics studies have arisen for the past decade, as shown in **Figure 41**.

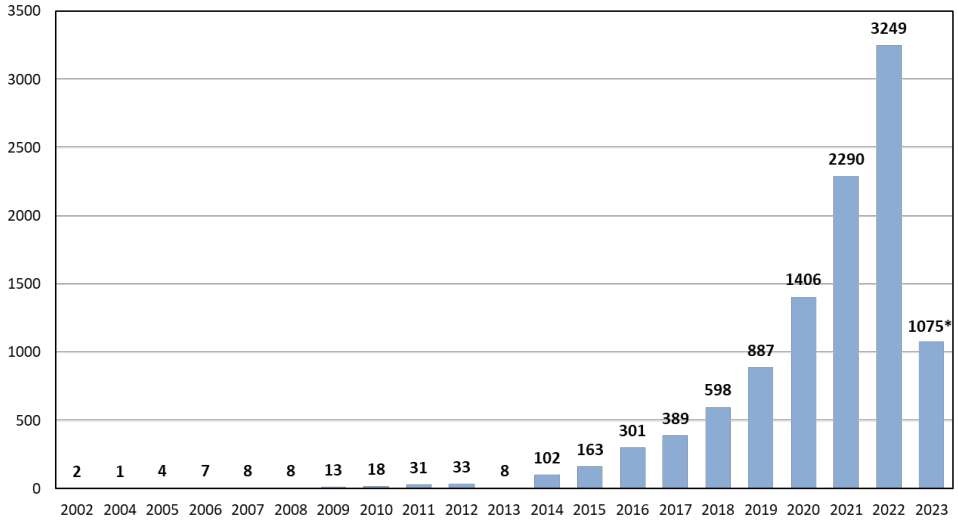


Figure 41: Number of publications whose abstract contains the term "multi-omics" in PubMed. *For 2023, the number of publications was collected on the 1st of April 2023.

Multi-omics, also called pan-omics, trans-omics, or vertical-omics, is defined as the use of at least three or more omic datasets coming from different layers of biological regulation. The integration of all multi-omics data have shown promising results to understand potential causative alterations leading to some disease or to potential biomarkers target that might lead to future treatments[8,339–343].

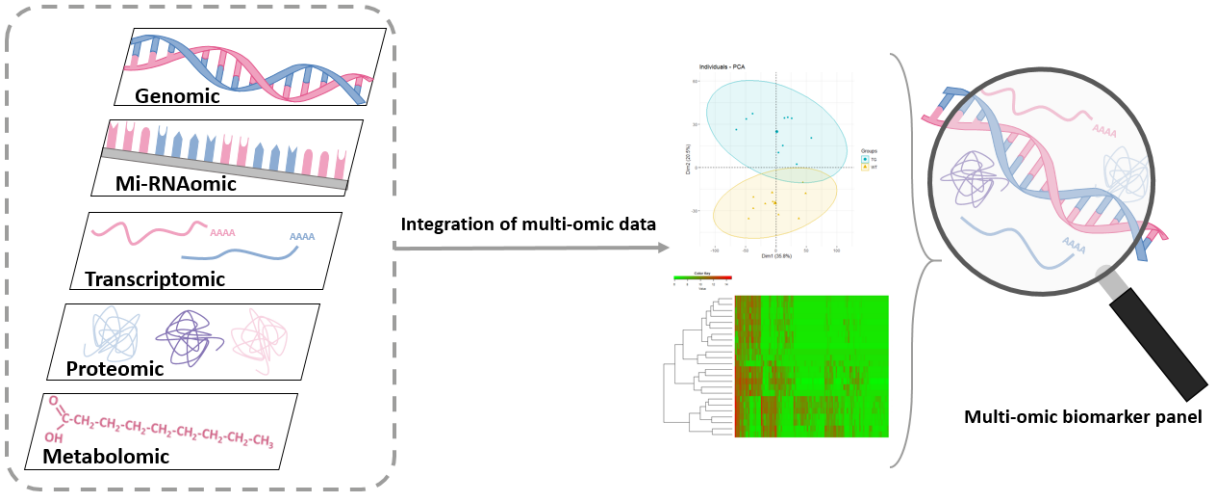


Figure 42: Non-exhaustive schematic representation of a multi-omics analysis.

1. The different omics data types

Some examples of the types of omics data that can be used in the context of a multi-omics project are discussed below. This list is not exhaustive as a comprehensive list of all omics type would not be appropriate as new omics technologies are rapidly emerging.

- Genomics : the genome is defined as the complete sequence of DNA in a cell or organism[344]. It was the first type of omics to emerge. In the field of medical research, genomics aims to identify genetic variants associated with disease, response to treatment, or future patient prognosis.
- Transcriptomics: the transcriptome is the complete set of RNA transcripts from DNA in a cell or tissue[344]. It includes ribosomal RNA (rRNA), messenger RNA (mRNA), transfer RNA (tRNA), micro RNA (miRNA) and other non-coding RNA. Transcriptomics examines RNA levels genome-wide, both qualitatively (which transcripts are present) and quantitatively (how much of each transcript is expressed). RNA is the molecular intermediate DNA and proteins.
- Proteomics: proteins are large and complex molecules made of thousands of amino acids, performing and enabling various tasks within biological systems and thus essential for life. The variety of protein is extensive and collectively, proteins catalyze and control nearly all cellular processes. They form a highly structure entity called proteome[200]. Proteomics is defined as the large-scale characterization of the entire protein complement of a cell, tissue or organism, at a specific time and location, and under given physiologic/pathologic conditions[58]. Proteomics analysis is used to quantify peptide abundance, modifications and interactions[340].
- Metabolomics: the metabolome is the complete set of metabolites found within a biological sample such as amino acids, fatty acids, carbohydrates or other products of cellular metabolic function. Metabolite levels and relative ratios are a reflection of metabolic function and abnormal perturbations may indicate some disease [340,344].

Other omics data types include miRNA-omics[345], epigenomics[56], microbiomics[346] and much others.

2. Challenges of omics studies

Each individual omics present different challenges of their own[347,348]. Multi-omics analysis therefore take on each of the challenges of the individual omics, and is faced with additional ones linked to treatment of non-uniform missing values, data integration, computation and visualization, as well as data annotation and storage[339,349]. The different multi-omics challenges are detailed in various publications[339,348–352], only some of them will be detailed here.

When setting up a multi-omics analysis, multiple aspects need to be considered. For example, as it was shown by Tarazona *et al.*[349], different omics platforms vary in the number of detected features. This is well represented with proteomics (**Figure 43-(A)**). Indeed, proteomics has an inherent bias of detecting more easily abundant proteins or targeting those with specific chemical properties, while this issue is almost non-existent in transcriptomics. Due to this differential feature coverage, analysis of the link between gene and protein expression is thus limited by the proteomics data. Moreover, different omics require a different number of samples to acquire reliable results as statistical power varies with sample size depending on the omic data[339,349], as shown in **Figure 43-(B)**. Reliability depends on the FDR which is linked to the number of measured entities. For example, proteomics might need around 14 samples per experimental group to achieve a power of 0.6 while with metabolomics, only 4 samples are needed to achieve the same statistical power[353].

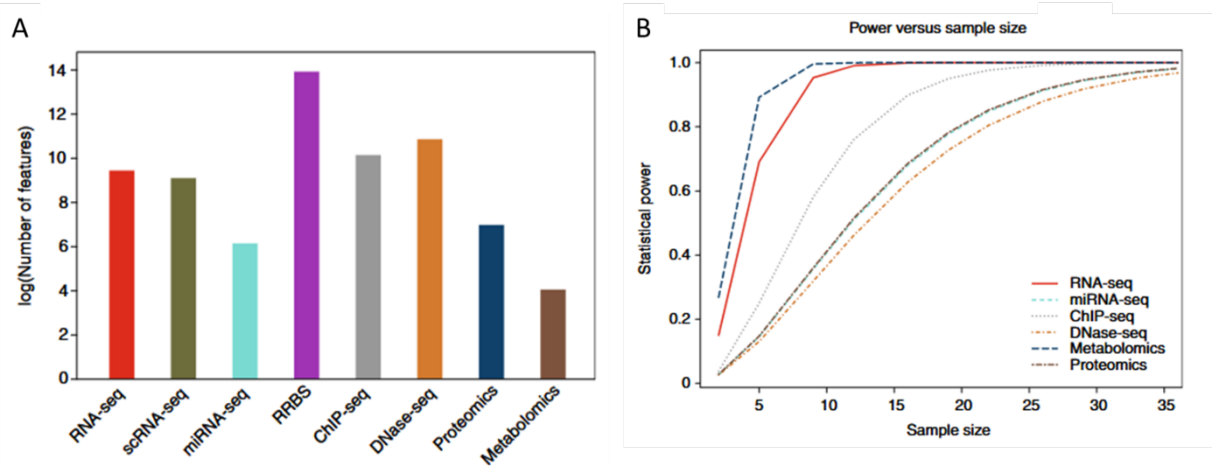


Figure 43: Comparison of the properties of some omics data type, from Tarazona *et al.*[349].(A) Number of features detected by each omic technology (B) Statistical power curves across omics data types as a function of sample size.

Another main challenge in multi-omics analysis is sample preparation. Indeed, most omics have their own sample preparation process, each facing their own hurdles[58,354]. Independent molecular extraction techniques, while allowing for the correlation of single molecular classes, add multiple limitations to the process with (i) additional experimental deviations (ii) more time consuming sample preparation and (iii) a need of high quantities of starting material, the latter being especially challenging for study of clinical samples[352]. In this context, some protocols have been developed, proposing a unified extraction procedure working for each omic, such as the Metabolite, Protein, and Lipid Extraction (MPLEx)[355] protocol or the Simultaneous Metabolite, Protein, Lipid Extraction (SIMPLEX)[356] and others[357,358]. More recently, the Bead-enabled Accelerated Monophasic Multi-omics (BAMM) sample preparation approach was proposed by Muehlbauer *et al.*[359]. This technique combines the addition of n-butanol based monophasic extraction solvent with the addition of unmodified magnetic beads. Then samples are incubated on ice for 5 minutes to allow proteins to aggregate onto the beads while metabolites and lipids remains in the supernatant. Unbound metabolites and lipids can be then removed for further analysis. Digestion is performed on the proteins for less than 1 hour and all three omics samples can be analyzed within the same day as sample preparation. Compared to other methods such as MPLEx or SIMPLEX, it allows for the generation of comparable data depthwhen applied to various type of sample (cell pellets, bio-fluids, cell culture plates) while saving an average of 19 hours[359].

One other main hurdle faced by multi-omics is the reproducibility of the analysis. It is quite challenging due to the diversity of methods and tools used for data analysis and statistical processing. Many development are done thus aimed at achieving the Findability, Accessibility, Interoperability and Reusability (FAIR) standards[339,360,361], described in **Figure 44**.

In an effort of to meet those principles, public platforms as GitHub are used to openly share scripts and codes of analysis. An example of an openly accessible omics data platform is the Omics Discovery Index (OmicsDI) which allows access and integration of genomics, transcriptomics, proteomics and metabolomics datasets[362]. This database contains more than datasets, covering almost 4000 different diseases, and more than 6000 species[348]. Other multi-omics data repositories exists but they focus on

one specific disease[350], such as The Cancer Atlas Genome (TCGA)[363] or the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)[364].

FINDABILITY	INTEROPERABILITY
• F1: (meta)data are assigned a globally unique and persistent identifier • F2: data are described with rich metadata (defined by R1 below) • F3: metadata clearly and explicitly include the identifier of the data it describes • F4: (meta)data are registered or indexed in a searchable resource	• I1: (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. • I2: (meta)data use vocabularies that follow FAIR principles • I3: (meta)data include qualified references to other (meta)data
ACCESIBILITY	**REUSABILITY**
• A1: (meta)data are retrievable by their identifier using a standardized communications protocol • A1.1: the protocol is open, free, and universally implementable • A1.2: the protocol allows for an authentication and authorization procedure, where necessary • A2: metadata are accessible, even if data are no longer available	• R1: meta(data) are richly described with a plurality of accurate and relevant attributes • R1.1: (meta)data are released with a clear and accessible data usage license • R1.2: (meta)data are associated with detailed provenance • R1.3: (meta)data meet domain-relevant community standards

Figure 44: The FAIR principles and main guidelines[360,361].

3. Multi-omics data integration

The integration of multi-omic data is a key step in the analysis, as it will help unravel and understand underlying mechanisms of the studied disease. It is however far from trivial because of data heterogeneity between the different –omic data. Indeed, they are mostly generated using various technologies and platforms, operating in different dynamic ranges of detection and quantification. Moreover, integration of multi-omic data also increases the difficulty to handle false positives in the combined datasets.

Currently, no unanimous workflow has been adopted by the scientific community for multi-omic data integration. However, some tools have been developed to improved data handling such as Similarity Network Fusion (SNF)[365], mixOmics[366], Multi-Omics Factor Analysis (MOFA)[367] for example. MOFA is an unsupervised computational method for integration of multi-omics data on the same (or at least partially) samples. MOFA allows analyzing the sources of heterogeneity in multi-omics data set, thus improving the identification of discrete subgroups of samples[350,367]. MixOmics is a R package based on supervised and unsupervised multivariate approach to perform data integration with focus on data exploration, dimension reduction and visualization. SNF is based on the construction of networks of samples (*ie* patients) for each data type and then fuses iteratively these networks into a single similarity network. This network represents a full vision of the data, and avoids dealing with different scales and noise, that will disappear with iterations[350,365]. Many other tools for multi-omics data integration, based on different approaches and with various applications, are reported in literature[339,350].

Machine learning tools, which are already used for single omics data, are now also being developed to investigate and integrate multi-omics data. Machine learning tools have been applied to a variety of multi-omics studies, from brain disease[8] study to single-cell analysis[368]. A plethora of different machine learning and deep learning techniques are used in the literature for multi-omics data analysis, all with various goals and characteritics[369,370].

PART III

Development of a fully automated high throughput phosphoproteomics workflow

Protein phosphorylation regulates many cellular processes, as discussed in **Part II, Chapter 2**. It is a fast and reversible reaction, modulating protein conformation, thus activating or inhibiting protein function, enzyme activity and influencing protein interactions[4]. Despite its widespread use in laboratories, and despite technological and instrumental evolutions for the past decade, phosphoproteomics still faces many challenges[7,180].

A significant number of quantitative phosphoproteomics studies are based on isotopic labeling techniques using TMT as they allow increased sample throughput, thus improved statistical power and limited missing values[186,207,231,371–373]. Labeling however greatly increases the cost of the analysis as well as the time of sample preparation and the quantification accuracy is reduced compared to label-free approaches. For these reasons, my goal was to develop a robust label free phosphoproteomic method allowing for the identification and quantification of phosphorylations.

Additionally, when working on clinical samples, the amount of starting material at disposal can be small, when working on tumor biopsies for instance, whereas the usual amount of sample required for phosphoproteomics analysis is relatively high, as the many steps of the workflow might generate sample losses. Moreover, in most multi-omics clinical studies, a high number of samples are used to be as representative as possible of the studied population. Therefore, the whole workflow needs to be highly reproducible and high throughput. This is why I focused on developing an automated and high throughput phosphoproteomics workflow, applicable on large cohorts of samples and compatible with small amounts of starting material.

To develop this method, I focused on key points of the workflow: protein extraction and digestion, nanoLC-MS/MS analysis, and data treatment. The analytical framework is represented in **Figure 45**.

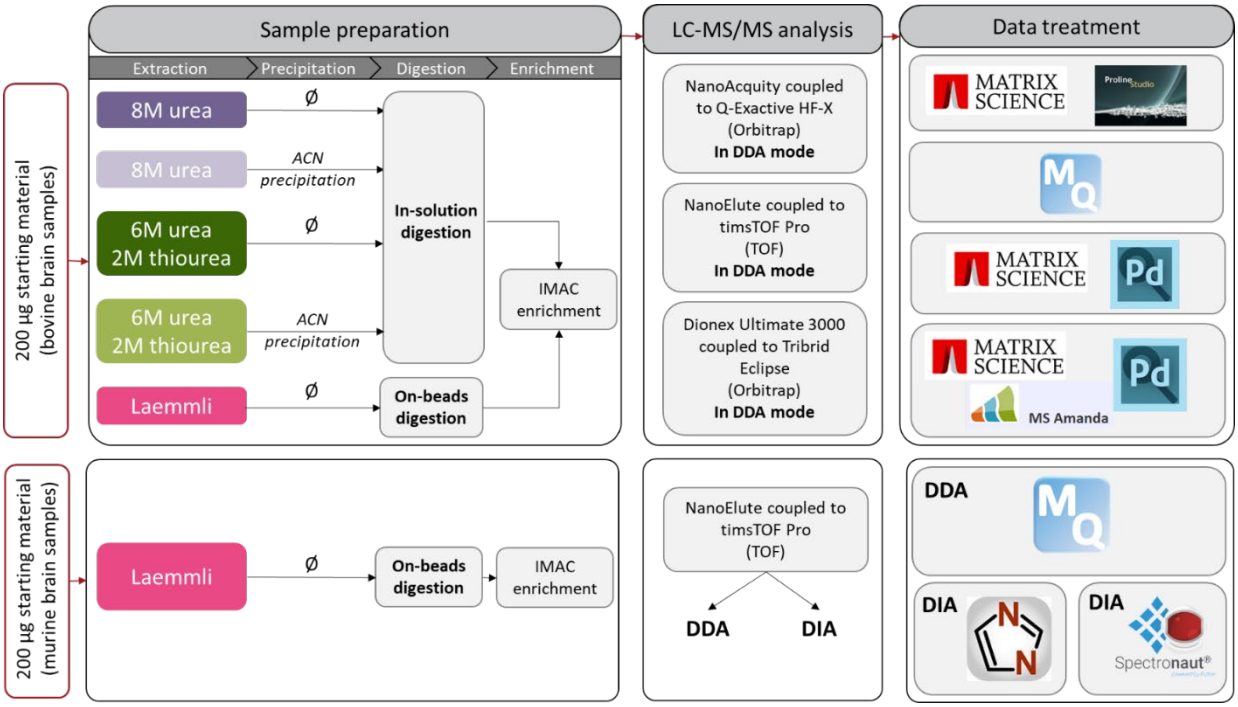


Figure 45: Analytical framework for the development of an automated and high throughput phosphoproteomics workflow.

A first experiment was conducted on bovine brain samples to optimize the different parts of the workflow:

- First, protein extraction from brain tissue was optimized by evaluating different lysis buffers. Two different digestion protocols were also compared. After protein digestion, phosphopeptide enrichment of the samples was performed using a standard IMAC method on an AssayMAP Bravo platform.
- For the nanoLC-MS/MS analysis, two different platforms were compared and evaluated: the Q-Exactive HF-X (Thermo) and the TimsTOF Pro (Bruker), both using HCD/CID fragmentation methods. Some tests were also conducted on a Tribrid Eclipse (Thermo, Orbitrap/Ion Trap) using an alternative fragmentation method, ETD.
- Finally, different pipelines were compared for the database search (Mascot (Matrix Science), Andromeda and MS Amanda), the localization and validation (Mascot, Andromeda, PhosphoRS), and the quantification (Proline, MaxQuant, Proteome Discoverer (Thermo)) of phosphorylation sites.

Taking the optimizations one step further, the developed method was applied on mouse brain tissues and both DDA and DIA methods for phosphopeptides analysis were optimized on the TimsTOF Pro. Finally, two different software were evaluated for phosphoproteomics DIA data analysis: Spectronaut (Biognosys) and DIA-NN[3] .

All experimental protocols are detailed in **Part IV: Experimental section.**

Chapter 1: Development of a high throughput and automated phosphoproteomics sample preparation workflow

1. Determination of the most adapted protein extraction protocol

An efficient protein extraction relies on releasing proteins by cell lysis and is a key step for (phospho)proteomic analysis. The lysis buffer must thus be able to solubilize as many proteins as possible without modifying or degrading them, and without introducing a bias of over- or under-expression of one subtype of proteins. Different protocols were compared for tissue protein extraction on triplicates of about 30 mg of bovine brain. A representation of all protocols is displayed in **Figure 46**.

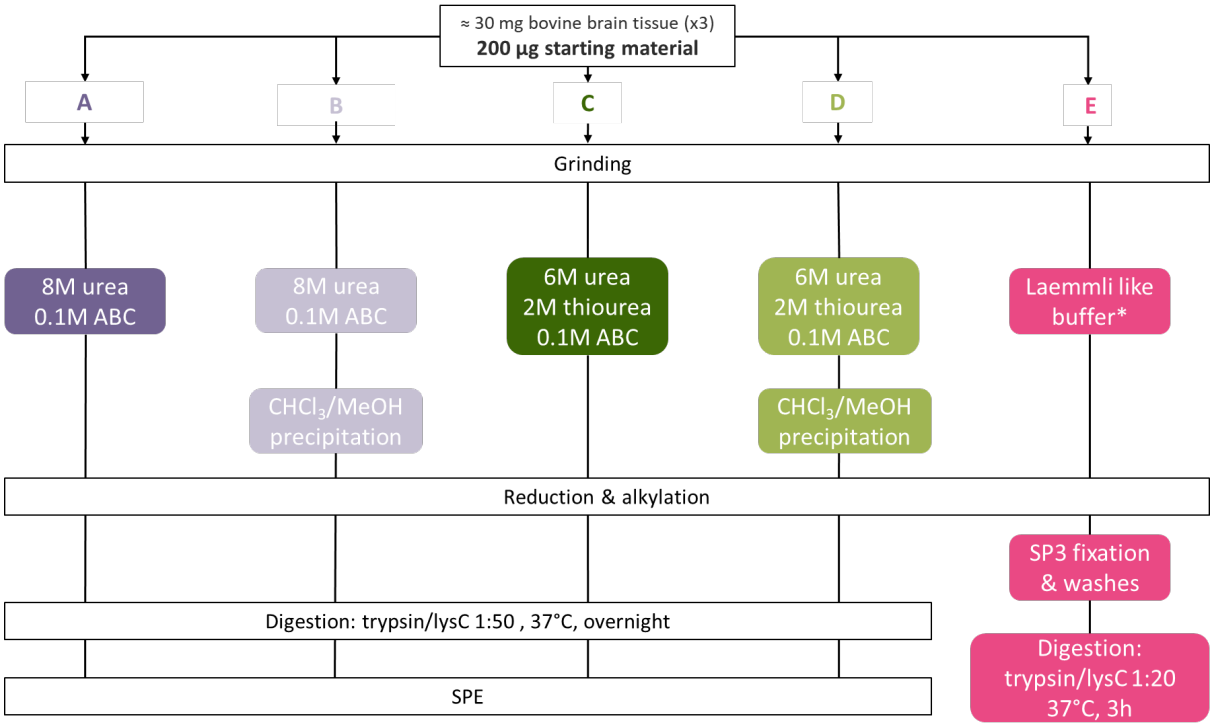


Figure 46: Method development for protein extraction and digestion. Five different protocols were compared: A = 8M urea; B = 8M urea with precipitation; C = 6M urea, 2M thiourea; D = 6M urea, 2M thiourea with precipitation; E = Laemmli-like buffer**.**

Three different lysis buffers described in literature[374] were used to extract proteins from the brain tissues: a buffer containing 8M of urea in 0.1M of ammonium bicarbonate (ABC), another with 8M urea and 6M thiourea in 0.1M of ABC, and finally a Laemmli-like buffer with 2.5% SDS in 50mM of TrisHCl. Each buffer contains proteases inhibitors and phosphatases inhibitors to prevent phosphorylation degradation. For the two urea-based buffers, a step of chloroform and methanol ($CHCl_3$/MeOH) precipitation was added to remove impurities from the sample such as salts or lipids that might interfere with the enzymatic digestion or the LC-MS/MS analysis.

First of all, we compared the yield of extraction for each protocol, as displayed in **Figure 47**. The two conditions with precipitation (conditions B and D) display the lowest results. Indeed, the addition of a

precipitation step is known to cause sample losses[58] but can be useful to remove contaminants before MS analysis. The two highest yields are obtained respectively with the urea and the Laemmli-like extraction but the Laemmli-like extraction appears to be more reproducible than the urea one. Using a combination of thiourea and urea does not seem to improve the extraction. Indeed conditions C and D both display lower results compared to their only-urea counterparts (respectively, conditions A and B).

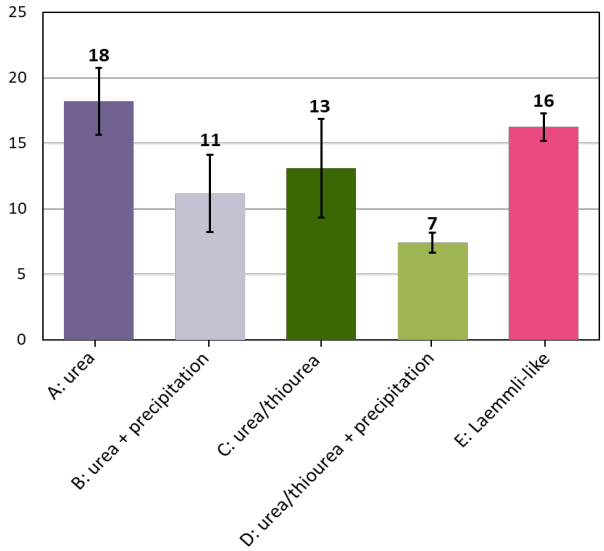


Figure 47: Average yield of extraction (in %) calculated from the theoretical initial mass of tissue and the protein amount in the sample measured for each extraction by proteic assay.

For all urea-based conditions, protein extracts were then digested using a mixture of trypsin and lysC before undergoing peptide clean-up using a SPE protocol to remove salts (especially urea) from the samples. For the Laemmli-like condition, a SP3 digestion was performed, which is compatible with SDS. A peptide clean-up step was not used for SP3 as washing steps are performed during the protocol before the digestion. The generated peptides were injected on a nanoAcquity coupled to a Q-Exactive Plus (Thermo) to check the performance of the protocols before phosphopeptide enrichment. Proteins and peptides identification was performed by Mascot and validation using Proline. Identification results for the different protocols are displayed in **Figure 48**.

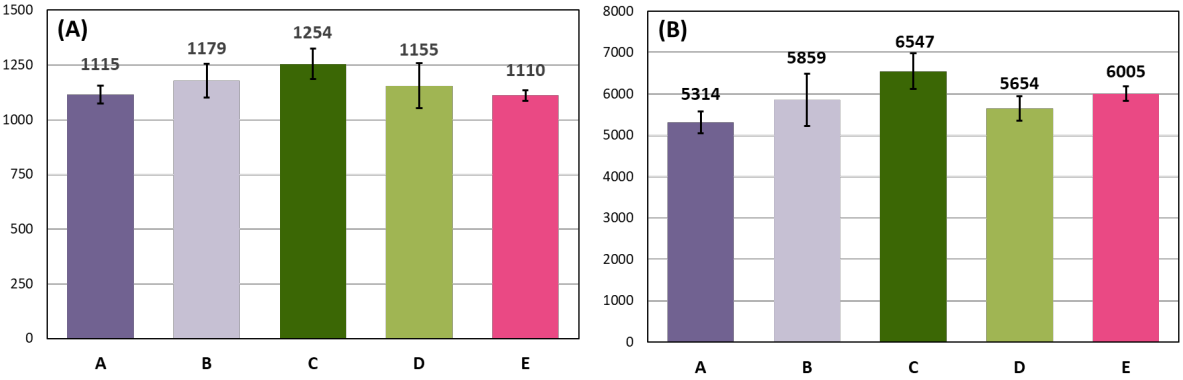


Figure 48: Average number of (A) proteins and (B) peptides identified and validated for each extraction. A = urea; B = urea with precipitation; C = urea/thiourea; D = urea/thiourea with precipitation; E = Laemmli-like.

Surprisingly, compared to the extraction yield results, the best identification results for both proteins and peptides are obtained with condition C with 1254 proteins and 6547 peptides identified. The addition of thiourea in the extraction buffer thus seems to solubilize or extract additional proteins compared to a urea-only buffer. These additional identifications are represented by the Venn diagram in **Figure 49** and 16% of proteins are identified exclusively using the urea/thiourea buffer (condition C).

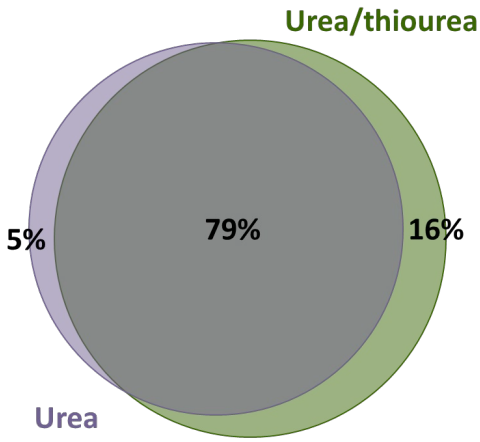


Figure 49: Venn diagram of the overlap of identified proteins between the urea condition (A) and the urea/thiourea condition (C).

The Laemmli-like condition, while not being the best in terms of protein identifications, has the second highest number of peptides identified after the urea condition with 6005 peptides identified (**Figure 48**) with the best reproducibility of both proteins and peptides identification. Like for extraction yield results, the precipitation step decreases the number of identifications for the urea/thiourea buffer. However, it surprisingly seems to improve the results for the urea buffer with a 6% increase in terms of proteins identified and an 11% increase for peptides identified using precipitation. However, these results need to be taken cautiously, as if we take into account the relatively high standard deviations of conditions A and B' results, the difference of identification between the two conditions is not representative.

Quantitation was performed with Proline[158] and a stringent filter was applied as quantitative values were required in 3 out of the 3 replicates, meaning no missing values were allowed. As shown by the **Figure 50 – (A)**, condition C with urea/thiourea and condition with Laemmli allows for the highest quantifications with respectively 4087 and 3956 quantified peptides. Coefficients of variation (CV) were then computed on the intensities of quantified peptides and their distribution by condition are represented by **Figure 50 – (B)**. The Laemmli extraction (condition B) displays the lowest CV with a median CV of 15.7%.

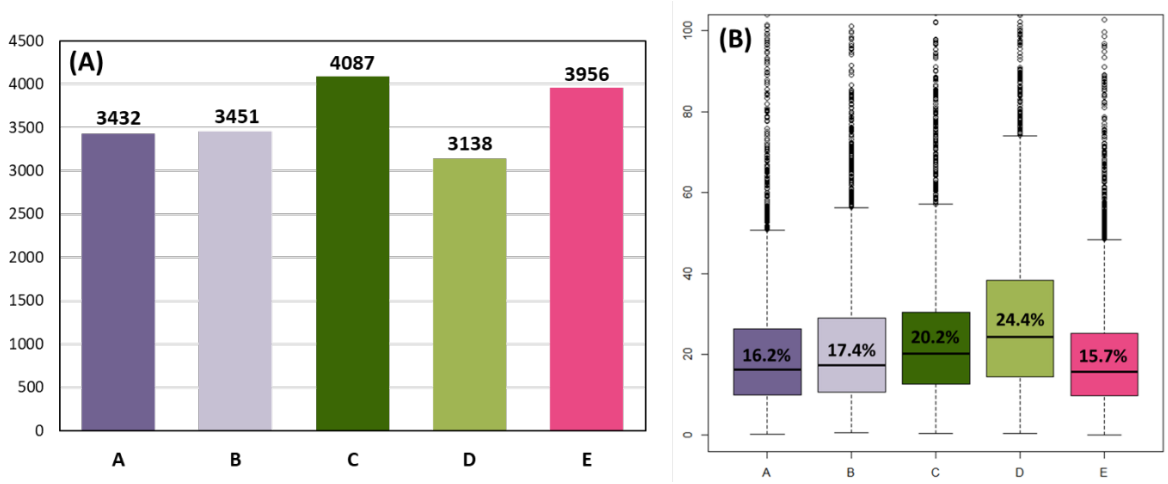


Figure 50 : (A) Number of quantified peptides in 3/3 replicates for each condition (B) Boxplots representing the distribution of CVs on the peptide intensities and the median CV per condition**.**

Overall, conditions 3 (urea/thiourea) and 5 (Laemmli-like) display the best results in terms of identification, quantification and reproducibility. Those two conditions are therefore chosen to further evaluate the phosphopeptide enrichment step.

2. Evaluation of the automated phosphopeptide enrichment protocol

Phosphopeptide enrichment is performed on 150 µg of digested peptides in a solution of 80% ACN, 0.1% TFA using IMAC cartridges filled with a 5 µL Fe(III)-NTA phase. Peptides are eluted by increasing the pH with a 1% NH_4OH solution and enriched phosphopeptides fractions were then injected on a NanoElute coupled to a TimsTOF Pro (Bruker).

i. Enrichment efficiency

To estimate the enrichment efficiency, a mixture of isotopically labeled (heavy or light) synthetic phosphopeptides can be added to the samples. These phosphopeptides, called Phosphomix, are derived from naturally occurring peptides in HeLa cells and commercialized by Thermo Fisher Scientific. Different mixes can be used, all of them containing 10 different synthetic phosphopeptides that are mono- or bi-phosphorylated. They were used to evaluate the efficiency and reproducibility of the phosphopeptide enrichment process, as they were added before (Phosphomix light, in their naturally occurring isotopic abundance) and after the enrichment (Phosphomix heavy, in their stable isotope enriched version). They are not widely used in literature as the ratio of phosphorylated peptides over all peptides is more commonly used. Other phosphopeptides mixtures are also used to monitor protein phosphorylation:

- SpikeMix™ PTM-Kit (JPT): pool of 100 proteotypic phosphoserine and phosphothreonine containing peptides.
- SureQuant™ Phosphopeptide suitability standard (ThermoFisher Scientific): The SureQuant Multipathway Phosphopeptide Standard contains an optimized mixture of 131 isotopically labeled phosphopeptides while the SureQuant Phosphopeptide Suitability Standard contains a mixture of 20 isotopically labelled phosphopeptides with increasing hydrophobic properties.
- SigPath[375]: 298 synthetic heavy-labelled phosphopeptides.

As described in the Phosphomix technical sheet (**Table 4**), some of those phosphopeptides are more easily detected than others.

Peptide	Number of phosphate group	Monoisotopic light mass weight	Monoisotopic heavy mass weight	Relative signal intensity
VLHSG**S**	1	834.37	844.38	Weak
RS**YS**RS	2	1070.41	1080.41	Weak
RDSLG**T**YSS	1	1220.52	1230.53	Medium
TKLI**T**QLRDA	2	1445.70	1453.72	Strong
EVQAEQPSS**S**SP	1	1480.62	1490.63	Medium
ADEP**S**SEESDLEID	1	1742.68	1750.69	Strong
ADEPS**S**EE**S**DLEID	2	1822.64	1830.66	Medium
FEDEGAGFEES**S**ETGDYEE	1	2333.84	2341.85	Strong
ELSN**S**PLRENSFG**S**PLEF	2	2338.00	2348.01	Medium
SPTEYHEPV**Y**ANPFYRPT**T**PQ	2	2809.19	2819.20	Strong

Table 4: Technical sheet of Phosphomix 1 with the peptide sequences and the phosphorylated sites, the associated heavy and light mass weight, and signal intensity. The relative signal intensity was obtained after reversed phase LC-MS/MS with ESI ionization.

Equivalent quantities of phosphomix light and heavy were added before and after the phopshopeptide enrichment. The ratio of the heavy peptides intensities over the light peptides intensities thus corresponds to the ratio of peptides enriched *ie* the enrichment efficiency. In the Proteomics Multicentric Experiment 11 (PME11) inter-laboratories study, every lab added phospomix 1 and 2 to yeast lysate before performing phosphopeptide enrichment (either IMAC or MOAC) and highlighted the varibility of enrichment process between differnet labs[376]. To our knowledge, the only study describing the use of phosphomix standards to evaluate IMAC enrichment is an application note from Agilent. In this study, they used phosphomix 1 and 2 to highlight the efficiency of the IMAC phosphopeptide enrichment using Fe(III)-NTA cartidges on an AssayMAP Bravo[220]. This study was however performed on cell lysates and not tissues, but as a comparison, only 6 out of the 10 phosphomix were identified in the experiment.

Phosphomix 1 was added to the peptidic mixture from the urea/thiourea condition and the corresponding phosphopeptides ion currents' were extracted using Skyline (v.20.2.0). Out of the 10 synthetic phosphopeptides, only 3 of them have a usable signal : EVQAEQPSS**S**SP, ADEPS**S**EE**S**DLEID, and FEDEGAGFEES**S**ETGDYEE. According to their technical sheet, they are expected to have from a medium to a strong relative signal intensity. As summarized in **Table 5**, they were some of the most frequently identified phosphomix standards among the 31 laboratories of the PME11 initiative[376]. As a comparison, the mean ratio heavy/light of those peptides in the Agilent Application[220] note is also added in **Table 5.**

Peptide	Mean heavy/light ratio	Mean heavy/light ratio in Agilent application note	Frequency of detection in PME11 study (in %)
EVQAEQPSS**S**SP	75 ± 35%	82.5 ± 6.5%	100
ADEPS**S**EE**S**DLEID	45 ± 15%	95.1 ± 4%	65
FEDEGAGFEES**S**ETGDYEE	38 ± 27%	97.1 ± 4%	85

Table 5: Mean heavy/light ratio of the 3 observed Phosphomix for the urea/thiourea condition and comparison with mean heavy/light ratio and frequency of detection from literature.

The enrichment efficiency appears to be quite low, especially when compared to the ratios obtained in the Agilent technical note. However, we need to take into account that our experiment was performed on tissue samples, much more complex than cell lines, which can explain the differences obtained. Moreover, in the application note, the synthetic peptides were followed in a targeted way, using MRM, greatly increasing the sensitivity and specificity of the analysis toward those specific peptides. We can also note that the ratio of enrichment seems greatly dependent on the peptide. For example, ADEPS**S**EE**S**DLEID peptide was detected whereas its mono-phosphorylated counterpart ADEPS**S**EESDLEID was not. This might be due to the bias of IMAC enrichment towards multi-phosphorylated peptides[377,378]. Considering the low efficiency of enrichment as well as the low number of synthetic phosphopeptides detected, one can question the overall quality of the Phosphomix batch used, as it was prepared a couple of years ago, suggesting that the phosphopeptides might have been degraded.

A more widespread method to evaluate phosphopeptide enrichment efficiency is to look at the ratio of the number of phosphorylated peptides over the total number of identified peptides and eventually comparing this ratio for non-enriched samples. For urea/thiourea, this ratio for enriched samples is equal to 25±9% whereas it is of 13±1% for the Laemmli-like buffer.

ii. Identification results

As shown in **Figure 51**, between 640 and 713 phosphoproteins were identified on bovine brain tissues using either Laemmli like or urea/thiourea buffer for protein extraction. Phosphoproteomics studies on bovine brain tissues are almost inexistent in literature, so comparison can be made, cautiously, with other species' brain tissues. For example, on mouse brain tissues in a urea-based buffer and enriched by IMAC, almost 400 phosphoproteins were identified[379]. In another study, human brain tissues were homogenized in SDS buffer, labeled by TMT, fractionated and finally enriched by IMAC, identifying 4631 phosphopeptides[227]. Those results are however difficult to compare with ours as both TMT labeling and fractionation steps were used, as in numerous phosphoproteomics studies[6,228,235], which both greatly improve identifications.

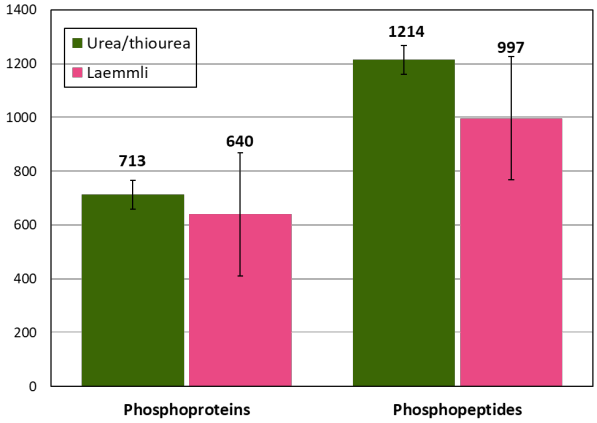


Figure 51: Mean number of phosphoproteins and phosphopeptides identified for the urea/thiourea and Laemmli extraction conditions.

Phosphoproteins and phosphopeptides are relatively well identified by both extraction buffers as they have around 45% to almost 60% of identification overlap (**Figure 52**). The two buffers seem quite complementary as around 20% of identifications are unique to the Laemmli-like buffer whereas approximately 30% are unique to the urea/thiourea buffer.

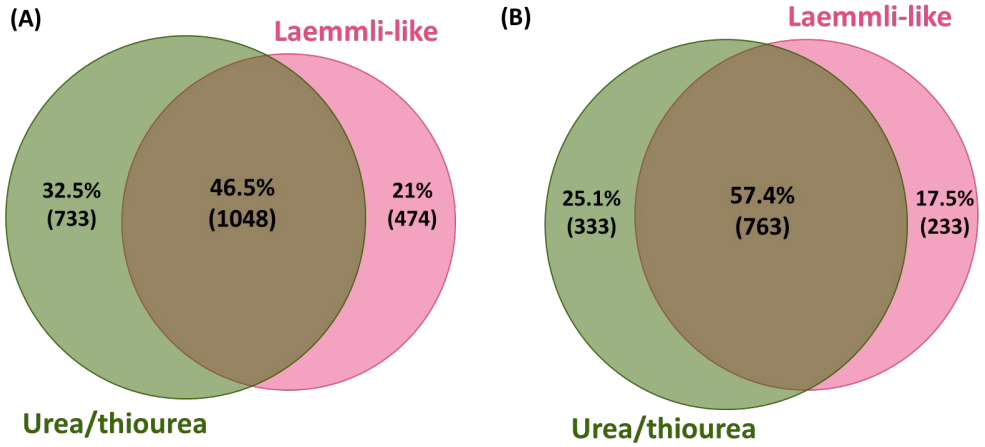


Figure 52: Venn diagrams of the (A) phosphopeptides and (B) phosphoproteins identified by the two extraction methods (urea/thiourea and Laemmli-like).

iii. Sample preparation reproducibility

Three technical replicates for each condition were prepared following the exact same steps of sample preparation in order to evaluate the reproducibility of the workflow. Almost 30% of the phosphopeptides are shared between the 3 technical replicates, corresponding to 423 phosphopeptides for the Laemmli condition and 460 for the urea/thiourea condition (**Figure 53**). A 30% overlap between the 3 technical replicates might seem low at first sight. However, compared to classical global proteomics, it was observed in the lab that on average only 50% of the peptides are common between 3 replicates due to the stochasticity of the DDA analysis. Here, we have an extra phosphopeptide enrichment step, adding another level of variability in sample preparation. Moreover, we are studying a very labile modification which induces another layer of variability in the analysis. Taking all these elements into the equation, a 30% recovery in phosphopeptides identified seems a reasonable result. Indeed, a 25% recovery was obtained on four replicates of lung cell lysate after TiO_2 enrichment[380]. On A431 lysate digest, 48% of recovery was obtained on unique phosphopeptides

identified across three technical replicates of Fe^{3+} IMAC enrichment[211]. More recently, 45% coverage was achieved between the phosphoproteome of 3 malignant melanoma samples after automated Fe(III)-NTA phosphopeptide enrichment[212].

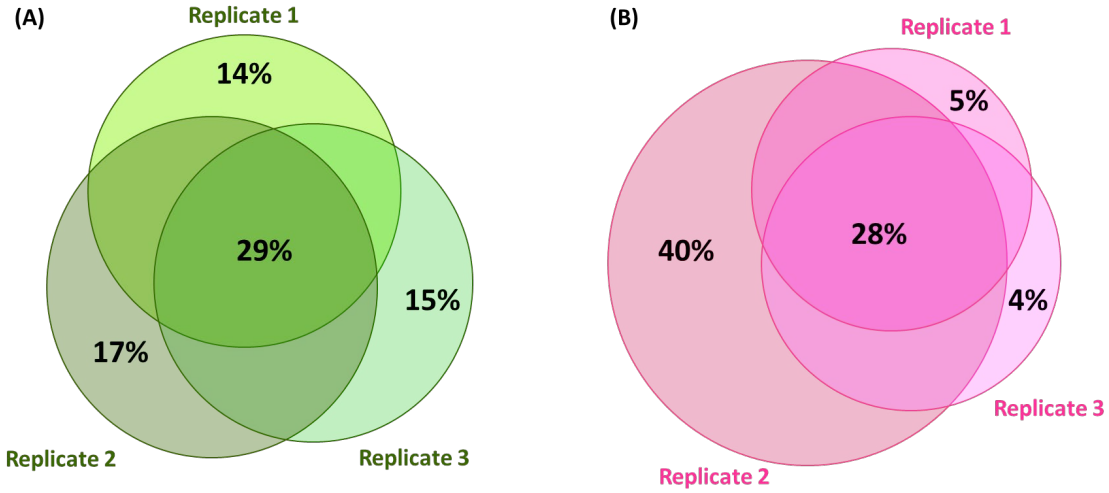


Figure 53: Venn diagrams of the phosphopeptides coverage between the 3 replicates for (A) the urea/thiourea extraction and (B) the Laemmli-like extraction.

Differences are observed for Replicate 2 of the Laemmli-like condition, compared to the two other replicates, as 40% of phosphopeptides are uniquely identified by this replicate. As the bovine brain samples were not clinically prepared but bought and home-made cut then only frozen, there can be a great variability from one sample to the other. Another explanation might be that the 2nd replicate was the first sample injected of the series, whereas Replicate 1 and Replicate 3 were stored in the auto-sampler at 4°C and injected almost 48h later in the sequence. We did not anticipate such a big effect of the time between different replicate injections. This effect is discussed further in the next paragraph on the reproducibility of the LC-MS/MS analysis.

iv. LC-MS/MS analysis reproducibility

To evaluate the reproducibility of the LC-MS/MS analysis, 3 injection replicates were analyzed for the urea/thiourea condition. The overlap of those injection replicates, injected at each 24h interval, is represented by **Figure 54**.

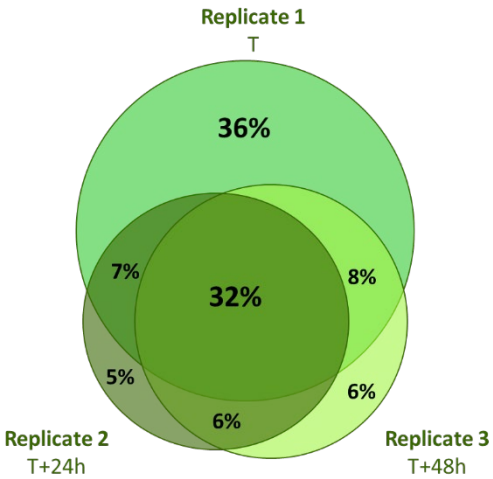


Figure 54: Venn diagram of the phosphopeptides overlap between 3 injection replicates.

We observe an almost 40% decrease in identification between T and T+24h. This loss of phosphopeptides might be explained by the thermal degradation of phosphopeptides but also by their potential adsorption onto the tube wall[206]. Thus, for all further phosphoproteomics experiments in this manuscript, samples after enrichment were frozen at -80°C and de-frozen by 24h-period before their injection on the LC-MS/MS platform.

v. Distribution of the phosphorylation on amino acids

The distribution of the phosphorylations on serine (S), threonine (T) and tyrosine (Y) residues for the two extraction conditions is represented in **Figure 55**.

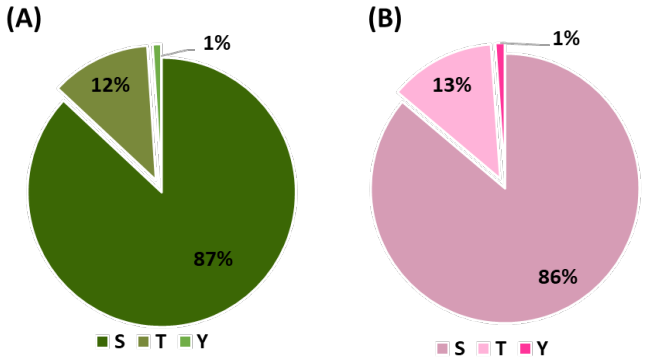


Figure 55: Distribution of the phosphorylation on serine (S), threonine (T) and tyrosine (Y) residues for (A) the urea/thiourea condition and (B) the Laemmli-like condition.

The repartition obtained is equivalent to the one in literature[181], as around 86-87% of phosphorylations were found on serine residues, 12-13% on threonine and 1% on tyrosine residues.

To conclude, the urea/thiourea condition allows for a greater number of phosphopeptide identifications with the highest efficiency of enrichment.

Chapter 2: Optimization of LC-MS/MS methods for the analysis of phosphopeptides

1. Evaluation of the best LC-MS/MS platform for phosphoproteomics

i. Comparison of two different nanoLC-MS/MS systems for phosphopeptides analysis: Q-Exactive HF-X and TimsTOF Po

Here, the goal was to compare a well-established Orbitrap instrument, the Q-Exactive HF-X (Thermo), to a latest generation instrument equipped with an ion mobility dimension, the TimsTOF Pro (Bruker), for the specific purpose of phosphopeptides analysis. Indeed, when I began my PhD, the acquisition methods were optimized for classical proteomics on the TimsTOF Pro but no evaluation had been performed for phosphopeptides analysis. In addition, very few publications described phosphoproteomics on a TimsTOF Pro in the literature. We thus started using "standard" parameters on the machine to compare its performances to the ones of the Q-Exactive HF-X. Later, we have further optimized those parameters (see **Chapter 2 – 2. Optimization of a DDA method on a TimsTOF Pro platform and 3.Development of a dia-PASEF method**).

The main differences between the two nanoLC-MS/MS platforms are the fragmentation modes and the additional dimension brought by the ion mobility in the TimsTOF. Moreover, the TimsTOF is coupled with the PASEF technology, greatly increasing the instrument's sensitivity (see **Part II: State of the art** for more details). The two mass spectrometers also use different fragmentation techniques as the Q-Exactive HF-X employs HCD whereas TimsTOF uses CID fragmentation. The additional energy brought with HCD might allow for a better fragmentation of phosphopeptides by reducing neutral losses of the phosphate group and thus generating more informative spectra[180].

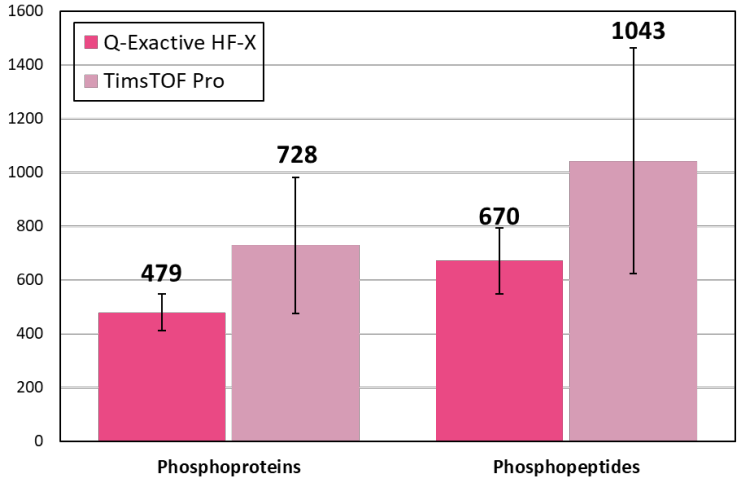


Figure 56: Mean numbers of identified phosphoproteins and phosphopeptides on the two instruments.

Identification of phosphopeptides was performed using MaxQuant, and results are displayed in **Figure 56.** Identifications are around 50% higher with the TimsTOF Pro compared to the Q-Exactive HF-X. However, the reproducibility of identifications is more than twice lower with the TimsTOF Pro than the

Q-Exactive HF-X. Indeed, for the Q-Exactive HF-X the standard deviation is of 14% and 18% respectively for phosphoproteins and phosphopeptides, compared to 35% and 40% for the TimsTOF Pro.
The recovery of the identified phosphopeptides between the two platforms is represented in **Figure 57**. Less than 30% of the phosphopeptides identified robustly (in 3 out of the 3 replicates) are common between the two platforms and more than 300 phosphopeptides are uniquely identified in all replicates on the TimsTOF Pro.

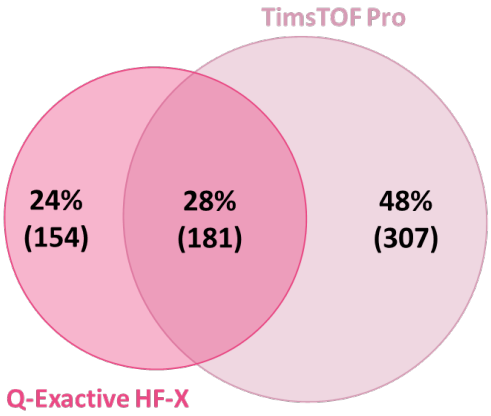


Figure 57: Venn diagram representing the overlap between the most reliable phosphopeptides identified by the two platforms in 3 out of 3 replicates.

In terms of quantification performances, the median CVs of the intensities of all phosphopeptides quantified in 3 out of 3 replicates were calculated for both platforms. The median CV was of 41% for Q-Exactive HF-X data and 50% for TimsTOF Pro data, highlighting a slightly more robust quantification of phosphopeptides with Q-Exactive HF-X.

All these results highlight that despite a slightly increased robustness of quantification with HCD fragmentation on Orbitrap instruments, the TimsTOF Pro offers promising identifications of phosphopeptides. Both platforms seem complementary as they both identify different populations of phosphopeptides. The TimsTOF Pro is not yet as widespread as the Q-Exactive HF-X for phosphoproteomics analysis. Optimizations are therefore necessary to improve the reproducibility of identified and quantified phosphopeptides on a TimsTOF Pro platform. Those optimizations will be developed later in this manuscript (see **Part III -Chapter 2: Optimization of a DDA method on a TimsTOF Pro platform**).

ii. Investigation of an alternative fragmentation technique for phosphopeptides: electron transfer dissociation on a Tribrid Eclipse instrument

Many publications emphasize the use of alternative fragmentation methods (in opposition to classical CID/HCD) to improve the analysis of modified peptides and especially phosphopeptides. Indeed, collision fragmentation (CID/HCD) often leads to the loss of the phosphate group, instead of the peptidic bond fragmentation, thus leading to non-informative spectra. Softer ionization techniques, such as ETD, allow keeping the phosphate group while fragmenting the peptidic bond, allowing to get information on the peptide sequence and the site localization[224].

We wanted to test this fragmentation method, available in the lab on an Orbitrap Eclipse Tribrid (Thermo) equipped with multiple fragmentation modes (HCD, CID, ETD, EThCD, UVPD). Samples enriched in phosphopeptides were thus injected in ETD and in HCD for comparison. Generated data were analyzed using Proteome Discoverer.

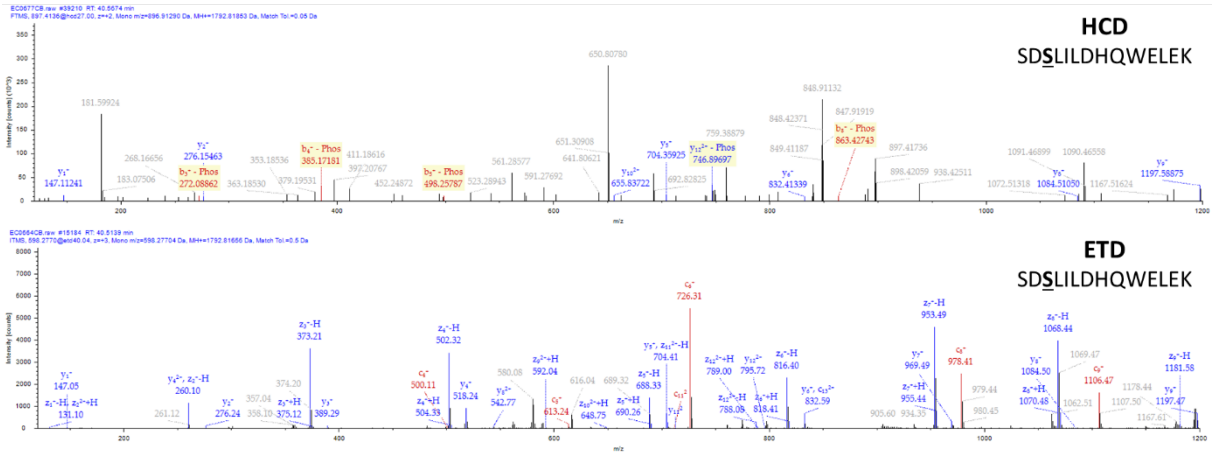


Figure 58: MS/MS spectra of peptide SD**S**LILDHQWELEK obtained by HCD (upper panel) or ETD (lower panel) fragmentation on the Tribrid Eclipse instrument.

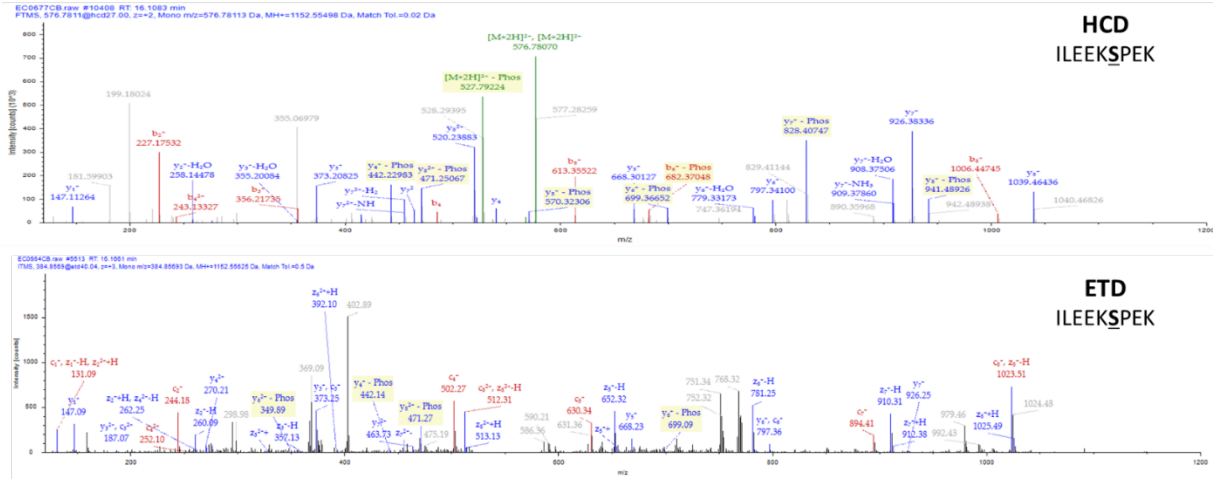


Figure 59: MS/MS spectra of peptide ILEEK**S**PEK obtained by HCD (upper panel) or ETD (lower panel) fragmentation on the Tribrid Eclipse instrument.

In **Figure 58**, fragmentation spectra of the SD**S**LILDHQWELEK phosphorylated peptide, obtained either by HCD or ETD are represented. In HCD, very few fragments are detected compared to the much richer spectrum of ETD. In **Figure 59**, spectra (by HCD and ETD) of the ILEEK**S**PEK phosphorylated peptide are represented. Here, in HCD, the most intense peak at 576.78070 corresponds to the mass of the non-fragmented peptide, while the peak at 527.79224 m/z corresponds to the non-fragmented peptide with a phosphate loss (-98 Da). On the other hand, the ETD spectrum displays lots of informative fragments that allow to assemble the peptide sequence and localize the phosphate group.

When comparing the phosphosites identified by both techniques (**Figure 60**), 500 additional phosphosites are identified using HCD. Despite an easier identification of phosphosites thanks to better fragmentation spectra, ETD identifies a smaller total number of phosphosites. Indeed, ETD is a slow fragmentation technique as it needs a consequent reaction time between positively charged peptides and fluoroanthene anion. This greatly increases the overall cycle time of the MS analysis,

leading to poorer coverage of the chromatographic peaks. This explains the lower identifications of phosphosites compared to ETD. Additionally, we generated different phosphosites with the two technics, suggesting that they are complementary. The same observation is described in literature, and ETD and HCD are depicted as complementary techniques that allow the identification of different populations of phosphorylations[381].

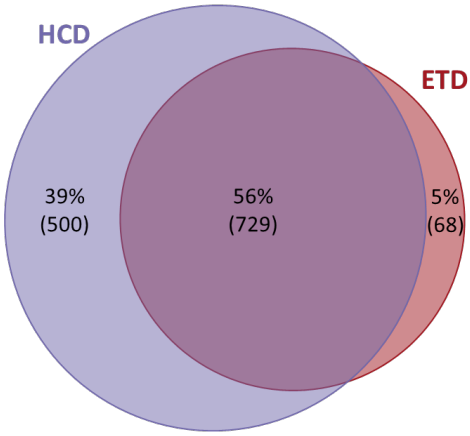


Figure 60: Overlap of the identified class I phosphosites between HCD and ETD fragmentation.

While ETD generates in theory promising results, it lacks for now optimizations and stays too slow for large scale phosphoproteomics.

2. Optimization of a DDA method on a TimsTOF Pro platform

In order to improve the performances of the TimsTOF Pro for the analysis of phosphopeptides, different parameters were optimized: the collision energy, the ion mobility window, the accumulation time, the number of PASEF scans and the LC gradient. All those optimizations were performed on murine brain tissues samples enriched in phosphopeptides and data analyzed using MaxQuant.

i. First evaluation of MS/MS method parameters for phosphoproteomics on TimsTOF Pro

The ion mobility value depends on the mass, the charge and the shape of the considered ion. It is also affected by the presence of a modification, as Ogata *et al*. showed that the reduced mobility coefficient (1/K0), also called Collision Cross Section (CCS), of phosphorylated peptides has a different value from its unmodified counterpart[132]. Thus, as the collision energy is applied as a function of the measured ion mobility, both parameters deserve being optimized for phosphoproteomics.
The value of the accumulation time of the ions in the mobility cell will also have an impact on the analysis as the longer it is, the more ions are accumulated and thus the more information we will obtained. This is particularly suitable for the analysis of low quantities of material. However, if accumulation time is too long, we risk impacting the duty cycle and increase the cycle time. The cycle time is also an important parameter to optimize. As represented in **Figure 61**, the longer the cycle time the smaller the number of points per chromatographic peak thus the less defined the chromatographic peak. Therefore, a balance needs to be found for an optimal cycle time in which there are enough points per peak (usually between 8 to 10) to achieve the best accuracy and resolution possible. The number of PASEF scans in a method will also have an impact on the cycle time. If not enough PASEF

scans are performed, the risk is that not all precursors will be fragmented, but the greater the number of PASEF scans, the longer the cycle time.

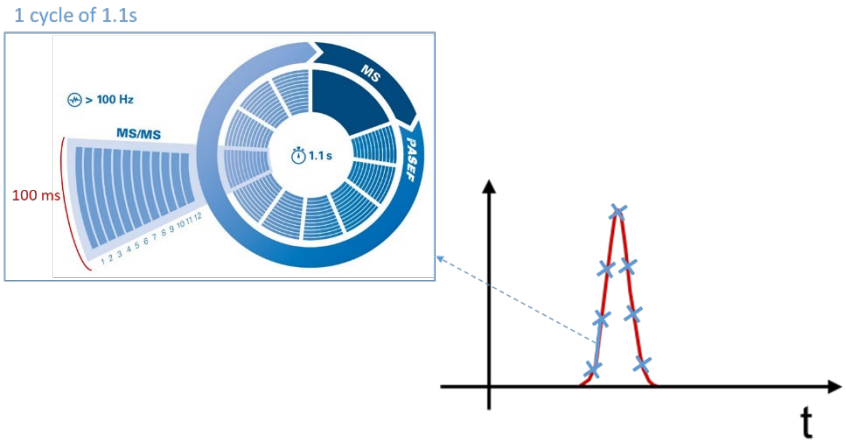


Figure 61: Description of one cycle across a chromatographic peak.

All these parameters were evaluated and optimized in order to achieve the best performances for phosphopeptides analysis. We first set up two methods (A and B) and compared them with the two default methods supplied by Bruker (C and D). All other methods developed and their parameters are detailed in **Figure 62 – (A)** and **Figure 62 – (B)**.

(A)	Ion mobility window (1/K0)	Collision energy range (eV)	Slope	Accumulation time (ms)	Number of PASEF scans	Cycle time (s)
Method A	0.7 → 1.25	22 → 62	Stepwise	166	10	1.89
					8	1.55
					6	1.2
Method B	0.7 → 1.25	22 → 62	Linear	166	10	1.89
					8	1.55
					6	1.2
Method C	0.6 → 1.6	20 → 80	Stepwise	100	10	1.17
Method D	0.6 → 1.6	20 → 60	Stepwise	100	10	1.17

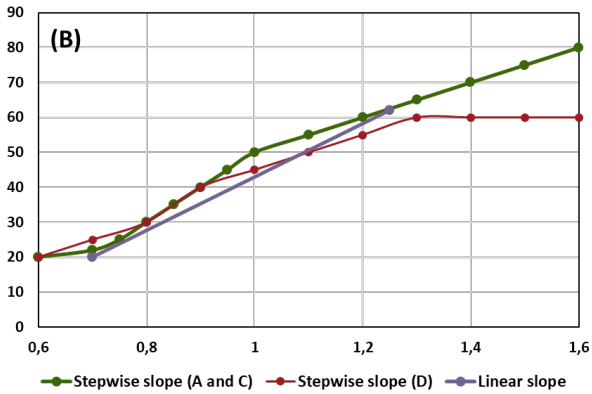


Figure 62: (A) Table describing the 4 different methods used with their respective ion mobility range, collision energy range, slope of collision energy range, accumulation time, number of PASEF scans and cycle time (B) Description of the stepwise and linear slopes of collision energy range used for the different methods.

As displayed in **Figure 63**, optimized methods A and B both outperform C and D methods in terms of numbers of class I phosphosites identified. Indeed, because of their longer accumulation time of 166 ms (in contrast to100 ms for methods C and D), methods A and B identify between 300 to 500 additional class I phosphosites compared to methods C and D. Methods with a linear slope of collision energy (methods B) give slightly better results than methods with a stepwise slope (methods A) with in average a 10% increase in identifications. No significant difference is however observed for a method depending on the number of PASEF scans.

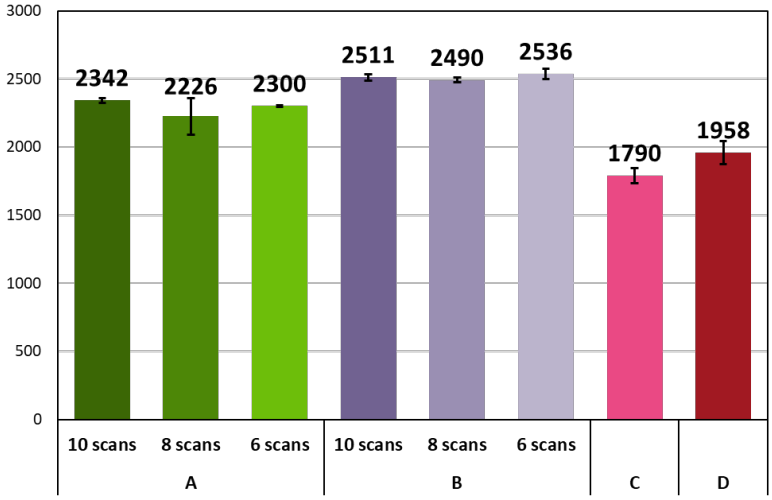


Figure 63: Mean number of class I phosphosites identified by each method.

Similarly to identifications, optimized methods A and B both quantify more class I phosphosites than methods C and D (**Figure 64 – (A)**). Method B, with a linear range of collision energy, allows for the quantification of more than 1800 class I phosphosites.

To evaluate the accuracy of quantification of the different methods, the coefficients of variation were calculated on the intensities of the class I phosphosites quantified in the two replicates (**Figure 64 – (B)**). Method C, with a median CV of 6.6%, appears to be the most robust for quantification, as it uses the higher collision energies and thus suggests a more efficient fragmentation of the phosphopeptides. However, every other methods seem to have similar reproducibility with all CV < 20%, except for method A with 8 scans which displays a CV of almost 30%.

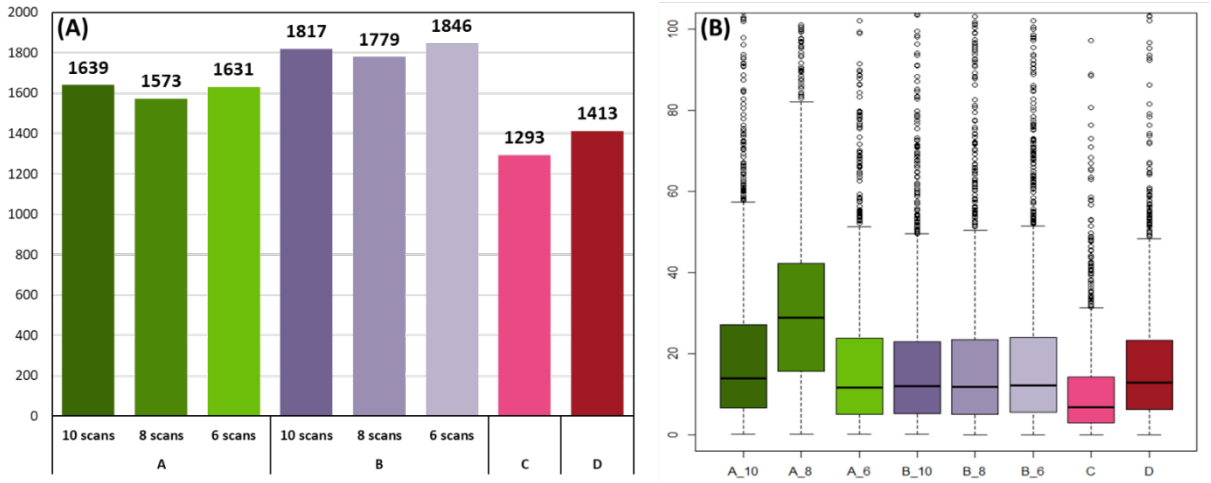


Figure 64: (A) Number of class I phosphosites quantified in 2 out of 2 replicates per method (B) Repartition of the CVs on class I phosphosites quantified in 2 out of 2 replicates per method.

These results are to be interpreted cautiously as only 2 injection replicates per method could be injected due to instrumental problems. For this reason, additional tests for optimization were performed.

ii. Fine tuning of an MS/MS method for phosphoproteomics on TimsTOF Pro

For this second experiment, previously tested methods A, B and C were evaluated again. Two other new methods were added for fine tuning of the parameters, they are described in details in **Figure 65-(A)**. In addition, five replicates per method were injected here (instead of two in the previous test) in order to evaluate more finely the quantification performances of the methods. Additionally, we used phospho-enriched murine tissues freshly prepared, while for the first experiment, already enriched samples from a previous experiment that were kept at -80°C had been used.

(A)	Ion mobility window (1/K0)	Collision energy range (eV)	Slope	Accumulation time (ms)	Number of PASEF scans	Cycle time (s)
Method A	0.7 → 1.25	22 → 62	Stepwise	166	10	1.89
					8	1.55
					6	1.2
Method B	0.7 → 1.25	22 → 62	Linear	166	10	1.89
					8	1.55
					6	1.2
Method C	0.6 → 1.6	20 → 80	Stepwise	100	10	1.17
Method E	0.6 → 1.6	20 → 80	Stepwise	166	10	1.89
Method F	0.7 → 1.4	22 → 62	Linear	166	10	1.89
					8	1.55
					6	1.2

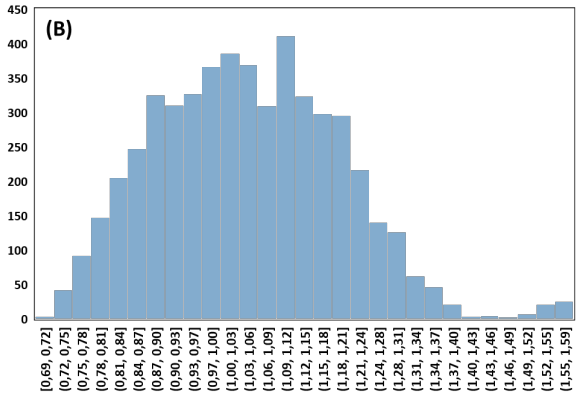


Figure 65: (A) Table describing the 5 different methods used with their respective ion mobility range, collision energy range, slope, accumulation time, number of PASEF scans and cycle time (B) Ion mobility distribution of the precursors of the phosphopeptides identified with method D**.**

Method E is the same as C except that it has a longer accumulation time (166 ms instead of 100 ms). Method F was designed to optimize the ion mobility window. Indeed, we looked at the ion mobility of the precursors of identified phosphopeptides of the method D of the first test, for which the ion mobility window was set from 0.6 to 1.6 1/K0. As shown by **Figure 65- (B)**, the distribution of the ion motilities is mainly centered between 0.7 and 1.4 1/K0. For this reason, we wanted to evaluate the effect of a reduce ion mobility window, from 0.7 to 1.4 1/K0.

The results of identification and quantification are represented in **Figure 66**.

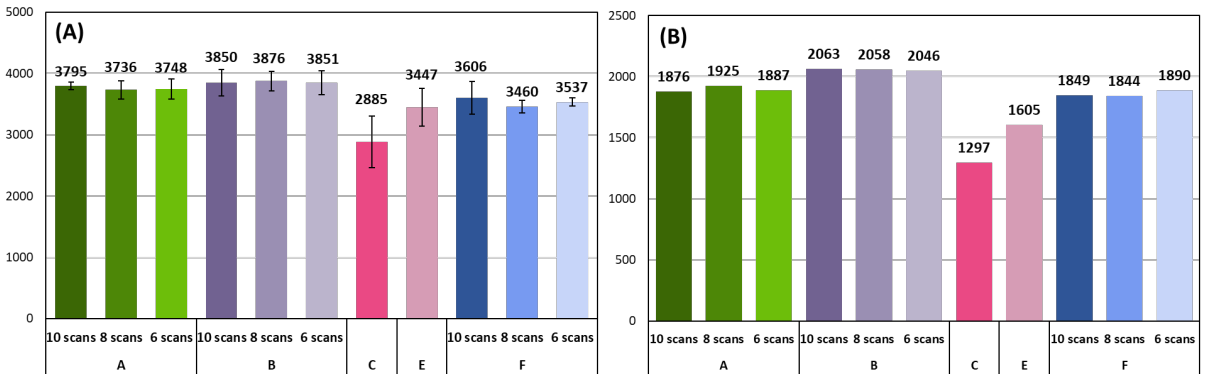


Figure 66: (A) Mean number of class I phosphosites identified by each method (B) Number of class I phosphosites quantified in 5 out of the 5 replicates by each method.

Both methods A and B display comparable results with around 3800 class I phosphosites identified, but method B slightly outperforms method A in terms of quantification. This suggests that a linear range of collision energy (method B) might be more adapted than a stepwise range (method A). The use of a reduced ion mobility window, even if adapted to the observed ion mobility of phosphopeptides precursors, does not allow for additional identifications. Indeed, method F (from 0.7 to 1.4 1/K0) quantifies on average 10% less class I phosphosites than the same method with a wider ion mobility range (method B, from 0.6 to 1.6 1/K0). Methods C and E both have the lowest identification (with respectively 2885 and 3447 class I phosphosites identified) and quantification (with respectively 1297 and 1605 class I phosphosites quantified) rates. Comparing the two of them, method E allows for a 19% increase in identification numbers and a 24% increase in quantification numbers, suggesting that a longer accumulation time (166 ms instead of 100 ms) is more adapted for the analysis. As for the previous test, the number of PASEF scans does not significantly impacts the identification and quantification results. It does however seems that for each method, 8 PASEF scans allows for the better results, suggesting that it is the balance between an efficient fragmentation and a short cycle time.

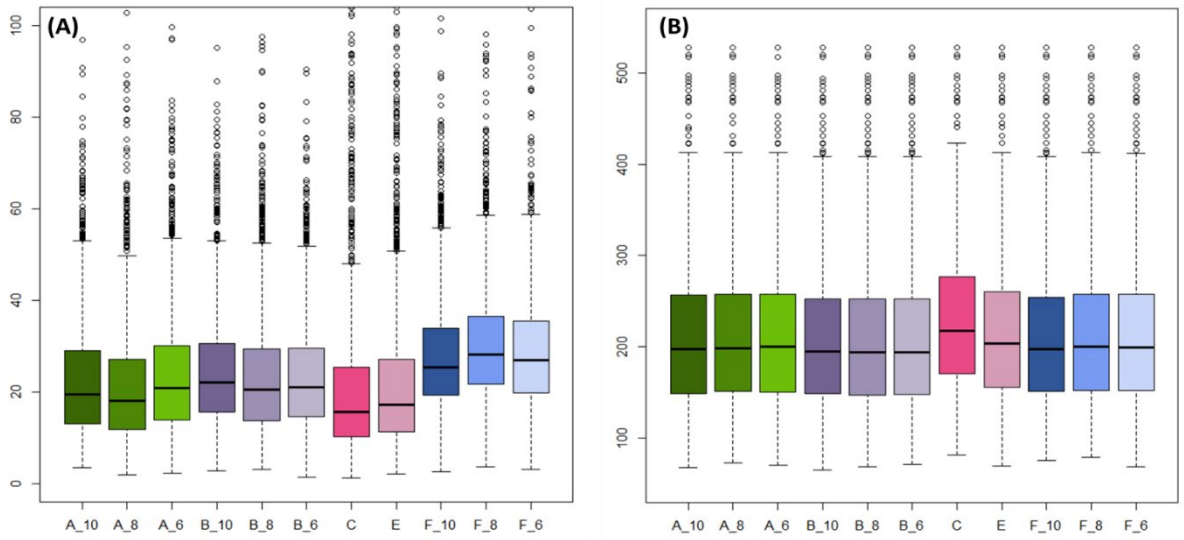


Figure 67: (A) Distribution of the CVs on class I phosphosites quantified in 5 out of 5 replicates for each method (B) Distribution of the scores of class I phosphosites quantified in 5 out of 5 replicates for each method.

However, when looking at the reproducibility of quantification results (**Figure 67 – (A)**), methods C and E seem to be the more robust with median CVs on class I phosphosites quantified in all replicates of 15.6% for C and 17.1% for E, whereas almost all other methods have median CVs > 20%. Indeed, as they are high collision energy methods (up to 80 eV), methods C and E generate more informative spectra and thus have a higher fragmentation efficienc, compared to lower collision energy methods (methods A, B and F, up to 62 eV). Moreover, when looking at the repartition of the scores of the best associated MS/MS spectrum, method C displays the highest median score (217.7), 10% higher in average than the one from any other "low –energy" method. Again, as the collision energies are higher in methods C and E, they generate better MS/MS spectra and thus scores for the best associated MS/MS spectrum are higher.

To conclude, our results show that higher collision energy methods, despite identifying and quantifying less phosphopeptides than other methods, allow for robust and reproducible quantification of those

phosphopeptides. A method with 8 PASEF scans seems optimal for the analysis as it is a good balance between an efficient fragmentation and a short cycle time. The cycle time comes out as one of the most –if not the most- important parameter to look for as longer cycle times decrease the precision of the analysis with not enough data to generate informative results. Comparing to our very first phosphopeptide analysis on the TimsTOF Pro on the bovine brain tissues (see **Figure 68**), our phosphopeptides identifications on mouse brain tissues were increased by more than 200% and quantification by 300%, highlighting the need for nanoLC-MS/MS method optimization to study phosphorylation.

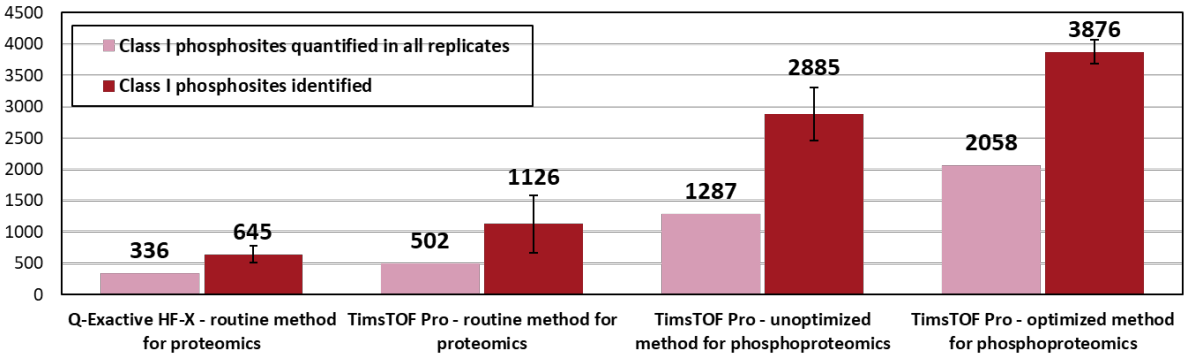


Figure 68: Identification and quantification of class I phosphosites through our different optimizations.

iii. Choice of the best suited LC gradient

Next, we wanted to evaluate the best chromatographic conditions for phosphopeptides' analysis. For this, we injected triplicates of our phosphopeptides-enriched samples of mouse brain tissues on the TimsTOF Pro, using 3 different LC gradients (30 minutes, 45 minutes, 80 minutes) and the MS method B with 8 PASEF scans. Identification and quantification results are displayed in **Figure 69**.

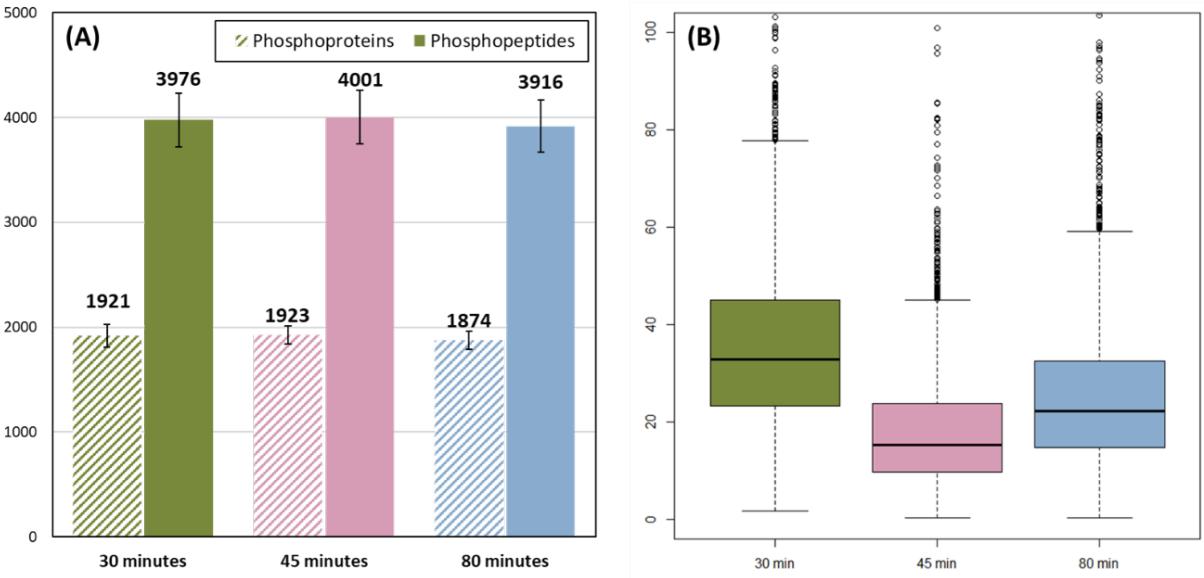


Figure 69: (A) Mean numbers of phosphoproteins and phosphopeptides identified for each gradient (B) Distribution of the CVs on the intensities of phosphosites quantified in 3 out of 3 replicates for each method.

In terms of phosphoproteins and phosphopeptides identification (**Figure 69 – (A)**), comparable results were obtained with the 3 conditions, with between 3916 phosphopeptides identified with the 80 minutes gradient to 4001 phosphopeptides identified with the 45 minutes gradient. However, looking into the quantification, and especially into the CVs on the intensities of phosphosites quantified in 3 out of the 3 replicates for each gradient, differences in the results appears. Indeed, the median CV for the 30 minutes gradient is 33%, 15% for the 45 minutes gradient and 22% for the 80 minutes gradient. The quantification seems much more reproducible with the 45 minutes gradient, the only one with a CV <20%.In conclusion, the best compromise between analysis time and quantification robustness is achieved with the 45 min gradient.

3. Development of a dia-PASEF method on the TimsTOF Pro

The dia-PASEF is a recent DIA acquisition method specific to TimsTOF Pro as it uses to its advantage the PASEF presented in the previous section. DIA should benefit from the PASEF with the increased acquisition speed of the instrument, the noise reduction, the improvement of the signal with the accumulation of ions and the better separation of co-eluting peptides from the LC thanks to ion mobility. Recently, a few publications have shown the application of dia-PASEF technology on phosphopeptides with promising results in terms of identification, quantification, throughput and sensitivity of analysis[133,241,334]. For these reasons, we decided to investigate the potential of this innovative method, and thus developed a dia-PASEF pipeline for the analysis of phosphopeptides and evaluated its performances.

i. Test 1: evaluation of isolation window width and accumulation time

Mouse brain tissue samples enriched in phosphopeptides and four replicates were injected per method on the TimsTOF Pro using a 45 minutes gradient. Generated data were analyzed using Spectronaut (Biognosys). The four different methods tested are described in **Figure 70 – (A)**. For all methods in this experiment, ion mobility range was set from 0.7 to 1.4 1/K0 and mass ranges from 400 to 1400 Da.

The first parameter to assess was the optimal isolation window width. Indeed, narrower isolation window results in less complex spectra but on the other hand increases the total number of windows needed to cover the whole mass range and thus the cycle time. As for DDA method optimization, we also evaluated two different accumulation time settings (100 ms and 166 ms).

	Isolation window width (Da)	Accumulation time (ms)	Cycle time (s)
Method 1	25	100	1.59
Method 2	30	100	1.06
Method 3	25	166	1.75
Method 4	30	166	1.55

Figure 70: (A) Description and parameters of the 4 different dia-PASEF methods.

We first evaluated the performances of the different methods in terms of identification. As shown in **Figure 70 – (B)**, method 1 with an isolation window of 25 Da and 100 ms accumulation time displays the best results with highest numbers of both phosphopeptides (4458) and class I phosphosites (4563) identified. Using a wider isolation window of 30 Da decreases identifications by approximately 5%.

Comparing the different accumulation times, 100 ms allows for an average 14% increase in the number of phosphopeptides identified.

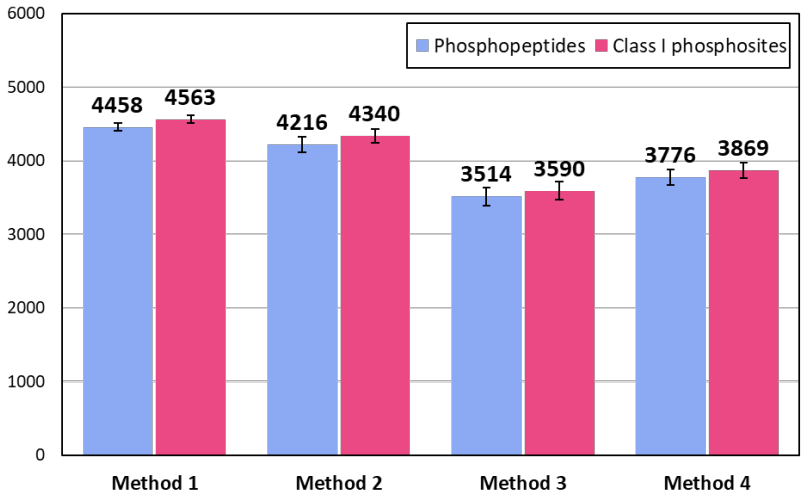


Figure 71: Mean number of phosphopeptides and class I phosphosites identified by each method.

For quantification, method A (25 Da window, 100 ms) displays the highest number of phosphopeptides quantified in all replicates with 2639 phosphopeptides quantified in all 4 replicates (**Figure 72– (A)**). However, if we apply a filter on the CV calculated on the intensities of the precursors of those quantified phosphopeptides, the drop of quantification for method A is of more than 80%. On the other hand, method 2 (30 Da window, 100 ms) quantifies 2850 phosphopeptides in all 4 replicates, 5% of them (1642 phosphopeptides) being quantified with a CV <20%. These results suggest that a 30 Da window width, while identifying and quantifying less phosphopeptides, allows for a better reproducibility of quantification. The same result is observed between methods 3 and 4: while they display similar results in number of quantified phosphopeptides, method 4 (30 Da window) has a higher proportion of quantified phosphopeptides with a CV < 20% (43% for method 4 opposed to 26% for method 3) and a lower median CV. This is further emphasized by the distribution of the CVs on the precursors' intensities for each method, represented in **Figure 72 – (B)**: method B has a median CV of 13.7% whereas method A has a median CV of 19.8%.

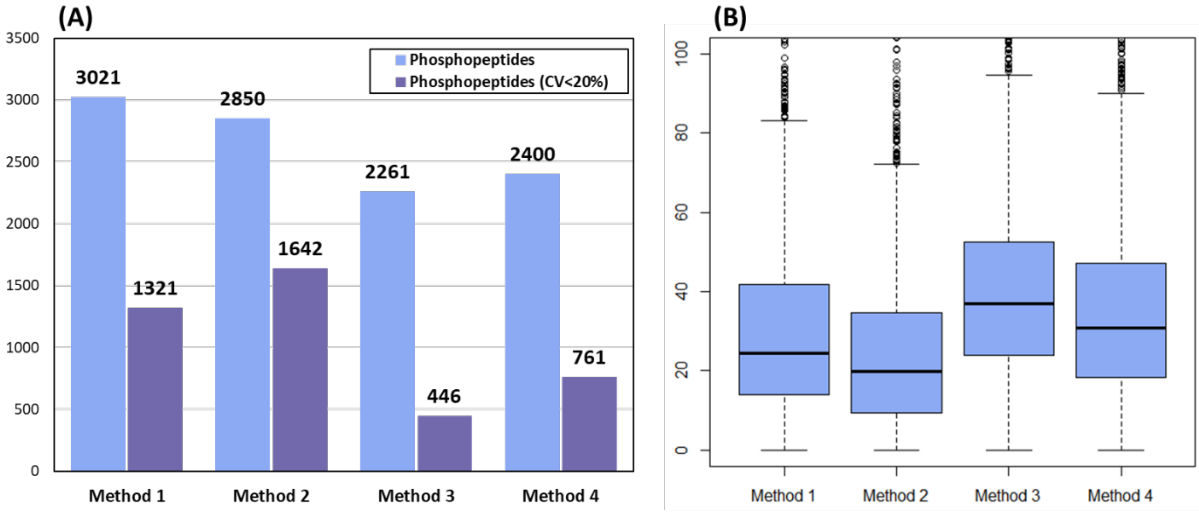


Figure 72: (A) Phosphopeptides quantified in 4 out of the 4 replicates and quantified in 4 out of the 4 replicates with a CV <20% for each method (B) Distribution of the CV on intensities of the quantified phosphopeptides.

Comparably to identification, both methods A and B (100 ms) outperform methods C and D (166 ms) for quantification. The drop in quantification with the CV <20% filter is quite massive for those two methods, with almost a 300% drop for method 3 and a 130% drop for method 4 (**Figure 72– (A)**). Both methods also have a median CV on the precursors' intensities higher than 20% with respectively a median CV of 33.1% for method 3 and 23.8% for method 4 (**Figure 72 – (B)**). These results highlight that a 100 ms accumulation time is optimal for phosphopeptides analysis by dia-PASEF. Indeed, a 166 ms accumulation time might be less effective as (i) it increases the cycle time and thus reduces the sensitivity of the analysis (ii) too much ions are accumulated and therefore the complexity of analysis increases. A 100 ms accumulation time thus seems the optimum balance between filling the TIMS cell to its capacity and achieving the highest ion mobility resolution possible. Most global proteomics dia-PASEF published studies and in the very few published work on phosphoproteomics dia-PASEF, a 100 ms accumulation time is systemically used[292,294,334].

This first experiment led to different conclusions: (i) a 100 ms accumulation time allows for robust identification and quantification of phosphopeptides and (ii) a 30 Da isolation window seems best adapted for reproducible quantification of phosphopeptides. However, these optimizations are not sufficient as many other dia-PASEF parameters might impact our phosphopeptides analysis. Hence, further optimizations were performed as followed.

ii. Test 2: optimization of the number of mobility steps and ion mobility range

For this second experiment, mouse brain tissue samples enriched in phosphopeptides were injected on the TimsTOF Pro in three replicates per method using a 45 minutes gradient and generated data were analyzed using Spectronaut (Biognosys). The six different methods tested and their corresponding parameters are described in **Table 6**. For all methods in this experiment, the accumulation time was of 100 ms and mass ranges were set from 400 to 1400 Da.

	Isolation window width (m/z)	Number of mobility steps	Cycle time (s)	Ion mobility range (1/K0)
Method 1	25	1	1.59	0.7 → 1.4
Method 2	30	1	1.06	0.7 → 1.4
Method 3	30	2	1.17	0.7 → 1.4
Method 4	30	3	1.38	0.7 → 1.4
Method 5	30	1	1.17	0.8 → 1.35
Method 6	30	2	1.17	0.8 → 1.35

Table 6: Description and parameters of the 6 different dia-PASEF methods**.**

For this second experiment, we assessed again the isolation window width by comparing a 25 Da window (method 1) to a 30 Da window (method 2). Except for method 1, all other methods were set with a 30 Da isolation window as first experiment suggested that it allowed a more robust quantification of phosphopeptides. We also evaluated the impact of different number of mobility steps. As represented in **Figure 73**, two possibilities are available when setting up a dia-PASEF experiment. In **Figure 73 – (A)**, the m/z range is split in consecutive fixed isolation windows but only one ion mobility step is used. Alternatively, multiple mobility windows can be used, as shown in **Figure 73 – (B)**, meaning the ion mobility range will also be split into windows (usually from 1 to 3 maximum). This can be useful to decomplexify the precursors so that less co-eluting peptides should be observed.

However, as the overall number of steps is increased, we risk increasing the variability but also the cycle time and thus the quantitative precision.

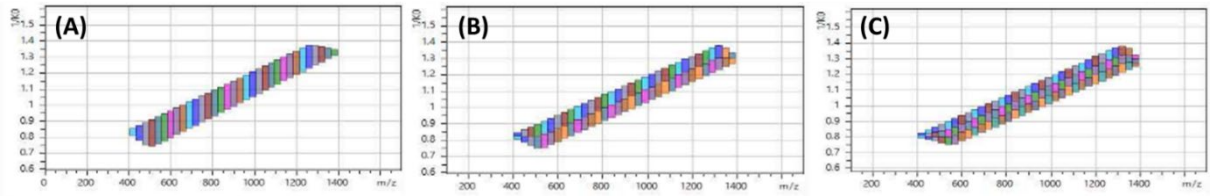


Figure 73: (A) A dia-PASEF polygon with 1 mobility step, (B) with 2 mobility steps and (C) with 3 mobility steps.

In addition, we evaluated a smaller ion mobility range, more adapted to phosphopeptides. Indeed, we looked at the ion mobilities of the precursors of identified phosphopeptides in the first experiment, and its distribution is represented in **Figure 74**. While an ion mobility range of 0.7 to 1.4 1/K0 was set, most precursors appear to have an IM value between 0.8 and 1.35. For this reason, we evaluated ***α*** method with an ion mobility range from 0.8 to 1.35 1/K0, optimally cover the area occupied by phosphopeptides and more adapted the to the peak density (method with 1 mobility step and method with 2 mobility steps).

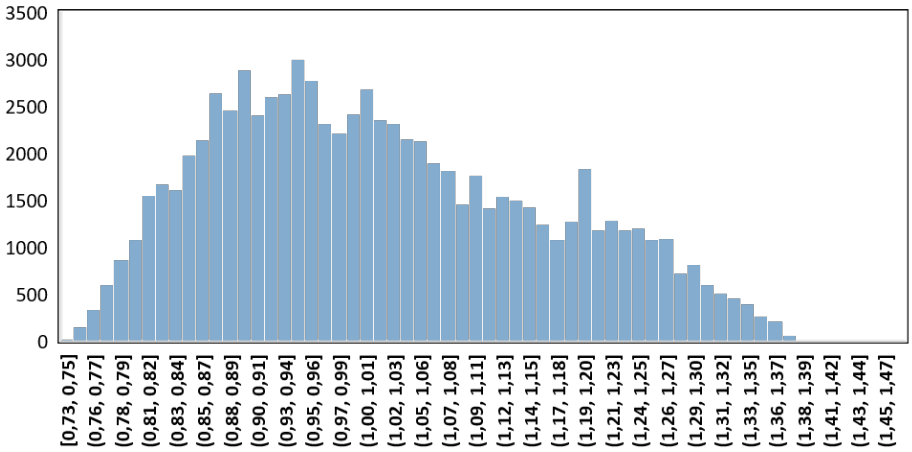


Figure 74: Distribution of precursors' ion mobility for identified phosphopeptides of the 1st experiment.

Identification results for this second experiment are presented in **Figure 75.** Consistently with what was observed in the first experiment, we confirm here that better results are obtained using a 30 Da isolation window as method 2 allows identifying more than 700 additional class I phosphosites compared to method 1. Results between method 2, 3 and 4, which have all the same parameters but increasing numbers of mobility steps, are similar as all the three methods identify between 7236 and 7518 phosphopeptides. It seems that the number of mobility steps does not have a great impact on phosphopeptides identification. This is further shown when comparing method 5 to method 6 with respectively 1 and 2 mobility steps. The smaller ion mobility window (from 0.8 to 1.35 1/K0, methods 5 and 6) appears to give lower identification results as they identify on average 20% less phosphopeptides compared to the same method with a wider ion mobility range (0.7 to 1.4 1/K0).

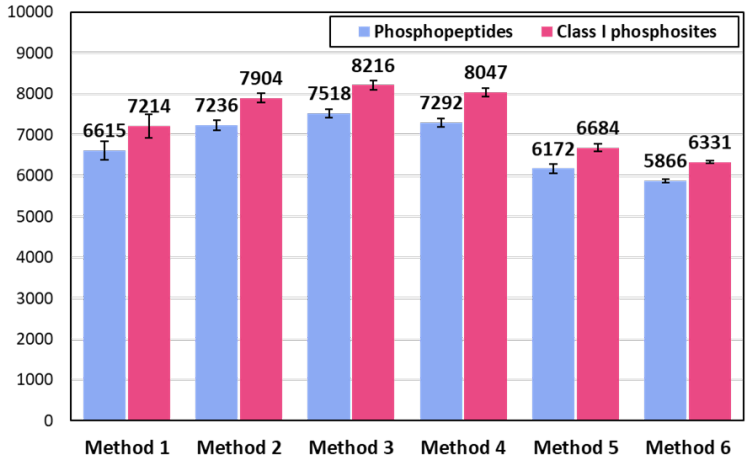


Figure 75: Mean number of phosphopeptides and class I phosphosites identified by each method.

As for identifications, methods 2, 3 and 4 display similar results for quantification with between 4138 to 4267 phosphopeptides quantified in all 3 replicates (**Figure 76 – (A)**). However, when applying a 20% filter on the CV of the quantified phosphopeptides, a clear distinction appears as method 3 (2 mobility steps) and method 4 (3 mobility steps) lose more than 50% of their quantified phosphopeptides. On the other hand, method 2 (with 1 mobility step) still robustly quantifies (with a CV lower than 20%) 2891 phosphopeptides, compared to 4226 without the CV filter. This suggests that while an increased number of mobility steps has no impact on the number of phosphopeptides identified and quantified, it greatly reduced the reproducibility of quantification. This is further shown when looking at the global distribution of the CVs of the different methods, where method 2 has a median CV of 14% compared to 26% and 25% (**Figure 76 – (B**)) respectively for method 3 with 2 mobility steps and method 4 with 3 mobility steps. This decreased reproducibility of quantification with the increase in the number of mobility steps is linked to the increasing cycle time, leading to poor quantitative precision.

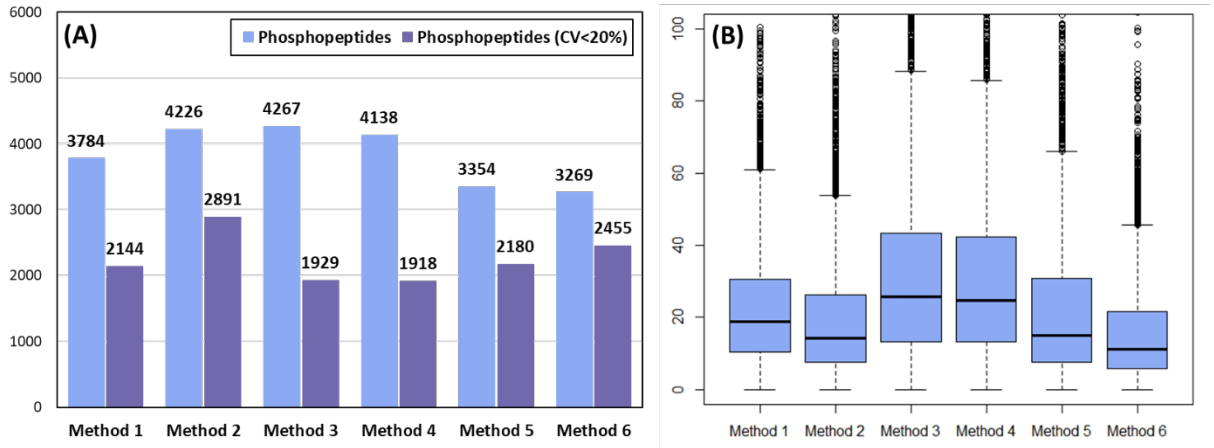


Figure 76: (A) Phosphopeptides quantified in 3 out of the 3 replicates and quantified in 3 out of the 3 replicates with a CV <20% for each method (B) Distribution of the CV on intensities of the quantified phosphopeptides.

Surprisingly, for the 0.8 to 1.35 1/K0 methods (methods 5 and 6), the opposite trend is observed. Both methods display comparable results in both numbers of phosphopeptides identified and quantified. However, after the CV <20% filter, method 6 (with 2 mobility steps) allows the quantification of almost 300 additional phosphopeptides. Method 6 also displays a lower median CV (11%) compared to method 5 (15%), suggesting that 2 mobility steps allows for more robust quantification when using a

smaller ion mobility window. These comparable results between the two methods can be explained as the cycle time between the two methods is the same, thus the addition of a 2nd ion mobility step does not impair the quantification.

Comparing our phosphoproteomics dia-PASEF results to the literature is not an easy task as they are very few published studies so far on the application of dia-PASEF to phosphoproteomics. Additionally, most of them are using the py_diAID package[241], a Python package for DIA with an automated isolation window design. This package allows defining variable isolation windows in the m/z versus ion mobility plane that are adjusted to the expected precursor ion density of phosphopeptides, leading to an almost complete phosphorylated precursor coverage.

Using this algorithm, Skowronek *et al.* identified almost 20 000 phosphopeptides on 100 µg of stimulated HeLa cells[241]. These results are however not comparable to ours for many reasons : (i) they analyzed HeLa cells which are much less complex than brain samples (ii) they used py_diAID for optimal isolation design and (iii) they used a library based approach while we used a library free approach to analyze our data.

Similarly, Oliinnyk *et al.* identified almost 9 000 class I phosphosites on 20 µg HeLa cells digest[133]. They used Spectronaut in its directDIA mode, but also used py_diAID algorithm to set up their isolation windows. On fungus samples, using isolation windows of 25 Da and 15 dia-PASEF scans per cycle, a study identified around 6000 phosphopeptides[334]. Data was however analyzed using DIA-NN with a spectral library.

Finally, from 20 µg of HeLa starting material, using latest developed µPhos protocol for phosphopeptide enrichment and py_diAID package, Oliinyk *et al*. identified more than 20 000 phosphopeptides[216]. These results are again not comparable to ours, as too many parameters differ from our experiment.

4. Summary of the improvements achieved via our phosphoproteomics method development

Through all MS/MS optimizations performed, we were able to greatly improve our phosphoproteomics analysis. Indeed, we started from scratch with phosphopeptides injections on a Q-Exactive HF-X with a MS/MS method not adapted at all to phosphoproteomics. Thanks to multiple optimizations on the TimsTOF Pro, first in DDA then in DIA modes, we were able to increase by a factor of 12 the phosphosites identified and by a factor of 14 the phosphosites quantified (**Figure 77**). Moreover, while identifications were increased, the time of analysis was reduced from an original 90 min gradient to a final 45 min gradient. By decreasing the time of analysis, we allow for high throughput phosphoproteomics.

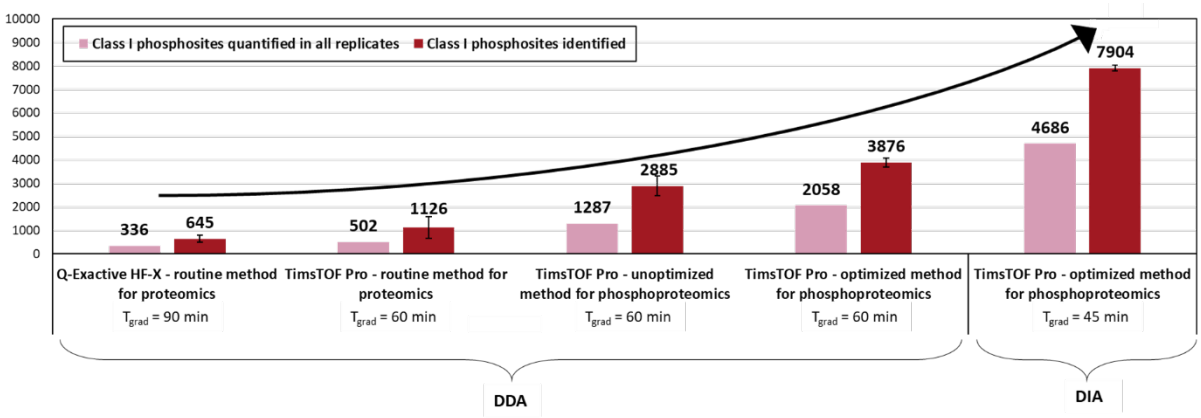


Figure 77: Summary of the different optimizations performed during my PhD, which led to an increase in phosphosites class I identification by a factor of 12.

Moreover, thanks to the additional IM mobility separation on the TIMS, co-eluting phosphopeptides that are isobaric can be resolved. **Figure 78** shows examples of co-eluting phosphopeptides, with the same sequence but different phosphorylation sites, which were separated based on their different ion mobilities. These co-eluting phosphopeptides thus generate two discrete MS/MS spectra that will be confidently assigned to different phosphopeptides by the search engine. This separation would not have been possible on a conventional non-TIMS mass spectrometer.

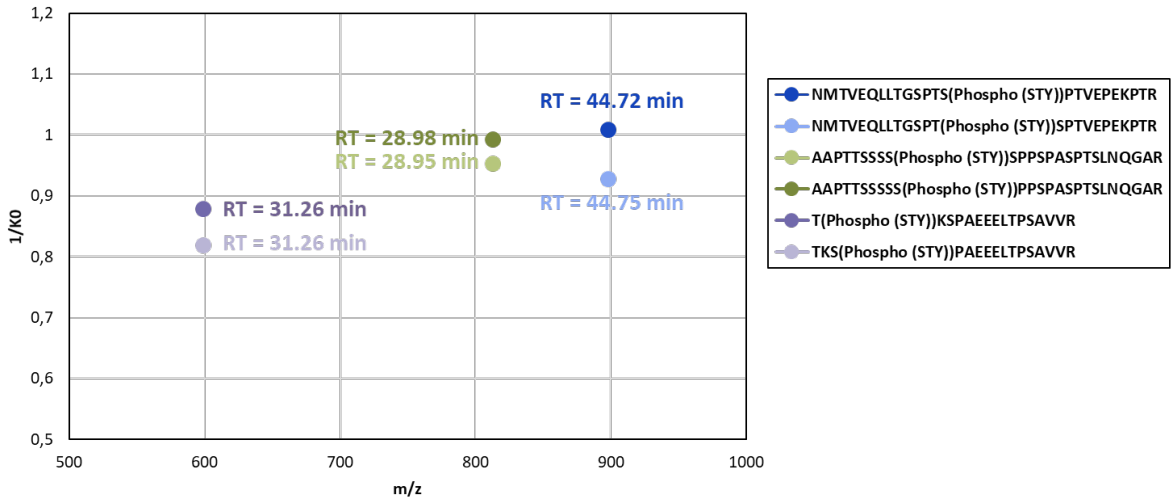


Figure 78: Co-eluting isobaric phosphopeptides that differ only by the phosphorylation localization site are separated by TIMS**.**

In addition, comparing DDA and DIA results, we were able to also greatly reduce the percentage of missing values across samples. As represented in **Figure 79**, while with DDA we have an overall percentage of missing values (MV) of almost 66%, this percentage of MVs is reduced by almost twice in DIA. These results highlight the known stochasticity of DDA, further enhanced in phosphoproteomics. Similar amount of MV have been obtained on a phosphoproteomics dataset by Weng *et al.* on T-cells samples, with around 40% of MV in DDA compared to 30% in DIA[382].

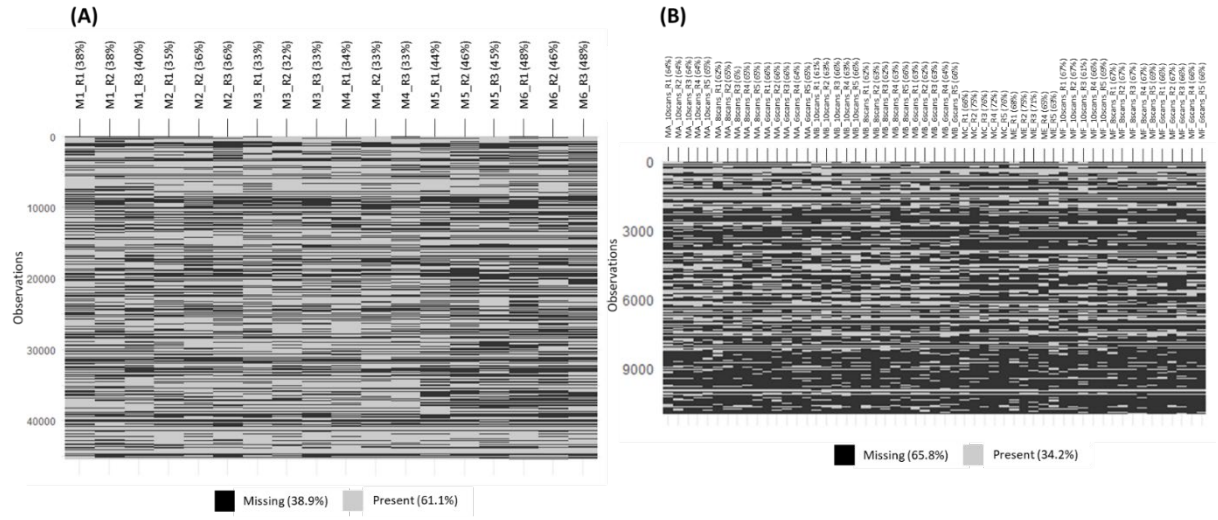


Figure 79: Representation of the amount of missing values across phosphopeptides identification for (A) DIA dataset and (B) DDA datasets, from method optimization on the TimsTOF Pro. MX_RX (X%) indicates that replicate X of method X has X% of missing values.

Strong of these optimizations and results, future high throughput phosphoproteomics analysis can now be performed on the TimsTOF Pro, either in DDA or DIA mode, and generate thousands of robust phosphosites' identification and quantification.

Chapter 3: Evaluation of different data treatment pipelines for phosphosites identification, quantification and localization

1. Benchmarking of different pipelines for DDA phosphoproteomics data analysis

Data analysis in phosphoproteomics is a challenging step as in addition to identify peptides, phosphorylation sites need to be localized. Identification and quantification of phosphorylation sites can be performed by various algorithms, which also generate various scores to evaluate the reliability of the localization of the phosphorylation. However, scores returned by different algorithms are not always directly comparable. For this reason, various software for data treatment were compared. The different combinations evaluated are summarized in **Table 7**.

Software	Proline	MaxQuant	Proteome Discoverer	Proteome Discoverer
Search engine	Mascot	Andromeda	Mascot	Mascot + MS Amanda
Validation	Target Decoy Proline	Target Decoy MaxQuant	Percolator PD	Percolator PD
Localization	MD-Score	PTM-Score	PhosphoRS	PhosphoRS

Table 7: Algorithms used to search, validate and localize phosphosites.

The evaluation of the different pipelines for data analysis was performed on triplicates of bovine brain tissues extracted with Laemmli-like buffer and analyzed on Q-Exactive HF-X platform (see **Chapter 1: Development of a high throughput and automated phosphoproteomics sample preparation** workflow), and results are described in the following paragraphs.

i. Definition of a phosphorylation site depending on the software

The definition of a phosphorylation site is specific and software-dependent. In general, a phosphorylation site refers to the localization of an amino acid in the peptidic sequence bearing a phosphorylation. The notion of site gives more information than the phosphopeptide as one peptide can bear multiple phosphorylations and multiple peptides can bear the same phosphorylation site. This is highlighted in **Figure 80**, when for example a peptide is miss-cleaved or also bears another modification.

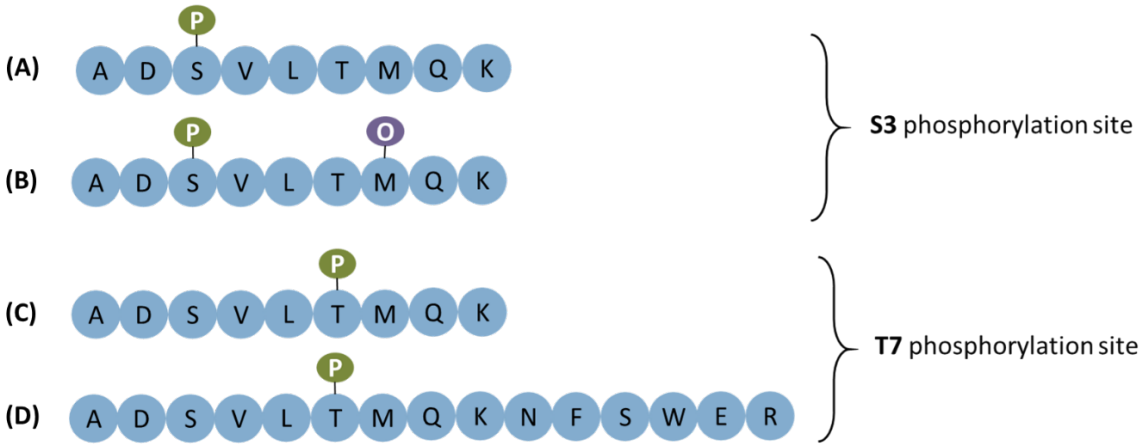


Figure 80: Schematic representation of different phosphorylation sites. A and B: the peptides have the same sequence and bear the same S3 phosphorylation site but the second one also has an oxidation in M7. A and C: the peptides have the same sequence but not the same phosphorylation site. C and D: they both have the same phosphorylation site and partially the same sequence but peptide D is miss-cleaved.

The functional effect of phosphorylation can be site-dependent meaning that they happen only if the phosphorylation is on a specific amino-acid or sometimes on multiple amino-acids of the same protein. It is thus important to be able to correctly localize the phosphosite and have information on its multiplicity. All studied software allow for identification at the phosphosite level. For MaxQuant, thanks to the "PhosphoSites" file that can be then processed through Perseus with the option "expand sites tables" to have access to the site information. For Proline, it is through "modification sites" or "modification clusters" tabs, and for Proteome Discoverer, through "Modification Sites" tab. However, each software has its own definition of a phosphosite, and they are illustrated in **Figure 81**.

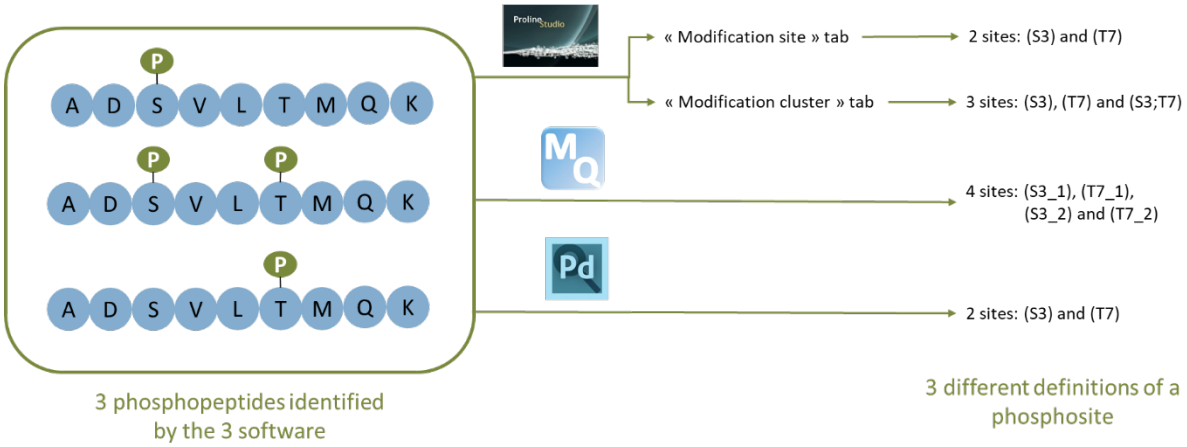


Figure 81: Schematic representation of the definition of a phosphorylation site depending on the software.

As showed in **Figure 81**, the different software also handle differently the multiplicity of the site, meaning if the peptide is mono- or multi-phosphorylated. In Proteome Discoverer or in the "modification cluster" tab from Proline, which have both the same site definition, we do not have access to the multiplicity information. On the other hand, it is accessible both in Proline in the "modification site" tab and in MaxQuant. However, keeping the multiplicity information gives rise to a redundancy as for one bi-phosphorylated site, two different lines are created in the output file. To

compare our results between the different software at the phosphosite level, we thus decided to get rid of the redundancy and chose the Proteome Discoverer definition of a phosphosite. As explained earlier, the multiplicity of a site can be a crucial information, and thus the ability to access this information can also be a criteria of evaluation between the software.

ii. Identification results

Shared peptides are also handled differently from one software to another. Both Proline and Proteome Discoverer attribute peptide/phosphorylation sites to the different isoforms of a protein. This creates a redundancy and leads to a bias, as one single PSM will wrongly identify multiple sites in multiple proteins. As displayed in **Figure 82**, DIESQVNKLR peptide, which is phosphorylated on serine 6, is associated to 8 different proteins, all myosine's isoforms. This leads, in the case of Proline and Proteome Discoverer to eight lines in the output file, as if 8 different phosphorylation sites were identified. On the other hand, MaxQuant groups all protein isoforms for a same PSM in a protein group, overcoming the redundancy issue. To get rid of the redundancy in Proline and Proteome Discoverer and be able to compare the results from the different software all together, Charlotte BRUN developed[383] a Python (v 3.8) script to attribute the phosphorylated peptide to the protein with the highest Mascot/MS Amanda score.

Proline

Protein	Peptide	Modification (site)	Localization probability	Score
XP_026376784.1	ADIAESQVNKLR	Phospho (S6)	100.00 %	34.81
XP_026376792.1	ADIAESQVNKLR	Phospho (S6)	100.00 %	34.81
XP_026342191.1	ADIAESQVNKLR	Phospho (S6)	100.00 %	34.81
XP_026376788.1	ADIAESQVNKLR	Phospho (S6)	100.00 %	34.81
XP_026342257.1	ADIAESQVNKLR	Phospho (S6)	100.00 %	34.81
XP_026376789.1	ADIAESQVNKLR	Phospho (S6)	100.00 %	34.81
XP_026376787.1	ADIAESQVNKLR	Phospho (S6)	100.00 %	34.81
XP_026376791.1	ADIAESQVNKLR	Phospho (S6)	100.00 %	34.81

Proteome discoverer

Modification	Amino acid	Position in the peptide	Peptide	Protein
Phospho	S	6	ADIAESQVNKLR	XP_026376784.1
Phospho	S	6	ADIAESQVNKLR	XP_026376791.1
Phospho	S	6	ADIAESQVNKLR	XP_026376789.1
Phospho	S	6	ADIAESQVNKLR	XP_026376787.1
Phospho	S	6	ADIAESQVNKLR	XP_026376792.1
Phospho	S	6	ADIAESQVNKLR	XP_026342191.1
Phospho	S	6	ADIAESQVNKLR	XP_026342257.1
Phospho	S	6	ADIAESQVNKLR	XP_026376788.1

MaxQuant

Amino acid	Multiplicity	Localization probability	Position in the peptide	Protein group	Protein	Peptide (localization probabiliy)
S	1	1	6	XP_026376785.1;XP_026	XP_026376792.1	ADIAES(1)QVNKLR

Figure 82: Example of how the different software handle shared peptides with the ADIESQVNKLR peptide

The identification results are displayed in **Figure 83**. Mascot search algorithm, used in both Proline and Proteome Discoverer pipelines, gives rise to much different results depending on the software. Indeed, Proteome Discoverer (with Mascot only) identifies almost twice more phosphosites and phosphopeptides than Proline. This increase in identifications is due to the use of Percolator as validation algorithm, compared to the target decoy approach of Proline. Locard-Paulet *et al.*[251] observed the same tendency of Percolator outperforming classical target decoy approaches when comparing different pipelines for phosphoproteomics data treatment. Percolator is an algorithm developed by Käll *et al.* in 2007, designed to increase confident peptides identification through semi-supervised machine learning[159].

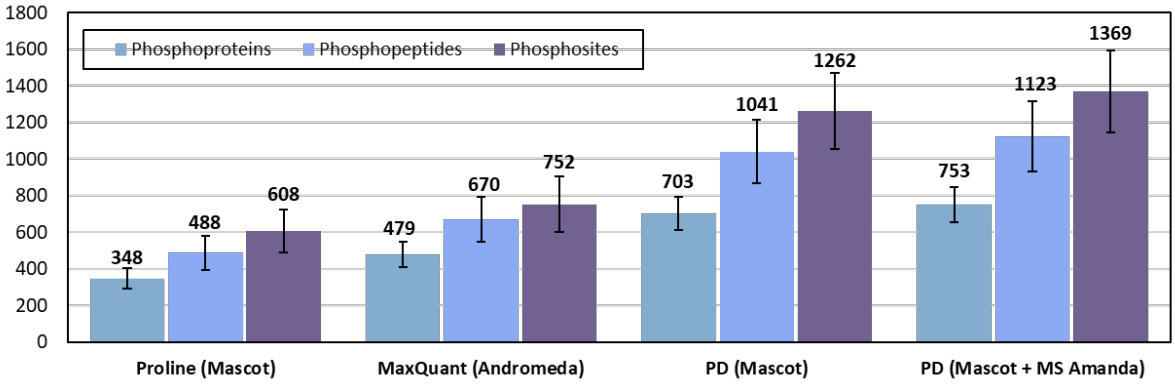


Figure 83: Phosphoproteins, phosphopeptides and phosphosites identification depending on the data treatment pipeline.

Highest numbers of identifications are obtained using the combination of Mascot and MS Amanda (**Figure 83**) with more than 1300 phosphosites, 8% more than using Mascot alone, highlighting the effect of the search algorithm. However, looking deeper into the data and into the spectra of some phosphopeptides identified uniquely by MS Amanda, it seems that those spectra are less informative than when using Mascot (**Figure 84**). We however only looked manually at a handful number of spectra, and thus this might not be a generality but only a trend.

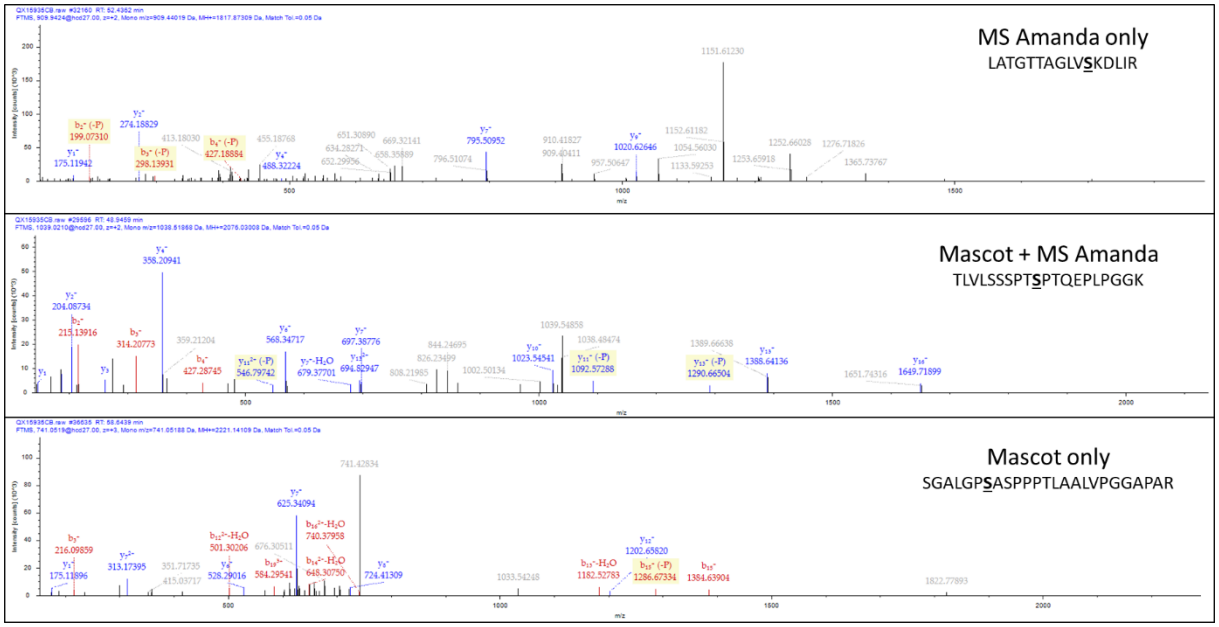


Figure 84: Spectra of a phosphopeptide identified with MS Amanda, with Mascot and with both Mascot and MS Amanda.

All these results display the great impact of the choice of both the search and validation algorithms to perform phosphoproteomics data analysis.

iii. Localization of the phosphorylation

As said previously, the correct localization of a phosphorylation on the peptide might be crucial for the understanding of biological processes, and the difficulty to correctly localize phosphosites has been widely discussed in the scientific community[224]. Another difficulty is added when we want to compare the performances of different localization algorithms, as most of them are tied to specific search engine. It is thus tricky to evaluate if observed differences are due to the localization algorithm or to the search engine. Several studies showed that, even when using synthetic peptides with known phosphorylation sites, the results obtained by the different algorithms are not directly comparable[251,384]. Here, we evaluated the performances of three different localization algorithms: MD-Score (Proline), PTM-Score (MaxQuant) and PhosphoRS (Proteome Discoverer). All those algorithms compute the localization probability of a phosphorylation on a peptide (between 0 and 1). In the case where the peptide only has one site that could be phosphorylated, the probability is 1. The usual threshold of probability is set up to 0.75 on the MaxQuant probability score, these phosphorylation sites are then defined as class I phosphosites. The localization probabilities obtained on our data by the different pipelines are represented in **Figure 85**.

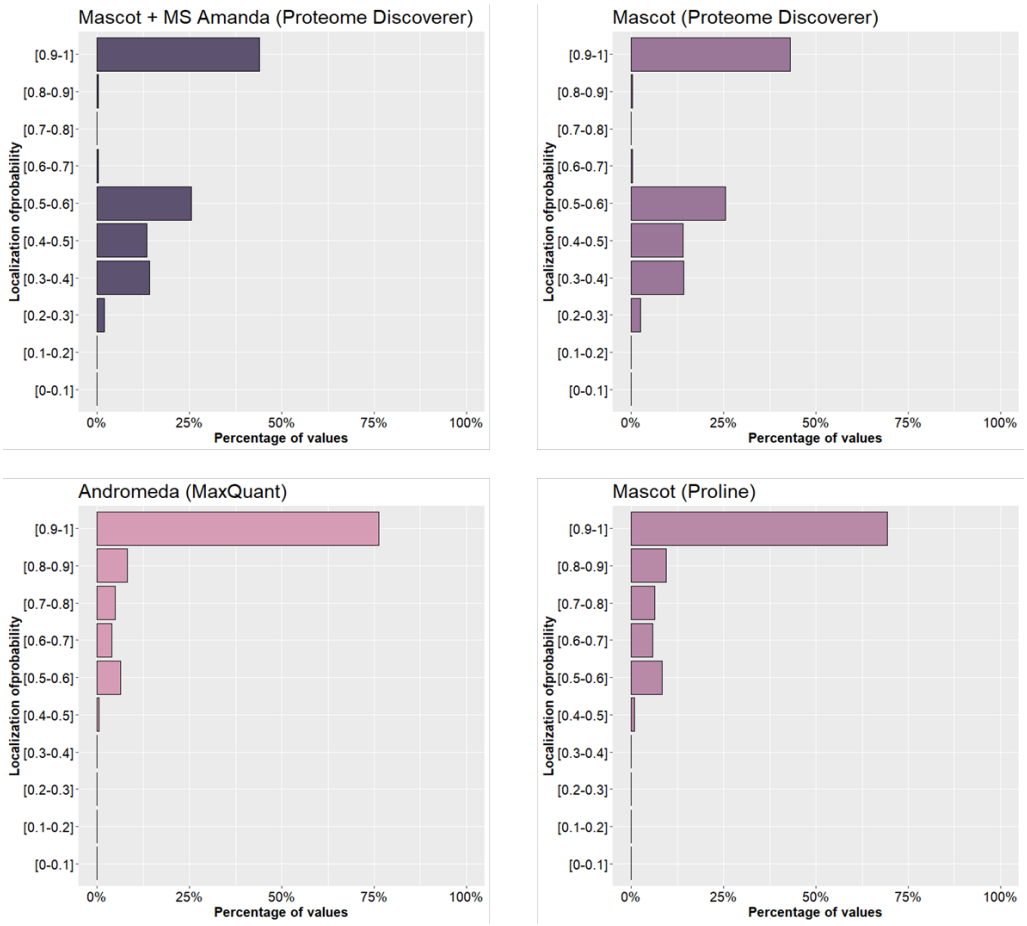


Figure 85: Frequency (in %) of the different localization probabilities for the identified phosphorylation sites depending on the pipeline.

MaxQuant's PTM Score displays the best results with more than 75% of localization score higher than 0.9. PhosphoRS algorithm (in Proteome Discoverer) however has the highest amount of low localization scores with more than 50% of them below 0.5. Surprisingly, almost no phosphorylation sites are localized with a probability between 0.6 and 0.9, the other half of identified phosphosites being localized with a score higher than 0.9. These results are nonetheless coherent with what Locard *et al.* reported in their pipelines comparison[251]. Proline MD Score displays similar results to the PTM Score, with 70% of localization score between 0.9 and 1, and more than 90% higher than 0.5.

iv. Global comparison of all pipelines

To have a general overview of the overall performances of the different pipelines, their main advantages and drawbacks are represented in **Table 8**. One need to keep in mind that these results highlight the performances of the software in the specific case of our analysis, and results might be different in another case.

	Proline (Mascot)	MaxQuant (Andromeda)	Proteome Discoverer (Mascot)	Proteome Discoverer (Mascot + MS Amanda)
+ features	• Site definition with or without the multiplicity • Scoring algorithm	• Site definition with the multiplicity information • Quantification at the site level possible	• Great number of identification (thanks to Percolator) • Scoring algorithm	• High number of identification by combining the two search engines • Scoring algorithm
- features	• Poor identifications due to a too stringent validation • No quantification at the site level	Scoring algorithm might not be stringent enough	• No multiplicity information at the site level • No quantification at the site level	• Additional identification seems to come from poor quality spectra • No multiplicity information at the site level • No quantification at the site level

Table 8: Main positive and negative features of the different pipelines

As previously explained, the definition of a site is crucial in phosphoproteomics, especially for quantitative phosphoproteomics. Indeed, if quantification is performed at the peptide level instead of the site level, a bias is introduced. Additional variable modifications or a missed-cleavage both can generate different phosphopeptides while they bear the same phosphorylation site, displaying the same biological phosphorylation event. While Proline and Proteome Discoverer both allow phosphosites identification, they do not give access to any quantitation information at the site level. To date, MaxQuant is the only software that allows for phosphosite quantification by summing the intensities of all the phosphopeptides carrying a specific phosphosite, thus containing miss-cleaved peptides and peptides with additional modifications. However, because MaxQuant keeps the multiplicity information, this quantification will led to two values of quantification for one same multi-phosphorylated peptide. This downside can lead to a bias if statistical analysis is performed afterwards. For these reasons, MaxQuant was used in all the phosphoproteomics work presented here for phosphosites identification, validation, localization and quantification.

2. Spectronaut and DIA-NN software for dia-PASEF data treatment

As for DDA data processing, a plethora of software tools for DIA data processing exist. The choice of the software tool and of the spectral library has a great impact on the resulting data. While many publications review the performances of those tools for DIA proteomics[318,319,385,386], only a handful have done the same for DIA phosphoproteomics or even for dia-PASEF phosphoproteomics[387,388]. For these

reasons, we decided to evaluate the performances of two of the most common DIA software (Spectronaut and DIA-NN) on our phosphoproteomics dia-PASEF dataset. Spectronaut version 17.1 and DIA-NN version 1.8 were used.

In some DIA phosphoproteomics workflows, a tailor made spectral library is constructed prior to data processing as it has shown to achieve higher coverage and quantification compared to library free approaches[10]. However, this usually comes at the price of time, samples and effort consuming library building. Indeed, building a high-quality spectral library for DIA phosphoproteomics usually requires DDA analysis of extensively pre-fractionated of repeatedly injected samples[322]. It is therefore much more accessible to implement a library free workflow for DIA phosphoproteomics. Library free approach has also extra advantages when it comes to phosphoproteomics. When building a DIA phosphoproteomics library, rare phosphorylation sites may get hindered by other more abundant ones. Additionally, a phosphorylation must be present in the library in order to be considered, while in library free approach, all possible phosphorylation sites combination for a given peptide are considered.

For all the reasons listed above, we decided in the following experiments to compare both software in directDIA/library-free modes.

i. Identification and quantification results

First of all, we evaluated the identification and quantification performances of both software. As displayed in **Figure 86 – (A)**, Spectronaut identifies in average between 5866 and 7518 phosphopeptides, while DIA-NN identifies between 6930 and 8994 phosphopeptides. Across the different methods, DIA-NN allows an overall increase in phosphopeptides identifications of 19%. As for quantification performances, DIA-NN unanimously outperforms Spectronaut (in **Figure 86 – (B)**), with up to 5527 phosphopeptides quantified robustly in 3 out of 3 replicates and with a CV < 20%.

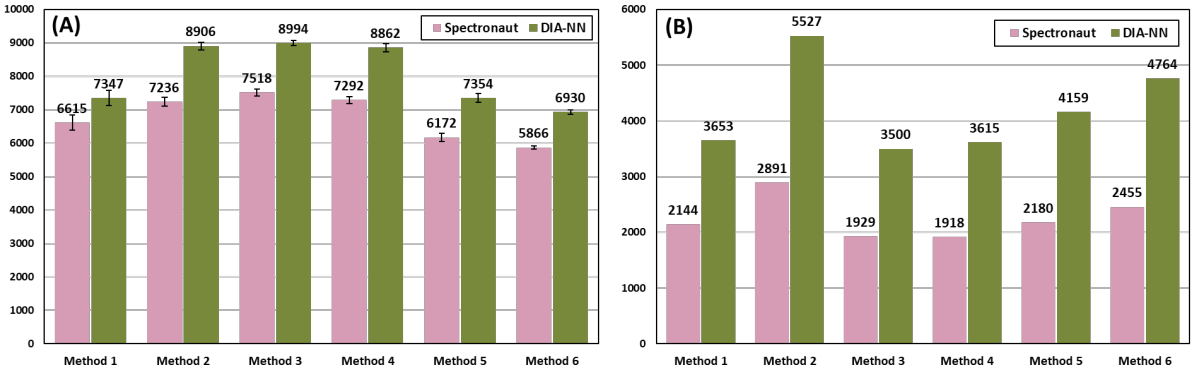


Figure 86: Comparison of Spectronaut and DIA-NN performances in terms of (A) average number of identified phosphopeptides and (B) number of phosphopeptides quantified in 3 out of 3 replicates with a CV < 20%.

In **Figure 87 – (A)** is represented the distribution of the CVs at the precursor level of all phosphopeptides quantified in the three replicates across all methods. We can notice that while the number of quantified phosphopeptides is much higher in DIA-NN than Spectronaut, the difference in the robustness of quantification is much less significant. The distribution of CV is slightly more spread for Spectronaut, but the overall median CV for Spectronaut is 18% while it is of 17% for DIA-NN. By method, the same observations can be drawn (**Figure 87 – (B)**). For most methods, Spectronaut's distribution of CVs is mildly wider, but no significant differences are observed in terms of median CV.

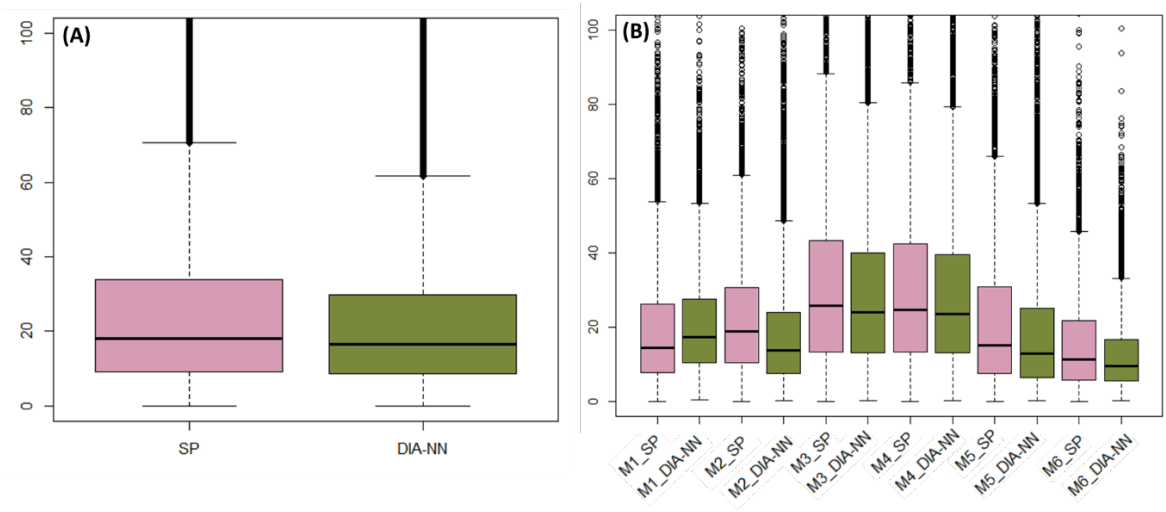


Figure 87: (A) Boxplots representing the distribution of the coefficients of variation (CVs) at the precursor level for all phosphopeptides quantified in 3 out of 3 replicates, for both Spectronaut (SP) and DIA-NN (B) Boxplots representing the distribution of the coefficients of variation (CVs) at the precursor level for all phosphopeptides quantified in 3 out of 3 replicates across the different methods, for both Spectronaut (SP) and DIA-NN.

In a recent publication Lou *et al*[388] also compared DIA-NN (v.1.8.1) and Spectronaut (v.17) for DIA phosphoproteomics analysis on both a TimsTOF Pro and a Q-Exactive HF-X using treated and stimulated cell lines enriched in phosphopeptides. Interestingly, while DIA-NN slightly outperformed Spectronaut for phosphopeptides identifications on the TimsTOF Pro, the opposite trend was observed on the Q-Exactive HF-X. While they did not share numbers of phosphopeptides quantified, they also obtained a similar quantification reproducibility between the two software on the dia-PASEF data.

On TiO_2 enriched THP1 cells analyzed by DIA on an Orbitrap Fusion Lumos, Wen *et al.*[387] showed the higher performances of directDIA Spectronaut (v.17.1) compared to DIA-NN (v.1.8.1), both for phosphopeptides and phosphosites identification. However, when analyzing enriched murine fibroblaste cells by dia-PASEF, the considerable advantage of Spectronaut compared to DIA-NN disappeared. Indeed, both software identified a similar number of phosphopeptides (around 12 000 identified phosphopeptides). They also highlighted that, for phospho-dia-PASEF data, directDIA workflow displayed a lower quantification reproducibility than the DIA-NN-based workflow.

Recently, Vashist *et al.*[389] presented results of the comparison of Spectronaut (v.17) and DIA-NN (v.1.8) on synthetic phosphopeptides analyzed by dia-PASEF. In this work, both software displayed similar identification performances as well as similar quantification CVs.

ii. Comparison of the phosphopeptides populations

We then investigated the different populations of phosphopeptides identified by the two software. As displayed in **Figure 88 – (A)**, Spectronaut surprisingly identifies a total number of phosphopeptides (11 829 phosphopeptides) higher than DIA-NN (9878 phosphopeptides). However, illustrated by **Figure 88 – (B)**, which displays the overlap between quantified phosphopeptides, DIA-NN allows the robust quantification of an additional 36% phosphopeptides compared to Spectronaut. These results, along with our previous identification and quantification results, suggest that while Spectronaut identifies more phosphopeptides, identifications and quantifications are sparser across the different methods and replicates. While DIA-NN identifies a lower total number of phosphopeptides, these phosphopeptides are identified and quantified robustly in most methods and replicates.

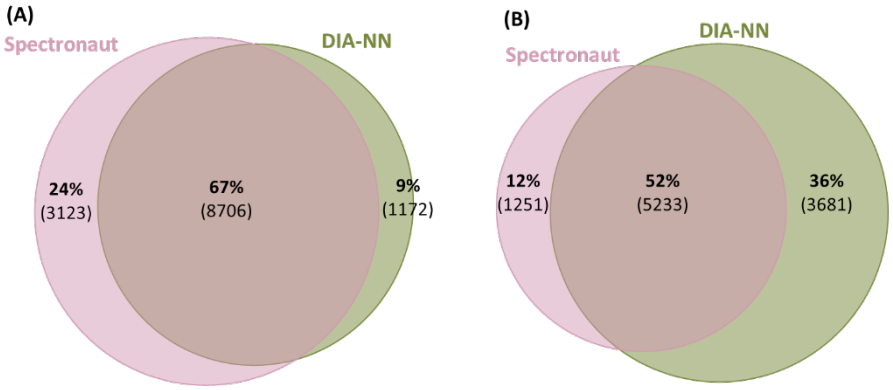


Figure 88: Overlap phosphopeptides between Spectronaut and DIA-NN (A) all identified phosphopeptides across the 6 methods (B) phosphopeptides quantified in all replicates in at least one method.

We then investigated the nature of the different populations of phosphopeptides. We looked into the additional 3681 phosphopeptides quantified only by DIA-NN. We wondered if those phosphopeptides were quantified in Spectronaut but solely did not passed the 0.1% p-value (precursor level) quantification filter. As represented in **Figure 89**, most of the DIA-NN quantified phosphopeptides are also found by Spectronaut but did not pass the quantification filter. This means that for a same set of peptides, DIA-NN seems to be less stringent in terms of validation than Spectronaut.

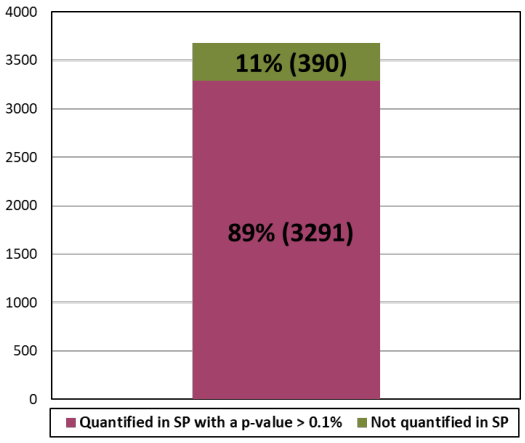
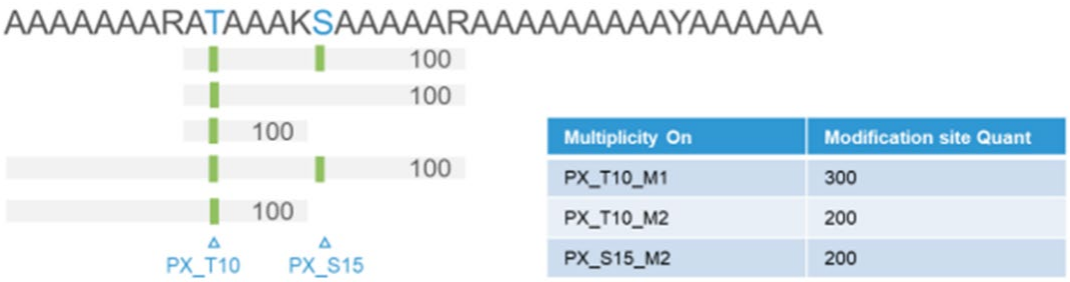


Figure 89: Percentage of phosphopeptides quantified by DIA-NN quantified by Spectronaut with a p-value > 0.1%.

iii. Definition of a phosphorylation site depending on the software

As already noticed for DDA software tools, the way of handling a phosphorylation site is software-dependent. As previously described, a phosphorylation site usually refers to the localization of an amino acid in the peptidic sequence bearing a phosphorylation. In Spectronaut, a phosphosite is defined as the combination of the protein carrying the site, the position of the site in the parent protein, the amino acid carrying the site and the multiplicity of the site. As explained in **1.Benchmarking of different pipelines for DDA phosphoproteomics data analysis ; i.Definition of a phosphorylation site depending on the software**, keeping the multiplicity information leads to an inherent redundancy for multi-phosphorylated sites. Spectronaut allows the quantification at the phosphosite level, thanks to an implemented algorithm developed by Bekker-Jensen *et al.*[10]. This algorithm's function is similar to the "expand site table" option in Perseus software for DDA analysis. As represented by **Figure 90**, the algorithm will sum (if "sum" is selected as "PTM consolidation" in Spectronaut's parameters) all identified peptides that correspond to this PTM site group and consolidate them into site quantities.



Multiplicity On	Modification site Quant
PX_T10_M1	300
PX_T10_M2	200
PX_S15_M2	200

Figure 90: Example of a quantitative site collapse of phosphorylated parent peptides, performed according to their multiplicity (Spectronaut user manual).

On the other hand, DIA-NN does not straightly give a site information, we only have access to a "Modified sequence" which is the combination of the peptide sequence with the modification inserted in the peptide sequence. While no redundancy is observed as we do not have the multiplicity information, another problem emerges with this definition of a phosphosite. In example 1 (**Figure 91**), the peptide is seen carrying the phosphosite on two different amino acids. They are therefore two distinct phosphosites, belonging to the same peptide. With the DIA-NN definition of a phosphosite, we keep the correct information as we will get two different lines for the two different phosphosites as followed: "AEEEGGS(Unimod:21)EEEGSDRSPQESK" and "AEEEGGEEEGSDRS(Unimod:21)PQESK". However, if as in example 2 (**Figure 91**), a phosphorylated peptide is miss-cleaved, a redundancy will appear. Indeed, as DIA-NN defines a phosphosites with its peptidic sequence, a same phosphosite on a miss-cleaved peptide will be considered as a different phosphosite. This will lead to the wrongful output of two different lines for one same phosphosite: "ALGLEES(Unimod:21)PEEEGK" and "ALGLEES(Unimod:21)PEEEGKAR". Additionally, DIA-NN does not give access to any quantitation information at the phosphosite level.

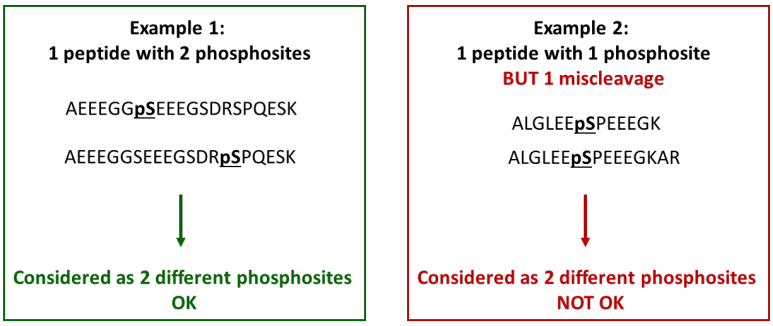


Figure 91: Two examples of phosphosites in DIA-NN.

iv. Overall comparison of the two software

To sum up, the main advantages and drawbacks of the two software are displayed in **Table 9**.

	Spectronaut directDIA v17.1	**DIA-NN** Library free v1.8.1
+ features	• Search time relatively fast • Allows for quantification at the phosphosite level • Lot of options for data visualization	• Highest numbers of phosphopeptides identified and quantified • Reproducibility of quantification
- features	• Sparse identifications across methods • Too stringent q-value calculations ?	• No quantification at the phosphosite level • No data visualization implemented • Overlong data search time

Table 9: Main positive and negative features of the different pipelines for phospho-dia-PASEF data treatment

One needs to keep in mind that this table results from the performances of the software observed in our specific context, *ie* the study protein phosphorylation in mouse brain tissues using dia-PASEF without spectral library on a TimsTOF Pro.

All the different results presented here highlight that there is no consensus yet on the better performing software for DIA phosphoproteomics data. Results vary depending on the type of sample (synthetic phosphopeptides or biological sample), the instrument used (classical DIA or dia-PASEF) and last but not least, on the version of the software used. Indeed, software are in constant evolution and new versions are frequently released. The use of one software or the other should thus be adapted to each specific project**.**

PART IV

Multi-omic analysis of axono-synaptic degeneration in motoneuron disease – application to the MAXOMOD project

Amyotrophic lateral sclerosis (ALS) is a rapidly progressive neurodegenerative motor neuron disorder causing progressing muscle weaknesses, muscle atrophy and cramps. These symptoms spread with the progression of the disease and usually lead to death due to respiratory failure within 3 to 5 years. Although being a rare disease, with an estimated incidence of 1.75-3 per 100 000 persons per year (10-12 per 100 000 in Europe), the lifetime risk to be affected is 1:400 in women and 1:350 in men[12]. In only a small amount of the ALS cases (10%) does the patient have a family history suggesting familial case of the disease (fALS). The majority of the cases are considered as sporadic (90%, sALS)[12]. ALS is a highly heterogeneous disease at the genetic level, with more than 30 genetic mutations that have been associated with it. Out of these, 4 gene variants account for 55% of fALS and more than 12% of sALS : SOD1, FUS, TARDBP and C9orf72[13,390].

ALS diagnosis is based on only clinical criteria and occurs relatively late as it takes from 8 to 15 months in average to confirm diagnosis since first symptoms' appearance[391]. Indeed, ALS symptoms at the onstage of the disease vary from patient to patient and are sometimes very similar to the symptoms of other diseases. Even though they are not yet fully integrated into standard clinical utilization, several biomarkers have been proposed to facilitate diagnosis and as potential therapeutic target. Neurofilaments levels in cerebrospinal fluid and serum is one of the most promising biomarker but seems however not to correlate with disease progression[190,392]. For the past two decades in most European countries, the only approved and licensed pharmacological treatment was the glutamate antagonist riluzole. It is shown to prolong patient survival by 3-6 months maximum, with various side effects including liver problems and nausea. It also seems to be only efficient on a subpopulation of patients[12,393]. Recently, the antioxidant edaravone has been approved for ALS treatment in US, Canada, Japan, South Korea and Switzerland. It is however not yet approved in the European Union because of the small size of patient panel, the short study duration, and the lack of improved survival data under edaravone treatment[12,394]. Even more recently, PB/Turso (co-formulation of sodium phenylbutyrate and taurursodiol) was approved by the FDA as it was shown to slow the decline rated associated with ALS and might provide survival benefit compared to placebo[395].

Because of its poor prognosis, rapid progression and limited therapeutic options, a better characterization of the onset events in ALS development is needed. Most studies on ALS focused only on an individual molecular subset such as transcripts[396], miRNA[397] or proteins[338,398], and have not yet been able to fully comprehend the disease. In this context, the European E-RARE MAXOMOD (Multi-omic analysis of axono-synaptic degeneration in motoneuron disease) consortium, which partially financed my PhD, was created. This project's aim was to develop and implement a large multi-omic investigation on both human (post-mortem brain tissues and cerebrospinal fluids) and mouse models samples to identify new disease-relevant pathways and biomarkers related to amyotrophic lateral sclerosis. The MAXOMOD project thus involves eight different teams, each of them with different tasks within the project:

- acquisition and preparation of the samples
- genomic, transcriptomic, miRNAomic and metabolomic analysis
- proteomic and phosphoproteomic analysis
- global multi-omic data analysis and integration with development of the FAIR-ALS data integration platform
- identification of molecular pathway targets for therapeutic validation and biomarkers target
- *in vitro* validation of molecular targets, pathways and pharmacological treatment

During my PhD, I was responsible for conducting all the proteomics and phosphoproteomics analyses of post-mortem brain tissues (N>100), mouse models (N>80) and human CSF samples (N>100). This resulted in the analysis of more than almost 400 samples (proteomics and phosphoproteomics included). First of all, proteomics analyses were performed on human post-mortem brain tissues and brain tissues from four different mouse models (SOD1, TDP43, FUS and C9). The optimized workflow for phosphoproteomics discussed previously in this manuscript (see **PART III- Chapter 1**) was then applied on the 80 mouse model samples. Then for CSF samples, a sample preparation workflow common for proteomics, phosphoproteomics and metabolomics analysis was developed. The efficiency and reproducibility of our high-throughput analysis of hundreds of samples was evaluated thanks to the implementation of different quality controls. Finally, the performances of an open modification search software (IonBot developed by the CompOmics group, Ghent, Belgium) were evaluated to improve identification and localization of phosphorylation sites.

Chapter 1: Proteomics and phosphoproteomics analysis of large cohorts of brain tissues

1. High throughput proteomics of large cohorts of mouse and human brain tissues

Protein aggregation of different proteins, such as SOD1, TDP-43 (encoded by gene TARDP), C9ORF72 or FUS in the brain and namely in the frontal cortex, is well reported as one of the central characteristic of ALS[13,14]. Therefore, the proteomic study of both human and transgenic mouse pre-frontal cortex (PFC) tissues may give the opportunity to identify disease-specific proteins that participate in key pathological processes and might be used as potential ALS biomarkers. Here, we worked with four established transgenic mouse models of ALS that each recapitulate different aspects of ALS, namely SOD1, TDP43, C9orf72 and FUS mouse models[399]. For each mouse model, we had 20 PFC samples, half of them from non-transgenic mice (wild type condition, WT) and the other half from transgenic (TG) mice. Among the two conditions, the same number of samples came from male and female mice. We thus had a total of 80 PFC mice samples. For human post-mortem brain tissues, frontal cortex samples came either from sALS patients (N = 52, male/female) and age-matched control without neurodegeneration symptoms (N = 50, male/female).

i. Sample preparation of mouse and human brain tissues

Sample preparation of brain tissues is especially crucial as brain tissues have a high lipid level (about one half of its dry weight) and these lipids might co-elute with analytes of interest[400,401]. For the proteomic analysis of both mice and human brain tissues, we therefore decided to go for an in-gel based (non-separated stacking gel) sample preparation, as it was at the time still the lab's reference method for tissue sample preparation[402] (**Figure 92**).

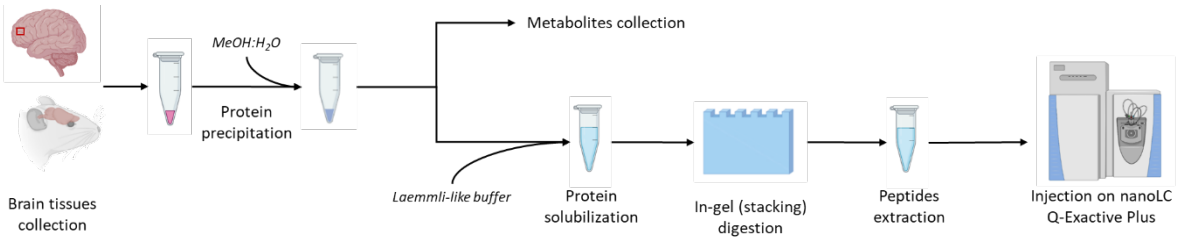


Figure 92: Schematic representation of sample preparation workflow for proteomics analysis of mouse and human brain tissues (Figure created with BioRender.com).

An additional step of protein MeOH/H_2O precipitation was added to the workflow. This precipitation step allowed to recover the metabolites of the sample for metabolomics analysis to be performed. It was also useful to concentrate and separate proteins from other cellular constituents such as the lipids from the brain tissues. MeOH precipitation is widely reported to perform metabolites extraction in tissues[17,18] and thus was used here, for further analysis of metabolites by our collaborators in Zurich (Switzerland).

A Laemmli-like buffer containing 3% of SDS was used for protein extraction, as it has been shown in different studies that a detergent-based buffer was more suitable for global proteomic analysis of brain tissues[401,403,404]. Indeed, Karpinski *et al.*[401] recently showed that detergent buffer displayed highest yield of protein extraction for brain tissues compared to chaotropic agent buffer or detergent-free buffer. By using a detergent buffer, they achieved higher protein concentration

compared to detergent-free, and their gene ontology analysis revealed that it allowed to identify more neuro-relevant proteins than detergent-free.

ii. Identification and quantification results

Identification and quantification were performed using MaxQuant software. We identified reliably and reproducibly more than 3400 proteins for each cohort, as displayed in **Figure 93**.

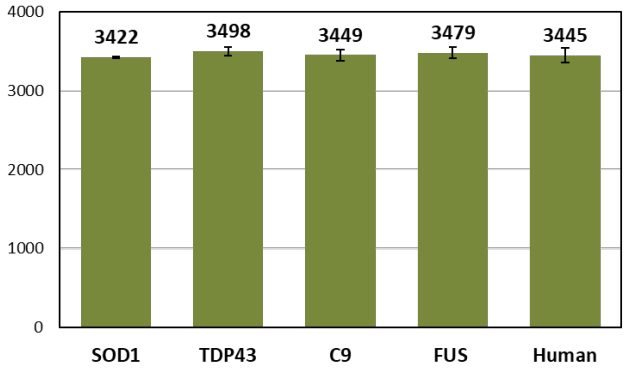


Figure 93: Mean number of proteins identified in each cohort of samples**.**

The goal was then to perform a differential analysis of the transgenic (TG) condition versus the wild-type (WT) *ie* the healthy condition. This analysis will allow to potentially find key proteins differentially expressed between our two conditions that could be later used as biomarkers. As ALS is slightly more prevalent in males than females[12], and because sex-specific differences were observed in blood of ALS patients[405,406] and on therapeutic responses in mouse models[407,408], we separated male and female conditions for the differential analysis. For each model, we thus divided the samples in 4 different conditions: TG_Male, TG_Female, WT_Male and WT_Female.

Before performing this differential analysis, we first need to make sure of the quality of our quantitative data. For this, we performed statistical validation using Prostar[409] software. First, we kept only proteins identified with at least one unique peptide. Because label-free untargeted proteomics data are sometimes filled with missing values, due mainly to the stochasticity of DDA analysis, we need to ensure that we get rid of the proteins that are only sparsely quantified. The percentage of missing values in the case of LC-MS/MS approaches can range from 10 to 50%, while the amount of proteins or peptides that present at least one missing value can range from 70 to 90%[410]. Missing values are a significant issue in proteomics because they are introducing a bias in quantification and lead to inaccurate representation of the samples. We thus kept only proteins, which had at least 80% of values (max 20% missing values), in at least one condition. Before proceeding to missing value imputation, we need to get rid of the technical variability between the analyses. For this, a normalization of the protein abundances was performed using the quantile normalization algorithm with a 15% quantile. Then, we got rid of the remaining missing values by performing imputation. Multiple methods for missing values exist and were compared in different studies[411,412]. As missing values might sometimes be due to proteins with low levels of expression in a specific condition, we imputed the missing values with small numbers. Using Prostar, we performed imputation using the *detquantile* algorithm, with a 2.5% quantile. Then we performed our differential analysis and set up a filter on p-value at 0.1% in order to best control our False Discovery Rate (FDR). As illustrated in **Figure 94** , we quantified more than 3000 proteins for each mouse model and for human samples.

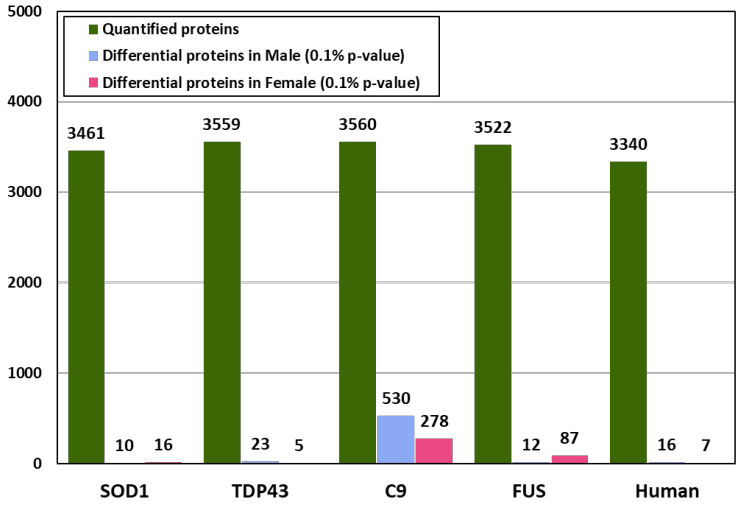


Figure 94: Total number of quantified proteins after statistical validation and number of differentially expressed proteins between TG vs WT conditions, for male and female (p-value < 0.1%).

For most studied models, the difference between the ALS (or TG) and the CTRL (or WT) conditions are minor, highlighting the high complexity of ALS disease.
In human samples, we detected 16 differentially expressed proteins (DEPs) in males and 7 in females (**Figure 95**). Among them, only 3 were common between the two sexes, all of them down-regulated in ALS condition. While not differentially expressed, several known neurodegeneration proteins were quantified such as matrin-3 (MATR3), spartin (SPART) or alpha synuclein (SYUA). MATR3 protein was shown involved in ALS[413], SPART protein in spastic paraplegia[414] and alpha synuclein is a known biomarker of Parkinson's disease[415].

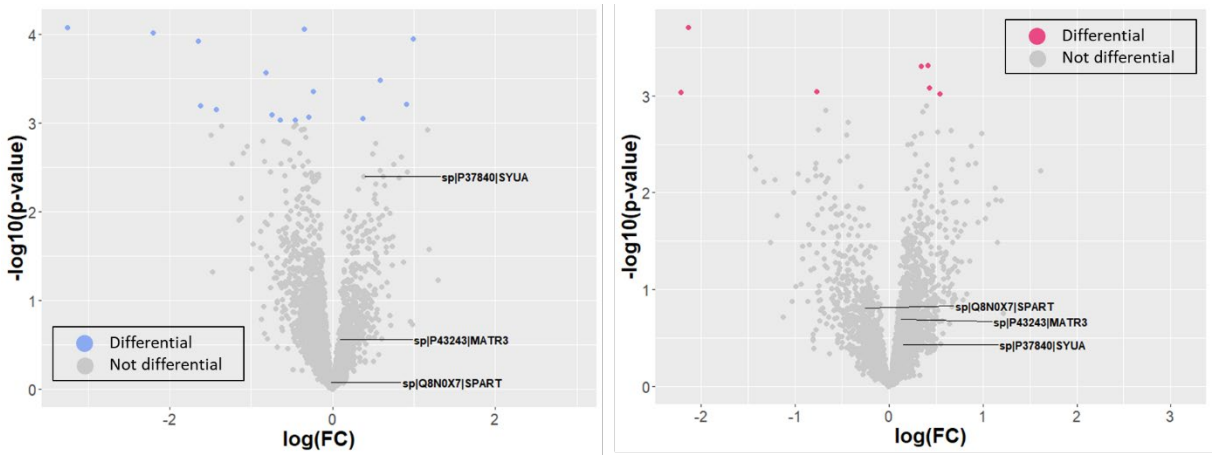


Figure 95: Volcano plots representing the differential analysis for human samples (ALS versus CTRL) for both sexes. All quantified proteins are represented by dots, colored dots representing DEPs.

For mouse models, between 0.3% to 0.7% of proteins were differentially expressed in males, and 0.1% to 2.3% in females. The C9 model however appears to be an exception and displays the strongest changes with more than 8% of DEPs in females and almost 15% of DEPs in males (**Figure 96**).

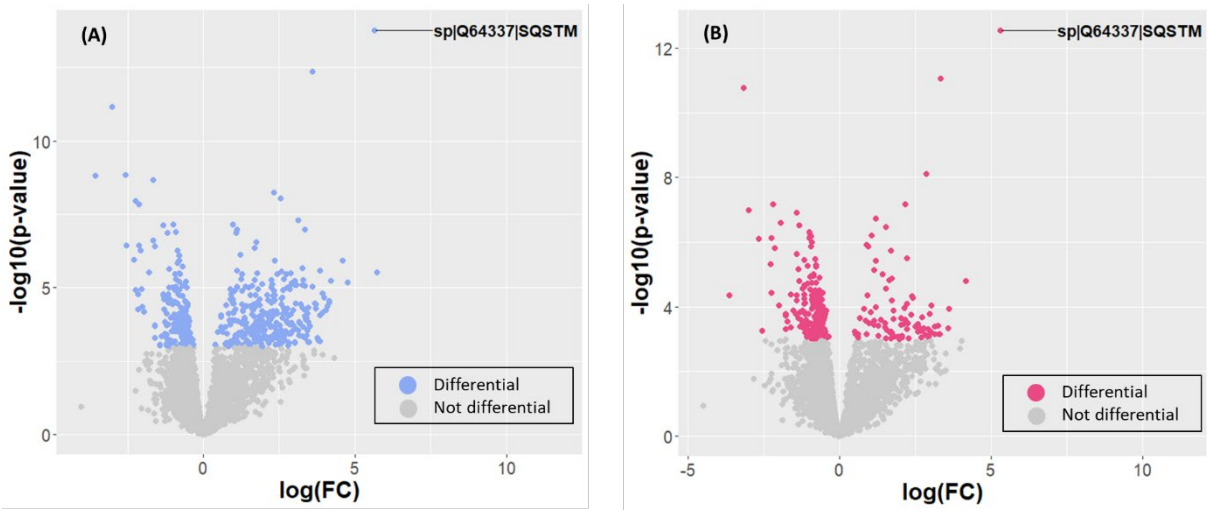


Figure 96: Volcano plots representing the differential analysis of C9 model (TG versus WT) for both sexes. All quantified proteins are represented by dots, colored dots representing DEPs.

One of the DEPs in C9 is sequestome-1 (SQSTM) protein, the highest up-regulated protein in both males and females (**Figure 96**). Sequestome-1 protein is the product of ALS-causing gene and its aggregation is reported in ALS patients as well as mouse models of fALS[416,417]. In both SOD1 and C9, male and female, differential analysis, exportin-1 (XPO1) appears up-regulated. XPO1 is a regulator of nuclear RNA transport and it has already been investigated as an ALS therapeutic target[418].

2. Application of the optimized phosphoproteomics workflow to study phosphorylation in mouse brain tissues

The analysis of mouse models' phosphoproteome is widespread to study various neurodegenerative diseases and especially Alzheimer's, but are scarcer for ALS. Phosphoproteomics analysis was not performed on human samples as most phosphoproteins are degraded under 72 hours in post-mortem tissues. Here, strong of the developments on phosphoproteomics described in **Part III**, we performed phosphoproteomics analysis on the four models of transgenic mouse PFC tissues. Analyses were performed on the protein extract from proteomic sample preparation of the mouse tissues. Samples enriched in phosphopeptides were injected on a Q-Exactive HF-X platform and identification and quantification were performed using MaxQuant. Identification results are displayed in **Figure 97.**

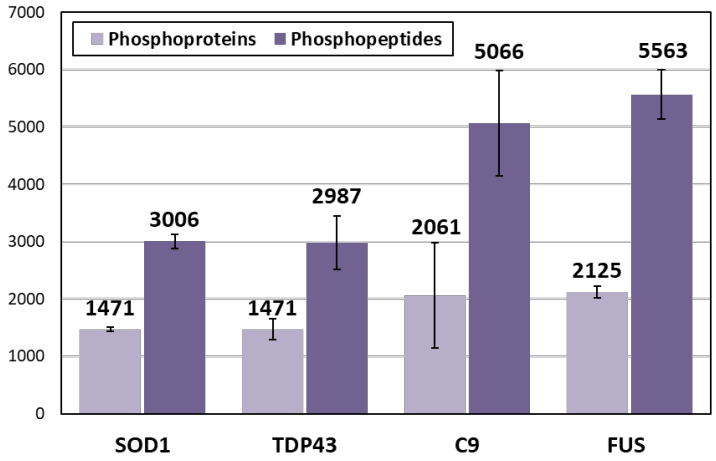


Figure 97: Mean number of phosphoproteins and phosphopeptides identified in each of the 4 mouse models.

Across the four different mouse models, we identified between 3000 and 5500 phosphopeptides, and from 1470 to 2125 phosphoproteins. In comparison, on C57B6 mice whole brains after phosphopeptide enrichment using both IMAC/C18 and TiO_2/C18, Nakamura *et al.*[419] identified by DDA 2938 phosphopeptides and 1567 phosphoproteins. More recently, 2124 phosphopeptides were identified in mice brain after SDS-Page sample preparation and TiO_2 phosphopeptide enrichment[420]. For most of our mouse models, both phosphoproteins and phosphopeptides identifications were reproducible with a CV <20%. However, the C9 model stands out with a CV on phosphoproteins identifications of 44% and on phosphopeptides identification of 19%. Overall, the reproducibility of identification within a mouse model but also the reproducibility across the different mouse models is much lower than it was for global proteomics. Once again, this is mainly due to the addition on an extra phosphopeptide enrichment step in the workflow but also to the lability of the phosphorylation, which both increase the variability in the results.

As for the global proteomics analysis, samples of each model were divided into 4 different conditions before performing differential analysis: TG_Male, TG_Female, WT_Male and WT_Female. Using Perseus and it's 'expand site stable' option, intensities of all peptides involved in a phosphosite were extracted from MaxQuant's Phospho(STY).txt file and merged to obtain quantification information at the phosphosite level. Only phosphosites with a probability of localization greater than 0.75 were kept. Using Prostar, quality filters were applied, normalization and imputation of the data were performed. The results of the different differential analysis, performed at the class I phosphosites level, are represented in **Figure 98**.

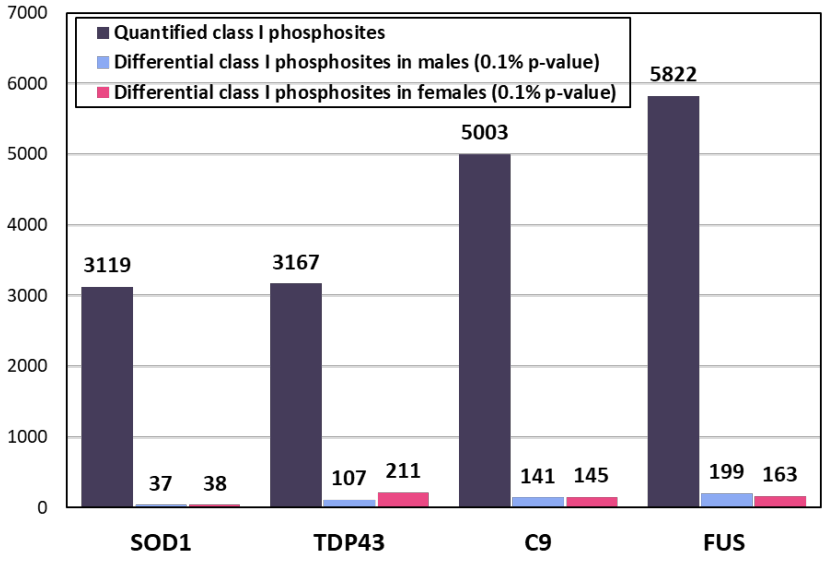


Figure 98: Total number of quantified class I phosphosites after statistical validation and numbers of differentially expressed class I phosphosites between TG vs WT conditions, for male and female (p-value < 0.1%).

We quantified around 3000 class I phosphosites for both SOD1 and TDP43, and up to more than 5000 for C9 and FUS models. Among these, between 1.2% to 6% of the class I phosphosites were differentially expressed (Differentially Expressed phosphosites, DEpS) between transgenic and wild type conditions.

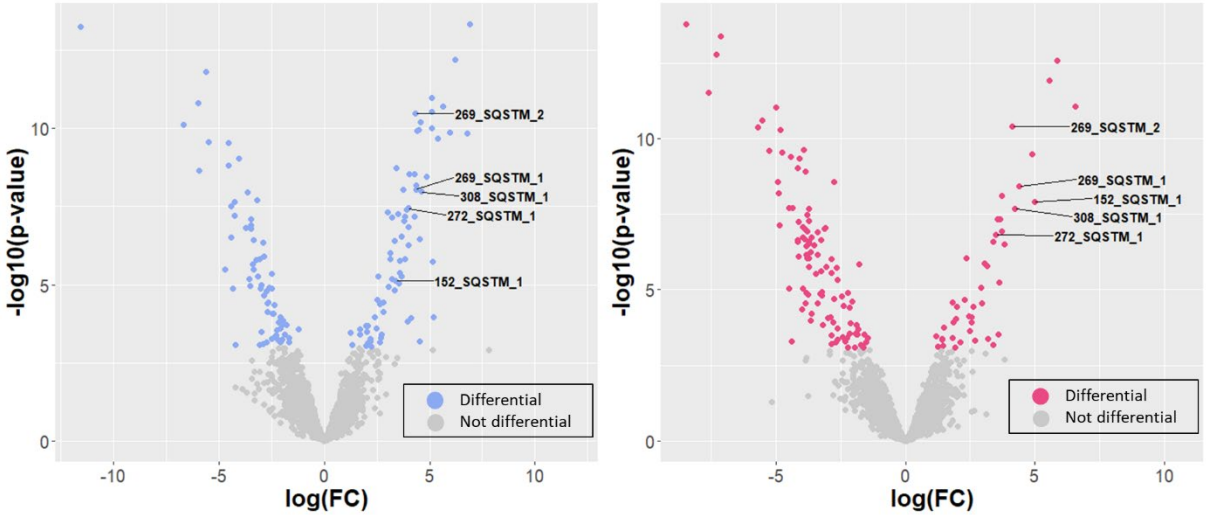


Figure 99: Volcano plots representing the differential analysis of C9 model (TG versus WT) for both sexes. All quantified class I phosphosites are represented by dots, colored dots representing DEpS.

Amongst the DEpS in the C9 model, five phosphosites are localized on the sequestome-1 (**Figure 99**) protein that was previously found as differentially expressed in the C9 model for global proteomics. These five sequestome-1 phosphosites are all highly up-regulated in the ALS condition, for both males and females.

3. Multiomic ALS signatures highlight sex differences, molecular subclusters and the MAPK pathway as therapeutic target

The (phopsho)proteomics analysis of mouse and human brain tissues were performed in a multiomic study. The integrated analysis of transcriptomes, (phospho)proteomes, and miRNAomes are detailed in a paper currently in submission in Nature Medicine journal (see **Appendices**). For this publication, all the results of the (phospho)proteomics analysis on tissue samples were made public through a complete submission of the data on PRIDE platform with the following datasets identifiers: : PXD043300 and PXD043297.

Chapter 2: Development of a protocol for both proteomics and phosphoproteomics analysis of cerebrospinal fluid

Cerebrospinal fluid (CSF) is produced in the brain at a rate of 500 mL a day, allowing to replace the 150 mL of circulating volume to be recycled three to four times a day[421]. CSF is the only body fluid in direct contact with the extracellular space of the brain, and is thus widely used to study central nervous system (CNS) diseases. Indeed, it is a valuable reporter of abnormal variation in the CNS, such as inflammation, infection, neurodegeneration or tumor growth[421]. It is thus a fluid of choice to investigate the discovery of potential biomarkers for neurodegenerative diseases such as ALS[85,422]. Composition of the CSF is close to plasma, with albumin and immunoglobulin constituting respectively around 50% and 15% of the total protein content[421,423]. However, CSF contains also its specific highly abundant proteins, such as cystatin C or prostaglandin D2 synthase, both synthetized in the CNS. Because of the large dynamic range of protein concentration in CSF (**Figure 100**), with its top 10 most abundant proteins accounting for about 80% of total concentration, identification and quantification of low abundant proteins is challenging. Additionally, total protein concentration in CSF is 50-100 times lower than in plasma. Indeed, protein concentration in CSF samples is quite low, with a mean CSF protein amount estimated at 0.42 µg/µL, and an overall protein concentration below 1 µg/µL[421,424].

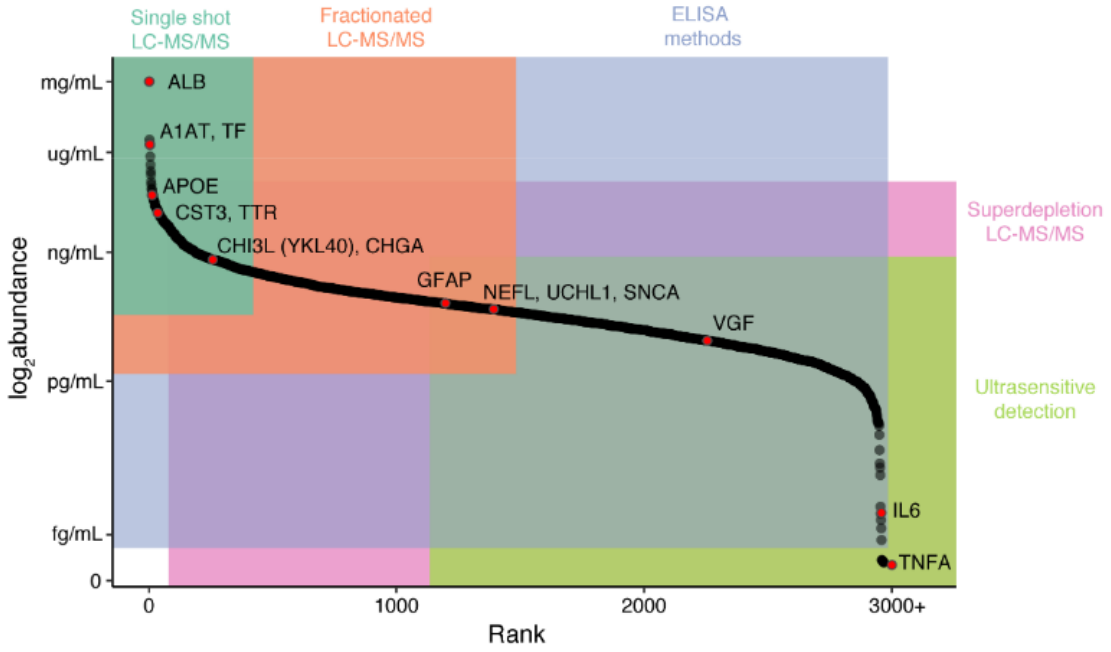


Figure 100: Cerebrospinal fluid protein dynamic range according to Carlyle *et al.*[425]. Proteins are ranked according to their abundance, with the location of specific proteins placed according to their concentrations in enzyme-linked immunoassays (ELISAs), Multiple-Reaction-Monitoring (MRM), and (unpublished) label-free experiments. One should note that there is disagreement between experiments on the exact concentration of these analytes and thus their place in the plot should be considered illustrative[425].

Sample preparation of CSF is thus a crucial step due to the high dynamic range of concentrations. Therefore, different pre-analytical technics have been implemented to enhance proteome coverage. One of the most commonly used technic is depletion of high abundant proteins[421,426,427], similarly to plasma sample preparation. Affinity depletion is based on the specific capture of target proteins thanks

to immobilized molecules or antibodies, such as protein A/G or Cibacron Blue, which have specific affinity and high affinity for the targets. Depending on the used depletion kit, between 2 to 20 of the most abundant proteins can be removed[423]. One of the risk with this technic is that low-abundant proteins might be co-depleted from the sample, making their detection even more difficult[428]. Other technics such as relative enrichment of low and medium abundant proteins[423], protein precipitation[429–431] or sample fractionation[427,429] are reported in various CSF studies. However, the addition of any of those steps might inherently introduce sample preparation variability. Moreover, enrichment of low abundant proteins was shown to lead to protein loss[423]. For fractionation, large amounts of CSF are required due to its low protein concentration, but usually only small amounts of CSF are available as it is a precious biological sample, collected through invasive lumbar puncture procedure[432]. During this CSF collection through lumbar puncture, unwanted vascular bleeding might happen (in up to 20% of the procedures), leading to peripheral blood to contaminate CSF. These blood contaminations may introduce variabilities and lead to unreliable detection of biomarkers. Among the most common, hemoglobin is a marker of blood contamination in CSF as it is highly expressed in red blood cells. Carbonic anhydrase and catalase were also already used as blood contamination markers in CSF. One thus needs to look for those potential blood contaminants, as quality of CSF sample is key to the analytical outcome. Some methods have been set up to estimate levels of blood contamination in CSF through mass spectrometry analysis[433].

Here, because of the low amount of starting material at our disposal (for the MAXOMOD project, 150 µL of CSF for proteomics and 500 µL for phosphoproteomics), a fractionation step was not considered as it usually requires more than 500 µL of sample. On a large cohort of samples such as the one for this specific project, a precipitation step appeared to be the most economic and suitable approach compared to the use of commercial kits either for protein depletion or enrichment. Protein precipitation was also necessary in our case in order to recover the metabolites that were analyzed by another MAXOMOD partner. Methanol precipitation was chosen as it is well reported for efficient metabolites extraction on CSF samples[434,435].
As for tissue (phospho)proteomics, in the following work, different protocols were evaluated to set up a common (phospho)proteomics and metabolomics sample preparation for high throughput CSF analysis. The optimized method was then applied on more than 100 clinical CSF samples for both proteomics and phosphoproteomics analysis.

1. Optimization of a sample preparation protocol for CSF proteomics and phosphoproteomics analysis

i. Evaluation of in-solution versus in-gel digestion

In a first approach, in-solution digestion (ISD) using RapiGest (Waters) surfactant was compared to classical in-gel digestion (IGD), as both technics are reported in literature for CSF sample preparation[427,428,436]. RapiGest is a reagent used to enhance the in-solution digestion. Both protocols were compared with and without a methanol precipitation step, to evaluate the impact of this additional step on protein concentration and recovery. The different protocols are represented in **Figure 101**. Two biological replicates and three technical replicates were used to evaluate each workflow.

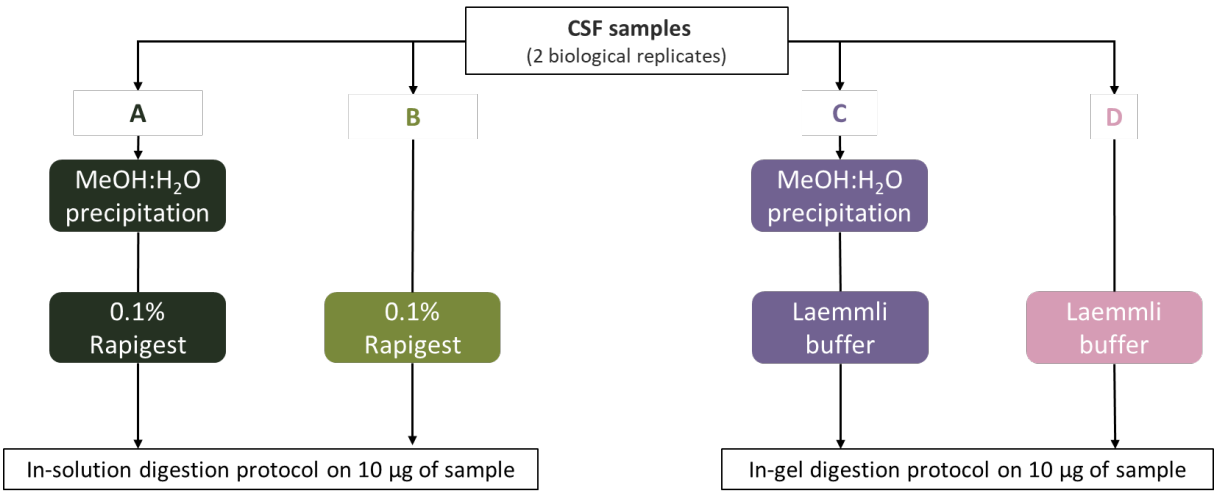


Figure 101: Method development for CSF sample preparation. Four different protocols were compared: A = MeOH precipitation + ISD ; B = ISD ; C = MeOH precipitation + IGD; D = IGD.

We assessed protein concentration before and after methanol precipitation. Before precipitation, the range of concentration was [0.3 – 0.7] µg/µL while after precipitation the range of protein concentration increased to [2.2 – 4.7] µg/µL, depending on the biological replicate and on the protocol. Thanks to this precipitation step, CSF protein concentration was successfully increased.

We also evaluated the protein recovery efficiency of the methanol extraction for the two protocols. Starting with 150 µg of material, we recovered between 57 µg – 117 µg (38% - 78%) of proteins with IGD (protocol C) and between 55 µg – 98 µg (37% - 65%) with ISD (protocol A). Protein loss appears similar between the two protocols but high variability is observed between the recoveries of the two replicates.

a. *Identification results*

We then compared the performances of the different protocols in terms of identification, and results are displayed in **Figure 102.**

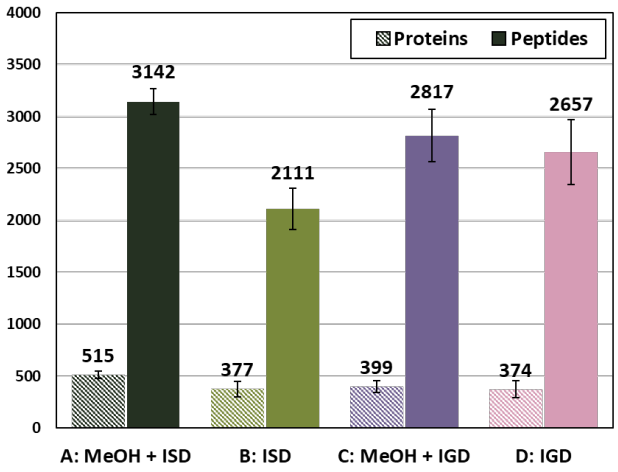


Figure 102: Average numbers of proteins and peptides identified by the different protocols.

Protocol A with a methanol precipitation step followed by in-solution digestion using RapiGest displays the best results with more than 500 proteins and 3000 peptides identified. It also seems to be the more reproducible protocol with a CV on peptides (proteins) identified of only 4% (9%). The precipitation step increases identifications for both protocols but appears to have a greater impact on in-solution digestion than on in-gel digestion. Indeed, the addition of methanol precipitation in the case of RapiGest liquid digestion improves identifications by 27% for proteins and 33% for peptides.

b. *Quantification results*

The number of peptides that were quantified in 6 out of 6 replicates (2 biological replicates x 3 injection replicates per method) are represented in **Figure 103**.

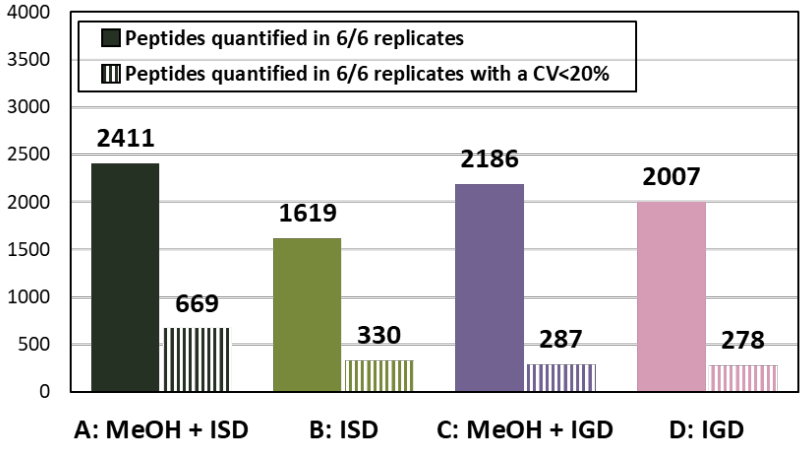


Figure 103: Number of peptides quantified in 6/6 replicates with and without a filter on CV<20%.

Protocol A (MeOH precipitation and in-solution digestion) and protocol C (MeOH precipitation and in-gel digestion) both display the highest number of quantified peptides (respectively, 2411 and 2186 quantified peptides). However, when we apply a filter on the CVs of the quantified peptides intensities,

we lose almost 90% of quantified peptides with method C. Less than 300 peptides are quantified reproducibly with a CV <20% with method C, while method A still allows to robustly quantify 669 peptides. This is further shown with the distribution of CVs on quantified (in 6/6 replicates) peptides intensities, represented in **Figure 104**. Here, protocol A displays a more reproducible quantification with a median CV of 25%, the lowest compared to other methods.

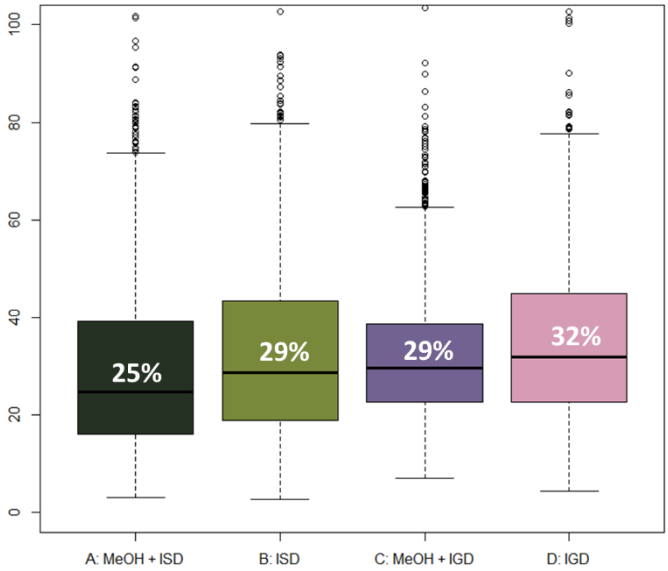


Figure 104: Distribution of the coefficient of variation (CVs) on the intensities of peptides quantified in 6 out of 6 replicates for the different protocols.

c. *Comparison of protein and peptide populations identified with the different protocols*

A represented in **Figure 105**, we obtain an overlap between the different protocols of almost 50% at the protein level and 27% at the peptide level. These numbers are quite correct considering that the average recovery between technical replicates for a DDA analysis is around 70% for proteins and 40-50% for peptides. By comparison, Neset *et al.* evaluated three different protocols (one of them being ISD) on HeLa cells and obtained a 67.1% overlap at the quantified protein level[102]. In another study from Ludwig *et al.*, a 55% protein overlap and a 38% peptide overlap were obtained comparing 5 different sample preparation protocols on colon cancer cells[104].

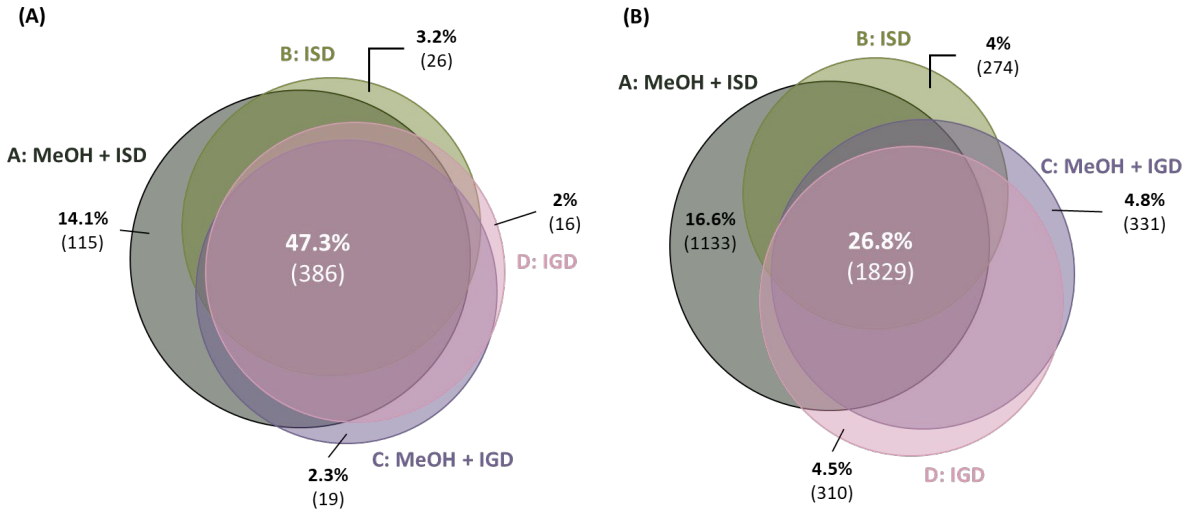


Figure 105: Venn diagrams representing the overlap between (A) identified proteins and (B) identified peptides between the different protocols.

Protocol A with methanol precipitation followed by in-solution digestion allows to uniquely identify more than 110 proteins and 1130 peptides, representing respectively 14% and almost 17% of all identified proteins and peptides (compared to less than 5% for other protocols).

The proportion of missed-cleaved peptides among all the identified peptides is also an element to evaluate the efficiency of digestion between protocols. They are represented in **Figure 106**. Protocol A (MeOH precipitation and ISD) displayed the lowest amount of missed-cleavages (MC) with 13% of missed-cleaved peptides. The addition of a precipitation step seems to decrease the number of miss-cleaved peptides compared to the sample protocol without precipitation. Indeed, protocol B has 18% of MC compared to protocol's A 13%, and protocol D has 27% compared to protocol's C 25%. Protein precipitation thus seems to increase digestion efficiency. In-gel digestion protocols displayed the highest percentages of missed-cleaved proteins, which can be explained by the use of only trypsin for the enzymatic digestion for this protocol, compared to the combined use of trypsin and lysC for in-solution digestion. This higher MC proportion for IGD was also observed by Yang *et al*., compared to ISD or filter-aided digestion[106]. In-solution digestion protocols (A and B) might also display better digestion efficiency thanks to the use of RapiGest surfactant, that is designed to improve protein digestion by facilitating their unfolding while retaining enzymatic activity.

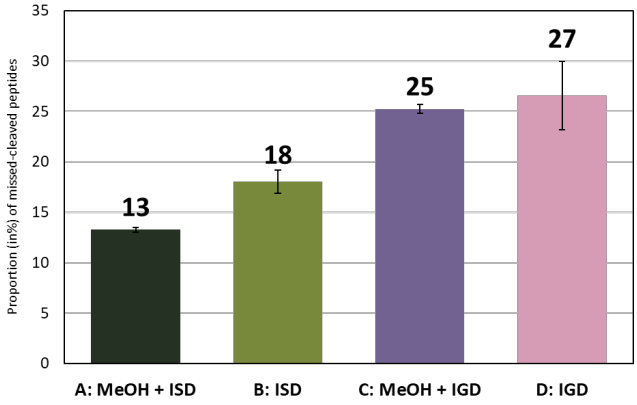


Figure 106: Proportion (in %) of missed-cleaved peptides for the different protocols.

ii. Evaluation of in-solution versus on-membrane digestion

Now that we stated that methanol precipitation improved identifications and that RapiGest in-solution digestion outperformed in-gel digestion, we wanted to benchmark this protocol (protocol A, RapiGest ISD) against another sample preparation method: on-membrane digestion. For this, we used PreOmics iST kit (PreOmics), which is also reported in literature for CSF proteomics sample preparation[85]. Sample preparation workflow is represented in **Figure 107**. Three biological replicates were used to evaluate each workflow. While there is no recommendation on the amount of protein material for RapiGest in-solution digestion, it is advised with the iST to apply the protocol on 1-100 µg of protein. Additionally, for less than 20 µg of starting material, the protocol needs to be slightly adapted. Thus, to be in the middle of the advised range, we performed iST protocol on 45 µg of protein. For in-solution digestion with RapiGest, protocol was applied on 10 µg of starting material.

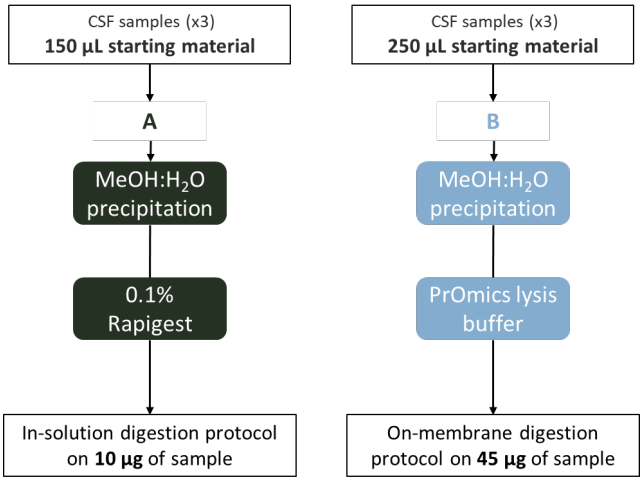


Figure 107: Comparison on two different protocols for CSF sample preparation. A = in-solution digestion using RapiGest ; B = PreOmics iST kit protocol with membrane digestion.

We measured protein concentration before and after protein resuspension to evaluate the percentage of protein recovery depending on the lysis buffer. For protocol A (RapiGest), protein recovery ranged depending the biological replicate between 68% to 87% while it was between 39% to 60% for protocol B (PreOmics).

a. Identification and quantification results

Next, we evaluated the performances of the two protocols in terms of peptides identification and quantification (**Figure 108**). Protocol 1 using 0.1% RapiGest with in-solution digestion allows to identify 3333 peptides and 593 proteins, which represents 32% more peptides and proteins than with the iST protocol. Results obtained using RapiGest are in accordance with Barkovits *et al.* results, as they identified on average over 3 CSF replicates 573 proteins and 3180 peptides on Q-Exactive HF in DDA mode[428].

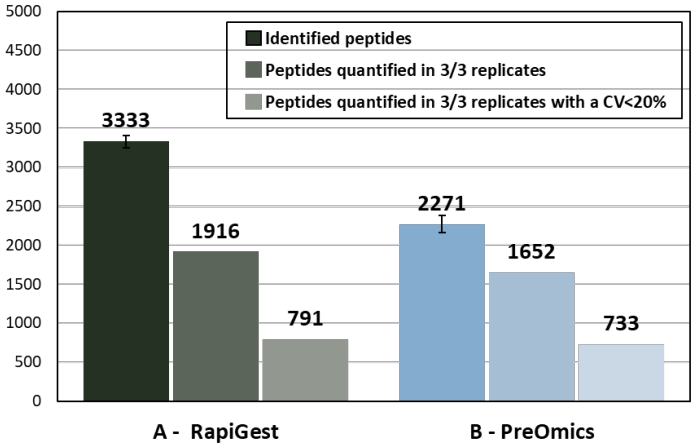


Figure 108: Average numbers of peptides identified and numbers of peptides quantified in 3 out of 3 replicates, with and without a filter on CV < 20%.

In terms of quantification, in-solution method also allows for the quantification of more peptides with almost 2000 quantified peptides. We then applied a filter on the CVs of the intensities of quantified peptides, to obtain only the more robustly quantified peptides. With this filter, the performances of the two methods are comparable, with respectively 791 and 733 quantified peptides for method A and method B. In addition, both protocols have similar reproducibility of quantification with a median CV of 23% for both methods, highlighting their similar robustness of quantification (**Figure 109**).

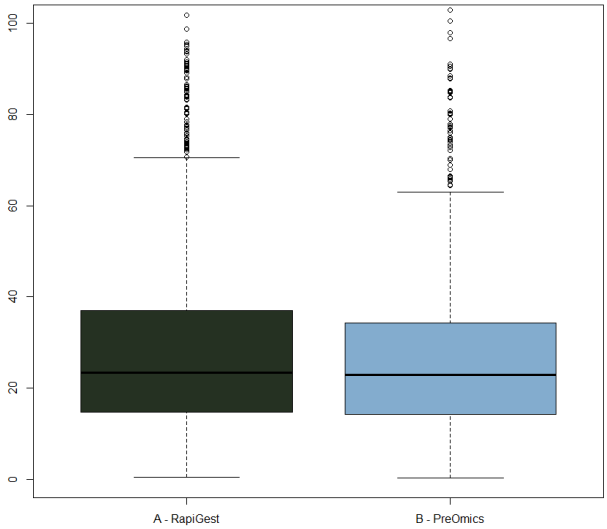


Figure 109: Distribution of the coefficient of variation (CVs) on the intensities of peptides quantified in 3 out of 3 replicates for the two protocols.

b. *Comparison of protein and peptide populations identified with the two protocols*

We also looked at the overlap between the peptides identified by the two protocols. **Figure 110** represents the recovery of all peptides identified in the three replicates for the RapiGest protocol with all peptides identified in the three replicates for the iST PreOmics protocol.

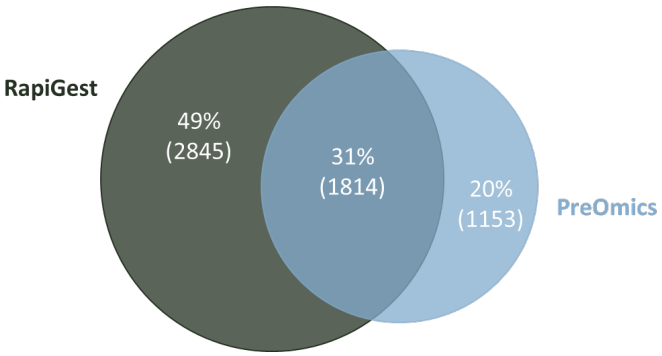


Figure 110: Venn diagram representing the overlap in peptides identification between the two protocols.

Only a 31% overlap is observed between the two protocols, meaning they both identify very different peptides. In total, over the three replicates, the RapiGest protocol identified more than 4600 peptides while PreOmics identifies in total 2967 peptides. The proportion of missed-cleaved peptides among all the identified peptides was also evaluated for the different protocol. They are represented in **Figure 111**.

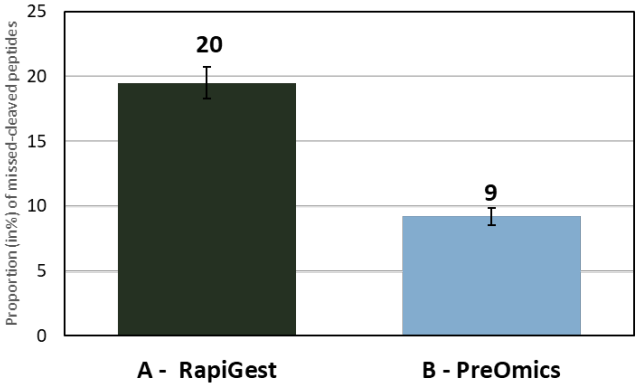


Figure 111: Proportion (in %) of missed-cleaved identified peptides for the different protocols**.**

PreOmics iST digestion (protocol B) displays twice less missed cleavages than RapiGest in-solution digestion (20% for RapiGest compared to 9% for iST). Comparable results were obtained by Ding *et al.* on urine samples comparing different sample preparation protocols including in-solution digestion and PreOmics iST protocols. In their work, in-solution digestion showed around 75% of full cleavages while iST method yielded the best digestion efficiency with around 92% of full cleavage[50]. This accrued digestion efficiency of iST method is suggested to be due to higher abundance of trypsin in the provided iST kit. To evaluate this theory, we looked at the contribution of trypsin to the overall MS1 peptide intensity. In **Figure 112** are represented the percentage of trypsin peptides' MS1 intensity over the total MS1 peptides intensity. Trypsin's peptides intensity is more than 10% higher with iST method compared to RapiGest method, suggesting that trypsin amount is indeed higher in iST kit. This is also

the potential reason for which PreOmics protocol displays less missed-cleavages compared to in-solution method.

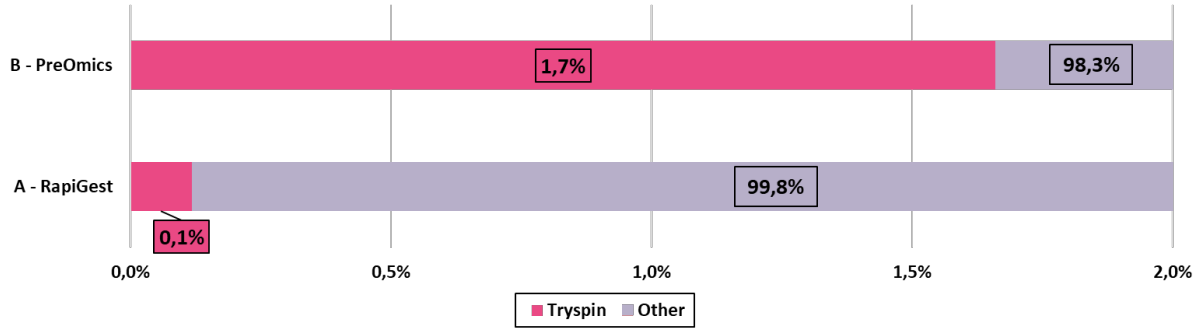


Figure 112: Percentage of trypsin and LysC peptides' MS1 intensity over the total MS1 peptides intensity for each protocol.

iii. Choice of the best suited CSF sample preparation protocol for proteomics analysis

To choose the best suited protocol for CSF sample preparation, we need to take into account not only the different results shown previously but also the time that each protocol takes. The description of the different steps of the different protocols and their corresponding timelines are detailed in **Figure 113**.

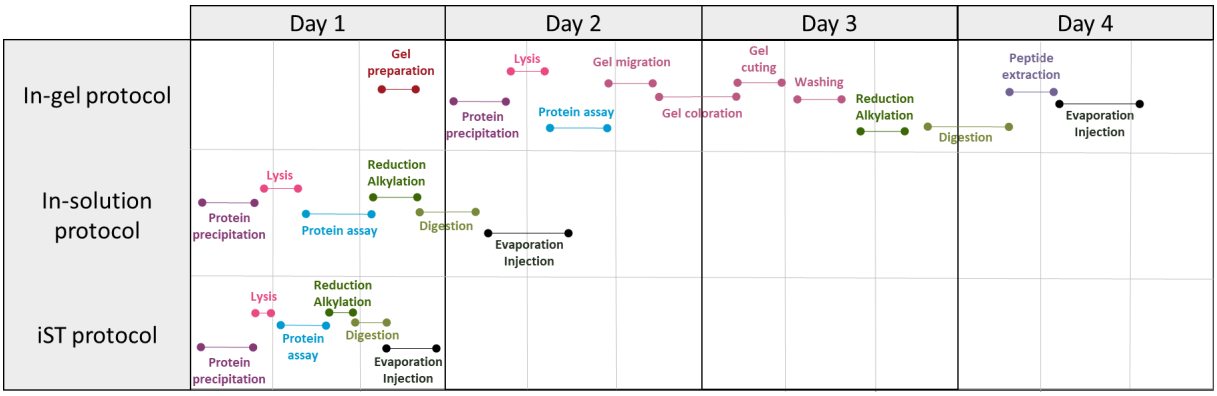


Figure 113: Description of the sample preparation workflows evaluated for CSF proteomics analysis.

In-gel protocol is the longest as it takes approximately three days to prepare samples for injection (a bit less if using pre-prepared commercial gels), while both in-solution and iST protocols are much faster with less than 1 day of sample preparation for iST protocol.

Main advantages and drawbacks of the different protocols were summed up in **Table 10** to have a better overall view of the comparisons. It is however important to keep in mind that the conclusion discussed here works for CSF samples but might be very different for other sample types.

	Protocol duration	Peptides identification	Missed cleavages	Peptides quantification	Quantification reproducibility
In-gel protocol	-	++	-	+	+
In-solution protocol	++	+++	++	+++	++
iST protocol	+++	++	+++	++	++

Table 10: Summary of the drawbacks and advantages of the different protocols according to our results.

As the goal was to develop a high-throughput proteomics analysis adapted for thousands of clinical CSF samples, the in-gel protocol was not adapted as it is the most time-consuming method. Moreover, in-solution digestion outperformed in-gel digestion both in terms of identification and quantification performances. While iST protocol was the fastest, RapiGest in-solution method exceeded iST protocol for peptides identification and quantification, while the two methods produced comparable results for quantification reproducibility. In addition, RapiGest performances were achieved on a smaller amount of material compared to iST, which is a key asset when working on limited amounts of clinical samples. Finally, one other important parameter to take into account for the analysis of large cohort of samples is the price of the protocol. PreOmics iST kit is the most expensive, costing (at the time these analysis were performed) around 200€ for the preparation of 10 samples. On the other hand, preparing samples by in-solution digestion with RapiGest costs around 80€ per 10 samples. For all these reason, we decided to keep RapiGest in-solution digestion as optimized sample preparation protocol for CSF samples.

iv. Evaluation of sample preparation for CSF phosphoproteomics analysis

The optimized sample preparation protocol using RapiGest was applied on the CSF samples and an additional automated phosphopeptide enrichment step was added, to evaluate its performances for phosphoproteomic analysis (**Figure 114**).

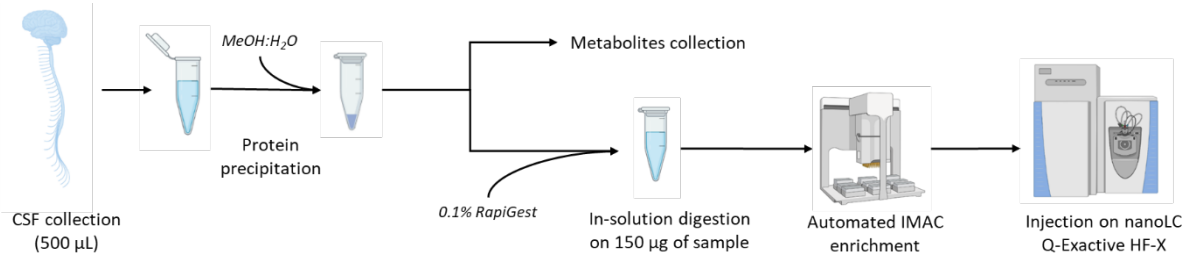


Figure 114: Schematic representation of sample preparation workflow for phosphoproteomics analysis of cerebrospinal fluid samples (Figure created with BioRender.com).

Identification results are displayed in **Figure 115.** With this protocol, a total of almost 200 phosphopeptides and close to 150 class I phosphosites were identified on the two CSF replicates (**Figure 115 - (A)**). Moreover, an average of 84 phosphoproteins and 146 phosphopeptides were identified. At the phosphopeptides level, a coverage of almost 50% was achieved between the two replicates (**Figure 115 - (B)**), which is around the expected value when studying phosphorylation. Very few studies of the CSF phosphoproteome can be found in literature, making it difficult to evaluate if our results match expected values or not. Among the few CSF phosphoproteomics studies published so far, Nakamura *et al.*[419], using an LTQ Orbitrap, reported 123 phosphopeptides and 70 phosphoproteins identified after sequential IMAC/C18 and TiO_2/C18 phosphopeptide enrichment. 44 phosphorylated proteins were identified by Bahl *et al.* on CSF samples after TiO_2 enrichment followed by LC-MS/MS analysis on a LTQ-Orbitrap[437]. More recently, 1200 phosphopeptides were quantified by PRM-PASEF on extracellular vesicles isolated from CSF and enriched in phosphopeptides[438].

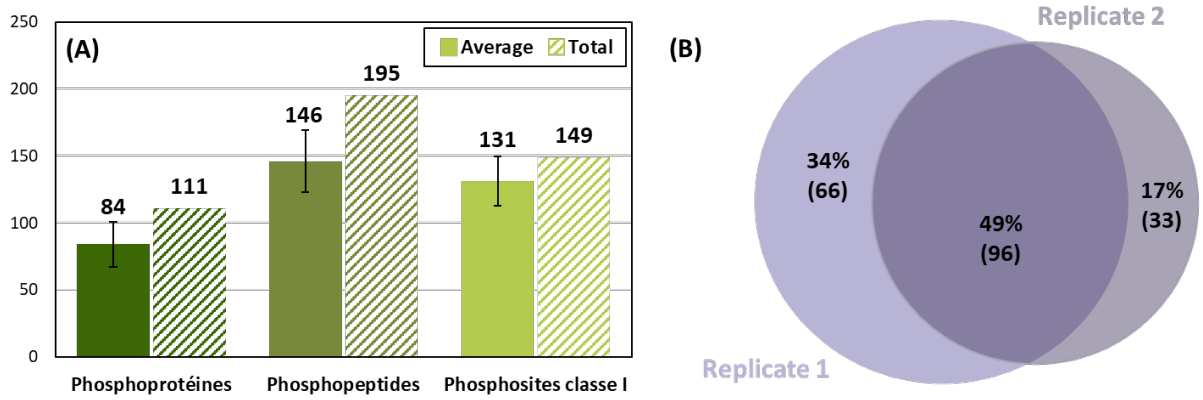


Figure 115: (A) Average and total numbers of phosphoproteins, phosphopeptides, and class I phosphosites identified in the two biological replicates (B) Overlap of the phosphopeptides identified in the two biological replicates.

In conclusion, a protocol starting with a methanol precipitation followed by in-solution digestion using RapiGest (with an additional phosphopeptides IMAC enrichment for phosphoproteomics) has showed efficient identification results for (phospho)proteomics analysis of CSF samples. Therefore, this protocol will be applied on the MAXOMOD cohort of a 110 CSF samples.

2. High throughput (phospho)proteomics analysis of ALS-cerebrospinal fluid samples

i. Quality of CSF samples

As stated previously, blood proteins can contaminate CSF during sample collection. In order to evaluate the quality of our CSF samples *ie* their contamination level, we used the classification established by Barkovits *et al.*[433]. In their publication, different methods were used to detect blood levels in CSF: ELISA, Combur10-Test® strips, and MS-based analysis. They then defined five different contamination values: (i) "negative" corresponds to 1, (ii) very low to 10, (iii) low to 20, (iv) high to 30, and (v) very high to 40. Specific thresholds were then set up to each categories from each methods. Focusing on MS, the identification pattern of three known blood contaminant proteins were used *ie* hemoglobin (HB), carbonic anhydrase 1 (CAH1) and catalase-A (CATA). The criteria for each category are presented in **Table 11**.

Contamination value	LC-MS criteria
1	No HB or <5 peptides
10	HB (≥5 peptides)
30	HB and CAH1
30	HB, CAH1, CATA (≤4 peptides)
10	HB, CAH1, CATA (≥5 peptides)

Table 11: Categorization of blood contamination in CSF. Five specific contamination levels were selected on the basis detection of specific blood proteins (for LC-MS analysis). Adapted from Barkovits *et al.*[433].

This categorization system was thus applied on our 103 CSF samples. As displayed in **Figure 116**, almost all samples contamination levels were below 10, and thus considered as very low contaminated.

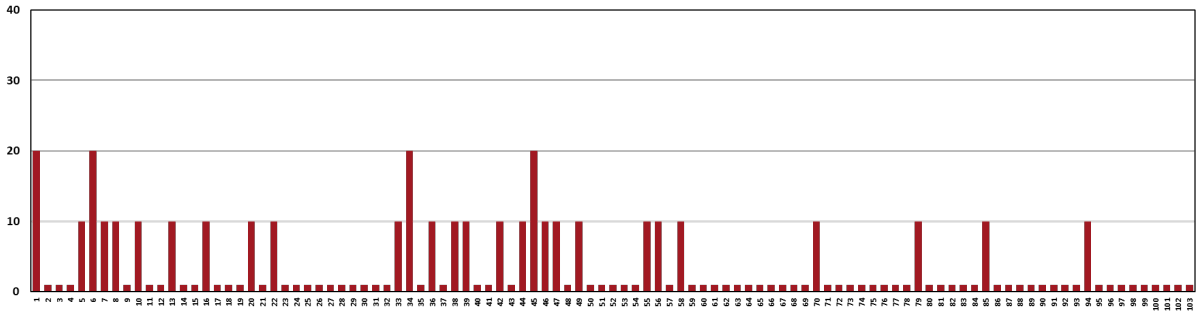


Figure 116: Results of the blood level categorization to our clinical CSF samples.

Thanks to these classifications, we were able to ensure the good quality of our CSF samples, and thus that no blood contaminant might affect our analysis.

ii. Global proteomics analysis of ALS-cerebrospinal fluid samples

After applying the optimized sample preparation protocol on the 103 samples of CSF, generated data were analysed using MaxQuant. Protein intensities were extracted and some quality filters were applied using Prostar. As for global proteomics on mouse and human brain tissues, we divided samples in 4 different conditions (ALS_Male, ALS_Female, CTRL_Male and CTRL_Female) and kept only the proteins that were identified in at least 80% of the samples in at least one condition. Quantile normalization and *detquantile* data imputation were performed. After this, we were able to perform differential analysis of the ALS (*ie* the "disease" condition) against the CTRL condition (*ie* the "healthy" condition). Results of these differential analyses are represented by volcano plots in **Figure 117**.

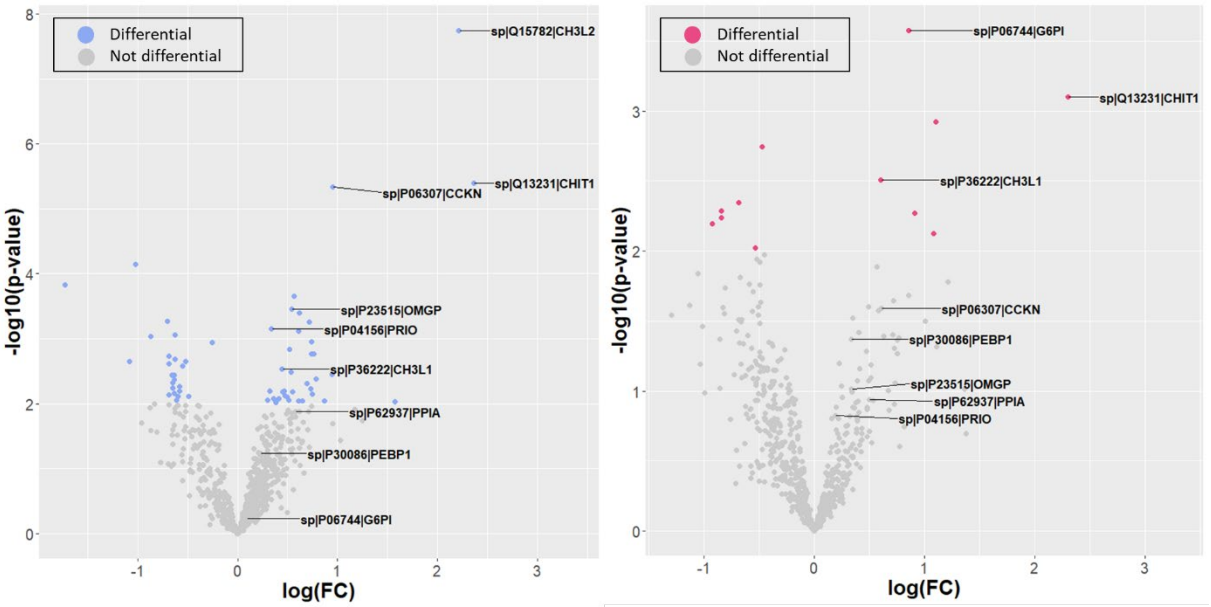


Figure 117: Volcano plots representing the differential analysis of CSF samples (ALS versus CTRL samples) for both sexes, with a 1% p-value filter. All quantified proteins are represented by dots, colored dots representing DEPs.

In total, we robustly quantified 669 proteins. For the comparison amongst females, only 12 proteins were differentially expressed while 59 DEPs were identified in the male condition.

Four proteins were differentially expressed in both males and females, and among them was chitotriosidase-1 (CHIT1, also known as chitinase 1), which is one of the highest up-regulated protein in ALS in both females and males. CHIT1 belongs to the chitinase protein family, which are secreted by activated glial cells and other cells of the immune system. CHIT1 was already reported to be involved in ALS disease[439,440] and other neurodegenerative diseases such as multiple sclerosis[441]. Indeed, CHIT1 levels in CSF were shown more elevated in ALS samples compared to disease controls and correlated with disease progression rate, thus representing a promising tool for both diagnostic and progression of ALS[439]. Two other chitinase proteins, the Chitinase-3-like protein 1 (CH3L1) and Chitinase-3-like protein 2 (CH3L2) were also differentially quantified in females and/or males. CH3L1 and CH3L2 are both reportedly linked to ALS and especially to spinal cord atrophy for CH3L1[440,442].
Some other DEPs found in the male condition are reported to be linked with various neurodegenerative diseases. Among them is oligodendrocyte (OMGP) protein. Oligodendrocytes (OLs) are proteins of the central nervous system whose primary function is to form the myelin, the structure wrapping the axons and ensuring protection and signal conduction to neurons. The impact of myelin deterioration (demyelination) is known in a variety of neurodegenerative diseases, including ALS and is considered a key factor of disability progression. Studies suggests that OLs are affected during disease progression and that their dysfunction may be a key factor of the disease, and thus could be a novel therapeutic target for ALS[443]. Prion protein PRIO was also found up-regulated in ALS in the males. Prion proteins are misfolded proteins in that can trigger other “normal” proteins to fold abnormally and thus aggregate, leading to neurodegenerative prion diseases. Similarities between prion diseases and ALS have been reported, opening another road to potential insights into new therapeutic strategies for ALS[444,445]. Another protein that was up-regulated in ALS for males is cholecystokinin (CCKN), hormone of satiety and is highly expressed in brain regions such as hippocampus.

Dysregulation of Glucose-6-phosphate isomerase (G6PI) is also particularly shown in females. This enzyme is supposedly involved in many neurodegenerative diseases (ALS, Parkinsons’, Alzheimers’, Huntingtons’) but its role especially in ALS has not yet been fully explored[446].
Other quantified proteins were found related to ALS and neurodegeneration. Cyclophlin A (PPIA) is an enzyme involved in protein folding and assembly and is mainly in the CNS. Cyclophin A’s activity has been shown to be linked with ALS[447,448]. Phosphatidylethanolamine binding protein 1 (PEBP1) was also quantified here, and its potential role as an Alzheimer’s disease biomarker has been suggested in many studies[449,450].

iii. Phosphoproteomics analysis of ALS-cerebrospinal fluid samples

As for the global proteomics analysis, phosphoproteomics samples were divided into 4 different conditions before performing differential analysis: ALS_Male, ALS_Female, CTRL_Male and CTRL_Female. Using Perseus and it's 'expand site stable' option, the intensities of all peptides involved in a phosphosite were extracted from Phospho(STY).txt file from MaxQuant and merged to obtain quantification information at the phosphosite level. Only the phosphosites with a probability of localization greater than 0.75 were kept. Using Prostar, we kept only the proteins that were identified in at least 50% of the samples in at least one condition. Normalization and imputation of the data were also performed. The results of the different differential analyses, performed at the class I phosphosites level, are represented in **Figure 118**.

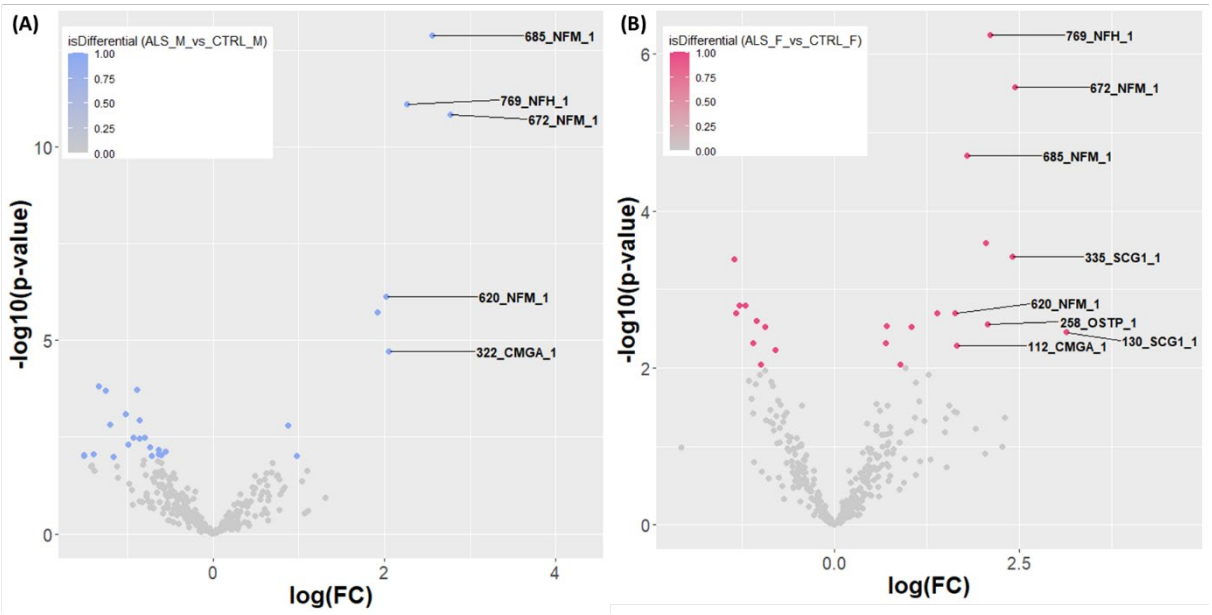


Figure 118: Volcano plots representing the differential analysis of phospho CSF samples (ALS versus CTRL samples) for both sexes, with a 1% p-value filter. All quantified class I phosphosites are represented by dots, colored dots representing DEpS.

In total, 365 class I phosphosites were quantified. Out of those, only a few were differentially quantified between the ALS and control condition with respectively 23 and 28 DEpS for females and males. Among those, five are common between both sexes. Four of them are phosphorylated neurofilament sites, highly over-expressed in the ALS condition. Neurofilaments (NF Heavy, Medium, or Light) phosphorylation is well reported to be linked to various neurological disorders, including ALS[5,190,442,451]. 620_NFM_1, 672_NFM_1 and 685_NFM_1 phosphorylated sites for example were described as up-regulated in Alzheimer's brain samples compared to control samples[452].
Other up-regulated phosphosites in females are 335_SCG1_1 and 130_SCG1_1, on Secretogranin-1 protein. SCG1 (also called CHGB) belongs to the family of granin neuropeptides located in the nervous system and that help modulate both neural activity and synaptic signalling. Granin proteins have been linked to various neurodegenerative diseases and especially in studies on brain and CSF of AD patients[453].

Osteopontin (OSTP), also known as secreted phosphoprotein 1 (SPP1), is expressed by various components including the immune system and the central system. It was shown to play a role in neurodegenerative diseases including multiple sclerosis, Parkinson's disease, AD, FTD and ALS[454,455]. Recently, De Luna *et al.* found that SPP1 levels were significantly higher in CSF of ALS patients compared to healthy controls[456]. Here, we found in females' analysis that one osteopontin phosphosite was up-regulated in our ALS samples (258_OSTP_1). As displayed in **Figure 119**, osteopontin is a protein with a large number of potential phosphorylation. Out of those, we quantified in total 59 osteopontin phosphosites.

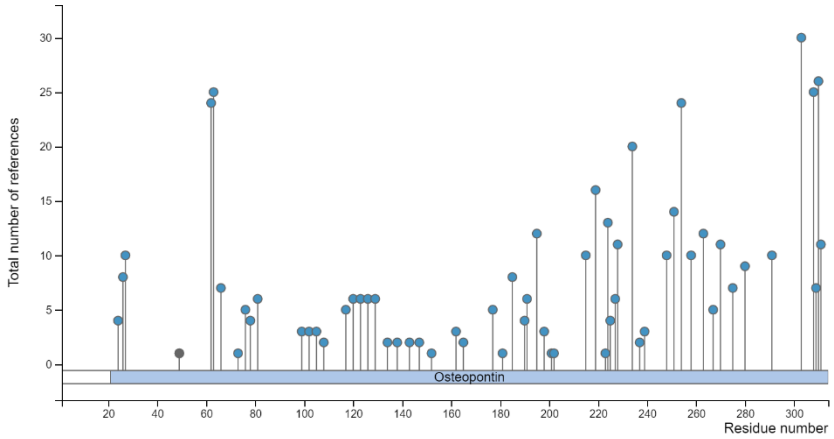


Figure 119: Potential phosphorylated residues of osteopontin and their corresponding number of references, from PhosphoSitePlus®.

In both males and females, the same protein, Chromogranin A (CMGA or CHGA), is phosphorylated and upregulated, while the phosphorylation is not on the same site. In females, 333_CMGA_1 and 112_CMGA_1 sites are phosphorylated, while in males it is 322_CMGA_1. Chromogranin is a protein found in neuroendocrine cells and neurons. It's overexpression was shown to accelerate AMS disease onset in mouse models and increased levels of CMGA were found in CSF of ALS patients[451,457].

In conclusion, we were able here to develop and optimize a complete sample preparation workflow for high throughput (phospho)proteomics analysis of cerebrospinal fluid. We then applied this protocol on large cohorts of clinical samples of CSF. We were able to robustly quantify more than 660 proteins and 365 class I phosphosites in more than 100 CSF samples. By performing a differential analysis on these samples, we found know targets of ALS that have a potential as disease biomarkers. Moreover, we also found differentially expressed proteins/phosphosites that are yet unknown as ALS targets, which could potentially be new biomarkers for ALS. All these potential targets are currently being validated *in-vitro* by a team of biologists from the MAXOMOD consortium to evaluate their potential as biomarkers. Targeted proteomics experiments are also envisioned to be performed on an independent validation cohort.

3. Msqrob2PTM: differential abundance and differential usage analysis of MS-based proteomics data at the post-translational modification and peptidoform levels

i. Context of the project

This side-project was carried out in collaboration with Nina Demeulemeester, Prof. Lieven Clement and Pr. Lennart Martens (CompOmics lab, Ghent University, Belgium).

Thanks to open-modification search engines, LC-MS/MS-based proteomics can detect more post-translational modifications than ever. These developments have the potential to take proteomics research to the next level, as PTMs are key in many cellular processes. However, despite these advances in modification identification, statistical methods for PTM-level quantification and differential analysis have yet to catch up. This is partially due to the inherently low abundance of many PTMs and the confounding of PTM intensities with its parent protein abundance. Therefore, our collaborators have developed msqrob2PTM, a new workflow in the msqrob2 universe capable of differential abundance analysis at the PTM, and at the peptidoform level. In this context, both our proteomics and phosphoproteomics CSF data were shared with them to be used to evaluate the msqrob2PTM method.

ii. Publication

The obtained results were submitted to Molecular & Cellular Proteomics and are currently in review (see **Appendices**).

Chapter 3: Quality controls for proteomics and phosphoproteomics analysis of large cohorts

Despite many technological and computational progresses, proteomic experiments are still prone to a considerable variability, especially when large cohorts of samples are involved. This variability comes from a combination of different factors, represented in **Figure 120**.

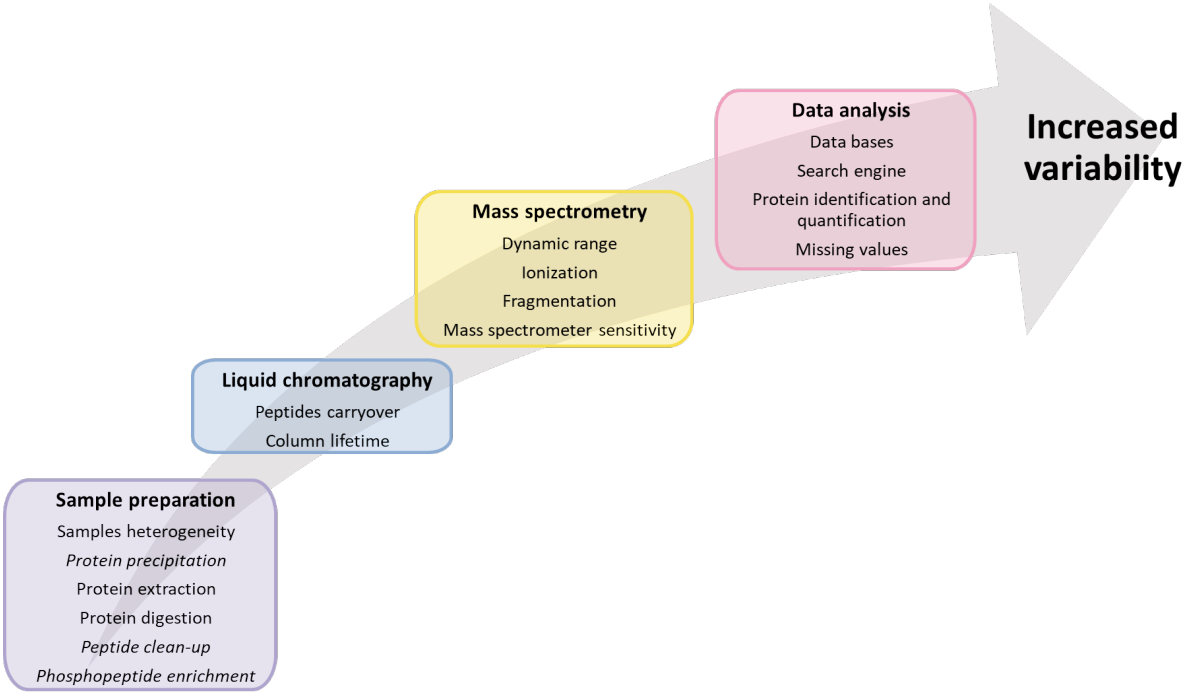


Figure 120: Representation of the different steps of a (phospho)proteomics workflow and their contribution to the global variability of the experiment (adapted from [458,459]).

First of all, part of the variability comes from the inherent samples' heterogeneity, which is further enforced on large cohorts of samples or when multiple biological conditions are considered. All the different steps of sample preparation then further increase this variability, especially when performed manually. Indeed, variations can be introduced through partial inefficiency and/or irreproducibility of the digestion, or through unexpected variable modifications[460]. After sample preparation, peptides are usually separated and analyzed by LC-MS/MS. Here, multiple factors can affect the accuracy of analysis: peptides with poor chromatographic behavior, overloading, cross-contamination due to sample carry-over, presence of contaminants…Moreover, when hundreds of samples are analyzed over a few days or even weeks, sample degradation might appear. Additionally, the state of both the chromatographic system and the mass spectrometer might be variable over such a long period of time. Finally, the accuracy of the final results is also dependent on the bioinformatics pipeline used for data processing. For example, the use of databases, spectral libraries, search engines, used parameters, statistical validation and normalization, all will have an impact on the generated results[460].

To produce reproducible and confident results on large-scale proteomics analysis, it is thus necessary to set up appropriate quality controls (QCs) to keep the control and evaluate the inherent variability of the analysis. Currently, there is no global consensus on a QC methodology and there were only a few attempts to review the different methodologies[458–461]. There are however, some tools developed to evaluate the quality of MS raw data (QC-ART[462], QCloud2[463], PACOM[464]…), proteomic quantitative

data, (pmartR[465,466]) or MaxQuant generated data[467]. These tools generate different degrees of quality assessment (raw data quality, identifications and quantification quality) but none of them includes all levels of assessment. Only recently MaCProQC was developed, a tool a priori enabling evaluation of data metric at all three levels[460].

Here, we will describe the different QCs that we have set up to ensure the reproducibility and robustness of both proteomics and phosphoproteomics analysis of hundreds of clinical samples.

1. Global quality control in mass spectrometry (phospho)proteomics

One of the most common QC samples are whole-cell lysates, used routinely to evaluate the performances of the chromatographic system and of the mass spectrometer. Performance criteria are set in terms of both chromatographic parameters (following the RT, mid-height width, area of a few peptides spread across the chromatographic gradient) and MS/MS acquisition (number of proteins, peptides and PSM spectra…). Once set up, validation criteria and thresholds are established. If the performances fail to achieve the set up thresholds, intervention either on the LC system (change of pre-column or column) or on the mass spectrometer (change of capillary or spray needle, cleaning) are conducted. These QC samples are usually injected before starting each new sequence of samples *ie* at least once a week. When injecting large cohorts of samples over a few weeks, it is recommended to also inject these QC during the sample sequence. It allows to evaluate the performances of the LC-MS/MS throughout the course of injections. Therefore, for our largest cohorts of >100 samples (namely proteomics for human PFC and CSF, and phosphoproteomics for CSF), we injected these QC lysate in between our samples. For proteomics on human brains and CSF, samples were injected on a Q-Exactive Plus platform, with a yeast lysate as QC (200 ng injected). For phosphoproteomics analysis of CSF, samples were injected on a Q-Exactive HF-X, with a HeLa lysate as QC (100 ng injected).
First of all, we evaluated chromatographic performances during the analysis by following four (or five for the Q-Exactive HF-X) peptides. In **Figure 121** are represented the different metrics evaluated for the followed peptides of the QCs, injected at different days. As displayed in **Figure 121**, no shift in retention time of the peptides is observed across days, nor any tailing of the chromatographic peaks as mid-height width are constant across QC injections. Additionally, area under the chromatographic peaks appears also to be repeatable over the different QC injections.

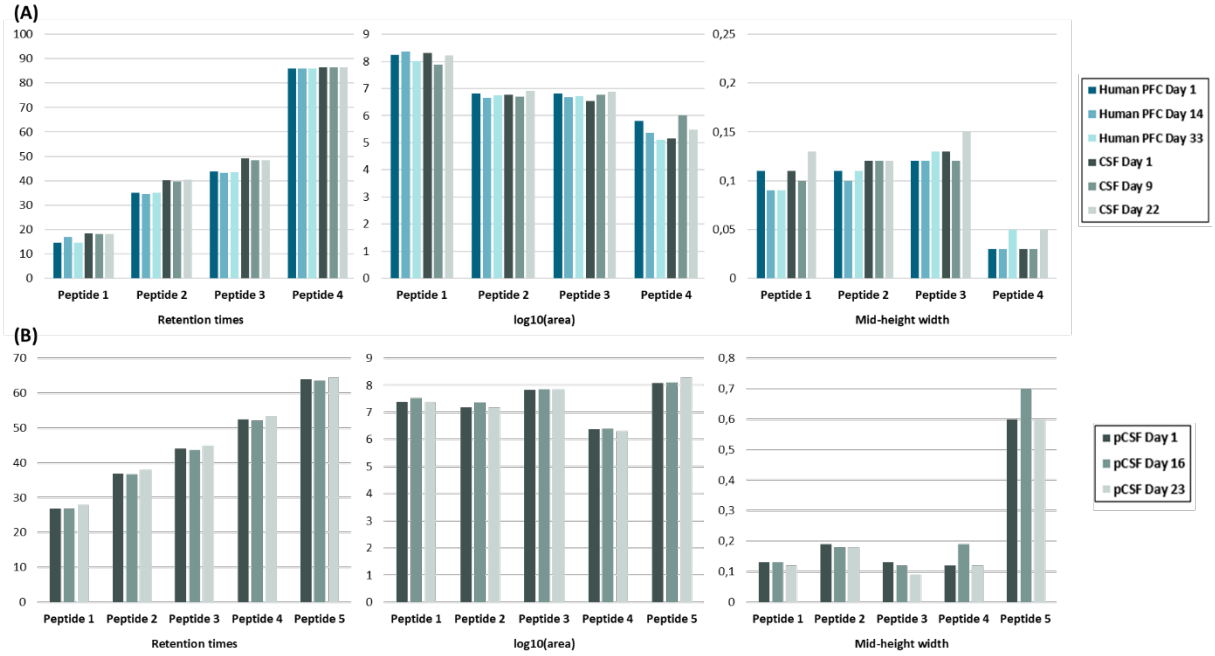


Figure 121: (A) Chromatographic performances of yeast digest quality control on NanoAcquity coupled to Q-Exactive Plus for human PFC and CSF cohorts over sequence analysis (B) Chromatographic performances of HeLa digest quality control on NanoAcquity coupled to Q-Exactive HF-X for phosphoCSF cohort over injection sequence.

Then, qualitative evaluation of the identifications was performed through the following metrics: number of proteins and peptides identified, and number of PSM spectra. As highlighted in **Figure 122**, numbers of identified proteins were almost perfectly reproducible over QC injections. The number of peptides identified and PSM spectra, while not as stable as the number of peptides, were still always above the set thresholds (red lines in **Figure 122**).

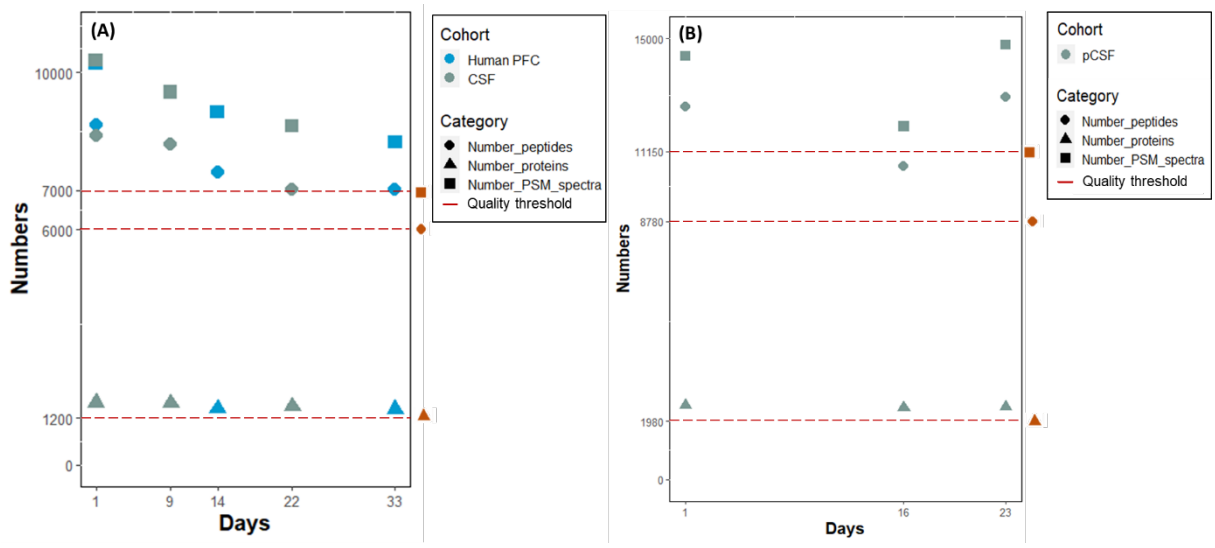


Figure 122: (A) MS/MS performances of yeast digest quality control on Q-Exactive Plus for human PFC and CSF cohorts over injection sequence (B) MS/MS performances of HeLa digest quality control on Q-Exactive HF-X for phosphoCSF cohort over injection sequence.

Thanks to this external control, we were able to assess the performances of the LC-MS/MS system and highlight its overall stability over weeks of injections. However, more complex and more adapted QC

are necessary to fully evaluate the robustness and reproducibility of our analysis. The other quality controls set up during our analyses are detailed in the next sections.

2. Internal and external quality controls for global proteomics analysis

i. External QC: pools of samples

For all proteomics cohorts, a fraction of all samples was pooled before protein reduction and alkylation. The pool was submitted to the same protocol as the biological samples. QC samples, depending on their type, can be incorporated in the injection sequence in various ways. Here, pool samples were interleaved with biological samples to detect potential decreases in performance and avoid any sample loss. In total, for the mouse models, five QC runs were injected. For the human brain cohort, 20 pools were injected and 16 for the human CSF cohort.
Identification numbers are often used as basic indicators of proteomics data quality. In **Figure 123 – (A)** are represented the identified protein counts for all pool samples in the different cohorts. As displayed by this figure, we obtain a reproducible rate of identifications across all different pools, even for larger cohorts. Protein identification count is almost perfectly identical across the pools of mouse models. C9 mouse model shows however slightly higher identification numbers and stands out from other mouse models, in coherence with previous results on this particular model. We also looked at the reproducibility in protein abundance levels (**Figure 123 – (B)**) across pools. Overall, within a dataset, proteins' abundances are quite stable over all pool samples.

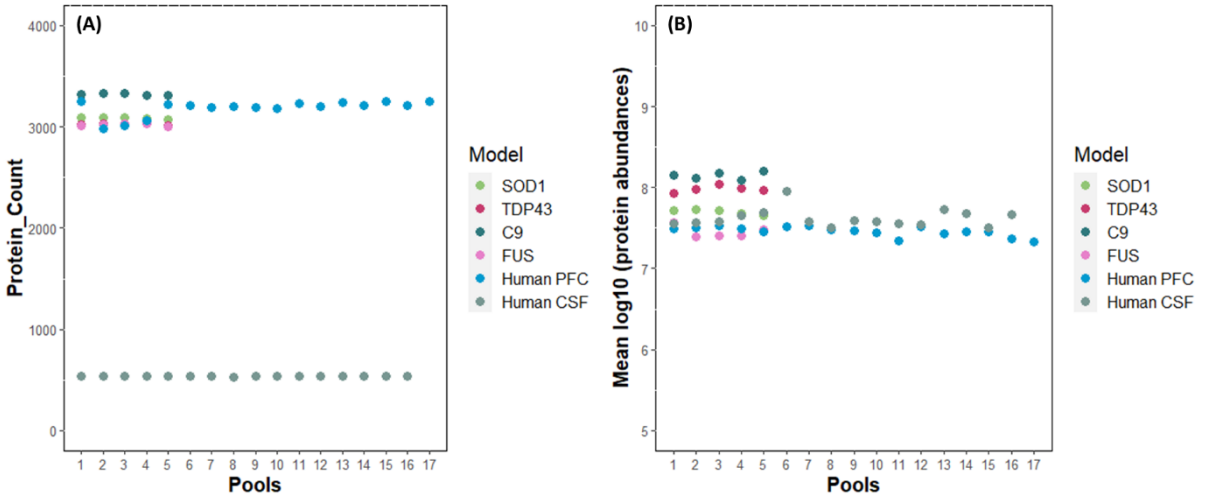


Figure 123: (A) Scatter plot of identified protein counts (B) Scatter plot of average protein abundances.

Proteins' abundances stability is further highlighted in **Figure 124**, in which are represented the distribution of the CVs on the intensities (A) and LFQ intensities (B) of all proteins quantified in all pool samples. By calculating the CV on the LFQ intensities of all pool injections, we are able to highlight the great stability of the system. Median CVs on the intensities ranged between [14 - 33%], with the higher median CVs of 26% and 33% for respectively human PFC and human cohorts, both cohorts larger than 100 samples. Label-Free Quantitation (LFQ) intensities are normalized to exclude potential outliers and best represent the ratio changes in different samples and are thus more reproducible that classical intensities. Median CVs on LFQ intensities were much smaller and ranged between [6 – 15%]. The distribution of CVs on the LFQ intensities is also really narrow, highlighting the great stability of the system even across hundreds of samples.

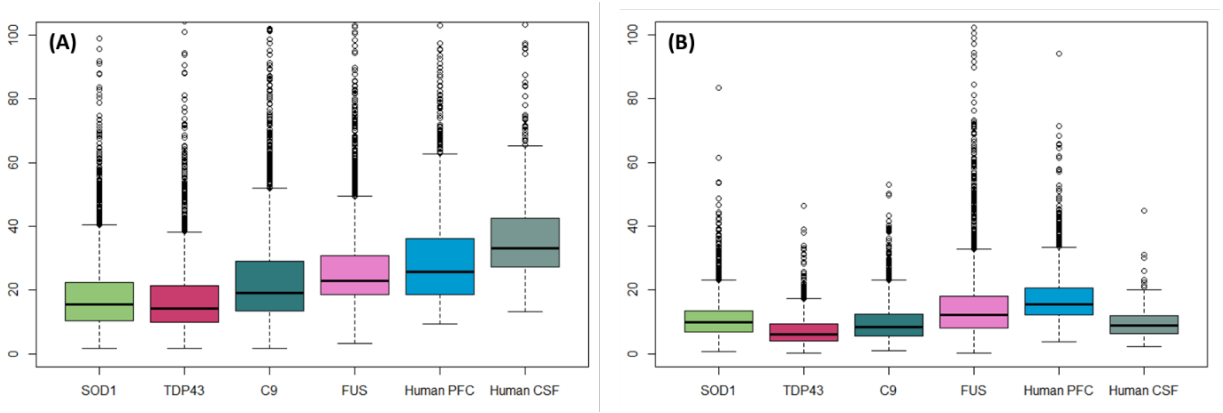


Figure 124: (A) Boxplots representing the distribution of CVs on the pools' intensities for the different cohorts (B) Boxplots representing the distribution of CVs on the pools' LFQ intensities for the different cohorts.

ii. Internal QC: iRT synthetic peptides

Another type of QC is the mixture of synthetic peptides. Theses mixtures can either be run individually or spiked into the QC pool samples and/or biological samples. One of the main advantages of spiking in those synthetic peptides is that quality control is performed in real-time with the analysis. A direct link between the qualitative information and the experimental data can thus be established[458]. Spiked-in peptides should not overlap with the signal of the biologically relevant peptides. For this reason, artificial, synthetic peptides or isotopically labelled peptides might be used. Here, we used indexed Retention Time (iRT) synthetic peptides (Biognosys), which are a mixture of 11 non-naturally occurring synthetic peptides. In comparison to experimental RT, the iRT value of a peptide is stable and enables accurate prediction of peptide retention for any chromatographic setup. These iRT peptides were added to each of our proteomic samples, including pool samples, to check retention time alignment over all our injections.

Out of the eleven synthetic peptides of the iRT mix, we were able to systematically detect nine of them in all of the proteomics cohorts (mouse brains, human brains and human CSF). For each of those peptides, we calculated the CV of their retention time over all the samples (**Figure 125**). For all peptides and across all cohorts, the CV was lower than 4%, and for most of them even lower than 2%. These results highlight the stability of the LC system over weeks and hundreds of injections.

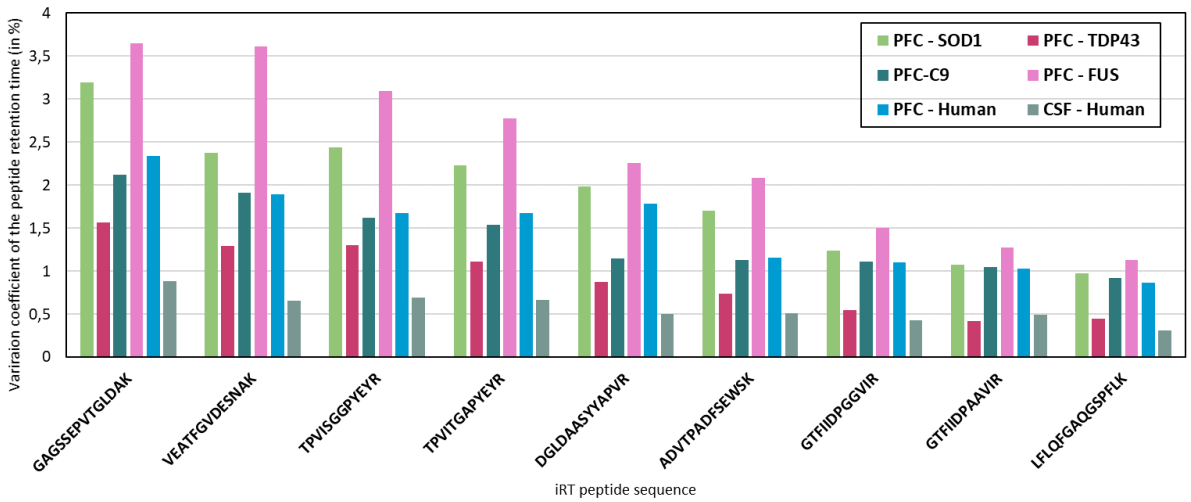


Figure 125: CVs (in %) on the retention time (RT) of the different iRT synthetic peptides in all samples of the different proteomics cohorts.

We also calculated the median CV of iRT peptides intensity values across all samples injections for the different cohorts of samples. We obtain median CV between 14% at the lowest (for FUS mouse model) and 38% at the highest (for CSF cohort). By comparison, on mouse liver tissues, Imbert *et al.*[468] obtained a 40% median CV on the 11 iRT synthetic peptides over 42 samples.

As all the synthetic peptides of the mix have indexed retention times that are known and referenced, we then evaluated the correlation between the reference iRT and the measured retention time of the synthetic peptides. For each cohort of samples, we averaged across samples the experimental retention time of each synthetic peptide and plot it against its iRT. In **Figure 126** the linear line represents the correlation between the two retention time for each synthetic peptides.

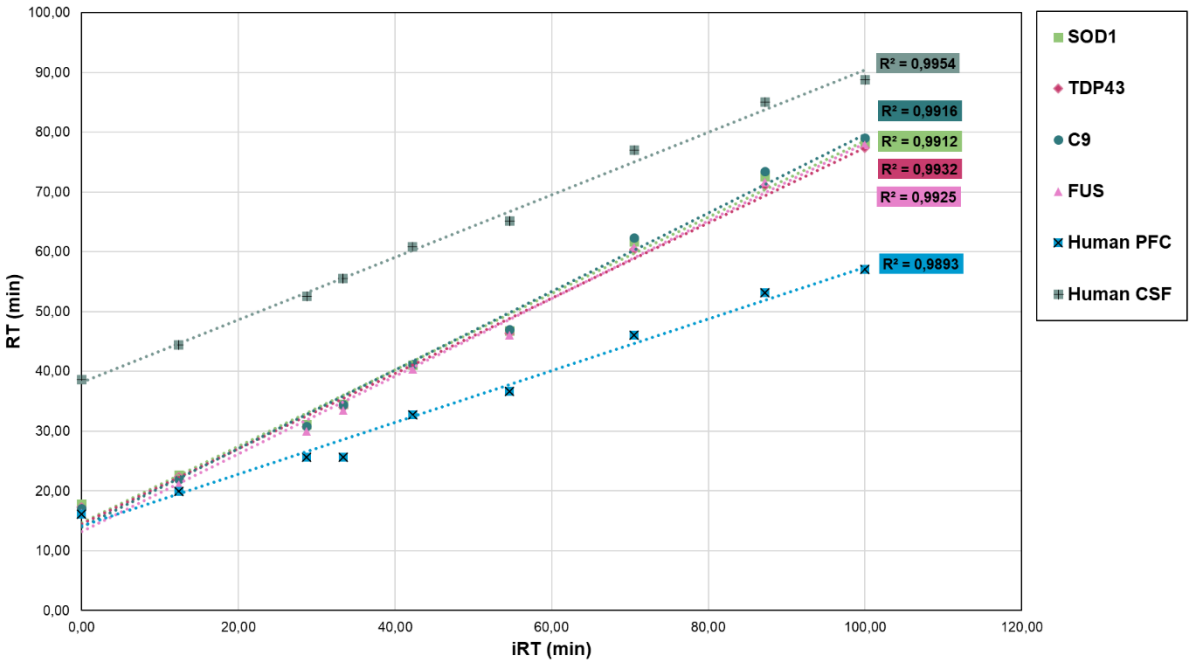


Figure 126: Correlation between Indexed Retention Time (iRT) of synthetic peptides and their experimental Retention Time (RT).

For all cohorts, we obtain excellent correlations, as all correlation coefficients (r) are above 0.99, from 0.9946 at the lowest for human PFC up to 0.9977 for human CSF. Determination coefficients (R^2) are also almost all greater than 0.99, demonstrating the quality of the linear regression. We notice that the human CSF cohort, while it as the highest correlation, also has the greatest shift between expected iRT and measured RT with an ordinate at origin of almost 40 minutes. This means that while retention times appears shifted from their expected values, this shift was consistent and reproducible across the whole gradient and across all samples.

We also can note that all mouse cohorts have really have similar correlations results. It is expected as the exact same chromatographic gradient was used for all mouse models and thus synthetic peptides are expected to have a similar behavior, while human PFC and CSF were analyzed with a slightly different gradient.

All these parameters together emphasize both stability and robustness of the chromatographic system across our different cohorts of samples.

3. Specific quality controls for phosphoproteomics analysis

i. *Performances of the phosphopeptide enrichment*

For the phosphoproteomics sample preparation, we used adapted phospho quality controls. The Phosphomixes are mixes of synthetic phosphopeptides derived from naturally occurring peptides in HeLa cells. They were used to evaluate the efficiency and reproducibility of the phosphopeptide enrichment process, as they were added before (Phosphomix light, in their naturally occurring isotopic abundance) and after the phosphopeptide enrichment process (Phosphomix heavy, in their stable isotope enriched version). Out of the ten phosphopeptides of the Phosphomix, we were able to detect five of them across our different cohorts, all of them described as having either a strong or medium relative signal intensity by instructor's information (**Table 12**, Sigma-Aldrich, Product information, MS Phosphomix data sheet).

Peptide*	Total # of Phosphates	MW Light (Monoisotopic)	MW Heavy (Monoisotopic)	Relative Signal Intensity***	# of Phosphates per amino acid		
					S	T	Y
PhosphoMix 1							
VLHSGpS[R]	1	834.37	844.38	Weak	1		
RSpYpSRS[R]	2	1070.41	1080.41	Weak	1		1
RDSLGpTYSS[R]	1	1220.52	1230.53	Medium		1	
pTKLpTQLRDA[K]	2	1445.70	1453.72	Strong		2	
EVQAEQPSSpSSP[R]	1	1480.62	1490.63	Medium	1		
ADEPpSSEESDLEID[K]	1	1742.68	1750.69	Strong	1		
ADEPSpSEEpSDLEID[K]	2	1822.64	1830.66	Medium	2		
FEDEGAGFEESpSETGDYEE[K]	1	2333.84	2341.85	Strong	1		
ELSNpSPLRENSFGpSPLEF[R]	2	2338.00	2348.01	Medium	2		
SPTEYHEPVpYANPFYRPTpTPQ[R]	2	2809.19	2819.20	Strong		1	1

Table 12: MS Phosphomix 1 data sheet (Sigma-Aldrich, Product information). Highlighted in green, the phosphopeptides detected in all 5 cohorts of samples. In blue, an additional phosphopeptide detected in both TDP43 and C9 mouse models. In pink, an additional phosphopeptide detected in FUS mouse model.

Using Skyline software (v.22.2.0.351), we extracted the intensities of the light and heavy versions of those peptides for each cohort and determined the ratio $\frac{light\ peptide\ intensity}{heavy\ peptide\ intensity}$ in order to look at the phosphopeptide enrichment efficiency.

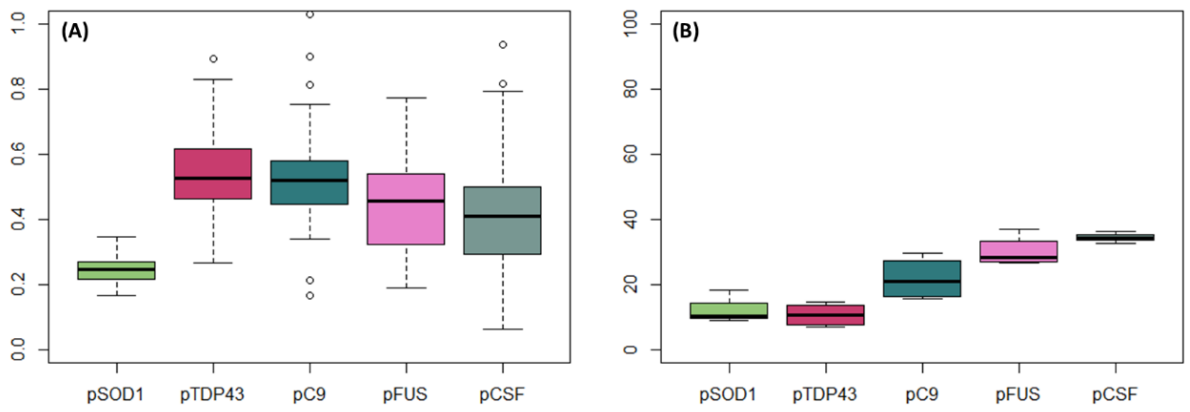


Figure 127: (A) Boxplots representing the ratio of phosphopeptide enrichment for each cohort (B) Boxplots representing the distribution of CVs on the ratio of phosphopeptide enrichment for each cohort.

In **Figure 127 – (A)** are represented the ratio of peptides that were enriched as the ratios $\frac{light\ peptide\ intensity}{heavy\ peptide\ intensity}$. For most cohorts, we were able to enrich between 41% to 53% (median values) of our peptidic material. Only the phosphopeptide enrichment of the SOD1 cohort seems less efficient with only a median ratio of enrichment of 24%. One explanation of this is the use of an old batch of Phosphomix for the SOD1 experiment, which was replaced by a new batch for other experiments once we saw the performances for the SOD1 model.

In a study on a human breast cancer cell line [220], they performed an automated IMAC phosphopeptide enrichment on an AssayMAP Bravo platform on four samples of tryptic digest of purified cancer cell lines, pre-spiked with Phosphomix I and II. Out of the ten Phosphomix peptides, they were able to detect only six peptides and obtained a recovery ratio of 0.63 over all Phosphomix peptides and samples. Considering the much larger size of our cohorts and our increased sample complexity, we can conclude our enrichment process was overall efficient.

To ensure that our phosphopeptide enrichment was reproducible, we also looked at the CVs of the ratio of enrichment (**Figure 127 – (B)**). All median CV were found lower than 40%, and lower than 20% for SOD1, TDP43 and C9 with median CVs of respectively 10%, 11% and 20%. These results highlight the great reproducibility of the phosphopeptide enrichment step, thanks to automation, even on large cohorts of samples.

ii. Stability of the LC-MS/MS system

We then used the three Phosphomix phosphopeptides that were detected reproducibly in every cohorts as internal standards.

In order to evaluate the stability of the chromatographic system, we represented the CVs on the retention times of the different synthetic phosphopeptides in each cohort (see **Figure 128 – (A)**). For all mouse models, almost all CVs were below 1%, stressing the great stability of the LC system. Moreover, even for the large CSF cohort, CVs are all below 2%. This leads to the conclusion that the chromatographic system is highly stable over hundreds of injections.

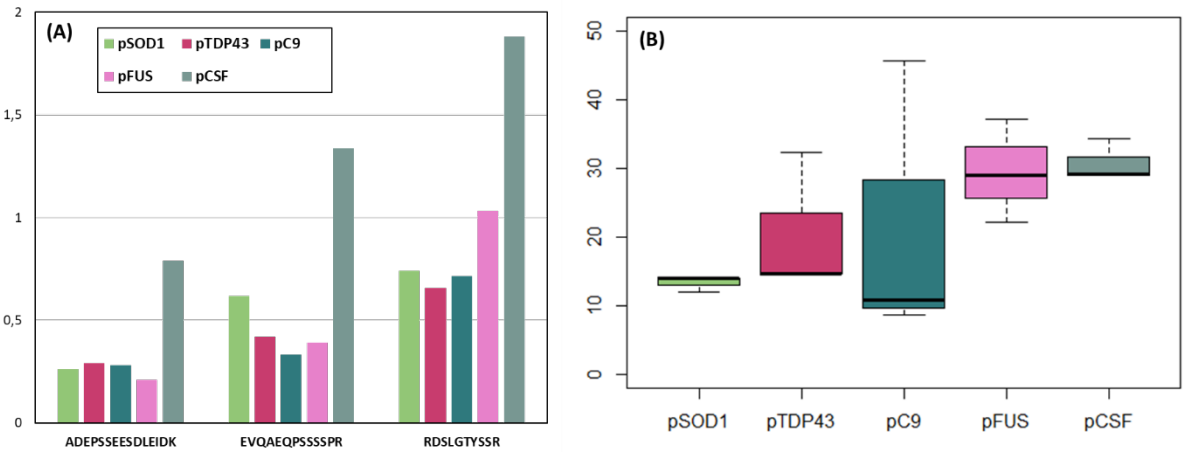


Figure 128: (A) CVs (in %) on the retention time of the different Phosphomix phosphopeptides in all samples of the different phosphoproteomics cohorts; (B) Boxplots representing the distribution of CVs on the heavy Phosphomix phosphopeptides intensities for the different cohorts.

Figure 128 – (B) represents the distribution of the CVs on the intensities of the heavy synthetic phosphopeptides across the different cohorts. For all cohorts of samples, except for C9, the distribution of CVs is quite narrow. The distribution for C9 is wider, which is coherent with the higher global heterogeneity observed for this mouse model. For SOD1, TDP43 and C9 mouse models, we obtain median CVs below 20%. While even for the largest cohort of CSF samples, median CV is below 30%. These results emphasize the robustness of the analysis and signal stability

Chapter 4: Open Modification Searches software evaluation to increase phosphorylation identifications

As detailed previously, a large amount of spectra usually are unmatched in a proteomics analysis. One of the reason for this is the restricted search space of current engines that can therefore not identify peptides with unexpected modifications. Indeed, if a particular modification has not been specified in the search settings, then spectra corresponding to peptides bearing this modification will be assigned to an incorrect amino acid sequence[15]. In this perspective, Open Modification Searching (OMS) tools have been developed to identify modified spectra. OMS software are thus especially promising for PTMs studies.

Here, we decided to evaluate the performances of one OMS tool, IonBot[16], on our different phospho-enriched mouse brain tissues datasets. IonBot, developed at the CompOmics laboratory (Ghent University, Belgium) combines machine learning-based algorithms in order to predict both peptide RT (with DeepLC[172]) and MS2 peak intensities (with MS2Pip[173]). Then through fully machine learning based re-scoring, it performs sensitive and accurate identifications of (modified) peptides. IonBot (v.0.10.0) performances will be benchmarked against previously used MaxQuant (v.1.6.14) at the identification and localization levels.

1. Improved identification with open modification searching

The total numbers of phosphoproteins, phosphopeptides and phosphosites identified with the two different pipelines and on the different mouse models are represented in **Figure 129**. Here, to have the same definition of a phosphorylation site between MaxQuant and IonBot, a phosphosite was defined as the protein carrying the site and the position of the site within the protein.

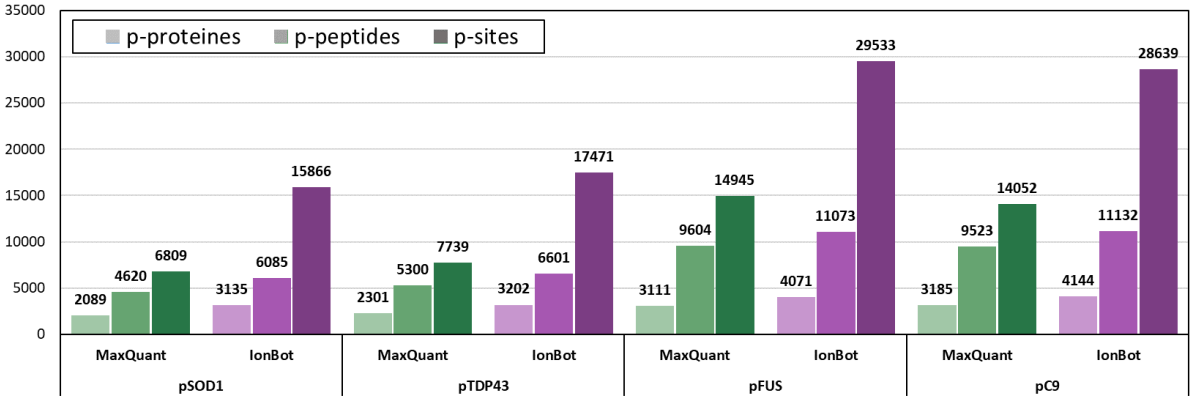


Figure 129: Phospho-identification results for the different mouse models using either Andromeda (MaxQuant) or IonBot for the search.

Depending on the model, IonBot allows to increase phosphoproteins identifications from 30% to 50%. Going at the phosphopeptides levels, the gap is a bit smaller between the two pipelines with between 15% and 30% of additional phosphopeptides identifications with IonBot. The increase is even more impressive at the site level with on average an additional 10 000 phosphosites identified thanks to IonBot. While IonBot increases in phospho-identifications are quite impressive, we wanted to investigate these additional identifications.

2. Populations of peptides identified

For this, we combined identifications from the different mouse models and looked at the coverage in identifications between the two pipelines.

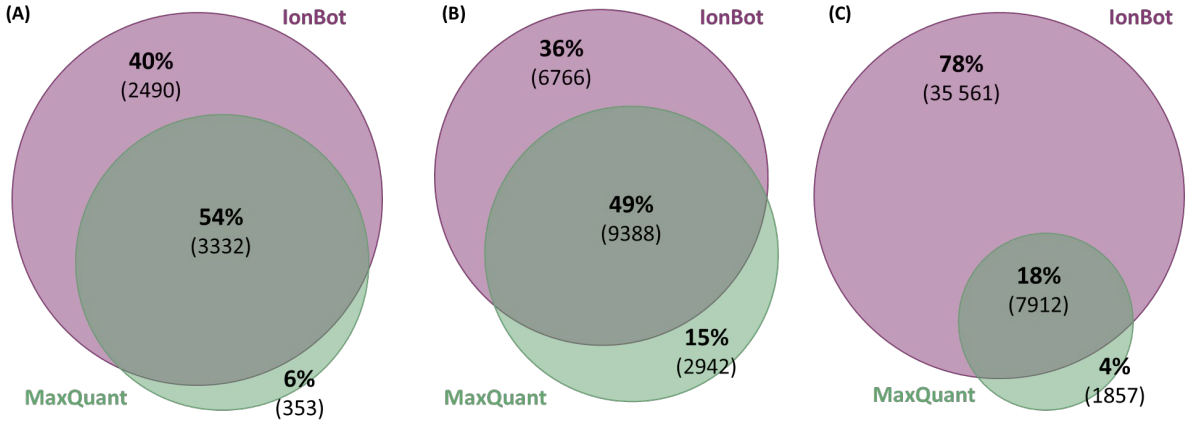


Figure 130: Overlap of phospho-identifications between the two pipelines at (A) the phosphoprotein level (B) the phosphopeptide level (C) the phosphosite level.

Figure 130 represents the identification overlap between IonBot and MaxQuant. While more than half of the phosphoproteins are shared between the two software, IonBot identified an additional 40% of phosphoproteins (**Figure 130 – (A)**). Results are similar at the peptidic level with also half of the phosphopeptides identified by both MaxQuant and IonBot, and 36% of additional phosphopeptides using IonBot (**Figure 130 – (B)**). However, less than 20% of all identified phosphosites are shared between the two pipelines (**Figure 130 – (C)**). IonBot allows for the impressive additional identification of more than 35 000 phosphosites.

We then focused on the 2942 phosphopeptides identified only by MaxQuant and tried to found out why they where not found by IonBot.

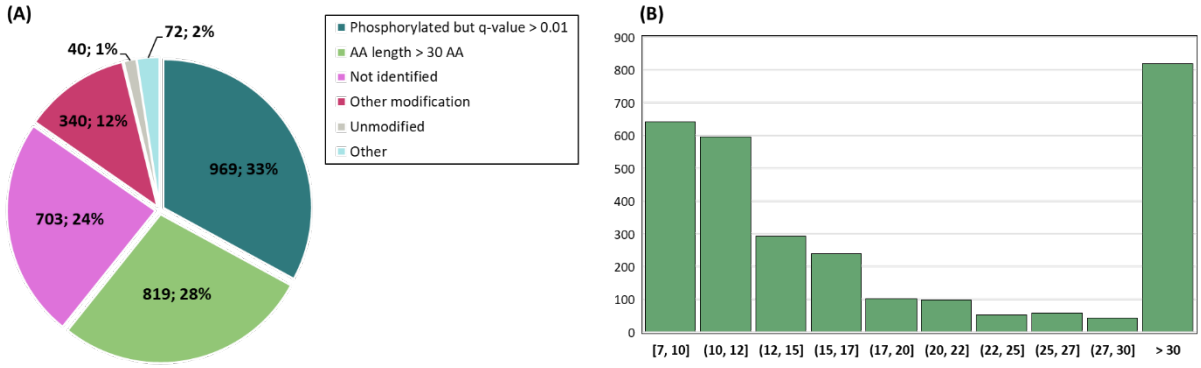


Figure 131: (A) Distribution of the characteristics of the phosphopeptides only identified by MaxQuant (B) Amino acid length of the phosphopeptides identified solely by MaxQuant.

As represented in **Figure 131 – (A)**, the majority (75%) of the phosphopeptides identified by MaxQuant only, are also found by IonBot but because of some other characteristics, did not make into the final list of identified phosphopeptides. First of all, we looked at the length of those 1332 phosphopeptides (**Figure 131 – (B)**). Out of them, 819 *ie* 28% have more than 30 amino acids (AA). While there is no upper limit of the number of amino acids in MaxQuant, it is limited to 30 AA maximum in IonBot. Although this length limitation prevents us here to identify those 819 phosphopeptides, it is set to limit the search space thus the overall open search time. Then, about one third of them were also identified

as phosphorylated peptides in IonBot but had a q-value > 1% so did not make it through the quality filter. Finally, 340 of them were identified by IonBot but with another modification, while 40 of them were found unmodified by IonBot.

3. Identified modifications in IonBot

IonBot allows identifying all UniMod modifications and any individual sequence variant. By comparison, MaxQuant will only identify modifications set up a priori in the search parameters. These modifications were in our case: Carbamidomethylation of cysteines, acetylation of protein N termini, oxidation of methionines, and phosphorylation of serines, threonine or tyrosine residues. This explains why IonBot identifies much more modified peptides and thus more modified peptides. In total, 87 815 different modifications were identified by IonBot. The top modifications and the number of peptides identified with these modifications are represented in **Figure 132**.

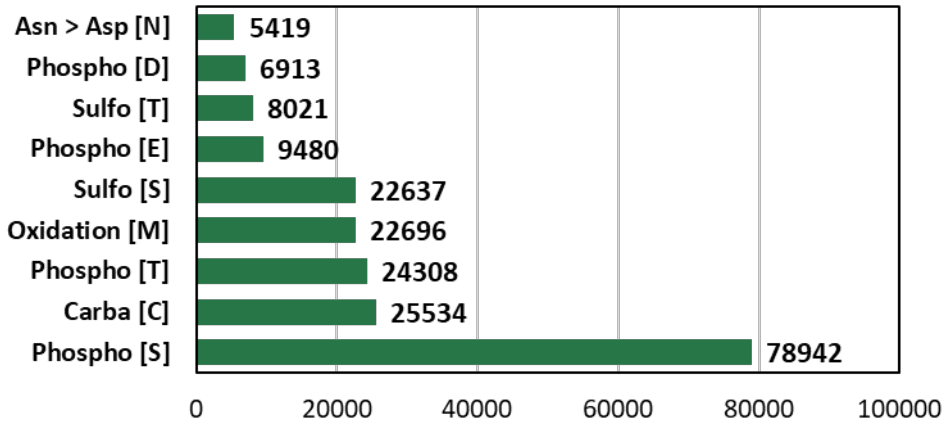


Figure 132: Nine most abundant modifications in IonBot and their corresponding number of identified modified peptides.

Reassuringly, phosphorylated peptides come at the top of le list. Carbamidomethylation and oxidation are also among the most frequent modifications on peptides. However, it is interesting to notice that it appears that some phosphorylated peptides are phosphorylated on other amino acids such as glutamic acid [E] or aspartic acid [D]. Phosphorylation happens mainly on serine [S], threonine [T] and tyrosine residues [Y] and they constitute together the family of phosphoesters or O-phosphate, as these all bind through the oxygen of the residues. While they happen less often and are less studied, six other naturally occurring amino acids can be phosphorylated, with 3 other types of linkages. Altogether, they are name the SONA, for S-, O-, N- and A- phosphate linkage families[469]. In particular, the A-phosphate (acyl phosphate) family is composed of phosphoaspartate and phosphoglutamate. It is thus interesting to look more in details on the different amino acids that are found phosphorylated thanks to IonBot.

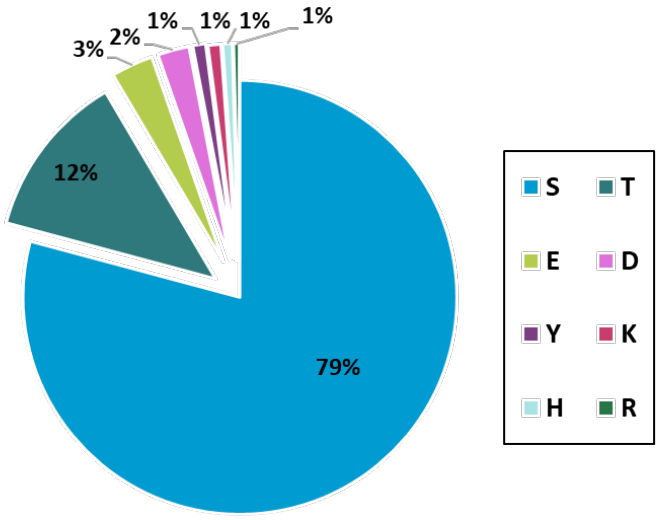


Figure 133: Repartition of phosphorylation on different amino acids at the PSM level.

As represented in **Figure 133**, the majority of identified phosphorylations are on serine residues (79%). Phosphorylations of the N-phosphate (phosphoramidates) family are also present, with 12% of phosphorylated arginines, and a lower presence of lysine and histidine phosphorylations. Nevertheless, aspargin and glutamin phosphorylations, observed previously, represent less than 2% of the total of phosphorylated PSMs.

While those additional phosphorylated residues might increase global phospho-identifications, their relevance needs to be further investigated. In this specific context, localization probability represents a variable of choice.

General conclusion

General conclusion

My PhD work intended to develop high-throughput proteomics and phosphoproteomics methods applicable on large cohorts of samples for biomarker discovery.

The opening part of this manuscript consisted in a bibliographic study on the current state-of-the-art of quantitative proteomics. The classical bottom-up label free workflow is presented, with specific focus on its key steps: sample preparation, peptides separation, mass spectrometry analysis, protein identification, validation and quantification. The specific challenges of phosphoproteomics are detailed, along with their potential answer. Emphasis was put on data independent acquisition mode, its principle, evolution and challenges, and its promises to improve phosphoproteomic analysis. An overview of multi-omic approaches and their associated hurdles was finally presented.

In this context, the objectives of this PhD work were the following:

- The development of an automated sample preparation workflow for phosphoproteomics
- The optimization of LC-MS/MS methods for phosphopeptides' analysis using different acquisition modes
- The benchmarking of both DDA and DIA software tools and their performances for identification quantification, and localization of phosphorylation events
- The reproducible and high-throughput analysis of large cohorts of various types of samples for the identification of biologically relevant ALS targets

The first part of the results presented was focused on the developments of an automated and high-throughput phosphoproteomics workflow. Starting from the beginning of the workflow, the influence of protein extraction and digestion protocols were evaluated on bovine brain samples. Results revealed that, while a SP3-based protocol was kept for practical reasons, a urea/thiourea-based protocol allowed for highest phosphoproteomics identifications.

I also benchmarked phosphoproteomics performances of different nanoLC-MS/MS platforms and fragmentation techniques. Because of its PASEF technology and 4-dimensions separation ability, allowing for increased sensitivity, almost perfect duty cycle and high depth of analysis, the TimsTOF Pro proved to be the ideal choice for phosphoproteomics analysis. Through the evaluation of various acquisition parameters of both a dda- and diaPASEF methods, an optimized MS/MS method was set up for phosphopeptides analysis. While the optimized ddaPASEF method allowed for improved identification of phosphopeptides, the optimized diaPASEF method indisputably outperformed with the highest depth and coverage of phosphoproteome achieved.

As data analysis is a critical step of every workflow, I then evaluated various data treatment pipelines for identification, quantification and localization of phosphorylation sites. For DDA data analysis, four different pipelines were benchmarked. While Proteome Discoverer performances were the highest in terms of identifications, better localization of phosphorylation sites was achieved with MaxQuant. For diaPASEF phosphoproteomics data, performances of Spectronaut and DIA-NN were compared. We showed that DIA-NN outperformed Spectronaut for phosphopeptides counts but both software displayed similar quantification reproducibility. However, one crucial point is to be noted concerning both MaxQuant (for DDA generated data) and Spectronaut (for DIA generated data). They

are, at the time this work was performed, the only software that allow for quantification of phosphorylation at the site level.

The second part of my PhD results were centered around the multi-omic E-Rare MAXOMOD project, focusing on ALS biomarker discovery. In the context of this project, I performed global proteomics analysis of large cohorts of mice and human brain tissues, which led to the robust quantification of thousands of proteins. The previously optimized phosphoproteomics sample preparation workflow was applied on mice samples and allowed for the robust quantification of up to almost 6000 class I phosphosites. The differential analysis, for both proteomics and phosphoproteomic tissues analysis, highlighted the complexity of ALS disease as only a few significant differences were observed.

For this project, a common sample preparation for proteomics, phosphoproteomics and metabolomics analysis of CSF samples was needed. I thus benchmarked different sample preparation protocols, and demonstrated that in-solution digestion using RapiGest was the most adapted protocol for (phospho)proteomics analysis of large cohorts of clinical CSF. This optimized protocol led to the robust quantification of thousands of proteins and class I phosphosites.

Then, as every (phospho)proteomics analysis is prone to various sources of variability, I set up different quality controls to ensure the control this inherent variability and the quality and reproducibility of our analyses. Different external quality controls (HeLa cell digest, pool of samples) and internal quality controls (iRT synthetic peptides) highlighted the stability and robustness of the LC-MS/MS systems over weeks and hundreds of injections. For phosphoproteomic analysis, synthetic phosphopeptides were used to evaluate the efficiency and reproducibility of the phosphopeptides enrichment step. They showed the great reproducibility of the enrichment step over thousands of samples thanks to automation.

Finally, I evaluated the performances of an open modification search tool, IonBot, to improve the phosphoproteome coverage. On average, we were able to increase identifications by 50% and identified up to 30 000 phosphosites. This is only a preliminary work and will need further investigations, but the results already achieved look very promising.

Overall, the work detailed in this manuscript highlights the relevance of analytical developments, at each level of the workflow, for both proteomics and phosphoproteomics analysis.

To conclude, I would like to insist on some points and perspectives about proteomics.

A huge part of my work was focused on developments of sample preparation. It is important for me to point out that, while I discussed here some optimized methods on different types of samples, one should know that there is no universal method for sample preparation. Depending on the analysis and the type of sample, sample preparation optimization should be performed, as one protocol might produce different performances on different types of samples. Additionally, emphasis needs to be put on automation, which is key for sample preparation on large cohorts of samples and low amount of starting material. Considerable efforts are done to reduce considerably quantities of starting material in (phospho)proteomics, with the aim of high depth analysis of single cells. I look forward to see more

and more applications and developments of single cell analysis, especially in the field of single cell phosphoproteomics and single cell multi-omics.

I also want to stress out the current popularity of DIA-based methods. One reason for this is that DIA continuously proves its ability to combine in a single analysis the performances of both global and targeted approach for detection and quantification of proteins. Moreover, taking advantage of the strengths of the TimsTOF Pro, diaPASEF methods have shown more than promising results for both proteomics and phosphoproteomics analysis. Indeed, thanks to the ion-mobility dimension, diaPASEF allows distinguishing signals from peptides that co-elute and would otherwise be co-fragmented, thus producing cleaner spectra. New DIA scan modes are emerging (synchroPASEF, slicePASEF, midiaPASEF…) with the promises to achieve even higher sensitivity and precision of analysis.

The third point I would like to mention is that, as of today, data treatment remains a bottleneck of proteomics analysis, especially with DIA and phosphoproteomics. While progresses are made with the latest versions of software, there is still major room for improvement. For phosphoproteomics especially, there is for instance no clear consensus on the definition of a phosphosite. This results in software and users with different definitions, and thus non-comparable results from one to another. In addition, while identification of phosphosites is routine, quantification at the site level is not yet possible with all software. Moreover, one huge challenge of phosphoproteomics is the localization of the phosphorylation site and its validation. Probability of localization needs to be computed to ensure reliability of the results, usually through False Localization Rate computation. However, only a few software and tools currently allow for this kind of calculations. I believe however that the numerous improvements emerging in bioinformatics, *e.g.* prediction algorithm, open search tools, deep learning and artificial intelligence, should greatly improve and facilitate data treatment in proteomics in a near future.

However, there is one matter I would like to discuss with the emergence of new bioinformatics tools and software solutions for proteomics. With more and more tools emerging, each based on different algorithms, with different parameters tunable by the user, with different validation methods, and different versions released frequently, it becomes more and more difficult for a user to compare the results. To choose the most adapted pipeline for data analysis, users thus should not be afraid of digging into the their data, to ensure they have comparable results. In addition, with the race for numbers that is sometimes observed, my advice keep a critical eye with the results that might be presented in publications or conferences. Indeed, there is a gap between the type of validation needed on data generated from classical cell digest for example, to the validation needed on more complex clinical samples. This difference in validation of the data might greatly impact the numbers and results obtained. Therefore, no matter how impressive the numbers might be, one should always keep an eye for the type of sample the experiment was performed on, but mostly on if and how the presented numbers were validated.

One last topic I would like to point out about proteomics data, or scientific data in general, is data sharing. While data sharing is more than useful to the scientific community and has been common practice for a few years in proteomics, it critically lacks metadata annotation. This lack of metadata annotation makes it a challenge for researchers to fully understand the context of the data and thus greatly limits data reanalysis. In the case of proteomics data sharing on PRIDE platform, most people go for a partial data submission over a complete one, for which less files and less data annotation are needed. From what I could experience, this is mostly because metadata annotation is quite time consuming, especially on large cohorts of samples. While it may be more work, I hope, with the tools

that are emerging to facilitate data sharing and annotation, that complete data submission will become the standard practice, to ensure data longevity and re-usability.

Last, I would like to bring up the challenges faced when working on large cohorts of samples. Indeed, there is no consensus amongst the scientific community for a global quality control for (phospho)proteomics analysis for instance. Additionally, most publications do not discuss the eventual QC samples and metrics they might have used. For phosphoproteomics, most studies do not give any information on either their efficiency or reproducibility of enrichment. However, I have shown in this work the imperative need to set up some appropriate QC to ensure the overall quality and reproducibility of analysis. I thus hope that sharing this kind of information will be of any use and become more usual in the future.

PART V

Experimental part

Chapter 1: Development of a fully automated high throughput phosphoproteomics workflow

1. Development of a high throughput and automated phosphoproteomics sample preparation workflow

i. Cell lysis, protein assay and protein precipitation

Three different lysis buffers were compared:

- "Urea": with 8 M urea, 0.1 M ABC, protease inhibitors (1:50 v:v) and 1 mM sodium orthovanadate
- "Urea/thiourea": with 6 M urea, 2 M thiourea, 0.1 M ABC, protease inhibitors (1:50 v:v) and 1 mM sodium orthovanadate
- "Laemmli-like": with 50 mM Tris-HCl pH=6.8, 10% glycerol, 2.5% SDS, 1 mM EDTA, protease inhibitors (1:50 v:v) and 1 mM sodium orthovanadate

The urea and the urea/thiourea buffers were used with and without a MeOH/CH_3Cl_3 precipitation step. In total, 5 conditions were tested as represented in **Figure 134**.

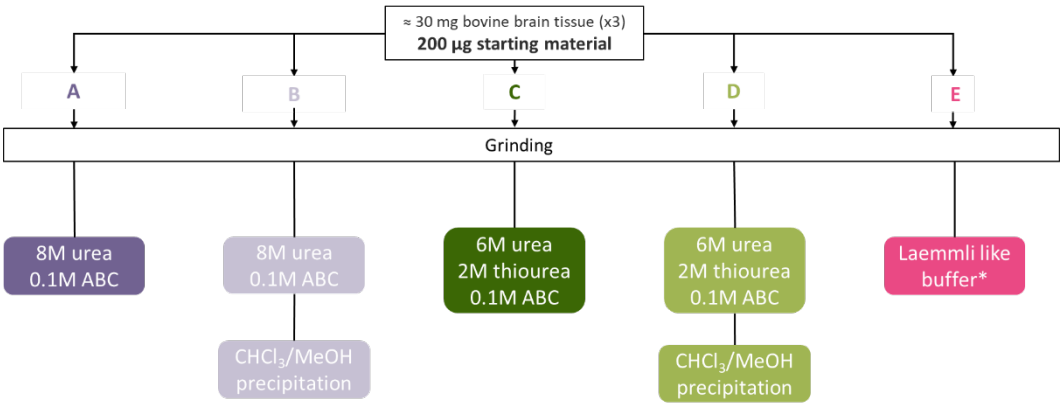


Figure 134: Five different protocols compared: A = 8M urea; B = 8M urea with precipitation; C = 6M urea, 2M thiourea; D = 6M urea, 2M thiourea with precipitation; E = Laemmli-like buffer.

Frozen bovine brain samples (N= 3 replicates per condition) were grinded using Kimble™ BioMasher II (Dutscher, Bernolsheim, France). Samples were then resuspended in 350 µL of lysis buffer. For conditions B and D, protein precipitation was performed adding 4 volumes of ice cold MeOH, 1 volume of CH_3Cl_3, and 3 volumes of H_2O. After centrifugation (4°C, 5000 g, 30 min), proteins were at the interface between two phases. The upper phase was removed and 3 volumes of ice cold MeOH were added. Samples were centrifuged (4°C, 5000 g, 10 min) and pellets washed with 1 mL of ice cold MeOH before being centrifuged again (4°C, 5000 g, 5 min). Protein pellets were then resuspended in 350 µL of lysis buffer. Protein concentration was assessed using RC-DC assay (Bio-Rad, Hercules, USA).

ii. Reduction and alkylation

For each replicate, 200 µg of proteins were reduced for 30 min at 37°C using 20 mM DTT, 0.1 M ammonium bicarbonate (ABC) to reach a final DTT concentration of 12 mM. Protein alkylation was performed for 1 h at room temperature in the dark using 700 mM IAM, 0.1 M ABC to reach a final IAM

concentration of 40 mM. For conditions A, B, C and D, samples were diluted with 0.1 M ABC buffer to decrease urea concentration under 1 M.

iii. Conditions A to D: in-solution digestion and peptide clean-up

For conditions A to D, protein digestion was performed overnight at 37°C by adding trypsin/Lys-C (Mass Spec Grade, Promega, Madison, WI, USA) at a 1:50 ratio (enzyme:protein, m:m). Digestion was stopped by TFA addition to reach a pH <3.
Generated peptides were purified by automated SPE: peptide clean-up v2.0 protocol was loaded on AssayMAP Bravo platform (Agilent Technologies, Santa Clara, CA, USA). Briefly, 5 µL C18 cartridges were washed and primed with 50% ACN, 0.1% TFA and equilibrated with 0.1% TFA acidic solution. Samples were then loaded on the cartridges at 5 µL/min, washed with 0.1% TFA and then peptides were eluted at 5 µL/min with 40 µL of 70 µL ACN, 0.1% FA.

iv. Condition E: SP3 digestion

On-bead digestion protocol was used for condition E with the Laemmli like buffer. Beads A (Sera-Mag Speed, Thermo Fisher Scientific, 45152105050250) and beads B (Sera-Mag Speed,Thermo Fisher Scientific, 65152105050250) were combined (ratio 1:1) and washed 3 times with ultra-pure water. A solution at 100 µg/µL of beads was then prepared. The bead mixture was added to the samples with a bead:protein ratio of 10:1 for each type of beads thus of 20:1 for the combination of beads. Pure ACN was added to reach a 50% final concentration and proteins bound to the beads for 18 min at room temperature. Samples are then incubated for 2 min on a magnetic rack to remove supernatant. Proteins were washed twice with 200 µL of 80% EtOH and once with 180 µL of 100% ACN. Proteins were then resuspended in 90 µL of 0.1M ABC and digested on the beads overnight at 37°C and 1 000 rpm using trypsin/Lys-C (Mass Spec Grade, Promega, Madison, WI, USA) at a 1:20 ratio (enzyme:protein, m:m). Digestion was stopped by TFA addition to reach a pH <2. After quick sonication, samples were incubated on the magnetic rack for 2 min and peptidic supernatant was collected.

v. NanoLC-MS/MS analysis of the samples before enrichment

For each replicate, 2 µL of sample were collected, vacuum dried and resuspended in 0.1% FA to be injected on NanoAcquity coupled to Q-Exactive Plus.

a. Chromatographic conditions

Samples (equivalent to 300 ng of proteins) were loaded on a Symmetry C18 precolumn (20 mm × 180 µm with 5 µm diameter particles, Waters) over 3 min at 5 µL/min with 99% of solvent A (H_2O, 0.1% FA) and 1% of solvent B (ACN, 0.1% FA). Peptides were separated on an ACQUITY UPLC BEH130 C18 column (250 mm × 75 µm with 1.7 µm diameter particles) at 400 nL/min with the gradient of solvent B detailed in **Table 13**. Two blank injections were realized between each sample and samples were injected in randomized order.

Time (min)	% in B solvant
0	1
2	8
79	35
80	90
85	90
87	1
105	1

Table 13: Chromatographic gradient used on the NanoAcquity for analysis before phosphopeptide enrichment.

b. *MS and MS/MS parameters*

The Q-Exactive Plus operated in positive ESI mode with the source temperature at 250°C and a 2.1 kV spray voltage. The system was operated DDA mode with automatic switching between MS and MS/MS modes. MS full scans (300-1800 m/z) were acquired with a 70 000 resolution at 200 m/z, a maximal injection time of 50 ms and an AGC target of 3.10^6. The ten most abundant precursor ions were selected on each MS spectrum for further isolation and higher energy collision dissociation fragmentation, excluding mono-charged and unassigned ions. The dynamic exclusion time was set to 60 s. MS/MS spectra were acquired with a 17 500 resolution at 200 m/z, a maximal injection time of 100 ms and an AGC target of 1.10^5.

vi. *Results before enrichment: identification and quantification*

The peaklist (mgf files) were generated from raw data using ProteoWizard MS Convert (v 3.0.11417). Peaks were assigned using Mascot (v 2.6.2) search engine with trypsin/P specificity against an in-house generated protein sequence database containing all Bos Taurus entries extracted from SwissProt (26th of June 2020, 12 220 entries). The precursor mass tolerance was set at 5 ppm and the fragment ion mass tolerance at 0.07 Da. A maximum of one missed cleavage was allowed. Methionine oxidation and acetylation of proteins' N-termini were set as variable modifications and Cysteine carbamidomethylation as a fixed modification. Generated data were validated using Proline (v 2.2.0). The maximum false discovery rate was set to 1% at PSM and protein levels with the use of a decoy strategy. A minimal peptide length of seven amino acids was required, as well as a protein pretty rank <1 and a protein score >25. A minimal of one specific peptide per protein was also required. Data normalization as well as protein quantification were also performed in Proline (v 2.2.0). The map alignment option was selected, all other parameters were set on default values.

vii. *Conditions C and E: phosphopeptide enrichment*

Peptidic material from conditions C and E was dried upon speed vacuum concentrator and resuspended in 170 µL of 80% ACN, 0.1% FA. Phosphomix I light (Thermo Fisher Scientific) were added to each sample (ratio peptide(µg):mix(fmol) = 1.6). Phosphopeptide enrichment was then performed on 5 µL phase Fe(III)-NTA cartridges on an AssayMAP Bravo platform following an IMAC protocol. Briefly, cartridges were washed and primed with 50% ACN, 0.1% TFA, then equilibrated with 80% ACN, 0.1% TFA. 160 µL of samples were loaded at 2 µL/min on the phase then washed with 80% ACN, 0.1%

TFA before being eluted in 20 µL 1% NH_4OH at 5 µL/min. After the enrichment, FA was added to each sample as well as phosphomix I heavy (Sigma Aldrich) (ratio peptide (µg)/mix(fmol) = 1.6).

viii. LC-MS/MS analysis of enriched samples

All enriched samples were dried upon speed vacuum concentrator and resuspended in 40 µL of 2% ACN, 0.1% FA. They were then analyzed on a nanoElute (Bruker) coupled to a TimsTOF Pro (Bruker).

a. Chromatographic conditions

Samples (5 µL) were loaded on an Acclaim™ PepMap™ 100 C18 precolumn (100 µm x 20 mm with 5 µm diameter particles) with 2% of solvent B (ACN, 0.1% FA). Phosphopeptides were separated on an Aurora C18 column (20 mm x 180 µm with 1.6 µm diameter particles; IonOpticks) at 300 nL/min with the following gradient of solvent B (**Table 14**). The column was then washed with 95% of B for 10 min and equilibrated for 10 min with 2% of B. Samples were injected in a randomized order with two blank injections between each sample.

Time (min)	% in B solvant
0	2
2	2
62	50
65	95
75	95

Table 14: Chromatographic gradient used on the NanoElute for samples after phosphopeptide enrichment.

b. MS and MS/MS parameters

The TimsTOF Pro was operated in DDA-PASEF mode with the CaptiveSpray nano-electrospray source temperature at 180°C and a 1.6 kV spray voltage. Mass spectra for MS and MS/MS were acquired between 100 to 1 700 m/z. Both accumulation and ramp time were set on 166 ms. The ion mobility range was fixed between 0.7 to 1.25 1/K0 and the collision energy was ramped stepwise from 20 eV to 52 eV depending on the ion mobility value. Data acquisition was performed using 10 PASEF scans per cycle with a near 100% duty cycle and total cycle time was of 1.88 s. A polygon filter was applied in the m/z and ion mobility space to exclude low m/z, singly charged ions from PASEF precursor selection. For ion precursor selection, the intensity threshold was set at 1 000 (arbitrary units) and the target intensity at 17 000 (arbitrary units) with a dynamic exclusion time of 0.4 min.

ix. Results after enrichment: identification and quantification

Raw data were processed using MaxQuant software (version 2.3.1). Peaks were assigned with the Andromeda search engine with trypsin/P specificity against an in-house generated protein sequence database containing all Bos Taurus entries extracted from UniProtKB-SwissProt (17th of December 2020, 73 032 entries). The minimal peptide length required was seven amino acids and a maximum of one missed cleavage was allowed. Methionine oxidation, acetylation of proteins' N-termini, and serine, threonine and tyrosine phosphorylation were set as variable modifications while cysteine carbamidomethylation as a fixed modification. For protein quantification, the "match between runs"

option was enabled. The maximum false discovery rate was set to 1% at peptide and protein levels with the use of a decoy strategy. All other parameters were set on default values. "Phospho(STY)Sites" output file was used and processed through Perseus (v 2.0.7) to obtain quantification information at the phosphosite level thanks to the "expand sites table" function.

2. Optimization of a LC-MS/MS method for the analysis of phosphopeptides

i. Evaluation of the best LC-MS/MS platform for DDA analysis of phosphopeptides

Phospho-enriched samples from the Laemmli-like buffer extraction condition were injected on 3 different systems:

- A NanoElute (Bruker) coupled to the TimsTOF Pro (Bruker), see **viii.LC-MS/MS analysis of enriched samples**.
- A nanoAcquity (Waters) coupled to a Q-Exactive HF-X (Thermo Fisher Scientific)
- A Dionex Ultimate 3000 (Thermo Fisher Scientific) Tribrid Eclipse (Thermo Fisher Scientific), to compare HCD to ETD fragmentation.

a. Q-Exactive HF-X

- Chromatographic conditions

Samples (8 µL) were loaded on an ACQUITY UPLC® M-Class Symmetry® C18 Trap Column (20 mm x 180 µm with 5 µm diameter particles; Waters) over 3 min at 5 µL/min with 99% of solvent A (H_20, 0.1% FA) and 1% of solvent B (ACN, 0.1% FA). Phosphopeptides were separated on an ACQUITY UPLC® Peptide BEH C18 Column (250 mm x 75 µm with 1.7 µm diameter particles) at 400 nL/min with the following gradient of solvent B (**Table 15**). Samples were injected in a randomized order with two blank injection between each sample.

Time (min)	% in B solvant
0	1
2	2
79	35
80	90
85	90
87	1
105	1

Table 15: Chromatographic gradient used on the NanoAcquity for samples after phosphopeptide enrichment.

- MS and MS/MS parameters

The Q-Exactive HF-X is operated in positive ESI mode with the source temperature at 250°C and a 2.0 kV spray voltage. The system was operated in DDA mode with automatic switching between MS and MS/MS modes. MS full scans (375-1 500 m/z) were acquired with a 120 000 resolution at 200 m/z, a maximal injection time of 60 ms and an AGC target of 3.10^6. The 20 most abundant precursor's ions were selected on each MS spectrum for further isolation and higher energy collision dissociation

fragmentation, excluding monocharged and unassigned ions. The dynamic exclusion time was set to 40 s. MS/MS spectra (200-2 000 m/z) were acquired with a 15 000 resolution at 200 m/z, a maximal injection time of 60 ms and an AGC target of 1.10^5.

- Identification and quantification of phosphorylation sites

Raw data were processed using MaxQuant software (version 2.3.1). Peaks were assigned with the Andromeda search engine with trypsin/P specificity against an in-house generated protein sequence database containing all Bos Taurus entries extracted from UniProtKB-SwissProt (17th of December 2020, 37 518 entries). The minimal peptide length required was seven amino acids and a maximum of one missed cleavage was allowed. Methionine oxidation, acetylation of proteins' N-termini, and serine, threonine and tyrosine phosphorylation were set as variable modifications while cysteine carbamidomethylation as a fixed modification. For protein quantification, the "match between runs" option was enabled. The maximum false discovery rate was set to 1% at peptide and protein levels with the use of a decoy strategy. All other parameters were set on default values. "Phospho(STY)Sites" output file was used and processed through Perseus (v 2.0.7) to obtain quantification information at the phosphosite level thanks to the "expand sites table" function.

b. Eclipse Tribrid

- Chromatographic conditions

6 µL of samples were loaded on Acclaim PepMap™ (100 µm x 20 mm, 5 µm : Thermo Fisher Scientific) at 10 µL/min of 1% B (0.1% FA in ACN) and 99% A (0.1% FA in H_2O) for 3 min. Phosphopeptides were separated on C18 Aurora (250 mm x 75 µm, 1.6 µm, IonOptics) at 300 nL/min and with the gradient detailed in **Table 16**.

Time (min)	% in B solvant
0	2
65	35
66	90
71	90
72	2
90	2

Table 16: Chromatographic gradient used on the Dionex Ultimate 3000 for phosphopeptides analysis.

- MS and MS/MS parameters

HCD: the Eclipse worked in positive ESI mode with the source temperature at 250°C and a 2.0 kV spray voltage. The system was operated in DDA mode with automatic switching between MS and MS/MS modes. MS full scans (375-1500 m/z) were acquired on an Orbitrap with a 120 000 resolution at 200 m/z, a maximal injection time of 50 ms and an AGC target of 4.10^6. The 20 most abundant precursor's ions were selected on each MS spectrum for further isolation and HCD fragmentation, excluding monocharged and unassigned ions. The dynamic exclusion time was set to 40 s. MS/MS spectra (120-1200 m/z) were acquired on Orbitrap with a 15 000 resolution at 200 m/z, a maximal injection time of 22 ms and an AGC target of 5.10^5.

ETD: All parameters were the same as for HCD fragmentation except for the following. ETD fragmentation was performed, excluding monocharged and unassigned ions. MS/MS spectra (120-

1 200 m/z) were acquired on ionic trap using "rapid" scan mode, with a maximal injection time of 35 ms and an AGC target of 1.10^5.

- Identification and quantification of phosphopeptides

Raw data were processed using Proteome Discoverer (version 2.5). Peaks were assigned with Mascot search engine with trypsin/P specificity against an in-house generated protein sequence database containing all Bos Taurus entries extracted from UniProtKB-SwissProt (17th of December 2020, 37 518 entries). The minimal peptide length required was seven amino acids and a maximum of one missed cleavage was allowed. Methionine oxidation, acetylation of proteins' N-termini, and serine, threonine and tyrosine phosphorylation were set as variable modifications while cysteine carbamidomethylation as a fixed modification. Localization probability of the phosphorylation sites was evaluated by phosphoRS algorithm. A 1% FDR was applied at the PSM, peptide and protein levels with Percolator. Only sites with a localization probability greater than 25% were kept.

ii. Optimization of a DDA method on a TimsTOF Pro platform

Optimizations of phosphopeptides LC-MS/MS method on the TimsTOF Pro were performed on murine brain tissues.

a. Sample preparation

Mouse brain protein extract from Laemmli-like extraction were used. Protease inhibitors (1:50 v:v) and 200 mM sodium orthovanadate (for a final concentration of 1 mM) were added to every sample. Protein concentration was assessed using RC-DC assay (Bio-Rad, Hercules, USA). Samples were diluted in Laemmli buffer (TrisHCl, 3% SDS) to reach a 50 µL final volume and 225 µg of proteins were reduced for 30 min at 37°C using 20 mM DTT, 0.1 M ABC to reach a final DTT concentration of 12 mM. Protein alkylation was performed for 30 min at room temperature in the dark using 700 mM IAM, 0.1 M ABC to reach a final IAM concentration of 40 mM. SP3 digestion was then performed, under the same conditions as those described in **1.Development of a high throughput and automated phosphoproteomics sample preparation workflow - iv.Condition E: SP3 digestion**. After peptides recovery, 225 µL of sample were dried upon speed vacuum concentrator and resuspended in 170 µL of 80% ACN and 0.1% TFA. Phosphomix I light (Thermo Fisher Scientific) was added to each sample (ratio peptide(µg):mix(fmol) = 1.6). Phosphopeptide enrichment was then performed on 5 µL phase Fe(III)-NTA cartridges on an AssayMAP Bravo platform following an IMAC protocol, previously described. After the enrichment, FA was added to each sample as well as Phosphomix I heavy (Sigma Aldrich) (ratio peptide (µg)/mix(fmol) = 1.6). Enriched samples were dried upon speed vacuum concentrator and resuspended in 20 µL of H_2O, 2% ACN, 0.1% FA and all samples were pooled together.

b. Chromatographic conditions

Samples (4 µL) were loaded on an Acclaim™ PepMap™ 100 C18 precolumn (100 µm x 20 mm with 5 µm diameter particles) with 2% of solvent B (ACN, 0.1% FA). Phosphopeptides were separated on an Aurora C18 column (20 mm x 180 µm with 1.6 µm diameter particles; IonOpticks) at 300 nL/min. The different gradient tested are described in **Table 17**. Samples were injected in randomized order and one blank was injected between every sample

30 min gradient

Time (min)	% in B solvant
0	2
1	2
19	15
26	23
30	30
33	85
40	85

45 min gradient

Time (min)	% in B solvant
0	2
1	2
19	15
26	23
40	30
43	85
55	85

60 min gradient

Time (min)	% in B solvant
0	2
2	2
62	50
65	95
75	95

80 min gradient

Time (min)	% in B solvant
0	2
2	2
15	10
80	35
82	95
90	95

Table 17: Chromatographic gradients tested on the NanoAcquity for phosphopeptides analysis.

c. *MS and MS/MS parameters*

- For MS/MS optimizations

For all methods, the TimsTOF Pro was used in DDA-PASEF mode with the source temperature at 180°C and a 1.6 kV spray voltage. Mass spectra for MS and MS/MS were acquired between 100 to 1 700 m/z. A polygon filter was applied in the m/z and ion mobility space to exclude low m/z, singly charged ions from PASEF precursor selection. For ion precursors selection, the intensity threshold was set at 1 000 (arbitrary units) and the target intensity at 17 000 (arbitrary units) with a dynamic exclusion time of 0.4 min. Other parameters that were variable from one method to the other are detailed in the following **Table 18**.

	Ion mobility window (1/K0)	Collision energy range (eV)	Slope	Accumulation time (ms)	Number of PASEF scans	Cycle time (s)
Method A	0.7 → 1.25	22 → 62	Stepwise	166	10	1.89
					8	1.55
					6	1.2
Method B	0.7 → 1.25	22 → 62	Linear	166	10	1.89
					8	1.55
					6	1.2
Method C	0.6 → 1.6	20 → 80	Stepwise	100	10	1.17
Method D	0.6 → 1.6	20 → 60	Stepwise	100	10	1.17
Method E	0.6 → 1.6	20 → 80	Stepwise	166	10	1.89
Method F	0.7 → 1.4	22 → 62	Linear	166	10	1.89
					8	1.55
					6	1.2

Table 18: Different parameters tested for each method.

The different collision energy range and their corresponding slope are represented in **Figure 135**.

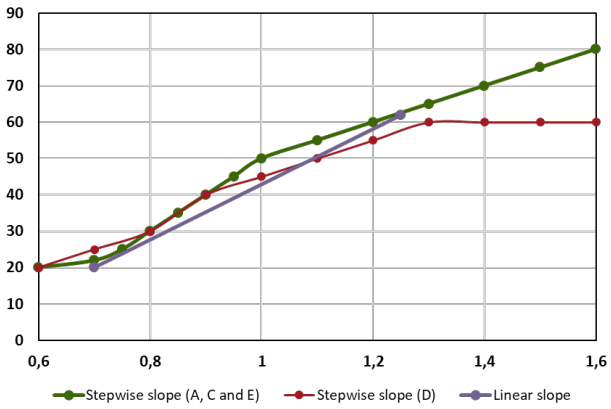


Figure 135: Description of the stepwise and linear slopes of collision energy range used for the different methods.

- For gradient optimizations

For all gradient tested, the TimsTOF operated with the same parameters than described for the MS/MS optimizations, and method B with 8 PASEF scans was used.

d. Identification and quantification of phosphopeptides

Raw data were processed using MaxQuant software (version 2.0.3). Peaks were assigned with the Andromeda search engine with trypsin/P specificity against an in-house generated protein sequence database containing all mouse entries extracted from SwissProt (19th of January 2020, 36 725 entries). The minimal peptide length required was seven amino acids and a maximum of one missed cleavage was allowed. Methionine oxidation, acetylation of proteins' N-termini, and serine, threonine and tyrosine phosphorylation were set as variable modifications while cysteine carbamidomethylation as a fixed modification. For protein quantification, the "match between runs" option was not activated. The maximum false discovery rate was set to 1% at peptide and protein levels with the use of a decoy strategy. All other parameters were set on default values. "Phospho(STY)Sites" output file was used and processed through Perseus (v 2.0.7) to obtain quantification information at the phosphosite level thanks to the "expand sites table" function.

iii. Development of a dia-PASEF method

Optimization of a dia-PASEF method for phosphoproteomics on the TimsTOF Pro were performed on murine brain tissues (see **Chapter 1 – 2.Optimization of a LC-MS/MS method for the analysis of phosphopeptides – ii.Optimization of a DDA method on a TimsTOF Pro platform- a.Sample preparation**).

a. Chromatographic conditions

Samples (4 µL) were loaded on an Acclaim™ PepMap™ 100 C18 precolumn (100 µm x 20 mm with 5 µm diameter particles) with 2% of solvent B (ACN, 0.1% FA). Phospho-peptides were separated on an Aurora C18 column (20 mm x 180 µm with 1.6 µm diameter particles; IonOpticks) at 300 nL/min with the gradient of B detailed in **Table 19 .** Samples were injected in randomized order and one blank was injected between every sample.

Time (min)	% in B solvant
0	2
1	2
19	15
26	23
40	30
43	85
55	85

Table 19: Chromatographic gradient used on the NanoAcquity for dia-PASEF phosphopeptides analysis.

b. *MS and MS/MS parameters*

The TimsTOF Pro was used in dia-PASEF mode with the source temperature at 180°C and a 1.6 kV spray voltage. Mass spectra for MS and MS/MS were acquired between 400 to 1400 m/z. Collision energy ranged from 20 eV to 80 eV. Other parameters that were variable from one method to the other are detailed in **Table 20**.

1st experiment

	Isolation window width (Da)	Accumulation time (ms)	Number of mobility steps	Cycle time (s)	Ion mobility range (1/K0)
Method 1	25	100	1	1.59	0.7 → 1.4
Method 2	30	100	1	1.06	0.7 → 1.4
Method 3	25	166	1	1.75	0.7 → 1.4
Method 4	30	166	1	1.55	0.7 → 1.4

2nd experiment

	Isolation window width (Da)	Accumulation time (ms)	Number of mobility steps	Cycle time (s)	Ion mobility range (1/K0)
Method 1	25	100	1	1.59	0.7 → 1.4
Method 2	30	100	1	1.06	0.7 → 1.4
Method 3	30	100	2	1.17	0.7 → 1.4
Method 4	30	100	3	1.38	0.7 → 1.4
Method 5	30	100	1	1.17	0.8 → 1.35
Method 6	30	100	2	1.17	0.8 → 1.35

Table 20: Different parameters tested for each dia-PASEF methods.

c. *Identification and quantification*

Generated DIA phospho-enriched data were analyzed in Spectronaut software (v.17.1 ; Biognosys) using directDIA™.

- Test 1: an in-house generated protein sequence database containing all mouse entries extracted from SwissProt (19th of January 2020, 36 822 entries) was used for the Pulsar search. Trypsin/P was used as digestion enzyme with two missed cleavages allowed. Carbamidomethylation of cysteine residues was set as a fixed modification. Oxidation of methionine residues, acetylation of proteins n-termini and phosphorylation of serine, threonine and tyrosine residues were set as variable modifications. Peptide length was set up between 7 to 52 amino acids. A maximum of 5 variables modifications per peptide were allowed. For quantitative data extraction, MS and MS/MS mass tolerances, Extracted Ion Chromatogram (XIC) and retention time windows were all set as dynamic. DirectDIA+ (deep) workflow was used. A false discovery rate of 1% was set at precursor and protein levels. A localization probability cutoff of 0.75 was set for PTMs. “PTM report” file was exported to extract phosphosites intensities.
- Test 2: an in-house generated protein sequence database containing all mouse entries extracted from SwissProt (15th of May 2023, 17 268 entries) was used for the Pulsar search. Trypsin/P was used as digestion enzyme with one missed cleavage allowed. Carbamidomethylation of cysteine residues was set as a fixed modification. Oxidation of methionine residues and phosphorylation of serine, threonine and tyrosine residues were set as variable modifications. Peptide length was set up between 7 to 30 amino acids. A maximum of 2 variables modifications per peptide and one missed cleavage were allowed. For quantitative data extraction, MS and MS/MS mass tolerances, XIC and retention time windows were all set as dynamic. DirectDIA+ (deep) workflow was used. A false discovery rate of 1%

was set at precursors and proteins levels. A localization probability cutoff of 0.75 was set for PTMs. “PTM report” file was exported to extract phosphosites intensities.

3. Evaluation of different data treatment pipelines for phosphosites identification, quantification and localization

i. *Benchmarking of different pipelines for DDA phosphoproteomics data analysis*

Phospho-enriched data generated by the Q-Exactive HF-X were analyzed by different pipelines for identification, quantification and localization of phosphorylation.

a. *Mascot (Proline)*

The peaklist (mgf files) were generated from raw data using ProteoWizard MS Convert (v 3.0.11417). Peaks were assigned using Mascot (v 2.6.2) search engine with trypsin/P specificity against the Bos Taurus database already described. The precursor mass tolerance was set at 5 ppm and the fragment ion mass tolerance at 0.05 Da. A maximum of one missed cleavage was allowed. Methionine oxidation, acetylation of proteins’ N-termini, serine, threonine and tyrosine phosphorylations were set as variable modifications. Cysteine carbamidomethylation was set as a fixed modification. Generated data were validated using Proline (v 1.0). The maximum false discovery rate was set to 1% at PSM and protein levels with the use of a decoy strategy. A minimal peptide length of seven amino acids was required, as well as a protein pretty rank <1 and a protein score >25. A minimal of one specific peptide per protein was also required.

b. *Mascot (Proteome Discoverer)*

Raw data were processed using Proteome Discoverer (version 2.5). Peaks were assigned with Mascot search engine. Proteic database and search parameters were the same as described in **3- i.Benchmarking of different pipelines for DDA phosphoproteomics data analysis – a.Mascot (Proline)**. Localization probability of the phosphorylation sites was evaluated by phosphoRS algorithm. A 1% FDR was applied at the PSM, peptide and protein levels with Percolator. Only sites with a localization probability greater than 25% were kept.

c. *Mascot + MS Amanda (Proteome Discoverer)*

Raw data were processed using Proteome Discoverer (version 2.5). Proteome Discoverer allows to perform two parallels searches at the same time. Here, search was performed by both Mascot and MS Manda (v.2.0) algorithms. For both searches, proteic database and search parameters were as described in **3- i.Benchmarking of different pipelines for DDA phosphoproteomics data analysis – a.Mascot (Proline)**. Localization probability of the phosphorylation sites was evaluated by phosphoRS algorithm. A 1% FDR was applied at the PSM, peptide and protein levels with Percolator. Only sites with a localization probability greater than 25% were kept.

d. *Andromeda (MaxQuant)*

See **1.Development of a high throughput and automated phosphoproteomics sample preparation workflow - b - ix.Results after enrichment: identification and quantification.**

ii. Spectronaut and DIA-NN software for dia-PASEF data treatment

a. Spectronaut

See **2.Optimization of a LC-MS/MS method for the analysis of phosphopeptides iii.Development of a dia-PASEF method c.Identification and quantification – Test 2.**

b. DIA-NN

DIA-NN (version 1.8.1) was used in a library free approach. The proteic database used for the DIA-NN search and search parameters were the same as described for Spectronaut Pulsar search. A 1% FDR at the precursor level was set. The “deep learning-based spectra, RTs and IMs prediction” as well as “heuristic inferences” were activated. Quantification strategy was set as robust LC (high precision). Low RAM and high speed mode was activated.

Chapter 2: Multi-omic analysis of axono-synaptic degeneration in motoneuron disease – application to the MAXOMOD project

1. Proteomics and phosphoproteomics analysis of large cohorts of brain tissues

i. Global proteomics analysis of human and mouse brain tissues

a. Sample preparation

Brain tissues coming both from human post-mortem frontal cortex and from pre-frontal cortex (PFC) of the 4 different mouse models (SOD1, TDP43, C9, FUS) were prepared as follows. Tissues were grinded with a biomasher using 350 µL of MeOH:H_2O (4:1). After incubation on ice for 20 min, they were centrifuged at 14 000 g at 4°C for 15 min, to extract metabolites for metabolomics analysis (conducted at the FGC Zurich). Protein pellets were resuspended in 200 µL Laemmli buffer (10% SDS, Tris 1M pH 6.8, glycerol) then centrifuged at 5 000 g at 4°C for 5 min. Protein concentration was determined using DC assay (BioRad, Hercules, CA, USA) according to the manufacturer's instructions. 100 µg of each sample were taken and diluted up to 50 µL with Laemmli. DTT was added to each sample to reach a final concentration of DTT of 50 mM. A Quality Control sample (QC) was prepared by mixing 5 µL of each protein extract. Sample mix followed exactly the same sample preparation steps than any other biological sample. Samples were then heated at 95°C for 5 min and stacked in an in-house prepared 5% acrylamide SDS-PAGE stacking gel. Gel bands were reduced and alkylated prior to overnight digestion (enzyme:protein ratio of 1:80) at 37°C using modified porcine trypsin (Mass Spec Grade, Promega, Madison, USA). The generated peptides were extracted with 60% ACN followed by a second extraction with 100% ACN. Vacuum dried peptidic samples were resuspended in 30 µL of H_20, 2% ACN, 0.1% FA and iRT peptides (Biognosys, Zurich, Switzerland) were added to each sample according to the manufacturer's instructions as an internal QC.

b. Chromatographic conditions

NanoLC-MS/MS analyses were performed on a nanoAcquity UltraPerformance LC® (UPLC®) device (Waters Corporation, Milford, MA) coupled to a Q-Exactive Plus mass spectrometer (Thermo Fisher Scientific, Waltham, MA). The solvent system consisted of 0.1% FA in water (solvent A) and 0.1% FA in ACN (solvent B). Tryptic digests (equivalent to 800 ng of proteins) were loaded on a Symmetry C18 precolumn (20 mm × 180 µm with 5 µm diameter particles, Waters) over 3 min at 5 µL/min with 99% of solvent A and 1% of solvent B. Peptides were separated on an ACQUITY UPLC BEH130 C18 column (250 mm × 75 µm with 1.7 µm diameter particles) at 400 nL/min with the gradient of solvent B detailed in **Table 21**. The samples of each cohort were injected in randomized order. The QC samples (mix of samples digest was injected regularly throughout the cohort, every 6 samples for the human cohort (total of 21 QCs) and every 5 samples for mouse cohorts (total of 5 QCs).

Time (min)	% in B solvant
0	1
2	8
79	35
80	90
85	90
87	1
105	1

Table 21: Chromatographic gradient used on the NanoAcquity Q-Exactive Plus for proteomics analysis of human and mouse brain tissues.

c. *MS and MS/MS parameters*

Q-Exactive Plus was operated in DDA mode with automatic switching between MS (mass range 300 – 1800 m/z with R = 70 000, AGC fixed at 3x 10^6 ions and a maximum injection time set at 50 ms) and MS/MS (mass range 200–2000 m/z with R = 17 500, AGC fixed at 1 x 10^5 and the maximal injection time set to 100 ms) modes. The ten most abundant precursor's ions were selected on each MS spectrum for further isolation and higher energy collision dissociation fragmentation, excluding monocharged and unassigned and ions. The dynamic exclusion time was set to 60 s.

d. *Identification and quantification*

Raw data were processed using MaxQuant software (version 1.6.14). Peaks were assigned with the Andromeda search engine with trypsin/P specificity against an in-house generated protein sequence database containing either all human entries extracted from UniProtKB-SwissProt (24th of August 2020, 20 421 entries) or all mouse entries extracted from UniProtKB-SwissProt (27th of March 2020, 17 134 entries for SOD1 & TDP43 models and 29th of September 2020, 17 061 entries for C9 and FUS models). The minimal peptide length required was seven amino acids and a maximum of one missed cleavage was allowed. Methionine oxidation and acetylation of proteins' N-termini were set as variable modifications and Cysteine carbamidomethylation as a fixed modification. For protein quantification, the "match between runs" option was enabled. The maximum false discovery rate was set to 1% at peptide and protein levels with the use of a decoy strategy. Intensities were extracted from the Proteingroup.txt file to perform the following statistical analysis.

e. *Statistical and differential analysis*

Protein intensities were uploaded in Prostar software (v.1.30.7 and DAPAR v.1.30.6) and data was split into 4 conditions for each cohort of samples: TG_Male, TG_Female, WT_Male and WT_Female. After log transformation of the intensities, contaminants and reverse proteins were removed as well as proteins identified with no unique peptide. Proteins with at least 80% of non-missing values, in at least one condition, were kept. Normalization of the data was performed using quantile centering normalization with a 15% quantile. Missing values were imputed using *det quantile* algorithm and a 2.5% quantile. Finally, a Limma statistical test was applied as well as Benjamini-Hochberg p-value calibration. Differential analysis of the TG versus WT conditions was performed, with a 0.1% p-value filter.

ii. Phosphoproteomic analysis of mouse brain samples

a. Sample preparation

For SOD1, TDP43 and FUS mouse models, protein extract from global proteomics analysis were used for sample preparation. For the C9 mouse model, proteins were extracted from brain tissues following the steps detailed in global proteomics sample preparation. For all mouse models, protease inhibitors (1:50 v:v ; Sigma, P8340) and 20 mM sodium orthovanadate (final concentration in Na_3VO_4 = 1 mM) were added to protein extract. Protein concentration was determined using RC-DC assay (BioRad, Hercules, CA, USA) according to the manufacturer's instructions. 250 µg of proteins for each sample were reduced and alkylated prior to an in-house SP3 protocol, previously described. Proteins were then resuspended in 95 µL NH_4HCO_3 prior to overnight on-beads digestion (enzyme:protein ratio of 1:20) at 1 000 rpm at 37°C using modified porcine trypsin/Lys-C (Mass Spec Grade mix ,Promega, Madison, USA). Digestion was stopped using TFA (final pH < 2). Recovered peptides were resuspended in 170 µL 80% ACN, 0.1% TFA and Phosphomix I light (Sigma Aldrich) was added to each sample (ratio peptide (µg)/mix(fmol) = 1.6).

Phosphopeptide enrichment was performed on 5 µL phase Fe(III)-NTA cartridges on an AssayMAP Bravo platform following an IMAC protocol, previously described. After the enrichment, FA was added to each sample as well as Phosphomix I heavy (Sigma Aldrich) (ratio peptide (µg)/mix(fmol) = 1.6). Phosphopeptides were vacuum dried and resuspended in 40 µL H_2O, 2% ACN, 0.1% FA.

b. Chromatographic conditions

Samples (8 µL) were analyzed on the nanoLC Q-Exactive HF-X platform with the same chromatographic condiitons as detailed in **Chapter 1 - LC-MS/MS analysis of enriched samples - a.Chromatographic conditions.**

c. MS and MS/MS parameters

Q-Exactive HF-X MS and MS/MS parameters were as described in **Chapter 1 - LC-MS/MS analysis of enriched samples – b.MS and MS/MS parameters**.

d. Identification and quantification

Raw data were processed using MaxQuant software (version 1.6.14). Peaks were assigned with the Andromeda search engine with trypsin/P specificity against an in-house generated protein sequence database containing all mouse entries extracted from UniProtKB-SwissProt (29th of September 2020, 17 061 entries). The minimal peptide length required was seven amino acids and a maximum of one missed cleavage was allowed. Methionine oxidation and acetylation of proteins' N-termini were set as variable modifications and Cysteine carbamidomethylation as a fixed modification. For protein quantification, the "match between runs" option was enabled. The maximum false discovery rate was set to 1% at peptide and protein levels with the use of a decoy strategy. Intensities were extracted from the Phospho(STY)sites.txt file to perform the following statistical analysis.

e. Statistical and differential analysis

Phospho(STY)sites.txt file was loaded into Perseus software (v 2.0.7). Contaminants and reverse protein were removed, and lines with null intensity values in all samples were removed. Then, using

the "expand site tables" option, the intensities of all phosphopeptides carrying the same phosphosite were combined to obtain intensities at the phosphosite level. Only phosphosites with a localization probability >0.75 were kept. Phosphosite intensities were then uploaded in Prostar software (v.1.30.7 and DAPAR v.1.30.6) and data split into 4 conditions for each mouse model: TG_Male, TG_Female, WT_Male and WT_Female. After log transformation of the intensities, phosphosites with at least 70% of non-missing values, in at least one condition, were kept. Normalization of the data was performed using quantile centering normalization with a 15% quantile. Missing values were imputed using *imp4p* algorithm. Finally, a Limma statistical test was applied as well as Benjamini-Hochberg p-value calibration. Differential analysis of the TG versus WT conditions was performed, with a 0.1% p-value filter.

2. Development of a protocol for both proteomics and phosphoproteomics analysis of cerebrospinal fluid

i. Test 1: in-solution digestion versus in-gel digestion

a. Sample preparation

The different protocols tested are represented in **Figure 136**.

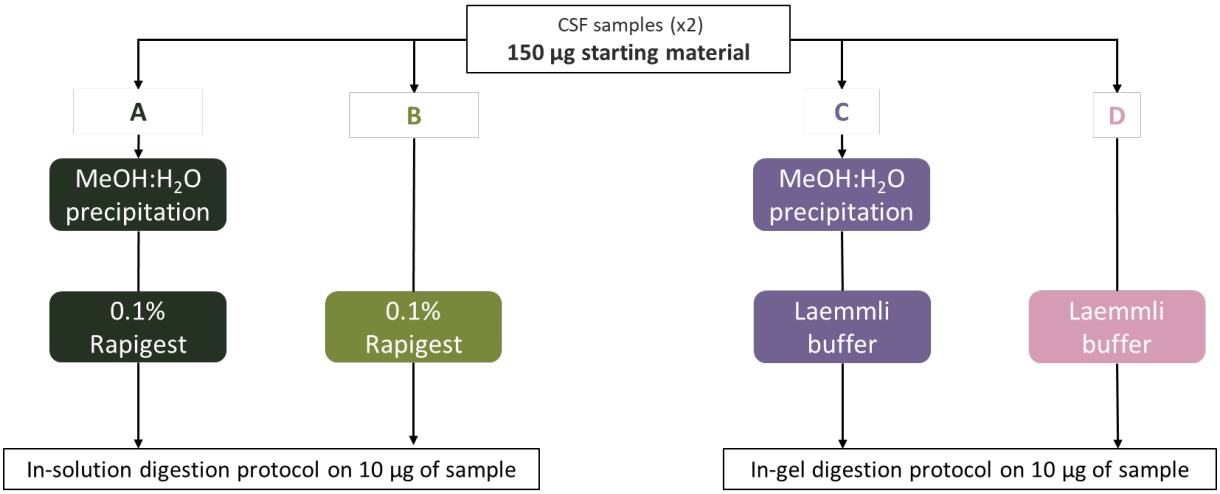


Figure 136: Method development for CSF sample preparation**.**

Two biological replicates of human CSF were used. Protein concentration was determined using DC assay (BioRad, Hercules, CA, USA) according to the manufacturer's instructions. For conditions with a precipitation step, 150 µg of sample were precipitated using MeOH:H_2O (4:1) , vortexed vigorously for 10 s and incubated for 20 min on ice. Samples were then centrifuged for 15 min at 16 000 g at 4°C. Metabolites were recovered and proteins resuspended in either 25 µL of 0.1% RapiGest in 50 mM ABC or in 25 µL of Laemmli buffer. Samples were sonicated for 3 min. Protein concentration after precipitation was determined using DC assay (BioRad, Hercules, CA, USA) according to the manufacturer's instructions.

- For in-solution digestion: DTT was added to 10 µg of each sample (final concentration in DTT = 10 mM). Samples were incubated for 30 min at 60°C. IAM was added to each sample to reach a concentration in IAM of 55 mM and samples were incubated at room temperature in the dark for 30 min. Overnight digestion (enzyme:protein ratio 1:25) was performed at 37°C using a mixture of Trypsin/Lys-C (Mass Spec Grade mix, Promega, Madison, USA). Digestion was

stopped using TFA (final concentration of TFA = 0.5%, pH < 2). Samples were then incubated at 37°C for 45 min followed by a centrifugation at 13 000 rpm for 10 min in order to get rid of RapiGest.

- For in-gel digestion: DTT was added to 10 µg of each sample to reach a final concentration of 50 mM. Samples were then heated at 100°C for 5 min and stacked in an in-house prepared 4% acrylamide SDS-PAGE stacking gel. Gel bands were reduced and alkylated prior to overnight digestion (enzyme:protein ratio of 1:40) at 37°C using modified porcine trypsin (Mass Spec Grade, Promega, Madison, USA). The generated peptides were extracted with 60% CAN followed by a second extraction with 100% ACN.

All samples were vacuum dried and resuspended in H_20, 0.1% FA.

b. *Chromatographic conditions*

Samples (2 µL) were loaded on a Symmetry C18 precolumn (20 mm × 180 µm with 5 µm diameter particles, Waters) over 3 min at 5 µL/min with 99% of solvent A (H_2O, 0.1% FA) and 1% of solvent B (ACN, 0.1% FA). Peptides were separated on an ACQUITY UPLC BEH130 C18 column (250 mm × 75 µm with 1.7 µm diameter particles) at 450 nL/min with the gradient of solvent B detailed in **Table 22.** One blank injection was realized between each sample.

Time (min)	% in B solvant
0	1
2	8
120	35
121	90
126	90
128	1
130	1

Table 22: Chromatographic gradient used on the NanoAcquity Q-Exactive Plus.

c. *MS and MS/MS parameters*

The Q-Exactive Plus operated in positive ESI mode with the source temperature at 250°C and a 2.1 kV spray voltage. The system was operated in DDA mode with automatic switching between MS and MS/MS modes. MS full scans (300-1800 m/z) were acquired with a 70 000 resolution at 200 m/z, a maximal injection time of 50 ms and an AGC target of 3.10^6. The ten most abundant precursor ions were selected on each MS spectrum for further isolation and HCD fragmentation, excluding monocharged and unassigned ions. The dynamic exclusion time was set to 60 s. MS/MS spectra were acquired with a 17 500 resolution at 200 m/z, a maximal injection time of 100 ms and an AGC target of 1.10^5.

d. *Identification and quantification*

The peaklists (mgf files) were generated from raw data using ProteoWizard MS Convert (v 3.0.11417). Peaks were assigned using Mascot (v 2.6.2) search engine with trypsin/P specificity against an in-house generated protein sequence database containing all human entries extracted from SwissProt (25^{th} of August 2021, 20 339 entries). The precursor mass tolerance was set at 5 ppm and the fragment ion mass tolerance at 0.05 Da. A maximum of one missed cleavage was allowed. Methionine oxidation and

acetylation of proteins' N-termini were set as variable modifications and Cysteine carbamidomethylation as a fixed modification. Generated data were validated using Proline (v 2.2.0). The maximum false discovery rate was set to 1% at PSM and protein levels with the use of a decoy strategy. A minimal peptide length of seven amino acids was required, as well as a protein pretty rank <1 and a protein score >25. A minimal of one specific peptide per protein was also required.

ii. Test 2: in-solution digestion versus on-membrane digestion

a. Sample preparation

The different protocols tested are represented in **Figure 137**.

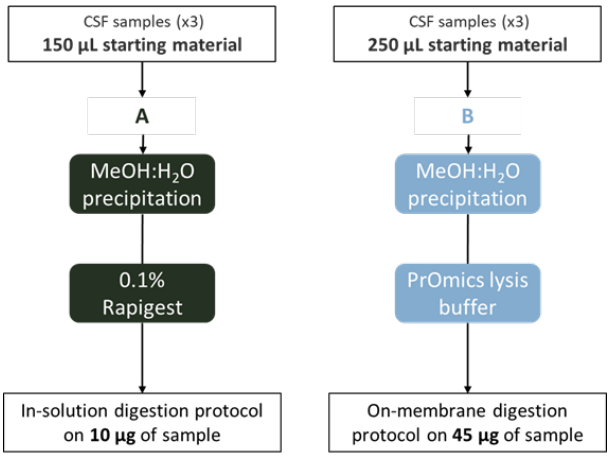


Figure 137: Method development for CSF sample preparation

Three biological replicates of human CSF were used.

- For Protocol A (in-solution digestion): protein concentration was determined using Pierce 660 nm assay (ThermoFisher Scientific) according to the manufacturer's instructions. 150 µL of CSF were precipitated using 100% MeOH, vortexed vigorously for 10 s and incubated for 10 min on ice. Samples were then centrifuged for 15 min at 16 000 g at 4°C. Metabolites were recovered and protein resuspended in 100 µL of 0.1% RapiGest in 50 mM. Samples were sonicated for 3 min. Protein concentration after precipitation was determined using Pierce 660 nm assay (ThermoFisher Scientific) according to the manufacturer's instructions. 10 µg of samples were diluted up to 30 µL with 0.1 M ABC and 20 mM DTT was added to reach a final concentration in DTT of 5 mM. Samples were incubated for 30 min at 60°C. 100 mM IAM was added to each samples to reach a final concentration in IAM of 15 mM and samples were incubated in the dark at room temperature for 30 min. Overnight digestion (enzyme:protein ratio 1:25) was performed at 37°C and 1 000 rpm using a mixture of Trypsin/Lys-C (Mass Spec Grade mix, Promega, Madison, USA). Digestion was stopped using TFA (final concentration of TFA = 0.5%, pH < 2). Samples were then incubated at 37°C for 45 min followed by a centrifugation at 13 000 rpm for 10 min in order to get rid of RapiGest. Samples were vacuum dried and resuspended in 100 µL of H_2O, 0.1% FA and injected on Q-Exactive HF-X platform.

- For Protocol B (on-membrane digestion): protein concentration was determined using BCA assay (ThermoFisher Scientific) according to the manufacturer's instructions. 250 µL of CSF were precipitated using 100% MeOH, vortexed vigorously for 10 s and incubated for 10 min on ice. Samples were then centrifuged for 15 min at 16 000 g at 4°C. Metabolites were

recovered and protein resuspended in 70 µL of PreOmics iST kit (GmbH, Germany) lysis buffer. Protein concentration after precipitation was determined using BCA assay (ThermoFisher Scientific) according to the manufacturer's instructions. 10 µg of samples were used to follow the PreOmics iST kit protocol. Samples were vacuum dried and resuspended in H_2O, 0.1% FA and injected on Q-Exactive Plus platform

b. *Chromatographic conditions*

- For in-solution digestion (Protocol A): tryptic digests (300 ng) were loaded on an ACQUITY UPLC® M-Class Symmetry® C18 Trap Column (20 mm x 180 µm with 5 µm diameter particles; Waters) over 3 min at 5 µL/min with 99% of solvent A (H_20, 0.1% FA) and 1% of solvent B (ACN, 0.1% FA). Peptides were separated on an ACQUITY UPLC® Peptide BEH C18 Column (250 mm x 75 µm with 1.7 µm diameter particles) at 400 nL/min with the following gradient of solvent B (**Table 23**). Samples were injected in a randomized order with two blank injection between each sample.

Time (min)	% in B solvant
0	1
2	2
79	25
89	35
90	90
95	90
100	2
105	1

Table 23: Chromatographic gradient used on the NanoAcquity Q-Exactive HF-X.

- For on-membrane digestion (Protocol B): samples (2 µL) were loaded on a Symmetry C18 precolumn (20 mm × 180 µm with 5 µm diameter particles, Waters) over 3 min at 5 µL/min with 99% of solvent A (H_2O, 0.1% FA) and 1% of solvent B (ACN, 0.1% FA). Peptides were separated on an ACQUITY UPLC BEH130 C18 column (250 mm × 75 µm with 1.7 µm diameter particles) at 450 nL/min with the gradient of solvent B previously detailed in **Table 22.** One blank injection was realized between each sample.

c. *MS and MS/MS parameters*

- For in-solution digestion (Protocol A): the Q-Exactive HF-X is operated in positive ESI mode with the source temperature at 250°C and a 2.0 kV spray voltage. The system was operated in DDA mode with automatic switching between MS and MS/MS modes. MS full scans (375 - 1 500 m/z) were acquired with a 120 000 resolution at 200 m/z, a maximal injection time of 60 ms and an AGC target of 3.10^6. The 20 most abundant precursor ions were selected on each MS spectrum for further isolation and higher energy collision dissociation fragmentation, excluding monocharged and unassigned ions. The dynamic exclusion time was set to 40 s. MS/MS spectra (200-2000 m/z) were acquired with a 15 000 resolution at 200 m/z, a maximal injection time of 60 ms and an AGC target of 1.10^5.

- For on-membrane digestion (Protocol B): see parameters described in **2.Development of a protocol for both proteomics and phosphoproteomics analysis of cerebrospinal fluid – i.Test 1: in-solution digestion versus in-gel digestion c.MS and MS/MS parameters**.

d. Identification and quantification

See **2.Development of a protocol for both proteomics and phosphoproteomics analysis of cerebrospinal fluid - i.Test 1: in-solution digestion versus in-gel digestion – d.Identification and quantification**.

iii. Evaluation of sample preparation for CSF phosphoproteomics analysis

a. Sample preparation

500 µL (equivalent 150 µg of proteic material) of CSF samples for phosphoproteomics analysis were prepared with the in-solution digestion protocol using 0.1% RapiGest described in **2. Development of a protocol for both proteomics and phosphoproteomics analysis of cerebrospinal fluid – ii.Test 2: in-solution digestion versus on-membrane digestion – a.Sample preparation**. A peptide clean-up step was performed using 5 µL C18 phase cartridges on AssayMAP Bravo (Agilent) platform. Briefly, cartridges were washed and primed with 50% ACN, 0.1% TFA, then equilibrated with H_2O, 0.1% TFA. 360 µL of samples were loaded at 5 µL/min on the phase then washed with 50% ACN, 0.1% TFA before being eluted in 20 µL 70% ACN, 0.1% TFA at 5 µL/min. Samples were then diluted up to 90 µL with 80% ACN, 0.1% TFA. Phosphomix I light (Thermo Fisher Scientific) was added to each sample (ratio peptide(µg):mix(fmol) = 1.6). 100 µL of samples were then loaded on AssayMAP Bravo platform to perform IMAC phosphopeptide enrichment, previously described. After the enrichment, FA was added to each sample as well as Phosphomix I heavy (Sigma Aldrich) (ratio peptide (µg)/mix(fmol) = 1.6). Enriched samples were dried upon speed vacuum concentrator and resuspended in 25 µL of H_2O, 2% ACN, 0.1% FA.

b. Chromatographic conditions

Samples (8 µL) were loaded on an ACQUITY UPLC® M-Class Symmetry® C18 Trap Column (20 mm x 180 µm with 5 µm diameter particles; Waters) over 3 min at 5 µL/min with 99% of solvent A (H_20, 0.1% FA) and 1% of solvent B (ACN, 0.1% FA). Peptides were separated on an ACQUITY UPLC® Peptide BEH C18 Column (250 mm x 75 µm with 1.7 µm diameter particles) at 400 nL/min with the following gradient of solvent B (**Table 15**). Samples were injected in a randomized order with two blank injection between each samples.

c. MS/MS and data analysis

Q-Exactive HF-X was operated in the same exact parameters as described for in-solution digestion. Identification and quantification parameters were the same as for in-solution digestion experiment. See **2.Development of a protocol for both proteomics and phosphoproteomics analysis of cerebrospinal fluid - i.Test 1: in-solution digestion versus in-gel digestion**.

iv. Global proteomics analysis of MAXOMOD CSF samples

a. Sample preparation

We analyzed 103 samples of human CSF coming from both ALS patients and control samples. Additionally, 20 CSF samples were submitted to the same protocol and used as external quality control. 150 µL samples were precipitated using 100% MeOH, vortexed vigorously for 10 s and incubated for 10 min on ice. Samples were then centrifuged for 15 min at 16 000 g at 4°C. Metabolites were recovered and protein resuspended in 100 µL of 0.1% RapiGest in 50 mM. Samples were sonicated for 3 min. Protein concentration after precipitation was determined using Pierce 660 nm assay (ThermoFisher Scientific) according to the manufacturer's instructions. 20 µg of samples were diluted up to 60 µL with 0.1 M ABC. Reduction and alkylation were performed on AssayMAP Bravo (Agilent) platform using In-solution digestion (v1.2) protocol. Briefly, 10 µL of 35 mM DTT was added to reach a final concentration in DTT of 5 mM. Samples were incubated for 30 min at 60°C. 10 µL of 120 mM IAM was added to each sample to reach a final concentration in IAM of 15 mM and samples were incubated at room temperature for 30 min. Overnight digestion (enzyme:protein ratio 1:25) was performed manually at 37°C and 1 000 rpm using a mixture of Trypsin/Lys-C (Mass Spec Grade mix, Promega, Madison, USA). Digestion was stopped using TFA (final concentration of TFA = 0.5%, pH < 2). Samples were then incubated at 37°C for 45 min followed by a centrifugation at 13 000 rpm for 10 min in order to get rid of RapiGest. 110 µL of samples were then loaded on AssayMAP Bravo to perform peptide clean-up protocol on C18 cartridges (described previously). Eluted samples were vacuum dried and resuspended in 150 µL (100 µL for pool samples) of H_2O, 0.1% FA. The 20 additional CSF samples were pooled and divided in 50 µL aliquots to be injected as external QCs.

b. Chromatographic conditions

Samples (300 ng equivalent) were loaded on a Symmetry C18 precolumn (20 mm × 180 µm with 5 µm diameter particles, Waters) over 3 min at 5 µL/min with 99% of solvent A (H_2O, 0.1% FA) and 1% of solvent B (ACN, 0.1% FA). Peptides were separated on an ACQUITY UPLC BEH130 C18 column (250 mm × 75 µm with 1.7 µm diameter particles) at 400 nL/min with the gradient of solvent B detailed in **Table 24**. Samples were injected in a randomized order with two blank injection between each sample.

Time (min)	% in B solvant
0	1
2	2
79	25
89	35
90	90
95	90
95.1	1
110	1

Table 24: Gradient used on the NanoAcquity for CSF sample analysis.

c. MS and MS/MS parameters

The Q-Exactive Plus operated in positive ESI mode with the source temperature at 250°C and a 2.1 kV spray voltage. The system was operated in DDA a 70 000 resolution at 200 m/z, a maximal injection

time of 50 ms and an AGC target of 3.10^6. The ten most abundant precursor ions were selected on each MS spectrum for further isolation and HCD fragmentation, excluding monocharged and unassigned ions. The dynamic exclusion time was set to 60 s. MS/MS spectra were acquired with a 17 500 resolution at 200 m/z, a maximal injection time of 100 ms and an AGC target of 1.10^5.

d. Identification and quantification

Raw data were processed using MaxQuant software (version 2.3.1.0). Peaks were assigned with the Andromeda search engine with trypsin/P specificity against an in-house generated protein sequence database containing all human entries extracted from UniProtKB-SwissProt (11th of January 2023, 20 428 entries). The minimal peptide length required was seven amino acids and a maximum of one missed cleavage was allowed. Methionine oxidation and acetylation of proteins' N-termini were set as variable modifications and Cysteine carbamidomethylation as a fixed modification. For protein quantification, the "match between runs" option was enabled. The maximum false discovery rate was set to 1% at peptide and protein levels with the use of a decoy strategy. Intensities were extracted from the ProteinGroup.txt file to perform the following statistical analysis.

e. Statistical and differential analysis

Protein intensity were uploaded in Prostar software (v.1.22.4) and data split into 4 conditions: ALS_Male, ALS_Female, CTRL_Male and CTRL_Female. After log transformation of the intensities, contaminants and reverse proteins were removed as well as proteins identified with only one unique peptide. Proteins with at least 80% of non-missing values, in at least one condition, were kept. Normalization of the data was performed using quantile centering normalization with a 15% quantile. Missing values were imputed using *det quantile* algorithm and a 2.5% quantile. Finally, a Limma statistical test was applied as well as Benjamini-Hochberg p-value calibration. Differential analysis of the ALS versus CTRL conditions was performed, with a 1% p-value filter.

v. Phosphoproteomics analysis of MAXOMOD CSF samples

a. Sample preparation

Samples used for CSF phosphoproteomics analysis were the same as for proteomics. Additionally, 20 CSF samples were also processed as every other samples to be later pooled together and used as external quality control. 500 µL samples were precipitated using 100% MeOH, vortexed vigorously for 10 s and incubated for 10 min on ice. Samples were then centrifuged for 15 min at 16 000 g at 4°C. Metabolites were recovered and protein resuspended in 300 µL of 0.1% RapiGest in 50 mM. Samples were sonicated for 3 min. Protein concentration after precipitation was determined using Pierce 660 nm assay (ThermoFisher Scientific) according to the manufacturer's instructions. 100 µg of samples were diluted up to 60 µL with 0.1 M ABC. Reduction and alkylation were performed on AssayMAP Bravo (Agilent) platform using In-solution digestion (v1.2) protocol, previously described. Overnight digestion (enzyme:protein ratio 1:25) was performed manually at 37°C and 1 000 rpm using a mixture of Trypsin/Lys-C (Mass Spec Grade mix, Promega, Madison, USA). Digestion was stopped using TFA (final concentration of TFA = 0.5%, pH < 2). Samples were then incubated at 37°C for 45 min followed by a centrifugation at 13 000 rpm for 10 min in order to get rid of RapiGest. Samples were then loaded on AssayMAP Bravo to perform peptide clean-up protocol on C18 cartridges (described previously). Eluted samples were diluted in 80% ACN, 0.1% TFA and Phosphomix I light (Sigma Aldrich) synthetic phosphopeptides were added to each samples (ratio peptide (µg)/mix(fmol) = 1.6). 100 µL of samples

were loaded on AssayMAP Bravo to perform IMAC phosphopeptide enrichment (described previously). Phospho-enriched peptidic material were eluted in 20 µL to which Phosphomix I heavy (Sigma Aldrich) synthetic phosphopeptides were added (ratio peptide (µg)/mix(fmol) = 1.6). Sampled were then vacuum dried and resuspended in 20 µL of H_2O, 0.1% FA.

b. Chromatographic and MS/MS conditions

Phosphopeptides were injected on a nanoLC Q-Exactive HF-X platform with the exact same chromatographic, Ms and MS/MS parameters as described in **iii.Evaluation of sample preparation for CSF phosphoproteomics analysis**.

c. Identification and quantification

Raw data were processed using MaxQuant software (version 2.3.1.0). Peaks were assigned with the Andromeda search engine with trypsin/P specificity against an in-house generated protein sequence database containing all human entries extracted from UniProtKB-SwissProt (11th of January 2023, 20 428 entries). The minimal peptide length required was seven amino acids and a maximum of one missed cleavage was allowed. Methionine oxidation, acetylation of proteins' N-termini and serine, threonine and tyrosine phosphorylations were set as variable modifications while Cysteine carbamidomethylation as a fixed modification. For protein quantification, the "match between runs" option was enabled. The maximum false discovery rate was set to 1% at peptide and protein levels with the use of a decoy strategy. Intensities were extracted from the Phospho(STY)sites.txt file to perform the following statistical analysis.

d. Statistical and differential analysis

Phospho(STY)sites.txt file was loaded into Perseus software (v 2.0.7). Contaminants and reverse protein were removed, and lines with null intensity values in all samples were removed. Then, using the "expand site tables" option, the intensities of all phosphopeptides involved in one phosphosites were combined to obtain intensities at the phosphosite level. Only phosphosites with a localization probability >0.75 were kept. Phosphosite intensities were then uploaded in Prostar software (v.1.22.4 and data split into 4 conditions: ALS_Male, ALS_Female, CTRL_Male and CTRL_Female. After log transformation of the intensities, phosphosites with at least 50% of non-missing values, in at least one condition, were kept. Normalization of the data was performed using quantile centering normalization with a 15% quantile. Missing values were imputed using *slsa* algorithm. Finally, a Limma statistical test was applied as well as Benjamini-Hochberg p-value calibration. Differential analysis of the ALS versus CTRL conditions was performed, with a 1% p-value filter.

3. Open modification searching

Phosphoproteomics mouse data generated in **Chapter 2: Multi-omic analysis of axono-synaptic degeneration in motoneuron disease – application to the MAXOMOD project – ii.Phosphoproteomic analysis of mouse brain samples** were used. MGF generated files were loaded in IonBot (version 0.10.0). In-house generated protein sequence database containing all mouse entries extracted from UniProtKB-SwissProt (11th of January 2023, 17 154 entries) was used, with a K|R cleavage pattern. Error tolerances were set on default values: MS precursors tolerance at 20 ppm and MS/MS fragment tolerance at 0.02 Da. Methionine oxidation and serine, threonine and tyrosine phosphorylations were

set as variable modifications while Cysteine carbamidomethylation as a fixed modification. Open modification search option was enabled.

References

References

1. Kuras, M. *et al.* Assessing Automated Sample Preparation Technologies for High-Throughput Proteomics of Frozen Well Characterized Tissues from Swedish Biobanks. *J. Proteome Res.* **18**, 548–556 (2018).

2. Kitata, R. B., Yang, J. & Chen, Y. Advances in data-independent acquisition mass spectrometry towards comprehensive digital proteome landscape. *Mass Spectrometry Reviews* e21781 (2022).

3. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods* **17**, 41–44 (2020).

4. Needham, E. J., Parker, B. L., Burykin, T., James, D. E. & Humphrey, S. J. Illuminating the dark phosphoproteome. *Sci. Signal.* **12**, eaau8645 (2019).

5. Gendron, T. F. *et al.* Phosphorylated neurofilament heavy chain: a biomarker of survival for C9ORF72-associated amyotrophic lateral sclerosis. *Ann Neurol.* **82**, 139–146 (2017).

6. Butterfield, D. A. Phosphoproteomics of Alzheimer disease brain: Insights into altered brain protein regulation of critical neuronal functions and their contributions to subsequent cognitive loss. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1865**, 2031–2039 (2019).

7. Solari, F. A., Dell'Aica, M., Sickmann, A. & Zahedi, R. P. Why phosphoproteomics is still a challenge. *Mol. BioSyst.* **11**, 1487–1493 (2015).

8. Morello, G., Salomone, S., D'Agata, V., Conforti, F. L. & Cavallaro, S. From Multi-Omics Approaches to Precision Medicine in Amyotrophic Lateral Sclerosis. *Front. Neurosci.* **14**, 577755 (2020).

9. Paulo, J. A. & Schweppe, D. K. Advances in quantitative high-throughput phosphoproteomics with sample multiplexing. *Proteomics* **21**, e2000140 (2021).

10. Bekker-Jensen, D. B. *et al.* Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. *Nat Commun* **11**, 787 (2020).

11. Meier, F. *et al.* diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nat Methods* **17**, 1229–1236 (2020).

12. Masrori, P. & Van Damme, P. Amyotrophic lateral sclerosis: a clinical review. *Eur J Neurol* **27**, 1918–1929 (2020).

13. Malik, R. & Wiedau, M. Therapeutic Approaches Targeting Protein Aggregation in Amyotrophic Lateral Sclerosis. *Front. Mol. Neurosci.* **13**, 98 (2020).

14. Blokhuis, A. M., Groen, E. J. N., Koppers, M., van den Berg, L. H. & Pasterkamp, R. J. Protein aggregation in amyotrophic lateral sclerosis. *Acta Neuropathol* **125**, 777–794 (2013).

15. Bittremieux, W., Meysman, P., Noble, W. S. & Laukens, K. Fast Open Modification Spectral Library Searching through Approximate Nearest Neighbor Indexing. *J. Proteome Res.* **17**, 3463–3474 (2018).

16. Degroeve, S. *et al.* ionbot: a novel, innovative and sensitive machine learning approach to LC-MS/MS peptide identification. *bioRxiv* (2022).

17. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).

18. Harper, J. W. & Bennett, E. J. Proteome complexity and the forces that drive proteome imbalance. *Nature* **537**, 328–338 (2016).

19. Leutert, M., Entwisle, S. W. & Villén, J. Decoding Post-Translational Modification Crosstalk With Proteomics. *Molecular & Cellular Proteomics* **20**, 100129 (2021).

20. Wang, Y., Zhang, J., Li, B. & He, Q. Advances of Proteomics in Novel PTM Discovery: Applications in Cancer Therapy. *Small Methods* **3**, 1900041 (2019).

21. Zhang, S. *et al.* Genome-wide identification of the genetic basis of amyotrophic lateral sclerosis. *Neuron* **110**, 992-1008.e11 (2022).

22. Humphrey, J. *et al.* Integrative transcriptomic analysis of the amyotrophic lateral sclerosis spinal cord implicates glial activation and suggests new risk genes. *Nature Neuroscience* **26**, 150–162 (2022).

23. Zhang, Y., Fonslow, B. R., Shan, B., Baek, M.-C. & Yates, J. R. Protein Analysis by Shotgun/Bottom-up Proteomics. *Chem. Rev.* **113**, 2343–2394 (2013).

24. Li, X., Wang, W. & Chen, J. Recent progress in mass spectrometry proteomics for biomedical research. *Sci. China Life Sci.* **60**, 1093–1113 (2017).

25. Richards, A. L. *et al.* One-hour proteome analysis in yeast. *Nat Protoc* **10**, 701–714 (2015).

26. Wilkins, M. R. *et al.* From Proteins to Proteomes: Large Scale Protein Identification by Two-Dimensional Electrophoresis and Amino Acid Analysis. *Nature Biotechnology* **14**, 61–65 (1996).

27. James, P. Protein identification in the post-genome era: the rapid rise of proteomics. *Quart. Rev. Biophys.* **30**, 279–331 (1997).

28. Meyer, J. G. & Schilling, B. Clinical applications of quantitative proteomics using targeted and untargeted data-independent acquisition techniques. *Expert Review of Proteomics* **14**, 419–429 (2017).

29. Doll, S., Gnad, F. & Mann, M. The Case for Proteomics and Phospho-Proteomics in Personalized Cancer Medicine. *Prot. Clin. Appl.* **13**, 1800113 (2019).

30. Valdés, A. *et al.* Foodomics: Analytical Opportunities and Challenges. *Anal. Chem.* **94**, 366–381 (2022).

31. Korte, R. & Brockmeyer, J. Novel mass spectrometry approaches in food proteomics. *TrAC Trends in Analytical Chemistry* **96**, 99–106 (2017).

32. Vu, L. D. *et al.* Up-to-Date Workflow for Plant (Phospho)proteomics Identifies Differential Drought-Responsive Phosphorylation Events in Maize Leaves. *J. Proteome Res.* **15**, 4304–4317 (2016).

33. Subba, P. & Prasad, T. S. K. Plant Phosphoproteomics: Known Knowns, Known Unknowns, and Unknown Unknowns of an Emerging Systems Science Frontier. *OMICS: A Journal of Integrative Biology* **25**, 750–769 (2021).

34. Cleland, T. P. Human Bone Paleoproteomics Utilizing the Single-Pot, Solid-Phase-Enhanced Sample Preparation Method to Maximize Detected Proteins and Reduce Humics. *J. Proteome Res.* **17**, 3976–3983 (2018).

35. Warinner, C., Korzow Richter, K. & Collins, M. J. Paleoproteomics. *Chem. Rev.* **122**, 13401–13446 (2022).

36. Vilanova, C. & Porcar, M. Art-omics: multi-omics meet archaeology and art conservation. *Microb. Biotechnol.* **13**, 435–441 (2020).

37. Galluzzi, F., Chaignepain, S., Arslanoglu, J. & Tokarski, C. Hydrogen-deuterium exchange mass spectrometry to study interactions and conformational changes of proteins in paints. *Biophysical Chemistry* **289**, 106861 (2022).

38. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science* **246**, 64–71 (1989).

39. Karas, Michael. & Hillenkamp, Franz. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal. Chem.* **60**, 2299–2301 (1988).

40. Nesvizhskii, A. I. & Aebersold, R. Interpretation of Shotgun Proteomic Data. *Molecular & Cellular Proteomics* **4**, 1419–1440 (2005).

41. Toby, T. K., Fornelli, L. & Kelleher, N. L. Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annual Rev. Anal. Chem.* **9**, 499–519 (2016).

42. Fornelli, L. *et al.* Top-down proteomics: Where we are, where we are going? *Journal of Proteomics* **175**, 3–4 (2018).

43. Cupp-Sutton, K. A. & Wu, S. High-throughput quantitative top-down proteomics. *Mol. Omics* **16**, 91–99 (2020).

44. LeDuc, R. D. *et al.* Accurate Estimation of Context-Dependent False Discovery Rates in Top-Down Proteomics. *Molecular & Cellular Proteomics* **18**, 796–805 (2019).

45. Ghezellou, P. *et al.* A perspective view of top-down proteomics in snake venom research. *Rapid Commun Mass Spectrom* **33**, 20–27 (2019).

46. Muntel, J. *et al.* Surpassing 10 000 identified and quantified proteins in a single run by optimizing current LC-MS instrumentation and data analysis strategy. *Mol. Omics* **15**, 348–360 (2019).

47. Reinders, J. & Sickmann, A. State-of-the-art in phosphoproteomics. *Proteomics* **5**, 4052–4061 (2005).

48. Khoonsari, P. E. *et al.* The human CSF pain proteome. *Journal of Proteomics* **190**, 67–76 (2019).

49. Shu, T. *et al.* Plasma Proteomics Identify Biomarkers and Pathogenesis of COVID-19. *Immunity* **53**, 1108-1122.e5 (2020).

50. Ding, H. *et al.* Urine Proteomics: Evaluation of Different Sample Preparation Workflows for Quantitative, Reproducible, and Improved Depth of Analysis. *J. Proteome Res.* **19**, 1857–1862 (2020).

51. Jiang, L. *et al.* A Quantitative Proteome Map of the Human Body. *Cell* **183**, 269-283.e19 (2020).

52. Hayoun, K. *et al.* Evaluation of Sample Preparation Methods for Fast Proteotyping of Microorganisms by Tandem Mass Spectrometry. *Front. Microbiol.* **10**, 1985 (2019).

53. Slavov, N. Single-cell protein analysis by mass spectrometry. *Current Opinion in Chemical Biology* **60**, 1–9 (2021).

54. Cañas, B., Piñeiro, C., Calvo, E., López-Ferrer, D. & Gallardo, J. M. Trends in sample preparation for classical and second generation proteomics. *Journal of Chromatography A* **1153**, 235–258 (2007).

55. Varnavides, G. *et al.* In Search of a Universal Method: A Comparative Survey of Bottom-Up Proteomics Sample Preparation Methods. *J. Proteome Res.* **21**, 2397–2411 (2022).

56. Shirvaliloo, M. Epigenomics in COVID-19; the link between DNA methylation, histone modifications and SARS-CoV-2 infection. *Epigenomics* **13**, 745–750 (2021).

57. Bissell, K. *et al.* Semi-automated, high-throughput homogenization technique for in-depth analysis of tissue proteome.

58. Feist, P. & Hummon, A. Proteomic Challenges: Sample Preparation Techniques for Microgram-Quantity Protein Analysis from Biological Samples. *IJMS* **16**, 3537–3563 (2015).

59. Pop, C., Mogosan, C. & Loghin, F. Evaluation of Rapigest efficacy for the digestion of proteins from cell cultures and heart tissues. *Medicine and Pharmacy Reports* **87**, 258–262 (2014).

60. Davalieva, K., Kiprijanovska, S., Dimovski, A., Rosoklija, G. & Dwork, A. J. Comparative evaluation of two methods for LC-MS/MS proteomic analysis of formalin fixed and paraffin embedded tissues. *Journal of Proteomics* **235**, 104117 (2021).

61. Shehadul Islam, M., Aryasomayajula, A. & Selvaganapathy, P. A Review on Macroscale and Microscale Cell Lysis Methods. *Micromachines* **8**, 83 (2017).

62. Novák, P. & Havlíček, V. Protein Extraction and Precipitation. in *Proteomic Profiling and Analytical Chemistry* 51–62 (Elsevier, 2016).

63. Yuan, X. & Desiderio, D. M. Proteomics analysis of human cerebrospinal fluid. *Journal of Chromatography B* **815**, 179–189 (2005).

64. Tubaon, R. M., Haddad, P. R. & Quirino, J. P. Sample Clean-up Strategies for ESI Mass Spectrometry Applications in Bottom-up Proteomics: Trends from 2012 to 2016. *Proteomics* **17**, 1700011 (2017).

65. Brownridge, P. & Beynon, R. J. The importance of the digest: Proteolysis and absolute quantification in proteomics. *Methods* **54**, 351–360 (2011).

66. Switzar, L., Giera, M. & Niessen, W. M. A. Protein Digestion: An Overview of the Available Techniques and Recent Developments. *J. Proteome Res.* **12**, 1067–1077 (2013).

67. Depuydt, M., Messens, J. & Collet, J.-F. How Proteins Form Disulfide Bonds. *Antioxidants & Redox Signaling* **15**, 49–66 (2011).

68. Cleland, W. W. Dithiothreitol, a New Protective Reagent for SH Groups. *Biochemistry* **3**, 480–482 (1964).

69. Getz, E. B., Xiao, M., Chakrabarty, T., Cooke, R. & Selvin, P. R. A Comparison between the Sulfhydryl Reductants Tris(2-carboxyethyl)phosphine and Dithiothreitol for Use in Protein Biochemistry. *Analytical Biochemistry* **273**, 73–80 (1999).

70. Müller, T. & Winter, D. Systematic Evaluation of Protein Reduction and Alkylation Reveals Massive Unspecific Side Effects by Iodine-containing Reagents. *Molecular & Cellular Proteomics* **16**, 1173–1187 (2017).

71. Vandermarliere, E., Mueller, M. & Martens, L. Getting intimate with trypsin, the leading protease in proteomics: trypsin in proteomics. *Mass Spec Rev* **32**, 453–465 (2013).

72. Tsiatsiani, L. & Heck, A. J. R. Proteomics beyond trypsin. *FEBS J* **282**, 2612–2626 (2015).

73. Perutka, Z. & Šebela, M. Pseudotrypsin: A Little-Known Trypsin Proteoform. *Molecules* **23**, 2637 (2018).

74. Dau, T., Bartolomucci, G. & Rappsilber, J. Proteomics Using Protease Alternatives to Trypsin Benefits from Sequential Digestion with Trypsin. *Anal. Chem.* **92**, 9523–9527 (2020).

75. Morsa, D. *et al.* Multi-Enzymatic Limited Digestion: The Next-Generation Sequencing for Proteomics? *J. Proteome Res.* **18**, 2501–2513 (2019).

76. Giansanti, P., Tsiatsiani, L., Low, T. Y. & Heck, A. J. R. Six alternative proteases for mass spectrometry–based proteomics beyond trypsin. *Nat Protoc* **11**, 993–1006 (2016).

77. Saveliev, S. *et al.* Trypsin/Lys-C protease mix for enhanced protein mass spectrometry analysis. *Nat Methods* **10**, i–ii (2013).

78. Glatter, T. *et al.* Large-Scale Quantitative Assessment of Different In-Solution Protein Digestion Protocols Reveals Superior Cleavage Efficiency of Tandem Lys-C/Trypsin Proteolysis over Trypsin Digestion. *J. Proteome Res.* **11**, 5145–5156 (2012).

79. Hakobyan, A., Schneider, M. B., Liesack, W. & Glatter, T. Efficient Tandem LysC/Trypsin Digestion in Detergent Conditions. *Proteomics* **19**, 1900136 (2019).

80. Waas, M., Pereckas, M., Jones Lipinski, R. A., Ashwood, C. & Gundry, R. L. SP2: Rapid and Automatable Contaminant Removal from Peptide Samples for Proteomic Analyses. *J. Proteome Res.* **18**, 1644–1656 (2019).

81. Wojtkiewicz, M., Berg Luecke, L., Kelly, M. I. & Gundry, R. L. Facile Preparation of Peptides for Mass Spectrometry Analysis in Bottom-Up Proteomics Workflows. *Current Protocols* **1**, (2021).

82. Granvogl, B., Plöscher, M. & Eichacker, L. A. Sample preparation by in-gel digestion for mass spectrometry-based proteomics. *Anal Bioanal Chem* **389**, 991–1002 (2007).

83. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat Methods* **11**, 319–324 (2014).

84. Gautam, S. S. *et al.* Label-free plasma proteomics for the identification of the putative biomarkers of oral squamous cell carcinoma. *Journal of Proteomics* **259**, 104541 (2022).

85. Bader, J. M. *et al.* Proteome profiling in cerebrospinal fluid reveals novel biomarkers of Alzheimer's disease. *Mol Syst Biol* **16**, e9356 (2020).

86. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat Methods* **6**, 359–362 (2009).

87. Elinger, D., Gabashvili, A. & Levin, Y. Suspension Trapping (S-Trap) Is Compatible with Typical Protein Extraction Buffers and Detergents for Bottom-Up Proteomics. *J. Proteome Res.* **18**, 1441–1445 (2019).

88. Zougman, A., Selby, P. J. & Banks, R. E. Suspension trapping (STrap) sample preparation method for bottom-up proteomics analysis. *Proteomics* **14**, 1006–1000 (2014).

89. Berger, S. T. *et al.* MStern Blotting–High Throughput Polyvinylidene Fluoride (PVDF) Membrane-Based Proteomic Sample Preparation for 96-Well Plates. *Molecular & Cellular Proteomics* **14**, 2814–2823 (2015).

90. Hughes, C. S. *et al.* Ultrasensitive proteome analysis using paramagnetic bead technology. *Mol Syst Biol* **10**, 757 (2014).

91. Hughes, C. S. *et al.* Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nat Protoc* **14**, 68–85 (2019).

92. Moggridge, S., Sorensen, P. H., Morin, G. B. & Hughes, C. S. Extending the Compatibility of the SP3 Paramagnetic Bead Processing Approach for Proteomics. *J. Proteome Res.* **17**, 1730–1740 (2018).

93. Müller, T. *et al.* Automated sample preparation with SP 3 for low-input clinical proteomics. *Mol Syst Biol* **16**, e9111 (2020).

94. Batth, T. S. *et al.* Protein Aggregation Capture on Microparticles Enables Multipurpose Proteomics Sample Preparation. *Molecular & Cellular Proteomics* **18**, 1027–1035 (2019).

95. Mikulášek, K. *et al.* SP3 Protocol for Proteomic Plant Sample Preparation Prior LC-MS/MS. *Front. Plant Sci.* **12**, 635550 (2021).

96. Araújo, M. J. *et al.* Comparison of Sample Preparation Methods for Shotgun Proteomic Studies in Aquaculture Species. *Proteomes* **9**, 46 (2021).

97. Gonzalez-Lozano, M. A., Koopmans, F., Paliukhovich, I., Smit, A. B. & Li, K. W. A Fast and Economical Sample Preparation Protocol for Interaction Proteomics Analysis. *Proteomics* **19**, 1900027 (2019).

98. Paulo, J. A., Navarrete-Perea, J. & Gygi, S. P. Multiplexed proteome profiling of carbon source perturbations in two yeast species with SL-SP3-TMT. *Journal of Proteomics* **210**, 103531 (2020).

99. Griesser, E. *et al.* Quantitative Profiling of the Human Substantia Nigra Proteome from Laser-capture Microdissected FFPE Tissue. *Molecular & Cellular Proteomics* **19**, 839–851 (2020).

100. Cagnetta, R., Frese, C. K., Shigeoka, T., Krijgsveld, J. & Holt, C. E. Rapid Cue-Specific Remodeling of the Nascent Axonal Proteome. *Neuron* **99**, 29-46.e4 (2018).

101. Sielaff, M. *et al.* Evaluation of FASP, SP3, and iST Protocols for Proteomic Sample Preparation in the Low Microgram Range. *J. Proteome Res.* **16**, 4060–4072 (2017).

102. Neset, L., Takayidza, G., Berven, F. S. & Hernandez-Valladares, M. Comparing Efficiency of Lysis Buffer Solutions and Sample Preparation Methods for Liquid Chromatography–Mass Spectrometry Analysis of Human Cells and Plasma. *Molecules* **27**, 3390 (2022).

103. Balotf, S., Wilson, R., Tegg, R. S., Nichols, D. S. & Wilson, C. R. Optimisation of Sporosori Purification and Protein Extraction Techniques for the Biotrophic Protozoan Plant Pathogen Spongospora subterranea. *Molecules* **25**, 3109 (2020).

104. Ludwig, K. R., Schroll, M. M. & Hummon, A. B. Comparison of In-Solution, FASP, and S-Trap Based Digestion Methods for Bottom-Up Proteomic Studies. *J. Proteome Res.* **17**, 2480–2490 (2018).

105. Muller, L., Fornecker, L., Van Dorsselaer, A., Cianférani, S. & Carapito, C. Benchmarking sample preparation/digestion protocols reveals tube-gel being a fast and repeatable method for quantitative proteomics. *Proteomics* **16**, 2953–2961 (2016).

106. Yang, Y., Anderson, E. & Zhang, S. Evaluation of six sample preparation procedures for qualitative and quantitative proteomics analysis of milk fat globule membrane. *Electrophoresis* **39**, 2332–2339 (2018).

107. Duong, V.-A. & Lee, H. Bottom-Up Proteomics: Advancements in Sample Preparation. *IJMS* **24**, 5350 (2023).

108. Yang, Z., Shen, X., Chen, D. & Sun, L. Toward a Universal Sample Preparation Method for Denaturing Top-Down Proteomics of Complex Proteomes. *J. Proteome Res.* **19**, 3315–3325 (2020).

109. Dagley, L. F., Infusini, G., Larsen, R. H., Sandow, J. J. & Webb, A. I. Universal Solid-Phase Protein Preparation (USP3) for Bottom-up and Top-down Proteomics. *J. Proteome Res.* **18**, 2915–2924 (2019).

110. Virant-Klun, I., Leicht, S., Hughes, C. & Krijgsveld, J. Identification of Maturation-Specific Proteins by Single-Cell Proteomics of Human Oocytes. *Molecular & Cellular Proteomics* **15**, 2616–2627 (2016).

111. Zecha, J. *et al.* Data, Reagents, Assays and Merits of Proteomics for SARS-CoV-2 Research and Testing. *Molecular & Cellular Proteomics* **19**, 1503–1522 (2020).

112. Johnston, H. E. *et al.* Solvent Precipitation SP3 (SP4) Enhances Recovery for Proteomics Sample Preparation without Magnetic Beads. *Anal. Chem.* **94**, 10320–10328 (2022).

113. Wang, S., Kojima, K., Mobley, J. A. & West, A. B. Proteomic analysis of urinary extracellular vesicles reveal biomarkers for neurologic disease. *EBioMedicine* **45**, 351–361 (2019).

114. Higginbotham, L. A. *et al.* Integrated proteomics reveals brain-based cerebrospinal fluid biomarkers in asymptomatic and symptomatic Alzheimer's disease. *Science advances* **6**, eaaz9360 (2020).

115. Alexovič, M., Urban, P. L., Tabani, H. & Sabo, J. Recent advances in robotic protein sample preparation for clinical analysis and other biomedical applications. *Clinica Chimica Acta* **507**, 104–116 (2020).

116. Rivera, K. D. *et al.* Automating UbiFast for High-throughput and Multiplexed Ubiquitin Enrichment. *Molecular & Cellular Proteomics* **20**, 100154 (2021).

117. Liu, L. *et al.* Automated Intact Glycopeptide Enrichment Method Facilitating Highly Reproducible Analysis of Serum Site-Specific N-Glycoproteome. *Anal. Chem.* **93**, 7473–7480 (2021).

118. Birk, M. S., Charpentier, E. & Frese, C. K. Automated Phosphopeptide Enrichment for Gram-Positive Bacteria. *J. Proteome Res.* **20**, 4886–4892 (2021).

119. Pollock, S., Wu, S., Han, J. & Murphy, S. Automated MHC-Associated Peptide Enrichment for Immunopeptidomics Analysis Using Agilent AssayMAP Bravo Large Capacity Cartridges. *Agilent Application Note* (2020).

120. Matzinger, M., Müller, E., Dürnberger, G., Pichler, P. & Mechtler, K. Robust and Easy-to-Use One-Pot Workflow for Label-Free Single-Cell Proteomics. *Anal. Chem.* **95**, 4435–4445 (2023).

121. Zhang, H. *et al.* Acoustic Ejection Mass Spectrometry for High-Throughput Analysis. *Anal. Chem.* **93**, 10850–10861 (2021).

122. Camerini, S. & Mauri, P. The role of protein and peptide separation before mass spectrometry analysis in clinical proteomics. *Journal of Chromatography A* **1381**, 1–12 (2015).

123. Gillet, L. C., Leitner, A. & Aebersold, R. Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. *Annual Rev. Anal. Chem.* **9**, 449–472 (2016).

124. Dams, M., Dores-Sousa, J. L., Lamers, R.-J., Treumann, A. & Eeltink, S. High-Resolution Nano-Liquid Chromatography with Tandem Mass Spectrometric Detection for the Bottom-Up Analysis of Complex Proteomic Samples. *Chromatographia* **82**, 101–110 (2019).

125. Šesták, J., Moravcová, D. & Kahle, V. Instrument platforms for nano liquid chromatography. *Journal of Chromatography A* **1421**, 2–17 (2015).

126. Hodge, K., Have, S. T., Hutton, L. & Lamond, A. I. Cleaning up the masses: Exclusion lists to reduce contamination with HPLC-MS/MS. *Journal of Proteomics* **88**, 92–103 (2013).

127. Dodds, J. N. & Baker, E. S. Ion Mobility Spectrometry: Fundamental Concepts, Instrumentation, Applications, and the Road Ahead. *J. Am. Soc. Mass Spectrom.* **30**, 2185–2195 (2019).

128. Zhou, M. *et al.* Higher-order structural characterisation of native proteins and complexes by top-down mass spectrometry. *Chem. Sci.* **11**, 12918–12936 (2020).

129. Bonneil, E., Pfammatter, S. & Thibault, P. Enhancement of mass spectrometry performance for proteomic analyses using high-field asymmetric waveform ion mobility spectrometry (FAIMS): Application of FAIMS to proteomics. *J. Mass Spectrom.* **50**, 1181–1195 (2015).

130. Meier, F. *et al.* Parallel Accumulation–Serial Fragmentation (PASEF): Multiplying Sequencing Speed and Sensitivity by Synchronized Scans in a Trapped Ion Mobility Device. *J. Proteome Res.* **14**, 5378–5387 (2015).

131. Meier, F. *et al.* Online Parallel Accumulation–Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer. *Molecular & Cellular Proteomics* **17**, 2534–2545 (2018).

132. Ogata, K., Chang, C.-H. & Ishihama, Y. Effect of Phosphorylation on the Collision Cross Sections of Peptide Ions in Ion Mobility Spectrometry. *Mass Spectrometry* **10**, A0093–A0093 (2021).

133. Oliinyk, D. & Meier, F. Ion mobility-resolved phosphoproteomics with dia-PASEF and short gradients. *Proteomics* **23**, e2200032 (2022).

134. Wysocki, V. H., Tsaprailis, G., Smith, L. L. & Breci, L. A. Mobile and localized protons: a framework for understanding peptide dissociation. *J. Mass Spectrom.* **35**, 1399–1406 (2000).

135. Penkert, M. *et al.* Electron Transfer/Higher Energy Collisional Dissociation of Doubly Charged Peptide Ions: Identification of Labile Protein Phosphorylations. *J. Am. Soc. Mass Spectrom.* **30**, 1578–1585 (2019).

136. Wiesner, J., Premsler, T. & Sickmann, A. Application of electron transfer dissociation (ETD) for the analysis of posttranslational modifications. *Proteomics* **8**, 4466–4483 (2008).

137. Williams, J. P. *et al.* Top-Down Characterization of Denatured Proteins and Native Protein Complexes Using Electron Capture Dissociation Implemented within a Modified Ion Mobility-Mass Spectrometer. *Anal. Chem.* **92**, 3674–3681 (2020).

138. Fort, K. L. *et al.* Implementation of Ultraviolet Photodissociation on a Benchtop Q Exactive Mass Spectrometer and Its Application to Phosphoproteomics. *Anal. Chem.* **88**, 2303–2310 (2016).

139. Frese, C. K. *et al.* Unambiguous Phosphosite Localization using Electron-Transfer/Higher-Energy Collision Dissociation (EThcD). *J. Proteome Res.* **12**, 1520–1525 (2013).

140. Li, Y. F. & Radivojac, P. Computational approaches to protein inference in shotgun proteomics. *BMC Bioinformatics* **13**, S4 (2012).

141. Cox, J. *et al.* Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J. Proteome Res.* **10**, 1794–1805 (2011).

142. Blueggel, M., Chamrad, D. & Meyer, H. E. Bioinformatics in Proteomics. *Current Pharmaceutical Biotechnology* **5**, 77–88 (2004).

143. Dorfer, V. *et al.* MS Amanda, a Universal Identification Algorithm Optimized for High Accuracy Tandem Mass Spectra. *J. Proteome Res.* **13**, 3679–3684 (2014).

144. Griss, J. *et al.* Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nat Methods* **13**, 651–656 (2016).

145. Pathan, M., Samuel, M., Keerthikumar, S. & Mathivanan, S. Unassigned MS/MS Spectra: Who Am I? in *Proteome Bioinformatics* (eds. Keerthikumar, S. & Mathivanan, S.) vol. 1549 67–74 (Springer New York, 2017).

146. Dorfer, V., Maltsev, S., Winkler, S. & Mechtler, K. CharmeRT: Boosting Peptide Identifications by Chimeric Spectra Identification and Retention Time Prediction. *J. Proteome Res.* **17**, 2581–2589 (2018).

147. Skinner, O. S. & Kelleher, N. L. Illuminating the dark matter of shotgun proteomics. *Nat Biotechnol* **33**, 717–718 (2015).

148. den Ridder, M., Daran-Lapujade, P. & Pabst, M. Shot-gun proteomics: why thousands of unidentified signals matter. *FEMS Yeast Research* **20**, foz088 (2020).

149. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat Methods* **14**, 513–520 (2017).

150. David, M., Fertin, G., Rogniaux, H. & Tessier, D. SpecOMS: A Full Open Modification Search Method Performing All-to-All Spectra Comparisons within Minutes. *J. Proteome Res.* **16**, 3030–3038 (2017).

151. Apweiler, R. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research* **32**, 115D – 119 (2004).

152. The UniProt Consortium *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49**, D480–D489 (2021).

153. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733–D745 (2016).

154. Armengaud, J. A perfect genome annotation is within reach with the proteomics and genomics alliance. *Current Opinion in Microbiology* **12**, 292–300 (2009).

155. Vaudel, M. *et al.* PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat Biotechnol* **33**, 22–24 (2015).

156. Vaudel, M. *et al.* Exploring the potential of public proteomics data. *Proteomics* **16**, 214–225 (2016).

157. Elias, J. E. & Gygi, S. P. Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics. in *Proteome Bioinformatics* (eds. Hubbard, S. J. & Jones, A. R.) vol. 604 55–71 (Humana Press, 2010).

158. Bouyssié, D. *et al.* Proline: an efficient and user-friendly software suite for large-scale proteomics. *Bioinformatics* **36**, 3148–3155 (2020).

159. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* **4**, 923–925 (2007).

160. Bogdanow, B., Zauber, H. & Selbach, M. Systematic Errors in Peptide and Protein Identification and Quantification by Modified Peptides. *Molecular & Cellular Proteomics* **15**, 2791–2801 (2016).

161. Na, S. & Paek, E. Software eyes for protein post-translational modifications: eyes for PTMs. *Mass Spec Rev* **34**, 133–147 (2015).

162. Ahrné, E., Nikitin, F., Lisacek, F. & Müller, M. QuickMod: A Tool for Open Modification Spectrum Library Searches. *J. Proteome Res.* **10**, 2913–2921 (2011).

163. Ye, D. *et al.* Open MS/MS spectral library search to identify unanticipated post-translational modifications and increase spectral identification rate. *Bioinformatics* **26**, i399–i406 (2010).

164. Burke, M. C. *et al.* The Hybrid Search: A Mass Spectral Library Search Method for Discovery of Modifications in Proteomics. *J. Proteome Res.* **16**, 1924–1935 (2017).

165. Yu, F., Li, N. & Yu, W. PIPI: PTM-Invariant Peptide Identification Using Coding Method. *J. Proteome Res.* **15**, 4423–4435 (2016).

166. Na, S., Bandeira, N. & Paek, E. Fast Multi-blind Modification Search through Tandem Mass Spectrometry. *Molecular & Cellular Proteomics* **11**, M111.010199 (2012).

167. Devabhaktuni, A. *et al.* Measuring proteomes with long strings: A new, unconstrained paradigm in mass spectrum interpretation. *bioRxiv* (2018).

168. Chi, H. *et al.* Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nat Biotechnol* **36**, 1059–1061 (2018).

169. Schulze, S. *et al.* Enhancing Open Modification Searches via a Combined Approach Facilitated by Ursgal. *J. Proteome Res.* **20**, 1986–1996 (2021).

170. Altenburg, T., Muth, T. & Renard, B. Y. yHydra: Deep Learning enables an Ultra Fast Open Search by Jointly Embedding MS/MS Spectra and Peptides of Mass Spectrometry-based Proteomics. *bioRxiv* (2021).

171. Arab, I., Fondrie, W. E., Laukens, K. & Bittremieux, W. Semisupervised Machine Learning for Sensitive Open Modification Spectral Library Searching. *J. Proteome Res.* **22**, 585–593 (2023).

172. Bouwmeester, R., Gabriels, R., Hulstaert, N., Martens, L. & Degroeve, S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nature Methods* **18**, 1363–1369 (2021).

173. Degroeve, S. & Martens, L. MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics* **29**, 3199–3203 (2013).

174. Neilson, K. A. *et al.* Less label, more free: Approaches in label-free quantitative mass spectrometry. *Proteomics* **11**, 535–553 (2011).

175. Blein-Nicolas, M. & Zivy, M. Thousand and one ways to quantify and compare protein abundances in label-free bottom-up proteomics. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1864**, 883–895 (2016).

176. Wang, X., Shen, S., Rasam, S. S. & Qu, J. MS1 ion current-based quantitative proteomics: A promising solution for reliable analysis of large biological cohorts. *Mass Spec Rev* **38**, 461–482 (2019).

177. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* **11**, 2301–2319 (2016).

178. Escher, C. *et al.* Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* **12**, 1111–1121 (2012).

179. Xu, H. *et al.* PTMD: A Database of Human Disease-associated Post-translational Modifications. *Genomics, Proteomics & Bioinformatics* **16**, 244–251 (2018).

180. von Stechow, L., Francavilla, C. & Olsen, J. V. Recent findings and technological advances in phosphoproteomics for cells and tissues. *Expert Review of Proteomics* **12**, 469–487 (2015).

181. Ardito, F., Giuliani, M., Perrone, D., Troiano, G. & Muzio, L. L. The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (Review). *International Journal of Molecular Medicine* **40**, 271–280 (2017).

182. Urban, J. A review on recent trends in the phosphoproteomics workflow. From sample preparation to data analysis. *Analytica Chimica Acta* **1199**, 338857 (2022).

183. Low, T. Y. *et al.* WIDENING THE BOTTLENECK OF PHOSPHOPROTEOMICS: EVOLVING STRATEGIES FOR PHOSPHOPEPTIDE ENRICHMENT. *Mass Spec Rev* **40**, 309–333 (2021).

184. Qiu, W., Evans, C. A., Landels, A., Pham, T. K. & Wright, P. C. Phosphopeptide enrichment for phosphoproteomic analysis - A tutorial and review of novel materials. *Analytica Chimica Acta* **1129**, 158–180 (2020).

185. Vlastaridis, P. *et al.* Estimating the total number of phosphoproteins and phosphorylation sites in eukaryotic proteomes. *GigaScience* **6**, (2017).

186. Paulo, J. A. & Schweppe, D. K. Advances in quantitative high-throughput phosphoproteomics with sample multiplexing. *Proteomics* **21**, 2000140 (2021).

187. Singh, V. *et al.* Phosphorylation: Implications in Cancer. *Protein J* **36**, 1–6 (2017).

188. Coopman, P. Protein Phosphorylation in Cancer: Unraveling the Signaling Pathways. *Biomolecules* **12**, 1036 (2022).

189. Wegmann, S., Biernat, J. & Mandelkow, E. A current view on Tau protein phosphorylation in Alzheimer's disease. *Current Opinion in Neurobiology* **69**, 131–138 (2021).

190. De Schaepdryver, M. *et al.* Comparison of elevated phosphorylated neurofilament heavy chains in serum and cerebrospinal fluid of patients with amyotrophic lateral sclerosis. *J Neurol Neurosurg Psychiatry* **89**, 367–373 (2018).

191. Gendron, T. F. *et al.* Phosphorylated neurofilament heavy chain: a biomarker of survival for C9ORF72-associated amyotrophic lateral sclerosis. *Ann Neurol.* **82**, 139–146 (2017).

192. Hasegawa, M. *et al.* Phosphorylated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Ann Neurol.* **64**, 60–70 (2008).

193. El-Armouche, A. *et al.* Decreased phosphorylation levels of cardiac myosin-binding protein-C in human and experimental heart failure. *Journal of Molecular and Cellular Cardiology* **43**, 223–229 (2007).

194. Chan, C. Y. X., Gritsenko, M. A., Smith, R. D. & Qian, W.-J. The current state of the art of quantitative phosphoproteomics and its applications to diabetes research. *Expert Review of Proteomics* **13**, 421–433 (2016).

195. Roskoski, R. Properties of FDA-approved small molecule protein kinase inhibitors: A 2023 update. *Pharmacological Research* **187**, 106552 (2023).

196. Olsen, J. V. & Mann, M. Status of Large-scale Analysis of Post-translational Modifications by Mass Spectrometry. *Molecular & Cellular Proteomics* **12**, 3444–3452 (2013).

197. Riley, N. M. & Coon, J. J. Phosphoproteomics in the Age of Rapid and Deep Proteome Profiling. *Anal. Chem.* **88**, 74–94 (2016).

198. Bubis, J. A., Gorshkov, V., Gorshkov, M. V. & Kjeldsen, F. PhosphoShield: Improving Trypsin Digestion of Phosphoproteins by Shielding the Negatively Charged Phosphate Moiety. *J. Am. Soc. Mass Spectrom.* **31**, 2053–2060 (2020).

199. Fíla, J. & Honys, D. Enrichment techniques employed in phosphoproteomics. *Amino Acids* **43**, 1025–1047 (2012).

200. Arrington, J. V., Hsu, C.-C., Elder, S. G. & Andy Tao, W. Recent advances in phosphoproteomics and application to neurological diseases. *Analyst* **142**, 4373–4387 (2017).

201. Leitner, A. Phosphopeptide enrichment using metal oxide affinity chromatography. *TrAC Trends in Analytical Chemistry* **29**, 177–185 (2010).

202. Yi, L. *et al.* Targeted Quantification of Phosphorylation Dynamics in the Context of EGFR-MAPK Pathway. *Anal. Chem.* **90**, 5256–5263 (2018).

203. Matheron, L., van den Toorn, H., Heck, A. J. R. & Mohammed, S. Characterization of Biases in Phosphopeptide Enrichment by Ti^{4+} -Immobilized Metal Affinity Chromatography and TiO_2 Using a Massive Synthetic Library and Human Cell Digests. *Anal. Chem.* **86**, 8312–8320 (2014).

204. Thingholm, T. E., Jensen, O. N., Robinson, P. J. & Larsen, M. R. SIMAC (Sequential Elution from IMAC), a Phosphoproteomics Strategy for the Rapid Separation of Monophosphorylated from Multiply Phosphorylated Peptides. *Molecular & Cellular Proteomics* **7**, 661–671 (2008).

205. Rogers, J. C. Sequential Enrichment from Metal Oxide Affinity Chromatography (SMOAC), a Phosphoproteomics Strategy for the Separation of Multiply Phosphorylated from Monophosphorylated Peptides. (2017).

206. Boersema, P. J., Mohammed, S. & Heck, A. J. R. Hydrophilic interaction liquid chromatography (HILIC) in proteomics. *Anal Bioanal Chem* **391**, 151–159 (2008).

207. Lombardi, B., Rendell, N., Edwards, M., Katan, M. & Zimmermann, J. G. Evaluation of phosphopeptide enrichment strategies for quantitative TMT analysis of complex network dynamics in cancer-associated cell signalling. *EuPA Open Proteomics* **6**, 10–15 (2015).

208. Batth, T. S., Francavilla, C. & Olsen, J. V. Off-Line High-pH Reversed-Phase Fractionation for In-Depth Phosphoproteomics. *J. Proteome Res.* **13**, 6176–6186 (2014).

209. Johnson, H. & White, F. M. Toward quantitative phosphotyrosine profiling in vivo. *Seminars in Cell & Developmental Biology* **23**, 854–862 (2012).

210. Abe, Y., Nagano, M., Tada, A., Adachi, J. & Tomonaga, T. Deep Phosphotyrosine Proteomics by Optimization of Phosphotyrosine Enrichment and MS/MS Parameters. *J. Proteome Res.* **16**, 1077–1086 (2017).

211. Ruprecht, B. *et al.* Comprehensive and Reproducible Phosphopeptide Enrichment Using Iron Immobilized Metal Ion Affinity Chromatography (Fe-IMAC) Columns. *Molecular & Cellular Proteomics* **14**, 205–215 (2015).

212. Murillo, J. R. *et al.* Automated phosphopeptide enrichment from minute quantities of frozen malignant melanoma tissue. *PLoS ONE* **13**, e0208562 (2018).

213. Liu, X. *et al.* Fe3+-NTA magnetic beads as an alternative to spin column-based phosphopeptide enrichment. *Journal of Proteomics* **260**, 104561 (2022).

214. Humphrey, S. J., Karayel, O., James, D. E. & Mann, M. High-throughput and high-sensitivity phosphoproteomics with the EasyPhos platform. *Nat Protoc* **13**, 1897–1916 (2018).

215. Humphrey, S. J., Azimifar, S. B. & Mann, M. High-throughput phosphoproteomics reveals in vivo insulin signaling dynamics. *Nat Biotechnol* **33**, 990–995 (2015).

216. Oliinyk, D., Will, A., Schneidmadel, F. R., Humphrey, S. J. & Meier, F. µPhos: a scalable and sensitive platform for functional phosphoproteomics. *bioRxiv* (2023).

217. Leutert, M., Rodríguez-Mias, R. A., Fukuda, N. K. & Villén, J. R2-P2 rapid-robotic phosphoproteomics enables multidimensional cell signaling studies. *Mol Syst Biol* **15**, (2019).

218. Voinov, V. G. *et al.* A Novel, Automated and Highly Selective Phosphopeptide Enrichment for Phosphopeptide Identification and Phosphosite Localization. *Agilent Application Note* (2020).

219. Post, H. *et al.* Robust, Sensitive, and Automated Phosphopeptide Enrichment Optimized for Low Sample Amounts Applied to Primary Hippocampal Neurons. *J. Proteome Res.* **16**, 728–737 (2017).

220. Wu, S. & Wu, L. Human Breast Cancer Cell Line Phosphoproteome Revealed by an Automated and Highly Selective Enrichment Workflow. *Agilent Application Note* (2018).

221. Zeneyedpour, L. *et al.* Phosphorylation Ratio Determination in Fresh-Frozen and Formalin-Fixed Paraffin-Embedded Tissue with Targeted Mass Spectrometry. *J. Proteome Res.* **19**, 4179–4190 (2020).

222. Gao, Y. & Wang, Y. A method to determine the ionization efficiency change of peptides caused by phosphorylation. *J. Am. Soc. Mass Spectrom.* **18**, 1973–1976 (2007).

223. Steen, H., Jebanathirajah, J. A., Rush, J., Morrice, N. & Kirschner, M. W. Phosphorylation Analysis by Mass Spectrometry. *Molecular & Cellular Proteomics* **5**, 172–181 (2006).

224. Potel, C. M., Lemeer, S. & Heck, A. J. R. Phosphopeptide Fragmentation and Site Localization by Mass Spectrometry: An Update. *Anal. Chem.* **91**, 126–141 (2019).

225. Robinson, M. R., Taliaferro, J. M., Dalby, K. N. & Brodbelt, J. S. 193 nm Ultraviolet Photodissociation Mass Spectrometry for Phosphopeptide Characterization in the Positive and Negative Ion Modes. *J. Proteome Res.* **15**, 2739–2748 (2016).

226. Tang, N., Perkins, P., Miller, C. & van de Goor, T. Protein Phosphorylation Sites Determination Using A Microfluidic Chip Interfaced With ETD Ion Trap And Q-TOF Mass Spectrometry. (2007).

227. Sathe, G. *et al.* Multiplexed Phosphoproteomic Study of Brain in Patients with Alzheimer's Disease and Age-Matched Cognitively Healthy Controls. *OMICS: A Journal of Integrative Biology* **24**, 216–227 (2020).

228. Ping, L. *et al.* Global quantitative analysis of the human brain proteome and phosphoproteome in Alzheimer's disease. *Sci Data* **7**, 315 (2020).

229. Qin, G. *et al.* iTRAQ-based quantitative phosphoproteomics provides insights into the metabolic and physiological responses of a carnivorous marine fish (Nibea albiflora) fed a linseed oil-rich diet. *Journal of Proteomics* **228**, 103917 (2020).

230. Baro, B. *et al.* SILAC-based phosphoproteomics reveals new PP2A-Cdc55-regulated processes in budding yeast. *GigaScience* **7**, (2018).

231. Koenig, C., Martinez-Val, A., Franciosa, G. & Olsen, J. V. Optimal analytical strategies for sensitive and quantitative phosphoproteomics using TMT-based multiplexing. *Proteomics* **22**, 19–20 (2022).

232. Hogrebe, A. *et al.* Benchmarking common quantification strategies for large-scale phosphoproteomics. *Nat Commun* **9**, 1045 (2018).

233. Li, J. *et al.* TMTpro-18plex: The Expanded and Complete Set of TMTpro Reagents for Sample Multiplexing. *J. Proteome Res.* **20**, 2964–2972 (2021).

234. Frost, D. C., Feng, Y. & Li, L. 21-plex DiLeu Isobaric Tags for High-Throughput Quantitative Proteomics. *Anal. Chem.* **92**, 8228–8234 (2020).

235. Wang, Z. *et al.* 27-Plex Tandem Mass Tag Mass Spectrometry for Profiling Brain Proteome in Alzheimer's Disease. *Anal. Chem.* **92**, 7162–7170 (2020).

236. Zhang, Y. *et al.* Comparative Assessment of Quantification Methods for Tumor Tissue Phosphoproteomics. *Anal. Chem.* **94**, 10893–10906 (2022).

237. Ogata, K., Tsai, C.-F. & Ishihama, Y. Nanoscale Solid-Phase Isobaric Labeling for Multiplexed Quantitative Phosphoproteomics. *J. Proteome Res.* **20**, 4193–4202 (2021).

238. Dephoure, N., Gould, K. L., Gygi, S. P. & Kellogg, D. R. Mapping and analysis of phosphorylation sites: a quick guide for cell biologists. *MBoC* **24**, 535–542 (2013).

239. Stepath, M. *et al.* Systematic Comparison of Label-Free, SILAC, and TMT Techniques to Study Early Adaption toward Inhibition of EGFR Signaling in the Colorectal Cancer Cell Line DiFi. *J. Proteome Res.* **19**, 926–937 (2020).

240. Michna, T. PASEF-DDA enables deep coverage single-shot phosphoproteomics and ion mobility-based elucidation of phosphosite isomers. (2020).

241. Skowronek, P. *et al.* Rapid and In-Depth Coverage of the (Phospho-)Proteome With Deep Libraries and Optimal Window Design for dia-PASEF. *Molecular & Cellular Proteomics* **21**, 100279 (2022).

242. Adoni, K. R., Cunningham, D. L., Heath, J. K. & Leney, A. C. FAIMS Enhances the Detection of PTM Crosstalk Sites. *J. Proteome Res.* **21**, 930–939 (2022).

243. Zhao, H., Cunningham, D. L., Creese, A. J., Heath, J. K. & Cooper, H. J. FAIMS and Phosphoproteomics of Fibroblast Growth Factor Signaling: Enhanced Identification of Multiply Phosphorylated Peptides. *J. Proteome Res.* **14**, 5077–5087 (2015).

244. Muehlbauer, L. K., Hebert, A. S., Westphall, M. S., Shishkova, E. & Coon, J. J. Global Phosphoproteome Analysis Using High-Field Asymmetric Waveform Ion Mobility Spectrometry on a Hybrid Orbitrap Mass Spectrometer. *Anal. Chem.* **92**, 15959–15967 (2020).

245. Fuseau, C. *et al.* Differential Phosphoproteomics Deciphers Physiopathology of High-Risk Mantle Cell Lymphoma. *Blood* **140**, 9300–9301 (2022).

246. Adams, C. *et al.* TIMScore with PaSER: Exploiting the CCS-dimension. *Bruker Application Note* (2022).

247. Srinivasan, A., Sing, J. C., Gingras, A.-C. & Röst, H. L. Improving Phosphoproteomics Profiling Using Data-Independent Mass Spectrometry. *J. Proteome Res.* **21**, 1789–1799 (2022).

248. López, E. *et al.* Technical phosphoproteomic and bioinformatic tools useful in cancer research. *J Clin Bioinformatics* **1**, 26 (2011).

249. Taus, T. *et al.* Universal and Confident Phosphorylation Site Localization Using phosphoRS. *J. Proteome Res.* **10**, 5354–5362 (2011).

250. Savitski, M. M. *et al.* Confident Phosphorylation Site Localization Using the Mascot Delta Score. *Molecular & Cellular Proteomics* **10**, S1–S12 (2011).

251. Locard-Paulet, M., Bouyssié, D., Froment, C., Burlet-Schiltz, O. & Jensen, L. J. Comparing 22 Popular Phosphoproteomics Pipelines for Peptide Identification and Site Localization. *J. Proteome Res.* **19**, 1338–1345 (2020).

252. Jiang, X. *et al.* Evaluation of search engines for phosphopeptide identification and quantitation. *Agilent Application Note* (2016).

253. Chalkley, R. J. & Clauser, K. R. Modification Site Localization Scoring: Strategies and Performance. *Molecular & Cellular Proteomics* **11**, 3–14 (2012).

254. Fermin, D., Walmsley, S. J., Gingras, A.-C., Choi, H. & Nesvizhskii, A. I. LuciPHOr: Algorithm for Phosphorylation Site Localization with False Localization Rate Estimation Using Modified Target-Decoy Approach. *Molecular & Cellular Proteomics* **12**, 3409–3419 (2013).

255. Baker, P. R., Trinidad, J. C. & Chalkley, R. J. Modification Site Localization Scoring Integrated into a Search Engine. *Molecular & Cellular Proteomics* **10**, M111.008078 (2011).

256. Ramsbottom, K. A. *et al.* Method for Independent Estimation of the False Localization Rate for Phosphoproteomics. *J. Proteome Res.* **21**, 1603–1615 (2022).

257. Ressa, A., Fitzpatrick, M., van den Toorn, H., Heck, A. J. R. & Altelaar, M. PaDuA: A Python Library for High-Throughput (Phospho)proteomics Data Analysis. *J. Proteome Res.* **18**, 576–584 (2019).

258. Kim, H. J. *et al.* PhosR enables processing and functional analysis of phosphoproteomic data. *Cell Reports* **34**, 108771 (2021).

259. Ramasamy, P. *et al.* Scop3P: A Comprehensive Resource of Human Phosphosites within Their Full Context. *J. Proteome Res.* **19**, 3478–3486 (2020).

260. Hornbeck, P. V. *et al.* 15 years of PhosphoSitePlus®: integrating post-translationally modified sites, disease variants and isoforms. *Nucleic Acids Research* **47**, D433–D441 (2019).

261. Gnad, F., Gunawardena, J. & Mann, M. PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Research* **39**, D253–D260 (2011).

262. Savage, S. R. & Zhang, B. Using phosphoproteomics data to understand cellular signaling: a comprehensive guide to bioinformatics resources. *Clin Proteom* **17**, 27 (2020).

263. Hu, A., Noble, W. S. & Wolf-Yadlin, A. Technical advances in proteomics: new developments in data-independent acquisition. *F1000Res* **5**, 419 (2016).

264. Venable, J. D., Dong, M.-Q., Wohlschlegel, J., Dillin, A. & Yates, J. R. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat Methods* **1**, 39–45 (2004).

265. Zhang, F., Ge, W., Ruan, G., Cai, X. & Guo, T. Data-Independent Acquisition Mass Spectrometry-Based Proteomics and Software Tools: A Glimpse in 2020. *Proteomics* **20**, 1900276 (2020).

266. Purvine, S., Eppel*, J.-T., Yi, E. C. & Goodlett, D. R. Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. *Proteomics* **3**, 847–850 (2003).

267. Silva, J. C. *et al.* Quantitative Proteomic Analysis by Accurate Mass Retention Time Pairs. *Anal. Chem.* **77**, 2187–2200 (2005).

268. Panchaud, A. *et al.* Precursor Acquisition Independent From Ion Count: How to Dive Deeper into the Proteomics Ocean. *Anal. Chem.* **81**, 6481–6488 (2009).

269. Geiger, T., Cox, J. & Mann, M. Proteomics on an Orbitrap Benchtop Mass Spectrometer Using All-ion Fragmentation. *Molecular & Cellular Proteomics* **9**, 2252–2261 (2010).

270. Carvalho, P. C. *et al.* XDIA: improving on the label-free data-independent analysis. *Bioinformatics* **26**, 847–848 (2010).

271. Weisbrod, C. R., Eng, J. K., Hoopmann, M. R., Baker, T. & Bruce, J. E. Accurate Peptide Fragment Mass Analysis: Multiplexed Peptide Identification and Quantification. *J. Proteome Res.* **11**, 1621–1632 (2012).

272. Gillet, L. C. *et al.* Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Molecular & Cellular Proteomics* **11**, O111.016717. (2012).

273. Geromanos, S. J., Hughes, C., Ciavarini, S., Vissers, J. P. C. & Langridge, J. I. Using ion purity scores for enhancing quantitative accuracy and precision in complex proteomics samples. *Anal Bioanal Chem* **404**, 1127–1139 (2012).

274. Egertson, J. D. *et al.* Multiplexed MS/MS for improved data-independent acquisition. *Nat Methods* **10**, 744–746 (2013).

275. Zabrouskov, V. *et al.* Large-Scale Targeted Protein Quantification Using Wide Selected-Ion Monitoring Data-Independent Acquisition. *LCGC* **12**, 19–25 (2014).

276. Prakash, A. *et al.* Hybrid Data Acquisition and Processing Strategies with Increased Throughput and Selectivity: pSMART Analysis for Global Qualitative and Quantitative Analysis. *J. Proteome Res.* **13**, 5415–5430 (2014).

277. Distler, U. *et al.* Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. *Nat Methods* **11**, 167–170 (2014).

278. Zhang, Y. *et al.* The Use of Variable Q1 Isolation Windows Improves Selectivity in LC–SWATH–MS Acquisition. *J. Proteome Res.* **14**, 4359–4371 (2015).

279. Bruderer, R. *et al.* Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues. *Molecular & Cellular Proteomics* **14**, 1400–1410 (2015).

280. Moseley, M. A. *et al.* Scanning Quadrupole Data-Independent Acquisition, Part A: Qualitative and Quantitative Characterization. *J. Proteome Res.* **17**, 770–779 (2018).

281. Meier, F., Geyer, P. E., Virreira Winter, S., Cox, J. & Mann, M. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat Methods* **15**, 440–448 (2018).

282. Bekker-Jensen, D. B. *et al.* A Compact Quadrupole-Orbitrap Mass Spectrometer with FAIMS Interface Improves Proteome Coverage in Short LC Gradients. *Molecular & Cellular Proteomics* **19**, 716–729 (2020).

283. Guan, S., Taylor, P. P., Han, Z., Moran, M. F. & Ma, B. Data Dependent–Independent Acquisition (DDIA) Proteomics. *J. Proteome Res.* **19**, 3230–3237 (2020).

284. Messner, C. B. *et al.* Ultra-fast proteomics with Scanning SWATH. *Nat Biotechnol* **39**, 846–854 (2021).

285. Cai, X. *et al.* PulseDIA: Data-Independent Acquisition Mass Spectrometry Using Multi-Injection Pulsed Gas-Phase Fractionation. *J. Proteome Res.* **20**, 279–288 (2021).

286. Salovska, B., Li, W., Di, Y. & Liu, Y. BoxCarmax: A High-Selectivity Data-Independent Acquisition Mass Spectrometry Method for the Analysis of Protein Turnover and Complex Samples. *Anal. Chem.* **93**, 3103–3111 (2021).

287. Guo, J., Shen, S., Xing, S. & Huan, T. DaDIA: Hybridizing Data-Dependent and Data-Independent Acquisition Modes for Generating High-Quality Metabolomic Data. *Anal. Chem.* **93**, 2669–2677 (2021).

288. Masselon, C. *et al.* Accurate Mass Multiplexed Tandem Mass Spectrometry for High-Throughput Polypeptide Identification from Mixtures. *Anal. Chem.* **72**, 1918–1924 (2000).

289. Bilbao, A. *et al.* Processing strategies and software solutions for data-independent acquisition in mass spectrometry. *Proteomics* **15**, 964–980 (2015).

290. Panchaud, A., Jung, S., Shaffer, S. A., Aitchison, J. D. & Goodlett, D. R. Faster, Quantitative, and Accurate Precursor Acquisition Independent From Ion Count. *Anal. Chem.* **83**, 2250–2257 (2011).

291. Demichev, V. *et al.* dia-PASEF data analysis using FragPipe and DIA-NN for deep proteomics of low sample amounts. *Nat Commun* **13**, 3944 (2022).

292. Huang, Z. *et al.* Proteomic datasets of HeLa and SiHa cell lines acquired by DDA-PASEF and diaPASEF. *Data in Brief* **41**, 107919 (2022).

293. Arenas-De Larriva, M. del S., Fernández-Vega, A., Jurado-Gamez, B. & Ortea, I. diaPASEF Proteomics and Feature Selection for the Description of Sputum Proteome Profiles in a Cohort of Different Subtypes of Lung Cancer Patients and Controls. *IJMS* **23**, 8737 (2022).

294. Mun, D.-G. *et al.* DIA-Based Proteome Profiling of Nasopharyngeal Swabs from COVID-19 Patients. *J. Proteome Res.* **20**, 4165–4175 (2021).

295. Skowronek, P. *et al.* Synchro-PASEF Allows Precursor-Specific Fragment Ion Extraction and Interference Removal in Data-Independent Acquisition. *Molecular & Cellular Proteomics* **22**, 100489 (2023).

296. Szyrwiel, L., Sinn, L., Ralser, M. & Demichev, V. Slice-PASEF: fragmenting all ions for maximum sensitivity in proteomics. *bioRxiv* (2022) doi:10.1101/2022.10.31.514544.

297. Distler, U. *et al.* midiaPASEF maximizes information content in data-independent acquisition proteomics. *bioRxiv* (2023) doi:10.1101/2023.01.30.526204.

298. Ting, Y. S. *et al.* Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. *Molecular & Cellular Proteomics* **14**, 2301–2307 (2015).

299. Ludwig, C. *et al.* Data-independent acquisition-based SWATH -MS for quantitative proteomics: a tutorial. *Mol Syst Biol* **14**, (2018).

300. Schubert, O. T. *et al.* Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nat Protoc* **10**, 426–441 (2015).

301. Penny, J., Arefian, M., Schroeder, G. N., Bengoechea, J. A. & Collins, B. C. A gas phase fractionation acquisition scheme integrating ion mobility for rapid diaPASEF library generation. *Proteomics* **23**, 2200038 (2023).

302. Lam, H. *et al.* Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**, 655–667 (2007).

303. Pino, L. K. *et al.* The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics. *Mass Spec Rev* **39**, 229–244 (2020).

304. Zhao, M. *et al.* Evaluation of Urinary Proteome Library Generation Methods on Data-Independent Acquisition MS Analysis and its Application in Normal Urinary Proteome Analysis. *Prot. Clin. Appl.* **13**, 1800152 (2019).

305. Desiere, F. The PeptideAtlas project. *Nucleic Acids Research* **34**, D655–D658 (2006).

306. Deutsch, E. W., Lam, H. & Aebersold, R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep* **9**, 429–434 (2008).

307. Wang, M. Proteomics data reuse with MassIVE-KB. *Cell Systems* (2018).

308. Martens, L. *et al.* PRIDE: The proteomics identifications database. *Proteomics* **5**, 3537–3545 (2005).

309. Rosenberger, G. *et al.* A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci Data* **1**, 140031 (2014).

310. Bruderer, R. *et al.* Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. *Molecular & Cellular Proteomics* **16**, 2296–2309 (2017).

311. Ammar, C. *et al.* Multi-Reference Spectral Library Yields Almost Complete Coverage of Heterogeneous LC-MS/MS Data Sets. *J. Proteome Res.* **18**, 1553–1566 (2019).

312. Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat Methods* **16**, 509–518 (2019).

313. Schmidt, T. *et al.* ProteomicsDB. *Nucleic Acids Research* **46**, D1271–D1281 (2018).

314. Tiwary, S. *et al.* High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nat Methods* **16**, 519–525 (2019).

315. Zhou, X.-X. *et al.* pDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. *Anal. Chem.* **89**, 12690–12697 (2017).

316. Yang, Y. *et al.* In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nat Commun* **11**, 146 (2020).

317. Meyer, J. G. Deep learning neural network tools for proteomics. *Cell Reports Methods* **1**, 100003 (2021).

318. Fröhlich, K. *et al.* Benchmarking of analysis strategies for data-independent acquisition proteomics using a large-scale dataset comprising inter-patient heterogeneity. *Nat Commun* **13**, 2622 (2022).

319. Gotti, C. *et al.* Extensive and Accurate Benchmarking of DIA Acquisition Methods and Software Tools Using a Complex Proteomic Standard. *J. Proteome Res.* **20**, 4801–4814 (2021).

320. Lou, R. *et al.* Benchmarking commonly used software suites and analysis workflows for DIA proteomics and phosphoproteomics. *Nat Commun* **14**, 94 (2023).

321. Cox, J. Prediction of peptide mass spectral libraries with machine learning. *Nature Biotechnology* **41**, 33–43 (2022).

322. Lou, R. *et al.* DeepPhospho accelerates DIA phosphoproteome profiling through in silico library generation. *Nat Commun* **12**, 6685 (2021).

323. Röst, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat Biotechnol* **32**, 219–223 (2014).

324. Röst, H. L. *et al.* TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. *Nat Methods* **13**, 777–783 (2016).

325. Gupta, S., Ahadi, S., Zhou, W. & Röst, H. DIAlignR Provides Precise Retention Time Alignment Across Distant Runs in DIA and Targeted Proteomics. *Molecular & Cellular Proteomics* **18**, 806–817 (2019).

326. Liu, Y. *et al.* DeepRTAlign: toward accurate retention time alignment for large cohort mass spectrometry data analysis. *bioRxiv* (2022).

327. Zhang, N. *et al.* ProbIDtree: An automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics* **5**, 4096–4106 (2005).

328. Wang, J. *et al.* MSPLIT-DIA: sensitive peptide identification for data-independent acquisition. *Nat Methods* **12**, 1106–1108 (2015).

329. Searle, B. C. *et al.* Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nat Commun* **9**, 5128 (2018).

330. Ting, Y. S. *et al.* PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nat Methods* **14**, 903–908 (2017).

331. Tsou, C.-C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods* **12**, 258–264 (2015).

332. Sinitcyn, P. *et al.* MaxDIA enables library-based and library-free data-independent acquisition proteomics. *Nat Biotechnol* **39**, 1563–1573 (2021).

333. Searle, B. C., Lawrence, R. T., MacCoss, M. J. & Villén, J. Thesaurus: quantifying phosphopeptide positional isomers. *Nat Methods* **16**, 703–706 (2019).

334. Bersching, K., Michna, T., Tenzer, S. & Jacob, S. Data-Independent Acquisition (DIA) Is Superior for High Precision Phospho-Peptide Quantification in Magnaporthe oryzae. *Journal of Fungi* **9**, 63 (2022).

335. Rosenberger, G. *et al.* Inference and quantification of peptidoforms in large sample cohorts by SWATH-MS. *Nat Biotechnol* **35**, 781–788 (2017).

336. Kitata, R. B. *et al.* A data-independent acquisition-based global phosphoproteomics system enables deep profiling. *Nat Commun* **12**, 2539 (2021).

337. Lanznaster, D. *et al.* Metabolomics: A Tool to Understand the Impact of Genetic Mutations in Amyotrophic Lateral Sclerosis. *Genes* **11**, 537 (2020).

338. Barschke, P., Oeckl, P., Steinacker, P., Ludolph, A. & Otto, M. Proteomic studies in the discovery of cerebrospinal fluid biomarkers for amyotrophic lateral sclerosis. *Expert Review of Proteomics* **14**, 769–777 (2017).

339. Krassowski, M., Das, V., Sahu, S. K. & Misra, B. B. State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. *Front. Genet.* **11**, 610798 (2020).

340. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol* **18**, 83 (2017).

341. Clark, C., Rabl, M., Dayon, L. & Popp, J. The promise of multi-omics approaches to discover biological alterations with clinical relevance in Alzheimer's disease. *Front. Aging Neurosci.* **14**, 1065904 (2022).

342. Douglas, G. M. *et al.* Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease. *Microbiome* **6**, 13 (2018).

343. Sathyanarayanan, A. *et al.* Multi-omics data integration methods and their applications in psychiatric disorders. *European Neuropsychopharmacology* **69**, 26–46 (2023).

344. Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials, Board on Health Care Services, Board on Health Sciences Policy, & Institute of Medicine. *Evolution of Translational Omics: Lessons Learned and the Path Forward*. (National Academies Press, 2012).

345. Li, C., Sullivan, R. E., Zhu, D. & Hicks, S. D. Putting the "mi" in omics: discovering miRNA biomarkers for pediatric precision care. *Pediatric Research* **93**, 316–323 (2023).

346. Lal, C. V., Bhandari, V. & Ambalavanan, N. Genomics, microbiomics, proteomics, and metabolomics in bronchopulmonary dysplasia. *Seminars in Perinatology* **42**, 425–431 (2018).

347. Schubert, O. T., Röst, H. L., Collins, B. C., Rosenberger, G. & Aebersold, R. Quantitative proteomics: challenges and opportunities in basic and applied research. *Nat Protoc* **12**, 1289–1294 (2017).

348. Misra, B. B., Langefeld, C., Olivier, M. & Cox, L. A. Integrated omics: tools, advances and future approaches. *Journal of Molecular Endocrinology* **62**, R21–R45 (2019).

349. Tarazona, S., Arzalluz-Luque, A. & Conesa, A. Undisclosed, unmet and neglected challenges in multi-omics studies. *Nat Comput Sci* **1**, 395–402 (2021).

350. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform Biol Insights* **14**, 117793221989905 (2020).

351. Palsson, B. & Zengler, K. The challenges of integrating multi-omic data sets. *Nat Chem Biol* **6**, 787–789 (2010).

352. Kopczynski, D. *et al.* Multi-OMICS: a critical technical perspective on integrative lipidomics approaches. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **1862**, 808–811 (2017).

353. Tarazona, S. *et al.* Harmonization of quality metrics and power calculation in multi-omic studies. *Nat Commun* **11**, 3092 (2020).

354. Vuckovic, D. Current trends and challenges in sample preparation for global metabolomics using liquid chromatography–mass spectrometry. *Anal Bioanal Chem* **403**, 1523–1548 (2012).

355. Nakayasu, E. S. *et al.* MPLEx: a Robust and Universal Protocol for Single-Sample Integrative Proteomic, Metabolomic, and Lipidomic Analyses. *mSystems* **1**, e00043-16 (2016).

356. Coman, C. *et al.* Simultaneous Metabolite, Protein, Lipid Extraction (SIMPLEX): A Combinatorial Multimolecular Omics Approach for Systems Biology. *Molecular & Cellular Proteomics* **15**, 1435–1466 (2016).

357. Valledor, L. *et al.* A universal protocol for the combined isolation of metabolites, DNA, long RNAs, small RNAs, and proteins from plants and microorganisms. *Plant J* **79**, 173–180 (2014).

358. Quinn, R. A. *et al.* From Sample to Multi-Omics Conclusions in under 48 Hours. *mSystems* **1**, e00038-16 (2016).

359. Muehlbauer, L. K. *et al.* Rapid Multi-Omics Sample Preparation for Mass Spectrometry. *Anal. Chem.* **95**, 659–667 (2023).

360. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).

361. Berrios, D. C., Beheshti, A. & Costes, S. V. FAIRness and Usability for Open-access Omics Data Systems. *AMIA Annu Symp Proc.* (2018).

362. Perez-Riverol, Y. *et al.* Discovering and linking public omics data sets using the Omics Discovery Index. *Nat Biotechnol* **35**, 406–409 (2017).

363. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).

364. METABRIC Group *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).

365. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* **11**, 333–337 (2014).

366. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol* **13**, e1005752 (2017).

367. Argelaguet, R. *et al.* Multi-Omics Factor Analysis : a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* **14**, (2018).

368. Ma, A., McDermaid, A., Xu, J., Chang, Y. & Ma, Q. Integrative Methods and Practical Challenges for Single-Cell Multi-omics. *Trends in Biotechnology* **38**, 1007–1022 (2020).

369. Reel, P. S., Reel, S., Pearson, E., Trucco, E. & Jefferson, E. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology Advances* **49**, 107739 (2021).

370. Feldner-Busztin, D. *et al.* Dealing with dimensionality: the application of machine learning to multi-omics data. *Bioinformatics* **39**, btad021 (2023).

371. McClatchy, D. B. *et al.* Global quantitative analysis of phosphorylation underlying phencyclidine signaling and sensorimotor gating in the prefrontal cortex. *Mol Psychiatry* **21**, 205–215 (2016).

372. Zecha, J. *et al.* TMT Labeling for the Masses: A Robust and Cost-efficient, In-solution Labeling Approach. *Molecular & Cellular Proteomics* **18**, 1468–1478 (2019).

373. Fang, B. *et al.* Lowering Sample Requirements to Study Tyrosine Kinase Signaling Using Phosphoproteomics with the TMT Calibrator Approach. *Proteomics* **20**, 2000116 (2020).

374. Iliuk, A. Identification of Phosphorylated Proteins on a Global Scale. *Current Protocols in Chemical Biology* **10**, (2018).

375. Keshishian, H. *et al.* A highly multiplexed quantitative phosphosite assay for biology and preclinical studies. *Molecular Systems Biology* **17**, e10156 (2021).

376. Colomé, N. *et al.* Multi-laboratory experiment PME11 for the standardization of phosphoproteome analysis. *Journal of Proteomics* **251**, 104409 (2022).

377. Yue, X., Schunter, A. & Hummon, A. B. Comparing Multistep Immobilized Metal Affinity Chromatography and Multistep TiO_2 Methods for Phosphopeptide Enrichment. *Anal. Chem.* **87**, 8837–8844 (2015).

378. Salovska, B., Tichy, A., Rezacova, M., Vavrova, J. & Novotna, E. Enrichment strategies for phosphoproteomics: state-of-the-art. *Reviews in Analytical Chemistry* **31**, (2012).

379. Huttlin, E. L. *et al.* A Tissue-Specific Atlas of Mouse Protein Phosphorylation and Expression. *Cell* **143**, 1174–1189 (2010).

380. Kettenbach, A. N. & Gerber, S. A. Rapid and Reproducible Single-Stage Phosphopeptide Enrichment of Complex Peptide Mixtures: Application to General and Phosphotyrosine-Specific Phosphoproteomics Experiments. *Anal. Chem.* **83**, 7635–7644 (2011).

381. Wiesner, J., Premsler, T. & Sickmann, A. Application of electron transfer dissociation (ETD) for the analysis of posttranslational modifications. *Proteomics* **8**, 4466–4483 (2008).

382. Weng, S. H. S. *et al.* Improved Phosphoproteomics Workflow with Automation Platform on KingFisher Flex and Data-Independent Acquisition (DIA) Analysis.

383. Brun, C. Développement de stratégies analytiques quantitatives pour l'étude des protéines, de leurs phosphorylations et glycations. (Université de Strasbourg, 2022).

384. Ferries, S. *et al.* Evaluation of Parameters for Confident Phosphorylation Site Localization Using an Orbitrap Fusion Tribrid Mass Spectrometer. *J. Proteome Res.* **16**, 3448–3459 (2017).

385. Kuharev, J., Navarro, P., Distler, U., Jahn, O. & Tenzer, S. In-depth evaluation of software tools for data-independent acquisition based label-free quantification. *Proteomics* **15**, 3140–3151 (2015).

386. An Staes *et al.* Benchmarking DIA data analysis workflows. *bioRxiv* 2023.06.02.543441 (2023).

387. Wen, C. *et al.* Evaluation of DDA Library-Free Strategies for Phosphoproteomics and Ubiquitinomics Data-Independent Acquisition Data. *J. Proteome Res.* acs.jproteome.2c00735 (2023).

388. Lou, R. *et al.* Benchmarking commonly used software suites and analysis workflows for DIA proteomics and phosphoproteomics. *Nat Commun* **14**, 94 (2023).

389. Vashist, T. D. *et al.* DIA Phosphoproteomics: Comparative Evaluation of Dynamic Range and Quantitative Linearity Across Multiple MS Platforms. (2023).

390. Ryan, M., Heverin, M., McLaughlin, R. L. & Hardiman, O. Lifetime Risk and Heritability of Amyotrophic Lateral Sclerosis. *JAMA Neurol* **76**, 1367 (2019).

391. Paganoni, S. *et al.* Diagnostic timelines and delays in diagnosing amyotrophic lateral sclerosis (ALS). *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* **15**, 453–456 (2014).

392. Oeckl, P. *et al.* Proteomics in cerebrospinal fluid and spinal cord suggests UCHL1, MAP2 and GPNMB as biomarkers and underpins importance of transcriptional pathways in amyotrophic lateral sclerosis. *Acta Neuropathol* **139**, 119–134 (2020).

393. Zoccolella, S. *et al.* Riluzole and amyotrophic lateral sclerosis survival: a population-based study in southern Italy: Riluzole and ALS survival in Puglia. *European Journal of Neurology* **14**, 262–268 (2007).

394. Breiner, A., Zinman, L. & Bourque, P. R. Edaravone for amyotrophic lateral sclerosis: barriers to access and lifeboat ethics. *CMAJ* **192**, E319–E320 (2020).

395. Paganoni, S. *et al.* Trial of Sodium Phenylbutyrate–Taurursodiol for Amyotrophic Lateral Sclerosis. *N Engl J Med* **383**, 919–930 (2020).

396. Tam, O. H. *et al.* Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia. *Cell Reports* **29**, 1164-1177.e5 (2019).

397. Figueroa-Romero, C. *et al.* Expression of microRNAs in human post-mortem amyotrophic lateral sclerosis spinal cords provides insight into disease mechanisms. *Molecular and Cellular Neuroscience* **71**, 34–45 (2016).

398. Umoh, M. E. *et al.* A proteomic network approach across the ALS-FTD disease spectrum resolves clinical phenotypes and genetic vulnerability in human brain. *EMBO Mol Med* **10**, 48–62 (2018).

399. Lutz, C. Mouse models of ALS: Past, present and future. *Brain Research* **1693**, 1–10 (2018).

400. Li, P. & Bartlett, M. G. A review of sample preparation methods for quantitation of small-molecule analytes in brain tissue by liquid chromatography tandem mass spectrometry (LC-MS/MS). *Anal. Methods* **6**, 6183–6207 (2014).

401. Karpiński, A. A. *et al.* Study on Tissue Homogenization Buffer Composition for Brain Mass Spectrometry-Based Proteomics. *Biomedicines* **10**, 2466 (2022).

402. Li, K. W., Ganz, A. B. & Smit, A. B. Proteomics of neurodegenerative diseases: analysis of human post-mortem brain. *J. Neurochem.* **151**, 435–445 (2019).

403. Shevchenko, G., Musunuri, S., Wetterhall, M. & Bergquist, J. Comparison of Extraction Methods for the Comprehensive Analysis of Mouse Brain Proteome using Shotgun-based Mass Spectrometry. *J. Proteome Res.* **11**, 2441–2451 (2012).

404. Ericsson, C., Peredo, I. & Nistér, M. Optimized protein extraction from cryopreserved brain tissue samples. *Acta Oncologica* **46**, 10–20 (2007).

405. Santiago, J. A., Quinn, J. P. & Potashkin, J. A. Network Analysis Identifies Sex-Specific Gene Expression Changes in Blood of Amyotrophic Lateral Sclerosis Patients. *IJMS* **22**, 7150 (2021).

406. Murdock, B. J., Goutman, S. A., Boss, J., Kim, S. & Feldman, E. L. Amyotrophic Lateral Sclerosis Survival Associates With Neutrophils in a Sex-specific Manner. *Neurol Neuroimmunol Neuroinflamm* **8**, e953 (2021).

407. Günther, R. *et al.* The rho kinase inhibitor Y-27632 improves motor performance in male SOD1G93A mice. *Front. Neurosci.* **8**, (2014).

408. Torres, P. *et al.* Gender-Specific Beneficial Effects of Docosahexaenoic Acid Dietary Supplementation in G93A-SOD1 Amyotrophic Lateral Sclerosis Mice. *Neurotherapeutics* **17**, 269–281 (2020).

409. Wieczorek, S. *et al.* DAPAR & ProStaR: software to perform statistical analyses in quantitative discovery proteomics. *Bioinformatics* **33**, 135–136 (2017).

410. Kong, W., Hui, H. W. H., Peng, H. & Goh, W. W. B. Dealing with missing values in proteomics data. *Proteomics* **22**, 2200092 (2022).

411. Liu, M. & Dongre, A. Proper imputation of missing values in proteomics datasets for differential expression analysis. *Briefings in Bioinformatics* **22**, bbaa112 (2021).

412. Lazar, C., Gatto, L., Ferro, M., Bruley, C. & Burger, T. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *J. Proteome Res.* **15**, 1116–1125 (2016).

413. Salem, A. *et al.* Matrin3: Disorder and ALS Pathogenesis. *Front. Mol. Biosci.* **8**, 794646 (2022).

414. Diquigiovanni, C. *et al.* A novel mutation in SPART gene causes a severe neurodevelopmental delay due to mitochondrial dysfunction with complex I impairments and altered pyruvate metabolism. *FASEB j.* **33**, 11284–11302 (2019).

415. Pedersen, C. C. *et al.* A systematic review of associations between common SNCA variants and clinical heterogeneity in Parkinson's disease. *npj Parkinsons Dis.* **7**, 54 (2021).

416. Hirano, M. *et al.* Mutations in the gene encoding p62 in Japanese patients with amyotrophic lateral sclerosis. *Neurology* **80**, 458–463 (2013).

417. Mitsui, S. *et al.* Systemic overexpression of SQSTM1/p62 accelerates disease onset in a SOD1H46R-expressing ALS mouse model. *Mol Brain* **11**, 30 (2018).

418. Wobst, H. J., Mack, K. L., Brown, D. G., Brandon, N. J. & Shorter, J. The clinical trial landscape in amyotrophic lateral sclerosis—Past, present, and future. *Med Res Rev* **40**, 1352–1384 (2020).

419. Nakamura, T., Myint, K. T. & Oda, Y. Ethylenediaminetetraacetic Acid Increases Identification Rate of Phosphoproteomics in Real Biological Samples. *J. Proteome Res.* **9**, 1385–1391 (2010).

420. Siino, V. *et al.* Impact of diet-induced obesity on the mouse brain phosphoproteome. *The Journal of Nutritional Biochemistry* **58**, 102–109 (2018).

421. Lachén-Montes, M., Gonzales-Morales, A., Fernandez-Irigoyen, J. & Santamaría, E. Determination of Cerebrospinal Fluid Proteome Variations by Isobaric Labeling Coupled with Strong Cation-Exchange Chromatography and Tandem Mass Spectrometry. in *Cerebrospinal Fluid (CSF) Proteomics: Methods and Protocols* 155–168 (Springer New York, 2019).

422. Collins, M. A., An, J., Hood, B. L., Conrads, T. P. & Bowser, R. P. Label-Free LC–MS/MS Proteomic Analysis of Cerebrospinal Fluid Identifies Protein/Pathway Alterations and Candidate Biomarkers for Amyotrophic Lateral Sclerosis. *J. Proteome Res.* **14**, 4486–4501 (2015).

423. Jankovska, E., Svitek, M., Holada, K. & Petrak, J. Affinity depletion versus relative protein enrichment: a side-by-side comparison of two major strategies for increasing human cerebrospinal fluid proteome coverage. *Clin Proteom* **16**, 9 (2019).

424. Yoshihara, T. *et al.* Cerebrospinal Fluid Protein Concentration in Healthy Older Japanese Volunteers. *IJERPH* **18**, 8683 (2021).

425. Carlyle, B., Trombetta, B. & Arnold, S. Proteomic Approaches for the Discovery of Biofluid Biomarkers of Neurodegenerative Dementias. *Proteomes* **6**, 32 (2018).

426. Macron, C., Núñez Galindo, A., Gahoi, N., Cominetti, O. & Dayon, L. A Versatile Workflow for Cerebrospinal Fluid Proteomic Analysis with Mass Spectrometry: A Matter of Choice between Deep Coverage and Sample Throughput. in *Cerebrospinal Fluid (CSF) Proteomics: Methods and Protocols* 129–154 (2019).

427. Macron, C., Lane, L., Núñez Galindo, A. & Dayon, L. Deep Dive on the Proteome of Human Cerebrospinal Fluid: A Valuable Data Resource for Biomarker Discovery and Missing Protein Identification. *J. Proteome Res.* **17**, 4113–4126 (2018).

428. Barkovits, K. *et al.* Characterization of Cerebrospinal Fluid via Data-Independent Acquisition Mass Spectrometry. *J. Proteome Res.* **17**, 3418–3430 (2018).

429. McKetney, J. *et al.* Pilot proteomic analysis of cerebrospinal fluid in Alzheimer's disease. *Proteomics Clinical Apps* **15**, 2000072 (2021).

430. Rao, A. A., Mehta, K., Gahoi, N. & Srivastava, S. Application of 2D-DIGE and iTRAQ Workflows to Analyze CSF in Gliomas. in *Cerebrospinal Fluid (CSF) Proteomics: Methods and Protocols* 81–110 (2019).

431. Birke, R., Krause, E., Schümann, M., Blasig, I. E. & Haseloff, R. F. Quantitative Evaluation of Different Protein Fractions of Cerebrospinal Fluid Using 18O Labeling. in *Cerebrospinal Fluid (CSF) Proteomics: Methods and Protocols* 119–128 (2019).

432. Barkovits, K., Tönges, L. & Marcus, K. CSF Sample Preparation for Data-Independent Acquisition. in *Cerebrospinal Fluid (CSF) Proteomics: Methods and Protocols* 61–67 (2019).

433. Barkovits *et al.* Blood Contamination in CSF and Its Impact on Quantitative Analysis of Alpha-Synuclein. *Cells* **9**, 370 (2020).

434. Hörmann, P., Barkovits, K., Marcus, K. & Hiller, K. Co-extraction for Metabolomics and Proteomics from a Single CSF Sample. in *Cerebrospinal Fluid (CSF) Proteomics: Methods and Protocols* (eds. Santamaría, E. & Fernández-Irigoyen, J.) 337–342 (Springer New York, 2019).

435. Yang, H. *et al.* Identification of cerebrospinal fluid metabolites as biomarkers for neurobrucellosis by liquid chromatography-mass spectrometry approach. *Bioengineered* **13**, 6996–7010 (2022).

436. Karayel, O. *et al.* Proteome profiling of cerebrospinal fluid reveals biomarker candidates for Parkinson's disease. *Cell Reports Medicine* **3**, 100661 (2022).

437. Bahl, J. M. C., Jensen, S. S., Larsen, M. R. & Heegaard, N. H. H. Characterization of the Human Cerebrospinal Fluid Phosphoproteome by Titanium Dioxide Affinity Chromatography and Mass Spectrometry. *Anal. Chem.* **80**, 6308–6316 (2008).

438. Sun, J. *et al.* Profiling phosphoproteome landscape in circulating extracellular vesicles from microliters of biofluids through functionally tunable paramagnetic separation. *Angewandte Chemie* e202305668 (2023).

439. Costa, J. *et al.* Cerebrospinal Fluid Chitinases as Biomarkers for Amyotrophic Lateral Sclerosis. *Diagnostics* **11**, 1210 (2021).

440. Thompson, A. G. *et al.* CSF chitinase proteins in amyotrophic lateral sclerosis. *J Neurol Neurosurg Psychiatry* **90**, 1215–1220 (2019).

441. Oldoni, E. *et al.* CHIT1 at Diagnosis Reflects Long-Term Multiple Sclerosis Disease Activity. *Ann Neurol* **87**, 633–645 (2020).

442. Schneider, R. *et al.* Chitinase 3–like 1 and neurofilament light chain in CSF and CNS atrophy in MS. *Neurol Neuroimmunol Neuroinflamm* **8**, e906 (2021).

443. Raffaele, S., Boccazzi, M. & Fumagalli, M. Oligodendrocyte Dysfunction in Amyotrophic Lateral Sclerosis: Mechanisms and Therapeutic Perspectives. *Cells* **10**, 565 (2021).

444. Lee, S. & Kim, H.-J. Prion-like Mechanism in Amyotrophic Lateral Sclerosis: are Protein Aggregates the Key? *Exp Neurobiol* **24**, 1–7 (2015).

445. McAlary, L. *et al.* Amyotrophic Lateral Sclerosis: Proteins, Proteostasis, Prions, and Promises. *Front. Cell. Neurosci.* **14**, 581907 (2020).

446. Tiwari, M. Glucose 6 phosphatase dehydrogenase (G6PD) and neurodegenerative disorders: Mapping diagnostic and therapeutic opportunities. *Genes & Diseases* **4**, 196–203 (2017).

447. Pasetto, L. *et al.* Defective cyclophilin A induces TDP-43 proteinopathy: implications for amyotrophic lateral sclerosis and frontotemporal dementia. *Brain* **144**, 3710–3726 (2021).

448. Pasetto, L. *et al.* Targeting Extracellular Cyclophilin A Reduces Neuroinflammation and Extends Survival in a Mouse Model of Amyotrophic Lateral Sclerosis. *J. Neurosci.* **37**, 1413–1427 (2017).

449. Wojdała, A. L. *et al.* Phosphatidylethanolamine Binding Protein 1 (PEBP1) in Alzheimer's Disease: ELISA Development and Clinical Validation. *JAD* **88**, 1459–1468 (2022).

450. Nilsson, J. *et al.* Cerebrospinal fluid biomarker panel for synaptic dysfunction in Alzheimer's disease. *Alz & Dem Diag Ass & Dis Mo* **13**, (2021).

451. Kaiserova, M. *et al.* Cerebrospinal fluid levels of chromogranin A and phosphorylated neurofilament heavy chain are elevated in amyotrophic lateral sclerosis. *Acta Neurol Scand* **136**, 360–364 (2017).

452. Rudrabhatla, P., Grant, P., Jaffe, H., Strong, M. J. & Pant, H. C. Quantitative phosphoproteomic analysis of neuronal intermediate filament proteins (NF-M/H) in Alzheimer's disease by iTRAQ. *FASEB j.* **24**, 4396–4407 (2010).

453. Quinn, J. P. *et al.* Cerebrospinal Fluid and Brain Proteoforms of the Granin Neuropeptide Family in Alzheimer's Disease. *J. Am. Soc. Mass Spectrom.* **34**, 649–667 (2023).

454. Mousavi, S. V., Agah, E. & Tafakhori, A. The Role of Osteopontin in Amyotrophic Lateral Sclerosis: A Systematic Review. *Arch Neurosci* **7**, (2020).

455. Yamamoto, T., Murayama, S., Takao, M., Isa, T. & Higo, N. Expression of secreted phosphoprotein 1 (osteopontin) in human sensorimotor cortex and spinal cord: Changes in patients with amyotrophic lateral sclerosis. *Brain Research* **1655**, 168–175 (2017).

456. De Luna, N. *et al.* Neuroinflammation-Related Proteins NOD2 and Spp1 Are Abnormally Upregulated in Amyotrophic Lateral Sclerosis. *Neurol Neuroimmunol Neuroinflamm* **10**, e200072 (2023).

457. Verde, F. *et al.* Chromogranin A levels in the cerebrospinal fluid of patients with amyotrophic lateral sclerosis. *Neurobiology of Aging* **67**, 21–22 (2018).

458. Bittremieux, W. *et al.* Quality control in mass spectrometry-based proteomics. *Mass Spec Rev* **37**, 697–711 (2018).

459. Köcher, T., Pichler, P., Swart, R. & Mechtler, K. Quality control in LC-MS/MS. *Proteomics* **11**, 1026–1030 (2011).

460. Rozanova, S. *et al.* Quality Control—A Stepchild in Quantitative Proteomics: A Case Study for the Human CSF Proteome. *Biomolecules* **13**, 491 (2023).

461. Patterson, K. L. *et al.* Establishing Quality Control Procedures for Large-Scale Plasma Proteomics Analyses. *J. Am. Soc. Mass Spectrom.* **34**, 1105–1116 (2023).

462. Stanfill, B. A. *et al.* Quality Control Analysis in Real-time (QC-ART): A Tool for Real-time Quality Control Assessment of Mass Spectrometry-based Proteomics Data. *Molecular & Cellular Proteomics* **17**, 1824–1836 (2018).

463. Olivella, R. *et al.* QCloud2: An Improved Cloud-based Quality-Control System for Mass-Spectrometry-based Proteomics Laboratories. *J. Proteome Res.* **20**, 2010–2013 (2021).

464. Martínez-Bartolomé, S. *et al.* PACOM: A Versatile Tool for Integrating, Filtering, Visualizing, and Comparing Multiple Large Mass Spectrometry Proteomics Data Sets. *J. Proteome Res.* **17**, 1547–1558 (2018).

465. Stratton, K. G. *et al.* pmartR : Quality Control and Statistics for Mass Spectrometry-Based Biological Data. *J. Proteome Res.* **18**, 1418–1425 (2019).

466. Degnan, D. J. *et al.* pmartR 2.0 : A Quality Control, Visualization, and Statistics Pipeline for Multiple Omics Datatypes. *J. Proteome Res.* **22**, 570–576 (2023).

467. Bielow, C., Mastrobuoni, G. & Kempa, S. Proteomics Quality Control: Quality Control Software for MaxQuant Results. *J. Proteome Res.* **15**, 777–787 (2016).

468. Imbert, A. *et al.* ProMetIS, deep phenotyping of mouse models by combined proteomics and metabolomics analysis. *Sci Data* **8**, 311 (2021).

469. Ardito, F., Giuliani, M., Perrone, D., Troiano, G. & Muzio, L. L. The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (Review). *International Journal of Molecular Medicine* **40**, 271–280 (2017).

470. Rainer, J. *et al.* A Modular and Expandable Ecosystem for Metabolomics Data Annotation in R. *Metabolites* **12**, 173 (2022).

471. Kohler, D. *et al.* MSstatsPTM: Statistical relative quantification of post-translational modifications in bottom-up mass spectrometry-based proteomics. *Molecular & Cellular Proteomics* 100477 (2022) doi:10.1016/J.MCPRO.2022.100477.

Appendices

Multiomic ALS signatures highlight sex differences, molecular subclusters and the MAPK pathway as therapeutic target

Lucas Caldi Gomes[1*], Sonja Hänzelmann[2,3,4*], Sergio Oller[2,3], Mojan Parvaz[1], Fabian Hausmann[2,3], Robin Khatri[2,3], Melanie Ebbing[2,3], Constantin Holzapfel[2,3], Laura Pasetto[5], Stefano Fabrizio Columbro[5], Serena Scozzari[5], Marie Gebelin[6], Johanna Knöferle[1], Isabell Cordts[1], Antonia F. Demleitner[1], Laura Tzeplaeff[1], Marcus Deschauer[1], Claudia Dufke[7], Marc Sturm[7], Qihui Zhou[8], Pavol Zelina[9], Emma Sudria-Lopez[9], Tobias B. Haack[7,10], Sebastian Streb[11], Magdalena Kuzma-Kozakiewicz[12], Dieter Edbauer[8,13], R. Jeroen Pasterkamp[9], Endre Laczko[11], Hubert Rehrauer[11], Ralph Schlapbach[11], Christine Carapito[6], Valentina Bonetto[5], Stefan Bonn[2,3 §], Paul Lingor[1,8,13 §,#]

1 Technical University of Munich, School of Medicine, Klinikum rechts der Isar, Department of Neurology; Ismaninger Str. 22, 81675 München, Germany.
2 Institute of Medical Systems Biology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany.
3 Center for Biomedical AI, University Medical Center Hamburg-Eppendorf, Hamburg, Germany.
4 III. Department of Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany.
5 Research Center for ALS, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milan, Italy
6 Laboratoire de Spectrométrie de Masse BioOrganique, IPHC UMR 7178, CNRS, Université de Strasbourg, Infrastructure Nationale de Protéomique ProFI - FR 2048, Strasbourg, France.
7 Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany
8 German Center for Neurodegenerative Diseases (DZNE), Feodor-Lynen-Straße 17, 81377 München, Germany
9 Department of Translational Neuroscience, University Medical Center Utrecht, Utrecht University, Universiteitsweg 100, 3584 CG, Utrecht, The Netherlands
10 Center for Rare Diseases, University of Tübingen, Tübingen, Germany
11 Functional Genomics Center Zurich (FGCZ), ETH Zurich and University of Zurich, Zurich, Switzerland
12 Department of Neurology, Medical University of Warsaw, Warsaw, Poland
13 Munich Cluster for Systems Neurology (SyNergy), Munich, Germany

*,§ These authors contributed equally.

#Corresponding Author: Paul Lingor; email: paul.lingor@tum.de; Department of Neurology, School of Medicine, University Hospital rechts der Isar, Technical University of Munich, Ismaninger Straße 22, 81675 Munich, Germany, Tel.: +498941408257

Introduction

Amyotrophic lateral sclerosis (ALS) is the most frequent motor neuron disease usually leading to paralysis and death within a few years after symptom onset. The vast majority of ALS patients do not have a family history for the disease and are considered sporadic (sALS). Approximately 10% of all ALS patients have a positive family history (fALS) and in approximately half of them, a genetic cause can be identified, most frequently due to variants in *C9orf72*, *SOD1*, *TARDBP,* or *FUS*. However, approximately 10% of all sALS patients also carry disease-causing mutations[1]. To date, sALS lacks effective disease-modifying treatments. Multiple disease mechanisms have been suggested for ALS, but the etiology of sALS in particular remains unclear [2]. An improved understanding of early disease mechanisms could facilitate the identification of diagnostic and prognostic biomarkers as well as the discovery of novel and potentially more efficient therapeutic drug targets.

Although direct analysis of affected nervous system tissue remains the gold standard for the understanding of neuropathology, patient material is only available post-mortem and in limited numbers. This carries the risk of describing disease end-stages, obscuring mechanisms that occur in earlier phases and may therefore represent more auspicious drug targets. Similar to what has been observed in other neurodegenerative disorders, such as Alzheimer's or Parkinson's disease, ALS pathology spreads from the motor cortex to other cortical brain areas over time[3,4]. Numerous studies have analyzed the motor cortex in ALS [5–7]. However, since this area is the most severely affected in ALS, it is likely that it primarily reflects the final stages of the disease. In contrast, the prefrontal cortex (PFC) in Brodmann area 6 typically exhibits only intermediate TDP43 pathology at the time of death [8]. This suggests that analyzing this area could provide insight into earlier disease-mediated alterations in post-mortem tissue.

Previous investigations of ALS brain tissue have primarily focused on individual molecular subsets, including transcripts [5,7,9], miRNAs [10], or proteins [11], suggesting that ALS is a complex and heterogeneous disease. Recent transcriptomic analysis identified potential distinct ALS populations, which were stratified in different subclusters based on gene set enrichment analyses (oxidative stress [ALS-Ox], retrotransposon activation [ALS-TE], and glial dysfunction [ALS-Glia])[7,12]. More recently, studies utilizing induced pluripotent stem cells (iPSC) [13] derived from ALS patients [14,15] and multiomic strategies in circulating biofluids have further deepened the comprehensive understanding of the molecular mechanisms underlying ALS [16].

In this study, we decipher early ALS disease mechanisms by profiling the transcriptome, miRNAome and (phospho-)proteome in the PFC of ALS patients and four mouse models of the disease. We identify strong sex differences and demonstrate that ALS is not a homogeneous disease, but consists of different molecular subtypes, which find their correlation in individual transgenic mouse models of the disease. Multiomic data integration identified MAPK as a target for early therapeutic intervention and we validate this target in models in vitro and in vivo.

Results

Cohort composition and depth of analysis

To obtain a comprehensive overview of the early molecular changes in sALS, we performed a deep multiomic characterization of human postmortem PFC (Brodmann area 6) from 51 neuropathologically confirmed ALS patients (35 males; 16 females) and 50 control subjects (22 males; 28 females) (**Fig. 1a,** Table 1, Supplementary Table 1**)**. On average, we detected 19.641 transcripts, 736 miRNAs (mature miRNAs and hairpin precursors), and 2.344 proteins per sample (Supplementary Fig. 1). We detected the C9orf72 repeat expansion in one ALS patient and another individual carried a previously reported pathogenic variant in NEK1 (c.3107 C>G, p.Ser1036Ter) [17] (Supplementary Fig. 2, Supplementary Table 2). Because two of our ALS cases had a known causative mutation and data on family history for ALS was not available from the brain banks, we will refer to all cases simply as "ALS" for clarity and accuracy. Four transgenic mouse models were analyzed to identify parallels to human ALS. PFC from presymptomatic or early disease stage was collected from C9orf72-, SOD1-, TDP43- or FUS-ALS transgenic mice with an equal distribution of wild-type (wt) and transgenic (tg) litter-mates, and sex (n=20 per model). In mouse tissue, we detected on average 17.020 transcripts, 842 miRNAs (mature miRNAs and hairpins), 2.568 proteins, and 6.755 phosphosites. Overall, sample quality was high (Supplementary Fig. 3).

Transcriptomic stratification of human ALS into four molecular subclusters

We first performed a principal component analysis (PCA) to assess the effects of disease, sex and sample origin on the transcriptome, revealing a moderate separation by condition (Silhouette score: 0.11 for ALS, -0.03 for CTR), but a strong separation by sex (Supplementary Fig. 4). We, therefore, analyzed differentially expressed genes (DEG) separately for males and females. The number of differentially expressed genes (DEGs) was much higher in males (72) compared to females (2) (**Fig. 1b,** Supplementary Table 3). Overrepresentation analysis revealed an upregulation of genes involved in *retinoic acid* and *lipid metabolism* in females, and a downregulation of *extracellular matrix (ECM), collagen,* and *vasculature* in males (Supplementary Fig. 5). KEGG pathway analysis showed a suppression of *complement/cytokine* and *ECM-receptor signaling* and an activation of *oxidative phosphorylation* and *glutamatergic synapse* in males, whereas in females *complement/coagulation* was decreased, and *ribosomal function* and *oxidative phosphorylation* were activated (**Fig. 1d**). Hierarchical clustering for enriched pathways revealed four distinct molecular subgroups: C1-C4. Regulation of *immune response* dichotomized ALS patients best (C1/2 vs. C3/4), whereas second-level arborization was mainly driven by *ECM* (C1 vs. C2), *synaptic function*, *RNA splicing, and protein folding* (C3 vs. C4, **Fig. 1c,** Supplementary Table 4). These clusters are reminiscent of previously proposed subclusters [7]**,** where C1 and C2 show similarities to “Oxidative stress” (ALS-Ox) (and less to “Retrotransposon reactivated” (ALS-TE)), whereas C3 and C4 correspond to the “Glial dysfunction” (ALS-Glia) cluster (Supplementary Fig. 6). To characterize the clusters by similarly regulated RNA networks, we performed a weighted gene co-expression network analysis (WGCNA [18]) resulting in 20 modules (M) (**Fig. 1e, S**upplementary Figs. 8-9, Supplementary Table 5). The module Mturquoise was enriched for *mitochondrial respiration* and upregulated in C1/2, particularly in males and driven by neuronal alterations, suggesting increased oxidative respiration in PFC neurons. The Myellow module was enriched for *synaptic function* and showed a similar regulation in C1/2. In contrast, the Mtan and Mlightcyan modules were enriched for *immune response* and *RNA splicing* were upregulated in modules C3/4 (**Fig. 1e,f**). In summary, molecular subclusters and sex-specific differences drive the heterogeneity in the PFC of ALS patients.

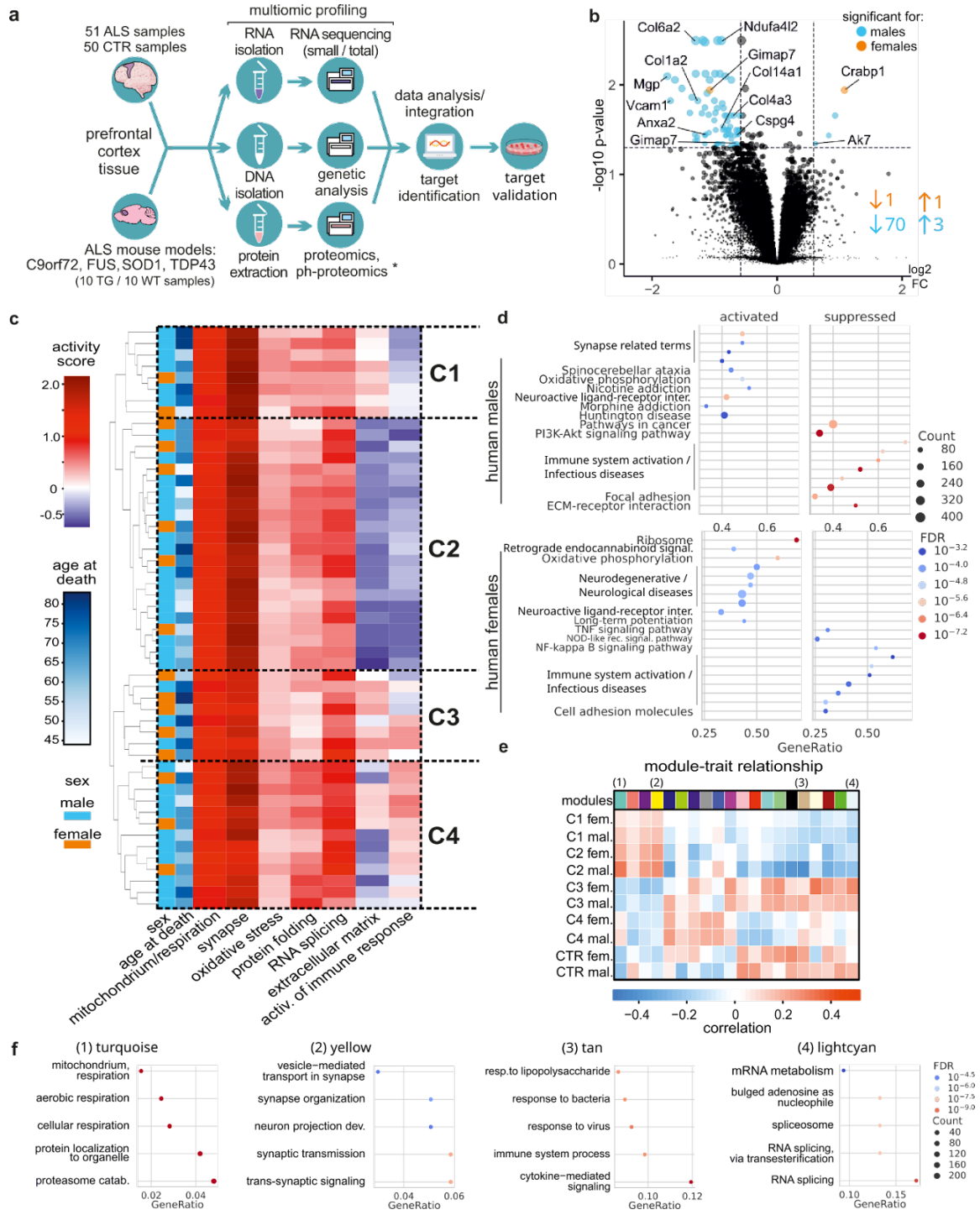


Fig. 1 |Study overview and transcriptomic landscape of ALS-affected brains.

a. Overview of the sample processing workflow. Prefrontal cortex samples were prepared for multiomics experiments from the human cohort (51 ALS/50 CTR samples), as well as from 4 selected ALS mouse models (C9orf72; FUS; SOD1; TDP43 - 10 TG / 10 CTR animals per group). RNA, DNA and protein lysates were prepared for mRNA / small RNA sequencing, targeted DNA sequencing and proteomics / phospho-proteomics (ph-proteomics) experiments, respectively. The latter was performed exclusively for animal tissue.

b. Volcano plot of deregulated genes in human samples. X-axis: log2 fold change (FC); Y-axis: negative log10 padj of each gene. Blue and orange circles indicate significant differential change: left side = decrease (low in ALS), right side = increase (high in ALS).

c. The heatmap depicts activity scores (calculated by Decoupler) of each pathway throughout the entire ALS human cort. Pathways were selected based on their increased occurrence across the different enrichment analyses. Pathways are indicated on the y-axis and the ALS samples on the x-axis (top) along with the metadata,

sex and age at death. Only selected pathways are shown. A full heatmap of all chosen pathways can be found in Supplementary figure 16.

d. Dotplot analysis of enriched KEGG pathways using ClusterProfiler in male and female samples.The dotplot displays significantly enriched pathways in males and females, represented by circles colored by their corresponding adjusted p-values (-log10 transformed) on the x-axis and gene count on the y-axis. The size of each circle corresponds to the number of genes annotated to the KEGG gene set. Similar pathways were summarized. Pathway clusters for human males were composed by the following terms: Retrograde endocannabinoid signaling, synaptic vesicle cycle, long-term potentiation, glutamatergic synapse (for synapse related terms); Staphylococcus aureus infection, Malaria, complement & coagulation cascades, cytokine-cytokine receptor interaction (for immune system activation/infectious diseases), in order of appearance. Pathway clusters for human females were composed by the following terms: Parkinson's disease, Huntington's disease, Spinocerebellar ataxia, pathway of neurodegeneration - multiple diseases, amyotrophic lateral sclerosis (for neurodegenerative/neurological diseases); Complement & coagulation cascades, viral protein interaction with cytokine & cytokine receptor, Staphylococcus aureus infection, Epstein-Barr virus infection, Influenza A, cytokine-cytokine receptor interaction (for immune system activation/infectious diseases), in order of appearance.

e. Heatmap showing Pearson's correlation of WGCNA modules with each sample group. On the y-achsis each cluster of the ALS samples and controls are shown (sub stratified by sex). Representative WGCNA modules were selected for the identified ALS clusters (i.e. turquoise (1) and yellow (2) for clusters C1 and C2; tan (3) and lightcyan (4) for clusters C3 and C4) and explored further by enrichment analyses.

f. Enrichment results (over-representation analysis) for each of the selected WGCNA modules (top 5 terms based on significance).

The transcriptome of murine ALS models reflects human ALS subclusters

The strongest transcriptomic changes were observed in the C9orf72 model, reflecting the fast disease progression present in this mouse model. Each mouse model was predominantly characterized by the deregulation of particular pathways, i.e., *immune/inflammatory response* in the C9orf72 mice, *ERK1/2 cascade, development* and *response to H2O2* in SOD1, *transcription* and *endopeptidase activity* in TDP43. No enrichment was found in the FUS model, which showed the smallest number of DEG (**Fig 2a-e,** Supplementary Fig. 8, Supplementary Table 6).

To assess the degree to which the observed alterations in gene expression reflect changes in the cellular composition of the PFC, we estimated cell type fractions for the subclusters of human ALS and mouse models using deep learning-based deconvolution [19]. Interestingly, human subclusters showed cell fraction changes that are reflected in specific mouse models. The SOD1 model and human C1 & C2 showed a decrease in glial and endothelial cells and a relative increase in excitatory neurons. The C9orf72 model and human C3 are dominated by a strong glial and endothelial cell increase and a decrease of excitatory neurons, suggesting strong neuroinflammation and neuronal loss. FUS and TDP43 models showed intermediate levels of glial and neuronal cell types (**Fig. 2f,** Supplementary Fig. 10). Evidently, transcriptional changes are partially driven by changes in cell composition. Interestingly, our data suggest that the neurovascular unit in ALS[20] may be differently affected in subgroups of ALS patients. Overall, our transcriptomic analyses revealed remarkable similarities between human clusters and mouse models: C1 and C2 showed the best correlation with the SOD1 model, whereas C3 correlated best with the C9orf72 and to a lesser extent with the TDP 43- and FUS-models. Finally, C4 showed a weak correlation with the FUS model (**Fig. 2g**). While we observed transcriptional and cell composition differences between mouse models and human ALS clusters, we found consistent changes in *MAPK signaling* in TDP43, SOD1, and C9orf72 models as well as in human ALS samples. Moreover, analysis of subclusters shows that the classical MAPK pathway is activated in C1 and C2, whereas it is downregulated in C3. On the other hand, C1 & 2 show a downregulation of the JNK and p38 MAPK pathway, which shows an activation in C3 (Supplementary Fig. 11). Mouse models of ALS thus reflect molecular subgroups of human ALS, partially driven by cell fraction changes. Deregulation of MAPK pathways is a common theme in both human ALS and mouse transcriptomes.

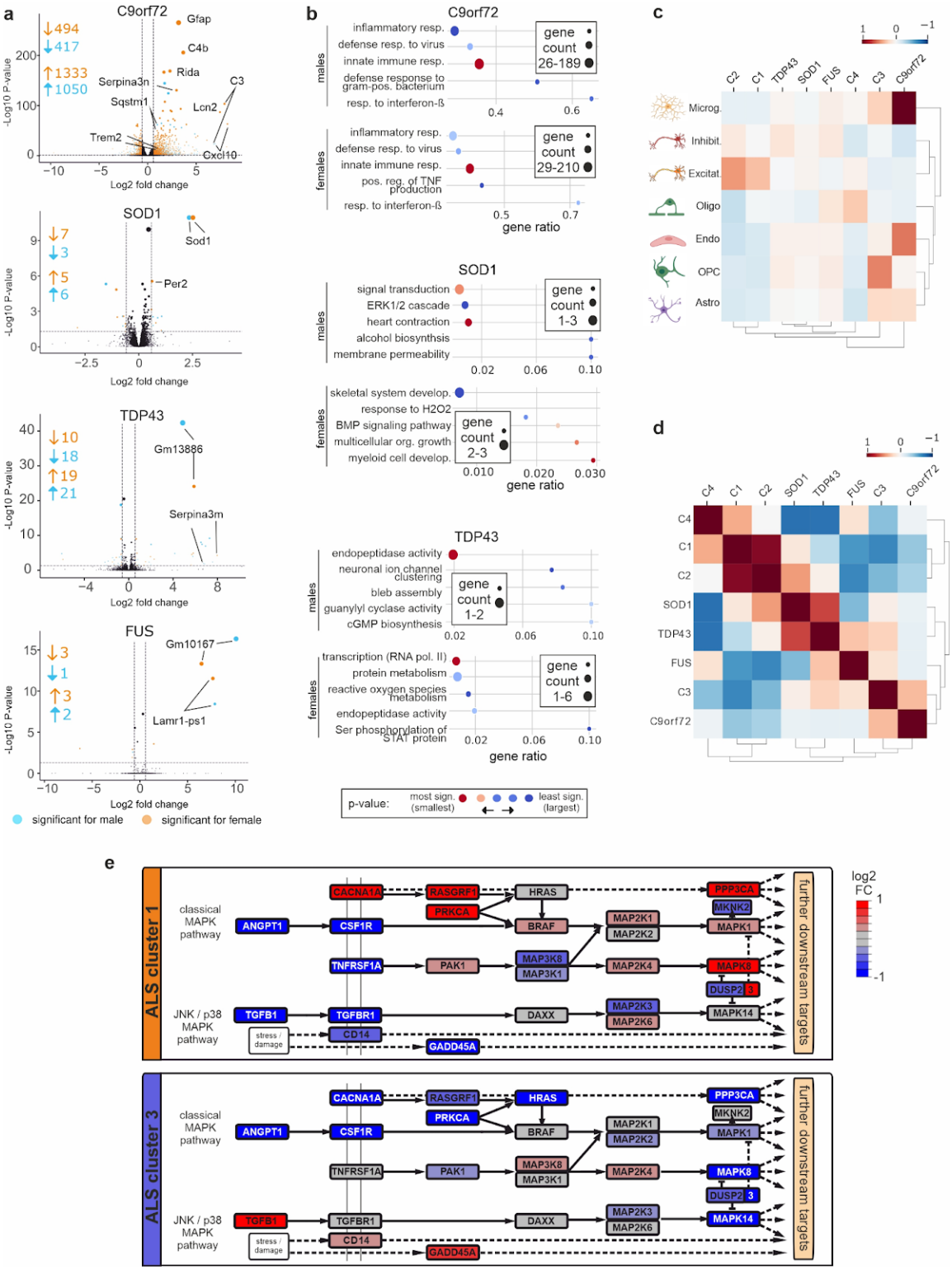


Fig. 2 | Overview of the results for the selected ALS mouse models

a-d Volcano plot of deregulated genes in the four transgenic mouse models, represented by dots colored by their corresponding sex. The x-axis shows the log2 FC between transgenic and wild-type mice, while the y-axis shows the negative log10 transformed adjusted p-values. Orange dots represent DEGs in females and blue dots in males.

e Dotplot analysis of enriched GO-BP using genetonic in male and female samples for ALS mouse models. The dotplot displays enriched pathways in males and females, represented by circles colored by their corresponding adjusted p-values on the x-axis and gene count on the y-axis. The size of each circle corresponds to the number of genes annotated to GO-BP gene sets.

f Heatmap showing summary of changes in cell type proportions across ALS mouse models and human clusters. The indicated changes are the relative difference between median cell type composition of ALS Cluster/TG and median cell type composition in corresponding CTR samples. Comparisons are grouped according to their similarity using hierarchical clustering with Euclidean distance.
g Heatmap showing pairwise Pearson correlation of relative changes in median cell type proportions of ALS mouse models and human clusters. Comparisons are grouped according to their similarity using hierarchical clustering as depicted by the dendrograms.

Differential alternative splicing events represent an early ALS disease mechanism
TARDBP/TDP43 and FUS regulate alternative splicing and transcript usage for hundreds of genes [21,22]. Recently, TDP43 was shown to repress cryptic exon splicing events in UNC13A shedding more light on its role in the ALS disease mechanism. However, these events are specific to neurons with TDP43 pathology and thus too rare to be detectable in bulk RNAseq at our sequencing depth and the sequences are not conserved in mice. Here, we identified significant differential alternative splicing (DAS) events that occur more frequently in male than in female ALS patients (**Fig. 3a**). DAS in males was observed for *CLTB*, *TPRN*, *NRN1*, *CAMK2N1*, and in females for *TPRN* and the gene encoding for *TMEM170A-CFDP1,* a novel readthrough protein (AC009163.5) (Supplementary Table 14). Genes involved in the regulation of *kinase activity* were particularly enriched in the DAS of male ALS patients (SI Table 7). Mouse DAS was more balanced between sexes (**Fig. 3b,** Supplementary Fig. 12) and we identified *FINB, CPLANE1* (involved in ciliogenesis and migration), and *ATP1B1* (a membrane-bound Na+/K+ ATPase) as differentially spliced in male and female models (**Fig. 3c**), showing enrichment in the terms *translation and myelin sheath* (C9orf72), *GTPase activity* and *myelin sheath* (SOD1), *translation* and *heat shock protein binding* (TDP43), and *purine metabolism* and *basal plasma membrane* (FUS) (SI Table 7). Early disease mechanisms thus appear to be influenced by splicing events, particularly in males. TPRN, a stereocilium-associated protein previously only described in nonsyndromic deafness, may have additional roles in the pathogenesis of ALS (**Fig. 3d**).

Downregulation of hairpin and mature miRNA in males characterize early ALS
To address the role of miRNA-mediated regulation in ALS disease mechanisms [23], we sequenced small RNA species. Male ALS patients showed a stronger (down-)regulation of mature miRNAs and miRNA-hairpins than females (**Fig. 3e,f**). The higher number of DE hairpins compared to mature miRNA (Supplementary Fig.1) points to early miRNA biogenesis defects as a potential early ALS disease mechanism. In males we observed 10 DE mature miRNAs (9 down-, 1 up-regulated), but none in females. For miRNA-hairpins, females presented 8 DE species (7 down-, 1 up-regulated), whereas males presented 82 DE hairpins (71 down-, 11 up-regulated). Top hits common for both males and females included the miRNA-hairpins let-7a-3, miR-26a-1, and let-7f-1 (all down-regulated), and miR-1227 (up-regulated) (**Fig. 3e-h**). Enrichment analysis for the targets of DE miRNAs revealed pathways involved in *cell growth*, *focal adhesion, integrin binding* and *kinase activity regulation* pointing to the ECM/neurovascular unit as an important target of miRNA regulation (Supplementary Fig. 13). Mouse models showed the most pronounced DE of miRNAs in the C9orf72 model (targets annotated for *cell survival/PI3K-Akt-, mTOR-, MAPK signaling*), followed by the SOD1 model (*cancer* and *lipid metabolism*) (Supplementary Fig. 13, Supplementary Table 8). Among the top 5 regulated miRNAs, miR-451a was found significantly deregulated across all mouse models and in human ALS patients (**Fig. 3g,h**). Targets of hsa-miR-451a, such as *MAPK1*, *AKT1* and *BCL2*, are known for their role in the regulation of cell growth, survival, and anti-apoptosis while others targets are involved in inflammatory signaling, such as *Il6R*, *IKBKB*, *MIF*, or extracellular matrix remodeling, e.g., *MMP2* and *MMP9* (**Fig. 3i,** Supplementary Fig. 14). In addition to pronounced sex-specific regulation of miRNAs, consistent miRNA changes related to the regulation of the MAPK pathway, ECM, and inflammatory signaling were identified and matched the findings for the transcriptomics dataset, underscoring their role in ALS.

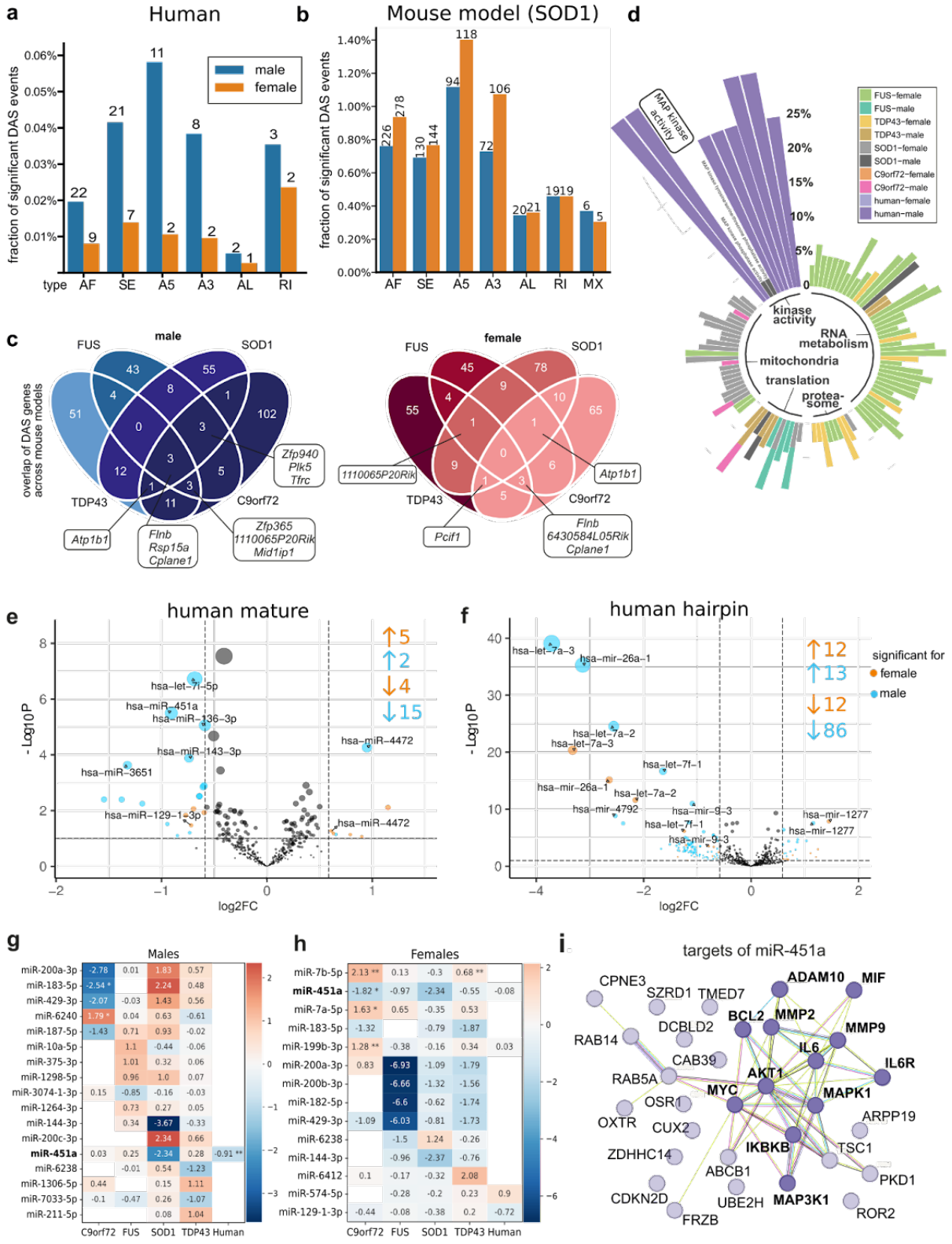


Fig. 3 | Differential alternative splicing and miRNAomic profiling

a-b Differential Alternative Splicing (DAS) analysis. The plot displays the results for human male and female samples for various splice events, i.e. alternative exon (AE), skipped exon (SE), Alternative 5' Splice Site (A5), Alternative 3' Splice Site (A3), Alternative Last Exon (AL) and retained intron (RI) events. Each event is represented by a separate bar, with the height of the bar representing the fraction of significant events in ALS vs.

CTRL. The bars are colored in blue for events with significant differential splicing in males, while orange bars represent events with significant differential splicing in females.
c Venn diagram showing the overlap of differentially alternatively spliced (DAS) genes in four transgenic mouse models (TDP43, FUS, C9orf72, and SOD1) for the ALS vs CTRL comparison. Each circle represents the number of genes with significant differential splicing in each model. The overlapping regions between the circles represent the number of genes with significant differential splicing in more than one model.
d Enrichment of selected themes for DAS genes in four transgenic mouse models and human, in male and female samples, and percentage of genes in each functional term. The plot displays two sets of information: the enrichment of DAS genes in different samples, and the percentage of genes in each functional term. The y-axis represents the percentage of genes in each term, while the x-axis displays the different functional terms. The bars are colored by sex and the different mouse models and human samples.
e - f Volcano plots showing the differential expression of miRNA in male and female human samples for mature miRNA (e) and hairpin miRNA (f). The plots display the log2 fold change (x-axis) and the negative log10 adjusted p-value (y-axis) for each miRNA between the male and female samples. The miRNAs are represented as dots in the plot, and the color of the dots indicates the significance of the differential expression (blue male and orange female).
g-h The heatmaps display the log2 fold change of the top miRNAs in female and male samples of mouse and human datasets. The miRNAs were selected based on the ordered absolute log2 fold change per dataset. Red color indicates upregulation of miRNAs, while blue color indicates downregulation of miRNAs. The statistical significance of differential expression of the miRNAs was inferred using DESeq2, and is indicated by stars. **P<0.01 and *P<0.05 represents statistical significance at a high and moderate level, respectively. Top 5 miRNAs, which are significant differentially expressed in at least two mouse models, were shown.
i Protein-protein interaction (PPI) network of the target genes of miRNA451a. MAPK1, AKT1, IKBKB, MYC, BCL2 and IL6 appear as important molecular hubs.

Human proteomic landscape most closely correlates with the TDP43 model

Next, we assessed proteomic signatures of early ALS using mass spectrometry. In human samples, we detected 49 (30) differentially expressed proteins (DEPs) in males (females), Annexin A2 (ANXA2) being the only protein down-regulated in both sexes (pvalue<0.1). Interestingly, we identified several neurodegeneration-related proteins, such as MATR3, SPART and SCNA (involved in genetic forms of ALS, spastic paraplegia-causing, and Parkinson's disease, respectively) (**Fig. 4a**, Supplementary Table 9). The projection of transcriptomic clusters onto the proteomic data did not reproduce the subclustering, likely because of the much smaller number of mapped entities (Supplementary Fig. 16). Functional enrichment and unsupervised clustering identified pathways with relevance in both sexes, such as *synaptic function*, *immune response*, and *ECM/cytoskeleton*. In contrast, *transmembrane transport, lipid metabolism*, *development*, and *catalytic activity* were enriched in females, whereas *cell metabolism* was captured for males (**Fig. 4b,c,** Supplementary Table 10). In mice, proteomic analyses revealed the strongest changes in the C9orf72 model, where sequestosome 1/p62, the product of the ALS-causing *SQSTM1* gene [24], showed the strongest upregulation pointing to reduced autophagic flux in this dipeptide accumulation model[25] (Supplementary Fig. 17-18). pSQSTM1 was also significantly up-regulated in the phospho-proteomic analysis (Supplementary Fig. 19). SOD1 males showed one up-regulated DEP, Exportin-1 (XPO1), which is a major regulator of nuclear RNA export. This corresponds to increased *XPO1* gene expression in this model and was also observed in TDP43 females, FUS-males and both sexes in the C9orf72 model (Supplementary Table 3). To compare GO term enrichments in human and animal proteomics, we used clustering in semantic space[26]. Human proteomics results showed a clustering for the mechanisms *differentiation and development* (females: clusters 3, 4, 7, 11; males: cluster 8), *synapse* and *immune* / *defense response* (**Fig. 4e,f**). In contrast, the enrichment in the C9orf72 model was shaped by clusters for *RNA processing, ribosome, translation*, *ATP synthesis*, *development*, *cell adhesion*, *transport*, and *synapse*. Remarkably, the SOD1 model showed a strong clustering for *ATP synthesis, mitochondrial respiration*, *translation*, and *vesicle-mediated transport*. Enrichment analysis thus underscored the pathways previously identified in our RNA sequencing data. Overall, the human proteome showed the strongest similarities with the TDP43 model. Several mouse models (especially males) showed increased XPO1 expression, a protein that is therapeutically targeted with the inhibitor BIIB100 in ALS[27] (**Fig. 4d**, Supplementary Fig. 20).

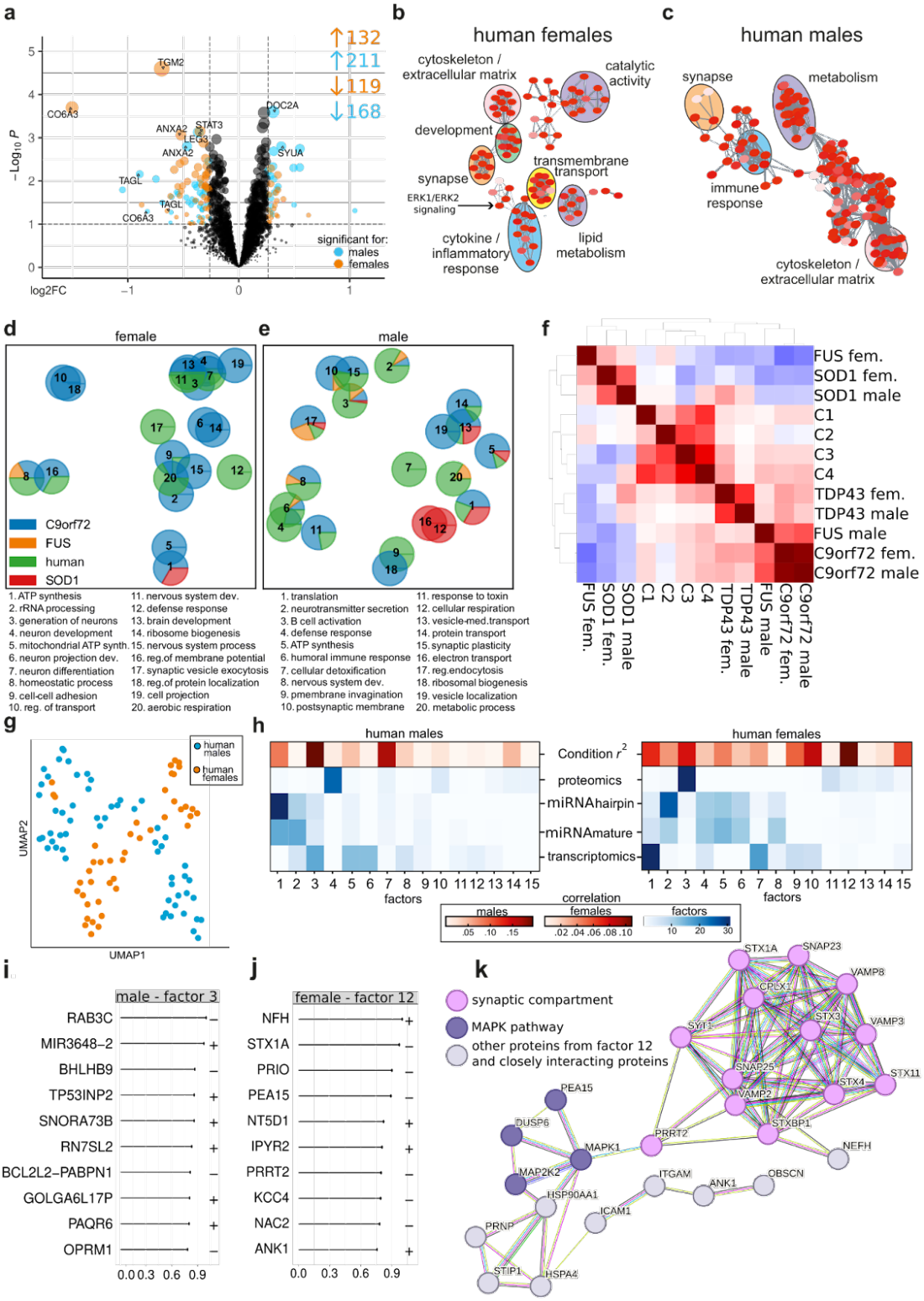


Fig. 4 | Proteomics overview and multiomics factor integration

a Volcano plot of deregulated proteins in human samples. X-axis: log2 fold change (FC); Y-axis: negative log10 padj of each protein. Blue and orange circles indicate significant differential change: left side = decrease (low in ALS), right side = increase (high in ALS).

b-c Revigo based summary of proteomics gene set enrichment results for the human samples (left females and right males). The plot summarizes the functional similarity of the proteins by reducing redundant GO BP terms and clustering the remaining non-redundant terms. The plot is divided into several areas, each of which represents a cluster of similar GO BP terms. The size of the circles represents the number of genes in the GO BP terms, and the color represents the significance

d-e Plots displaying the results of a GO enrichment analysis of the proteins identified in the proteomics data for the four mouse models (TDP43, FUS, C9, and SOD1) humans separated by sex. The GO terms were summarized by clustering pathways in the semantic space based on their descriptions with the GO-Figure clustering tool.
f Pearson correlation of log2 fold change of proteins found to be significantly deregulated (p-value < 0.1) in any mouse model or human samples. Comparisons are grouped according to their similarity using hierarchical clustering as depicted by the dendrograms.
g The MOFA analysis was performed to integrate and visualize the transcriptomics, proteomics, mature & hairpin miRNA data in male and female samples. UMAP representation showing the distribution of the human male (blue) and female (orange) samples.
h Variance decomposition plot showing the contribution of each factor in the MOFA analysis to the overall variance of the integrated dataset separated by sex. The plot displays the variance explained by transcriptomics, proteomics, mature and hairpin miRNA for human male and female samples. The y-axis represents the different omics, and the x-axis represents the different factors in the analysis (left: males, right: females). Blue color indicates the proportion of explained variance, while red color indicates Pearson correlation with the condition.
i The MOFA weight plot for miRNA human males data (Factor 3) displays the contribution of each gene to the MOFA model. Each line represents a gene, the x-axis shows the MOFA weight for that gene, reflecting its importance in the factor. T.
j The MOFA weight plot for proteomics human females data (Factor 12) displays the contribution of each protein to the MOFA model. Each line represents a protein, the x-axis shows the MOFA weight for that protein, reflecting its importance in the factor.
k Protein-protein interaction (PPI) network of the genes in factor 12 female. MAPK1 is an important molecular hub and interacts with PEA15, PRRT2, MEK2, DUSP6, HSPA4 and HSP90AA1. Further proteins of interest are highlighted (bold/dark purple).

Integration of multiomic data reveals sex-specific molecular networks of ALS

We made use of various omics modalities to uncover molecular pathways associated with early ALS-related changes. Here, we employed a biologically motivated approach that focused on identifying valid interaction triplets involving transcripts, miRNAs, and proteins. We found the miR-769-3p-Anxa2-ANXA2 triplet particularly intriguing, as ANXA2 was the only common differentially expressed protein in both sexes. The triplet miR-484-Lamb2-LAMB2 indicated the downregulation of the ECM-component Laminin subunit Beta 2, which is consistent with the identified alterations in the ECM/neurovascular junction compartment (see **Fig. 1**). We also identified triplets related to the regulation of the cytoskeleton component Vimentin, involving potential miRNA regulators such as miR-1301-3p, miR-138-5p, miR-16-5p, miR-26b-5p, and miR-320a (Supplementary Fig. 21, Supplementary Table 11). Valid quadruplets in mouse models, including phosphoproteomic data, highlighted GFAP, SQSTM1, ATXN2L, and XPO1 as salient target proteins (Supplementary Fig. 22, Supplementary Table 11). Importantly, GFAP has recently emerged as a potential marker for neurodegenerative disorders, and XPO1 is a therapeutically targeted protein in ALS.
Following the biologically motivated triplet and quadruplet analysis, we conducted an unsupervised integration of transcriptomic, small RNA and proteomic data using Multi-Omics Factor Analysis (MOFA)[28]. As sex was an important differentiating factor (**Fig. 4g**), MOFA was performed for each sex independently (**Fig. 4h**). To identify overarching terms of relevance beyond the transcriptomic subclusters, we included all patients in this analysis. In males, factor 1, mainly driven by hairpin sRNA, explained 23.7% of the variance. A downregulation in ALS of hsa-miR-7851, -1285-1, -5096 and a cluster of hsa-miR-1273 isoforms strongly contributed to its weight (Supplementary Fig. 24). MAPK1 took a central role among the miR-1273-targets (Supplementary Fig. 26). The transcriptome-based factor 3 correlated best with the disease condition and was driven by genes responsible for *vesicular function* (*RAB3C, NSF*), *oxidative phosphorylation* (*ND1, ND2*), *cell survival* (*BCL2, BHLHB9*) and *RNA metabolism* (*SNORA73B, RN7SL2*). This factor also contained miR-3648-2, which was shown to recruit amyloid precursor protein intracellular domains, regulating survival and neurogenesis[29]. Proteome-dominated factor 4 contains ZO2 and CD44, which are involved in myelination and BCNSB formation. Finally, factor 7, which also showed a strong correlation with disease condition, was dominated by neurofilament isoforms (NFH, NFM, NFL) as well as proteins involved in Ca-binding (HPCL4) and ECM formation (PGCA) (Supplementary Fig. 24). In females, factors 1-3 explained 42.6% of the variance, but factors 10 and 12 correlated best with the disease condition (**Fig. 4h**). Synaptic genes, such as *RAB3C*, *NAPB*, *SNAP25*, contributed to factor 1 and were upregulated in ALS. hsa-miR-1285-1, miR-5096 and the miR-1273 cluster were also contributing to factor 2, similar to factor 1 in males. Factors 3 and 7 were similar to male factor 4, including oligodendrocyte-/myelin markers as well as CD44 and ZO2 (**Fig. 4i**). Factor 10 contained antiproteases SERPINA1 and SERPINA3[30] as well as chitinases CHI3L1 and CHI3L2, which are known biomarkers for ALS[31] (Supplementary Fig. 25). Factor 12 showed downregulation of neurofilament heavy isoform (NFH) as well as the MAPK-ERK1/2-regulator PEA15

(**Fig. 4j**). MOFA analysis on one hand, underlined mechanisms that were identified in individual omics analyses, such as the downregulation of miRNA clusters (particularly in males, Supplementary Fig. 27), ECM components and oligodendrocyte/myelin markers. This unbiased data integration strategy highlighted known ALS biomarkers (neurofilaments, chitinases) and the MAPK pathway, especially in females, as important molecular hubs (**Fig. 4k**).

MAPK pathway emerges as therapeutic target based on multi-omic data integration

Single omic analyses revealed several molecular signaling pathways suitable as pharmacological targets, which could be more appropriate for different subclusters or sexes of ALS patients. Currently licensed drugs target *glutamatergic synapse function* (by Riluzole), *oxidative stress* (by the antioxidant Edaravone), *mitochondrial function* (by TUDCA/Phenylbutyrate), or SOD1 itself (by the antisense oligonucleotide Tofersen). Major mechanisms revealed in our analysis, such as *immune response* and *ECM/BCNSB function* are not yet addressed by licensed drugs. Although multiple molecular pathways identified in our analysis would merit therapeutic validation, we decided to focus on the MAPK pathway, which was altered consistently in humans (albeit stronger in females) and mouse models, across several data types, integration methods, and the different subclusters. More specifically, we concentrated on mitogen-activated protein kinase kinase 2 (MAP2K2 or MEK2) since it appeared upregulated in human PFC and multiple mouse models and is central in the female-specific ALS-associated cluster 12 (**Fig. 4k,** Supplementary figure 11). Furthermore, MEK2 can be modulated by the FDA-approved inhibitor trametinib.(PMID: 23237773).

Validation of MEK2 inhibition by trametinib *in vitro* and *in vivo*

First, we used primary cortical neuronal cultures (PCNC) from P0-1 B57/Bl6 mice, which were treated with glutamate as an in *vitro* model of excitotoxicity in ALS [32]. Treatment with 5mM (6h) glutamate did not affect MEK2 protein expression but significantly increased phospho-Erk1/2 levels. Application of trametinib attenuated glutamate-induced phosphorylation of Erk1/2. Glutamate intoxication also increased cell death (caspase-positive neurons) and reduced the average neurite length, both of which could be counteracted by trametinib (**Fig. 5a-f**, Supplementary Fig. 28). Our data suggests that trametinib attenuates MEK2 activity, reduces Erk1/2 phosphorylation resulting in decreased cell death and increased neurite outgrowth under excitotoxic stress.

To validate the importance of the MAPK pathway *in vivo*, we selected the SOD1 mouse model, which shows the strongest similarities with the largest human ALS subcluster (C1/2) (**Fig. 2h**). We observed strong MEK2 phosphorylation in the motor neurons of the spinal cord, the main tissue involved in the pathology in this animal model [33] (**Fig. 5g,h**). In females, pMEK2 substantially increased with disease progression, while in male pMEK2 returned to control levels after week 14 (**Fig. 5i**). We treated SOD1 mice for 7 weeks with trametinib, starting from week 9 (presymptomatic stage) and we observed a reduction of ERK1/2 phosphorylation compared to vehicle-treated mice in female and male mice (**Fig. 5j**). Trametinib significantly reduced the autophagy receptor p62 in the spinal cord of female, but not male mice, in agreement with its previously described neuroprotective role by increasing autophagy through TFEB activation[34] (**Fig. 5k**). p62 also co-localizes with ubiquitin and mutant SOD1 in protein aggregates[35]. Accordingly, we detected a significant reduction of detergent-insoluble SOD1 and ubiquitin in trametinib-treated females, but not in males (**Fig. 5l,m**). Finally, we investigated whether trametinib has an effect on neurodegeneration. Neurofilament light chain (NfL) plasma concentration [36] was significantly reduced after trametinib-treatment in female SOD1 mice (**Fig. 5n**). The striking sex difference in the trametinib response correlates with the increase of MEK2 phosphorylation during disease progression in female, but not in male SOD1 mice. In conclusion, trametinib has shown marked effects on the clearance of protein aggregates leading to neuroprotection in female SOD1 mice, suggesting that MEK2 is a promising therapeutic target for ALS, particularly in females.

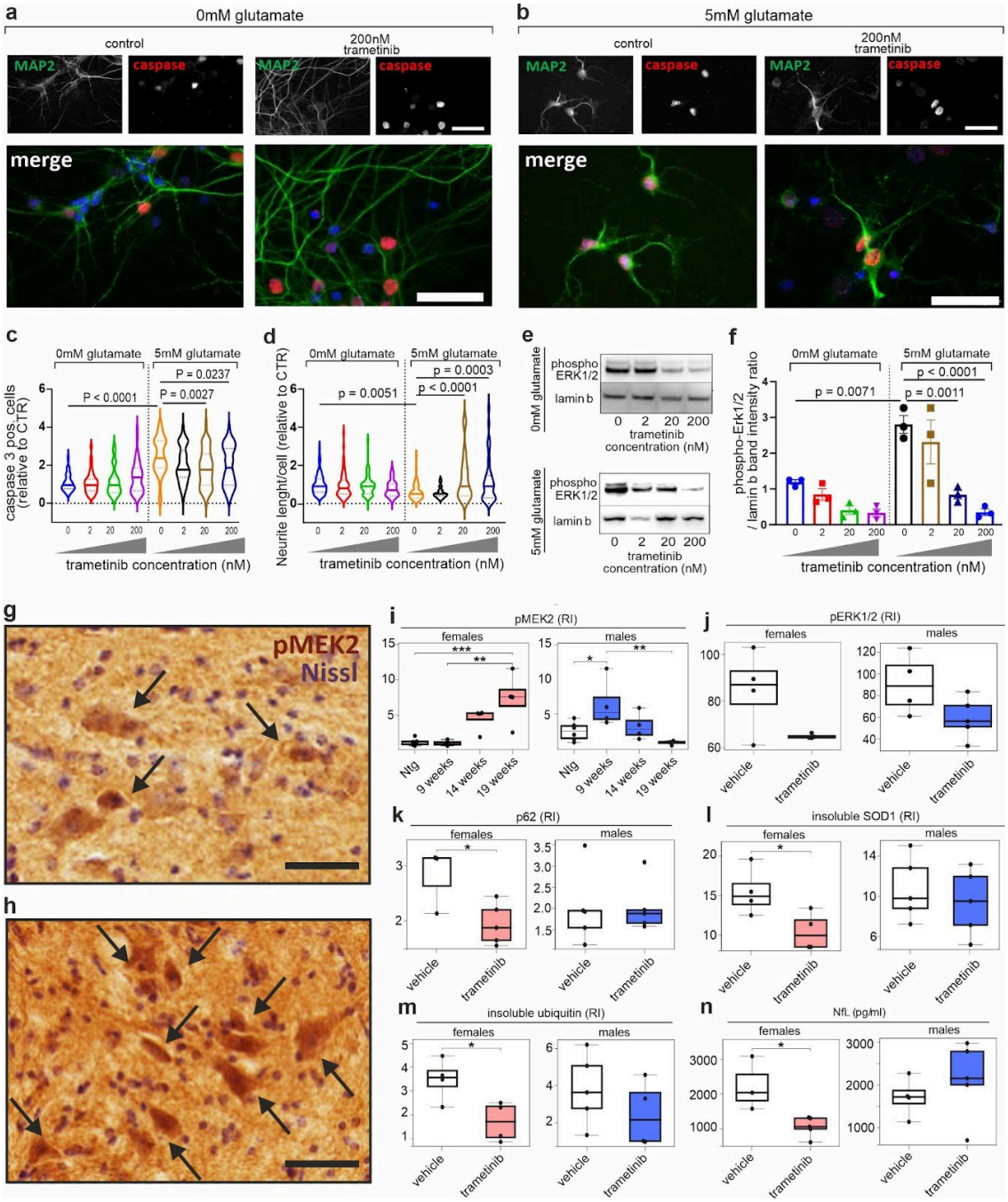


Fig. 5 | Validation experiments

a-b Effects of different concentrations of Trametinib on apoptosis in glutamate (5 mM) and non-glutamate treated cells analyzed by immunostaining. Representative photomicrographs for two of the analyzed conditions (control [vehicle]; 200 nM Trametinib). Scale bar: 40µm. c-d Quantification plots showing effects of the treatment with Trametinib on cell survival (caspase 3 staining) (c) and on neurite outgrowth (d), for glutamate-treated / vehicle-treated (control) treated cells, for all analyzed conditions (control [vehicle]; 2 nM Trametinib; 20 nM Trametinib; 200 nM Trametinib). Data are the mean ± SEM of at least 5 different cultures and tested by one-way Anova.
e-f. Western blot analysis and quantification of Trametinib effects on Phospho-Erk1/2 with and without glutamate treatment. Data are the mean ± SEM of at least 3 different cultures and tested by one-way Anova, $P < 0.05$.
g-h Diffuse pMEK2 immunostaining shown in the lumbar spinal cord of non-transgenic (g) and SOD1G93A mice **(h)** at 19 weeks of age. In ventral horns, pMEK2 staining is mainly present in motor neurons cells. Scale bar: 50 µm.
i Western blot analysis for pMEK2 in lumbar spinal cord of SOD1G93A and Ntg mice at 9, 14 and 19 weeks of age. Data are mean ± SEM (n=6-4 in each experimental group) and are expressed as relative immunoreactivity (RI). *p < 0.05 by pairwise TukeyHSD after one-way ANOVA. Dot blot analysis for pERK1/2 **(j),** p62 **(k)**, insoluble SOD1 **(l)** and ubiquitin **(m)** in spinal cord of SOD1G93A female and male mice, treated with Trametinib or vehicle,

at 16 weeks of age. Data are mean ± SEM (n=3/5 in each experimental group) and are expressed as relative immunoreactivity (RI). *p < 0.05 by Student's t test. NFL plasma levels were analysed in female and male (n) SOD1G93A mice treated or not with Trametinib at 16 weeks of age. Data are mean ± SEM (n =3/5 in each experimental group). *p < 0.05 by Student's t test.

Discussion

In this study, we conducted individual and combined analyses of multiple omics data types to obtain a comprehensive understanding of the molecular architecture of ALS in the PFC, an area that is affected at later stages by TDP43 pathology and has therefore the potential to reveal early disease mechanisms [8].

Almost all of our analyses were characterized by marked sex-specific differences, which were often more pronounced in males than in females. ALS has a slightly higher prevalence in males[37] and sex-specific differences in blood of ALS patients have been previously identified[38,39]. Sex-dependent therapeutic responses were previously observed in ALS mouse models[40,41], but current therapeutic options do not consider patient sex in the differential therapy and neither do clinical trials or the guidelines of the Food and Drug Administration for clinical trial design (FDA, https://www.fda.gov/media/130964/download). Our data, especially the sex-specific differences in MAPK signaling and therapeutic tractability, argues for a stronger consideration of sex as covariate in clinical trials for ALS.

Phenotypic heterogeneity, as reflected by stratum of onset or speed of disease progression, is clearly recognized in ALS[42], and previous transcriptomic analyses suggested molecular subtypes [12]. We identified four clusters in our human ALS cohort based on the transcriptome, which partially mirror previous findings that suggested a clustering of ALS patients into ALS-Ox, ALS-Glia and ALS-TE subgroups (Tam et al.). Whereas the ALS-Glia and ALS-Ox clusters can be correlated to our clusters C1/2 and C3/4, respectively, the ALS-TE cluster finds only marginal representation in our data. As we analyzed the PFC, an area in which TDP43 aggregation is observed later than in the motor cortex, our data supports the finding that ALS-TE is driven by TDP43 dysfunction [7]. Moreover, this suggests that molecular subclusters in ALS may evolve as a function of time and are subject to change in the course of the disease. This should be further explored on the level of liquid biomarkers and could have implications for patient stratification and inclusion into personalized clinical trials.

Furthermore, we suggest that the molecular phenotype in the four mouse models analyzed here can be approximated to these clusters and that these models can therefore serve as surrogates for the molecular subgroups in human ALS. Interestingly, the oldest and so-far most frequently used model, the SOD1.G93A mouse, correlated best with the largest cluster C1/2. Although the SOD1 mouse is not representative of all human ALS cases, our analyses suggest that the pathways dysregulated in this model represent an important subgroup in this population.

The integrated analysis of multiple omics data types identified several deregulated disease-relevant pathways that were previously attributed to ALS, i.e. mitochondrial respiration/oxidative stress, transcriptional regulation/splicing, and protein misfolding[2]. In addition, we also identified pathways that have been less prioritized previously, including dysregulation of the ECM and BCNSB, or the MAPK pathway. The deregulation of multiple pathways, which are in part only distantly related, suggests that future therapeutic approaches should consider the combination of multiple drugs in order to address human ALS mechanistically.

While we found evidence for a female-predominant deregulation of the MAPK pathway in the integrated MOFA analysis, it did not contribute to the separation of clusters C1-C4. Therefore, the MAPK pathway could be an interesting therapeutic target for all human ALS subgroups. MAPKs are fundamental signal transducers that are involved in proliferation, differentiation, cell survival and death[43]. Extracellular and intracellular signals are integrated by MAPK and an overactivation of MAPK signaling, e.g., abnormal phosphorylation of ERK1/2, has been reported in human ALS and ALS mouse models[44]. Increased phosphorylation of ERK1/2 is also observed in our glutamate toxicity model and phospho-MEK2 is increased in the SOD1 mouse model, reproducing the aberrant activation of this pathway, which could be restored with the MEK2-inhibitor trametinib. However, this effect was most pronounced in female animals, arguing for sex-specific efficacy. Sex-specific differences in ALS pathomechanisms constitute a relevant but yet unadressed point in ALS research. Currently, one clinical phase I/II trial examines the safety, tolerability and efficacy of trametinib in patients with ALS (https://clinicaltrials.gov/ct2/show/NCT04326283) and our data argues for an independent evaluation of male and female patients. In addition, our results propose the ECM, the immune response and the RNA processing machinery as potential targets for therapeutic intervention, which need to be explored in more detail in subsequent studies.

Despite the limitations of our study (e.g. analysis of postmortem-tissue limits the assessments to a retrospective view only; limited number of well characterized postmortem brain samples), we aimed at

overcoming them to provide valuable insights into the molecular architecture of ALS. Although more than one hundred brain samples were studied and allowed us to identify molecular subclusters, a larger number could yield an even more granular analysis of molecular subgroups. Nevertheless, analyzing PFC tissue allowed us to study the disease in earlier stages, providing a unique perspective compared to previous studies. Our robust multiomic, computational and integrative approaches led to the identification of subclusters that are reminiscent of previous ones, but show clear differences. Our findings also highlight the importance of splicing and transcript usage as early regulators of disease, with several ALS-related genes involved in these mechanisms. Overall, our study represents a significant step forward in unraveling the complexity of ALS and provides a foundation for further research in this field.

An important goal of our study was the comparison of human brain tissue and samples from transgenic mouse models of ALS. Although we find clear correlations of mouse models and molecular clusters in sALS, we acknowledge that the four mouse models analyzed here, represent four particular scenarios and that for each of the four genes studied in these models also other mouse models exist, which could potentially yield other results[45]. Our study also does not consider the analysis of the DNA methylation status and we did not study other post-translational modifications, such as glycation, methylation, or acetylation all of which could yield additional information about ALS-specific dysregulation, as much as a single cell-based analysis could yield another layer of detail.

In this deep multiomic data integration of human ALS tissue and animal models, we highlight clear sex-specific differences in the pathology of ALS, identify molecular clusters and advocate for their stronger consideration in clinical trials and the development of novel therapeutic strategies. Validation of omics results should rely on multiple systems as each individual mouse model system only reproduces parts of human pathology. Our data suggest additional molecules and pathways for validation as therapeutic targets for ALS and support further exploration of the MAPK pathway for ALS treatment.

Methods

Human postmortem prefrontal cortex samples

Human prefrontal cortex samples were provided by four different brain banks: the Netherlands Brain Bank, King's College London Brain Bank (London Neurodegenerative Diseases Brain Bank), Parkinson's UK Brain Bank, Oxford Brain Bank. In total, 51 ALS and 50 CTR samples (without signs of neurodegeneration) were included. Frozen tissues were shipped on dry ice to the Department of Neurology at the Klinikum rechts der Isar of the Technical University of Munich and stored at -80°C (Supplementary Table 1). Ethical approval for the use of human tissue was obtained from the Ethics Commission (EC) of the University Medical Center Göttingen (2/8/18 AN) and the EC of the Technical University Munich (145/19 S-SR). For sampling, prefrontal cortex blocks were transferred to a cryostat chamber at -20°C and punched with a 20-G Quincke Spinal Needle (Becton Dickinson). ~20 mg tissue was collected into RNAse/DNAse free tubes and kept at -80°C until further use.

ALS animal models

Four transgenic mouse models covering the most frequent ALS-causing genes were used for multiomic studies. Animal handling and all animal experiments in this study were performed in accordance with the applicable animal welfare laws and approved by respective regulatory organs from the involved research centers. Mice were housed in standard cages in a pathogen-free facility in a 12-h light/dark cycle with ad libitum access to food and water. B6;129S6-Gt(ROSA)26Sortm1(TARDBP*M337V/Ypet)Tlbt/J mice [46] (here simply called TDP43-mice) were provided by the Department of Translational Neuroscience of the University Medical Center Utrecht. This model was generated by inserting an 80 kb genomic fragment carrying the human TDP43 locus (including a patient-derived M337V mutation). TDP43 transgenic and control wild-type animals were sacrificed at the age of 26 weeks (presymptomatic stage) for biomaterial collection. B6SJL-Tg(SOD1*G93A)1Gur/J mice [33] (here simply called SOD1-mice) were provided by the Laboratory of Translational Biomarkers, IRCCS-Istituto di Ricerche Farmacologiche "Mario Negri" (IRFMN) Milano. High-copy number B6 congenic Tg(SOD1*G93A)1Gur/J SOD1*G93A male mice from Jackson Laboratory were bred with C57BL/6 female mice to obtain non-transgenic and mutant transgenic G93A*SOD1-expressing mice. SOD1 transgenic and control animals were sacrificed 14 weeks after birth (presymptomatic stage). (poly)GA-NES/C9orf72(R26(CAG-Isl-175GA)-29xNes-Cre) mice [47] (here simply called C9orf72-mice) were provided by the German Center for Neurodegenerative Diseases in Munich. Animals were generated by electroporating plasmids for conditional expression of DPRs which were produced by inserting GFP-(GA)175 genes (encoded using non-repeating alternate codons)

downstream of a floxed stop-cassette in the pEX CAG stop-bpA vector, into murine RMCE embryonic stem cells at the Rosa26 Safe Harbour. Mouse lines GAstop with germ-line transmission were backcrossed to the C57Bl6N background until >98% purity was confirmed using SNP genotyping. C9orf72 transgenic and control animals were sacrificed at 4.5 weeks after birth (early symptomatic stage). Tg(Prnp-FUS)WT3Cshw/J mice [48] (here simply called FUS-mice) were provided by the Laboratory of Translational Biomarkers, Istituto di Ricerche Farmacologiche Mario Negri IRCCS Milano, and were sacrificed at 4 weeks after birth. For each model, 10 transgenic and 10 non-transgenic mice, balanced for sex, were selected.

Preparation of prefrontal cortex from ALS mouse models
Mice were perfused with 50 mL ice-cold PBS prior to the microdissection. The head was removed by cutting at the base of the skull followed by removing the skin. The skull was removed through the small cuts and mouse brains were microdissected in order to isolate the prefrontal cortex region from both hemispheres. The olfactory bulb and the cerebellum were removed by cutting at the cerebellar peduncle, starting from the olfactory bulb (OB) and continuing along the interhemispheric fissure using tweezers with fine tips. Cortex was then lifted from the rest of the brain and removed. Incisions were made in the middle of the cortex in order to remove the PFC (Supplementary Fig. 25). Freshly prepared PFC were collected into nuclease-free tubes, and kept at -80°C until RNA and protein isolation experiments.

RNA and DNA isolation from human and mouse tissue samples
Total RNA was isolated from human and animal prefrontal cortex samples using Trizol Reagent (Sigma Aldrich, Taufkirchen, Germany). All RNA-related experiments were performed under an RNA-workstation fume hood. Briefly, 500 µl of TRI Reagent was added to each sample and tissues were homogenized using plastic homogenizer, followed by addition of 50µl of 1-Bromo-3-Chlor-Propane (Sigma Aldrich, Taufkirchen, Germany). The reaction tubes were mixed by inversion for 10 - 15 seconds and incubated at room temperature for 3 minutes. The lysates were centrifuged at 12.000 x g for 15 minutes / 4°C, leading to phase separation. The RNA-containing aqueous phase was collected and transferred to a fresh Nuclease-free tube. RNA precipitation was performed by adding 250 µl of 2-propanol (AppliChem, Darmstadt, Germany) and 2 µl GlycoBlue Co-precipitant (15 mg/ml) (ThermoFisher, Waltham, MA, USA), followed by mixing and overnight incubation at -20°C. The day after, samples were centrifuged at 12.000 x g for 30 minutes / 4°C, the supernatant was discarded and the RNA pellets were washed three times with 75% ice-cold ethanol (AppliChem, Darmstadt, Germany). The pellets were air-dried for a few minutes under the fume hood. Pellets were reconstituted with 15-20 µl of Nuclease-free water (Sigma Aldrich, Taufkirchen, Germany) followed by 2 minutes of incubation at 55°C in a thermoshaker in order to completely dissolve the RNA. After the RNA isolation, a DNAse treatment was performed in order to remove any DNA contamination from the samples. For that, 5µl of 10X DNAse I Incubation buffer (LifeTechnologies, Carlsbad, CA, United States), 5µl DNase I (2U/µL) and 0.5µl - RNase OUT (40U/µl) were added to each sample. Samples volume was filled up to 50 µl by the addition of Nuclease-free water, followed by incubation at 37°C for 20 minutes. Finally, the RNA samples were cleaned and concentrated with the RNAClean & Concentrator-5 KIT (Zymo Research, Irvine, CA, USA), following the manufacturer's instructions.
DNA isolation from human midbrain samples was performed with the QIAamp DNA Mini Kit following the manufacturer's instructions. Directly after RNA / DNA isolation, nucleic acid concentration and purity were measured in the NanoDrop One spectrophotometer (ThermoFisher, Waltham, MA, USA). RNA integrity was assessed with the Agilent 6000 NanoKit in the 2100 Bioanalyzer (Agilent).

DNA Sequencing Experiments and C9orf72 repeat expansion analysis
Prior to DNA sequencing experiments, the quality of the isolated DNA was determined with a TapeStation 4200 (Agilent, Santa Clara, California, USA). The Ampliseq protocol (Illumina, Inc, California, USA) was used in the succeeding steps. We used a target panel of 566 amplicons covering 30 ALS related genes: *TARDBP; DCTN1; ALS2; ERBB4; TUBA4A; CHMP2B; NEK1; MATR3; SQSTM1; FIG4; C9orf72; SIGMAR1; VCP; GLE1; SETX; OPTN; HNRNPA1; KIF5A; TBK1; ANG; SPG11; CCNF; FUS; PFN1; MAPT; VAPB; SOD1; CHCHD10; NEFH; UBQLN2.* Using this panel, the DNA samples (50-100 ng) were amplified, 14 PCR cycles were used. The amplicons were digested and the AmpliSeq CD indexes were ligated for multiplexing. The quality and quantity of the enriched libraries were validated using the TapeStation 4200 (Agilent, Santa Clara, California, USA). The average fragment size of the amplified product was approximately 480 bp. The libraries were normalized to 9 nM in Tris-Cl 10 mM, pH8.5 with 0.1% Tween 20. The MiSeq (Illumina, Inc, California, USA) was used for cluster generation and sequencing according to standard protocol. Loading concentration was 9pM and 15% of phiX was added. Sequencing configuration was paired-end 250 bp. C9orf72 repeat expansion

analysis was performed using the Asuragen Amplidex® PCR/CE C9orf72 Kit (Asuragen a Biotechne Brand, 2150 Woodward St., Suite 100 Austin, TX 78744 USA). In brief DNA samples (n=51, 40ng each) were amplified using the three-primer GGGCC -Repeat Primed (RP) configuration (combining flanking primers and GGGCC-repeat specific Primer). This configuration allows sizing of GGGCC alleles up to 145 repeats and the detection of expanded GGGCC Alleles > 145 Repeats simultaneously. PCR conditions were as follows. Initial denaturation 5min/98°C; 37 cycles: 35s/ 97°C, 35s/ 62°C, 3min/ 72°C; final elongation 10min/ 72°C). Capillary electrophoresis was carried out on an ABI 3730 (Applied Biosystems, Thermo Fisher Scientific Inc., Waltham, Massachusetts, USA) using the ROX 1000 sizing ladder (Asuragen), followed by analysis with GeneMapper 4.0 software (Applied Biosystems) and conversion of peak size to GGGGCC repeat length via calibration curve method according to manufacturer's instructions.

Preparation of RNA libraries

mRNA and small RNA sequencing experiments were performed in the Functional Genomics Center in Zurich. For the mRNA sequencing, total RNA libraries were prepared using either the TruSeq Stranded mRNA (Illumina, Inc, California, USA)(Short Read Sequencing), or the SMARTer® Stranded Total RNA-Seq Kit v2 -Pico Input Mammalian (A Takara Bio Company, California, USA)(Short Read Sequencing). Briefly, for the TruSeq protocol, total RNA samples (100-1000 ng) were poly-A enriched and then reverse-transcribed into double-stranded cDNA. The cDNA samples were fragmented, end-repaired and adenylated before ligation of TruSeq adapters containing unique dual indices (UDI) for multiplexing. Fragments containing TruSeq adapters on both ends were selectively enriched with PCR. This produces a smear with an average fragment size of approximately 260 bp. The libraries were normalized to 10nM in Tris-Cl 10 mM, pH8.5 with 0.1% Tween 20. For the SMARTer® Stranded Total RNA-Seq Kit v2 -Pico Input Mammalian protocol, total RNA samples (0.25–10 ng) were reverse-transcribed using random priming into double-stranded cDNA in the presence of a template switch oligo (TSO). This results in a cDNA fragment that contains sequences derived from the random priming oligo and TSO. PCR amplification using primers binding to these sequences adds full-length Illumina adapters, including the index for multiplexing. Ribosomal cDNA is cleaved by ZapR in the presence of the mammalian-specific R-Probes. Remaining fragments are enriched with a second round of PCR amplification using primers designed to match Illumina adapters. The product was a smear with an average fragment size of approximately 360 bp. These libraries were normalized to 5nM in Tris-Cl 10 mM, pH8.5 with 0.1% Tween 20. The quality and quantity of the isolated RNA and the enriched libraries were validated using a Fragment Analyzer (Agilent, Santa Clara, California, USA) (for the TruSeq kit) and the Tapestation (Agilent, Waldbronn, Germany) (for the SMARTer® Kit).

Cell-type deconvolution analyses

We performed single cell reference based cell-type deconvolution on RNAseq of mouse models using Scaden[19]. We used adult healthy scRNA-seq datasets for mouse[85] and human[86]). Scaden uses a fully-connected deep neural network ensemble trained on pseudo-bulks simulated from the reference scRNA-seq data. Before deconvolution, we filtered scRNA-seq data using Scanpy[87] to maintain at least 200 genes expressed per cell and at least 5 cells to have one gene's expression. For Scaden, counts per million (CPM) of simulated pseudo bulks and TPMs of the data to be deconvolved was used. Here, CPM was used for scRNA-seq instead of TPMs because the scRNA-seq consists of UMI counts and does not include gene-length bias. We used a variance cutoff of 0.01, and mean squared error calculated over each batch as loss.

Differential Alternative Splicing analysis (DAS)

The study used the splicing tool SUPPA2 (version 2.3) to calculate the differential alternative splicing for seven alternative splicing events, including exon skipping, mutually exclusive exons, intron retention, alternative 3' splice site, alternative 5' splice site, alternative first exon, and alternative last exon. SUPPA2 was used with multipleFieldSelection() to select the TPM values of transcripts, followed by generateEvents with the parameters `-f ioe -e SE SS MX RI FL` and annotation files GENCODE v37 for human data and GENCODE vM26 for mouse data. The inclusion values (PSupplementary) were calculated with psiPerEvent and the differences in PSupplementary (deltaPSupplementary/ΔPSupplementary) between mutant and control conditions were determined using diffSplice with parameters `-m empirical -l 0.05 -gc –save tpm_events` to detect anomalies in the splicing landscape.

Multi-Omics Factor Analysis (MOFA)

We used Multi-Omics Factor Analysis (MOFA) 2 v1.4.0 to integrate data from multiple omics levels, including transcriptomics, miRNA, proteomics, and phosphoproteomics, for mouse models of neurodegenerative diseases. The MOFA model was trained on the data and downstream analyses were performed. Each omics type was preprocessed in its own way and the default training parameters were used. The MOFA models were initialized with 15 initial factors and convergence was reached when the ELBO value did not change more than a deltaELBO value of 1e-4%.

RNA Sequencing

SequencingSequencing was done using the Illumina platforms NovaSeq 6000 (for transcriptomics) and HiSeq 2500 (for small RNA sequencing) (Illumina, Inc, California, USA) according to the standard protocols. Small RNA sequencing was processed with the use of the RealSeq-AC miRNA (SomaGenics, California, USA) (Short Read Sequencing). All samples were quantified and quality controlled with the Fragment Analyzer (Agilent, Santa Clara, California, USA). Briefly, RNA samples (1ng-1ug) were adaptor ligated and circularized followed by reverse-transcription into cDNA. The cDNA samples were amplified using PCR that also incorporated sample barcodes. The library product, a peak with a fragment size of approximately 149 bp, was normalized to 10nM in Tris-Cl 2 mM, pH8.5 with 0.1% Tween 20. The quality and quantity of the enriched libraries were also validated using a Fragment Analyzer. Transcriptomics data has been processed using the NextFlow Core RNASeq pipeline, version 3.0 described at[49]. The data has been demultiplexed with bcl2fastq, and the fastq files have undergone several quality checks including FastQC[50]. Salmon was used for pseudo alignment and quantitation, with a salmon index built using GRCm39 with annotations from GENCODE vM26 for the mice data and GRCh38 with annotations from GENCODE v37 for the human data. Count matrices from Salmon were used in downstream analyses. The count matrices were filtered, keeping genes with at least ten counts in 50% of the samples of any condition and sex. We used the clusterProfiler R package[51] and GO biological processes and molecular functions for gene set enrichment analysis, filtering terms by size between 10 and 500 genes, and correcting for multiple testing (Benjamini-Hochberg correction).

Global proteomics on mouse & human PFC brain tissue samples

Brain tissues coming from both human PFC and the 4 different mouse models were prepared as follows. Tissues were grinded with a biomasher using 350 µL of MeOH:H_2O (4:1). Protein pellets were resuspended in 200 µL Laemmli buffer (10% SDS, Tris 1M pH 6.8, glycerol) then centrifuged at 11.135 rpm at 4°C for 5 minutes. Protein concentration was determined using DC assay (BioRad, Hercules, CA, USA) according to the manufacturer's instructions. 100 µg of protein lysate for each sample were heated at 95°C for 5 minutes and stacked in an in-house prepared 5% acrylamide SDS-PAGE stacking gel. Gel bands were reduced and alkylated prior to overnight digestion (enzyme:protein ratio of 1:80) at 37°C using modified porcine trypsin (Mass Spec Grade, Promega, Madison, USA). The generated peptides were extracted with 60% acetonitrile followed by a second extraction with 100% acetonitrile (ACN). Peptides were resuspended in 30 µL of H20, 2% ACN, 0.1% FA and iRT peptides (Biognosys, Schlieren, Switzerland) were added to each sample according to the manufacturer's instructions as an internal Quality Control. NanoLC-MS/MS analyses were performed on a nanoAcquity UltraPerformance LC® (UPLC®) device (Waters Corporation, Milford, MA) coupled to a Q-Exactive Plus mass spectrometer (Thermo Fisher Scientific, Waltham, MA). The solvent system consisted of 0.1% FA in water (solvent A) and 0.1% FA in ACN (solvent B). Samples (equivalent to 800 ng of proteins) were loaded on a Symmetry C18 precolumn (20 mm × 180 µm with 5 µm diameter particles, Waters) over 3 min at 5 µL/min with 99% of solvent A and 1% of solvent B. Peptides were separated on an ACQUITY UPLC BEH130 C18 column (250 mm × 75 µm with 1.7 µm diameter particles) at 400 nL/min with the following gradient of solvent B: from 1 to 8 % over 2 min, from 8 to 35% over 77 min, from 35 to 90 % over 1 min, at 90% for 5 minutes and from 90 to 1% over 2 minutes. The samples of each cohort were injected in randomized order. The system was operated in Data Dependent Acquisition mode with automatic switching between MS (mass range 300–1800 m/z with R = 70,000, Automatic gain control (AGC) fixed at 3.10^6 ions and a maximum injection time set at 50 ms) and MS/MS (mass range 200–2000 m/z with R = 17,500, AGC fixed at 1.10^5 and the maximal injection time set to 100 ms) modes.The ten most abundant precursor ions were selected on each MS spectrum for further isolation and higher energy collisional dissociation, excluding monocharged and unassigned and ions. The dynamic exclusion time was set to 60 s. A sample pool was injected as an external QC every 6 samples for the human cohort and every 5 samples for mouse cohorts. The MaxQuant software (version 1.6.14) was used to process raw data. Andromeda search engine was used to assign peaks with trypsin/P specificity against a protein sequence database generated in-house containing all human (24th of August 2020, 20 421 entries) or mouse entries (27th of March 2020, 17 016 entries for SOD1 & TDP43 models and 29th of September 2020, 17 061 entries for C9 and FUS models) extracted from UniProtKB-SwissProt.

Methionine oxidation and acetylation of proteins' N-termini were set as variable modifications and cysteine carbamidomethylation as a fixed modification. The "match between runs" option was enabled for protein quantification. The maximum false discovery rate was set to 1% at peptide and protein levels using a decoy strategy. Intensities were extracted from the Proteingroup.txt file for statistical analysis. The MaxQuant protein vs. samples table was used for downstream analyses, including Label Free Quantitation (LFQ) intensities, and only Swiss-Prot proteins were kept while TrEMBL proteins were removed for higher reliability. After filtering out low abundant proteins, i.e. proteins which were detected in less than 50% of the samples in any combination of condition and sex, and imputing missing values using the missForest [52], the intensities were log2 transformed and used for principle component exploration, heatmaps, and differential abundance analysis. The limma package was used for linear modeling and p-values were adjusted with Benjamini-Hochberg correction. The protein names were mapped to corresponding genes and searched for enriched biological processes and molecular functions using the criteria described in the transcriptomics data, with a $p<0.01$ threshold for functional annotation analyses.

Phospho-proteomics on mouse PFC brain tissue samples

Starting from the protein extracts from the global proteomics experiments, proteases inhibitors (Sigma, P8340) and phosphatases inhibitors (final concentration in Na3VO4 = 1 mM) were added to all samples. Protein concentration was determined using RC-DC assay (BioRad, Hercules, CA, USA) according to the manufacturer's instructions. 250 µg of proteins for each sample were reduced and alkylated prior to an in-house optimized single-pot, solid-phase-enhanced sample preparation (SP3) protocol (adapted from Hughes et al., Nat Protoc, 2019). Briefly, beads A (Sera-Mag Speed beads, Fisher Scientific, Germany, 45152105050250) and beads B (Sera-Mag Speed beads, Fisher Scientific, Germany, 65152105050250) were combined (ratio 1:1) and, after 3 washing steps with H_2O, were added to the samples (ratio beads:protein of 10:1 for each type of beads, meaning a 20:1 ratio for the combination of beads). After inducing protein binding to the beads with 100% ACN for 18 minutes, the beads/proteins mixtures were washed twice with 80% EtOH and once with 100% ACN before being resuspended in 95 µL NH_4HCO_3 prior to overnight on-beads digestion (enzyme:protein ratio of 1:20) at 1 000 rpm at 37°C using modified porcine trypsin/lys-C (Mass Spec Grade mix, Promega, Madison, USA). Digestion was stopped using TFA (final pH < 2). Recovered peptides were resuspended in 170 µL of 80% ACN, 0.1% TFA and phosphomix I light (Sigma Aldrich) was added to each sample (ratio peptide (µg)/mix(fmol) = 1.6). Phosphopeptide enrichment was performed on 5 µL phase Fe(III)-NTA cartridges on an AssayMAP Bravo platform following an IMAC protocol. Briefly, cartridges were washed and primed with 50% ACN, 0.1% TFA, then equilibrated with 80% ACN, 0.1% TFA. 100 µL of samples were loaded at 2 µL/min on the phase then washed with 80% ACN, 0.1% TFA before being eluted in 20 µL 1% NH_4OH at 5 µL/min. After the enrichment, FA was added to each sample as well as phosphomix I heavy (Sigma Aldrich) (ratio peptide (µg)/mix(fmol) = 1.6). Dried phosphopeptides were resuspended in 40 µL H_2O, 2% ACN, 0.1% FA. Sample preparation steps for C9 & FUS mouse models were identical to those previously described for SOD1 & TDP43, except that proteins were extracted from new tissue samples just before the ph-proteomics experiment.

Nano-LC-MS/MS analyses were performed on a nanoAcquity UPLC devise (Waters) coupled to a Q-Exactive HF-X mass spectrometer (Thermo Scientific, Bremen, Germany) equipped with a Nanospray Flex™ ion source. The solvent system consisted of 0.1% FA in water (solvent A) and 0.1% FA in ACN (solvent B). Samples were loaded on an ACQUITY UPLC® Peptide BEH C18 Column (250 mm x 75 µm with 1.7 µm diameter particles) over 3 min at 5 µL/min with 99% of solvent A and 1% of solvent B. Phosphopeptides were separated on an ACQUITY UPLC® M-Class Symmetry® C18 Trap Column (20 mm x 180 µm with 5 µm diameter particles; Waters) at 400 nL/min with the following gradient of solvent B: from 1 to 2 % over 2 min, from 2 to 35% over 77 minutes, and from 35 to 90% over 1 minute. The samples of each cohort were injected in randomized order. The system was operated in Data Dependent Acquisition mode with automatic switching between MS (mass range 375–1500 m/z with R = 120,000, Automatic gain control (AGC) fixed at 3.10^6 ions and a maximum injection time set at 60 ms) and MS/MS (mass range 200–2000 m/z with R = 15,000, AGC fixed at 1.10^5 and the maximal injection time set to 60 ms) modes. The ten most abundant ions were selected on each MS spectrum for further isolation and higher energy collisional dissociation, excluding monocharged and unassigned ions. The dynamic exclusion time was set to 40 s.

Raw ph-proteomics data were processed using MaxQuant software (version 1.6.14). Peaks were assigned with the Andromeda search engine with trypsin/P specificity against an in-house generated protein sequence database containing all mouse entries extracted from UniProtKB-SwissProt (29th of September 2020, 17 061 entries). The minimal peptide length required was seven amino acids and a maximum of one missed cleavage was allowed. Methionine oxidation, acetylation of proteins' N-termini

and serine, threonine and tyrosine phosphorylation were set as variable modifications and Cysteine carbamidomethylation as a fixed modification. For protein quantification, the “match between runs" option was enabled. The maximum false discovery rate was set to 1% at peptide and protein levels with the use of a decoy strategy. Intensities were extracted from the Phospho(STY).txt file and processed through Perseus software (version 2.0.7.0) in which contaminants and reversed proteins, as well as proteins with a negative score were removed. Using the “expand sites table” option, the intensities of the different phosphopeptides involved in one phosphosite were summed and phosphosites with a localization probability below 75% were removed. The Perseus output table was then used for further statistical analysis.
Only those proteins that were detected in more than 50% of the mice samples of any combination of sex and condition were kept for each dataset. i Intensities were log2-transformed, quantile normalization was applied and missing values were imputed by lowest quartile. The rest of the data processing for the phospho proteomics follows the procedure described above for proteomics data.

small RNA sequencing data processing and miRNA target prediction
The small RNA data has been processed using the NextFlow Core smRNASeq pipeline, version 1.0[49]. Reads trimmed and aligned against miRBase (version 22.1) using Bowtie1, both for mature miRNA and hairpins. miRNAs with at least ten counts in 50% of the samples of any condition and sex were kept and the rest were filtered out. The unnormalized count matrices were used for subsequent DESeq2[53] differential expression analysis, stratified by sex, and adjusting the p-values using the Benjamini-Hochberg correction. The mature miRNA were mapped to their corresponding genes using miRDB v6.0[54], excluding matches with scores lower than 60 or more than 800 targets, as recommended. For each miRNA present in the miRNA expression matrices, we obtained experimentally validated targets from miRTarBase 8.0 and predicted targets from miRDB v22. miRTarBase provides the most extensive curated database of validated miRNA-target interactions (MTI) collected from literature using natural language processing (NLP) to select functional miRNA studies. Additionally, the miRDB database includes MTIs predicted from MirTarget, which uses a support vector machine (SVM) to analyze thousands of high-throughput sequencing experiments; each final prediction has a probability score attached to it, which is the output of SVM. A higher probability score indicates a higher likelihood of accurate target predictions. Therefore, we set a threshold of 0.6 on output probabilities to select only very likely MTIs. Finally, we joined miRNA-target pairs from both sources for further analysis.

Gene ontology, pathway enrichment analyses and protein interaction networks

The study performed gene set enrichment analysis using gseGO and gseKEGG from the clusterProfiler R package, with biological processes and molecular function chosen as background databases for GO enrichment. The p-value cutoff was set at 0.05. Differential expression was presented using volcano plots generated with the Enhanced Volcano plot package in R, and ORA (Over-Representation Analysis) was performed on genes that showed at least one significant DAS event using the clusterProfiler function enrichGO(), with a p-value cutoff of 0.1 and Benjamini Hochberg correction for multiple hypothesis testing. Protein protein interaction networks were created using the STRING protein interaction network database v11.0[58,59] under standard settings.

Weighted Correlation Network Analysis (WGCNA)
We performed weighted gene co-expression analysis using WGCNA[18]. Pairwise Pearson correlations between each pair of genes were used to create signed regulatory networks in WGCNA. We computed an adjacency matrix and used a soft-thresholding approach to approximate a scale-free topological network. Eigengenes or eigen proteins were calculated as the first principal component for each module. This resulted in several modules. To merge modules, we merged similar modules based on hierarchical clustering (RNAseq - SOD1: 0.4, C9orf72: 0.4, FUS: 0.5, TDP43: 0.4, Proteomics: 0.25). We calculated relationships between WGCNA modules and traits. Sex was also included in traits resulting in four traits - male ALS, female ALS, male CTR, female CTR. First we filtered modules based on the significance of this module-gene relationship ($p<0.05$) and then selected modules that were highly correlated with either male or female ALS. The correlation cutoffs differed between mouse models (RNAseq - SOD1: 0.5, C9orf72: 0.3, FUS: 0.3, TDP43: 0.5, Proteomics - SOD1: 0 , C9orf72: 0, FUS: 0, TDP43: 0). Using WGCNA, we also analyzed co-expression networks in the mouse models. The minimal module size was set to 30 with a merge height of 0.4-0.5 and a correlation threshold between 0.3-0.5 (Supplementary Table 14).

Reproducible Data Pipeline

Given the size and complexity of all the datasets, data processing and data analysis methods, we integrated all the analyses and raw data into a Data Version Control pipeline[80]. The pipeline used both nextflow and docker images to guarantee a reproducible execution environment. This enabled us to keep all results consistent, as any change or exploration to a data processing method automatically triggered the recalculation of all the dependent analyses and integrations.

Primary cortical culture from mice

Cortical neuronal cell cultures were prepared from mouse C57Bl/6J pups on postnatal day 0-1 (P0-P1). The protocol for generating the primary neuronal cell cultures was in accordance with local and international guidelines on the ethical use of animals. Animal care followed official governmental guidelines and all efforts were made to minimize the number of animals used and their suffering. Pups were decapitated and the brains were collected in dissection media containing the 10X Hanks balanced salt solution HBSS and sodium bicarbonate. Cortex was dissected, meninges were removed and small pieces of the cortex were collected in a falcon tube. Tissues were trypsinized at 37 °C in a water bath for 12 min and was treated with 200 µL of DNAse I (10 mg/mL). Tissues were triturated in fetal bovine serum with fire polished pasteur pipette until all tissue pieces dissolved, centrifuged for 4 min at 800 g and the cell pellet was suspended and maintained in neurobasal medium supplemented with B27 and antibiotics (0.06 µg/mL penicillin and 0.1 µg/mL streptomycin). The cells were seeded at a density of $3x10^5$ cells per well in 24-well plates, respectively. Before cell seeding, coverslips were acid washed, rinsed many times with water, sterilized with ethanol and UV light and placed in the well plate. The plates were then coated with poly-L-ornithine (0.05 mg/mL) overnight and Laminin (10 µg/mL) for 2 h in the incubator before use. The cells were cultured at 37 °C in a humidified atmosphere containing 5% CO_2 for 7 days prior to experimentation with medium exchange every 3 days. For glutamate excitotoxicity induction, L-glutamic acid (Tocris, UK) was dissolved in 50mM sodium hydroxide (NaOH), and the stock solution of 50 mM was prepared prior to use. Appropriate concentration of glutamate was prepared in maintenance media (neurobasal medium supplemented with B27 and antibiotics). Cells were exposed to 5mM glutamate by exchanging 1:3 of the media at DIV7. After 6h of incubation, glutamate was washed out thoroughly and the cells were fixed for immunocytochemistry or lysed for protein extraction.

Immunocytochemistry and microscopy

Cells were cultured on coverslips following the described methods and were immunostained at DIV7 according to standard techniques. To this, cells were fixed with (4 % paraformaldehyde in PBS at room temperature for 10 min. For quenching the free aldehyde groups, cells were treated with 50 mM ammonium chloride for 15 min and then washed with PBS. For permeabilization of the cell membrane, PBS with 0.25 % Triton X-100 was applied for 10 min at RT. Non-specific binding sites were blocked by applying 10% Goat serum in PBS for at least 20 min. Dilutions of primary antibodies were prepared in blocking solution to a final volume of 180 µl per 18-mm coverslip and cells were incubated for 90 min at 37 °C shaking. The following primary antibodies were used: mouse anti-MAP2 (Invitrogen,Catalog # MA5-12826, MA, USA) 1:500, rabbit anti-cleaved caspase 3 (cell signaling, Catalog #9661, MA, USA) 1:250 . Cells were washed 3 times for 5 min with PBS before applying the secondary antibodies. Secondary antibodies were applied to cells, and incubated for 30 min followed by repeating the washing steps. For double staining, a second primary antibody was added and the same steps were repeated. Coverslips were mounted on slides using a mounting medium with DAPI. Images were captured by a 63x oil objective with an inverted fluorescence microscope (Zeiss, Jena, Germany) and analyzed by image J software. Fifteen random images from each coverslip were analyzed for cell death by counting the number of cleaved caspase-3 positive cells. Neurite lengths were measured using simple neurite tracing (SNT) plugin in image J software. Statistical analyses were conducted with the GraphPad Prism software version 9.4.1 (GraphPad, SanDiego, CA, USA). Outliers were identified and removed using Grubbs test (Alpha = 0.1). Comparisons were done using One-Way Anova test and data plotted as mean ± standard error of the mean (SEM) of at least five independent experiments. Differences were considered significant when $p < 0.05$.

Protein extraction and western blotting

For protein analysis, cells were washed once with 1X PBS and after adding the lysis buffer RIPA, protease inhibitor cocktail 1:25 and phosphatase inhibitor 1:20 incubated on ice for 5 min. The cells were scratched with a cell scraper on ice and transferred in 1 mL reaction tubes and homogenized by passing through the U-100 Insulin syringes a few times. Protein concentration was determined using Pierce™ BCA Protein Assay Kit (Thermo Fisher Scientific) following the manufacturer's instructions. 1 µl of a protein sample was used in the assay. The prepared colorimetric reactions were analyzed in an ELISA

plate reader (Tecan's Infinite® M200 PRO). 20 g of the samples were loaded on the gel (NuPAGE™ 4 to 12%, Bis-Tris, Invitrogen, Carlsbad, CA). NuPAGE LDS-sample buffer 1:4 and sample reducing buffer 1:10 was added to lysed protein before loading on the gels and incubated shaking at 75 °C for 13 min and they centrifuged at 12.000 g at 4 °C. Proteins were separated by gel electrophoresis at 200 V. Proteins were transferred to a nitrocellulose membrane using the iBlot2 gel transfer device and transfer stack (Thermo Fisher Scientific). Membranes were blocked for 30 min at room temperature with 5% nonfat milk in PBST followed by incubation with primary antibodies (diluted in blocking buffer) overnight at 4 °C under rotation. After washing 4 times with PBST (5 min each time) the membranes were incubated with highly-sensitive HRP-labeled secondary antibodies (1:10.000 diluted in blocking buffer) at room temperature for 1 h followed by intensive washing with PBST. Blots were incubated with ECL reagent and were imaged on a BioRad Molecular Imager ChemiDoc™. Band signal intensities were quantified with Image J Software and were normalized to housekeeping protein and control condition. Statistical analyses were conducted with the GraphPad Prism software version 9.4.1 (GraphPad, SanDiego, CA, USA). Outliers were identified and removed using Grubbs test (Alpha = 0.1). Comparisons were done using One-Way Anova test and data plotted as mean ± standard error of the mean (SEM) of at least three independent experiments. Differences were considered significant when $p < 0.05$.

Tissue protein extraction for immunoblot analysis

Spinal cords were homogenized in 5 volumes (w/v) of 1% boiling SDS[61]. Protein homogenates were further sonicated, boiled for 10 min and centrifuged at 13.500 g for 5 min. Supernatants were analyzed by dot blot analyses. For detergent-insoluble protein extraction, mouse tissues were homogenized in 10 volumes (w/v) of buffer, 15 mM Tris-HCl pH 7.6, 1 mM DTT, 0.25 M sucrose, 1 mM $MgCl_2$, 2.5 mM EDTA, 1 mM EGTA, 0.25 M sodium orthovanadate, 2 mM sodium pyrophosphate, 25 mM NaF, 5 µM MG132, and a protease inhibitors cocktail (Roche), as described[62]. Briefly, the samples were centrifuged at 10.000 g and the pellet was suspended in an ice-cold homogenization buffer with 2% Triton-X100 and 150 mM KCl. The samples were then centrifuged at 10.000 g to obtain the Triton-insoluble fraction (insoluble).

Immunohistochemistry

Mice were anesthetized and perfused transcardially with 50 mL of phosphate-buffered saline (PBS) followed by 100 mL of 4% paraformaldehyde (Sigma-Aldrich) in PBS. Spinal cord was rapidly removed, postfixed for 3h, transferred to 20% sucrose in PBS overnight and then to 30% sucrose solution until they sank, frozen in N-pentane at 45°C and stored at ± 80°C. Before freezing, spinal cord was divided into cervical, thoracic, and lumbar segments and included in Tissue-tec OCT compound (Sakura). Coronal sections (30 µm) of lumbar spinal cord were then sliced and immunohistochemistry was done. Antibody used for immunohistochemistry is rabbit monoclonal anti-phospho-MEK (Ser221) (pMEK2) antibody (1:50, Cell Signaling; RRID: AB_490903). Briefly, slices were incubated for 1h at room temperature with blocking solutions (0.2% Triton X100 plus 2% normal goat serum (NGS)), then overnight at 4°C with the primary antibodies. After incubation with biotinylated secondary antibodies (1:200; 1 h at room temperature; Vector Laboratories) immunostaining was developed using the avidin–biotin kit (Vector Laboratories) and diaminobenzidine (Sigma). Sections were counterstained with 0.5% cresyl violet. Stained sections were collected at 20 X and 40 X with an Olympus BX-61 Virtual Stage microscope so as to have complete stitching of the whole section, with a pixel size of 0.346 µm. Acquisition was done over 6-µm-thick stacks with a step size of 2 µm. The different focal planes were merged into a single stack by mean intensity projection to ensure consistent focus throughout the sample. Finally, signals were analyzed for each slice with ImageJ and OlyVIA software.

Immunoblotting

Protein levels were determined using the BCA protein assay (Pierce) and analyzed by western blot and dot blot, as described previously[63]. Membranes were blocked with 3% (w/v) BSA (Sigma-Aldrich) and 0.1% (v/v) Tween 20 in Tris-buffered saline, pH 7.5, and incubated with primary antibodies and then with peroxidase-conjugated secondary antibodies (GE Healthcare). Antibodies used for immunoblot were the following: rabbit monoclonal anti-phospho-MEK (Ser221) (pMEK2) antibody (1:2000, Cell Signaling; RRID: AB_490903), rabbit monoclonal anti-phospho-p44/42 MAPK (Erk1/2) (Thr202/Tyr204) antibody (1:2000, Cell Signaling; RRID: AB_2315112), rabbit polyclonal anti-human SOD1 antibody (1:1000, StressMarq Biosciences; RRID: AB_2704217), rabbit polyclonal anti-ubiquitin antibody (1:1000, Abcam; RRID: AB_306069), mouse monoclonal anti-SQSTM1/p62 (p62) antibody (1:500, Abcam); goat anti-mouse or anti-rabbit peroxidase-conjugated secondary antibodies (respectively 1:20000 and 1:10000, GE Healthcare). Blots were developed with the Luminata Forte Western

Chemiluminescent HRP Substrate (Millipore) on the ChemiDoc™ Imaging System (Bio-Rad). Densitometry was done with Image Lab 6.0 software (Bio-Rad). The immune reactivity of the different proteins was normalized to Ponceau Red staining (Fluka).

Preclinical study in the SOD1^{G93A} mouse
Starting at a presymptomatic stage (9 weeks of age), SOD1^{G93A} female and male mice received 3 mg/kg dose of trametinib (n=10) or vehicle (PBS) (n=10) through intranasal delivery. The drug was administered twice per week for 7 weeks (from 9 to 16 weeks of age). At 16 weeks of age, mice were sacrificed, spinal cord and plasma were collected for subsequent biochemical analysis. The Mario Negri Institutional Animal Care and Use Committee and the Italian Ministry of Health (Direzione Generale della Sanità Animale e dei Farmaci Veterinari, Ufficio 6) prospectively reviewed and approved the animal research protocols of this study (prot. no. 9F5F5.143 and 9F5F5.250) and ensured compliance with international and local animal welfare standards.

NFL measurements
Plasma samples were collected from mice in K2-EDTA BD Microtainer blood collection tube and centrifuged at 5.000 g for 5 minutes to isolate plasma samples. The plasma NFL concentration was measured using the Simoa® NF-light™ Advantage (SR-X) Kit (#103400) on the Quanterix SR-X™ platform with reagents from a single lot, according to the protocol issued by the manufacturer (Quanterix Corp, Boston, MA, USA).

Statistical analysis for in vivo experiments
Prism 7.0 (GraphPad Software Inc., San Diego, CA) was used. For each variable, the differences between experimental groups were analyzed by Student's *t* test, or one-way ANOVA followed by post-hoc tests. *P* values below 0.05 were considered significant.

Data availibility

Mouse raw RNA-seq data and processed gene expression data can be accessed via the National Center for Biotechnology Information's Gene Expression Omnibus database (GSE234246). Encrypted raw RNA-seq data for the human cohort can be accessed via the European Genome-Phenome Archive (registered study: EGAS00001007318). Proteomics and phosphoproteomics datasets have been deposited in the ProteomeXchange Consortium database with the identifiers PXD043300 and PXD043297, respectively.

Code availability
For RNASeq, we used the NextFlow pipeline for alignment and DESeq2 for differential analysis:
https://www.nextflow.io/, https://www.bioconductor.org/packages//2.10/bioc/html/DESeq.html
Alternative splicing was done with SUPPA, DRIMSeq and DEXSeq:
https://github.com/comprna/SUPPA,
https://bioconductor.org/packages/release/bioc/html/DRIMSeq.html,
https://bioconductor.org/packages/release/bioc/html/DEXSeq.html
Enrichment was done with: https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html,
http://revigo.irb.hr/
Gene gene interaction network analysis was done with:
https://cran.r-project.org/web/packages/WGCNA/index.html
Deconvolution was done with Scaden:
https://scaden.readthedocs.io/en/latest/

References

1. Suzuki, N., Nishiyama, A., Warita, H. & Aoki, M. Genetics of amyotrophic lateral sclerosis: seeking therapeutic targets in the era of gene therapy. *J. Hum. Genet.* **68**, 131–152 (2023).
2. Mead, R. J., Shan, N., Reiser, H. J., Marshall, F. & Shaw, P. J. Amyotrophic lateral sclerosis: a neurodegenerative disorder poised for successful therapeutic translation. *Nat. Rev. Drug Discov.* **22**, 185–212 (2023).
3. Braak, H. & Braak, E. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol. (Berl.)* **82**, 239–259 (1991).
4. Braak, H. *et al.* Staging of brain pathology related to sporadic Parkinson's disease. *Neurobiol. Aging* **24**, 197–211 (2003).
5. Aronica, E. *et al.* Molecular classification of amyotrophic lateral sclerosis by unsupervised clustering of gene expression in motor cortex. *Neurobiol. Dis.* **74**, 359–376 (2015).
6. Morello, G. *et al.* Integrative multi-omic analysis identifies new drivers and pathways in molecularly distinct subtypes of ALS. *Sci. Rep.* **9**, 9968 (2019).
7. Tam, O. H. *et al.* Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia. *Cell Rep.* **29**, 1164-1177.e5 (2019).
8. Brettschneider, J. *et al.* Stages of pTDP-43 pathology in amyotrophic lateral sclerosis. *Ann. Neurol.* **74**, 20–38 (2013).
9. Prudencio, M. *et al.* Distinct brain transcriptome profiles in C9orf72-associated and sporadic ALS. *Nat. Neurosci.* **18**, 1175–1182 (2015).
10. Figueroa-Romero, C. *et al.* Expression of microRNAs in human post-mortem amyotrophic lateral sclerosis spinal cords provides insight into disease mechanisms. *Mol. Cell. Neurosci.* **71**, 34–45 (2016).
11. Umoh, M. E. *et al.* A proteomic network approach across the ALS-FTD disease spectrum resolves clinical phenotypes and genetic vulnerability in human brain. *EMBO Mol. Med.* **10**, 48–62 (2018).
12. Eshima, J. *et al.* Molecular subtypes of ALS are associated with differences in patient prognosis. *Nat. Commun.* **14**, 95 (2023).
13. Catanese, A. *et al.* Multiomics and machine-learning identify novel transcriptional and mutational signatures in amyotrophic lateral sclerosis. *Brain J. Neurol.* awad075 (2023) doi:10.1093/brain/awad075.
14. NeuroLINCS Consortium *et al.* An integrated multi-omic analysis of iPSC-derived motor neurons from C9ORF72 ALS patients. *iScience* **24**, 103221 (2021).
15. Straub, I. R., Weraarpachai, W. & Shoubridge, E. A. Multi-OMICS study of a CHCHD10 variant causing ALS demonstrates metabolic rewiring and activation of endoplasmic reticulum and mitochondrial unfolded protein responses. *Hum. Mol. Genet.* **30**, 687–705 (2021).
16. Mitropoulos, K., Katsila, T., Patrinos, G. P. & Pampalakis, G. Multi-Omics for Biomarker Discovery and Target Validation in Biofluids for Amyotrophic Lateral Sclerosis Diagnosis. *Omics J. Integr. Biol.* **22**, 52–64 (2018).
17. Kenna, K. P. *et al.* NEK1 variants confer susceptibility to amyotrophic lateral sclerosis. *Nat. Genet.* **48**, 1037–1042 (2016).
18. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
19. Menden, K. *et al.* Deep learning-based cell composition analysis from tissue expression profiles. *Sci. Adv.* **6**, eaba2619 (2020).
20. Månberg, A. *et al.* Altered perivascular fibroblast activity precedes ALS disease onset. *Nat. Med.* **27**, 640–646 (2021).
21. Buratti, E. & Baralle, F. E. The multiple roles of TDP-43 in pre-mRNA processing and gene expression regulation. *RNA Biol.* **7**, 420–429 (2010).
22. Polymenidou, M. *et al.* Misregulated RNA processing in amyotrophic lateral sclerosis. *Brain Res.* **1462**, 3–15 (2012).

23. Rinchetti, P., Rizzuti, M., Faravelli, I. & Corti, S. MicroRNA Metabolism and Dysregulation in Amyotrophic Lateral Sclerosis. *Mol. Neurobiol.* **55**, 2617–2630 (2018).

24. Hirano, M. *et al.* Mutations in the gene encoding p62 in Japanese patients with amyotrophic lateral sclerosis. *Neurology* **80**, 458–463 (2013).

25. Beckers, J., Tharkeshwar, A. K. & Van Damme, P. C9orf72 ALS-FTD: recent evidence for dysregulation of the autophagy-lysosome pathway at multiple levels. *Autophagy* **17**, 3306–3322 (2021).

26. Reijnders, M. J. M. F. & Waterhouse, R. M. Summary Visualizations of Gene Ontology Terms With GO-Figure! *Front. Bioinforma.* **1**, 638255 (2021).

27. Biogen. *A Phase 1, Double-Blind, Placebo-Controlled, Single-Ascending-Dose Study to Evaluate the Safety, Tolerability, Pharmacokinetics, and Pharmacodynamics of BIIB100 Administered Orally to Adult Participants With Amyotrophic Lateral Sclerosis*. https://clinicaltrials.gov/ct2/show/NCT03945279 (2023).

28. Argelaguet, R. *et al.* Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).

29. Shu, R. *et al.* APP intracellular domain acts as a transcriptional regulator of miR-663 suppressing neuronal differentiation. *Cell Death Dis.* **6**, e1651 (2015).

30. Kenigsbuch, M. *et al.* A shared disease-associated oligodendrocyte signature among multiple CNS pathologies. *Nat. Neurosci.* **25**, 876–886 (2022).

31. Costa, J. *et al.* Cerebrospinal Fluid Chitinases as Biomarkers for Amyotrophic Lateral Sclerosis. *Diagn. Basel Switz.* **11**, 1210 (2021).

32. Dong, X., Wang, Y. & Qin, Z. Molecular mechanisms of excitotoxicity and their relevance to pathogenesis of neurodegenerative diseases. *Acta Pharmacol. Sin.* **30**, 379–387 (2009).

33. Gurney, M. E. *et al.* Motor neuron degeneration in mice that express a human Cu,Zn superoxide dismutase mutation. *Science* **264**, 1772–1775 (1994).

34. Chun, Y. S. *et al.* MEK1/2 inhibition rescues neurodegeneration by TFEB-mediated activation of autophagic lysosomal function in a model of Alzheimer's Disease. *Mol. Psychiatry* **27**, 4770–4780 (2022).

35. Gal, J., Ström, A.-L., Kilty, R., Zhang, F. & Zhu, H. p62 accumulates and enhances aggregate formation in model systems of familial amyotrophic lateral sclerosis. *J. Biol. Chem.* **282**, 11068–11077 (2007).

36. Benatar, M., Wuu, J. & Turner, M. R. Neurofilament light chain in drug development for amyotrophic lateral sclerosis: a critical appraisal. *Brain J. Neurol.* awac394 (2022) doi:10.1093/brain/awac394.

37. Mehta, P. *et al.* Prevalence of amyotrophic lateral sclerosis in the United States using established and novel methodologies, 2017. *Amyotroph. Lateral Scler. Front. Degener.* **24**, 108–116 (2023).

38. Santiago, J. A., Quinn, J. P. & Potashkin, J. A. Network Analysis Identifies Sex-Specific Gene Expression Changes in Blood of Amyotrophic Lateral Sclerosis Patients. *Int. J. Mol. Sci.* **22**, 7150 (2021).

39. Murdock, B. J., Goutman, S. A., Boss, J., Kim, S. & Feldman, E. L. Amyotrophic Lateral Sclerosis Survival Associates With Neutrophils in a Sex-specific Manner. *Neurol. Neuroimmunol. Neuroinflammation* **8**, e953 (2021).

40. Günther, R. *et al.* The rho kinase inhibitor Y-27632 improves motor performance in male SOD1(G93A) mice. *Front. Neurosci.* **8**, 304 (2014).

41. Torres, P. *et al.* Gender-Specific Beneficial Effects of Docosahexaenoic Acid Dietary Supplementation in G93A-SOD1 Amyotrophic Lateral Sclerosis Mice. *Neurother. J. Am. Soc. Exp. Neurother.* **17**, 269–281 (2020).

42. Tahedl, M. *et al.* Propagation patterns in motor neuron diseases: Individual and phenotype-associated disease-burden trajectories across the UMN-LMN spectrum of MNDs. *Neurobiol. Aging* **109**, 78–87 (2022).

43. Sahana, T. G. & Zhang, K. Mitogen-Activated Protein Kinase Pathway in Amyotrophic Lateral Sclerosis. *Biomedicines* **9**, 969 (2021).

44. Ayala, V. *et al.* Cell stress induces TDP-43 pathological changes associated with ERK1/2 dysfunction: implications in ALS. *Acta Neuropathol. (Berl.)* **122**, 259–270 (2011).

45. Bonifacino, T. *et al.* Nearly 30 Years of Animal Models to Study Amyotrophic Lateral Sclerosis: A Historical Overview and Future Perspectives. *Int. J. Mol. Sci.* **22**, 12236 (2021).

46. Gordon, D. *et al.* Single-copy expression of an amyotrophic lateral sclerosis-linked TDP-43 mutation (M337V) in BAC transgenic mice leads to altered stress granule dynamics and progressive motor dysfunction. *Neurobiol. Dis.* **121**, 148–162 (2019).

47. LaClair, K. D. *et al.* Congenic expression of poly-GA but not poly-PR in mice triggers selective neuron loss and interferon responses found in C9orf72 ALS. *Acta Neuropathol. (Berl.)* **140**, 121–142 (2020).

48. Mitchell, J. C. *et al.* Overexpression of human wild-type FUS causes progressive motor neuron degeneration in an age- and dose-dependent fashion. *Acta Neuropathol. (Berl.)* **125**, 273–288 (2013).

49. Ewels, P. A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.* **38**, 276–278 (2020).

50. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. (2010).

51. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innov. Camb. Mass* **2**, 100141 (2021).

52. Stekhoven, D. J. & Bühlmann, P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinforma. Oxf. Engl.* **28**, 112–118 (2012).

53. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

54. Chen, Y. & Wang, X. miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res.* **48**, D127–D131 (2020).

55. Bhattacherjee, A. *et al.* Cell type-specific transcriptional programs in mouse prefrontal cortex during adolescence and addiction. *Nat. Commun.* **10**, 4169 (2019).

56. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).

57. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

58. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

59. Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2021).

60. Kuprieiev, R. *et al.* DVC: Data Version Control - Git for Data & Models. (2022) doi:10.5281/zenodo.7083235.

61. Filareti, M. *et al.* Decreased Levels of Foldase and Chaperone Proteins Are Associated with an Early-Onset Amyotrophic Lateral Sclerosis. *Front. Mol. Neurosci.* **10**, 99 (2017).

62. Pasetto, L. *et al.* Defective cyclophilin A induces TDP-43 proteinopathy: implications for amyotrophic lateral sclerosis and frontotemporal dementia. *Brain J. Neurol.* **144**, 3710–3726 (2021).

63. Luotti, S. *et al.* Diagnostic and prognostic values of PBMC proteins in amyotrophic lateral sclerosis. *Neurobiol. Dis.* **139**, 104815 (2020).

Acknowledgements
We thank all members of the Lingor, Bonn, Carapito, Schlapbach, Pasterkamp and Bonnetto laboratories for their feedback on the manuscript. This work was performed by the research consortium “Multi-omic analysis of axono-synaptic degeneration in motoneuron disease (MAXOMOD)”, which was funded in the scope of the E-Rare Joint Transnational Call for Proposals 2018 “Transnational research projects on hypothesis-driven use of multi-omic integrated approaches for discovery of disease causes and/or functional validation in the context of rare diseases.” Consortium members: S.B., J.P., V.B., C.C., R.S., M.K., coordination by P.L.
L.C.G., M.P., P.L. are supported by the Bundesministerium für Bildung und Forschung (01GM1917A), P.L. and D.E. were further supported by the Munich Cluster for Systems Neurology (SyNergy). S.H. was supported by SFB1192 B8, S.O. by SFB1286 SP02 and KFO296 P8, R.K. by FOR5068 P9, F.H. by the M3I excellence initiative and an UKE postdoctoral stipend, C.H. by KFO306 P11, and S.B. by SFB1286 SP02 and SFB1192 PB8 and PC3. J.P. was supported by Stichting ALS Nederland (TOTALS). D.E. has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101057649 and the Association for Frontotemporal Degeneration (AFTD).
The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions
P.L., L.C.G. conceived the project and S.H., S.O., S.B. conceptualized the computational analysis. P.L., L.C.G., J.K., A.D. designed the sample collection methodology, reviewed sample and data quality and coordinated the acquisition of human tissue samples from the brain banks, as well as pathological, genetic and clinical information. L.C.G, M.P., Q.Z. processed mouse/human brain samples for multiomics experiments with infra-structure provided by V.B., C.C., D.E and P.L.
C.D., M.S., T.B.H., I.C., and M.D. contributed to genetic data analyses.
S.H., S.O., R.K., F.H. were responsible for the computational data analysis and conceived and planned statistical analyses. L.C.G, S.H., F.H., R.K., P.L., M.P., and L.P, contributed to figure generation and assembling.
M.P, L.P, L.C.G, S.S., S.C. conducted validation experiments with conceptual input from P.L and V.B.
L.C.G, S.H., S.B. and P.L. wrote the manuscript, with input from all co-authors.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information
The online version contains supplementary material available at XXX.
Correspondence and requests for materials should be addressed to Paul Lingor.

Title

msqrob2PTM: differential abundance and differential usage analysis of MS-based proteomics data at the post-translational modification and peptidoform level

Authors

Nina Demeulemeester[1,2,3], Marie Gébelin[4], Lucas Caldi Gomes[5], Paul Lingor[5], Christine Carapito[4], Lennart Martens[1,2,] Lieven Clement[3]

Affiliations

1. VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium
2. Department of Biomolecular Medicine, Ghent University, Ghent, Belgium
3. StatOmics, Department of Applied Mathematics, Computer science and Statistics, Ghent University
4. Laboratoire de Spectrométrie de Masse BioOrganique, IPHC UMR 7178, CNRS, Université de Strasbourg, Infrastructure Nationale de Protéomique ProFI - FR2048, Strasbourg, France
5. Department of Neurology, Technical University Munich, Munich 80333, Germany

Corresponding author: lieven.clement@ugent.be

Running title

msqrob2PTM: differential PTM abundance and usage analysis

Abbreviations

DPA: differential PTM abundance
DPU: differential PTM usage
fdp: false discovery proportion
FDR: false discovery rate
fpr: false positive rate
LC: liquid chromatography
MS: mass spectrometry
PTM: post-translational modification
tpr : true positive rate
ROC: receiver operating characteristic

Abstract

In the era of open-modification search engines, more post-translational modifications than ever can be detected by LC-MS/MS-based proteomics. This development can switch proteomics research into a higher gear, as PTMs are key in many cellular pathways important in cell proliferation, migration, metastasis and ageing. However, despite these advances in modification identification, statistical methods for PTM-level quantification and differential analysis have yet to catch up. This absence can partly be explained by the inherently low abundance of many PTMs and the confounding of PTM intensities with its parent protein abundance.

Therefore, we have developed msqrob2PTM, a new workflow in the msqrob2 universe capable of differential abundance analysis at the PTM, and at the peptidoform level. The latter is important for validating PTMs found as significantly differential. Indeed, as our method can deal with multiple PTMs per peptidoform, there is a possibility that significant PTMs stem from one significant peptidoform carrying another PTM, hinting that it might be the other PTM driving the perceived differential abundance.

Our workflows can flag both Differential Peptidoform (PTM) Abundance (DPA) and Differential Peptidoform (PTM) Usage (DPU). This enables a distinction between direct assessment of differential

abundance of peptidoforms (DPA) and differences in the relative usage of peptidoforms corrected for corresponding protein abundances (DPU). For DPA, we directly model the log2-transformed peptidoform (PTM) intensities, while for DPU, we correct for parent protein abundance by an intermediate normalisation step which calculates the log2-ratio of the peptidoform (PTM) intensities to their summarized parent protein intensities.
We demonstrated the utility and performance of msqrob2PTM by applying it to datasets with known ground truth, as well as to biological PTM-rich datasets. Our results show that msqrob2PTM is on par with, or surpassing the performance of, the current state-of-the-art methods. Moreover, msqrob2PTM is currently unique in providing output at the peptidoform level.

Introduction

Mass-spectrometry-based proteomics allows the identification and quantification of a myriad of posttranslational modifications (PTMs) which reveal additional complexity and diversity of the proteome. Indeed, PTMs greatly extend the number of different forms of a protein, i.e., proteoforms, that can be found. More importantly, these PTMs can impact protein functions (1-4) and are linked to a variety of diseases and developmental disorders (5-8). Aberrant PTM status can cause a number of detrimental effects ranging from the alteration of protein folding to the dysregulation of cell signalling. It is thus of great importance to study these PTMs in detail, not only through their correct identification but also by their correct quantification and subsequent statistical analysis.

In recent years, there has been a significant improvement in the identification of PTMs with the advent of open-modification search engines such as MsFragger (9), Open-pFind (10) and ionbot (11). Yet, bespoke statistical methodologies for differential PTM analysis are lacking. To our knowledge, the only dedicated tool released at the time of writing is MSstatsPTM (12). This can be partly attributed to the complexity of PTM-rich data. Peptides can contain multiple PTM sites, sites are not always modified and modified peptides are usually harder to detect than their non-modified counterparts (4). This means that enrichment methods are most often needed for sufficient detection, which increases technical variability and experimental complexity, time and cost, which in turn leads to less available replicates (13,14). As a result, PTM-rich data are characterised by a high amount of missingness and variability, complicating statistical analysis.

Moreover, the parent proteins on which the PTMs occur can also change in abundance regardless of the PTM. Any changes in abundance of a PTM are then confounded with changes in protein abundance (15). It is therefore crucial that any proposed statistical methodology for PTMs can take this into account.
Here, we introduce the concept of differential PTM abundance (DPA) and differential PTM usage (DPU) to enable a clear distinction between directly assessing differential abundance of PTMs (DPA) on the one hand, and differences in relative PTM abundance upon correction for the overall protein abundance (DPU), on the other hand.

In the current state-of-the-art, MSstatsPTM, DPU is achieved through an adjustment based on the model estimates of a separate PTM model as well as a protein model. We argue that this approach is suboptimal as it fails to leverage the inherent correlation between the parent protein and PTMs or peptidoforms, i.e., a specific peptide with its corresponding modifications. Additionally, the separate modelling and adjustment process in MSstatsPTM can artificially amplify small differences. This phenomenon is demonstrated in figure 1. Here, we can see a PTM for which the PTM intensities closely mimic the protein intensities across the samples. Although no significant differences are observed at the PTM or protein level when comparing the "Combo" and "Ctrl" conditions in the dataset, the adjustment inflates the difference, causing MSstatsPTM to return a significant PTM. Hence, in msqrob2PTM we employ a different normalisation strategy that directly accounts for this correlation between peptidoform and protein.

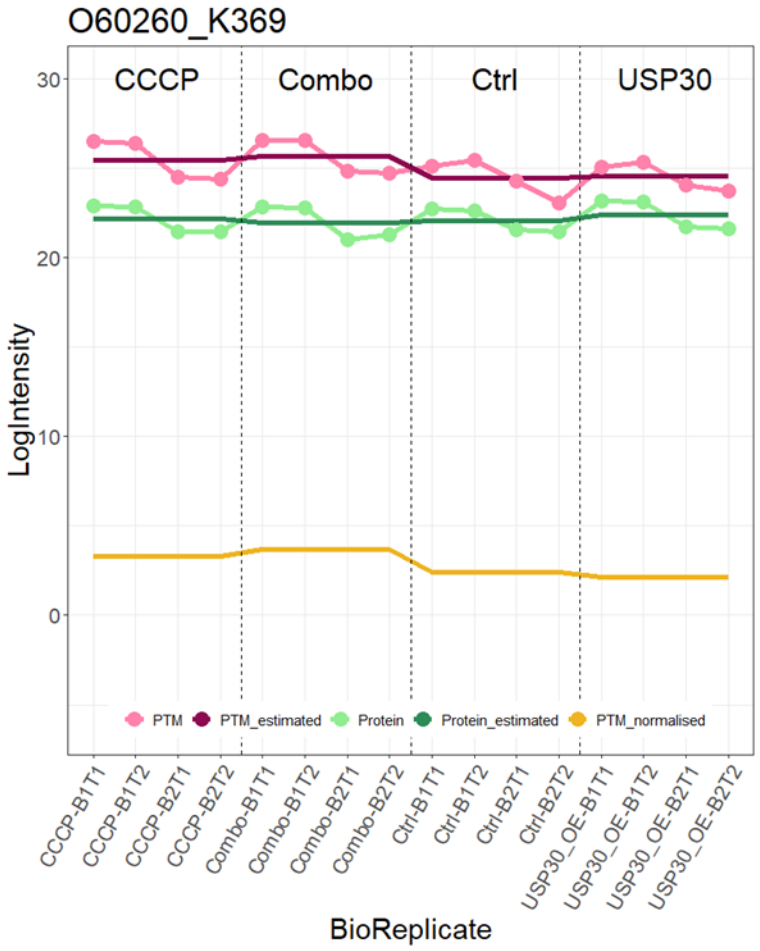


Figure 1: line plot displaying the PTM $\log_2$ intensity values (pink dotted line) and $\log_2$ intensity values of its parent protein (light green dotted line) in each sample. MSstatsPTM first fits a model to the PTM (dark pink line) and to the protein intensities (dark green line) to estimate the average intensity in each condition. Subsequently, the fitted average protein abundances are subtracted from the fitted average PTM intensities to obtain the average PTM abundances in each condition corrected for protein abundance (yellow line). MSstatsPTM corrected PTM abundances seem to indicate differential PTM usage. Moreover, the comparison between "Combo" vs "Ctrl" is returned by MSstatsPTM as statistically significant. This, however, appears to be an artifact of MSstatsPTM as the correction for protein abundance does not account for the link between protein and PTM intensities within samples. Indeed, when comparing "Combo" and "Ctrl" sample level intensities, the pattern at the PTM-level closely follows that of its parent protein.

Additionally, we will not limit ourselves to the analysis of the PTMs. Indeed, our method can manage the analysis of peptidoforms as well. In many studies, each distinct PTM will likely not be characterized by a myriad of peptidoforms. It is therefore possible that a significant PTM effect can be attributed to only one or two strongly significant associated peptidoforms, which may be significant for another reason, i.e. a different PTM occurring on that (those) peptidoform(s). We think it is crucial that potential users thus do not restrict their analysis to the PTM alone, but also assess the individual peptidoforms that carry the specific PTM.

We here present a statistical, R-based workflow, based on the msqrob2 R package (16), to carry out differential abundance as well as differential usage analysis at the peptidoform and PTM level. We apply this workflow to simulated datasets, a spike-in study, and to biological datasets, and use these to compare our method to MSstatsPTM. We show that our approach does not suffer from the artifacts that are introduced by uncoupling the within-sample correlation between PTM and parent protein, while maintaining good sensitivity and FDR control. The approach is freely available and can be consulted on https://github.com/statOmics/msqrob2PTMpaper.

Experimental procedures

In this section, we first introduce the msqrob2 workflow for differential peptidoform/PTM abundance and usage analysis. Next, we introduce the datasets that were used to test and validate the workflow and benchmark it to MSstatsPTM.

Workflow

The general workflow for the differential abundance analysis on PTM and peptidoform level was developed in R (17) (version 4.2) and is mainly based on two R packages: msqrob2 (https://www.bioconductor.org/packages/release/bioc/html/msqrob2.html, version 1.6.0) and QFeatures (18) (https://www.bioconductor.org/packages/release/bioc/html/QFeatures.html, version 1.8.0).
QFeatures provides an infrastructure to store and manage mass spectrometry data across different levels (e.g. peptidoform and protein level) whilst keeping links between the levels where possible. For each preprocessing step a novel, linked assay is constructed. In this way, the original data is not overwritten, and preprocessed data can be traced back to its origin. msqrob2 is a package with updated and modernised versions of the MSqRob (16) and MSqRobSum (19)tools and builds upon the QFeatures class infrastructure. It provides a robust statistical framework for differential analysis of label-free LC-MS proteomics data to infer on differential abundance on the peptide (peptidoform) and/or protein level. Here, we add workflows that provide inference on differential abundance and usage at the PTM and peptidoform level.

We make a distinction between differential abundance and differential usage. This is the difference between directly assessing differential abundance (DA) on the one hand, and differences in relative abundance upon correction for the overall protein abundance (DU), on the other hand. Essentially, this relates to a difference in normalisation (see point 3 below).

We first provide an overview of the workflow before going over each step in detail.

1. Conversion of input data and construction of the QFeatures object
2. Pre-processing
3. Normalisation
4. Peptidoform level analysis
5. Summarisation of peptidoforms to PTM level
6. PTM level analysis
7. Results exploration plus visualisation

1. Conversion of input data and construction of the QFeatures object

As input data, we require the output of a quantification algorithm (in txt or csv format) that contains all peptidoform identifications, parent protein(s) and per sample intensities. This should be in wide format: each unique peptidoform should be on one line that contains (at least) the information on its parent protein, modification (plus location), and intensities for each sample. As quantitative proteomics data can be readily transformed into this format, we have no restrictions on search engines or quantification algorithms users want to adopt.
Once the data are in the right format they are imported as a QFeatures object. Next, information on the experimental design can be added in the *colData* instance of the object.

2. Pre-processing

First, the peptidoform data can be filtered. Each peptidoform should have measured intensity values in at least two samples, or else are filtered out. Intensities are log-transformed if not already the case. Of course, decoys and contaminants should be removed.

The pre-processing steps are not limited to those above, as, depending on the nature of the dataset and user knowledge, more filtering steps can be added.

3. Normalisation

Distinct normalisation steps should be adopted for inferring on differential abundance and differential usage. For DA, only median centring or mean centring can be used, e.g. via the *normalise* function from the QFeatures package. DU requires an additional normalisation to correct for changes occurring in the parent protein. Indeed, changes in the overall protein abundance between conditions can trigger the associated PTM(s) to be detected as differentially abundant. To infer on PTM(s) for which the effect of the treatment differs from that of the overall protein, we first summarise the protein intensity value per sample for each unique protein, e.g., via robust regression using the *robustSummary* function in the MsCoreUtils[470] R package, and we subsequently subtract it from the intensity values corresponding to all peptidoforms derived from that protein, i.e.

$$y^*_{i,p,P} = y_{i,p,P} - \mu_{i,P} \quad (1)$$

With $y^*_{i,p,P}$ the normalised log2-transformed intensity for peptidoform *p* in sample *i* with parent protein *P*, $y_{i,p,P}$ the log2-transformed intensity for peptidoform *p* in sample *i* with parent protein *P* before normalisation and $\mu_{i,P}$ the summarised intensity for protein *P* in sample *i*.

It is possible to calculate the summarised protein intensity value directly from the PTM dataset itself. However, when the experiment includes both an enriched and non-enriched (global profiling) dataset we recommend using the non-enriched dataset to calculate the summarised protein values. Of note, steps one and two should also be applied to the non-enriched data.

4. Peptidoform level analysis

Before transitioning to the PTM level, it is possible to directly assess differential usage or expression on peptidoform level. The steps to take are exactly the same as step 6 below, but instead of using the PTM assay obtained in step 5, we use the normalised peptidoform assay obtained in step 3 as input to the *msqrob* function.
This allows the user to assess associated peptidoforms underlying significant PTMs of interest.

5. Summarisation of peptidoforms to PTM level

For each unique PTM (i.e. unique protein – modification – location combination), we need a summarised intensity value per sample. This is done by taking a subset of the dataset with all peptidoforms containing a specific PTM and summarising all corresponding intensity values into one value per sample. When peptidoforms contain multiple PTMs, these are used multiple times. Here we apply robust regression using the *robustSummary* function in the MsCoreUtils (20) R package by default to summarise the peptidoform level data at the PTM-level. In this way, we obtain an intensity assay on the PTM level. This assay can then be added to the existing QFeatures object.

6. PTM level analysis

We use the functionalities of the msqrob2 package for this step. Msqrob2 (16,19,21,22) provides a robust linear (mixed) model framework for assessing differential abundance in proteomics experiments. To assess differential abundance on the protein level, the workflows can start from raw peptide intensities or summarised protein abundance values. The model parameter estimates can be stabilized by ridge regression, empirical Bayes variance estimation and robust M-estimation. Here we assess differential abundance on the PTM level by first summarising peptidoform expression values (step 5).
When one predictor (e.g. *condition*) is present in the dataset, we perform an msqrob analysis on PTM intensities with the following model:

$$y_{cs} = \beta_0 + \beta_c^{condition} + \varepsilon_{cs}$$

With y_{cs} the summarised log2-transformed PTM intensity in sample *s* of condition *c*, β_0 the intercept, and $\beta_c^{condition}$, the effect of a condition *c*. The error term ε_{cs} is assumed to be normally distributed with mean 0 and variance σ2.
When multiple predictors are present, the model can be expanded as needed, with the additional possibility of using mixed models. The user needs to specify the model formula themselves using lm or lme4 (23) R syntax.
The contrast matrix for contrasts of interest can be specified via the *makeContrast* function present in msqrob2, which are subsequently assessed using the *hypothesisTest* function. By default, the results of the latter function are corrected for multiple testing using the Benjamini-Hochberg false discovery rate (FDR) method.

The model results are stored in the existing QFeatures object together with the raw data and the pre-processed data.

7. Results exploration plus visualisation

The abovementioned model results contain a significance table with (adjusted) p-values, log fold changes, standard errors, degrees of freedom and test statistics.
Different visualisations can easily be made based on this table and the links to the underlying intensity data in the QFeatures object, such as volcano plots, heatmaps and line plots at the peptidoform, PTM and/or protein level.

Data

Our novel msqrob2 workflow is tested and benchmarked to MSstatsPTM using two computer simulations developed by the MSstatsPTM team, the spike-in dataset from the MSstatsPTM paper, and data from two real experiments.

The computer simulations were specifically developed for testing differential PTM workflows, and also allowed us to directly compare our method to MSstatsPTM. The first simulation produced “perfect” datasets with no missing values and many modified features per PTM, while the second simulation incorporated missing values and limited modified features, producing more lifelike datasets.

The spike-in dataset consists of fifty human ubiquitinated peptides that were spiked into four background mixtures in known amounts. Hence, the true log-fold changes and identity of the truly changing PTMs are known.

The biological case studies consist of a biological, label-free LC-MS/MS ubiquitination experiment by Cunningham *et al*. to study the role of USP30, a deubiquitylase enzyme, in mitophagy regulation (24); and a label-free phosphorylation experiment with a two-factor design where both the total proteome and phosphoproteome were measured.

More details on each dataset are given below.

8. Computer simulations

We used the two computer simulations from the MSstatsPTM team that were found on https://github.com/devonjkohler/MSstatsPTM_simulations/tree/main/data (simulation1_data.rda and simulation2_data.rda). The first simulation consists of data without any missing values, while in

the second simulation, missing data is introduced. For each simulation, 24 datasets were created with different experimental designs and intensity variance. In each dataset, 1000 PTMs were simulated.

Half of the PTMs were simulated to have a fold change between conditions. However, of the half with differential fold changes on the PTM level, 250 could be confounded with differential fold changes of the parent protein. For further details on the creation of the datasets, we refer to the MSstatsPTM paper (12) and to their GitHub page.

Both simulations contain an enriched PTM dataset as well as its non-enriched protein counterpart. From each of the 24 datasets, the FeatureLevelData was extracted from the PTM and the protein dataset. These two datasets were then used as input to the workflow and all seven steps were followed. The protein dataset was used for the normalisation step.
Because it is known which PTMs are differentially abundant and/or differentially used, we can readily evaluate the performance of a method in terms of the false positive rate (fpr), sensitivity, specificity, precision and accuracy, and true positive rate (tpr) - false discovery proportion (fdp) plots. Note, that tpr is the fraction of the truly differentially abundant PTMs picked up by the method and fdp is the fraction of false positives in the total number of PTMs flagged as differentially abundant. On the tpr-fdp plot we also indicate the observed fdp at 5% FDR cut-off, which is expected to be close to 5%.

We compared our results with the results obtained with the MSstatsPTM method, on their GitHub page https://github.com/devonjkohler/MSstatsPTM_simulations/tree/main/data (adjusted_models_sim1.rda and adjusted_models_sim2.rda) and included these in the tpr-fdp plots.

6. Spike-in dataset

The MSstatsPTM team also developed a biological spike-in dataset with known ground truth to test their approach. Fifty human ubiquitinated peptides were spiked into four background mixtures consisting of human and E.coli proteins in different amounts. These four mixtures represent four different conditions and for each, two replicates were created. An overview of the experimental design can be seen in figure 2. Because the amount of spiked-in peptides is known, the true log-fold changes between the conditions is known and it is possible to assess whether the method can pick up these fold changes, and if these fold changes differ from the fold change of the corresponding protein in the background. Note, however, that as opposed to real experiments, the ubiquitinated peptides in the spike-in study are not correlated to their corresponding protein in the background. Further technical details can be found in the MSstatsPTM paper (12). The dataset can be found on MassIVE: MSV000088971.The true log fold changes (before and after protein adjustment) are depicted in table 1.

Comparison	**True log2 FC without adjustment**	**True log2 FC after adjustment**
mix2 vs mix1	-1	-1
mix3 vs mix1	0	1
mix4 vs mix1	-1	0
mix3 vs mix2	1	2
mix4 vs mix2	0	1
mix4 vs mix3	-1	-1

Table 1: True log2 fold changes of the spike-in peptides in the different comparisons between the mixtures.

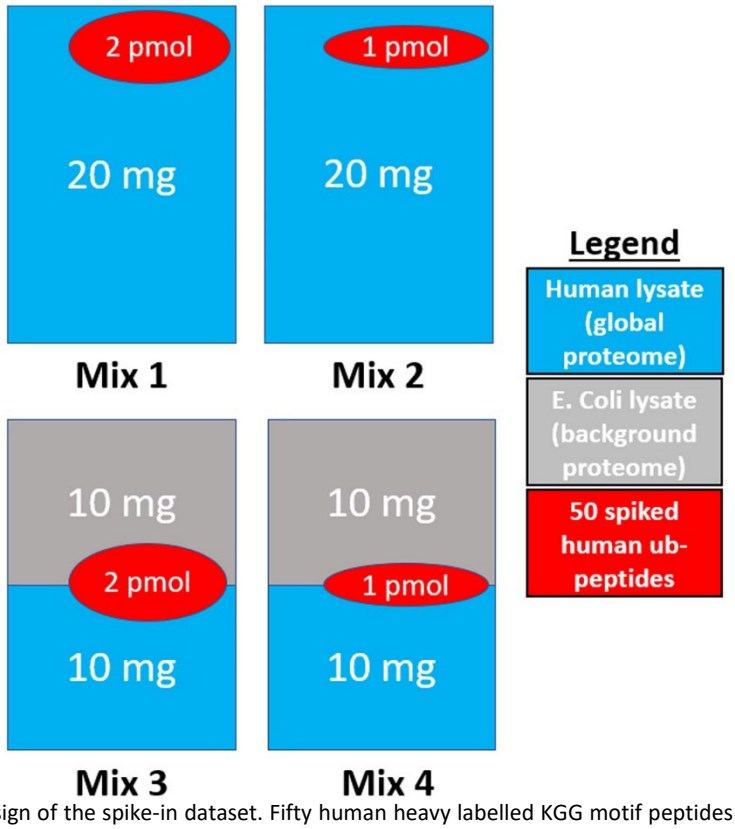


Figure 2: Experimental design of the spike-in dataset. Fifty human heavy labelled KGG motif peptides were spiked into four background mixtures in different amounts. Mixes 3 and 4 consist of a mix of E. Coli and human proteins. Only the human proteome was utilised as the global proteome. Figure adapted from [471]

As input to our workflow, we used the MSstatsPTM_Summarized.rda object provided on MassIVE. In the FeatureLevel data part of the object, the spiked-in peptides were not annotated and were irretrievable because the heavy peptides can also be present as their light counterparts. However, they were annotated in the ProteinLevel part. Hence, we could not use the low-level data, and had to start from the data that had already been pre-processed and summarised to PTM (for the PTM dataset) and Protein level (for the global profiling dataset) by MSstatsPTM, thus omitting step 4 and 5 from the workflow. We therefore could not assess our entire workflow based on these data, and moreover do not know which preprocessing steps were conducted.

We employed various methods to analyse this dataset. Our primary approach was the msqrob2PTM workflow as described in the workflow section, as well as the normal MSstatsPTM workflow. We also assessed the differential abundance of the PTMs with the standard msqrob2 workflow: *DPA-nonNorm*, which does no normalisation and hence skips step 3 of the standard workflow entirely and *DPA*, which only applies median centring in step 3.

Because we know the ground truth of this dataset, we can again use the same metrics to assess the performance. Here, we also make ROC (tpr-fpr) curves. Furthermore, the log fold changes estimated by msqrob2PTM and MSstatsPTM were used to generate boxplots showing the observed and expected FCs for each mixture. For MSstatsPTM, the log fold changes were derived from the MSstatsPTM_Model.rda object and Spike-in_Vizualization.Rmd contained R code for the boxplots. Both these files were found on MassIVE RMSV000000669.

7. Ubiquitination dataset

Details of the experimental set-up can be found in reference(24). The dataset itself is available on MassIVE (25) as MSV000078977.

The dataset consists of four conditions: carbonyl cyanide 3-chlorophenylhydrazone (CCCP) treatment, USP30 overexpression (USP30-OE), a combination of both (Combo), and a control group. Per condition, two biological replicates with two technical replicates each were generated. All pairwise comparisons were tested using msqrob2PTM.

This dataset has also been used in the MSstatsPTM paper, hence, we can compare our results to theirs for a biological case with unknown ground truth. As input to the msqrob2PTM workflow, we used the usp30_input_data.rda object found in the MassIVE MSstatsPTM analysis container RMSV000000358, which was also used as input to the MSstatsPTM workflow. This ensures compatibility of the results with those in the MSstatsPTM paper. In this container, the analysis file MSstatsPTM_USP30_Analysis.R can also be found, which was used for the MSstatsPTM results.

All steps of the workflow were followed as described above. The normalisation step made use of the available PTM dataset, given the lack of a non-enriched counterpart. Because each condition consists of two biological replicates which in turn consists of two technical replicates, we used the *msqrob* function with a mixed model as input.

The results of both analyses were used to generate line plots with input as well as our normalised PTM level-data and the estimated effects for each condition. The detailed model results in the MSstatsPTM model object allowed us to inspect the model output for each PTM and protein as well as those for PTM upon correction for protein.

8. Phospho dataset

The human phosphorylation datasets consist of 47 samples from condition A and 43 from condition B. Ethical approval for the use of human tissue was obtained from the Ethics Commission (EC) of the University Medical Center Göttingen (2/8/18 AN) and the EC of the Technical University Munich (145/19 S-SR). Two aliquots were processed for each sample: one dedicated to total proteome analysis, and the other one to the phosphoproteome analysis. The main sample preparation steps were identical for proteomics and phosphoproteomics apart from the additional phosphopeptide enrichment step. Briefly, MeOH precipitation was performed on all samples and protein pellets were resuspended with 0.1% RapiGest™ surfactant (Waters). Either 20 µg (proteomics) or 100 µg (phosphoproteomics) of samples were subjected to overnight trypsin/lysC (Mass Spec Grade mix, Promega, Madison, USA) digestion at 37°C with an enzyme:protein ratio of 1:25. Peptide samples were then incubated for 45 minutes at 37°C and centrifuged to remove RapiGest.

For total proteome analysis, collected supernatants were loaded on an AssayMAP Bravo (Agilent) for automated peptide clean-up using C18 cartridges. Desalted peptides were injected on a nanoAcquity UltraPerformance LC® (UPLC®) device (Waters Corporation, Milford, MA) coupled to a Q-Exactive Plus mass spectrometer (Thermo Fisher Scientific, Waltham, MA) and analysed using Data Dependent Acquisition (DDA).

For phosphoproteomics, collected supernatants were loaded on an AssayMAP Bravo (Agilent) for automated Fe(III)-NTA phosphopeptides enrichment. Enriched samples were then analysed on a nanoAcquity UPLC devise (Waters) coupled to a Q-Exactive HF-X mass spectrometer (Thermo Scientific, Bremen, Germany) using DDA.

Generated raw data files were searched against a database containing all human entries extracted from UniProtKB-SwissProt (25/08/2021, 20 339 entries) using MaxQuant (v.1.6.17). The minimal peptide length required was seven amino acids and a maximum of one missed cleavage was allowed. For proteomics data, methionine oxidation and acetylation of proteins' N-termini were set as variable modifications and cysteine carbamidomethylation as a fixed modification. For phosphoproteomics

data, serine, threonine and tyrosine phosphorylations were added as variable modifications. For protein quantification, the "match between runs" option was enabled. The maximum false discovery rate was set to 1% at peptide and protein levels with the use of a decoy strategy. Intensities were extracted from the Evidence.txt file to perform the following statistical analysis. All seven steps of the workflow were performed. The dataset can be found on PRIDE (PXD043476). Further technical details can be found in the supplementary information.

9. Mock analyses

For the phospho dataset a mock analysis was included, that is an analysis where we only take one treatment arm of the data, so none of the PTMs (peptidoforms) are expected to be differential. We then assign the samples at random to a mock treatment with two levels and assess differential usage between the two conditions (mock vs control). In this way, correct control of the type I error by the statistical method can be assessed. Indeed, every PTM that is called as differentially abundant is a false positive. Hence, we expect the method to return uniform p-values.

From the phospho dataset, only the samples from factor 1 condition B, and factor 2 condition y were withheld, i.e. 26 samples. Upon step 4, 13 out of the 26 samples were randomly assigned to condition "mock", the other 13 were assigned as condition "control". Step 5 was then carried out by testing for a condition effect and the calculated p-values were retained. The randomisation to the mock treatment and step 5 in the analysis was repeated 5 times and a histogram was made for the p-values for each mock simulation.

This mock analysis was done for different workflows: we assessed the effect of using robust regression in the modelling step, the use a non-enriched counterpart for normalisation and normalisation based on the enriched dataset, itself. Moreover, we conducted the analysis both on peptidoform as well as PTM-level.

Results

The performance of our novel PTM and peptidoform msqrob2 based workflows will be compared to MSstatsPTM based on computer simulations, the spike-in dataset, the ubiquitination and phospho datasets.

Computer simulations

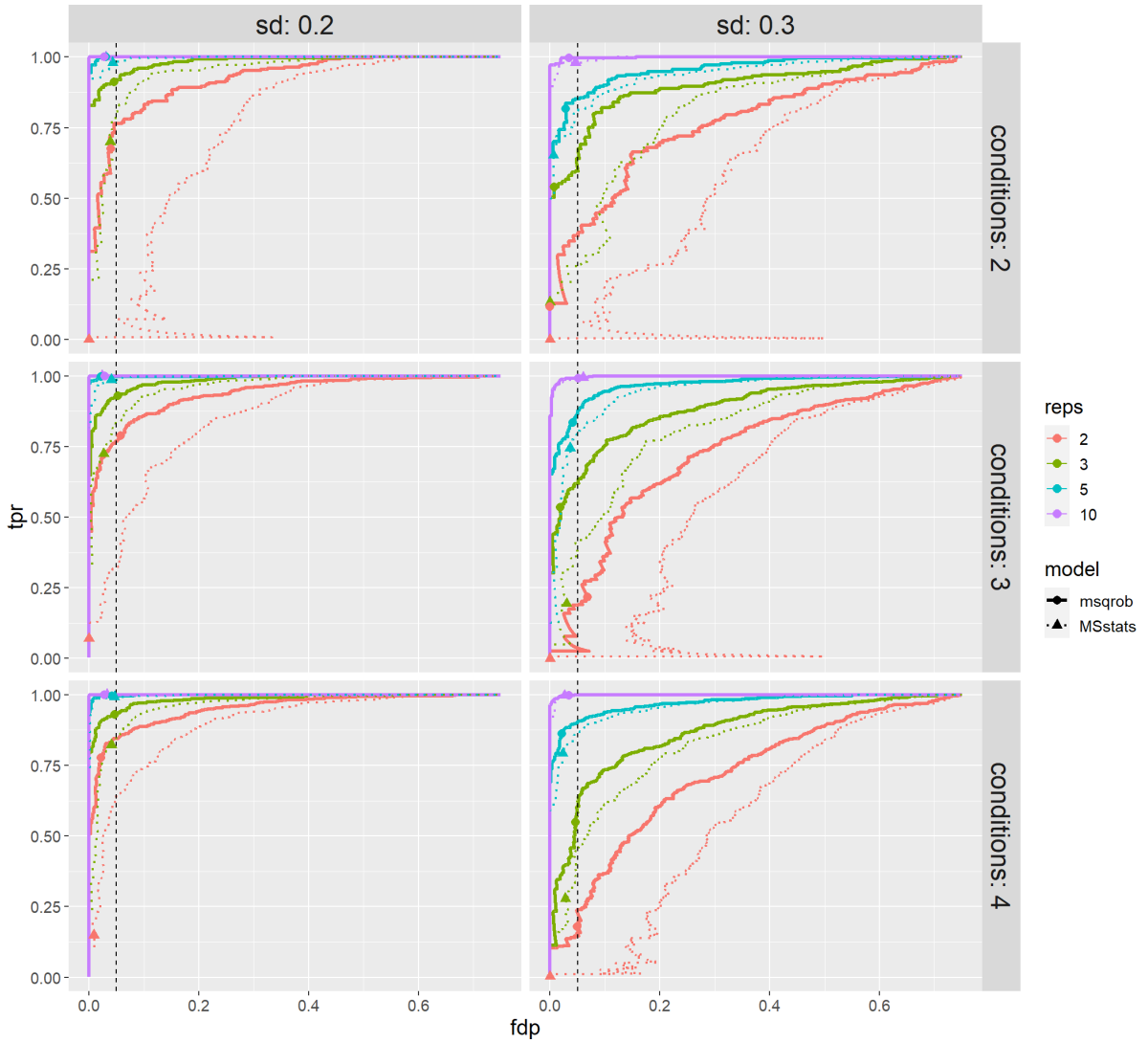
PTM-level

We first evaluated our method using the two computer simulations mentioned above. The first simulation consisted of 24 "perfect" datasets with no missing data and ten distinct peptidoforms carrying a specific PTM. Half of the datasets were simulated with a standard deviation of the difference in log-intensities between modified and unmodified peptidoforms of 0.2, the other half had a standard deviation of 0.3. The datasets differ in the number of replicates as well as in the number of conditions.

Figure 3 shows the true positive rate (tpr, the fraction of the truly differentially abundant PTMs picked up by the method) - false discovery proportion (fdp, the fraction of false positives in the total number of PTMs flagged as differentially abundant) curve for simulation 1 for all 24 datasets. As expected, both msqrob2PTM and MSstatsPTM perform better in datasets with lower variability and/or a higher number of replicates. Indeed, the true positive rate or sensitivity is higher for the same level of the false discovery proportion when the number of repeats increases while keeping the sd fixed (or when reducing the sd while keeping the number of repeats fixed). msqrob2PTM (solid line) clearly outperforms MSstatsPTM in all datasets (dotted line). Furthermore, MSstatsPTM in particular seems

to have issues when the number of replicates is low. Indeed, in four out of six datasets with two replicates, the dotted line immediately veers right instead of up, indicating that non-DU PTMs are returned among the most significant features. This particularly affects datasets with higher variation (sd 0.3). msqrob2PTM, however, does not suffer from a poor ranking of the PTMs for these four datasets and is still able to report (a few) true positive results at the 5% FDR level. Moreover, the fdp at the 5% FDR level for msqrob2PTM is close to 5% for most datasets, indicating a good control of false positives.

Figure 3: True positive rate (tpr) - false dicovery proportion (fdp) plots for datasets simulated under first scenario (no missingness). msqrob2PTM (full lines) is compared to MSstatsPTM (dotted lines). Observed fdp at a 5% FDR cut-off is denoted by dots for msqrob2PTM and by triangles for MSstatsPTM. msqrob2PTM uniformly outperforms MSstatsPTM for all datasets.



Indeed, MSstatsPTM is less sensitive, i.e. its tpr-fdp curve is always below the corresponding one of msqrob2PTM.

Figure 4 shows the tpr - fdp curves for simulation 2 for all 24 datasets. As expected, the higher number of missing values induces a slight drop in performance overall. However, for the larger sample sizes the performance remains very good for msqrob2PTM. Again, msqrob2PTM uniformly outperforms MSstatsPTM and the fdp is close to 5% when adopting a 5% FDR threshold. For two datasets, we see that the far end of the tpr-fdp curve for msqrob2PTM veers straight up (two conditions, two replicates sd 0.2 and sd 0.3), which reflects msqrob2's inability to fit the models for a number of PTMs. This happens because these PTMs have too few observations to fit the models due to the missingness introduced in this simulation scenario.

For further comparison, ROC curves (true positive rate vs false positive rate) are shown in Supplementary figure 1 and 2. These plots give less weight to a few top-ranked false positives. Again, these ROC curves demonstrate superior msqrob2PTM performance.

In supplementary table 1 and 2, the performance metrics (false positive rate, sensitivity, specificity, precision and accuracy) that were reported in the MSstatsPTM paper are also given for all datasets for comparison.

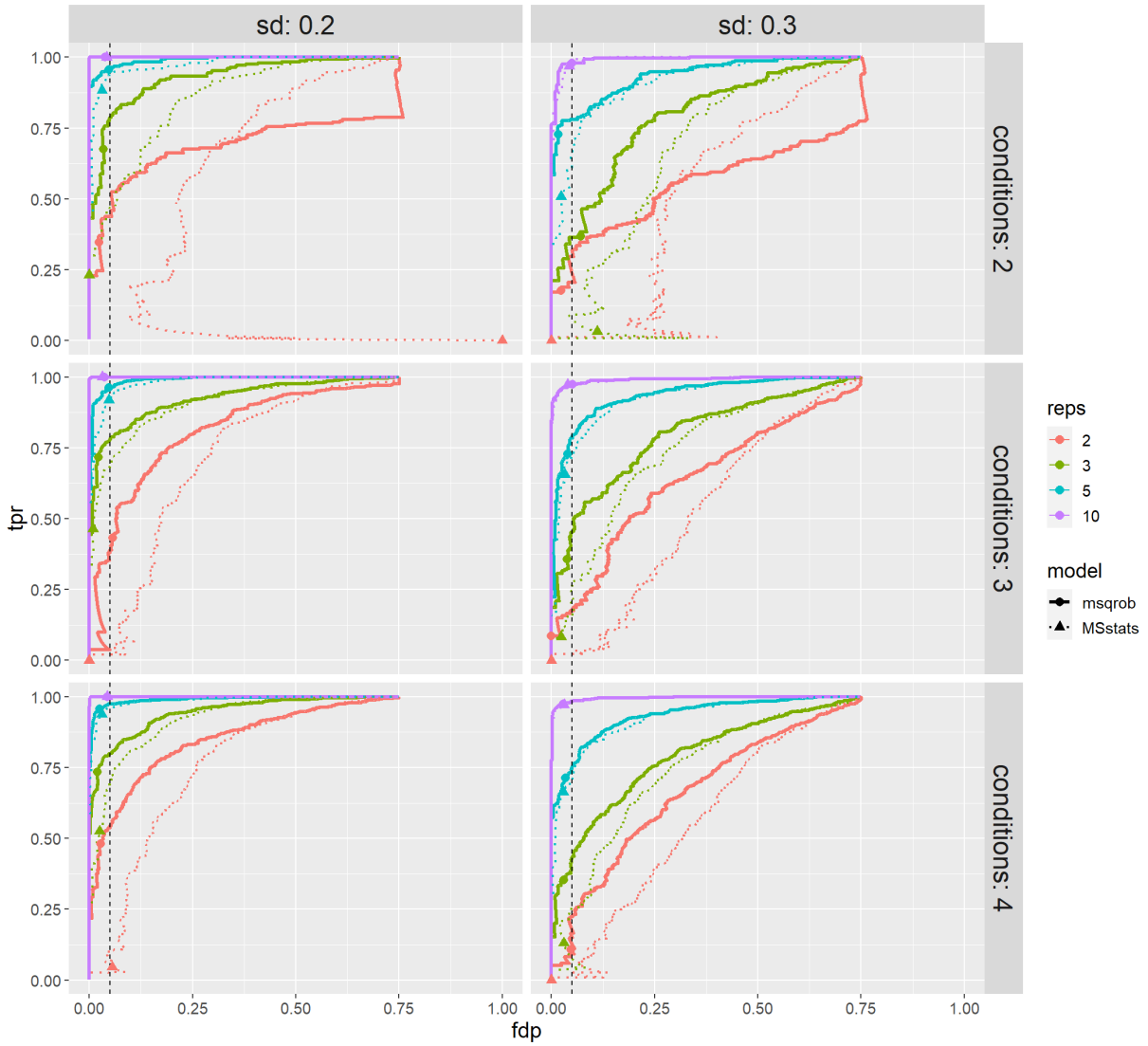


Figure 4: tpr-fdp plot for datasets simulated under second scenario (with missingness). msqrob2PTM (full lines) is compared to MSstatsPTM (dotted lines). Observed fdp at a 5% FDR cut-off is denoted by dots for msqrob2PTM and by triangles for MSstatsPTM. Here again, msqrob2PTM outperforms MSstatsPTM for all datasets.

Peptidoform level

Our msqrob2PTM workflow can also infer on differential usage at the peptidoform level, which we consider to be very important. Indeed, not all peptidoforms that carry the same PTM will necessarily follow the same abundance pattern. Therefore, it can occur that a significant effect at the PTM-level stems only from one or a few associated peptidoforms while the other associated peptidoforms remain unchanged between conditions. This might indicate that the underlying biology is not only affected by a single PTM, but rather by a combination of PTMs and/or sequence variation. We thus recommend adding a peptidoform analysis by default to the overall workflow.

Peptidoform level information was available in both simulations, hence the performance of our method can be evaluated at this level as well. The peptidoform level tpr-fdp plots are given in figures 5 and 6, and the underlying data in supplementary table 3 and 4. These show that msqrob2PTM also performs well on the peptidoform level and maintains good control of false positives. However, on peptidoform level, the method performance seems to be more affected by a lower number of replicates, increased variability, and missingness. This can be expected as there is inherently less information, but more variation, present at the peptidoform level. This variability is reduced by averaging over peptidoforms when summarising the data to the PTM level. However, because PTMs are not directly quantified, but averaged out over peptidoforms, they can lead to more ambiguous results.

Note that, as MSstatsPTM does not offer a peptidoform level analysis, no comparison could be included for this workflow.

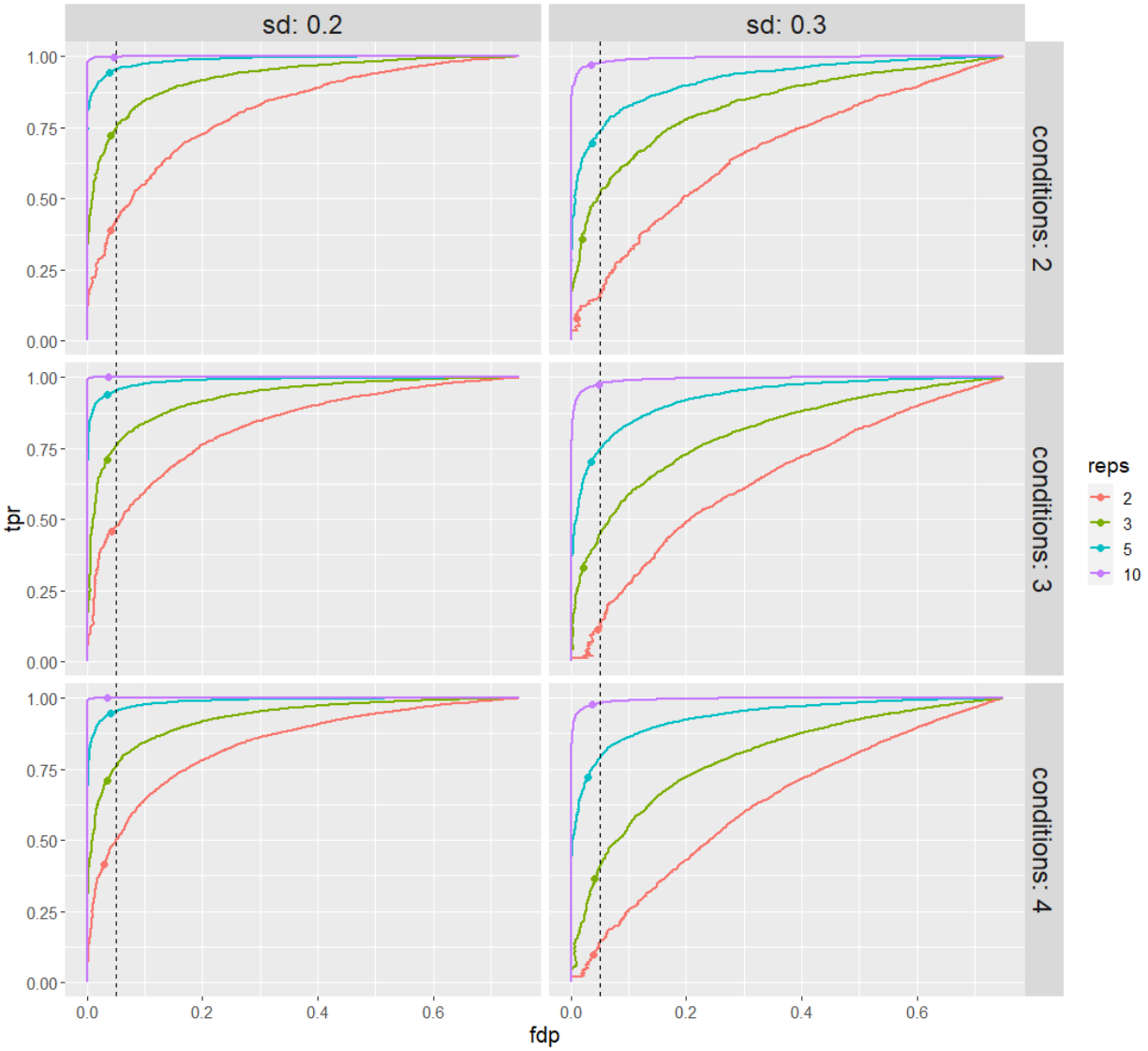


Figure 5: tpr-fdp plot for datasets simulated under the first scenario (no missingness). Performance of msqrob2PTM is assessed at peptidoform level. Observed fdp at a 5% FDR cut-off is denoted by dots.

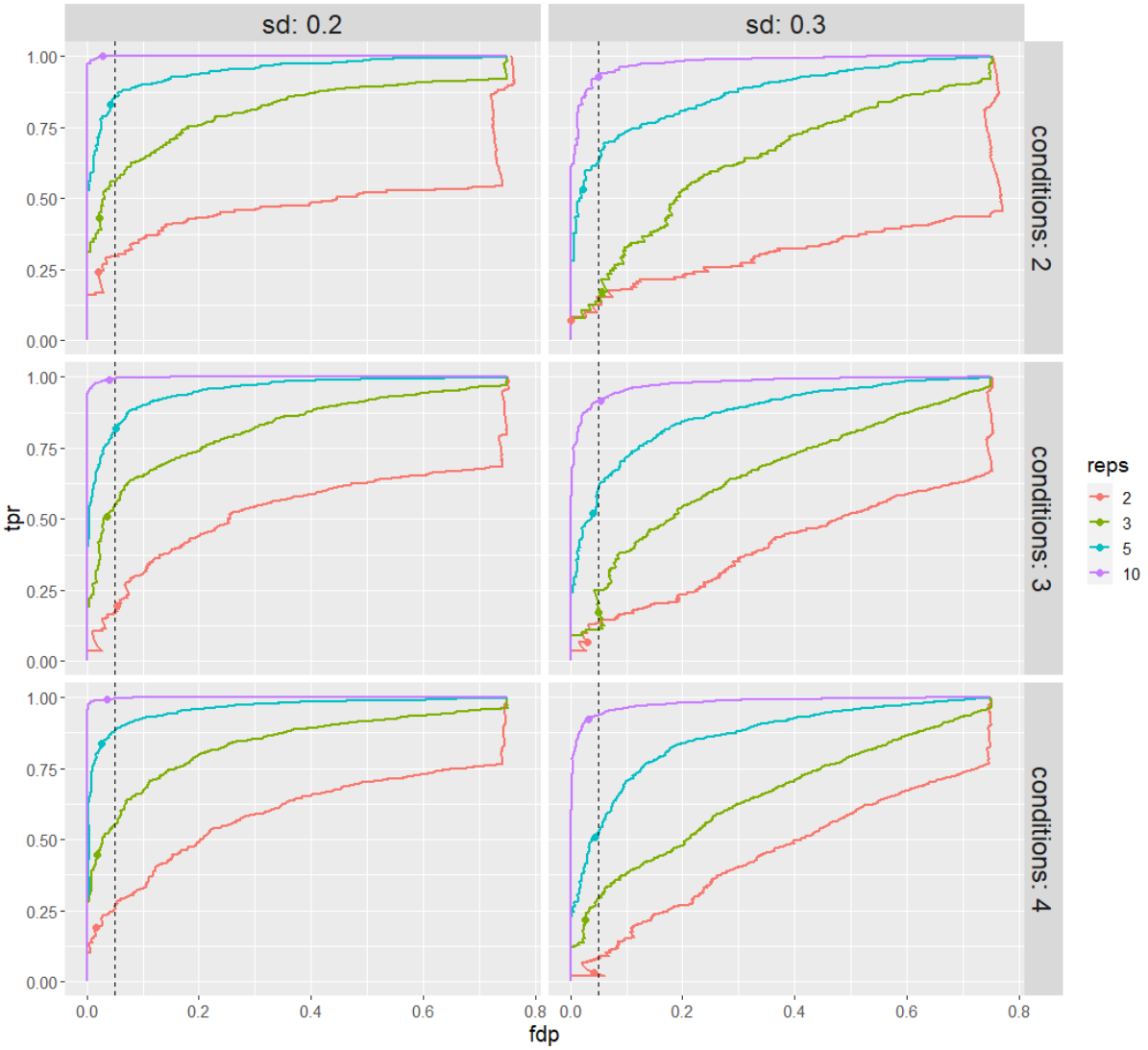


Figure 6: tpr-fdp plot for datasets simulated under the second scenario (with missingness). Performance of msqrob2PTM is assessed at peptidoform level. Observed fdp at a 5% FDR cut-off is denoted by dots. For datasets with only 2 or 3 replicates, the method starts to suffer from lack of information, making it harder to report significant peptidoforms, especially for datasets with sd 0.3.

Biological spike-in dataset

The design of the spike-in dataset (see also figure 2) is suboptimal to assess the performance of methods inferring differential PTM usage. This is because the spiked-in peptides and their corresponding protein abundance in the background proteome are not correlated as they would be in real experiments. Indeed, the latter does not contain the actual parent proteins of the spike-in peptides. Moreover, the E.coli proteins in mixes 3 and 4 induce loading differences present across the samples (see also supplementary figure 3), which brings additional normalisation issues. We illustrate these issues using ROC curves that compare the performance of different approaches: differential PTM abundance by adopting a conventional msqrob2 workflow directly on the summarized PTM-level intensities without normalisation (DPA-NonNorm), the same workflow upon normalisation with the median peptidoform log-intensity (DPA), the default workflow for msqrob2PTM (default msqrob2PTM workflow assessing DPU), and MSstatsPTM (default MSstatsPTM workflow) (figure 7). Every pairwise comparison between mixes is shown. Because all methods report many false positives for this dataset, the tpr-fdp plots quickly became unreadable (see supplementary figure 4).

When comparing mix 4 to mix 1 (mixmix4), the log2FC after adjustment should be 0, hence, no method should report any differential PTMs. Indeed, this comparison is an internal control, and the ROC curves are expected to lie along the diagonal. Here, *DPA-NonNorm* and *MSstatsPTM* show the largest deviations from the diagonal.

In the other comparisons, DPA always outperforms the other methods. Note that DPA assesses differential PTM abundance rather than differential usage as it does not normalise for parent protein intensity. This superior performance of the DPA method as compared to the DPA-NonNorm method indicates that it is very important to correct for technical variability resulting from the experimental design, i.e. the loading differences, rather than correcting for parent protein abundance. In the mix 2 vs mix 1 (mixmix2) and mix 4 vs mix 3 comparison, DPA-NonNorm also performs very well, because in these comparisons, the adjusted and unadjusted fold changes are the same. However, the loading differences for the other comparisons cause a breakdown of DPA-NonNorm. MSstatsPTM and msqrob2PTM always have a lower performance than DPA, but never break down. For the mix 2 vs mix 1 and the mix 4 vs mix 3 comparisons, MSstatsPTM performs slightly better than the default msqrob2PTM workflow, while the latter performs better in the remaining three comparisons. The decrease in performance by msqrob2PTM as compared to DPA can be explained by the increase in variability that is introduced in the workflow by subtracting the unrelated “parent protein intensities” from the spiked-in peptidoform intensities. In other words, the design is not suited to benchmark the performance of methods developed to quantify differential peptidoform usage. However, the design is useful for assessing the performance of methods that quantify differential PTM abundance. This can easily be obtained with standard msqrob2 workflows, but is not returned by default by MSstatsPTM. However, because the msqrob2 suite builds upon the QFeatures architecture, the results of a DPA and DPU workflow can both be stored in the same object, thus providing more transparency and reproducibility across the workflows.

For completeness, we also plotted the log2 fold changes for all PTMs in supplementary figures 5 and 6, which illustrate that both msqrob2PTM as well as MSstatsPTM provide good estimates for these.

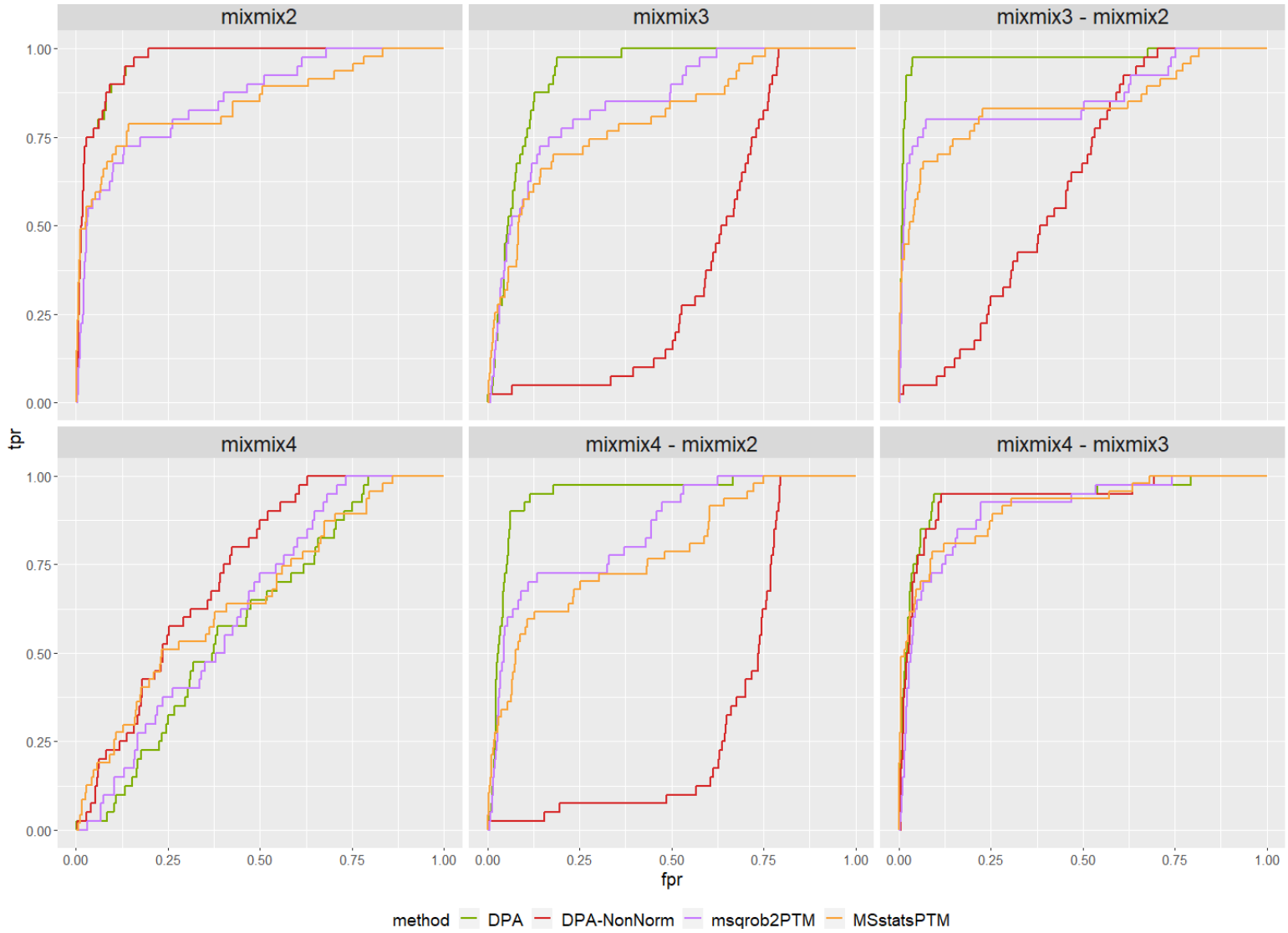


Figure 7: ROC curves of the different approaches for all pairwise comparisons. Mix 4 vs 1 (mixmix4) serves as internal control, thus the curves should follow the diagonal as closely as possible. DPA performs very well in all comparisons and outcompetes all other methods. DPA-NonNorm has good performance in the two comparisons where adjusted and unadjusted fold changes are the same, but breaks down for the other comparisons. The performance of MSstatsPTM and msqrob2PTM (the default differential PTM usage workflow) is similar, with performance dependent on the comparison being made.

Ubiquitination dataset

msqrob2(PTM) is capable of handling more complex designs that require mixed model analysis, as well as datasets that lack a non-enriched version of the dataset. These two aspects apply to the ubiquitination dataset. Note that this is an experimental, biological dataset, and therefore does not come with a known ground truth.
Despite the two abovementioned complexities, the standard msqrob2PTM workflow could find differentially abundant ubiquitin sites in most comparisons, except for the USP30_OE vs control comparison. However, table 2 shows that msqrob2PTM generally reports much fewer significant PTMs than MSstatsPTM.

Contrast	MSstatsPTM	Msqrob2PTM
Combo vs Ctrl	424	30
CCCP vs Ctrl	359	12
USP30_OE vs Ctrl	40	0
Combo vs CCCP	31	1
Combo vs USP30_OE	407	24
CCCP vs USP30_OE	364	13

Table 2: The number of significant PTMs reported for each contrast for both methods.

Upon closer inspection of the PTMs reported as significant by MSstatsPTM, it was discovered that this large discrepancy can be explained by several reasons.

First, both methods have a different way of dealing with missing data. Upon inspecting multiple line plots, we observed PTMs that were flagged as significant by MSstatsPTM despite having only one bio-repeat, or even only a single data point available in one of the conditions. In figure 8, for instance, line plots are shown for two PTMs that are significant in MSstatsPTM when comparing the combination condition (Combo) *versus* the control condition (Ctrl), but not in msqrob2PTM. Notably, PTM O00154_K205 only presents PTM information for the first biological replicate, while PTM O00159_K0578 contains just one data point within the entire control condition. For these features, msqrob2 therefore did not return a model fit.

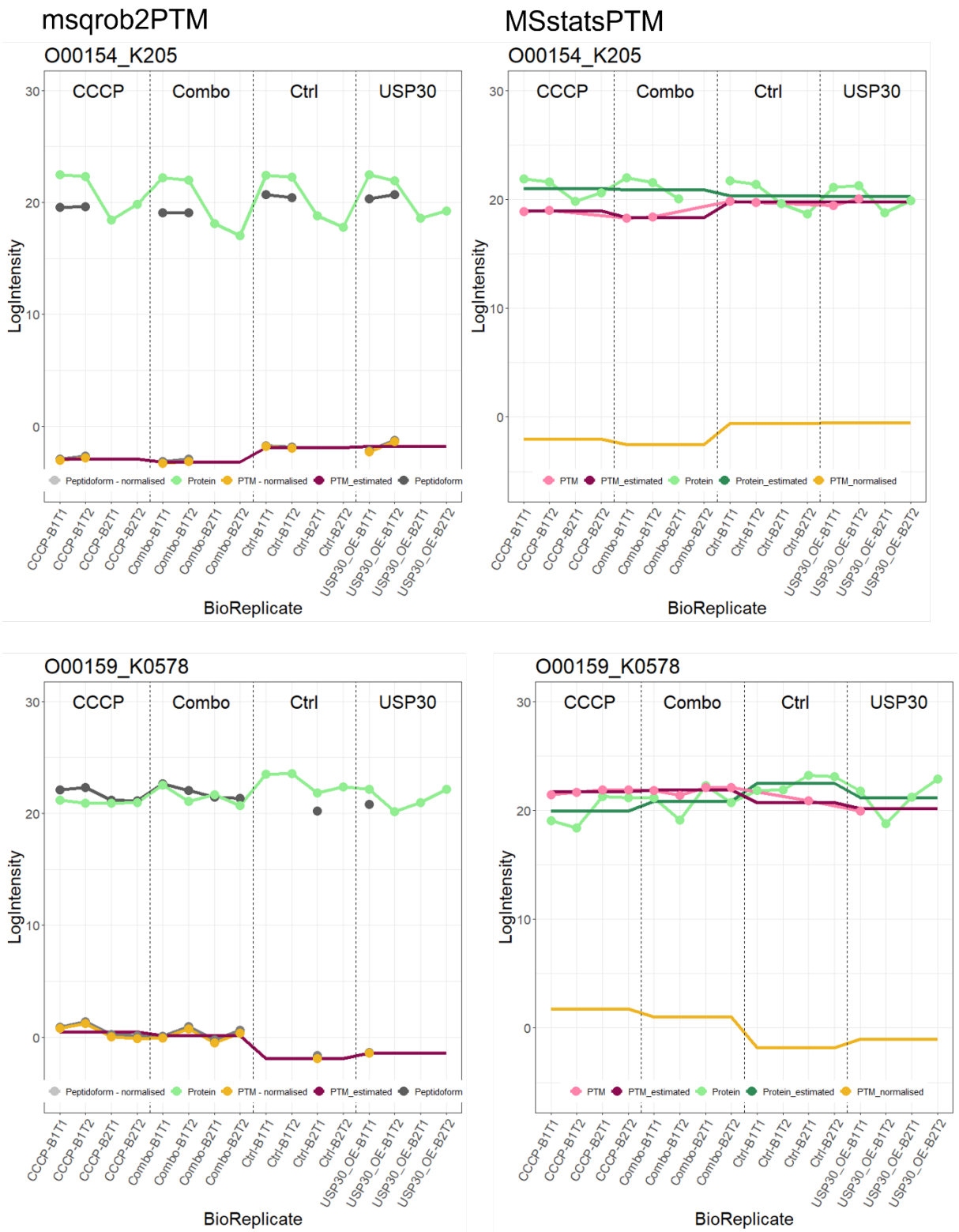


Figure 8: line plots displaying estimated log_2 intensity values of the PTM (dark pink) for each sample, its normalised intensity values (yellow), log_2 intensity values of its parent protein (green), for MSstatsPTM estimated log_2 intensity values of that parent protein (dark green), and for msqrob2PTM, log_2 intensity values of the peptidoforms (grey) on which the PTM occurs. On the left, line plots for PTM O00154_K205 and O00159_K0578 for msqrob2PTM, on the right for MSstatsPTM. Both PTMs were deemed significant by MSstatsPTM when comparing the control condition to the combination condition (combo), but not by msqrob2PTM. O00154_K205 only contains intensity information for bio replicate B1. O00159_K0578 only has 1 associated intensity value in the control condition. Hence, both of these PTMs contain too few datapoints for msqrob2PTM to determine significance.

When examining the results more closely, we noticed that MSstatsPTM uses three different models to fit the data and that the model choice is based on the available data points for each PTM (see UbiquitinationBioData exploration of results file on https://github.com/statOmics/msqrob2PTMpaper for detailed examples), i.e. a full mixed model was employed when no data was missing, using a fixed effect for group and random effects for subject (1 | SUBJECT) and subject x group (1 | GROUP:SUBJECT), as soon as a single data point is missing, the (1 | GROUP:SUBJECT) term is dropped, and when data is missing for one of the bio repeats in all conditions, a linear model is employed with only a fixed group effect. This adaptability to missing data comes with a price, however. Notably, the second model, without the (1 | GROUP:SUBJECT) term, ignores the between bio repeat variability. Indeed, bio repeat 1 in the control group is not the same as bio repeat 1 in the combination group. However, they are treated as such, resulting in underestimated standard errors.

Across comparisons, 15-27% of PTMs deemed significant were modelled with an incorrect mixed model (% differs according to comparison). Moreover, 44-75% of significant PTMs were modelled using a linear model, which represents features for which msqrob2 does not fit any model at all because biological repeats are lacking. Moreover, when examining the significant PTMs together with their parent proteins, it became apparent that for most features the PTM and protein intensities were modelled with a different model. This can lead to artifacts such as shown in figure 8 (top panels), where the protein data contains information about only one of the two bio repeats, but is still used to make the adjustment for the other bio repeat! To avoid these ambiguities, we conducted an MSstatsPTM-like analysis while enforcing the use of the full mixed model. Only PTMs with associated parent proteins were included in the analysis. Subsequently, the full mixed model was applied to both the PTM and protein-level data. The adjustment for protein abundance followed the standard MSstatsPTM procedure, and the resulting p-values were adjusted using the Benjamini-Hochberg method. Using the native MSstatsPTM implementation the “CCCP” vs “Ctrl” comparison identified 359 significant PTMs. However, when solely employing the full mixed model, only 55 PTMs remained significant, which is in line with our msqrob2 results.

Second, the two methods employ distinct conceptual approaches. In msqrob2PTM, within-sample normalisation according to protein level abundance is performed first, followed by statistical analysis. MSstatsPTM, however, uses the modelled PTM and protein results for normalisation, ignoring the inherent biological correlation between PTMs and their parent proteins within a sample. Analysing these separately can sometimes generate ambiguities. Figure 9 illustrates this issue, demonstrating a PTM that was flagged as significant for the “Combo” vs "Ctrl” comparison by MSstatsPTM, but not by msqrob2PTM. Specifically, the peptidoform carrying PTM O60260_K369 closely mirrors the intensity pattern of its parent protein, resulting in minimal differences, and therefore no significant regulation, in PTM intensities after normalisation for protein abundance in our msqrob2PTM workflow. However, as MSstastPTM first fits models to the PTM and protein level data separately, and only afterwards uses these model estimates to correct for the difference in protein abundance, differences in PTM usage are artificially enlarged, leading to a significant PTM according to MSstatsPTM in this comparison.

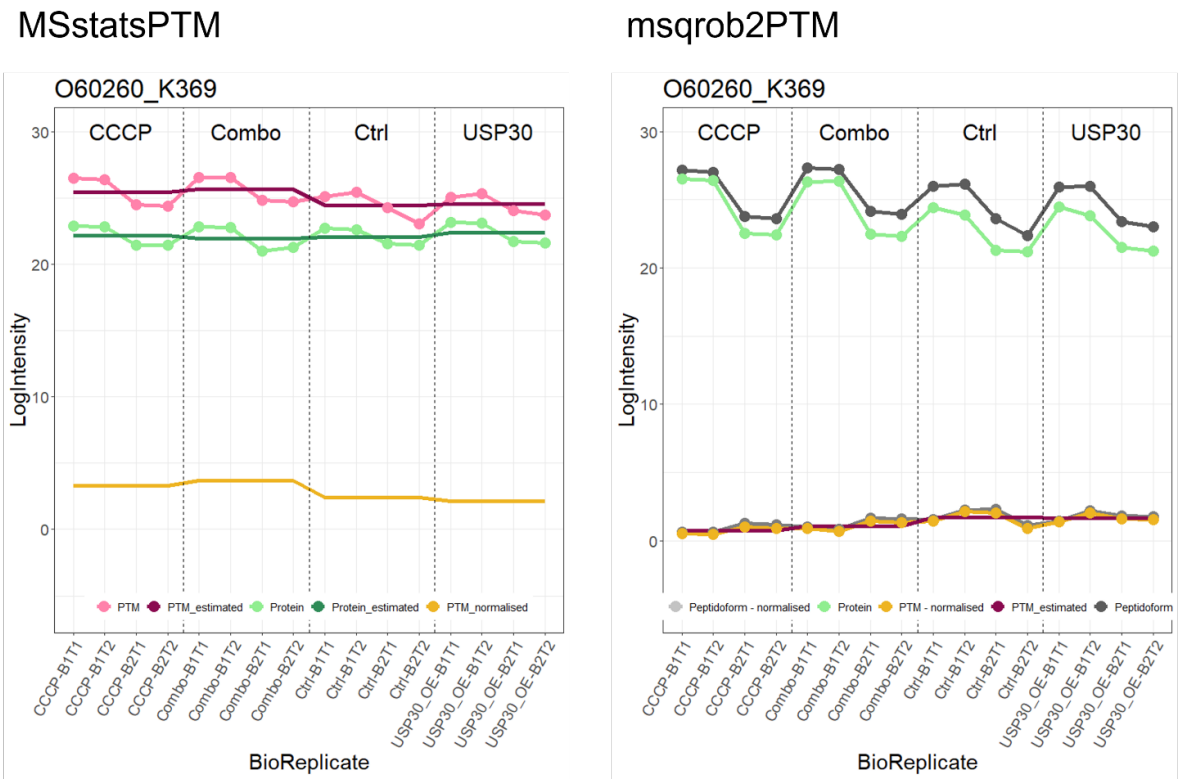


Figure 9: line plot displaying PTM log_2 intensity values (pink dotted line) or peptidoform log2 intensity values (dark grey dotted line) and log_2 intensity values of its parent protein (light green dotted line) in each sample. MSstatsPTM first fits a model to PTM (dark pink line) and to protein intensities (dark green line) to estimate average intensity in each condition. Subsequently, fitted average protein abundances are subtracted from fitted average PTM intensities to obtain average PTM abundances in each condition corrected for protein abundance (yellow line). Conversely, msqrob2PTM first normalises peptidoform intensities using parent protein abundance, resulting in a normalised peptidoform (light grey dotted line). From normalised peptidoforms, normalised PTM intensities are calculated (yellow dotted line). Estimated log2 intensity values of the PTM are depicted in dark pink. MSstatsPTM corrected PTM abundances seem to indicate differential PTM usage. Moreover, the comparison between "Combo" vs "Ctrl" is returned by MSstatsPTM as statistically significant. This, however, appears to be an artifact of MSstatsPTM as the correction for protein abundance does not account for the link between protein and PTM intensities within-sample. Indeed, when comparing "Combo" and "Ctrl" sample level intensities, the pattern at PTM-level closely follows that of its parent protein.

Phospho dataset

Two different workflows were employed for this dataset. The first workflow uses the non-enriched counterpart dataset to normalise for differences in protein abundance, while the second workflow only used the enriched dataset, also for the normalisation step. It is important to note that two distinct instrument platforms were used to analyse the total proteome and phosphoproteome samples. The chromatographic conditions were identical as well as the MS instrument geometry but two consecutive generations of Q-Orbitraps were used (Q-Exactive Plus versus Q-Exactive HF-X). This partly explains the observed heterogeneity between enriched and non-enriched datasets. Indeed, we observed a substantial proportion (approximately 25%) of proteins present in the enriched dataset that were absent in the non-enriched one. This led to some PTMs that could not be normalised, which we opted to exclude from subsequent analysis in workflow 1.

Both workflows involved testing multiple contrasts based on two factors: condition (A or B), and subset (x or y). In the first workflow (utilising both datasets), 31 unique differential PTMs were found, of which 25 phosphorylations. Most of these PTMs exhibited significant downregulation in condition A compared to B within subset y.

In the second workflow (using only the enriched dataset), fourteen unique significant PTMs were identified, of which eight phosphorylations. The majority of phosphorylations showed significant differential usage between condition A and B within subset y and/or exhibited significant differential usage between condition A and B averaged over subsets x and y. Supplementary tables S5 and S6 provide detailed results.
Interestingly, the results differ between the two workflows. Of the 31 PTMs identified in workflow 1, ten were also found in workflow 2.

Instead of solely focusing on significant PTMs, our method is capable of detecting differentially used peptidoforms as well. For this dataset, the first workflow detected twelve peptidoforms as differentially abundant, predominantly showing downregulation in condition A for subset y.
In the second workflow, which lacked a global profiling dataset, seven significant peptidoforms were detected across the different comparisons. LPIVNFDYS[Phospho (STY)]M[Oxidation (M)]EEK was picked up as DU by both workflows and is particularly interesting, because both PTMs present on this peptidoform are also returned as significant in the differential PTM usage analysis. Hence, one of the PTMs might have been detected as differential because the other PTM is also present on the same peptidoform, potentially influencing its significance upon averaging with the remaining peptidoforms carrying this PTM. To assess the contribution of different peptidoforms to a single PTM, line plots can be used to visualise both the PTM intensities across the samples as well as the intensities of its contributing peptidoforms. Figure 10 illustrates this issue. Indeed, the top panel shows a phosphorylation that occurs in two peptidoforms, the bottom panel shows an oxidation that also occurs on one of these peptidoforms. The peptidoform with both modifications was significant, while the second peptidoform that did not carry the oxidation was not significantly DU. The intensity for the phosho-PTM is obtained upon summarisation over both peptidoforms, and was reported significant when assessing the data at the PTM-level. However, the significance of the phospho-PTM might be an artifact triggered by the presence of additional oxidation in one of its underlying peptidoforms.

Figure 10: Line plots of normalised intensity values per sample for significant peptidoform (LPIVNFDYS[Phospho (STY)]M[Oxidation (M)]EEK) and its corresponding PTMs for the phospho dataset. At the top, the significant peptidoform is depicted in pink. In green is the PTM occurring on that peptidoform, in this case phosphorylation. In grey any other peptidoform carrying that same PTM, and in yellow, the PTM intensity value as estimated by the model. The PTM is represented by two peptidoforms that roughly follow the same pattern, resulting in a PTM that resides in the middle. At the bottom we see the other PTM occurring on that peptidoform, the oxidation. No other peptidoform carries that same modification, resulting in perfect overlap between the line of the significant peptide and that of the PTM. Here, it is possible that the oxidation is only significant because the phosphorylation is. Indeed, the driving force of the significance of this particular peptidoform could be coming from the phosphorylation (which has two associated peptidoforms). Note that, while these particular line plots were derived using the workflow without a non-enriched dataset, the corresponding plots from workflow 1 are extremely similar.

Some PTMs are also significant because they enable aggregating evidence over multiple non-significant peptidoforms that all have a similar expression pattern. An example of this can be seen in figure 11 for sp|P10451|OSTP_HUMAN (Phospho (STY)) 280.

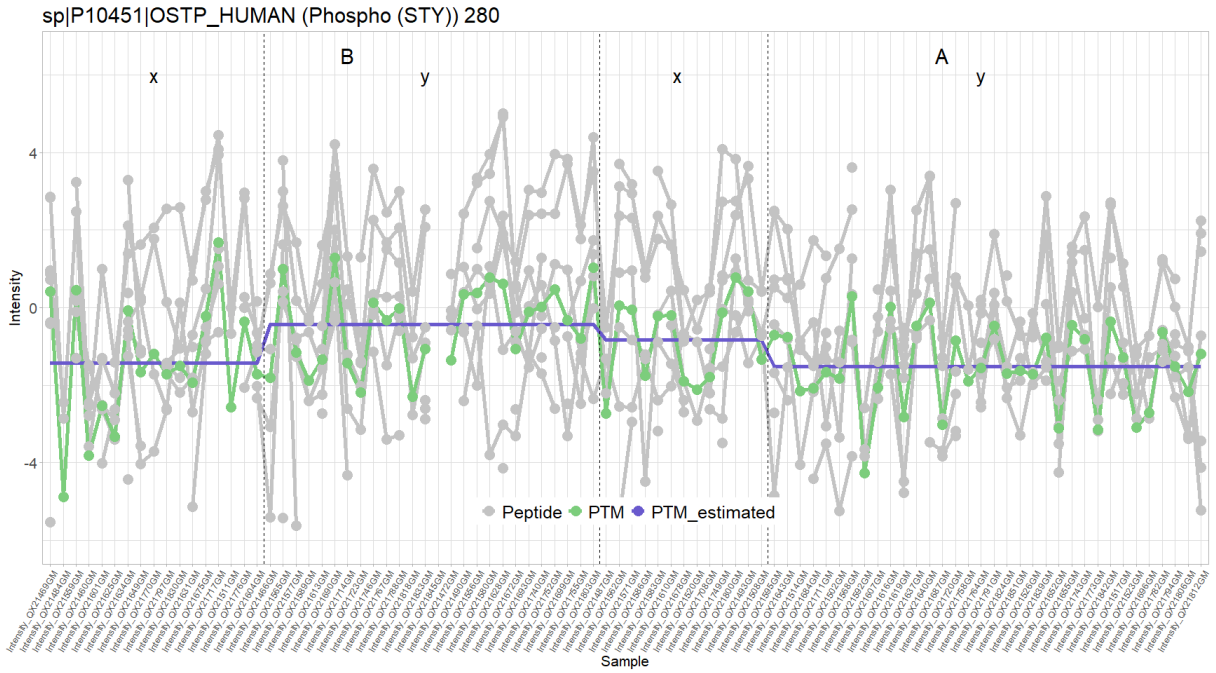


Figure 11: Line plot of normalised intensity values of significant PTM sp|P10451|OSTP_HUMAN (Phospho (STY)) 280 and its associated peptidoforms. In green the summarised and normalised intensity value of the PTM, in grey all peptidoforms (normalised) containing this PTM, in purple the PTM intensity values as estimated by the model. While none of the peptidoforms are individually significant, these all contribute to a PTM that can be picked up as differentially abundant (downregulated in condition A for samples from subset y).

Mock analyses

As the phospho datasets are biological experiments, the ground truth is unknown. Therefore, we cannot assess the performance of each method. We also do not know if the method provides reliable false positive control. To assess if our workflows provide good type I error control for the case study, we therefore perform a mock analysis. In particular, we introduce a factor for a non-existing effect, implying that all features that are returned significant upon testing for this factor are false positives. Here, we focus on subset y from condition B, so that ample samples remain. When the method provides good false positive control, the p-values upon assessing the mock effect will be uniform.

The p-value distribution for the workflow that only uses the enriched dataset is given in Figure 12. The top panels show the results for the PTM-level analysis and the bottom panels for peptidoform analysis. Both workflows with and without robust regression provide fairly uniform p-values. Supplementary figures 7-10 show similar plots for four other random mock datasets, showing consistency of performance.

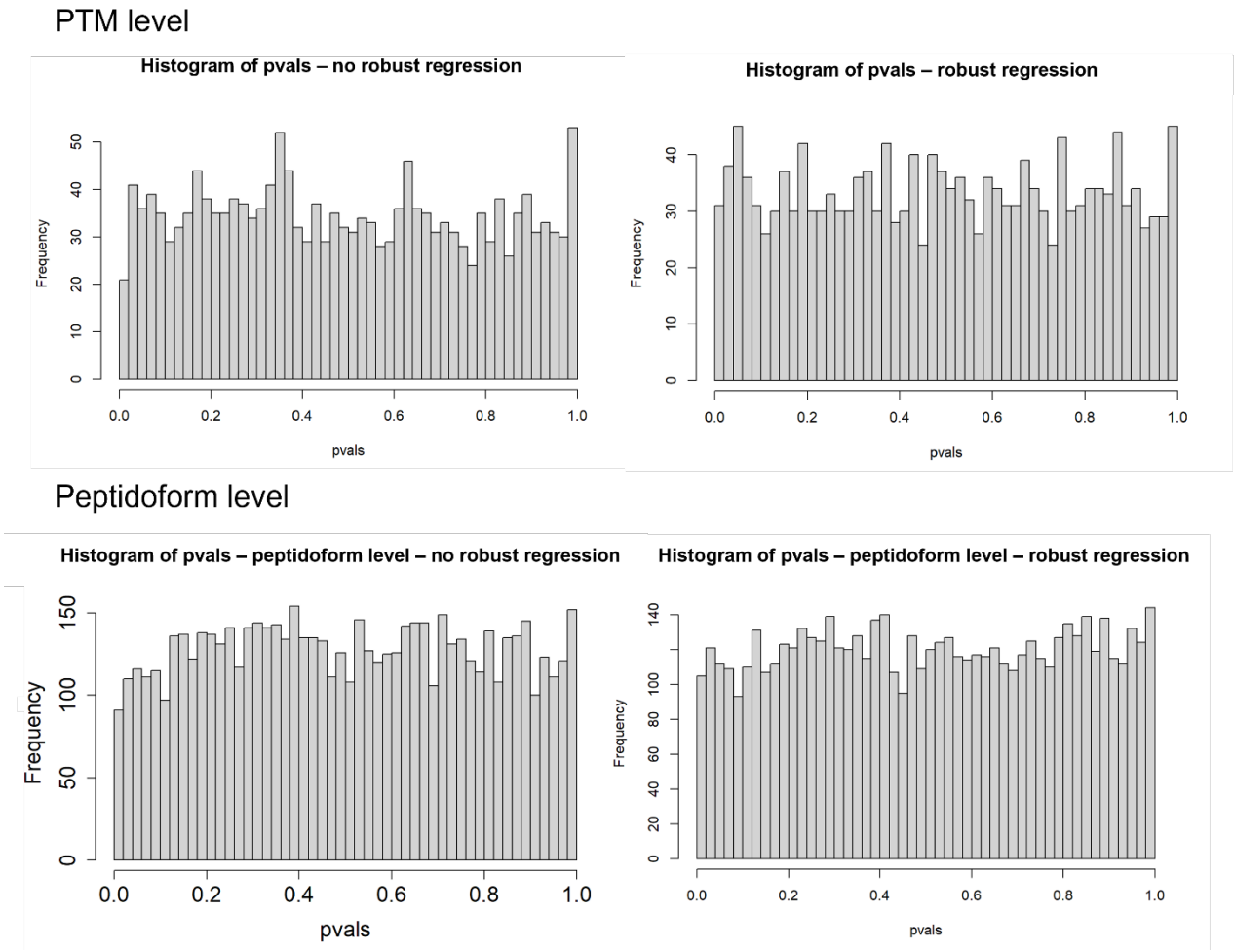


Figure 12: Distribution of p-values for mock analysis of the phospho dataset without global profiling run, for analysis on PTM level (top) as well as peptidoform level (bottom). Left panels are for workflows without robust regression in the modelling step; Right panels correspond to workflows with robust regression in the modelling step. All p-values are fairly uniform, indicating acceptable type I error control.

We did a similar mock analysis for the workflow that uses the non-enriched dataset for usage calculation (Figure 13). The workflow on peptidoform level using robust regression showed a slight increase in low p-values, which is also observed in some other random mock datasets (Supplementary Figures 11-14). The remaining workflows generated fairly uniform p-values for all random mock datasets (Figure 13 and Supplementary Figures 11-14). We therefore did not adopt robust regression for the peptidoform analysis.

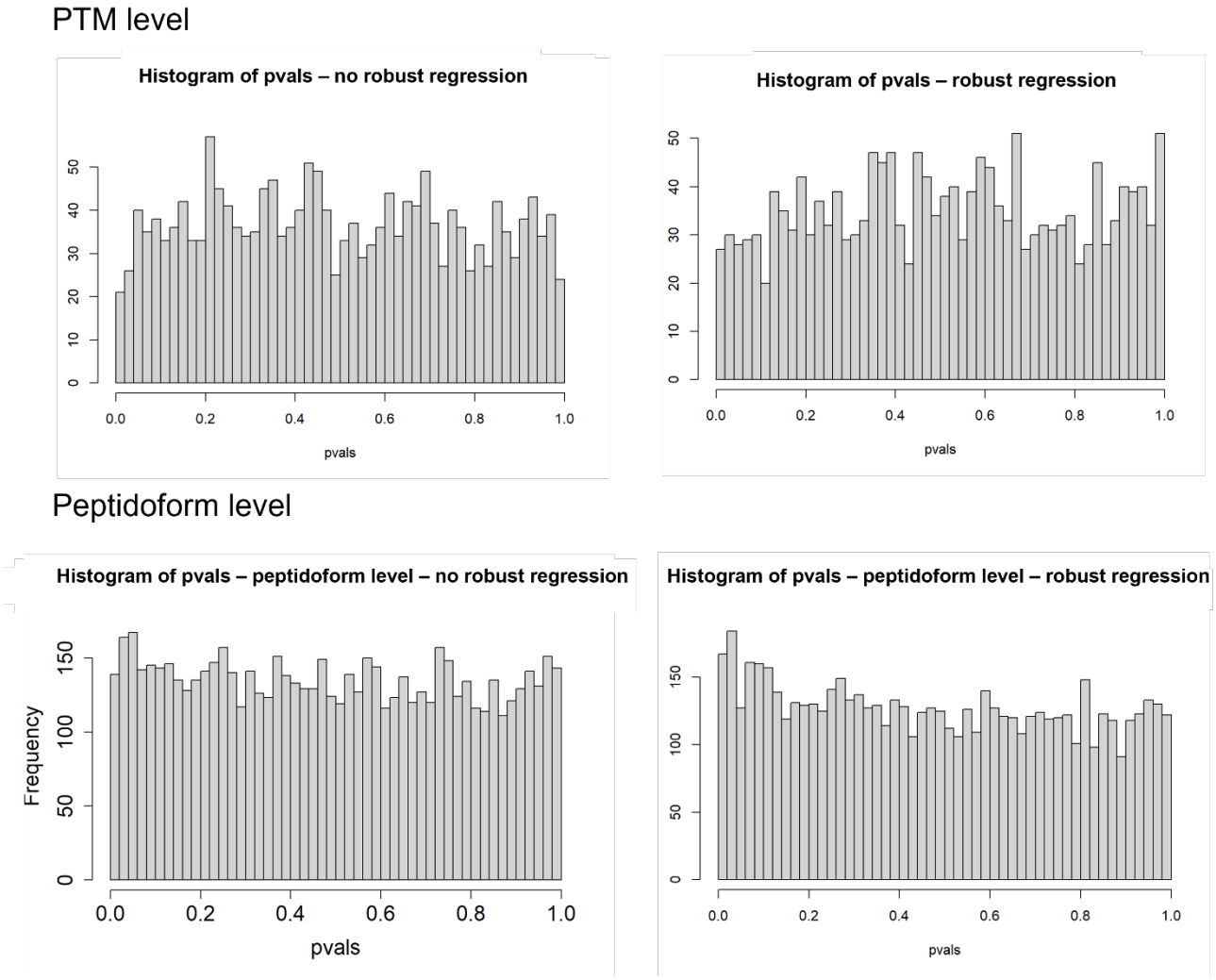


Figure 13: Distribution of p-values for mock analysis of the phospho dataset using the non-enriched dataset to estimate the usages. Results at PTM level (top panels) as well as at peptidoform level (bottom panels). Left panels are based on a workflow without robust regression; right panels on a workflow with robust regression.

Discussion

We here introduced msqrob2PTM, a novel workflow in the msqrob2 universe, designed for performing differential abundance as well as usage analysis on PTM and peptidoform level. These two analyses are distinguished by their normalisation strategies. In abundance analysis, only a normalisation to reduce technical variation is included, while the novel usage workflow incorporates normalisation against parent protein intensities. Both approaches have their relevance in PTM research. DPU enables the discovery of differential PTMs that respond differently than their parent protein. However, in certain scenarios, DPA might be of interest instead. Indeed, when an increase in total protein concentration leads to a corresponding increase in PTM concentration, there may be biological implications associated with this elevation in PTMs, regardless of whether it is driven by changes in parent protein levels or not. Therefore, the choice between DPA and DPU depends on the specific research question at hand, or they can both be performed to complement each other.

Through analysis of simulated and biological datasets, we have demonstrated that our workflows improve upon the state-of-the-art MSstatsPTM. We showed the advantage of first normalising the peptidoform intensities by the parent protein abundance before conducting the differential analysis. In this way, we can immediately model the usages as opposed to MSstatsPTM that estimates the fold changes for the PTM and protein values separately before differencing these to estimate DPU. Indeed,

the peptidoform and protein values from the same sample are correlated, which is explicitly accounted for in our DPU workflow but is ignored by MSstatsPTM. We showed for the latter method that this can lead to artifacts in the estimated fold change for some PTMs upon correction for the fold change in the parent protein. Moreover, MSstatsPTM also ignores the correlation when calculating the variance on the difference in fold change leading to incorrect inference.

Another key distinction between both packages is how they handle PTMs that cannot be fitted with the desired model. MSstatsPTM prioritises automation and aims to infer on as many PTMs as possible. However, this leads to reporting on PTMs for which the fit is based on different models and often on insufficient data to draw reliable inference on the contrast of interest. Moreover, for PTMs that lack a corresponding protein expression fold change, results are returned based on the PTM fold change alone. Hence, MSstatsPTM silently combines inference on differential usage with inference on differential abundance in one output list depending on the degree of missingness at the protein-level. In general, a standard user is not fully aware of these issues, and the subtleties of interpretation that these require. In contrast, our msqrob2PTM workflow emphasises transparency and reproducibility. While this choice may lead to some PTMs that cannot be estimated using the default workflow, it does ensure that users are fully aware of what was modelled for each PTM. Moreover, we feel that PTMs for which no results are returned due to missingness require the intervention of a skilled data analyst to develop tailored solutions to infer on differential abundance and/or usage; solutions that are moreover supported by the msqrob2 universe. Indeed, we showed that automatic approaches can lead to biased results, and especially in experiments with more complex designs.

These differences in normalisation approach and design concept elucidate the variations in performance across the different datasets that were used in our benchmark. In the simulated datasets, msqrob2PTM capitalises on the within-sample correlation between peptidoforms and proteins that is present in the data, resulting in superior performance compared to MSstatsPTM. However, in the spike-in dataset, where this correlation is absent due to its unrealistic design, the default msqrob2PTM workflow exhibits similar performance to MSstatsPTM. However, for this dataset we show that our workflow for assessing differential PTM abundance analysis uniformly outperforms both the msqrob2PTM and MSstatsPTM workflows assessing differential PTM usage. Indeed, the spike-in study is suited for assessing the performance on differential PTM abundance rather than on differential PTM usage, as the spiked PTMs were not correlated to their corresponding protein in the background. In the biological ubiquitination dataset, the high amount of missing data, and the absence of a global profiling dataset leads to a high number of PTMs that cannot be fitted with the required model. MSstatsPTM will then resort to other, simpler models that are often suboptimal or even mismatched, while msqrob2PTM will simply not return results for these PTMs, leading to a lower number of reported significant PTMs.

These datasets bring to attention a broader issue in the field, specifically the scarcity of suitable datasets for accurately assessing Differential Peptidoform Usage (DPU). When designing such experiments, it is favourable to incorporate a global profiling dataset along with an adequate number of biological replicates. This comprehensive approach not only enables a more thorough evaluation of DPU but also enhances statistical power, yielding more reliable and robust results. Indeed, the approach benefits from multiple replicates per feature. As PTMs usually appear low abundantly, this is often challenging to achieve in practice (26).

Although we recommend the addition of a global profiling counterpart to an enriched PTM dataset, this is conceptually not required as normalisation can be done using all peptidoforms mapping to the same protein. However, we showed that this approach has the risk of partially diluting the effect of the PTM as their underlying peptidoforms are now involved in the calculation of the PTM usage.

As opposed to MSstatsPTM we do not make use of converters. Hence, msqrob2 input is not restricted to certain search engines or quantification algorithms, providing the user with full flexibility. However, this does require the user to convert their data into appropriate input format, which is a simple flat text file format (as exportable from a spreadsheet) or a data frame in R that can be used by the constructor for QFeatures objects. Furthermore, our workflows are modular and provide the user with the flexibility to use custom pre-processing steps. Default workflows are presented in our package vignettes, but these can easily be altered by building upon methods in the QFeatures package. Moreover, the use of the QFeatures infrastructure also guarantees that input data is never lost during processing, but remains linked to the pre-processed and normalised assays as well as to the model output, insuring transparency, traceability, and reproducibility. This allows the user to perform differential usage (and/or abundance) analysis on both PTM and peptidoform (or even protein) level, while storing and linking all these different results in a structured manner in the same object.

Another advantage of msqrob2PTM is that it can manage multiple modification sites per peptidoform. The peptidoform will then simply be used in the summarisation of multiple PTMs. This is particularly useful when using open modification search engines, which can often find multiple PTMs per peptide. Moreover, we also include workflows on differential abundance and usage analysis on the peptidoform level. Indeed, as shown in figures 10 and 11, it can be relevant to know whether a significant PTM stems from multiple (slightly) significant associated peptidoforms, or whether it is driven by one or a few very strongly significant associated peptidoform(s). In the latter case, it could be possible that these significant peptidoforms carry another modification that is driving the differential usage. Hence, we always advise users to conduct a peptidoform level analysis as well.

Overall, we have shown that our msqrob2PTM workflow is a sensitive and robust approach compared to the state-of-the-art, while providing good fpr control and high accuracy. Our modular implementation offers our users full flexibility with respect to the search engine and pre-processing steps, while still offering a comprehensive, transparent, and reproducible workflow that covers the entire differential PTM analysis.

Code and data availibility

The analysis files and data are available on https://github.com/statOmics/msqrob2PTMpaper and PRIDE PXD043476.

Supplementary materials

This article contains supplemental data.

Acknowledgements

This research was funded by the Research Foundation Flanders (FWO) as a mandate awarded to **ND** (1S77220N), and as project funding awarded to **LM** (G010023N, G028821N) and **LC** (G062219N), funding from the European Union's Horizon 2020 Programme to **LM** (H2020-INFRAIA-2018-1) [823839], and funding from a Ghent University Concerted Research Action to **LM** [BOF21/GOA/033] and **LC** [BOF20/GOA/023]. **LCG** and **PL** were supported by the Bundesministerium für Bildung und Forschung (01GM1917A), the research consortium "Multi-omic analysis of axono-synaptic degeneration in motoneuron disease (MAXOMOD)", which was funded in the scope of the E-Rare Joint Transnational Call for Proposals 2018 "Transnational research projects on hypothesis-driven use of multi-omic integrated approaches for discovery of disease causes and/or functional validation in the context of rare diseases." Phosphoproteomics experiments were supported by the french proteomics infrastructure (ProFI FR2048, ANR-10-INBS-08-03).

Author contributions

N.D.: writing, method development, data analysis
M.G.: writing, phosphorylation experiment and analysis
L.C.G.: phosphorylation experiment
P.L.: supervision phosphorylation experiment
C.C.: supervision phosphorylation experiment
L.C.: supervision, writing, method development
L.M.: supervision, writing

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Doll S, Burlingame AL. 2015. Mass spectrometry-based detection and assignment of protein posttranslational modifications. ACS Chem Biol 10:63–71.

2. Virág D, Dalmadi-Kiss B, Vékey K, Drahos L, Klebovich I, Antal I, Ludányi K. 2020. Current Trends in the Analysis of Post-translational Modifications. Chromatographia 83:1–10.

3. Mann M, Jensen ON. 2003. Proteomic analysis of post-translational modifications. Nat Biotechnol 21:255–261.

4. Olsen J V., Mann M. 2013. Status of large-scale analysis of posttranslational modifications by mass spectrometry. Mol Cell Proteomics 12:3444–3452.

5. Santos AL, Lindner AB. 2017. Protein Posttranslational Modifications: Roles in Aging and Age-Related Disease. Oxid Med Cell Longev 2017.

6. Wende AR. 2016. Post-translational modifications of the cardiac proteome in diabetes and heart failure. PROTEOMICS – Clin Appl 10:25–38.

7. Ramesh M, Gopinath P, Govindaraju T. 2020. Role of Post-translational Modifications in Alzheimer's Disease. ChemBioChem 21:1052–1079.

8. Samanta L, Swain N, Ayaz A, Venugopal V, Agarwal A. 2016. Post-Translational Modifications in sperm Proteome: The Chemistry of Proteome diversifications in the Pathophysiology of male factor infertility. Biochim Biophys Acta - Gen Subj 1860:1450–1465.

9. Kong AT, Leprevost F V., Avtonomov DM, Mellacheruvu D, Nesvizhskii AI. 2017. MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nat Methods 14:513–520.

10. Chi H, Liu C, Yang H, Zeng WF, Wu L, Zhou WJ, Wang RM, Niu XN, Ding YH, Zhang Y, Wang ZW, Chen ZL, Sun RX, Liu T, Tan GM, Dong MQ, Xu P, Zhang PH, He SM. 2018. Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. Nat Biotechnol 2018 3611 36:1059–1061.

11. Degroeve S, Gabriels R, Velghe K, Bouwmeester R, Tichshenko N, Martens L. 2021. ionbot: a novel, innovative and sensitive machine learning approach to LC-MS/MS peptide identification. bioRxiv 2021.07.02.450686.

12. Kohler D, Tsai T-H, Verschueren E, Huang T, Hinkle T, Phu L, Choi M, Vitek O. 2022. MSstatsPTM: Statistical relative quantification of post-translational modifications in bottom-up mass spectrometry-based proteomics. Mol Cell Proteomics 100477.

13. Schwämmle V, Verano-Braga T, Roepstorff P. 2015. Computational and statistical methods for high-throughput analysis of post-translational modifications of proteins. J Proteomics 129:3–15.

14. Schwämmle V, Vaudel M. 2017. Computational and statistical methods for high-throughput mass spectrometry-based PTM analysis. Methods Mol Biol 1558:437–458.

15. Wu R, Dephoure N, Haas W, Huttlin EL, Zhai B, Sowa ME, Gygi SP. 2011. Correct Interpretation of Comprehensive Phosphorylation Dynamics Requires Normalization by Protein Expression Changes. Mol Cell Proteomics 10.

16. Goeminne LJE, Gevaert K, Clement L. 2016. Peptide-level robust ridge regression improves estimation, sensitivity, and specificity in data-dependent quantitative label-free shotgun proteomics. Mol Cell Proteomics 15:657–668.

17. R Core Team. 2023. R: A language and environment for statistical computing.

18. Gatto L VC. 2022. QFeatures: Quantitative features for mass spectrometry data. 1.8.0 https://doi.org/10.18129/B9.bioc.QFeatures.

19. Sticker A, Goeminne L, Martens L, Clement L. 2020. Robust summarization and inference in proteome-wide label-free quantification. Mol Cell Proteomics 19:1209–1219.

20. Rainer J, Vicini A, Salzer L, Stanstrup J, Badia JM, Neumann S, Stravs MA, Hernandes VV, Gatto L, Gibb S, Witting M. 2022. A Modular and Expandable Ecosystem for Metabolomics Data Annotation in R. Metabolites 12:173.

21. Goeminne LJE, Argentini A, Martens L, Clement L. 2015. Summarization vs peptide-based models in label-free quantitative proteomics: Performance, pitfalls, and data analysis guidelines. J Proteome Res 14:2457–2465.

22. E Goeminne LJ, Sticker A, Martens L, Gevaert K, Clement L. 2020. MSqRob Takes the Missing Hurdle: Uniting Intensity-and Count-Based Proteomics. Anal Chem 92:6287.

23. Bates D, Mächler M, Bolker BM, Walker SC. 2015. Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw 67:1–48.

24. Cunningham CN, Baughman JM, Phu L, Tea JS, Yu C, Coons M, Kirkpatrick DS, Bingol B, Corn JE. 2015. USP30 and parkin homeostatically regulate atypical ubiquitin chains on mitochondria. Nat Cell Biol 2014 172 17:160–169.

25. Choi M, Carver J, Chiva C, Tzouros M, Huang T, Tsai TH, Pullman B, Bernhardt OM, Hüttenhain R, Teo GC, Perez-Riverol Y, Muntel J, Müller M, Goetze S, Pavlou M, Verschueren E, Wollscheid B, Nesvizhskii AI, Reiter L, Dunkley T, Sabidó E, Bandeira N, Vitek O. 2020. MassIVE.quant: a community resource of quantitative mass spectrometry–based proteomics datasets. Nat Methods 2020 1710 17:981–984.

26. Zubarev RA. 2013. The challenge of the proteome dynamic range and its implications for in-depth proteomics. Proteomics 13:723–726.

Université de Strasbourg

Marie GEBELIN

Développement de méthodes d'analyse protéomique et phosphoprotéomique à haut débit et leur application pour la recherche de biomarqueurs de pathologies sur de larges cohortes

Résumé

L'analyse protéomique par spectrométrie de masse permet l'identification, la quantification et la caractérisation structurale des protéines impliquées dans de nombreux processus biologiques. Cette approche peut également être appliquée à l'étude des modifications post-traductionnelles des protéines, telle que la phosphorylation. Ce travail de thèse est axé sur le développement de méthodes analytiques pour l'analyse protéomique et phosphoprotéomique à haut débit. Ces développements ont été réalisés à différents niveaux : la préparation automatisée des échantillons, l'analyse LC-MS/MS avec l'évaluation de différentes méthodes d'acquisition et le traitement des données par diverses solutions algorithmiques. Ils ont ensuite été appliqués pour l'étude de large cohortes d'échantillons cliniques dans le but d'identifier et quantifier de potentiels marqueurs de la sclérose latérale amyotrophique. Enfin, des échantillons et métriques de contrôle qualité ont été implémentés à la fois pour l'analyse protéomique et phosphoprotéomique adaptés à des études de grands nombres d'échantillons.

Mots-clés : Protéomique, Spectrométrie de masse, Phosphoprotéomique, Acquisition indépendante des données (DIA), Multi-omique

Résumé en anglais

Mass spectrometry-based proteomic analysis enables the identification, quantification and structural characterisation of proteins involved in numerous biological processes. This approach can also be applied to the study of post-translational modifications of proteins, such as phosphorylation. This PhD work focuses on the development of analytical methods for high-throughput proteomic and phosphoproteomic analysis. Developments were carried out at different levels: automated sample preparation, LC-MS/MS analysis with the evaluation of different acquisition methods and data processing using various algorithmic solutions. They have then been applied to the study of large cohorts of clinical samples with the aim of identifying and quantifying potential markers of amyotrophic lateral sclerosis. Finally, samples and quality control metrics have been implemented for both proteomic and phosphoproteomic analysis adapted to studies of large numbers of samples.

Keywords: Proteomics, Mass spectrometry, Phosphoproteomics, Data Independent Acquisition (DIA), Multi-omics