Computational methods for identifying structural protein alterations from limited proteolysis-coupled mass spectrometry data

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Luise Nagel

geboren in Freiburg im Breisgau

angenommen im Jahr 2023

All cannot be lost when there is still so much being found.

– Lemony Snicket

Abstract

The structure of proteins plays a central role in the functional state of cells and organisms, incorporating multilayered mechanisms that contribute to the complexity of cellular processes. Understanding protein structures is essential for elucidating their functions, interactions with other molecules, and contributions to various biological processes, ultimately providing crucial insights into cellular mechanisms, metabolic pathways, pathological processes, and organismal physiology. An emerging method for quantifying structural changes on a proteome-wide scale is combining limited proteolysis with mass spectrometry (LiP-MS). LiP-MS utilizes unspecific proteinase K (PK) digest of accessible protein regions to identify structural alterations, such as conformation changes, misfolding or altered protein-protein interactions, in thousands of proteins in their native environment.

We investigated whether global, *in situ* analysis of protein structural changes could serve as structural biomarkers to reflect pathophysiological processes. Parkinson's disease (PD) is a prevalent neurodegenerative disease that lacks reliable biomarkers for diagnosis and, as it is a prominent example of a disease associated with structural proteome changes, it is an ideal candidate for identifying structural candidate biomarkers. To define these structural changes associated with disease, we analyzed samples from both healthy individuals and those with PD using LiP-MS. Our newly developed bioinformatic pipeline allowed us to identify 76 proteins with structural alterations in the cerebrospinal fluid (CSF) of individuals with PD compared to healthy donors. Multivariate feature selection models showed that CSF protein structural information is more effective than protein abundance information in distinguishing between healthy participants and those with PD. This study suggests that LiP-MS can identify novel structural biomarker candidates for disease and help generate hypotheses about underlying disease processes.

Identifying structural protein alterations from LiP-MS data requires novel analyses approaches, as it remains challenging to deconvolute the different signal contributions. To tackle this challenge, we developed a new computational processing and analysis workflow based on the PD work, which can be applied to various LiP-MS datasets. We, for the first time, compared different approaches that deal with the LiP-MS specific challenges, and propose a two-step approach that first removes unwanted variations from the LiP signal and then infers the effects of variables on the structural accessibility of proteins. Our framework provides a uniquely powerful approach for deconvolving LiP-MS signals by separating the contribution of structural changes, protein abundance, post-translational modifications and alternative splicing.

Zusammenfassung

Die dreidimensionale Struktur von Proteinen ist von zentraler Bedeutung für die Funktion von Zellen und Organismen. Sie spiegelt eine Vielzahl molekularer Mechanismen wider, die zur Komplexität zellulärer Prozesse beitragen. Daher ist es entscheidend, die Strukturen von Proteinen und ihre Veränderungen zu entschlüsseln, um ihre Funktionen, ihre Wechselwirkungen mit anderen Molekülen und ihre Beteiligung an verschiedenen biologischen Prozessen zu verstehen. Dies ermöglicht entscheidende Einblicke in zelluläre Mechanismen, Stoffwechselwege, pathologische Prozesse und die Physiologie von Organismen. Eine neue Methode zur Identifizierung von Strukturänderungen im gesamten Proteom ist die Anwendung von limitierter Proteolyse in Kombination mit Massenspektrometrie (LiP-MS). Der unspezifische Verdau von Proteinen durch Proteinase K (PK) in strukturell zugänglichen Regionen ermöglicht die massenspektroskopische Charakterisierung struktureller Veränderungen von Proteinen wie Konformationsänderungen, Fehlfaltungen oder veränderte Protein-Protein-Wechselwirkungen in tausenden von Proteinen in situ.

Da sich zelluläre pathophysiologische Prozesse in Proteinstrukturen widerspiegeln, haben wir getestet, ob globale in situ Analysen von strukturellen Veränderungen in Proteinen als strukturelle Biomarker dienen können. Morbus Parkinson (PD) ist eine weit verbreitete neurodegenerative Erkrankung, für die es bisher keine zuverlässigen diagnostischen Biomarker gibt. PD geht mit strukturellen Veränderungen im Proteom einher und ist daher eine geeignete Erkrankung für die Suche nach strukturellen Biomarker-Kandidaten. Um Biomarker-Kandidaten zu identifizieren, wurden Proben von gesunden und an Parkinson erkrankten Personen mittels LiP-MS analysiert. Ein neu entwickeltes bioinformatisches Analyseverfahren ermöglichte die Identifizierung von 76 strukturell veränderten Proteinen in der Zerebrospinalflüssigkeit (CSF) von Personen mit Parkinson im Vergleich zu gesunden Spender*innen. Modelle, die auf strukturellen Veränderungen von Proteinen basierten, um Personen mit Parkinson von gesunden Spendern zu unterscheiden, hatten eine höhere Erfolgsrate als Modelle, die auf der Häufigkeit von Proteinen basierten. Unsere Studie zeigt, dass LiP-MS verwendet werden kann, um strukturelle Biomarker-Kandidaten zu identifizieren, und dass auf dieser Basis Hypothesen über die zugrunde liegenden Krankheitsprozesse entwickelt werden können.

Die Extraktion struktureller Proteinmodifikationen aus LiP-MS Daten erfordert neue analytische Ansätze, da die Entschlüsselung der verschiedenen Signalbeiträge eine Herausforderung darstellt. Um dieses Problem zu lösen, haben wir einen neuen Datenverarbeitungs- und Analyseworkflow für die Anwendung auf verschiedene LiP-MS Datensätze entwickelt, der auf den Analysen des PD-Biomarker-Kandidaten-Projektes basiert. Zum ersten Mal vergleichen wir verschiedene Ansätze, um die spezifischen Herausforderungen in der LiP-MS Datenanalyse zu adressieren, und schlagen einen zweistufigen Vorgehen vor. In dessen Rahmen werden zuerst unerwünschte Variationen aus dem LiP-Signal entfernt und anschließend die Auswirkungen von Variablen auf die strukturelle Zugänglichkeit von Proteinen berechnet. Unsere Methode bietet einen einzigartigen robusten Ansatz für die Dekonvolution von LiP-MS Signalen, indem sie zwischen den Beiträgen von strukturellen Veränderungen, Proteinabundanzen, posttranslationalen Modifikationen und alternativem Spleißen unterscheidet.

Contents

Acknowledgements

This thesis is the culmination of more than four years of research in the Beyer lab, which was only possible with the tremendous support of many people. I would like to take this opportunity to acknowledge some of them.

First and foremost, I would like to express my deepest gratitude to Andreas Beyer for welcoming me into his lab and for being an excellent supervisor over the years. You have created an amazing lab environment, both scientifically and socially. You gave me the opportunity to participate in many exciting projects, motivated me to think outside the box, and always took the time to supervise and guide me. Your scientific curiosity and persistence continue to amaze and inspire me.

I am indebted to Paola Picotti for hosting me during a crucial six months of my Ph.D. journey, for the great collaborations, and for serving on my thesis advisory committee. Her support and valuable input in many projects, together with her positive attitude, have significantly shaped my work, as well as me as a scientist.

In the realm of this thesis, I would like to express my sincere gratitude to Jan Grossbach, Philipp Antczak, Daniel Sieme, and Christian Doerig. Their ongoing discussions and feedback throughout the writing process have been invaluable to this work.

I would like to express my gratitude to the entire Beyer lab for the excellent atmosphere, all the support, the scientific and non-scientific discussions, and the memories we collected. I would especially like to thank Jan Grossbach for all his help and support over the years. Thank you for all the patience, especially in the early years of my Ph.D., for the guidance and valuable ideas, as well as the countless coffee break discussions. Special thanks also to Carolina Leote. You have gone from being a colleague to one of my best friends. Thank you for always philosophizing with me about science, for all the time we spent together inside and outside the lab, and for being there at every step of the way – the good and the bad. I am grateful to have you in my life. Thank you to everyone who has been with me along the way, Antonis Papadakis, Tim Padvitski, Ronja Johnen, Fabian Titz-Teixeira, Dennis Gadalla, and many others. You make this lab an amazing place to be.

I would also like to thank all the members of the Picotti lab, where I was lucky enough to spend half a year during my Ph.D. A special thanks to Christian Doerig, Marie-Therese Mackmull, Aleš Holfeld, and Monika Pepelnjak for the warm and hearty welcome in their office. Thank you for introducing me to the world of mass spectrometry and all the ongoing discussions about science. You made Zurich feel like home for the time being. Special thanks to Marie-Therese Mackmull, you were an exceptional collaborator and greatly influenced my early development as a scientist. I would like to especially thank Natalie de Souza for the discussions and scientific insights, as well as for her inspiration and guidance.

Finally, my deepest appreciation goes to my friends and family outside of academia. To Doro, thank you for your unwavering support and constant encouragement. To my dear friends – Anni, Daniel, Kiki, Luisa, Thomas, Ragna, Djamila and others – your company has been a constant source of strength. And to my family, my parents and siblings, your continued support has been the foundation of my journey.

Abbreviations

α-Syn	α-Synuclein
AA	amino acid
AD	Alzheimer's disease
Cryo-EM	cryogenic electron microscopy (Cryo-EM)
CSF	cerebrospinal fluid
CFS	chronic fatigue syndrome
DLB	dementia with Lewy Bodies (DLB)
DNA	deoxyribonucleic acid
IDPs	intrinsically disordered proteins
IDRs	intrinsically disordered protein regions
LC-MS	liquid chromatography-mass spectrometry
LiP	limited proteolysis
LiP-MS	limited proteolysis coupled with mass spectrometry
mRNA	messenger RNA
MS	mass spectrometry
MSA	multiple system atrophy
NMR	nuclear magnetic resonance
PCA	principal component analysis
PD	Parkinson's disease
PDD	Parkinson's disease dementia
PK	proteinase K
PPI	protein-protein interaction
pre-mRNA	precursor mRNAs
PRM	parallel reaction monitoring
PTM	post-translational modification
RNA	ribonucleic acid
SRM	selected reaction monitoring
TPP	thermal proteome profiling
tRNA	transfer RNA
Trp	trypsin-only
trpCS	trypsin cleavage site
XL-MS	cross-linking combined with mass spectrometry

1 Introduction

1.1 Proteomic diversity reflects the functional landscape of an organism

The remarkable diversity and functionality of organisms is the result of the interplay between their genetic blueprint and the dynamic orchestra of cellular components. While the number of protein-coding genes encoded in an organism's genome largely defines its functional diversity, a cell's ability to exhibit diverse phenotypes and metabolic states significantly depends on variation within the transcriptome and beyond[1–3]. Proteins, proteoforms, and their intricate interactions are the final actors in shaping cellular diversity, metabolic states, and functional dynamics in the realm of cellular complexity[4,5]. Therefore, proteostasis, which encompasses the tight control and dynamic balance that regulates protein synthesis, folding, trafficking, and degradation, is necessary to maintain a steady state in healthy cells or organisms[1,6]. The structure and interactors of proteins mirror their functional state and incorporate multilayered mechanisms that contribute to the complexity of cellular processes.

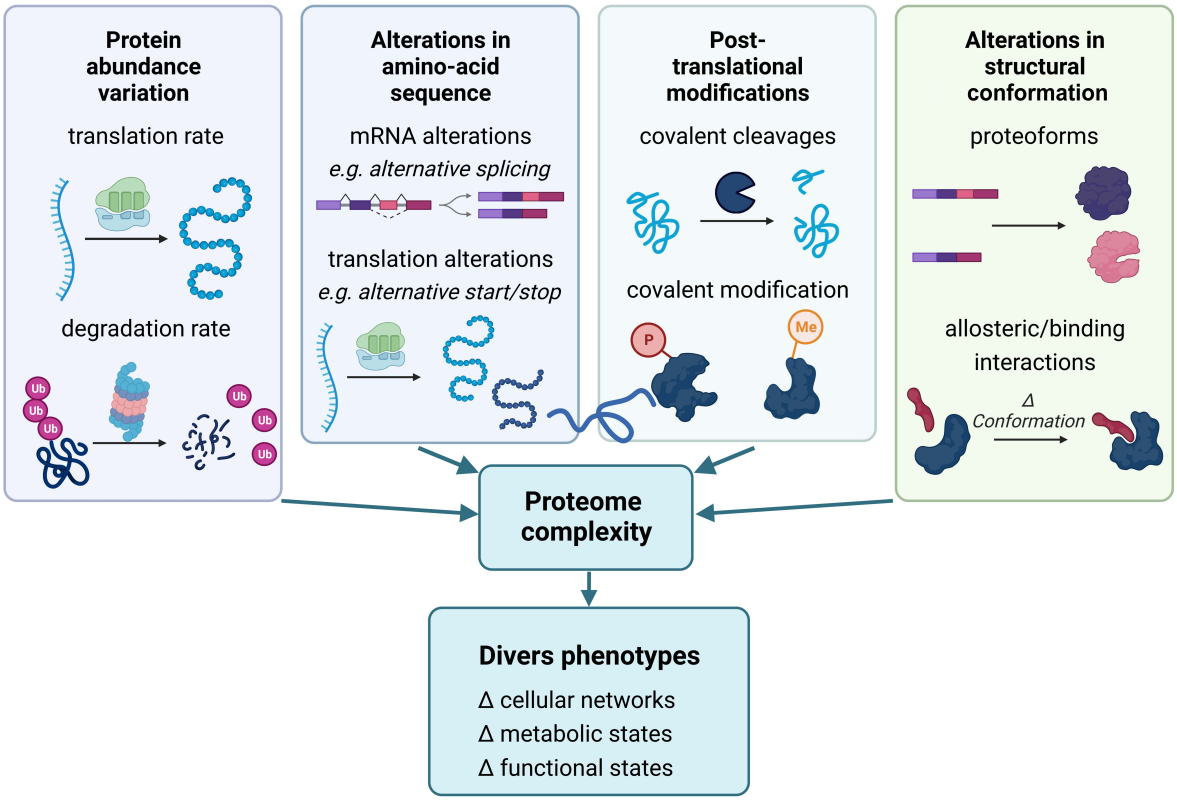


Figure 1: Factors contributing to the complexity of the proteome. *(Created with BioRender.com)*

A major contributor to the variation of the proteome, the complete set of proteins within a cell or organism, is the modulation of protein expression levels through the dynamic interplay of protein synthesis and degradation rates[2]. The abundance levels of all proteins in a cell significantly define the phenotype and thus the functional state of the cell[4]. A significant portion of the translation rate, the speed at which ribosomes convert messenger ribonucleic acid (mRNA) into proteins, can be attributed to fluctuations in mRNA levels[3,7,8], that are affected by both cis effects (modifications occurring on the same molecule) and trans effects (modifications caused by different molecules)[9]. In addition, mRNA transcripts themselves control the expression of their protein by influencing the rate of translation

through various mechanisms such as codon bias, interaction with regulatory elements or epitranscriptomic modifications[4,10]. Cells further modulate protein synthesis rates by altering transfer RNA (tRNA) and ribosomal availability[10]. The accumulation and thus the expression level of a protein is also dependent on its degradation rate, which reflects the rate at which proteins are degraded, primarily by the ubiquitin-proteasome system or autophagy[11].

Further protein variability is introduced by modulating the amino acid (AA) sequence of a protein. Such variations in protein sequence can affect not only the protein abundance, but also activity, binding partners, localization, three-dimensional conformation and, ultimately, function. One possibility for such variation is the presence of protein modulating sequence variation within the mRNA transcripts, which is then transferred to the protein level through translation. For instance, deoxyribonucleic acid (DNA) contains alternative transcription start and stop sites, which enable the production of precursor mRNAs (pre-mRNAs) of varying lengths from the same gene[12,13]. Another major contributor to protein variability, especially in eukaryotes, is alternative splicing of transcripts, which results in different forms of the same protein, by subselecting specific exons, that may be enriched for binding motifs and post-translational modification (PTM) sites required in tissue-specific segments[14,15]. Protein variability can also occur due to rare events, such as the modulation of alternative start and stop codons. This can result in the production of different polypeptides from the same mRNA transcript[16,17].

Much of the diversity within the proteome results from post-transcriptional modifications, i.e., cleavages or covalent modifications that occur after protein translation[18]. Post-translational proteolytic processing can modify protein function through covalent cleavage mediated by specific proteases or autocatalytic cleavage[19]. Covalent modifications induce enzyme-catalyzed changes on the side chains of proteins, including phosphorylation, acylation, and oxidation[18]. These covalent cleavages and modifications can regulate the protein function by altering protein activity, changing protein localization, and influencing interactions with other proteins and are a crucial component of cell signaling pathways[4,20].

The structural proteome, which refers to the complete set of three-dimensional arrangements and interactions of proteins within a cell or organism, reflects all of the aforementioned changes and adds another layer of complexity. Most simply, changes in the AA sequence of a protein alter its chemical properties and thus its three-dimensional structure[4]. However, even within the same proteoform, differences in three-dimensional conformation can exist[4]. After translation and folding, a protein can change conformation through allosteric interactions and/or binding partners, altering its structure and function to adapt to external cues[21]. It is important to note that protein structure not only defines but also reflects protein function[4,22]. Therefore, cells closely monitor the folding of proteins to ensure the generation and maintenance of their specific structural conformation[22].

The functional proteome of a cell or organism reflects its diversity, forming a biomolecular network. The structural proteome itself mirrors the state of a cell, incorporating multilayered cues of cellular diversity that reflect the functional and metabolic state.

1.2 Protein structure as a readout of the functional state of a cell or organism

Quantifying the structural state of a proteome is critical for analyzing the functional landscape and current state of a cell or organism. The structural proteome both defines and reflects protein function and thus the phenotype of a cell.

1.2.1 Experimental approaches to assess protein structure and conformation

The three-dimensional conformation of a protein is central to its function, defining its activity, stability, and interaction partners. Therefore, quantifying the structure of a protein and of protein complexes has been a main focus of the field for decades[23].

X-ray crystallography is one of the earliest methods used to determine protein structure. In this technique, proteins are crystallized and then exposed to X-rays. The diffraction patterns produced by the interaction between the crystallized protein and the X-rays can then be used to deduce the three-dimensional structure of the protein[23,24]. While this method is well-established and provides high-resolution protein structures, it relies on the crystallization of highly purified proteins, resulting in a static representation of the tested molecule. It is important to note that this method is therefore limited to static representations and does not provide information on dynamic changes in protein structure over time[24]. Both the production of sufficient amounts of purified protein and its crystallization are challenging and time-consuming, especially for membrane proteins or highly flexible and dynamic proteins. Obtaining the structure of a single protein typically requires weeks or months of work[25,26].

Nuclear magnetic resonance (NMR) spectrometry is an alternative method for determining the three-dimensional structure of proteins. The protein sample is subjected to a strong magnetic field and radio frequency pulses. By detecting and analyzing the signal emitted by the nuclei of the atoms in the protein, structural information can be derived from their chemical shifts and interactions[27]. Compared to X-ray, NMR spectrometry is not reliant on a crystalline sample and is highly sensitive to subtle changes in the chemical environment, allowing for the recapitulation of the dynamic motion of the protein and its interactions[28]. In the past there were severe limitations on the size of a biomolecule that could be analyzed. Although these limitations have recently been pushed towards 100kDa, they still persist[29]. While NMR spectrometry can be performed in more complex environments than X-ray[30], it is still limited to a single protein or a very small number of proteins, with the analysis of a single biomolecule typically requiring weeks to months of work by highly trained specialists[31].

Recently, cryogenic electron microscopy (Cryo-EM) has become extremely popular for quantifying protein structure, as it has overcome its previous limitations of being restricted to large complexes and low resolution[32]. Cryo-EM generates high-resolution structures by analyzing electron measurements taken from protein samples frozen in a thin layer of vitrified ice. This produces detailed three-dimensional models of protein molecules and even large complexes without the need for crystallization, making it accessible for the study of dynamic structures[33]. The sample is flash-frozen to capture proteins in their near-native configuration, preserving their biological state and allowing the visualization of transient or dynamic conformations[34]. Cryo-EM is a less time-consuming technique compared to X-ray crystallography or NMR spectrometry, typically taking only days up to a few weeks. However, purchasing and maintaining the necessary equipment is expensive, and operating it requires expert users[35].

While experimental approaches to protein structure elucidation have improved greatly over the past decades, they are limited to the analysis of a single protein or a small number of proteins simultaneously.

1.2.2 AlphaFold as the computational approach for predicting protein structure

To make protein structures more ubiquitous, computational approaches have been developed to generate protein structures *in silico*. These promise to complement experimental approaches, providing solutions where experimental techniques reach their limits, and offer techniques that can be applied to all known proteins in a very short time. Several protein structure modeling methods have been developed over the years[36], but the breakthrough came in 2020 with the release of DeepMind's AlphaFold[37].

AlphaFold uses a deep learning neural network to predict the three-dimensional structure of a protein from its sequence. The neural network itself is trained on known protein structures that scientists have elucidated over decades using a variety of experimental approaches, and recognizes patterns between amino acid sequences and their corresponding folded three-dimensional structure. This information is used to predict the folding patterns and spatial arrangements of a protein based solely on its sequence. Although AlphaFold still has limitations, such as poor performance on unstructured regions and currently mainly predicts single states of proteins, it clearly outperforms all other currently available computational approaches[37,38]. Additionally, PTMs, which can critically affect protein folding, are currently not taken into account by AlphaFold. Nevertheless, it provides for the first time a reliable approach to predicting the three-dimensional structure of proteins that can be rapidly applied to any known protein sequence[37]. AlphaFold has since been further developed and applied to many problems related to protein folding, including the prediction of protein-protein interactions (PPIs)[39] and protein-complexes[40], cellular responses to missense mutations[41], and de novo protein design[42].

The structural proteome is not a fixed state, as AlphaFold considers it, but is highly dynamic, with the structure of proteins constantly changing in response to various biological and environmental stimuli. Measuring these dynamic changes in protein structure can provide a functional readout of a biological system and its metabolic and, in the case of disease, pathological state[43]. It is important to note that the current structural state of a protein is significantly influenced by protein-binding interactions, as protein binding or unbinding events can mediate large conformational changes[21]. There are various experimental and computational approaches to define a single structural ground state and/or the current structural conformation of proteins. However, none of the existing methods provide a proteome-wide approach for high-throughput quantification of changes in the structural proteome in its complex biological context[44]. A novel approach that fulfills all of these requirements was introduced in 2014 in the form of coupling limited proteolysis with mass spectrometry (LiP-MS)[43,45].

1.3 Measuring protein structural changes on a proteome-wide scale using mass spectrometry

Since the identification of proteins as functional entities within cellular systems, researchers have sought to identify and measure proteins in experimental systems to better understand the observed differences in phenotypes. The technology has undergone several significant developments to increase the accuracy and speed of peptide/protein detection with mass spectrometry (MS) being the most widely used high-throughput, proteome-wide method today.

1.3.1 Mass-spectrometry for proteome-wide quantification

Mass spectrometry has become an indispensable technology for protein quantification on a proteome-wide scale, with liquid chromatography-mass spectrometry (LC-MS) being the most commonly used technique.

In a standard experimental setup, proteins are extracted from the biological sample of interest, such as cell lysate or tissue and then subjected to enzymatic digest resulting in multiple peptides per protein. Trypsin is commonly used for this step, as it cleaves specifically at arginine and lysine residues, producing distinct peptides with a protonated amino acid at the C-terminus, as the basic side chains can carry positive charges, allowing for reliable peptide detection in the MS. The peptide mixture is then injected into a high-performance liquid chromatography, where it is separated based on chemical affinities. The eluted peptides are ionized and introduced into the mass spectrometer as precursor ions. During the first MS scan (MS1), precursor ions are selected based on their chemical properties and then fragmented into product ions. The mass spectrometer then quantifies these ions based on their mass-to-charge ratio, resulting in an MS/MS spectrum. In the final step, the MS1 and MS/MS spectra are matched against the protein sequence database to identify and quantify fragments, precursors, peptides ,and corresponding proteins.[46]

Mass spectrometry can quantify the presence and abundance of proteins in a biological sample and is used to compare protein expression levels under different conditions. While this is the most common application of MS in proteomics, approaches such as the detection and characterization of PTMs[47] or PPIs[48] have also been developed and are widely used in the field.

1.3.2 Combining limited-proteolysis with mass spectrometry

Standard mass spectrometry can be used to study various properties of the proteome, but only a novel approach that combines limited proteolysis with mass spectrometry (LiP-MS) has made it possible to the study structural aspects of the proteome using MS[43,45,49]. Unlike trypsin digest, limited proteolysis does not depend on specific amino acid residues, but rather on the flexibility and surface accessibility of the protein[50]. LiP-MS utilizes these changes in digest patterns resulting from structural alterations in the proteome to identify differences in structural accessibility under different conditions in a proteome-wide, high-throughput manner.

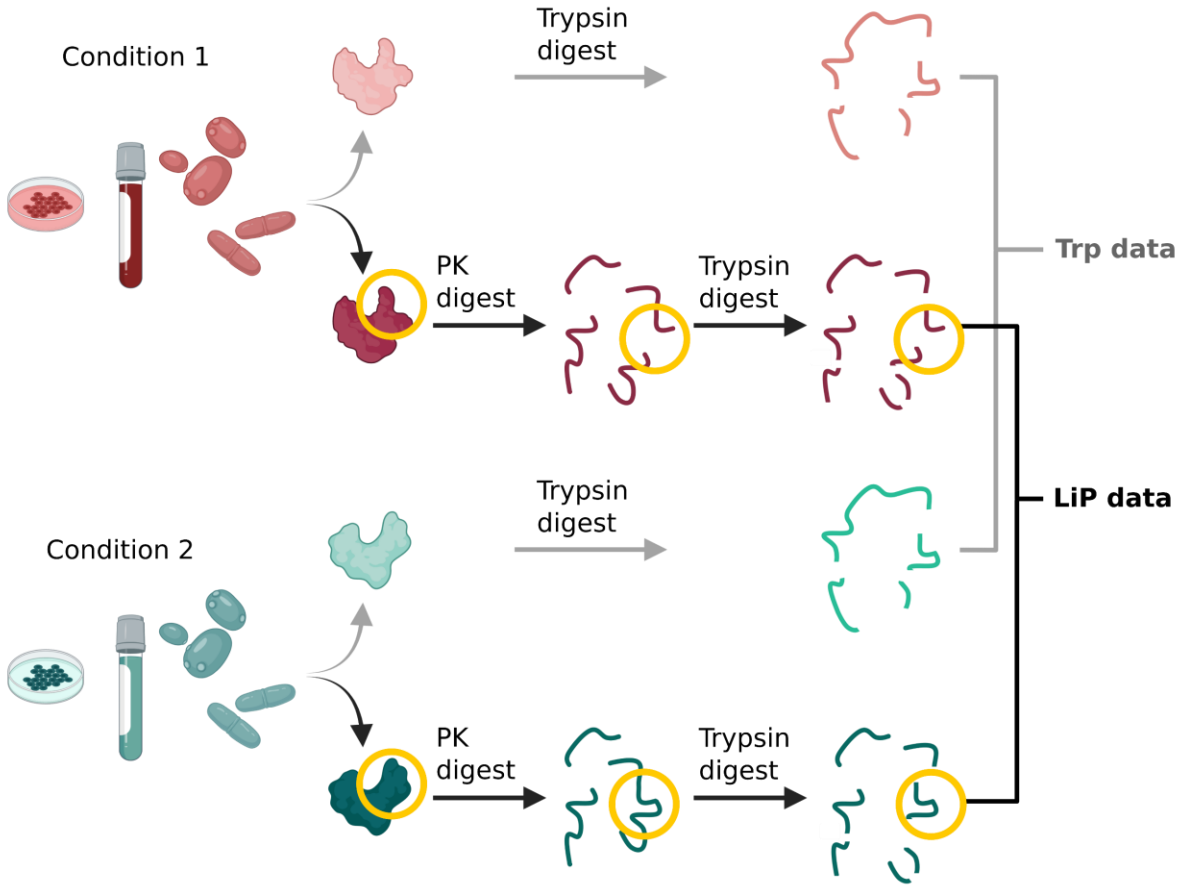


Figure 2: Experimental LiP-MS workflow. *The proteome of samples with different conditions is extracted under non-denaturing, native conditions and digested with (1) trypsin-only (Trp) (2) proteinase K (PK) (short time, limited proteolysis, LiP) followed by trypsin. Regions with a different structural PK accessibility (yellow) induce different digest patterns during the limited proteolysis. The intensities of LiP and Trp intensities are quantified by mass spectrometry. (Modified from Chapter II: LiPAnalyzeR Figure 2, parts are created with BioRender.com)*

To quantify differences in the structural accessibility of proteins, i.e., structural changes in the proteome, native proteomes from different conditions are first subjected to a short (i.e., limited) proteinase K (PK) digest. This sequence-unspecific digest targets the protein regions that are most accessible to PK. The PK digest is then stopped, and the samples are subjected to a conventional complete trypsin digest (Figure 2). Distinct PK digest patterns are likely to result from differences in protein accessibility between conditions. These patterns are typically reflected in both the half-tryptic peptides (digested by PK at one terminus and trypsin at the other) and fully-tryptic peptides (digested by trypsin at both the C- and N-terminus), providing a readout for structural change (Figure 2, yellow circled region). Samples that undergo this double digest are referred to as LiP samples. Following the double digest, MS is used to quantify differences in peptide abundance across the entire proteome.

The intensities of LiP peptides (i.e., the MS intensities of peptides subjected to the double digest with PK and trypsin) are not solely dependent on the variation in the PK accessibility, but are also driven by PK-independent effects such as the abundance of their protein, and the occurrence of peptide-specific modifications (i.e., alternative splicing or PTMs). LiP-MS studies may perform a second control digest without the limited PK digest by only using trypsin digest to quantify protein variation independent of the PK accessibility (Figure 2). The resulting trypsin-only digested samples are also analyzed using MS. Throughout this work, the measurements from the trypsin-only digested samples are referred to as 'Trp'. The use of 'Trp' as the abbreviation for the trypsin-only run has become established in the field[49],

although this may be potentially misleading as the processing of LiP samples also involves trypsin digest. Trp peptides (i.e., trypsin-only peptides) include variation due to differences in protein abundance as well as PK-independent peptide variation (i.e., alternative splicing or PTMs). Trp protein abundances are estimated from the Trp peptide intensities and therefore include variation introduced by differences in the protein abundance. To provide a more reliable estimate of protein abundance per protein per sample, peptide intensities from the same protein are typically combined. This is because a single peptide is expected to have a higher signal-to-noise ratio[51].

The LiP workflow generates two types of peptides: (1) LiP peptides, which are MS-measured peptide intensities from the double-digested samples that contain variations caused by changes in the structural accessibility, PK-independent peptide effects such as PTMs, and differences in the protein abundance; (2) Trp peptides, which are MS-measured peptide intensities of the trypsin-only digest that contain the variation induced by PK-independent peptide effects such as PTMs and differences in the protein abundance; and (3) Trp proteins, which are protein abundances estimated from the MS-measured Trp peptide intensities.

1.3.3 Thermal proteome profiling as an alternative to LiP-MS

Thermal Proteome Profiling (TPP) is a mass spectrometry-based method that can detect proteome-wide structural changes in proteins and assess their thermal stability as an alternative to LiP-MS[52,53].

In TPP, MS is used to measure denaturation curves of proteins on a proteome-wide scale, which unfold as temperature increases. This determines the thermal stability of proteins and allows for identification of ligand-induced shifts[52,53]. TPP has since been applied to study various topics, including protein complex dynamics[54,55], protein modifications[56], protein-protein interactions[57], and missense protein variants[58]. TPP has a high proteome coverage and generally produces reliable results, but is limited to detecting structural changes that affect the thermal stability of the protein or protein complex[53]. Although TPP offers a slightly greater proteome coverage than LiP-MS[59], it lacks the ability to provide structural resolution, since it only detects the overall thermal stability of a protein. In contrast, LiP-MS, enables the identification of the protein regions undergoing structural alteration by approximately 10 amino acids[49], allowing for statements on both the protein regions that undergo structural changes and those that remain stable. If, for example, a PTM alters the three-dimensional conformation of a protein and exposes a previously buried binding site, TPP would only reveal the overall temperature-related stability change. However, LiP-MS would identify the modification and the specific location of the structural alteration. This information can be used to predict how the structural alteration in the protein binding site might alter downstream effects in a cell.

1.4 Application of limited proteolysis-coupled mass spectrometry

LiP-MS data provides a snapshot of the structural proteome of a cell or organism, capturing the three-dimensional rearrangements and interactions of proteins. However, it has limitations in inferring high-resolution features of specific protein structures, such as defining their secondary structure. Furthermore, it does not provide information about the nature of detected structural changes. Instead, it provides the ability to simultaneously monitor structural changes in thousands of proteins between different conditions, such as cell lysate without interactors versus with interactors (e.g., small molecules, drug candidates), samples from cells under different nutritional conditions, or healthy versus diseased samples. LiP-MS can detect structural alterations beyond conformational changes of proteins, such as protein misfolding, aggregation, or variable shielding of surface areas due to ligand

binding. This also allows for the detection of protein interactions. Detailed follow-up experiments are necessary to determine the type of structural alteration detected by LiP-MS.[43,49,60]

1.4.1 Targeted analysis of proteins with LiP-MS

LiP-MS can be used to quantify structural features of specific proteins, such as structural transitions under different conditions. This application requires purified protein samples instead of the entire cell lysate. It can, for example, be used to quantify alternative structural conformations, such as of the human protein α-Synuclein (α-Syn). Different structural conformations of α-Syn, including the monomeric conformation and amyloid aggregates, result in distinct PK digest patterns[43] and can even be used to generate structure-specific proteolytic fingerprints[61]. This method can also detect more subtle conformational transitions and even allow for the calculation of the amounts of specific conformational states in mixed samples[43]. LiP-MS has been used to identify folded domains of the same protein empirically, quantify differences in stability, and define the most stable folded regions of the protein[62]. Similarly, the conformational stability of different protein variants can be quantified, which can be of great interest when aiming to develop a stable protein product for therapeutic purposes[63]. The identification of binding interfaces of proteins can be achieved by comparing the LiP patterns of the protein of interest in bound and unbound conformation. This has been utilized to investigate various types of binding events, including small molecule binding[64–66], hormone binding[65], protein-protein binding[67] as well as conformational rearrangements driven by different protein-ligand interactions[68].

1.4.2 Quantifying protein-binding events with LiP-MS

Limited proteolysis can also facilitate the study of the interactome of (bio-)molecules. For this type of application, LiP-MS is commonly performed on cell lysates, with some samples being subjected to the (bio-)molecule of interest, while others remain without the potential binding partner. Comparing the LiP patterns of the two samples allows the detection of binding partners, for example, to find PPIs and to define the interactome of the protein of interest[67,69]. Recently, this approach has been extended and applied to measure structure-specific protein-protein binding interactions. As different conformations of a protein have the potential to bind to different parts of the proteome, LiP-MS was used to study the PPIs of different conformations of α-Syn, revealing proteoform-specific interaction networks[70]. LiP-MS has also been utilized to define small molecule binding partners, enabling the identification of target proteins of known drugs[71] and characterizing new drug targets, as well as their potential off-targets[59]. The systems aspect of LiP-MS is particularly valuable in this context, as it allows for the identification of all proteins that bind a small molecule in the entire proteome subjected to it. Furthermore, LiP-MS can also be used to study metabolite-protein interactions in their native cellular environment. The application of this approach to *Escherichia coli* allowed the localization of metabolite binding sites on a proteome-wide scale. This provided new insights into catalytic, allosteric and metabolite-induced PPIs[72].

1.4.3 Applying LiP-MS for a global analysis of the structural proteome

While alternative methods such as cross-linking can also address some of the applications mentioned above[60,73], LiP-MS has unique properties that allow for the quantification of various types of structural changes in the native cellular environment on a proteome-wide scale.

Studying an organism's reactions and adaptations to environmental changes is crucial for basic biological research. These changes can affect various biological layers, such as gene expression or protein abundance, and are also reflected in the structural proteome. LiP-MS is a valuable tool for quantifying and monitoring these changes in detail. The quantification of the entire structural proteome can be used to assess protein binding interactions on a proteome-wide scale[72] or to monitor proteome-wide conformational changes in samples treated with small molecules[74]. In this context, LiP-MS has been used to identify nutrient-related changes in the structural proteome of both *Saccharomyces cerevisiae*[45] and *Escherichia coli*[75], shedding new light on the metabolic transition occurring in the samples. Similarly, LiP-MS was employed to investigate the acute stress response in *Saccharomyces cerevisiae*[75].

Another global application of LiP-MS is characterizing various yeast, bacterial strains, or cell lines. Although organisms within the same family may have quite similar genomes, differences in their metabolism should be reflected at the structural proteome level and can thus be accessed by LiP-MS. LiP-MS can also be used to study the downstream effects of a single knockout or mutation in the genome that propagates into the proteome of a cell[76]. The application of LiP-MS to cell lines as disease models demonstrates the ability of LiP-MS to deduce disease-related alterations in the structural proteome of cells. It is widely accepted that conformational changes in the proteome can be causal for a pathological phenotype, as well as result from a downstream effect in disease states[77]. Therefore, identifying changes in protein structure in disease states can provide insights into pathological processes in a cell. For instance, LiP-MS was used to analyze the structural proteome of two distinct breast cancer cell lines that represent varying degrees of disease severity[78]. The study revealed significant conformational changes that were found to be unique to different breast cancer phenotypes, and largely independent of the expression level of the affected proteins. Cell lines generated from samples of both diseased and healthy individuals have also been shown to exhibit distinct digest patterns and thus important differences in the structural proteome when analyzed globally by LiP-MS, even revealing structural variation between different cell types[79,80].

One of the most novel applications of LiP-MS is to whole organisms or mammalian samples, providing access to pathological changes in the global proteome to better understand altered functional states. For instance, LiP-MS has been used in aging research to investigate decreased protein homeostasis, which is a key focus in this field. Aging is often accompanied by protein misfolding and aggregation, which are present in many age-related diseases, including Alzheimer's disease (AD) and Parkinson's disease (PD)[81]. These structural changes related to aging have been since characterized in *Saccharomyces cerevisiae*[82], *Caenorhabditis elegans*[83], and mouse cerebrospinal fluid[84]. LiP-MS has also been applied to human CSF samples from individuals with AD, mild cognitive impairment, and healthy controls, identifying proteins undergoing structural changes, some of which are known to be associated with AD[85]. In another study, LiP-MS analysis of saliva samples from oral cancer patients revealed structural changes that are correlated with clinical features and may serve as new prognostic markers[86].

LiP-MS provides a novel approach for systematically obtaining proteome-wide structural changes. It has numerous applications for investigating various biological questions in different sample types and organisms.

1.5 Applying LiP-MS to define candidate biomarkers in Parkinson's disease

The human proteome is a complex system that requires a delicate balance. Any disruptions to this balance can cause stress on cells and organisms, contributing to or causing disease. Pathological phenotypes are reflected in the proteome, as it, for example, links disease-causing genetic mutations to the corresponding phenotype. While a disrupted protein balance can cause disease, it can also indicate a pathologically altered phenotype. Therefore, the proteomic landscape of cells and organisms can be assessed to quantify such modified functional states. Thus, proteomic data is often used as disease biomarkers, contributing to molecular diagnostics and monitoring disease progression and treatment.

1.5.1 Protein-based biomarkers for disease and health

A biomarker is a biological measurement that indicates normal or abnormal biological processes and can be utilized to quantify different states and responses[87]. They are typically used for disease diagnosis, monitoring disease progression and quantifying response to treatment. Biomarkers should be both sensitive and specific. Sensitivity refers to the biomarker's ability to detect a wide range of values, while specificity refers to its ability to accurately identify the intended biological process or condition. To minimize the risk of causing permanent damage when measuring biomarkers, it is beneficial to use non-invasive or minimally invasive methods. Biofluid samples, such as plasma, serum, urine, and cerebrospinal fluid, meet these criteria. While there are limited options for quantifying gene expression on the mRNA level, proteomic technologies, such as mass spectrometry, allow for the identification of thousands of proteins in these samples[88]. Therefore, proteomics is a promising resource for finding biomarkers of disease state and disease progression.

The process of discovering a novel protein biomarker for a disease typically involves multiple steps. The discovery phase is an unbiased and semi-quantitative process, in which differences between disease states are quantified utilizing *in vivo* or *in vitro* models, including mice, rats, zebrafish, a variety of immortalized cell lines or primary cells, as well as samples from human individuals. Biomarker identification can then be accomplished through classical statistical univariate approaches that test the difference in abundance between disease states or through multivariate feature selection methods. The latter aim to identify subsets of proteins that predict a given outcome and often result in higher predictive performance due to integrating multiple measurements and enabling interactions between different features in a multivariate model. The selected candidates must undergo multiple rounds of validation to ensure their robustness, sensitivity, and specificity. This is typically accomplished through targeted data generation or the use of existing public datasets. The selection of experimental approaches for validation varies depending on the type and size of the model generated. For example, single proteins may be characterized using sensitive techniques such as immunoaffinity peptide enrichment to test the precision across thousands of samples to define a better estimate of the variation within the population. Commercialization of biomarkers is then evaluated based on its efficiency, cost effectiveness, and accuracy.[89,90]

Protein biomarkers are often the presence or abundance of a specific protein. One example of such a biomarker is cardiac troponin 1, which is a sensitive and specific biomarker that can be measured in the blood to detect damage to the heart muscle[91]. When used in combination with other cardiac biomarkers, such as creatine kinase, it can aid in the diagnosis of a heart attack or myocardial infarction[92]. Proteins involved in the immune response, such as C-reactive protein, can serve as biomarkers for infection and internal organ damage as their levels increase during (acute)

inflammations[93,94]. Similarly, hormone or signaling molecules, such as growth hormones can be used to diagnose and monitor various growth disorders, acting as protein abundance-based biomarkers[95]. Enzyme activity is another type of protein biomarker, measuring the level of activity instead of the abundance of a specific protein. For example, the genetic disorder Gaucher's disease is difficult to diagnose by sequencing as it can result from many different gene mutations. However, measuring the activity of the beta-glucocerebrosidase enzyme in the blood provides a reliable diagnosis and is the standard test used when the disease is suspected[96]. Protein biomarkers are commonly used in clinical practice to identify, classify, and monitor various diseases. Biomarkers based on the presence or abundance of a protein or enzymatic activity are commonly used as they are easily measurable. Protein biomarkers based on PTMs or PPIs on the other hand are much rarer, but are still applied. An example of a PTM-based biomarker is the HbA1c test, which quantifies the glycosylation level of hemoglobin and is used to monitor the non-acute presence and severity of hyperglycemia in people with diabetes[97].

Protein biomarkers are therefore an essential component of clinical practice. They impact current biology and can represent an individual's current health state. As a result, they can be used for diagnosis, progression prediction, and treatment monitoring.

1.5.2 LiP-MS as a novel tool for discovering structural protein biomarkers

Robust biomarkers are commonly used to diagnose disease, measure disease progression, or monitor response to treatment. However, the identification and development of new biomarkers remains crucial. For instance, many diseases like Parkinson's disease (PD), Alzheimer's disease (AD), and chronic fatigue syndrome (CFS) currently lack reliable biomarkers for early and definitive diagnosis. This not only complicates the diagnostic process, but also leads to numerous incorrect diagnoses[98]. While some rare diseases may not have been investigated for biomarkers yet, there has been a lack of success in identifying robust ones for others, such as PD, AD, and CFS despite extensive research. For certain diseases, traditional biomarkers based on ribonucleic acid (RNA) transcript levels, PTMs or protein abundance may not have sufficiently strong and robust distinctions between healthy and diseased individuals to be useful as biomarkers.

As structures of proteins closely reflect their functional state, measuring dynamic alterations in protein structure can provide valuable insights into pathological states that may not be evident in other omics data. Currently, there are only a few biomarkers or candidate biomarkers that depend on the three-dimensional structure of proteins. These biomarkers were not discovered through an unbiased, semiquantitative process in the typical discovery phase. Instead they rely on prior knowledge about the pathology they intend to classify. For example, sporadic Creutzfeldt-Jakob disease is caused by the misfolding of the prion protein and the recently developed RT-QuIC test diagnoses this disease by detecting the pathological misfolded form of the prion protein in an individual's CSF with sufficient sensitivity and specificity[99].

Methods used in the discovery phase of protein biomarker development are typically blind to structural events such as altered protein conformation, binding of small molecules, PPIs, or protein misfolding. Therefore, the high-throughput detection of structural biomarkers has been unachievable, constraining structural (candidate) biomarkers to rely on already known pathological processes. With LiP-MS there is a tool that can be applied to identify disease-related changes in the structural accessibility of proteins. Its high-throughput nature allows for the inference of candidate biomarkers

from the complete proteome of samples. However, a tailored bioinformatic approach is necessary to deal with the specific challenges of LiP-MS analysis (see section 1.6). These novel biomarker candidates rely on structural changes in the proteome and may measure pathological changes that were previously not considered , as they are not reflected in other omics data.

1.5.3 Parkinson's disease – a poster child for structural protein biomarker discovery

We used LiP-MS to quantify novel candidate biomarkers that rely on alterations in the structural accessibility of proteins in individuals with PD.

PD is the second most common neurodegenerative disease, affecting over 8.5 million people worldwide as of 2019 – a prevalence that has doubled in the past 25 years[100]. Despite its high prevalence, diagnosing PD remains challenging due to the lack of reliable biomarkers and the heterogeneity of its phenotypes. This results in a relatively high number of misclassified individuals, particularly in the early stages of the disease[98,101]. Discovering biomarkers to predict, diagnose, and monitor disease progression and response to treatment in PD would be highly valuable, but has so far been unsuccessful despite many efforts[102]. The rationale for searching for structural candidate biomarkers in PD is based on the fact that PD is a prominent example of a disease associated with structural proteome changes. Although the pathology of PD remains largely unclear, it is known that the three-dimensional conformation of the presynaptic protein α-Syn plays a central role in the PD pathology[103]. Misfolding and aggregation of α-Syn into so-called Lewy bodies in specific brain regions (substantia nigra & cortex) are key pathological driver of PD. However, this is not a feasible biomarker for diagnosis of PD as collecting brain samples is not feasible in practice.

LiP-MS was utilized to identify potential biomarkers in the early stages of PD by analyzing proteome-wide structural accessibility changes in the CSF of affected individuals. CSF is an easily accessible biofluid, as samples can be obtained with minimally invasive techniques, making it a promising source of (candidate) biomarkers. CSF is a unique medium for detecting biochemical changes in the central nervous system as it is in direct contact with the extracellular space of the brain. It is involved in the clearance of toxic molecules from the brain and it has been suggested that CSF physiology may be associated with the development of neurodegenerative diseases[88]. Changes in the CSF protein composition can indicate pathological changes in the central nervous system (CNS)[104,105], making it a valuable source for PD biomarkers, as the brain is expected to harbor strong signals of PD pathology. In this context, previous studies have identified oligomeric and aggregated states of α-Syn in the CSF of individuals with PD and have demonstrated that the ratio of oligomeric to total α-Syn in the CSF is linked to the disease state.

We hypothesized that changes in the brain's proteome structure could be detected in cerebrospinal fluid (CSF) samples. To test this hypothesis, we applied LiP-MS (double-digest and trypsin-only digest) to CSF samples from a well-characterized cohort of individuals with early-diagnosed Parkinson's disease (PD) (n = 52) and age-matched healthy donors (n = 51)[106–110]. We also subjected a second small, independent cohort with post-mortem brain samples to LiP-MS to validate the signals derived from the CSF data. This is the first human cohort to undergo LiP-MS analysis, and the first study with the goal to quantify disease-related structural alterations from mammalian in vivo samples. Therefore, it was crucial to develop a tailored bioinformatic pipeline, dealing with the challenges in LiP-MS analysis (see section 1.6), to identify structural changes in the CSF of individuals with PD. This study constitutes the first chapter of my thesis.

1.6 Challenges in the analyses of LiP-MS data

Analyzing LiP-MS data poses many challenges including those commonly encountered with molecular high-throughput data. Challenges such as the need to account for technical variations, including batch effects or biological variables, that are irrelevant to the current research question, such as the sex of the sampled individuals, or multiple testing correction are frequently faced in omics analysis. Importantly, LiP-MS presents unique challenges due to its data structure, requiring novel computational approaches.

1.6.1 Separating signals of protein structural changes from non-structural variation

Differences in the LiP peptide signals cannot be solely attributed to the PK accessibility of the proteins. This opposes challenges when aiming to define structural accessibility changes as PK-independent signals due to underlying molecular events (i.e., changes in protein abundance or PK-independent peptide effects resulting from, for example, alternative splicing or PTMs) may also vary between conditions and contribute to the measured LiP peptide intensities. Differentiating between the LiP signal resulting from changes in the structural accessibility of the proteome and LiP variation generated by biological signals is crucial.

In mass spectrometry experiments, differences in protein abundance between two conditions result in variations in the intensity of the corresponding peptides[111], and protein abundance differences will therefore be reflected in the LiP peptide measurements. This does not play into account when both conditions of a LiP-MS experiment are derived from the same sample. For instance, in LiP-MS experiments that measure protein-small molecule binding interactions, a single batch of cell lysate is divided into two groups: one group receives a small molecule, while the other does not. Since the cell lysate used in both conditions is identical, the protein abundance remains constant. Therefore, any differences in LiP peptide intensities between the conditions are independent of protein abundance variation. However, in an increasing number of LiP-MS experiments performed today, different conditions are not derived from the same biological sample. In these, differences in the protein abundance frequently occur between these different conditions, such as healthy vs diseased[112]. It is crucial to take these differences into account when deriving structural accessibility variations between conditions from LiP peptide intensities to avoid mistaking changes in protein abundance for structural variations in the proteome.

The same principle applies to PK-independent peptide variation, i.e., peptide variation between conditions that is not due to difference in protein abundance or altered PK cleavage, and is included in the LiP peptide measurements. These effects can be caused by covalent modifications, such as PTMs. While the presence or absence of PTMs can affect a protein's abundance and its three-dimensional conformation, it also alters the intensity of the specific peptide with the PTM itself. Mass spectrometry results typically only quantify the version of a peptide without the PTM. Therefore, an increase in the occurrence of a PTM at a specific peptide will result in a decrease of the relative measured intensity of that peptide[113]. Another possible cause of variation in PK-independent peptides is alternative splicing. Different peptides may be present in different isoforms of a protein. If the relative abundance of different isoforms of a protein changes, peptides that are not present in all isoforms will also be more or less abundant, even if the protein abundance remains stable. LiP-MS experiments with different samples in different condition groups are expected to have condition-specific PK-independent peptide effects that may not be reflected in the abundance of the corresponding protein. For instance, when using LiP-MS to study disease-related structural changes,

disease-associated peptide effects caused by PTMs and alternative splicing will also occur[114,115]. To accurately infer structural changes based on the variation in PK accessibility of peptides, it is crucial to account for variation in the LiP peptides that can be attributed to PK-independent peptide effects.

1.6.2 Occurrence and overlapping of fully-tryptic and half-tryptic peptides

While the trypsin digest produces only fully-tryptic peptides (i.e., peptides digested by trypsin at both the C- and N-terminus), the PK digest in the LiP samples also produces half-tryptic peptides (i.e., peptides digested by PK at one terminus and trypsin at the other). For many proteins, both fully and half-tryptic peptides can be measured but these types vary in their sequence coverage. Therefore, it is possible that certain protein regions may not be covered by either type of peptide. Other regions may only be covered by fully or half-tryptic peptides or may be covered in complex patterns of both types of peptides.

A single fully-tryptic peptide (Figure 3a, peptide 1) may have several corresponding half-tryptic peptides (Figure 3a, peptides 2-7). These can result from the same PK cleavage event, as a single PK cleavage of the fully-tryptic peptide results in two half-tryptic peptides (Figure 3a, peptides 2-3). However, in LiP-MS data it is rather rare that both half-tryptic peptides are quantified, more often only one of the two is present in the data (Figure 3a, peptides 4-7). This is often the half-tryptic peptide cleaved by trypsin at the C-terminus (Figure 3a, peptides 4-6), since this peptide contains at least one protonatable amino acid side chain and can therefore be quantified in the mass spectrometer. Half-tryptic peptides with trypsin at the N-terminus (Figure 3a, peptide 7) are quantifiable only if they have at least one protonatable AA in their sequence. IF the half-tryptic peptides resulting from the double digest are too short, they will not be listed in the LiP-MS data due to their inability of being uniquely assigned to their parent protein. Furthermore, PK digest can also result in non-tryptic peptides (i.e., peptides digested by trypsin at both the C- and N-terminus) and these are typically not detected.

Further adding to the complexity of the peptide pattern in LiP-MS data, is the fact that trypsin does not cleave perfectly, resulting in peptides with missed cleavage sites (Figure 3b). A region of a protein may be covered by multiple fully-tryptic peptides (Figure 3b, peptides 8-9), some of which have missed cleavage sites. In addition, a protein may only be covered by a fully-tryptic peptide with a missed cleavage site as the fully tryptic peptide without missed cleavage sites is not present in the data (Figure 3b, peptide 8, positions 0-5). This further increases the complexity of the half-tryptic peptides (Figure 3b, peptides 10-15), since they can also contain missed trypsin cleavage sites (Figure 3b, peptides 14-15). Thus, some half-tryptic peptides (Figure 3b, peptides 11-13) overlap with several fully-tryptic peptides (Figure 3b, peptides 8-9), while other half-tryptic peptides only overlap with one fully tryptic peptide that may additionally not contain the same used trypsin-cleavage site (Figure 3b, peptide 8-10). Half-tryptic peptides can also only partially overlap with fully-tryptic peptides (Figure 3b, peptides 15 vs 8-9).

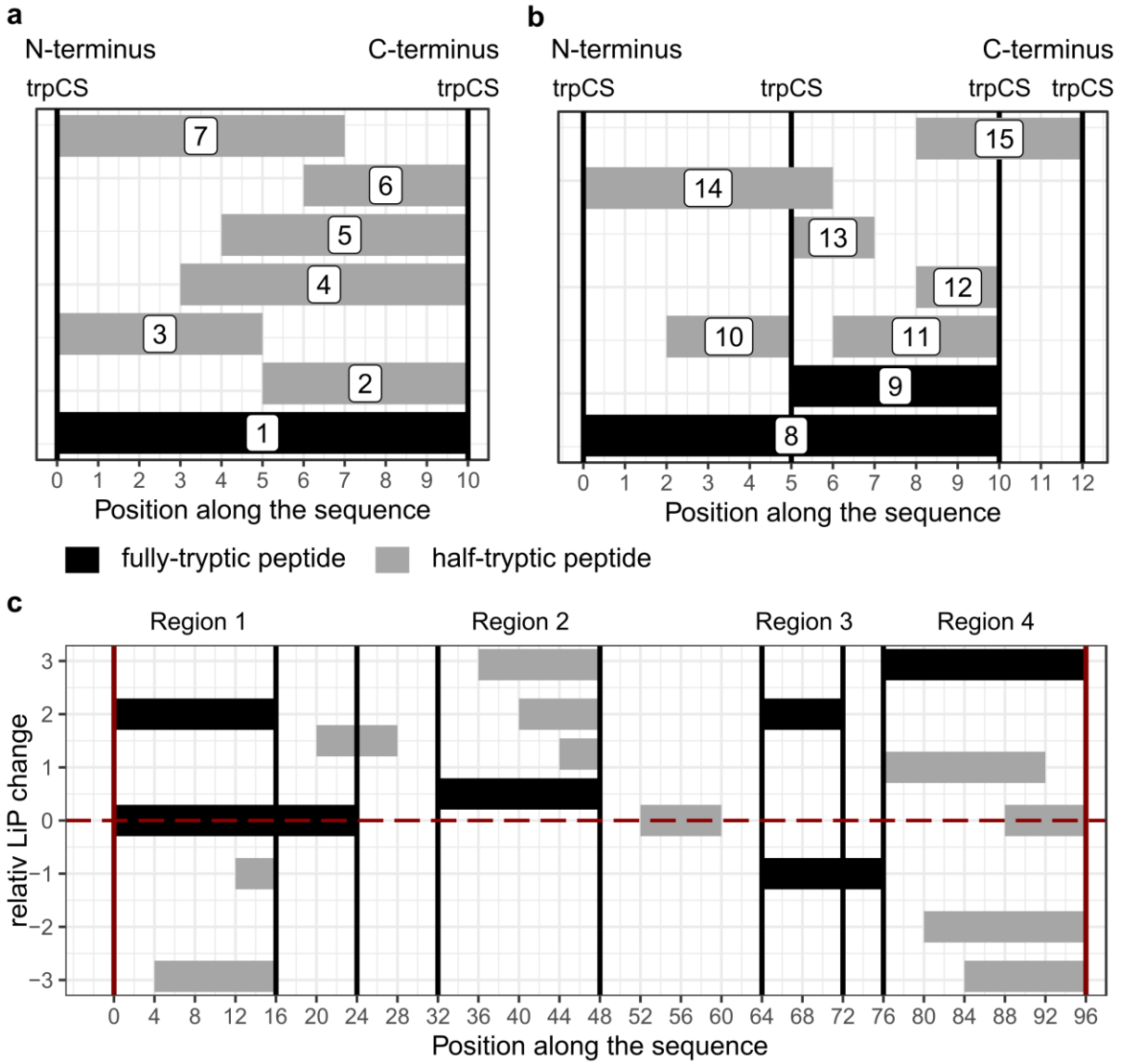


Figure 3: Examples of how fully-tryptic and half-tryptic peptides map over protein regions. a) *Schematic overview of fully-tryptic (black) and half-tryptic (gray) peptides spanning a protein region (N-terminus: left, C-terminus: right) with two trypsin cleavage sites (trpCS, black vertical lines).* ***b)*** *Schematic overview of fully-tryptic (black) and half-tryptic (gray) peptides spanning a protein region (N-terminus: left, C-terminus: right) with multiple trypsin cleavage sites (trpCS, black vertical lines).* ***c)*** *Schematic overview of the relative LiP change in fully-tryptic (black) and half-tryptic (gray) peptides spanning a complete protein with multiple trypsin cleavage sites (trpCS, black vertical lines). Peptides at the start and end (red vertical lines) of the protein sequence that have one trypsin cleavage site, are annotated as fully-tryptic as there is no PK-cleavage involved in the emergence of these peptides. A dashed red line at a relative LiP change of 0 is displayed.*

The structural accessibility of a given LiP peptide can be estimated by a simple log2 fold change approach between two conditions, resulting in relative LiP changes. A relative LiP change around 0 indicates that there is no difference in the intensity of the LiP peptide between the conditions, which can be interpreted as no difference in accessibility between the conditions, regardless of whether the peptide is fully or half-tryptic. A positive relative LiP change results, if a peptide is more abundant in condition 1 compared to condition 2. In case of a fully-tryptic peptide this means that more PK cleavage has occurred in condition 2, resulting in fewer fully-tryptic versions of the peptide. This may indicate an increased accessibility of this protein region in condition 2. A positive relative LiP change

in a half-tryptic peptide, on the other hand, means that there was an increase in PK cleavage in Condition 1 and therefore a decrease in accessibility in Condition 2. For peptides with a negative relative LiP change, the opposite is true; fully-tryptic peptides show a signal of decreased accessibility and half-tryptic peptides show an increased accessibility in condition 2 compared to condition 1. The general expectation is that half-tryptic peptides in a region where the fully-tryptic peptide, for example, shows a positive relative LiP change, will change in the opposite direction, showing a negative relative LiP change (Figure 3c, region 1). However, the pattern resulting from the relative LiP changes is often less straightforward and more ambivalent. For example, several half-tryptic and fully-tryptic peptides in the same region of a protein may all show a relative LiP change in the same direction (Figure 3c, regions 2-4). Surprisingly, the measured half-tryptic peptides are more abundant in the same condition as their corresponding fully-tryptic peptides. This indicates a more complex pattern of PK cleavage; differences in the structure of a protein could induce both more and less PK cleavage at different AA residues within the same fully-tryptic peptide regions. Some of the half-tryptic peptides may undergo further PK digest or be too small to quantify from the MS data, therefore parts of the LiP signal which explain the directionality of the fully-tryptic peptides may be missing. It is also possible that overlapping fully tryptic peptides are not consistent in their directionality (Figure 3c regions 1-3).

1.6.3 Approaching challenges in the analysis LiP-MS data

LiP peptide intensities are influenced by variations in PK accessibility, PK-independent peptide levels, and protein abundance. Therefore, a tailored computational approach is necessary to separate the signal of interest, i.e., the PK accessibility variation, from the other factors. The protein abundance can be accounted for by removing the variation introduced by the Trp protein abundance, estimated from the corresponding Trp peptides, from the LiP peptide intensities. This has been performed in many but not in all LiP-MS studies[49,75,84,104,116]. Correction for protein abundance can be applied to both the fully and half-tryptic peptides, as both are dependent on the protein abundance and can be matched to a specific protein.

Current analyses of LiP-MS data often fail to account for variation in Trp peptides that is independent of protein abundance. As a result, structural accessibility variation in peptides cannot be distinguished from PK-independent peptide variation. To address this issue, Trp peptide intensities should be taken into consideration. PK-independent peptide effects present in a fully-tryptic LiP peptide are expected to also be reflected in the peptide intensity in the corresponding Trp data, i.e., in the same fully-tryptic peptide in the trypsin-only digest of the same sample. Therefore, correcting the LiP peptide intensities for the variation present in the corresponding Trp peptide intensities has the potential to remove PK-independent peptide variation for the LiP signal. This would enable the differentiation of LiP variation caused by structural differences from other peptide-specific variations caused by factors such as PTMs and splicing.

This approach cannot be applied to half-tryptic peptides since they do not exist in the trypsin-only controls. Matching half-tryptic to fully-tryptic peptides is not a trivial task because some half-tryptic peptides have no corresponding fully-tryptic peptides, while others match to multiple fully-tryptic peptides. In addition, a PTM can occur in a part of the peptide that is only covered by either the fully-tryptic or the half-tryptic peptide. The variation in PK-independent peptides is present either (1) exclusively in the fully-tryptic peptide or (2) exclusively in the half-tryptic peptide, but not in both. This presents a risk that (1) the PK-independent signal may be mistakenly added to the half-tryptic peptide

signal during the correction step or (2) the signal cannot be removed from the half-tryptic peptide, resulting in a falsely identified structural change in the half-tryptic peptide.

To the best of our knowledge, no studies have investigated and compared different existing approaches that deal with these challenges. Therefore, the impact of not correctly accounting for, e.g., the protein abundance effect in LiP peptide intensities on the study results remains unknown. Additionally, there is overall a general lack of a comprehensive and accessible computational pipeline that is robust and is applicable to various LiP-MS experiments. To ensure accurate analysis of LiP-MS data, it is important to account for additional factors such as technical effects and biological cofactors like sex that can introduce unwanted variation in the LiP signal. Therefore, a computational approach used for analyzing LiP-MS data must include a method to remove this unwanted variation. The comparison of approaches for analyzing LiP-MS data, and the implementation of a novel computational pipeline to distinguish the origin of variation in the LiP signal and subsequently identify the signal caused by differences in PK accessibility in fully-tryptic and half-tryptic peptides, are both documented in the second chapter of this thesis.

2 Chapter I: Global, in situ analysis of the structural proteome in individuals with Parkinson's disease to identify a new class of biomarker

Contribution statement

I designed the bioinformatic pipeline, inferred differences in the proteome of the CSF and brain cohort and performed all the statistical tests on the data, such as GO enrichment. I built and evaluated the classification models based on the CSF proteome. All these analyses were performed with conceptional input from Jan Grossbach and Andreas Beyer. I was involved in the design of all Figure panels and created the majority of them myself (Figure 2a-f,h; Figure 3; Figure 4a,d,f,g,i,j,l; Figure 5, Ext. Data Figure 1b,c; Ext. Data Figure 2a-g, Ext. Data Figure 3, Ext. Data Figure 4c,f,h,i,k,l,n,o; Ext. Data Figure 5). Sample preparation, limited proteolysis and LC-MS/MS data acquisition was performed by Marie-Therese Mackmull. I, together with other co-authors, wrote and reviewed the manuscript.

Further contributions can be found in the contribution statement of the paper.

Data and code availability

All data not included in the thesis such as the peer review file, supplementary tables, and source data files can be accessed over:
https://www.nature.com/articles/s41594-022-00837-0#Sec52.

Code for the main analyses has been deposited on GitHub:
https://github.com/beyergroup/Global-analyses-of-the-human-structural-proteome-to-identify-a-new-type-of-disease-biomarker.

nature structural & molecular biology

Article

https://doi.org/10.1038/s41594-022-00837-0

Global, in situ analysis of the structural proteome in individuals with Parkinson's disease to identify a new class of biomarker

Received: 30 April 2021

Accepted: 18 August 2022

Published online: 12 October 2022

Check for updates

Marie-Therese Mackmull[1,9], Luise Nagel[2,9], Fabian Sesterhenn[1], Jan Muntel[3], Jan Grossbach[2], Patrick Stalder[1], Roland Bruderer[3], Lukas Reiter[3], Wilma D. J. van de Berg[4,5], Natalie de Souza[1,6], Andreas Beyer[2,7,8] ✉ and Paola Picotti[1] ✉

Parkinson's disease (PD) is a prevalent neurodegenerative disease for which robust biomarkers are needed. Because protein structure reflects function, we tested whether global, in situ analysis of protein structural changes provides insight into PD pathophysiology and could inform a new concept of structural disease biomarkers. Using limited proteolysis–mass spectrometry (LiP–MS), we identified 76 structurally altered proteins in cerebrospinal fluid (CSF) of individuals with PD relative to healthy donors. These proteins were enriched in processes misregulated in PD, and some proteins also showed structural changes in PD brain samples. CSF protein structural information outperformed abundance information in discriminating between healthy participants and those with PD and improved the discriminatory performance of CSF measures of the hallmark PD protein α-synuclein. We also present the first analysis of inter-individual variability of a structural proteome in healthy individuals, identifying biophysical features of variable protein regions. Although independent validation is needed, our data suggest that global analyses of the human structural proteome will guide the development of novel structural biomarkers of disease and enable hypothesis generation about underlying disease processes.

The rise of chronic diseases in aging populations necessitates a deeper understanding of disease pathophysiology and robust biomarkers for early disease detection and stratification of individuals with disease[1]. Studies of disease by MS-based proteomics typically identify varying protein abundances, post-translational modifications (PTMs), or isoform prevalence[2–6]. These global analyses are, however, blind to many molecular events that affect protein function, such as the binding of small molecules, protein-protein interactions, misfolding, and protein conformational changes. Reasoning that most molecular events that affect protein function would affect protein structure[7,8], we tested

[1]Institute of Molecular Systems Biology, Department of Biology, ETH Zurich, Zurich, Switzerland. [2]Cluster of Excellence Cellular Stress Responses in Aging-associated Diseases (CECAD), University of Cologne, Cologne, Germany. [3]Biognosys AG, Schlieren, Switzerland. [4]Amsterdam UMC location Vrije Universiteit Amsterdam, Section Clinical Neuroanatomy and Biobanking, Department Anatomy and Neurosciences, Amsterdam, the Netherlands. [5]Amsterdam Neuroscience, Neurodegeneration, Amsterdam, the Netherlands. [6]Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland. [7]Faculty of Medicine and University Hospital of Cologne, and Center for Molecular Medicine Cologne, University of Cologne, Cologne, Germany. [8]Institute for Genetics, Faculty of Mathematics and Natural Sciences, University of Cologne, Cologne, Germany. [9]These authors contributed equally: Marie-Therese Mackmull, Luise Nagel. ✉e-mail: andreas.beyer@uni-koeln.de; picotti@imsb.biol.ethz.ch

whether global analyses of the human structural proteome reflected pathological alterations and whether this could yield a novel type of structural biomarker of disease.

To detect protein structural alterations directly in biofluids of participants in clinical cohorts, we have used LiP–MS, our previously developed structural-proteomics approach. LiP–MS probes structural alterations on a proteome-wide scale in complex, near-native contexts[9,10], and captures a variety of functional alterations (for example, allostery, enzyme activation, active-site occupancy, chemical modification, aggregation) with the resolution of single functional sites[11]. In brief, LiP–MS uses a non-specific protease to generate structure-specific proteolytic patterns that can be analyzed by MS.

We used LiP–MS to identify structural alterations in the CSF of individuals with Parkinson's disease. PD affects around 1% of the world population over age 60, lacks early diagnostic biomarkers, and presently cannot be cured. Additionally, PD is a prominent example of a disease associated with altered protein structure. The protein α-synuclein plays a central role in PD pathology and forms oligomers and proteinaceous deposits, called Lewy bodies, in the brain[12–14]. Transcriptomics and classical proteomics studies report relatively sparse changes in body fluids of people with PD, some of which are non-specific to PD[15,16]. Previous studies have identified oligomeric and aggregated states of α-synuclein and altered activation states of endolysosomal enzymes in the CSF of people with PD, which show potential as PD protein markers[17–19], in support of our hypothesis that analysis of protein structures might be informative about the disease and may lead to biomarker identification. Altogether, given the lack of early biomarkers and the potential for structural changes accompanying disease onset, we chose PD to test whether global protein structural analysis can identify potential biomarkers.

We applied LiP–MS to CSF from a well-characterized cohort of individuals with early-diagnosed PD relative to age-matched healthy donors[17,19–22]. CSF is in constant exchange with brain tissue and therefore may harbor brain-derived structurally altered proteins. We used a statistical model to account for age, sex, protein abundance, and technical confounding factors, and identified 76 proteins that were structurally altered in individuals with PD relative to healthy donors. These candidate structural biomarkers were enriched for processes that are known to be misregulated in PD, such as synapse maintenance and acetylcholine metabolism. Notably, a combinatorial subset of structurally informative CSF peptides distinguished individuals with PD from healthy participants with better performance than protein abundance data, and provided complementary information to measures of the PD-associated protein α-synuclein[23]. Further, more than half of the detected proteins with structural changes in the CSF were also structurally changed in brain samples from a small, independent cohort, indicating that the approach has the potential to capture pathological processes; several of these proteins have been previously linked to PD. Lastly, analysis of the inter-individual variability of the human structural proteome using data from healthy individuals showed that more than 60 human CSF proteins had regions with high variability between individuals. Variable regions were more disordered, were more accessible to solvents, and had a higher propensity for mediating protein-protein interactions than non-variable regions.

Taken together, our findings identify structurally altered proteins that distinguish individuals with PD from healthy donors. A subset of these proteins might inform on disease mechanisms in the brain and help to define the molecular profile of PD, although clinical validity will require testing in larger independent cohorts. We propose that protein structures have great potential to reveal pathological processes in PD and could serve in the future as a novel type of biomarker of disease.

Results

Identification of structural changes in human cerebrospinal fluid

To probe the concept of structural biomarkers of disease, we systematically investigated whether structural changes are detectable in the CSF proteome of individuals with PD relative to that of healthy individuals. Our cohort consists of 52 individuals diagnosed with sporadic PD (PD group, or PDG), with an average disease duration of 5.8 years, and 51 healthy individuals (HG) (Extended Data Fig. 1a) covering the same age range. Both sexes are represented in each cohort group (Extended Data Fig. 1b). Quantitative clinical parameters were available for the PDG, and levels of previously identified PD-associated biochemical markers (levels of total (t-α-Syn), phosphorylated (p-α-Syn), and oligomeric α-synuclein (o-α-Syn)) had been previously collected for the complete cohort (Extended Data Fig. 1a)[17,19–22,24,25].

CSF samples were collected by lumbar puncture[19] under near-native conditions and were subjected to LiP–MS (Fig. 1a). Each sample was split into two; one aliquot underwent limited proteolysis (LiP) followed by trypsin digestion prior to MS, and the other underwent only trypsin digestion (trypsin-only). We used the trypsin-only data to measure protein-abundance changes and peptide-level changes that could be due to covalent modifications such as PTMs or peptide sequence changes from alternative splicing, RNA editing, endogenous proteolysis, or mutations. The LiP data were used to identify alterations in protein structure.

We used a data-independent acquisition (DIA) strategy combined with high-quality spectral libraries and label-free quantification to monitor protein abundance and structural changes across the cohort. We identified more than 2,100 proteins and more than 48,000 peptides for both LiP and trypsin-only data before filtering (Supplementary Tables 1 and 2), consistent with previous studies[26] but without depletion of abundant proteins. The number of peptides and proteins at various processing and analysis steps is shown in Supplementary Table 3. α-Synuclein was only sparsely detected owing to low endogenous levels, consistent with previous proteomic data[27], and was therefore not included in our analyses. Gene Ontology (GO) enrichment analysis of all identified proteins showed the enrichment of gene sets expected in the CSF (Extended Data Fig. 1c)[28]. We used these data to study the variability of the structural proteome between healthy individuals and to identify structural biomarker candidates by comparing healthy individuals and those with disease (Fig. 1b).

Variability of the healthy human cerebrospinal fluid structural proteome

We first assessed the variability of protein structures in the CSF of the 52 healthy participants. To disentangle changes in LiP peptide intensities due to structural variation from those due to protein abundance variation or to peptide-level variation caused by covalent modification, cleavage, or sequence changes, hereafter collectively termed proteinase K (PK)-independent peptide variation, we first compared the s.d. of LiP peptide intensities with that of the trypsin-only peptide intensities across the healthy samples (Fig. 2a). We then estimated a variability score for each peptide, which provides a measure of variation in LiP peptide intensities due to structural variation, corrected for PK-independent variation. The strong increase in positive variability scores indicates that there is a signal for protein structural variability in addition to protein abundance variability (Fig. 2b).

We classified 117 peptides as structurally highly variable (1.2%, variability score > 1 with $P < 1 \times 10^{-5}$), 386 peptides as medium-variable (3.9%, variability score of 0.5–1 with $P < 1 \times 10^{-5}$), and 9,385 peptides (all others), derived from 994 unique proteins, as non-variable across healthy individuals (Supplementary Tables 4 and 5). The 117 highly variable peptides originated from 64 unique proteins, suggesting that ~6.5% of the detected CSF proteins show at least one structurally highly variable region across healthy individuals, in addition to any changes due to covalent modifications, cleavage, or sequence alterations, which were removed in this analysis. Functional enrichment analysis of these structurally variable proteins identified terms across a range of cellular and neuronal functions.

https://doi.org/10.1038/s41594-022-00837-0

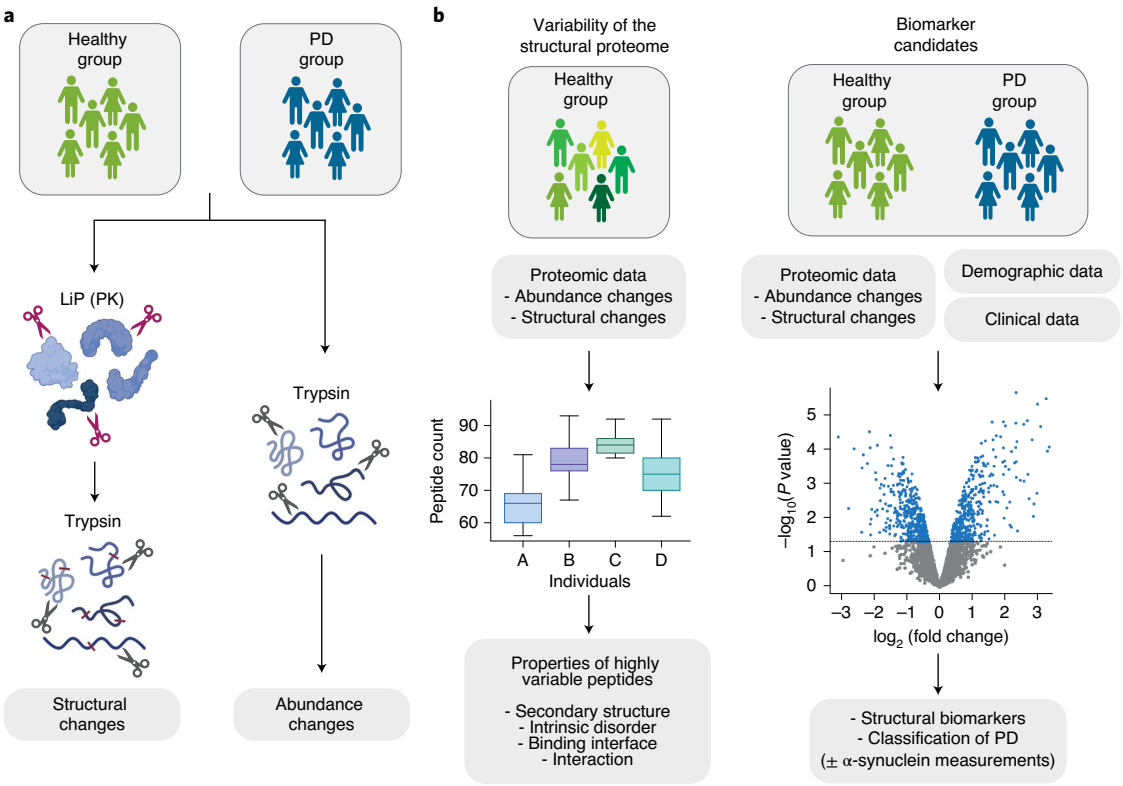


Fig. 1 | Schematic overview of the study. a, Schematic overview of the experimental pipeline for probing structural changes in the CSF of healthy and PD cohort groups. **b**, Overview of the data analysis pipeline for the identification of variable peptides in the CSF of healthy individuals, and for the identification of structural biomarker peptides that vary between healthy individuals and those with PD.

Regions associated with highly variable peptides had a slightly lower propensity to form beta strands and a higher propensity to form alpha-helices than did non-variable regions, but had the same propensity to be part of loops (Extended Data Fig. 2a). All variable peptides had a higher predicted propensity to bind other proteins (Fig. 2c)[29], but not nucleic acids (Extended Data Fig. 2b,c), consistent with their larger solvent-accessible surface area (Extended Data Fig. 2d). Proteins with at least one highly variable peptide showed more physical and functional high-confidence interactions (STRING confidence score > 0.9, Extended Data Fig. 2e)[30], but did not differ in their sequence length or number of domains compared with all other analyzed proteins (Extended Data Fig. 2f,g). Variable protein regions showed significantly higher predicted disorder than non-variable regions (Fig. 2d)[30], providing a structural rationale for their variability in situ. We also noted that 15 of the 117 highly variable peptides showed a bimodal distribution of their LiP intensities (for example, fructose bisphosphate aldolase, Fig. 2f); the remaining peptides showed a unimodal distribution (for example, semaphorin 7A, Fig. 2h) (Supplementary Table 5). These bimodal variable peptides did not show elevated levels of disorder (Fig. 2e). The bimodal distributions potentially reflect two structural states of the corresponding protein, for example, the ligand-bound and ligand-unbound state, and within an individual the protein would exist predominantly in one state. For example, fructose-bisphosphate aldolase showed two highly variable peptides, one bimodal and the other unimodal, in both the healthy and PD groups (Fig. 2f). Although the bimodal peptide mapped close to the enzyme active site (4.8 Å) (Fig. 2g), LiP–MS analysis of the in vitro enzyme suggested that the bimodal distribution does not reflect the substrate-bound versus unbound state (Extended Data Fig. 2h and Supplementary Table 6). The in situ-measured unimodal peptide, however, did significantly change upon substrate addition in vitro, so inter-individual variability at this site could reflect different levels of substrate occupancy. The variable unimodal peptide on semaphorin 7A (Fig. 2h) mapped to an unstructured region (Fig. 2i).

In summary, our structural approach enabled the unprecedented in situ identification and global analysis of CSF proteins with structurally varying regions across healthy individuals. Predicted disorder, surface accessibility, and the propensity for protein-protein interactions are associated with high baseline structural variability. Note that similar variation might be detectable for the same individual over time.

Structural proteomic changes in cerebrospinal fluid of people with Parkinson's disease

We went on to probe differences between CSF proteins of healthy individuals and those with PD. Given the complexity of human cohorts, robust identification of disease-specific changes requires accounting for several covariates. We designed a data analysis pipeline to achieve this (Fig. 3a). We performed multiple linear regression, fitting independent models to predict three types of data: LiP peptide intensities (Supplementary Tables 4 and 7), which probe structural variation, trypsin-only peptide intensities (Supplementary Tables 4 and 8), which indicate PK-independent peptide variation, and protein abundance variation, estimated from the trypsin-only samples (Supplementary Tables 4 and 9). In addition to cohort group membership (HG versus PDG), we included variables representing age and sex for each sample, and the two-way interactions of all the three demographic variables. Furthermore, we corrected for technical covariates, such as total protein concentration measured before sample preparation, and batch effects.

We modeled the intensities of LiP peptides, trypsin-only peptides, and protein abundance as a function of the demographic and technical covariates described above, to correct for these covariates. In the case of LiP peptides, we also included trypsin-only peptides and protein abundance as additional covariates to correct for

https://doi.org/10.1038/s41594-022-00837-0

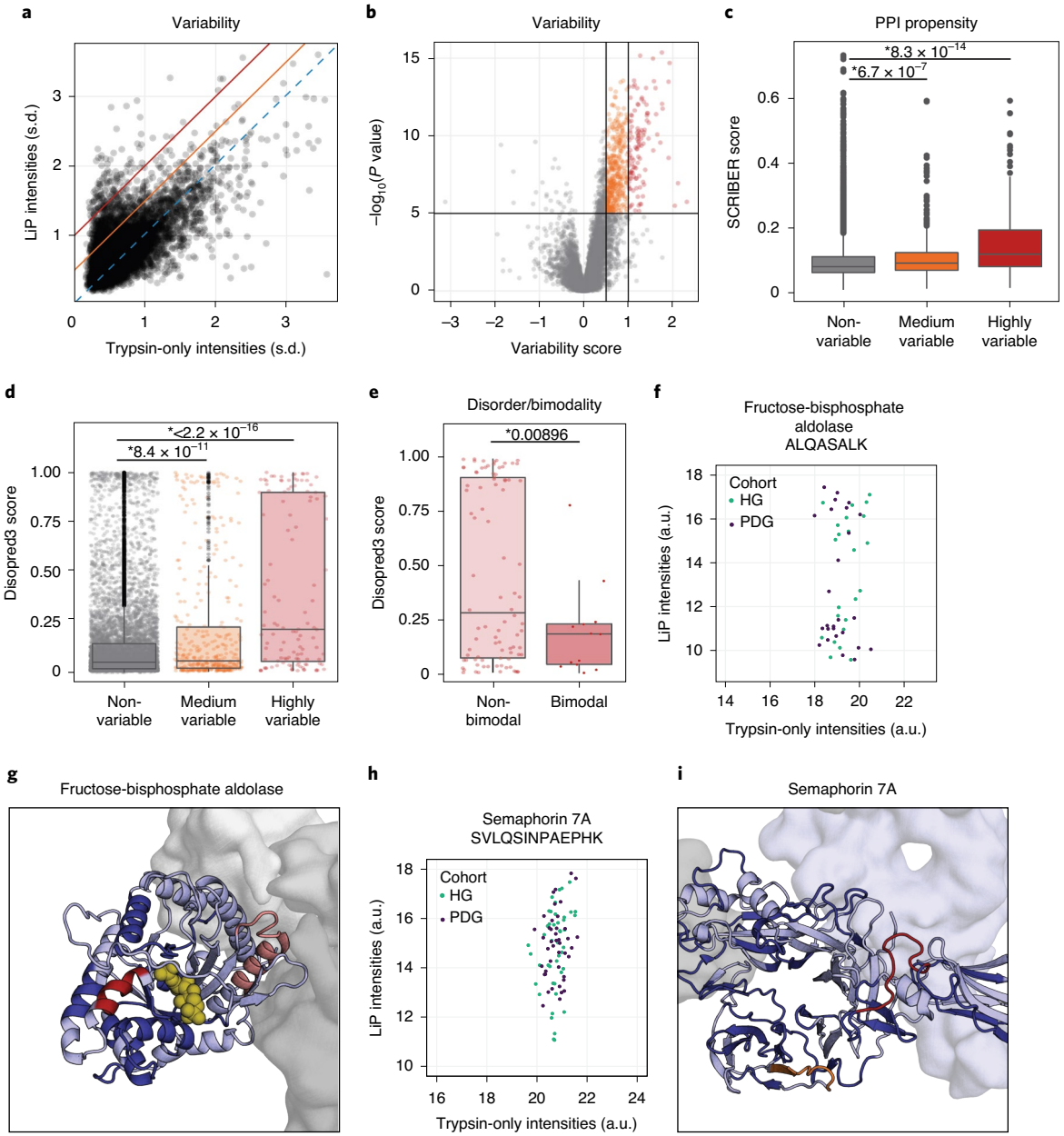


Fig. 2 | Structural variability of the proteome in healthy human CSF. a, The s.d. of LiP peptide intensities against the s.d. of trypsin-only peptide intensities. Each point represents a single peptide. Cut-offs for highly variable peptides (red line) and medium-variable peptides (orange line) are shown. **b**, Volcano plot showing the distribution of variability scores. Each point represents a single peptide. Red, highly variable peptides; orange, medium-variable peptides; gray, non-variable peptides. *P* values were estimated using Levene's test. **c**, SCRIBER scores for the indicated peptide classes. Box plots in all panels: median, center; first and third quantile, lower and upper hinges; largest/smallest value no further than 1.5 × inter-quantile range of the hinge, whiskers; data points beyond are defined as outliers and plotted individually. *P* values are indicated (Wilcoxon rank-sum test; $n = 9{,}385$ non-variable, 386 medium-variable, 117 high-variable peptides, from 51 participants). **d**, Disopred3 disorder scores for the indicated peptide classes. Each point represents a single peptide. *P* values are indicated (Fisher's exact test; *n* values are as in **c**). **e**, Disopred3 scores of highly variable peptides, comparing those with a bimodal distribution and those with a non-bimodal distribution of LiP intensities. *P* values are indicated (Fisher's exact test; $n = 102$ non-bimodal, 15 bimodal peptides from 51 participant). **f**, Distribution of the LiP intensities ($\log_2$) versus trypsin-only intensities ($\log_2$) for the indicated peptide (ALQASALK) from fructose bisphosphate aldolase. Each point represents one participant. **g**, Structure of human brain fructose bisphosphate aldolase (PDB 1XFB) showing one subunit of the homotetramer in light blue (the other three subunits are shown as gray surface). Yellow spheres represent the substrate, based on alignment of PDB 1XFB with the ligand-bound muscle isoform (PDB 4ALD). Highly variable peptides in dark red (bimodal peptide) and salmon (unimodal peptide). **h**, Distribution of LiP intensities ($\log_2$) versus trypsin-only intensities ($\log_2$) for a highly variable peptide (SVLQSINPAEPHK) from semaphorin 7A. Each point represents one participant. **i**, Structure of semaphorin 7A (light blue), in complex with Plexin C1 (gray, PDB 3NVQ). Non-variable peptides are in dark blue, highly variable peptide from **f** are in red, and medium-variable peptides are in orange.

intensity variation that was not due to peptide accessibility changes; this would identify LiP peptide intensity changes due only to protein structural alterations. Likewise, we corrected trypsin-only peptide intensities for protein abundance to distinguish PK-independent peptide-specific effects (for example, PTMs) from whole-protein abundance changes. Finally, we statistically tested the coefficients of all three models to determine significant effects of the cohort, age, and sex variables.

We analyzed the distributions of *P* values of the cohort effect for the three types of models (Fig. 3b) and found that the *P* value distribution for structural variation, but not for PK-independent effects or protein abundance, had more small *P* values for the cohort effect. This suggests that structural changes correlated more strongly with PD state than did the other two measures.

We similarly asked whether the distribution of the *P* values for age or sex showed a signal for structural variation. There was no increase of small *P* values for the sex variable, indicating that we did not detect significant protein structural differences between the CSF of men and women (Fig. 3c and Extended Data Fig. 3a). In contrast, there was an increase in the count of LiP peptides with small *P* values in the distributions representing age, indicating that we detected protein structural changes in CSF upon aging.

To identify potential structural biomarkers of disease, we focused on changes between the healthy and disease cohort groups. After correcting for a potential bias for large proteins (Methods), we selected all peptides with a significant cohort effect (corrected *P* values < 0.05) as candidate structural biomarkers, yielding 88 LiP peptides corresponding to 76 proteins (out of 859 proteins and 7,144 peptides included in the data analysis) (Fig. 3d and Supplementary Table 10). Two of these 76 structurally altered proteins in the CSF of individuals with PD, the alpha-1 (III) chain of collagen (COL3A1) and a peptidyl-glycine alpha-amidating monooxygenase (PAM), correspond to PD-linked genes from genome-wide association studies[31–33]. GO enrichment analysis on the structurally altered proteins yielded terms linked to known PD mechanisms (Fig. 3e), including, for example, several synapse-related terms, which could reflect the synaptic dysfunction and loss associated with numerous motor and non-motor symptoms of PD[34–42]. We note that 87/88 of the peptides indicating structural alterations in PD were classified as non-variable in our analysis of healthy CSF and are therefore not likely to be confounded by background variability.

Using a data-analysis method designed to correct for multiple covariates, we have identified 76 structurally altered proteins between the CSF of healthy individuals and individuals with PD.

Structural proteomic changes in brains of individuals with Parkinson's disease

To assess whether the detected structural changes in CSF reflect pathological events in the brain, we applied our approach (see Methods for analytical differences) to postmortem temporal cortex samples from a small independent cohort (5 individuals with clinically defined and pathologically confirmed PD, 5 age-matched healthy individuals, all female) and asked whether proteins with structural changes in the PD CSF have those changes in the brain as well. As expected, the overlap of detected proteins between the two datasets was low because the CSF has many proteins with extracellular functions (Extended Data Fig. 3b). Out of the 76 proteins with a structural change in CSF, we detected 27 proteins in the brain data. Strikingly, 16 of these 27 proteins showed a structural change in both the brains and CSF of individuals with PD, relative to healthy individuals. At the peptide level, 11 peptides with a structural change in PD CSF compared with healthy CSF were also present in the brain dataset. Most of these structural peptides (8/11) changed accessibility in the same direction (more versus less accessible) in the brain and CSF (Extended Data Fig. 3c), indicating that these structural changes in peptides in PD were conserved between the CSF and brain. Notably, as in the CSF, peptides with structural changes in the PD brain were more strongly correlated with the disease state than were peptides indicating protein abundance changes or other PK-independent peptide changes (Fig. 3f). This confirms that a global structural analysis is a powerful approach to characterize disease states, and that we could validate structural changes detected in the PD CSF in the brains of individuals with PD.

Three proteins that showed structural changes in the PD brain corresponded to PD-linked genes in genome-wide association studies: AP2-associated protein kinase 1 (AAK1)[43], CYFIP-related Rac1 interactor B (CYRIB)[32], and apolipoprotein E (APOE)[44,45], the last of which is a known component of Lewy bodies in PD brains[46,47]. Unlike in the CSF, α-synuclein is detected in the brain data, but we did not detect a statistically significant structural change in α-synuclein, due to a strong outlier individual in the PD group of this small dataset.

Structural alterations of proteins in Parkinson's disease cerebrospinal fluid and brain are linked to disease

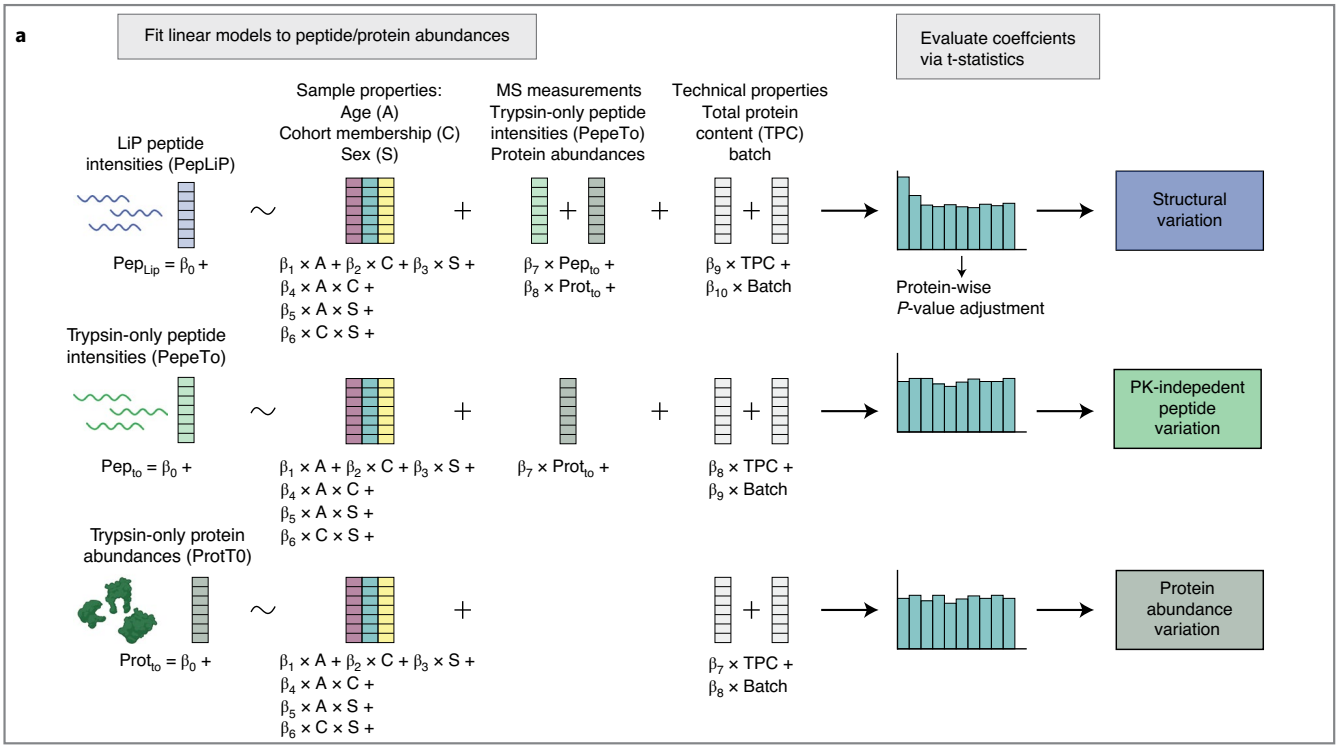
We examined specific structurally altered proteins in more detail, probing their disease link and asking whether structurally altered peptides map to functionally informative sites on the protein structure. We focused initially on proteins that changed structurally in both PD CSF and brain, as these are more likely to be linked to pathological processes. We identified 16 such proteins: almost all (15/16) have been previously associated with PD, and more than half of the proteins show clear links to disease (Supplementary Table 11). For instance, this group includes proteins that regulate neurotransmitter signaling, including that of dopamine (synaptotagmin 1)[48,49], and nicotinic acetylcholine receptors (lymphocyte antigen 6H)[50], cause PD-like phenotypes in animal models (stathmin 2)[51], or are components of Lewy bodies (peroxiredoxin-6)[52].

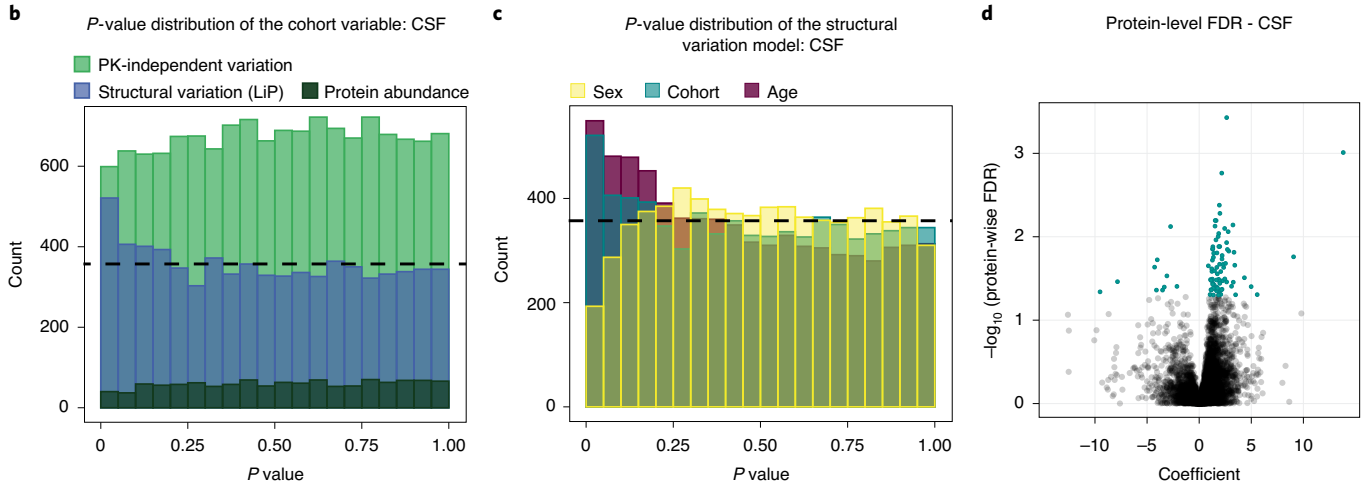
The neuronal cell adhesion molecule (NrCAM) and the PITH domain-containing protein (PITHD1) had reasonable sequence coverage. In the brain and CSF, NrCAM had structural alterations in similar regions (NrCAM, Fig. 4a). For PITHD1, the peptides were identical between the brain and CSF (Fig. 4b–d). In the brain and CSF, both proteins showed accessibility changes in the same direction, further supporting that the structural changes in PD were preserved between these samples. NrCAM, an important protein in neuronal development, is transcriptionally upregulated in the substantia nigra of people with PD[53], PITHD1 is linked to synaptic functions in the olfactory bulb[54,55], and olfactory dysfunctions are early signs of PD. For the Lewy body component PRDX6[52], which has been linked to the PINK1–parkin pathway[56] and worsens dopaminergic neurodegeneration in a PD mouse model[57], we observed a single changed peptide in each of the brain and CSF samples (Extended Data Fig. 4d–f).

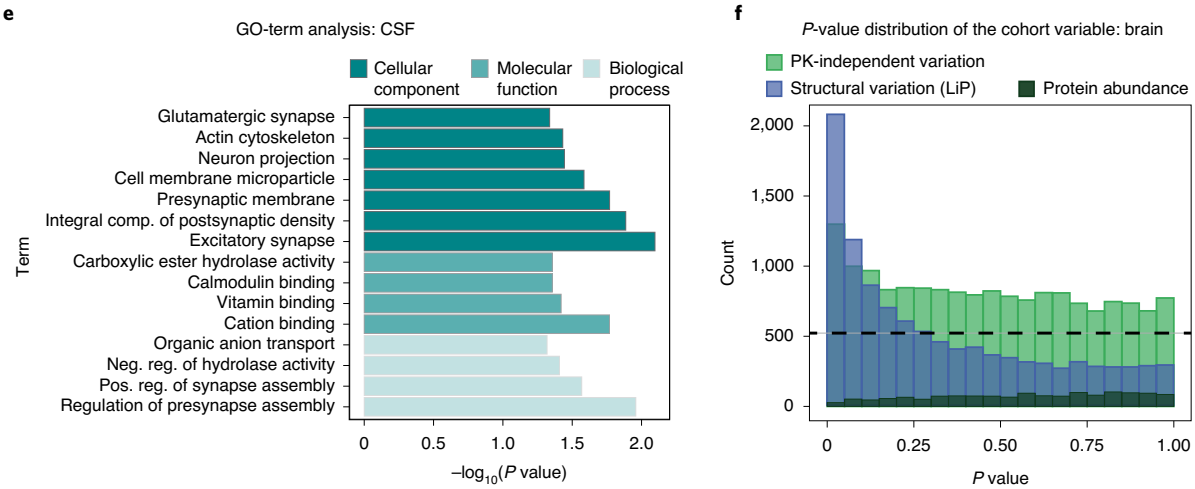
We additionally analyzed structurally altered CSF proteins with good sequence coverage and for which a three-dimensional structure was available. For cerebellin-1 (CBLN1) (Fig. 4e–g), the single altered

Fig. 3 | Identification of proteome structural variations between the healthy and PD cohort groups. a, Data analysis workflow to identify structural peptide variation, PK-independent peptide variation, and protein abundance variation in the CSF between the cohort groups (β = coefficient(s) of the linear model, ~ = equates.). **b**, Histogram showing the results of the analysis of the CSF data, visualizing the *P* values of the cohort variables estimated via *t* statistics from the coefficients of three different types of linear models, indicated by color. Effects based on structural variations (blue), PK-independent peptide variations (light green), and protein abundance variations (dark green) are shown. For all models, the first bar (extreme left) indicates significant (<0.05) *P* values. **c**, Histogram showing the results of the analysis of the CSF data, visualizing the *P* values of the age, cohort, and sex variables, indicated by color, estimated from the coefficients of the structural variations model. Effects based on cohort membership (blue), age (magenta), and sex (yellow) are plotted. Significant *P* values are as in **b**. **d**, Coefficients of each peptide, reflecting cohort membership in the linear model assessing structural variation (compare with the blue histograms in **b** and **c**), are plotted against their corresponding *P* value after applying protein-wise false-discovery rate correction (Benjamini–Hochberg procedure). Each point represents a single peptide; blue indicates candidate biomarker peptides that vary between the healthy and PD cohort groups in the CSF. **e**, GO term analysis of the proteins corresponding to the candidate biomarker peptides in **d**. The enrichment is computed relative to all proteins included in the data analysis. **f**, Histogram corresponding to the analysis of the brain data visualizing the *P* values of the cohort variables estimated via *t* statistics from the coefficients of three different types of linear model, indicated by color. Color and significant *P* values are as in **b**.

https://doi.org/10.1038/s41594-022-00837-0







peptide in PD maps to the center of a homotrimeric complex required for synapse integrity and plasticity[58] and could thus reflect complex disassembly and an alteration of synapse integrity upon PD. Notably, in addition to CBLN1, we identified structural variation in many other synapse-organizing proteins, for example SLIT and NTRK-like protein 5 (SLITRK5) and 1 (SLITRK1), neuronal pentraxin-1 (NPTX1), leucine-rich alpha-2-glycoprotein (LRG1), cell adhesion molecule 1 (CADM1) and 3 (CADM3/NECL-1), neuronal cell adhesion molecule (NrCAM), and disintegrin and metalloproteinase domain-containing protein 11 (ADAM11)[59–61]. Because PD is a known synaptopathy[62], the enrichment for structurally altered synaptic proteins strongly suggests that our analysis captures PD-relevant changes.

Vitamin D deficiency has been reported in PD, including a relationship between the severity of motor symptoms and the level of deficiency[63,64]. To assess whether the structural change we observed for the vitamin-D-binding protein (GC) (Fig. 4h–j) could reflect substrate binding, we compared LiP–MS patterns of the purified enzyme with and without its known binder 1,25 dihydroxy vitamin D (calcitriol), the active form of vitamin D. The single in situ-measured structural peptide also changed upon calcitriol addition in vitro (Fig. 4k–l and Supplementary Table 6), although other calcitriol-dependent peptides were not well covered in situ. These data are therefore consistent with an in situ structural change upon vitamin D (calcitriol) binding, but further validation is needed. We mapped structural changes for several other interesting proteins, including the enzymes phosphoserine aminotransferase (PSAT1) and butyrylcholinesterase (BCH), afamin (AFM), and serum amyloid P-component (SAMP) (Extended Data Fig. 4).

In summary, we observe structural variation in various classes of proteins, including enzymes and proteins involved in synapse organization and function, that are strongly linked to PD. Structurally altered proteins in the CSF of individuals with PD are, when detected, often altered in the PD brain as well, suggesting that some of our candidate structural biomarkers might be linked to pathological processes. Mapping the altered peptides to 3D protein structures suggests hypotheses for functional changes between the healthy and diseased states that could be followed up in mechanistic studies.

Cerebrospinal fluid structural peptides can be used to classify healthy versus Parkinson's disease groups

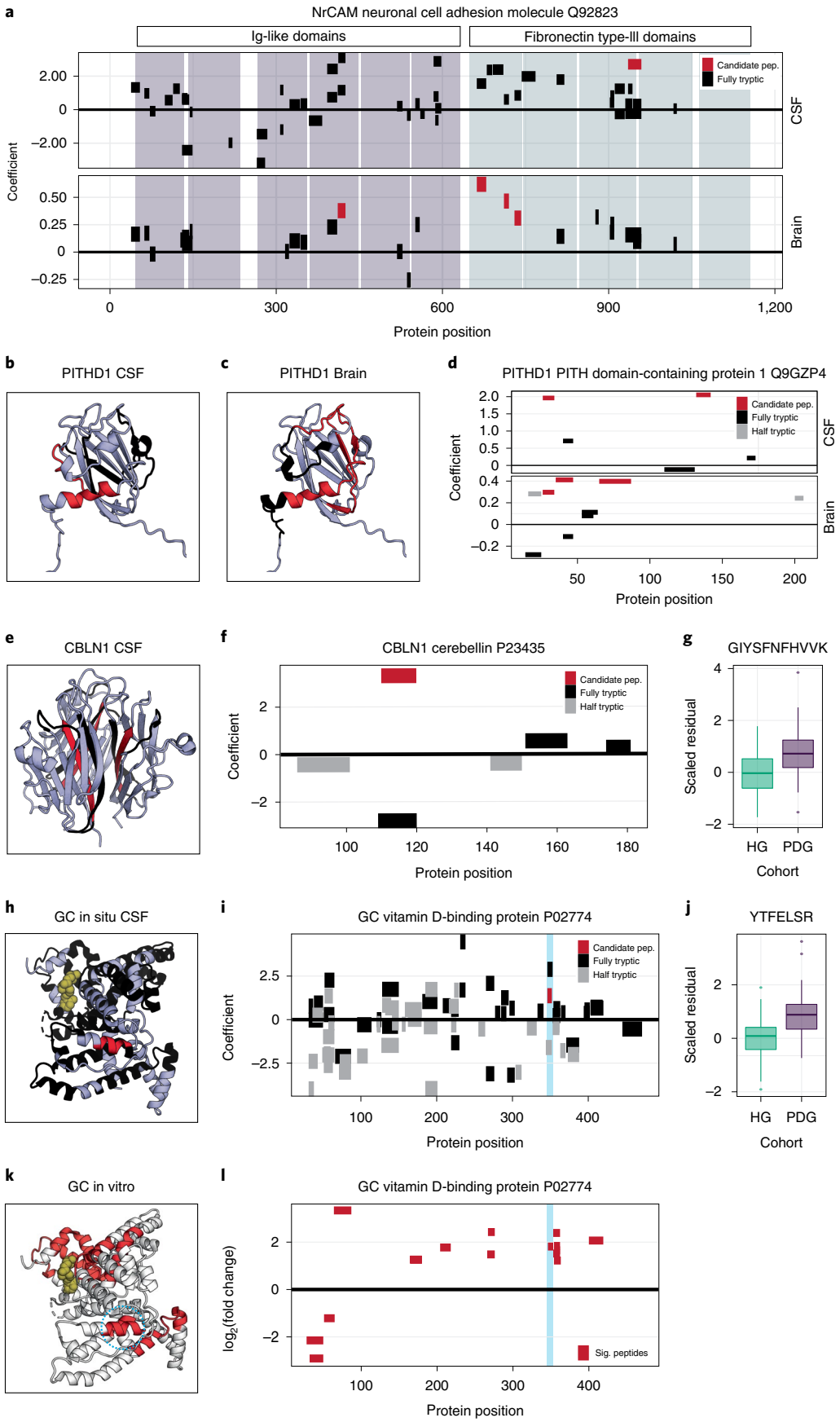
We next asked whether combinations of structural peptides could be used to classify individuals as belonging to the healthy or PD group, because biomarker combinations can improve sensitivity and specificity[65]. We used 5-fold cross-validation with evenly distributed healthy and PD samples, generated training and test sets, and defined potential biomarkers by fitting linear models to the training set, using peptide combinations. We devised a method that removes the effects of covariates (age, sex, total protein content, batch) on the signal of each peptide and fed these corrected peptide signals into a regularized regression algorithm for finding 1, 5, or 10 peptides that together were predictive of PD status (Fig. 5a). In each run of this analysis, the models were built using a subset of the data, and their performance was subsequently evaluated with test samples that were not used for training. We applied this analysis to LiP peptides (Fig. 5b), to trypsin-only peptides (Fig. 5c), and to protein abundances (Fig. 5d). Although all models performed better than expected by chance, the models using combinations of structural changes outperformed the models using non-structural changes (Fig. 5b,c; AUC of 0.752, 0.675, and 0.673 for models with 5 optimal predictors selected from either LiP peptides, PK-independent peptides, or protein abundances). A model based on a single LiP peptide (AUC 0.684) performed as well as those based on 10 PK-independent peptide or protein abundance features (AUC 0.674, 0.688). We also combined structural and abundance changes in a single model while only marginally increasing discriminatory power (Extended Data Fig. 5a–c). The superior performance of the structural model is unlikely to be due to overfitting (Methods), and the relative discriminatory power of structural peptides versus protein abundance is likely to be underestimated because our dataset is small and single peptide measurements are much noisier than protein abundance measurements. Together, these data confirm that protein structural differences between healthy and PD CSF contain more discriminatory information than protein abundance information.

Next, we compared the performance of our structure-based classification to ELISA-based measurements of different species of CSF α-synuclein, the best currently available potential biomarker[66,67]. α-Synuclein peptides were not included in the MS-based analysis because they were not well detected. Classifications on the basis of LiP peptides (AUC of 0.752, Fig. 5b; AUC of 0.756, 0.761, and 0.783, Extended Data Fig. 5a–c) slightly outperformed those on the basis of total α-synuclein (AUC of 0.733), but not on oligomeric α-synuclein (AUC of 0.799) or the ratio of oligomeric to total α-synuclein (AUC of 0.838) (Extended Data Fig. 5d).

We investigated whether a combination of α-synuclein and LiP peptide measures would further enhance the classification. We computed log (odds ratios) using the oligomeric to total α-synuclein and a model of five LiP peptides (using cross-validation; Fig. 5e). The overall accuracy (that is, individuals correctly classified as healthy or having PD) of each model was similar (73%, 75%), and was greatly improved when both models agreed (91%) (Fig. 5e). Notably, the α-synuclein model misclassified people with PD ($n = 12$) with low oligomer/total α-synuclein ratios, which were due to low oligomeric α-synuclein levels (Extended Data Fig. 5e,f); most of these participants ($n = 10$) were correctly classified by the LiP model. Donors correctly classified as having PD by LiP, but not by α-synuclein measures, had disease that had progressed relatively far (Extended Data Fig. 5g); this group included two donors with unusually long disease duration (16 years and 21 years). Overall, for the 38 individuals classified as having PD by one of the two models, about half were true positives (8/15 for the α-synuclein model, 10/23 for the LiP model). Hence, neither the oligomeric/total α-synuclein ratio nor the combined structural peptides classified all individuals with PD correctly. However, the two sets of measurements harbor complementary information on disease status and, together, correctly classified almost all participants.

Fig. 4 | Structural changes in selected proteins that are altered between healthy individuals and those with PD. a, Peptide coverage plot for NrCAM in CSF (top) and brain (bottom). Black indicates all analyzed peptides; red indicates significantly altered structural peptides between PD and healthy samples, after correction for covariates. Half-tryptic peptides were not visualized. Ig-like (dark gray) and fibronectin type-III domains (light gray) are highlighted. **b**,**c**, AlphaFold-predicted structure of PITH1, colored according to peptides in the CSF (**b**) and brain (**c**) data. Peptide colors are as in **a**. **d**, Peptide coverage plot for PITH1 in CSF (top) and brain (bottom). Peptide colors as in **a**; half-tryptic peptides in gray. **e**,**f**, Structure (**e**) and peptide coverage plot (**f**) of homotrimeric CBLN1 (PDB 5KC6), colored as in **a**, according to peptides in CSF data; half-tryptic peptides are in gray (**f**). **g**, Scaled residuals for healthy individuals and for those with PD for the indicated peptide of CBLN1. Box plots: median, center; first and third quantile, lower and upper hinges; largest/smallest value no further than 1.5 × inter-quantile range of the hinge, whiskers; data beyond are as outliers and are plotted individually ($n = 51$ participants in HG and $n = 49$ participants in PDG). **h**, Structure of vitamin-D-binding protein GC (PDB 1J78) with bound vitamin D (yellow spheres), colored as in **a**, according to peptides from in situ CSF data. **i**, Peptide coverage plot for GC in CSF. Colors are as in **d**. The light blue box indicates the protein region of the altered structural peptide. **j**, Scaled residual for the indicated peptide of GC, as in **g** ($n = 51$ participants each in HG and PDG). **k**, Structure of the vitamin-D-binding protein (as in **h**), colored according to the peptides that change upon substrate (calcitriol) addition in vitro. Red, significantly changed peptides (adjusted P values < 0.05, $\log_2$(fold change) < −1 or > 1, P values estimated with a two-sided t-test); white, detected but unchanged peptides. The blue circle encloses the peptide overlapping with the in situ-identified altered peptide. **l**, Coverage plot for significant peptides from **k**.

https://doi.org/10.1038/s41594-022-00837-0



https://doi.org/10.1038/s41594-022-00837-0

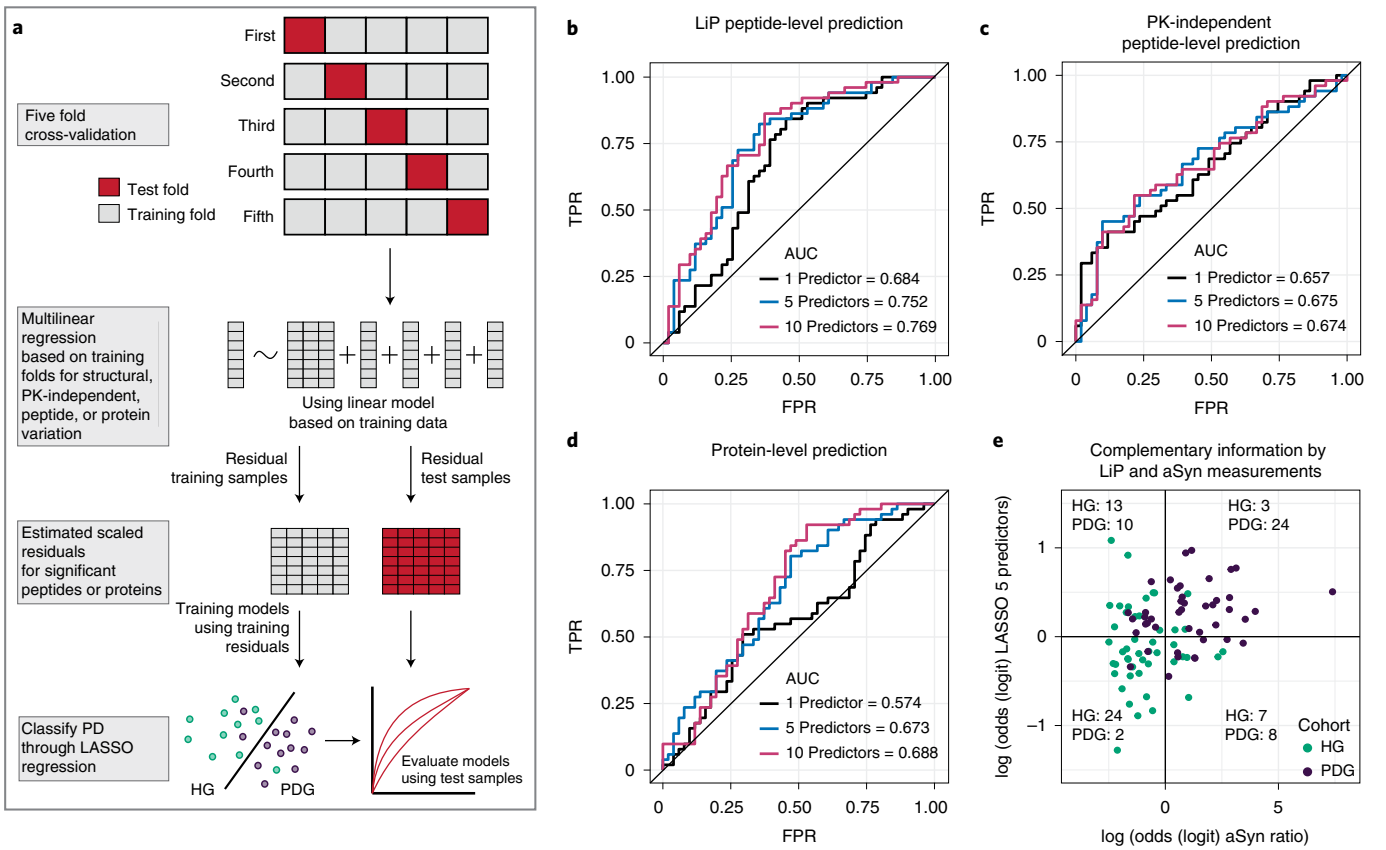


Fig. 5 | Classification of Parkinson's disease on the basis of CSF proteome information. a, Schematic of the analysis workflow for classification of samples as PD or healthy. **b**, Receiver operating characteristic (ROC) curves corresponding to the classification of PD on the basis of LiP peptide variation. **c**, ROC curves corresponding to the classification of PD on the basis of the PK-independent peptide variation. **d**, ROC curves corresponding to the classification of PD on the basis of the trypsin-only protein abundance. LASSO models were built using 1, 5, or 10 predictors, and the average classification across multiple cross-validations is used for the ROC curves. **e**, Comparison of classification using the ratio of oligomeric to total a-synuclein (log (odds) plotted on *x* axis) and using a combination of five LiP peptide levels (log (odds) plotted on *y* axis). Each point represents an individual, and the cohort membership is indicated by color. Samples with a log (odds) below zero were classified as healthy and those with a log (odds) above zero as having PD. The numbers in each quadrant indicate the results of the classification.

Overall, our data provide strong support for the hypothesis that protein structural changes provide more information about the PD state than protein abundance information. Further, our results suggest that global structural analyses of the CSF proteome may identify PD biomarkers complementary to existing α-synuclein-based measures, which may be combined with these measures to improve power to classify individuals with the disease.

Discussion

We have demonstrated that global in situ analyses of protein structures from a body fluid can generate a novel type of molecular readout of a human disease state. Protein structural alterations in human CSF better distinguished healthy individuals from those with PD than did protein abundance changes. We thus show evidence, to our knowledge for the first time, for the concept that global protein structural analyses could identify a new type of structural biomarker of a human disease, with potentially improved performance over classical CSF biomarkers based on protein levels.

We have identified proteins that are structurally altered between the CSF of healthy individuals and those with PD, and have substantially enlarged the set of proteins known to be structurally changed in PD. Performance of our candidate markers alone approached that of the ratio of oligomeric/total α-synuclein, which is the main known biochemical player in PD and is currently under investigation as a potential PD biomarker. We further showed that our structurally informative peptides provided complementary information to oligomeric/total α-synuclein levels and that combining these measures classified healthy individuals and those with PD with better performance (91% accuracy) than either measure alone (75% accuracy). In future studies, the signal of our candidate markers could be increased by optimizing enzyme-to-substrate ratios and incubation times in the LiP step. After correction for covariates, we observed only few changes in protein abundance between the CSF of healthy individuals and those with PD, some of which have been previously reported[15,18,68–77].

Translation of this concept to a clinical setting will nevertheless require additional work. The structural changes require validation in further independent cohorts and must be tested for specificity to PD versus other neurodegenerative diseases. Because bottom-up proteomics is challenging to implement in a clinical setting, validated markers could in the future form the basis for targeted proteomics or conformation-specific antibody-based assays to distinguish between the structural states of proteins in biofluids of people with PD. Although such developments face many challenges, including identification of stable structural states against which conformation-specific antibodies can be raised, our work provides a proof of concept for the use of a global structural analysis as a new type of readout to distinguish between the healthy and the disease states.

PD is a heterogeneous disease in which individuals either present only motor symptoms or also show various cognitive problems[12]. A better understanding of disease manifestation, early diagnosis, and

subtyping of individuals with the disease is essential to direct the choice of treatment. Our structural peptides performed well at classifying those with low oligomer/total α-synuclein ratios, which were misclassified by the α-synuclein-based measures, although these individuals with PD met the diagnostic criteria of clinically probable or established PD[78]. Interestingly, it has previously been shown that only about 89% of individuals with PD classified by the International Parkinson and Movement Disorder Society clinical diagnostic criteria for PD show Lewy body pathology at autopsy[79]. Individuals classified as having PD by our new structural signature, but not by α-synuclein measures, may represent those lacking Lewy body pathology, and this will require more detailed studies. In general, multidimensional peptide-based structural markers may have the potential to distinguish between disease subtypes, and therefore to stratify individuals according to prognosis. On the basis of our data, multidimensional structure-based markers are likely to be more sensitive than abundance-based ones. These predictions, however, remain to be tested.

Our data also show that a global in situ structural analysis of disease sample provides information about pathological mechanisms. Although our study was designed to identify potential structural biomarkers in CSF, whereas PD pathology manifests in brain tissue, the set of structurally altered CSF proteins in PD captures pathways known to be involved in the disease. Specifically, we identified multiple proteins involved in synapse organization, known altered enzymes, hormone and vitamin transport proteins implicated in PD, and proteins related to amyloid-clearance pathways. The enrichment of synaptic proteins is particularly intriguing because PD is a known synaptopathy. This suggests that, even in the CSF, we have identified structurally altered proteins that may reflect pathology in the PD brain. Indeed, analysis of postmortem brain samples from a small independent cohort showed that, for the structurally changed CSF proteins that we could detect in the brain, more than half showed structural changes in both samples, and many were strongly linked to the disease. It was expected that the set of overlapping detected proteins was small and that very few identical structural changes were observed in the PD brain and CSF, given that we were comparing samples with differing proteomes, interactomes, metabolomes, and biophysical properties. Nevertheless, these data are encouraging and suggest that the application of this approach to brain samples from a larger cohort of people with PD will provide insight into structural changes in the disease state.

We also present the first global analysis of a structural proteome in healthy individuals. Our dataset of variable and invariable CSF proteins across healthy humans is an important starting point for future structural biomarker development. It should inform study design to achieve sufficient statistical power for the identification of new biomarkers, and it indicates proteins (that is, those with high inter-individual variability) that are not likely to be useful for this purpose. It should also serve as a useful resource for structural biology because it identifies those structurally variable proteins for which in vitro structural data may be particularly difficult to extrapolate to in vivo state and function. Finally, our dataset should enable further study of protein structural changes during human aging.

Although our work has focused on PD, the approach is applicable to any disease involving dysfunctional proteins with altered structural states. For instance, constitutively active kinases or non-functional tumor suppressors in cancer would be expected to show altered structures and possibly interactomes. Structural-proteomics screens could capture such changes and prove to be a powerful tool for a more systemic understanding of human disease states.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41594-022-00837-0.

References

1. Kennedy, B. K. et al. Geroscience: linking aging to chronic disease. *Cell* **159**, 709–713 (2014).
2. Cilento, E. M. et al. Mass spectrometry: a platform for biomarker discovery and validation for Alzheimer's and Parkinson's diseases. *J. Neurochem.* **151**, 397–416 (2019).
3. Crutchfield, C. A., Thomas, S. N., Sokoll, L. J. & Chan, D. W. Advances in mass spectrometry-based clinical biomarker discovery. *Clin. Proteom.* **13**, 1 (2016).
4. Jiang, R. et al. Differential proteomic analysis of serum exosomes reveals alterations in progression of Parkinson disease. *Medicine* **98**, e17478 (2019).
5. Macklin, A., Khan, S. & Kislinger, T. Recent advances in mass spectrometry based clinical proteomics: applications to cancer research. *Clin. Proteomics* **17**, 17 (2020).
6. Thygesen, C., Boll, I., Finsen, B., Modzel, M. & Larsen, M. R. Characterizing disease-associated changes in post-translational modifications by mass spectrometry. *Expert Rev. Proteom.* **15**, 245–258 (2018).
7. Tzeng, S. R. & Kalodimos, C. G. Protein activity regulation by conformational entropy. *Nature* **488**, 236–240 (2012).
8. Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* **450**, 964–972 (2007).
9. Schopper, S. et al. Measuring protein structural changes on a proteome-wide scale using limited proteolysis-coupled mass spectrometry. *Nat. Protoc.* **12**, 2391–2410 (2017).
10. Feng, Y. et al. Global analysis of protein structural changes in complex proteomes. *Nat. Biotechnol.* **32**, 1036–1044(2014).
11. Cappelletti, V. et al. Dynamic 3D proteomes reveal protein functional alterations at high resolution in situ. *Cell* **184**, 545–559.e22 (2021).
12. Spillantini, M. G. et al. α-synuclein in Lewy bodies. *Nature* **388**, 839–840 (1997).
13. Braak, H. et al. Staging of brain pathology related to sporadic Parkinson's disease. *Neurobiol. Aging* **24**, 197–211 (2003).
14. Brás, I. C., Xylaki, M. & Outeiro, T. F. Mechanisms of alpha-synuclein toxicity: an update and outlook. *Prog. Brain. Res.* **252**, 91–129 (2020).
15. Maass, F., Schulz, I., Lingor, P., Mollenhauer, B. & Bähr, M. Cerebrospinal fluid biomarker for Parkinson's disease: an overview. *Mol. Cell. Neurosci.* **97**, 60–66 (2019).
16. Borrageiro, G., Haylett, W., Seedat, S., Kuivaniemi, H. & Bardien, S. A review of genome-wide transcriptomics studies in Parkinson's disease. *Eur. J. Neurosci.* **47**, 1–16 (2018).
17. Majbour, N. K. et al. Oligomeric and phosphorylated alpha-synuclein as potential CSF biomarkers for Parkinson's disease. *Mol. Neurodegener.* **11**, 7 (2016).
18. Parnetti, L. et al. CSF and blood biomarkers for Parkinson's disease. *Lancet Neurol.* **18**, 573–586 (2019).
19. van Dijk, K. D. et al. Changes in endolysosomal enzyme activities in cerebrospinal fluid of patients with Parkinson's disease. *Mov. Disord.* **28**, 747–754 (2013).
20. van Steenoven, I. et al. α-Synuclein species as potential cerebrospinal fluid biomarkers for dementia with lewy bodies. *Mov. Disord.* **33**, 1724–1733 (2018).
21. Van Dijk, K. D. et al. Cerebrospinal fluid and plasma clusterin levels in Parkinson's disease. *Park. Relat. Disord.* **19**, 1079–1083 (2013).
22. van Dijk, K. D. et al. Reduced α-synuclein levels in cerebrospinal fluid in Parkinson's disease are unrelated to clinical and imaging measures of disease severity. *Eur. J. Neurol.* **21**, 388–394 (2014).

23. Abdi, I. Y. et al. Preanalytical stability of CSF total and oligomeric α-synuclein. *Front. Aging Neurosci.* **13**, 85 (2021).
24. El-Agnaf, O. M. A. et al. Detection of oligomeric forms of α-synuclein protein in human plasma as a potential biomarker for Parkinson's disease. *FASEB J.* **20**, 419–425 (2006).
25. Oosterveld, L. P. et al. CSF biomarkers reflecting protein pathology and axonal degeneration are associated with memory, attentional, and executive functioning in early-stage Parkinson's disease. *Int. J. Mol. Sci.* **21**, 1–12 (2020).
26. Macron, C., Lane, L., Núnez 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).
27. Barkovits et al. Blood contamination in CSF and its impact on quantitative analysis of α-synuclein. *Cells* **9**, 370 (2020).
28. Macron, C. et al. Exploration of human cerebrospinal fluid: a large proteome dataset revealed by trapped ion mobility time-of-flight mass spectrometry. *Data Brief* **31**, 105704 (2020).
29. Zhang, J. & Kurgan, L. SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. *Bioinformatics* **35**, i343–i353 (2019).
30. Jones, D. T. & Cozzetto, D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **31**, 857–863 (2015).
31. Beecham, G. W. et al. PARK10 is a major locus for sporadic neuropathologically confirmed Parkinson disease. *Neurology* **84**, 972–980 (2015).
32. Nalls, M. A. et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **18**, 1091–1102 (2019).
33. Chang, D. et al. A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat. Genet.* **2017**, 1511–1516 (2017).
34. Hoxha, E., Tempia, F., Lippiello, P. & Miniaci, M. C. Modulation, plasticity and pathophysiology of the parallel fiber-purkinje cell synapse. *Front. Synaptic. Neurosci.* **8**, 35 (2016).
35. Lozovaya, N. et al. GABAergic inhibition in dual-transmission cholinergic and GABAergic striatal interneurons is abolished in Parkinson disease. *Nat. Commun.* **9**, 1–14 (2018).
36. Zheng, X. et al. Increase in glutamatergic terminals in the striatum following dopamine depletion in a rat model of Parkinson's disease. *Neurochem. Res.* **44**, 1079–1089 (2019).
37. Gardoni, F., Ghiglieri, V., Luca, M. di & Calabresi, P. Assemblies of glutamate receptor subunits with post-synaptic density proteins and their alterations in Parkinson's disease. *Prog. Brain Res.* **183**, 169–182 (2010).
38. Błaszczyk, J. W. Parkinson's disease and neurodegeneration: GABA-collapse hypothesis. *Front. Neurosci.* **10**, 269 (2016).
39. Kayakabe, M. et al. Motor dysfunction in cerebellar Purkinje cell-specific vesicular GABA transporter knockout mice. *Front. Cell. Neurosci.* **7**, 286 (2014).
40. Murueta-Goyena, A., Andikoetxea, A., Gómez-Esteban, J. C. & Gabilondo, I. Contribution of the GABAergic system to non-motor manifestations in premotor and early stages of Parkinson's disease. *Front. Pharmacol.* **10**, 1294 (2019).
41. Surmeier, D. J. et al. Calcium and Parkinson's disease. *Biochem. Biophys. Res. Commun.* **483**, 1013–1019 (2017).
42. Pchitskaya, E., Popugaeva, E. & Bezprozvanny, I. Calcium signaling and molecular mechanisms underlying neurodegenerative diseases. *Cell Calcium* **70**, 87–94 (2018).
43. Latourelle, J. C. et al. Genomewide association study for onset age in Parkinson disease. *BMC Med. Genet.* **10**, 98 (2009).
44. Blauwendraat, C. et al. Parkinson's disease age at onset genome-wide association study: defining heritability, genetic loci, and α-synuclein mechanisms. *Mov. Disord.* **34**, 866–875 (2019).
45. Tan, M. M. X. et al. Genome-wide association studies of cognitive and motor progression in Parkinson's disease. *Mov. Disord.* **36**, 424–433 (2021).
46. Wilhelmus, M. M. M. et al. Short communication apolipoprotein E and LRP1 increase early in Parkinson's disease pathogenesis. *Am. J. Pathol.* **179**, 2152–2156 (2011).
47. Troy T, R. & Jacob M, M. Apolipoprotein E fragmentation within lewy bodies of the human Parkinson's disease brain. *Int. J. Neurodegener. Disord.* **1**, 002 (2018).
48. Xu, J., Mashimo, T. & Südhof, T. C. Synaptotagmin-1, -2, and -9: Ca^{2+} sensors for fast release that specify distinct presynaptic properties in subsets of neurons. *Neuron* **54**, 567–581 (2007).
49. Delignat-Lavaud, B. et al. The calcium sensor synaptotagmin-1 is critical for phasic axonal dopamine release in the striatum and mesencephalon, but is dispensable for basic motor behaviors in mice. Preprint at bioRxiv https://doi.org/10.1101/2021.09.15.460511 (2021).
50. Wu, M., Puddifoot, C. A., Taylor, P. & Joiner, W. J. Mechanisms of inhibition and potentiation of α4β2 nicotinic acetylcholine receptors by members of the Ly6 protein family. *J. Biol. Chem.* **290**, 24509 (2015).
51. Wang, Q. et al. The landscape of multiscale transcriptomic networks and key regulators in Parkinson's disease. *Nat. Commun.* **2019**, 1–15 (2019). 101 10.
52. Power, J. H. T., Shannon, J. M., Blumbergs, P. C. & Gai, W. P. Nonselenium glutathione peroxidase in human brain: elevated levels in Parkinson's disease and dementia with Lewy bodies. *Am. J. Pathol.* **161**, 885–894 (2002).
53. Corradini, B. R. et al. Complex network-driven view of genomic mechanisms underlying Parkinson's disease: Analyses in dorsal motor vagal nucleus, locus coeruleus, and substantia nigra. *Biomed. Res. Int.* 543673 (2014).
54. Lachén-Montes, M. et al. Unveiling the olfactory proteostatic disarrangement in Parkinson's disease by proteome-wide profiling. *Neurobiol. Aging* **73**, 123–134 (2019).
55. Lachén-Montes, M. et al. Smelling the dark proteome: functional characterization of PITH domain-containing protein 1 (C1orf128) in olfactory metabolism. *J. Proteome Res.* **19**, 4826–4843 (2020).
56. Ma, S. et al. Peroxiredoxin 6 is a crucial factor in the initial step of mitochondrial clearance and is upstream of the PINK1–parkin pathway. *Antioxid. Redox Signal.* **24**, 486–501 (2016).
57. Yun, H. M., Choi, D. Y., Oh, K. W. & Hong, J. T. PRDX6 exacerbates dopaminergic neurodegeneration in a MPTP mouse model of Parkinson's disease. *Mol. Neurobiol.* **52**, 422–431 (2015).
58. Elegheert, J. et al. Structural basis for integration of GluD receptors within synaptic organizer complexes. *Science* **353**, 295–300 (2016).
59. Chipman, P. & Goda, Y in *Dendrites: Development and Disease* 425–465 (Springer Japan, 2016).
60. Won, S. Y., Lee, P. & Kim, H. M. Synaptic organizer: Slitrks and type IIa receptor protein tyrosine phosphatasess. *Curr. Opin. Struct. Biol.* **54**, 95–103 (2019).
61. Lee, S. J. et al. Presynaptic neuronal pentraxin receptor organizes excitatory and inhibitory synapses. *J. Neurosci.* **37**, 1062–1080 (2017).
62. Longhena, F., Faustini, G., Spillantini, M. G. & Bellucci, A. Living in promiscuity: the multiple partners of α-synuclein at the synapse in physiology and pathology. *Int. J. Mol. Sci.* **20**, 141 (2019).
63. Fullard, M. E. & Duda, J. E. A review of the relationship between vitamin D and Parkinson disease symptoms. *Front. Neurol.* **11**, 454 (2020).

64. Lawton, M. et al. Blood biomarkers with Parkinson's disease clusters and prognosis: the oxford discovery cohort. *Mov. Disord.* **35**, 279–287 (2020).
65. Li, T. & Le, W. Biomarkers for Parkinson's disease: how good are they? *Neurosci. Bull.* **36**, 183–194 (2020).
66. Kang, U. J. et al. Comparative study of cerebrospinal fluid α-synuclein seeding aggregation assays for diagnosis of Parkinson's disease. *Mov. Disord.* **34**, 536–544 (2019).
67. Rossi, M. et al. Ultrasensitive RT-QuIC assay with high sensitivity and specificity for Lewy body-associated synucleinopathies. *Acta Neuropathol.* **140**, 49–62 (2020).
68. Rotunno, M. S. et al. Cerebrospinal fluid proteomics implicates the granin family in Parkinson's disease. *Sci. Rep.* **2020**, 1–11 (2020).
69. Eusebi, P. et al. Cerebrospinal fluid biomarkers for the diagnosis and prognosis of Parkinson's disease: protocol for a systematic review and individual participant data meta-analysis. *BMJ Open* **7**, e018177 (2017).
70. Simrén, J., Ashton, N. J., Blennow, K. & Zetterberg, H. An update on fluid biomarkers for neurodegenerative diseases: recent success and challenges ahead. *Curr. Opin. Neurobiol.* **61**, 29–39 (2020).
71. Dixit, A., Mehta, R. & Singh, A. K. Proteomics in human Parkinson's disease: present scenario and future directions. *Cell. Mol. Neurobiol.* **39**, 901–915 (2019).
72. Parnetti, L. et al. Parkinson's and Lewy body dementia CSF biomarkers. *Clin. Chim. Acta* **495**, 318–325 (2019).
73. Heywood, W. E. et al. Identification of novel CSF biomarkers for neurodegeneration and their validation by a high-throughput multiplexed targeted proteomic assay. *Mol. Neurodegener.* **10**, 64 (2015).
74. Magdalinou, N. K. et al. Identification of candidate cerebrospinal fluid biomarkers in parkinsonism using quantitative proteomics. *Parkinsonism Relat. Disord.* **37**, 65–71 (2017).
75. Magdalinou, N., Lees, A. J. & Zetterberg, H. Cerebrospinal fluid biomarkers in parkinsonian conditions: an update and future directions. *J. Neurol. Neurosurg. Psychiatry* **85**, 1065–1075 (2014).
76. Sarkar, A., Rawat, N., Sachan, N. & Singh, M. P. Unequivocal biomarker for Parkinson's disease: a hunt that remains a pester. *Neurotox. Res.* **36**, 627–644 (2019).
77. Liu, W. et al. Role of exosomes in central nervous system diseases. *Front. Mol. Neurosci.* **12**, 240 (2019).
78. Postuma, R. B. et al. MDS clinical diagnostic criteria for Parkinson's disease. *Mov. Disord.* **30**, 1591–1601 (2015).
79. Geut, H. et al. Neuropathological correlates of parkinsonian disorders in a large Dutch autopsy series. *Acta Neuropathol. Commun.* **8**, 39 (2020).

Methods

Study cohorts

Cerebrospinal fluid study cohort. The study population consisted of 52 individuals with PD (Age: 41–84, Male: 39.2%) that attended the outpatient clinic for movement disorder of the VU University Medical Center Amsterdam, the Netherlands, and CSF was collected in the period of September 2008 to February 2012, as described previously[22]. Individuals with known monogenic forms of PD were excluded. Those with PD fulfilled the United Kingdom Parkinson's Disease Society Brain Bank clinical diagnostic criteria[80]. Severity of parkinsonism and disease state were rated using the Unified Parkinson's Disease Rating Scale (UPDRS) motor subscale and the modified Hoehn and Yahr (H&Y) classification[81,82]. The participants underwent a mini-mental-state examination to exclude dementia. The disease duration was calculated from the first motor symptoms as reported by the participants to the time of diagnosis. An age- and sex-matched healthy control cohort group consisted of 51 individuals (age: 51–82, male: 65.4%). For the healthy cohort group, dementia was excluded using the Cambridge Cognitive Examination scale (CAMCOG)[83]. All participants underwent an extensive, standardized clinical assessment, including medical history and a neurological examination. The levels of previously identified PD-associated biochemical markers, such as t-α-Syn, p-α-Syn, and oligomeric α-Syn, were measured by ELISA using monoclonal epitope-specific antibodies or monoclonal conformation-specific antibodies, as previously described[17,19,25]. The study was approved by the local ethics committee of the VU University Medical Center, and all participants gave written informed consent. Participants were reimbursed for travel costs and offered a lunch. No further compensation was offered. The number of samples per cohort group was restricted by the availability of suitable biofluid samples. No statistical methods were used to pre-determine sample size, but our sample sizes are similar to those reported in previous publications[68].

Brain study cohort. The postmortem brain tissue samples were collected by the Netherlands Brain Bank (www.brainbank.nl) in Amsterdam, the Netherlands, and further characterized by Wilma van de Berg[79]. For all donors, a written informed consent for brain autopsy and the use of the material and clinical information for research purposes had been obtained from the donor or the next of kin. Demographic features and clinical symptoms were extracted from the clinical files, including sex, age at symptom onset, age at death, disease duration, presence of dementia, and core and supportive clinical features for PD according to the International Parkinson and Movement Disorder Society clinical diagnostic criteria for PD[78]. Braak and McKeith αSyn stages were determined using the BrainNet Europe criteria[84]. On the basis of Thal amyloid-β phases scored on the medial temporal lobe[85], Braak neurofibrillary stages[86] and CERAD neuritic plaque scores[87], levels of AD pathology were determined according to National Institute on Aging and Alzheimer's Association consensus criteria[88]. Additionally, Thal CAA stages[89], presence of aging-related tau astrogliopathy (ARTAG)[90], microvascular lesions and hippocampal sclerosis were assessed. We selected 5 brain tissue samples of the medial temporal gyrus from female individuals with sporadic PD and 5 non-neurological controls aged 81–84 years. The study was approved by the local ethics committee of the VU University Medical Center, Amsterdam, the Netherlands. Donors and their relatives did not receive any compensation. By signing the informed consent form, donors gave permission for postmortem brain autopsy and use of their brain material and medical records for research purposes. Donors and their relatives did not receive any compensation. No statistical methods were used to pre-determine sample sizes.

Statistics and reproducibility

We conducted LiP–MS on CSF samples from a well-characterized clinical cohort consisting of 52 people with PD and 51 healthy age-matched individuals, with both sexes represented in both cohort groups. Statistical analyses to correct for covariates as well as for protein abundance and non-structural effects have been described in detail below. Our analysis was expected to provide unprecedented proteome-wide information on structural changes in PD, such that a prior estimate of effect size was not possible. Therefore, no statistical method was used to determine sample size, but our sample sizes are similar to those reported in other publications[68]. All LiP–MS experiments were conducted in triplicate and assessed with standard quality-control measures in the Picotti laboratory, and mass spectrometry runs were randomized between cohorts to minimize batch effects. Apart from data from a single individual in the healthy group who lacked protein abundance measurements and was therefore excluded from the comparison between healthy and PD cohort groups, no data were excluded. Data collection and analysis were not performed blind.

Cerebrospinal fluid collection

CSF was collected from individuals by lumbar puncture and collected in polypropylene collection tubes. The CSF was analyzed for red blood cell contamination (see Extended Data Fig. 1a), centrifuged at 1,800*g* at 4 °C for 10 minutes, aliquoted, and stored at −80 °C within 2 hours of collection until processing, in line with published guidelines[91].

Limited proteolysis of cerebrospinal fluid samples

The CSF samples were thawed in batches of seven on ice. Each batch was designed to have a similar sample distribution with respect to age, sex, and cohort group. For each sample, the protein concentration was determined using the Pierce BCA Protein Assay Kit (Thermo Fisher Scientific). Each sample was split into a trypsin-only and a LiP sample, containing 80 µl of CSF. Both samples were heated to 25 °C for 5 minutes. Proteinase K from *Tritirachium album* (Sigma Aldrich) was added to the LiP sample at an enzyme/substrate (E:S) ratio of 1:100 (wt/wt) and incubated for 5 minutes at 25 °C. The trypsin-only samples were treated with the corresponding amount of water under the same conditions. The digestion by Proteinase K was stopped by heating all samples to 95 °C for 5 minutes, samples were cooled down to 4 °C, and sodium deoxycholate (Sigma Aldrich) was added to a final concentration of 5%. Cysteine residues were reduced by adding TCEP (Pierce) to a final concentration of 5 mM and incubated for 30 minutes at 37 °C with shaking at 600 r.p.m., using an Eppendorf ThermoMixer C. The reduced cysteines were alkylated by the addition of IAA (Sigma Aldrich) to a final concentration of 4 mM and were incubated at room temperature for 30 minutes. Samples were then diluted with 0.1 M ammonium bicarbonate (AmBic) (Sigma Aldrich) to 2.5% sodium deoxycholate (DOC) before the first digestion step with LysC (Wako Chemical), at a 1:100 enzyme to substrate ratio (wt/wt) for 120 minutes at 37 °C at 600 r.p.m. After a further dilution step of DOC to 1% by 0.1 AmBic, samples were digested with trypsin (Promega), added at a 1:100 enzyme to substrate ratio (wt/wt) for 16 hours at 37 °C at 600 r.p.m. The digestion was stopped by the addition of formic acid (Sigma Aldrich) to a final concentration of 2% to result in a pH below 3. The DOC precipitate was filtered out using a 0.2-µm PVDF hydrophilic membrane (Corning). The filtered peptides mixture was loaded on 96-well MACROSpin plates (The Nest Group), desalted, and eluted with 80% acetonitrile and 0.1% formic acid. The peptides were dried using vacuum centrifugation and resuspended in 0.1% formic acid. The peptide concentration was determined using the Pierce™ Quantitative Colorimetric Peptide Assay (Thermo Fisher Scientific).

Limited proteolysis of brain tissue samples

The brain tissue samples were thawed on ice and transferred to a 1.5-ml tube using 300–400 µl ice cold lysis buffer (100 mM HEPES pH 7.4, 150 mM KCl, 1 mM $MgCl_2$; one cOmplete protease inhibitor cocktail tablet (Roche) was added per 50 ml lysis buffer), depending on the starting weight of the sample. The samples were not randomized because all brain tissue samples were processed in one batch. The cells were lysed using a pellet mixer. Samples were homogenized for 8 cycles of 15 strokes each, with a 1-minute break on ice between the cycles. The

lysate was then centrifuged at 9,390*g* for 10 minutes at 4 °C. The supernatant was transferred to a new tube and the protein concentration was determined using the Pierce BCA Protein Assay Kit (Thermo Fisher Scientific). For each brain tissue sample, a trypsin-only and LiP sample, in triplicate, was prepared using 40 µg of lysed material. Each sample was topped up to 50 µl using lysis buffer. The LiP and trypsin-digestion steps were performed as for the CSF samples, described above.

Liquid chromatography–mass spectrometry methods for cerebrospinal fluid samples

All chemicals were purchased from Sigma unless otherwise mentioned. For data collection of all raw MS files, Xcalibur (4.1) was used.

Liquid chromatography–mass spectrometry analysis for library generation (data-dependent acquisition). For library generation, two condition pools (HG and PDG) of the LiP-treated and the trypsin-only samples were prepared (in total 4 pools, ~200 µg each pool). Each pooled sample was fractionated by high pH reverse phase (HPRP) fractionation using a Dionex Ultimate 3000 LC (Thermo Fisher Scientific) on an ACQUITY UPLC CSH 1.7-µm C18 column (2.1 × 150 mm, Waters). The peptides were separated by a non-linear gradient from 1% HPRP buffer B (100% acetonitrile)/99% HPRP buffer A (20 mM ammonium formiate, pH 10) to 40% buffer B. A fraction was taken every 45 seconds and fractions were pooled into 15 final fractions. Afterward, peptides were dried completely in a speed vac and resuspended in solvent A (1% acetonitrile, 0.1% formic acid) and spiked with iRT peptides, according to the manufacturer's recommendations (Biognosys). Peptide concentrations were determined using nano-drop (Spectrostar Nano, BMG labtech).

Fractionated samples (2 µg) were separated by a non-linear gradient from 1% buffer B (85% acetonitrile, 0.1% formic acid in water)/95% buffer A (1% acetonitrile, 0.1% formic acid in water) to 44% buffer B in 2 hours on an in-house-packed 60-cm column (PicoFrit emitters, inner diameter 75 µm, New Objective; CSH 1.7-µm column material, Waters) using an Easy nLC 1200 (Thermo Fisher Scientific) coupled online to a Q Exactive HF-X mass spectrometer (Thermo Fisher Scientific). The flow rate was set to 250 nl/min. The Q Exactive HF-X mass spectrometer was operated in data-dependent acquisition Top15 mode with following settings: MS1 scan range: 350–1,650 Th; resolution 60,000; MS1 AGC target: 3×10^6; MS1 maximum injection time (IT): 25 ms; MS2 scan resolution: 15,000; MS2 AGC target: 2×10^5; MS2 maximum IT: 25 ms; isolation window: 4 Th; scan range: 200 to 2000 Th; NCE: 27; minimum AGC target: 1×10^3; only charge states 2 to 5 considered; peptide match: preferred; exclude isotopes: on; dynamic exclusion: 20 seconds.

Liquid chromatography–mass spectrometry analysis (data-independent acquisition). The DIA acquisition method was adapted from Bruder, R. et al.[92]. Peptides (2 µg) were separated by a non-linear gradient using the same LC–MS setup and gradient as described before. The mass spectrometer was operated in DIA mode using the following settings for the MS1 scan: range: 350 to 1650 Th, resolution: 120,000; AGC target: 5×10^6; maximum IT: 20 ms.The MS1 scan was followed by 45 DIA scans using the following settings: resolution: 30,000; AGC target: 3×10^6; maximum IT: 55 ms; fixed first mass: 200 Th; stepped NCE: 25.5, 27, 30. The window widths were adjusted to the precursor density.

Data analysis: hybrid library generation. The DDA and DIA raw files were searched separately with SpectroMine 2.0.190613.43665 (Biognosys) against the human UniProt FASTA including isoforms (downloaded on 1 July 2019). For the trypsin-control samples the following search settings were applied: acetyl (protein N-term) and oxidation (M); enzyme: trypsin/P with up to two missed cleavages. For the LiP-treated samples, the digestion specificity was changed to semi-tryptic. Mass tolerances were automatically determined by SpectroMine, and other settings were set to default. Search results were filtered by a 1% false-discovery rate (FDR) on precursor, peptide, and protein level[93,94]. For hybrid library generation[95], the search archives from the above-mentioned fractionated samples were used for library generation in SpectroMine. Default settings were applied during library generation (1% FDR).

Data-independent acquisition data analysis. Prior to analysis of the DIA data, the raw files were converted into htrms files using the htrms converter (Biognosys). MS1 and MS2 data were centroided during conversion. In Spectronaut, imputing was not performed. The other parameters were set to default. The htrms files were analyzed with Spectronaut 13 (version: 13.5.190812, Biognosys)[96] using the previously generated hybrid library and default settings. The results were filtered by a 1% FDR on precursor and protein level (Q value < 0.01) (Supplementary Tables 1 and 2).

Liquid chromatography–mass spectrometry methods for brain samples

Liquid chromatography–mass spectrometry analysis for library generation (data-dependent acquisition). The library for the brain tissue samples was prepared by pooling all replicates from two or three individuals into a new sample. In total, four trypsin-only and LiP-treated pooled samples were generated. Each sample (0.75 µg) was separated by a linear gradient from 3% buffer B (B: 100% acetonitrile, 0.1% formic acid; A: 0.1% formic acid in water) to 35% buffer B in 2 hours on an in-house-packed 40 cm column (PicoFrit emitters, inner diameter 75 µm, New Objective; Reprosil-Pur 120 C18-AQ, 1.9 µm, Dr. Maisch) using an ACQUITY UPLC M-Class (Waters) coupled online to a Orbitrap Fusion Lumos Tribrid (Thermo Fisher Scientific). The flow rate was set to 300 nl/minute. The mass spectrometer was operated in data-dependent acquisition mode with the following settings: MS1 scan range: 350–1,400 *m/z*; resolution: 120,000; MS1 AGC target: 200%; MS1 maximum injection time: 100 ms; MS2 scan resolution: 30,000; MS2 AGC target: 200%; MS2 maximum injection time: 54 ms; isolation window: 1.6 *m/z*; scan range: normal; dynamic exclusion: 60 seconds.

Liquid chromatography–mass spectrometry analysis analysis (data-independent acquisition). The peptides (0.75 µg) were separated by a linear gradient using the same LC–MS setup and gradient as described before. Prior to injection, iRT peptides were spiked-in, as for the CSF samples. The mass spectrometer was operated in DIA mode using the following setting for MS1 scan: range: 350–1400 *m/z*; resolution: 120,000, maximum injection time: 100 ms. The MS1 scan was followed by 41 DIA scans using the following settings: resolution: 30,000 *m/z*, AGC target: 200%; maximum injection time: 54 ms. The window widths were adjusted to the precursor density.

Data analysis: hybrid library generation and data-independent acquisition data analysis. Preparation of the hybrid library and data analysis were performed similarly as for the CSF data. The DDA and DIA files of all experiments were searched together with SpectroMine (2.8.210609.47784, Biognosys) with the same UniProt FASTA file as used for the CSF. The digestion specificity was set to semi-tryptic, and all other settings were set to default. All DIA files were analyzed in Spectronaut 15 (15.1.210713.50606, Biognosys), using the generated hybrid library, and default settings were applied.

Preprocessing of peptide intensities and protein abundances

Filtering and normalization of peptide intensities. Peptide filtering and normalization steps were applied to the trypsin-only treated data and to the proteinase K-treated (LiP) data individually before further processing steps were performed.

Peptides that were detected in less than 50% of the samples and/or fewer than 10 samples in each cohort group were excluded from any further analysis. Additionally, all peptides belonging to albumin were removed. The numbers of peptides and proteins at various processing and analysis steps are shown in Supplementary Table 3.

https://doi.org/10.1038/s41594-022-00837-0

After log-transforming the data, signals of remaining peptides were normalized by applying sample-specific correction factors. These correction factors were estimated on the basis of the observed intensities of the peptides that were detected across samples. For this, the mean intensity (across all samples) of every peptide that fell in the top and bottom 15% of peptide intensities was excluded from the further estimation of the correction factors. The remaining peptide intensities were centered over all samples. For retrieving the centered intensity (*cPI*) of each peptide $i \in (1, \ldots, N)$ in sample $k \in (1, \ldots, K)$, the mean peptide intensity (*PI*) of peptide *i* was subtracted from the peptide intensity of peptide *i* in sample *k*:

$$cPI_{ik} = PI_{ik} - \frac{1}{K}\sum_{k=1}^{K} PI_{ik}$$

Because the PI_{ik} values are log-transformed intensities, this corresponds to dividing the intensities by their geometric mean. For estimating the correction factor for sample $k \in (1, \ldots, K)$, a trimmed mean intensity (*tPI*) was computed from the centered peptide intensities, excluding the top 10% of peptides with the highest intensities and the bottom 10% of peptides with the lowest intensities in each sample.

Last, these sample-specific correction factors were subtracted from the measured intensities sample-wise to retrieve the normalized peptide intensities (*nPI*) which were used in the following analysis:

$$nPI_{ik} = PI_{ik} - tPI_{k}$$

Note that this correction was performed on all peptides, not only those that were used to compute the correction factor.

Summarizing triplicates in brain data. After filtering peptides and normalizing the peptide signals of the individual samples in the brain data, the triplicates were summarized in both the LiP and trypsin-only peptide data. This step was not necessary for the CSF data, for which we did not have sufficient material for replicates. A mean peptide intensity was estimated if a peptide was detected in at least two of the triplicates. If a mean peptide intensity could be estimated in all ten donors, the peptide remained in the analysis. All following analyses were performed with these summarized replicate intensities.

Estimating protein abundances from trypsin-only peptide levels. Protein abundances were estimated from the normalized peptide intensities of the trypsin-only data using mapDIA[97]. Half-tryptic peptides and peptides mapping to multiple proteins were excluded from this preprocessing step. If a protein had multiple isoforms, only one protein abundance was estimated. Protein abundances were then estimated using the following settings in mapDIA and subsequently log-transformed:

Parameter	Input CSF data	Input brain data
LEVEL	2	2
LOG2_TRANSFORMATION	false	false
EXPERIMENTAL_DESIGN	IndependentDesign	IndependentDesign
SDF	2	2
MIN_CORREL	0.3	0.3
MIN_OBS	5 5	5 5
MIN_PEP_PER_PROT	2	2
LABELS	HG PDG	HG PDG
SIZE	53 52	5 5
MIN_DE	.01	.01
MAX_DE	.99	.99
CONTRAST	−0 1 -	−0 1 -
MAX_PEP_PER_PROT	inf	inf

Assessing structural variability in cerebrospinal fluid of healthy individuals

Identifying peptides with increased structural variability between samples. In order to identify protein regions that show relatively large inter-individual structural variation, we focused on peptides with large variation in the LiP data, but relatively small variation in the trypsin-only data. Therefore, we are identifying peptides that vary specifically after proteinase K digestion, but not due to other factors, such as technical variability or variation in protein abundance.

In a first step, we removed outlier measurements using Cook's distance. We fitted the peptide intensities of each of the LiP and the trypsin-only data to the corresponding trypsin-only protein abundances. Subsequently, we estimated the influence that each peptide measurement has on the fitted response value using Cook's distance. A common threshold, equal to four divided by the number of observations, was applied, and all peptide measurements with a Cook's *D* above this value were removed from further analysis.

Subsequently, the s.d. of each peptide was estimated within the LiP and within the trypsin-only samples. A variability score (VarS) was computed by subtracting the s.d. of the peptide in the trypsin-only data from its s.d. in the LiP data. *P* values corresponding to the difference in s.d. values were estimated using the robust Brown–Forsythe Levene-type procedure from the R package lawstat (version 3.4)[98]. Peptides with a score above 1 and a *P* value below 1×10^{-5} were defined as highly variable peptides. Scores between 0.5 and 1 with a *P* value below 1×10^{-5} were defined as medium-variable peptides. If the VarS was below 0.5 or the *P* value was above 1×10^{-5}, peptides were defined as non-variable. We observed that some peptides had a bimodal distribution of their signals in the LiP data, indicating a 'switch-like' structural change of the protein, that is, a protein populating two states. In order to systematically identify such peptides, we used the function modetest from the R package multimode (version 1.5) on the normalized LiP intensities[99] (Supplementary Tables 4 and 5).

Disorder prediction. Intrinsically disordered regions were predicted from full-length protein sequences using DISOPRED version 3.1 (ref. [30]), relying on BLAST version 2.2.26 and the uniref90 database. The mean disorder score was computed over every peptide. Out of 9,888 peptides that were included in the analysis of variable peptides, mean disorder scores were successfully computed for 9,235 peptides. Statistical significance for differences in the degree of disorder for highly-variable, medium-variable, and non-variable peptides was assessed using Fisher's exact test; peptides with a mean disorder score >0.5 were considered disordered, whereas peptides with a mean disorder score <0.5 were considered ordered.

Secondary structure prediction. Secondary structures were predicted from protein sequences using Psipred v3.5 (ref. [100]). The secondary structure type (helix, beta strand, or loop) with the highest likelihood was assigned to every residue of the protein sequence. On the basis of the assigned secondary structure for every residue, the content of helical, beta strand, or loop secondary structure was calculated for all 9,888 peptides included in the analysis of inter-individual variability using a custom-made python script, facilitated by the rstoolbox library[101]. Statistical significance for differences in how often different secondary structures are present in highly-variable, medium-variable, and non-variable peptides was assessed using the Wilcoxon rank-sum test. All Wilcoxon rank-sum tests in this paper were performed as two-tailed tests.

Other structural descriptors. Sequence-level structural and functional annotations were obtained from DescribePROT (accessed October 2020), a recently released database of structural and functional annotations of proteins[102]. DescribePROT contains a collection of pre-computed scores from previously developed predictors,

such as SCRIBER (protein-protein interaction propensity)[29], VSL2B (disorder)[103], ASAquick (solvent-accessible surface area)[104], and DRNApred to predict DNA- and RNA-binding propensity[105]. Scores were obtained for 9,754 out of 9,888 peptides included in the analysis of inter-individual variability. Statistical significance for differences in sequence-level structural and functional annotations of highly-variable, medium-variable, and non-variable peptides was assessed using the Wilcoxon rank-sum test.

STRING analysis. Physical and functional protein-protein interactions were obtained from the STRING database, version 11.0 (ref. [106]) (accessed in December 2020). To capture both physical and functional interactions, we focused on the STRING combined score that is a weighted combination of probabilities in different evidence channels. Because the STRING analysis is carried out on proteins, rather than peptides, it was necessary to classify proteins as variable or invariable. Proteins that contained at least one highly variable peptide were classified as variable (64 out of 994 proteins in total), whereas all other proteins were classified as invariable (930 proteins). High-confidence interactions, defined by a cut-off of STRING combined score > 0.9, were counted for each protein. Statistical significance between the number of STRING interactors per protein defined as variable and invariable was assessed using the Wilcoxon rank-sum test.

Domain analysis. Protein domains, as annotated in the PFAM database, were extracted from UniProt (December 2020). Unique domains were counted for each protein, and the number of domains compared between variable and invariable proteins, defined as described above. The statistical significance of the different number of domains in variable and invariable proteins was assessed using the Wilcoxon rank-sum test.

Structural figures. Crystal structures of selected proteins were obtained from the Protein Data Bank[107], accessed in January 2021. Only structures that contained the peptide sequences of interest (that is, candidate biomarker peptides or variable peptides) were considered. If multiple PDB structures were found, structures of the wild-type sequence with the highest resolution were preferred. The peptide sequences were aligned to the PDB sequence and colored according to their significance using a custom-made python script. Similarly, UniProt annotations of functional sites and binding sites were mapped to the PDB structure, and their Euclidean distance to the peptides of interest was computed. All structural images were generated in PyMOL version 2.4.1.

Inferring structural variation for Parkinson's disease in cerebrospinal fluid data

Estimating effect sizes of covariates on peptide intensities and protein abundances. Cohort effects and effects of other covariates on the intensities of LiP peptides were estimated using linear models. The aim was to identify peptides that were significantly differently digested by proteinase K between the two cohort groups (PDG versus HG) and therefore indicate structural rearrangement (accessibility change) of the corresponding protein. An increase in accessibility will lead to increased production of half-tryptic peptides resulting from proteinase K digestion, while the relative abundance of the respective fully tryptic peptides from the same protein region declines. Here, we used only changes in the abundance of fully tryptic peptides as indicators of accessibility changes because the fully tryptic peptide signal has lower noise than that of half-tryptic peptides.

Normalized LiP peptide intensities (Pep_{LiP}) were modeled as a function of various technical and biological factors, specifically the level of the host protein, the intensity of the peptide in the trypsin-only data, the total protein content (*TPC*) of the sample, batch (*Batch*), age (*A*), sex (*S*), and disease status (*C*, healthy or PD). In addition, we modeled interactions between some of these factors, as detailed below. We observed that intensities of LiP peptides were correlated with the total protein abundance estimated from the trypsin-only intensities ($Prot_{to}$) of each sample, and thus included this confounder in the linear models. By including peptide intensities from the trypsin-only treated samples (Pep_{to}) in the model, we accounted for potential peptide-specific intensity changes that were not due to variable proteinase K accessibility, such as post-translational modifications, protein isoforms, or endogenous cleavage. Such changes would equally affect peptide intensities in the trypsin-only and LiP samples and would thus be PK-independent. Because we lack the TPC measurement of one of the healthy samples, this sample was removed from all of the following analyses, resulting in 51 healthy and 51 PD samples being included.

The resulting full model is:

$$Pep_{LiP} = \beta_0 + \beta_1 * A + \beta_2 * C + \beta_3 * S + \beta_4 * A * C + \beta_5 * A * S + \beta_6 * C * S + \beta_7 * Pep_{to} + \beta_8 * Prot_{to} + \beta_9 * TPC + \beta_{10} * Batch$$

The age (*A*), cohort group membership (healthy versus PD; *C*), and sex (*S*), as well as their two-way interactions, were modeled, β_1 - β_{10} are the regression coefficents. Cohort and sex are categorical variables that need to be modeled as factors using so-called 'dummy variables.' Different encodings of categorical variables are possible, and the encoding choice affects the interpretability of the resulting coefficients (contrasts). In order to aid the interpretation for this study, we have chosen the following encodings: cohort group was modeled using a dummy coding defining the cohort variable *C* for 'healthy' as 0 and for 'PD' as 1. Thus, coefficients of the cohort term reflect the deviation of the peptide signal in the diseased cohort group relative to healthy as a reference. Sex, however, was encoded using deviation coding, setting the variable *S* for 'male' to −1 and for 'female' to +1. Batch was modeled as a factor using dummy coding, with the first measured batch being used as the reference group. Note that the choice of the encoding also affects the interaction terms. We never tested for the statistical significance of the interaction terms, because we deemed the cohort size too small to enable detection of interactions. However, we included interaction terms in the model as a conservative approach to avoid overestimating marginal effects. After retrieving all linear models, we exclude those which were not able to model at least 80 of the 103 samples (owing to missing data) from any further analysis. The statistical significance of coefficients (contrasts) was determined using *t* statistics (two-tailed test of hypothesis). The *t* value $t_{\hat{\beta}}$ was calculated by dividing the coefficient $\hat{\beta}$ of a variable *i* by its standard error $S_{\hat{\beta}}$:

$$t_{\hat{\beta}} = \frac{\hat{\beta}}{S_{\hat{\beta}}}$$

The number of peptides detected per protein strongly varies depending on the protein and its abundance. Thus, proteins with many peptides have a higher probability that one of their peptides shows a significant effect by chance compared with proteins with few peptides. Hence, we corrected for this multiple testing effect by computing corrected *P* values using the Benjamini–Hochberg approach over all peptides from the same protein ('protein-wise'). Peptides with a corrected *P* value ≤ 0.05 were defined as candidate biomarker peptides and used for subsequent steps (Supplementary Tables 4 and 7).

Investigation of the *P* values over all peptides corresponding to age, cohort, and sex showed a depletion of small *P* values for the sex (Fig. 3b). To test whether the modeling of sex and interactions with sex was biasing *P* values of the cohort variable, three additional linear models were calculated with slight alterations to the model above: (1) sex was excluded as a main factor and all interactions with sex were removed from the model; (2) sex was excluded as a main factor but interactions with sex were kept in the model; (3) sex was included as a main factor but all interactions with sex were removed from

the model. No clear difference in the number of candidate peptides for the cohort effects resulted from this, suggesting that we are not inferring a bias for smaller *P* values for the cohort variable (Extended Data Fig. 3a).

Intensity variation of peptides in the trypsin-only samples results from effects other than protein accessibility changes, such as variation in protein abundance, protein isoform changes, or post-translational modifications. To model those effects, we created a corresponding model for the trypsin-only data, including (as above) two-way interactions between age (*A*), sex (*S*), and cohort (*C*):

$$Pep_{to} = \beta_0 + \beta_1 * A + \beta_2 * C + \beta_3 * S + \beta_4 * A * C + \beta_5 * A * S + \beta_6 * C * S + \beta_7 * Prot_{to} + \beta_8 * TPC + \beta_9 * Batch$$

The coefficients of the peptide variation model were evaluated using the same analysis as that applied to the coefficients of the LiP model. This model can also be adapted to model the half-tryptic peptides from the LiP data. Because there are no corresponding peptides in the trypsin-only data, this term is removed, leading to the model shown above (Supplementary Tables 4 and 8).

Last, we also modeled protein abundances. These were fit to batch, TPC, age, cohort, sex, and the two-way interactions of the last three:

$$Prot_{to} = \beta_0 + \beta_1 * A + \beta_2 * C + \beta_3 * S + \beta_4 * A * C + \beta_5 * A * S + \beta_6 * C * S + \beta_7 * TPC + \beta_8 * Batch$$

The evaluation was again performed using *t* statistics. Because protein abundances were used, no protein-wise *P* value adjustment was necessary (Supplementary Tables 4 and 9).

Estimating scaled residuals to reflect the impact of the cohort group membership on a single LiP peptide. Peptides with a significant difference of the cohort coefficient in the LiP model are predictive of the disease status of a sample. To actually predict the disease status of a sample given the signal of a specific peptide, we first estimated the expected intensity of that peptide if the person was healthy. By computing the expected intensity, we correct for confounders such as age, sex, or batch. We then compared the observed intensity with the expected intensity in order to predict the disease status.

We used the coefficients of the peptide-specific linear model to estimate the expected intensity of the LiP peptide P_{HG} in sample *i* if the donor was healthy:

$$P_{HG,i} = \beta_0 + \beta_1 * A_i + \beta_2 * HG + \beta_3 * S_i + \beta_4 * A_i * HG + \beta_5 * A_i * S_i + \beta_6 * HG * S_i + \beta_7 * Pep_{to,i} + \beta_8 * Prot_{to,i} + \beta_9 * TPC_i + \beta_{10} * Batch_i$$

Because the cohort variable is 0 for HG samples, the formula can also be re-written as:

$$P_{HG,i} = \beta_0 + \beta_1 * A_i + \beta_3 * S_i + \beta_5 * A_i * S_i + \beta_7 * Pep_{to,i} + \beta_8 * Prot_{to,i} + \beta_9 * TPC_i + \beta_{10} * Batch_i$$

Subsequently, by subtracting the predicted intensity for a healthy sample $P_{HG,i}$ from the observed LiP intensity of the sample $Pep_{LiP,i}$ the residual (R_i) was retrieved:

$$R_i = Pep_{LiP,i} - P_{HG,i}$$

The healthy samples result in residuals that are, on average, closer to zero than the PD samples. Because our model contains two-way interactions, an additional scaling step was applied. The two-way interactions not only include information about the cohort group but also about the age and the sex of the PD donors. If we completely remove the two-way interaction terms from the predictions, the residuals estimated for the PD donors will also include variation due to the age and sex of the samples. At the same time, this is not the case for the healthy samples (because 'healthy' was encoded as 0, any interaction with 'healthy' will be 0). To prevent this difference between predictions for HG and PDG samples, we also computed the expected intensity if the sample came from a PD individual ($P_{TO,i}$) and then scaled the residuals by the difference between the expected intensities assuming healthy versus diseased (that is, $P_{PDG,i} - P_{HG,i}$):

$$P_{PDG,i} = \beta_0 + \beta_1 * A_i + \beta_2 * PDG + \beta_3 * S_i + \beta_4 * A_i * PDG + \beta_5 * A_i * S_i + \beta_6 * PDG * S_i + \beta_7 * Pep_{to,i} + \beta_8 * Prot_{to,i} + \beta_9 * TPC_i + \beta_{10} * Batch_i$$

thus

$$P_{PDG,i} - P_{HG,i} = \beta_2 * PDG + \beta_4 * A_i * PDG + \beta_6 * PDG * S_i$$

and

$$RS_i = \frac{Pep_{LiP,i} - P_{HG,i}}{P_{PDG,i} - P_{HG,i}}$$

Note that this scaling also corrects for the direction (sign) of the cohort effect: after the scaling, residuals will be positive (on average) for samples from individuals with PD, irrespective of whether β_2, β_4, and β_6 are positive or negative. It is possible that, owing to specific combinations of age and sex, the predicted intensities for HG and PDG are almost identical, thus $P_{PDG,i} - P_{HG,i}$ is numerically close to zero. Without any limitation on the denominator, this would lead to very large scaled residuals (RS_i). Therefore, samples for which the difference between these two predictions was smaller than the s.d. of the difference over all samples were instead scaled by the s.d. while preserving the algebraic sign. Taken together, this approach is a compromise between accounting for personal confounders, such as age and sex, while at the same time avoiding extreme outlier prediction values. In case of a predictive peptide, scaled residuals of HG samples will be close to 0, and scaled residuals of PDG samples will be close to +1.

Gene Ontology enrichment analysis. To obtain an overview of the proteins detected in our CSF samples, we performed a GO enrichment analysis using all proteins we detected as the foreground distribution and all proteins we searched for as the background distribution (UniProt FASTA, July 2019). We performed GO enrichment using the classic algorithm of topGO (version 2.36.0)[108]. The minimum node size was set to 10.

We additionally performed GO enrichment analysis for all proteins containing at least one biomarker candidate peptide ($P \leq 0.05$). This set of proteins could be defined as structurally affected because the protein-wise FDR was already applied to the *P* values of the candidate biomarker peptides. All proteins for which we assess structural variation (that is, all proteins with at least one peptide included in the LiP model) were used as the background distribution. GO enrichment was then performed using the weight01 algorithm of topGO, and the minimum node size was set to 10.

Classifying healthy and Parkinson's disease samples using LASSO models. We used fivefold cross-validation to test how well healthy and PD samples could be classified in independent test datasets. The data were randomly divided into five folds, with HG and PDG samples evenly distributed between the individual folds. Then, a training dataset

was defined containing four of the folds, and the last fold was used as the test dataset. Peptides that were not measured in all samples were excluded from this analysis. We then trained the LiP model (which models LiP peptide intensity variation as a function of several factors, as described above) on the training dataset and selected peptides with a significant cohort effect ($P < 0.05$; after protein-wise FDR; see above). Subsequently, we estimated scaled residuals of the candidate biomarker peptides for both the training and the test datasets based on the linear model estimated from the training data.

The scaled residuals of the training dataset were then used to build logistic regression LASSO models (R package glmnet (version 4.1-1 (ref. [109])) for classifying healthy and PD samples. For this step, missing values in the scaled residuals were imputed using the mean scaled residual of the respective training or test dataset over all samples. The regularization parameter λ was chosen such that LASSO models using one, five, or ten predictors were built. If multiple alphas were possible, nested cross-validation was used to determine the best option. In case λ could not be chosen, such that models with 5 or 10 predictors are estimated, the closest possible option was chosen, hence models with 4/6 and 9/11 predictors were calculated. Thus, we tested how well disease status could be predicted using individual peptides or combinations of peptides. Missing values were imputed as the mean value of the respective peptide (that is, assuming no effect).

This approach was applied five times, iteratively using all five folds as the test dataset once. Additionally, the separation into five folds was also repeated five times to make the results more stable, resulting in 25 models per number of defined predictors. For further interpretation of the data, the log (odds) were estimated. Scores for each sample resulting from models with the same number of predictors were then summarized by estimating the mean, leading to a weighted ensemble approach.

In order to obtain log (odds ratios) based on the oligomeric/total α-synuclein ratio that were comparable to the log (odds ratios) originating from the peptide models, we performed logistic regression of the disease status versus the oligomeric/total α−synuclein ratio in our cohort while applying the same cross-validation scheme as above. Samples for which no α-synuclein measurements were available were removed from the α-synuclein models. Thus, those models were trained on a slightly smaller cohort than were the models not using α-synuclein levels.

The same modeling approach was taken to predict disease status using the trypsin-only intensities, as well as the protein abundances.

Additionally, alternative linear models fitting LiP intensities were built. The first did not include the trypsin-only peptides intensities or the protein abundances. Hence, not only are structural variations fitted, but also PK-independent peptide variation as well as protein abundance variation, resulting in:

$$Pep_{LiP} = \beta_0 + \beta_1 * A + \beta_2 * C + \beta_3 * S + \beta_4 * A * C + \beta_5 * A * S + \beta_6 * C * S + \beta_9 * TPC + \beta_{10} * Batch$$

Also, models including only one of trypsin-only peptide intensities or protein abundances as coefficients were estimated, hence:

$$Pep_{LiP} = \beta_0 + \beta_1 * A + \beta_2 * C + \beta_3 * S + \beta_4 * A * C + \beta_5 * A * S + \beta_6 * C * S + \beta_7 * Pep_{to} + \beta_9 * TPC + \beta_{10} * Batch$$

and

$$Pep_{LiP} = \beta_0 + \beta_1 * A + \beta_2 * C + \beta_3 * S + \beta_4 * A * C + \beta_5 * A * S + \beta_6 * C * S + \beta_7 * Prot_{to} + \beta_9 * TPC + \beta_{10} * Batch$$

This leads to three additional types of linear models used for modeling LiP peptides, including or excluding different combinations of PK-independent peptides and protein abundance variation from the fitted value.

We note that the used cross-validation scheme guards against overfitting, because model performance is only evaluated with data or samples not used for model estimation. In addition, the models using LiP intensities and trypsin-only intensities were trained using the same set of peptides so that any overfitting due to the number of peptides (potential predictors) would be expected in both models – hence a superior performance of the structural model cannot be based on this.

Inferring structural variation in Parkinson's disease in brain data

To assess structural variation between the HG and PDG in the brain data, a linear modeling approach as for the CSF data was applied. Normalized LiP peptide intensities (Pep_{LiP}) of all fully tryptic peptides were modeled as a function of various factors, including the protein abundance ($Prot_{to}$) and peptide intensities of the trypsin-only samples (Pep_{to}). Additionally, the cohort group membership (C) was modeled as a factor using dummy coding. Age and sex were not modeled, because all donors were female and between 81 and 84. All samples were processed in one batch, and all samples were adjusted to the same starting protein concentration, so there was no need to correct for TPC. The following linear model results for the brain data:

$$Pep_{LiP} = \beta_0 + \beta_1 * C + \beta_2 * Pep_{to} + \beta_3 * Prot_{to}$$

The statistical significance of coefficients (contrasts) was determined using t statistics (two-tailed test of hypothesis), again using the same approach as for the CSF data. When validating candidates previously found in the CSF, these P values were used to compare the brain with the CSF data, that is without correcting for the number of peptides per protein. For defining overall structural changes between health and PD in the brain data, we determined the protein-wise FDR.

To infer intensity variation of peptides in the trypsin-only samples resulting from effects other than protein accessibility changes, models for the trypsin-only data were created:

$$Pep_{to} = \beta_0 + \beta_1 * C + \beta_2 * Prot_{to}$$

The coefficients of the peptide variation models were evaluated using the same analysis as applied to the coefficients of the LiP models. (Data corresponding to candidate CSF biomarker peptides, which were also analyzed in the brain data, can be found in Supplementary Table 11.)

Additionally, protein abundances were modeled:

$$Prot_{to} = \beta_0 + \beta_1 * C$$

The evaluation was again performed using t statistics. Because protein abundances were used, no protein-wise P value adjustment was necessary.

In vitro experiments using purified proteins

Limited proteolysis for in vitro experiments. Aldolase A human (SRP6370), D-fructose 1,6-bisphosphate trisodium salt hydrate (F6803), and 1α,25-dihydroxyvitamin D3 (D1530) were purchased from Sigma Aldrich. Native human vitamin-D-binding protein (ab90920) was purchased from Abcam. Fructose bisphosphate aldolase and D-fructose 1,6-bisphosphate trisodium salt hydrate was resuspended in LiP-buffer (100 mM HEPES, 150 mM KCl, 1 mM MgCl2, pH 7.4) to a final concentration of 20 µM and 20 mM (1:1,000 molar ratio), respectively. Vitamin-D-binding proteins and 1α,25-dihydroxyvitamin D3 were resuspended in LiP buffer to a final concentration of 15 µM and 50 µM (1:3.3 molar ratio), respectively. For each in vitro experiment, the purified protein and substrate were mixed and incubated

for 2 minutes at 25 °C. The subsequent LiP and trypsin-digestion steps were performed as for the CSF and brain samples, described above.

Liquid chromatography–mass spectrometry methods for in vitro experiments. The in vitro experiments the samples were measured in DIA mode. The acquisition settings were identical to the ones used for the brain samples, except the gradient length was reduced to 60 minutes.

Data analysis of in vitro experiments. The DIA raw files were searched using the default directDIA pipeline of Spectronaut 15.6.211220.50606 (Biognosys) against the human UniProt FASTA (downloaded in March 2020), including the sequence of proteinase K. The digestion specificity was changed to semi-tryptic, and all other settings were set to default. Peptide precursor abundance comparison between treated and untreated conditions was done using a moderated *t* test and Benjamini–Hochberg adjustment after median normalization (Supplementary Table 6). The significantly changed peptides were mapped on the 3D structure using the R package protti (0.2.0)[110].

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The mass spectrometry proteomics dataset generated in this study is available in the PRIDE database[111] (accession number PXD034120). Source data are provided with this paper.

Code availability

Code for the main analyses (Figs. 2b, 3, and 5) has been deposited on GitHub at https://github.com/beyergroup/Global- analyses-of-the-human-structural-proteome-to-identify-a-new-type-of-disease-biomarker. Further code for plots and other analyses is available upon request. Supplementary Table 12 contains all necessary data to use with the provided scripts.

References

80. Hughes, A. J., Daniel, S. E., Kilford, L. & Lees, A. J. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: aclinico-pathological study of 100 cases. *J. Neurol. Neurosurg. Psychiatry* **55**, 181–184 (1992).
81. Hoehn, M. M. & Yahr, M. D. Parkinsonism: onset, progression, and mortality. *Neurology* **17**, 427–442 (1967).
82. Fahn, S. et al. The Unified Parkinson's Disease Rating Scale. In Fahn, S. et al. (eds.) Recent Developments in Parkinson's Disease, Vol. 2, 153-163 (1987).
83. Roth, M. et al. CAMDEX. A standardised instrument for the diagnosis of mental disorder in the elderly with special reference to the early detection of dementia. *Br. J. Psychiatry* **149**, 698–709 (1986).
84. Alafuzoff, I. et al. Staging/typing of Lewy body related alpha-synuclein pathology: a study of the BrainNet Europe Consortium. *Acta Neuropathol.* **117**, 635–652 (2009).
85. Thal, D. R. et al. Sequence of Aβ-protein deposition in the human medial temporal lobe. *J. Neuropathol. Exp. Neurol.* **59**, 733–748 (2000).
86. Alafuzoff, I. et al. Staging of neurofibrillary pathology in Alzheimer's disease: a study of the BrainNet Europe Consortium. *Brain Pathol.* **18**, 484–496 (2008).
87. Mirra, S. S. et al. The consortium to establish a registry for Alzheimer's disease (CERAD). Part II. Standardization of the neuropathologic assessment of Alzheimer's disease. *Neurology* **41**, 479–486 (1991).
88. Montine, T. J. et al. National institute on aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease: a practical approach. *Acta Neuropathol.* **123**, 1–11 (2012).
89. Thal, D. R., Griffin, W. S. T., de Vos, R. A. I. & Ghebremedhin, E. Cerebral amyloid angiopathy and its relationship to Alzheimer's disease. *Acta Neuropathol.* **115**, 599–609 (2008).
90. Kovacs, G. G. et al. Aging-related tau astrogliopathy (ARTAG): harmonized evaluation strategy. *Acta Neuropathol.* **131**, 87–102 (2016).
91. Teunissen, C. E. et al. A consensus protocol for the standardization of cerebrospinal fluid collection and biobanking. *Neurology* **73**, 1914–1922 (2009).
92. Bruderer, R. et al. Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results. *Mol. Cell. Proteom.* **16**, 2296–2309 (2017).
93. Reiter, L. et al. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry*. *Mol. Cell. Proteom.* **8**, 2405–2417 (2009).
94. Savitski, M. M., Wilhelm, M., Hahne, H., Kuster, B. & Bantscheff, M. A scalable approach for protein false discovery rate estimation in large proteomic data sets. *Mol. Cell. Proteom.* **14**, 2394–2404 (2015).
95. 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).
96. Bruderer, R. et al. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell. Proteom.* **14**, 1400–1410 (2015).
97. Teo, G. et al. MapDIA: preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry. *J. Proteom.* **129**, 108–120 (2015).
98. Hui, W., Gel, Y. R. & Gastwirth, J. L. Lawstat: an R package for law, public policy and biostatistics. *J. Stat. Softw.* **28**, 1–26 (2008).
99. Ameijeiras, J., Rosa, A., Crujeiras, M. & Rodríguez-Casal, A. multimode: an R package for mode assessment. *J. Stat. Softw.* **97**, 1–32 (2018).
100. Buchan, D. W. A. & Jones, D. T. The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res.* **47**, W402–W407 (2019).
101. Bonet, J., Harteveld, Z., Sesterhenn, F., Scheck, A. & Correia, B. E. Rstoolbox—a Python library for large-scale analysis of computational protein design data and structural bioinformatics. *BMC Bioinform.* **20**, 240 (2019).
102. Zhao, B. et al. DescribePROT: database of amino acid-level protein structure and function predictions. *Nucleic Acids Res.* **49**, D298–D308 (2021).
103. Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K. & Obradovic, Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinform.* **7**, 208 (2006).
104. Faraggi, E., Zhou, Y. & Kloczkowski, A. Accurate single-sequence prediction of solvent accessible surface area using local and global features. *Proteins* **82**, 3170–3176 (2014).
105. Yan, J. & Kurgan, L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA-and RNA-binding residues. *Nucleic Acids Res.* **45**, e84 (2017).
106. 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).

107. Berman, H. M. et al. The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
108. Alexa, A. & Rahnenfuhrer, J. topGO: Enrichment analysis for gene ontology. Bioconductor R package (2020).
109. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
110. Quast, J.-P., Schuster, D. & Picotti, P. protti: an R package for comprehensive data analysis of peptide- and protein-centric bottom-up proteomics data. *Bioinform. Adv.* **2**, 1 (2022).
111. Perez-Riverol, Y. et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* **50**, D543–D552 (2022).
112. Chen, H. M., Lin, C. Y. & Wang, V. Amyloid P component as a plasma marker for Parkinson's disease identified by a proteomic approach. *Clin. Biochem.* **44**, 377–385 (2011).
113. Dong, M. X. et al. Serum butyrylcholinesterase activity: A biomarker for Parkinson's disease and related dementia. *Biomed Res. Int.* **2017**, 1524107 (2017).

Acknowledgements

We gratefully acknowledge all individuals who donated samples used in this project. We thank: K. van Dijk and L. Oosterveld for help collecting CSF samples and clinical datasets; N. Majbour and O. El-Agnaf for collection of the alpha-synuclein datasets; and the Netherlands Brain Bank for postmortem brain tissue samples. M.-T. M. was supported by a long-term EMBO postdoctoral fellowship (ALTF 522-2019). L. N. was funded by DFG (grant CRC 1310 and grant agreement no. 398882498) and the German Academic Exchange Service (Forschungsstipendium fuer Doktorandinnen und Doktoranden). J. G. was funded by DFG (grants CRC 680 and CRC 1310). A. B. acknowledges funding by DFG (grant CRC 1310 and grant agreement no. 398882498). P. P. was funded by a Personalized Health and Related Technologies (PHRT) grant (PHRT-506), a Sinergia grant from the Swiss National Science Foundation (SNSF grant CRSII5_177195), the Peter Bockhoff Stiftung and the ETH Zurich Foundation, Parkinson Schweiz, the European Research Council (866004), and the EPIC-XS Consortium (823839), the last two under the EU Horizon 2020 program. W. D. J. v. d. B. was financially supported by grants from Amsterdam Neuroscience, Dutch Research council (ZonMW 70-73305-98-106; 70-73305-98-102; 40-46000-98-101), Michael J. Fox foundation (17253), and Dutch Parkinson Association (2020-G01). Some figures were created with BioRender.com.

Author contributions

P. P. conceived the project. M.-T. M. conceived the experimental pipeline with input from P. P. and A. B. M.-T. M., A. B. and P. P. designed the experiments. M.-T. M. performed the experiments. M.-T. M., L. N. and F. S. analyzed the data. J. M., R. B., L. R. and W. D. J. v. d. B. collected the data. P. S. and M.-T. M. designed and analyzed in vitro experiments. W. D. J. v. d. B. provided the clinical samples. L. N. and J. G. performed the statistical analysis with input from A.B. N. d. S. supervised writing of the manuscript. M.-T. M., L. N., F. S., N. d. S., A. B. and P. P. wrote the manuscript. A. B. and P. P. supervised the project. All authors discussed and revised the final manuscript prior to submission.

Competing interests

The authors RB, JM and LR are full-time employees of Biognosys AG (Zurich, Switzerland). Spectronaut is a trademark of Biognosys AG. PP is an inventor of a patent licensed by Biognosys AG that covers the LiP–MS method used in this manuscript. WvdB performed contract research and consultancy for Hoffmann-La Roche, Roche Tissue Diagnostics, Crossbeta Sciences, Discoveric Bio and received research consumables from Hoffmann-La Roche and Prothena. The remaining authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41594-022-00837-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41594-022-00837-0.

Correspondence and requests for materials should be addressed to Andreas Beyer or Paola Picotti.

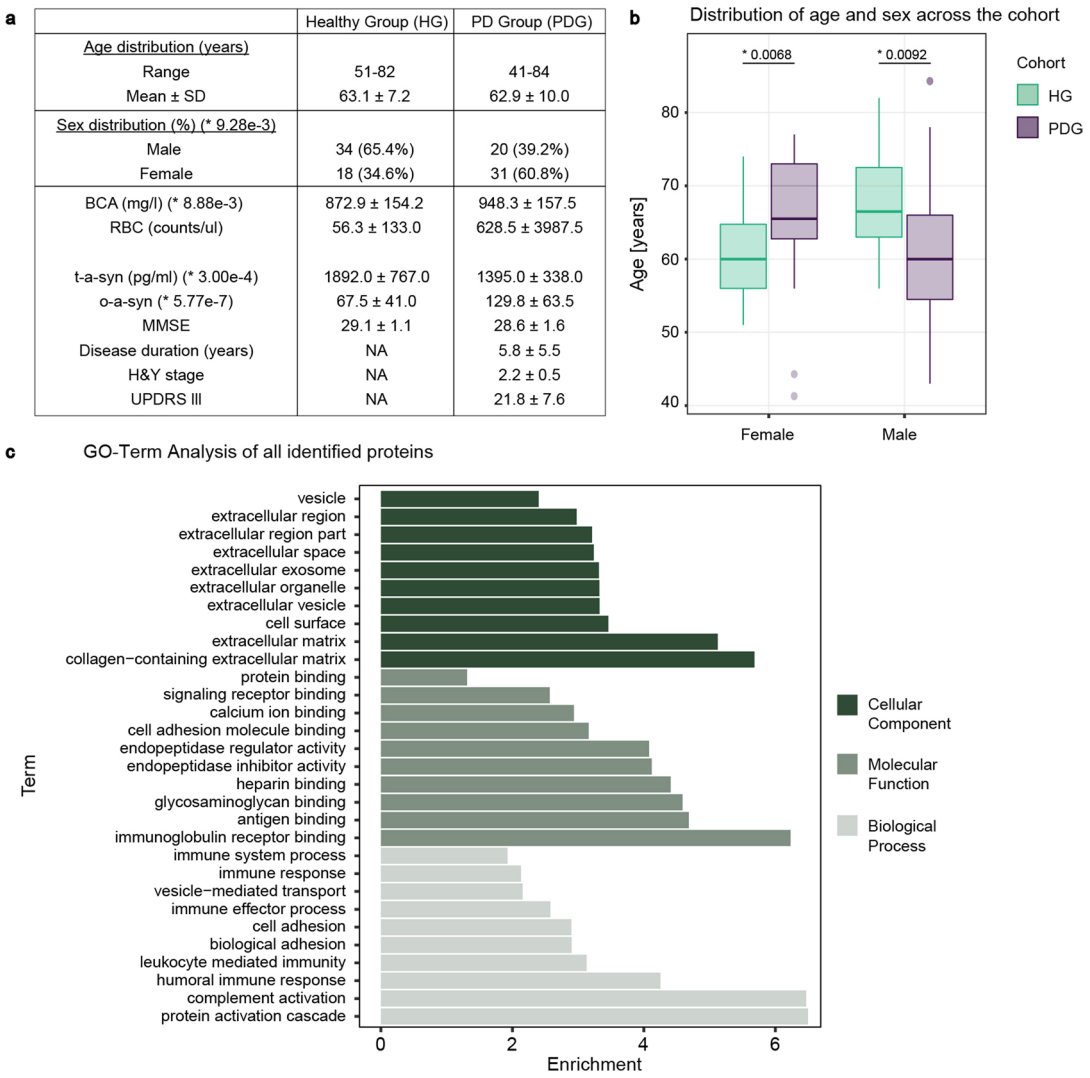
Peer review information *Nature Structural and Molecular Biology* thanks Tiago Outeiro, Marcus Bantscheff, and Laura Parkkinen for their contribution to the peer review of this work. Primary Handling editors: Anke Sparmann and Florian Ullrich, in collaboration with the Nature Structural & Molecular Biology team. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

https://doi.org/10.1038/s41594-022-00837-0

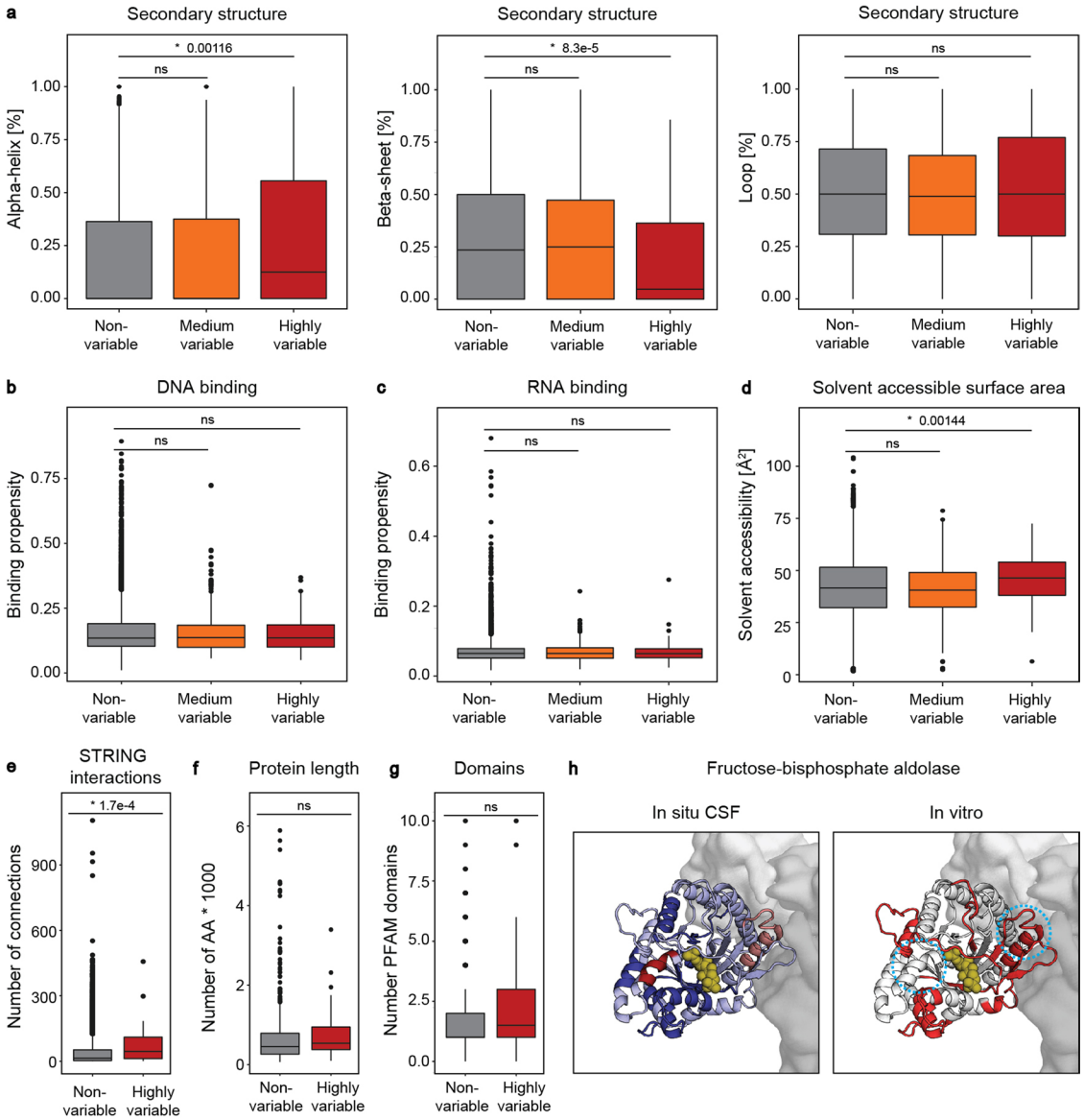
a

	Healthy Group (HG)	PD Group (PDG)
Age distribution (years)		
Range	51-82	41-84
Mean ± SD	63.1 ± 7.2	62.9 ± 10.0
Sex distribution (%) (* 9.28e-3)		
Male	34 (65.4%)	20 (39.2%)
Female	18 (34.6%)	31 (60.8%)
BCA (mg/l) (* 8.88e-3)	872.9 ± 154.2	948.3 ± 157.5
RBC (counts/ul)	56.3 ± 133.0	628.5 ± 3987.5
t-a-syn (pg/ml) (* 3.00e-4)	1892.0 ± 767.0	1395.0 ± 338.0
o-a-syn (* 5.77e-7)	67.5 ± 41.0	129.8 ± 63.5
MMSE	29.1 ± 1.1	28.6 ± 1.6
Disease duration (years)	NA	5.8 ± 5.5
H&Y stage	NA	2.2 ± 0.5
UPDRS III	NA	21.8 ± 7.6



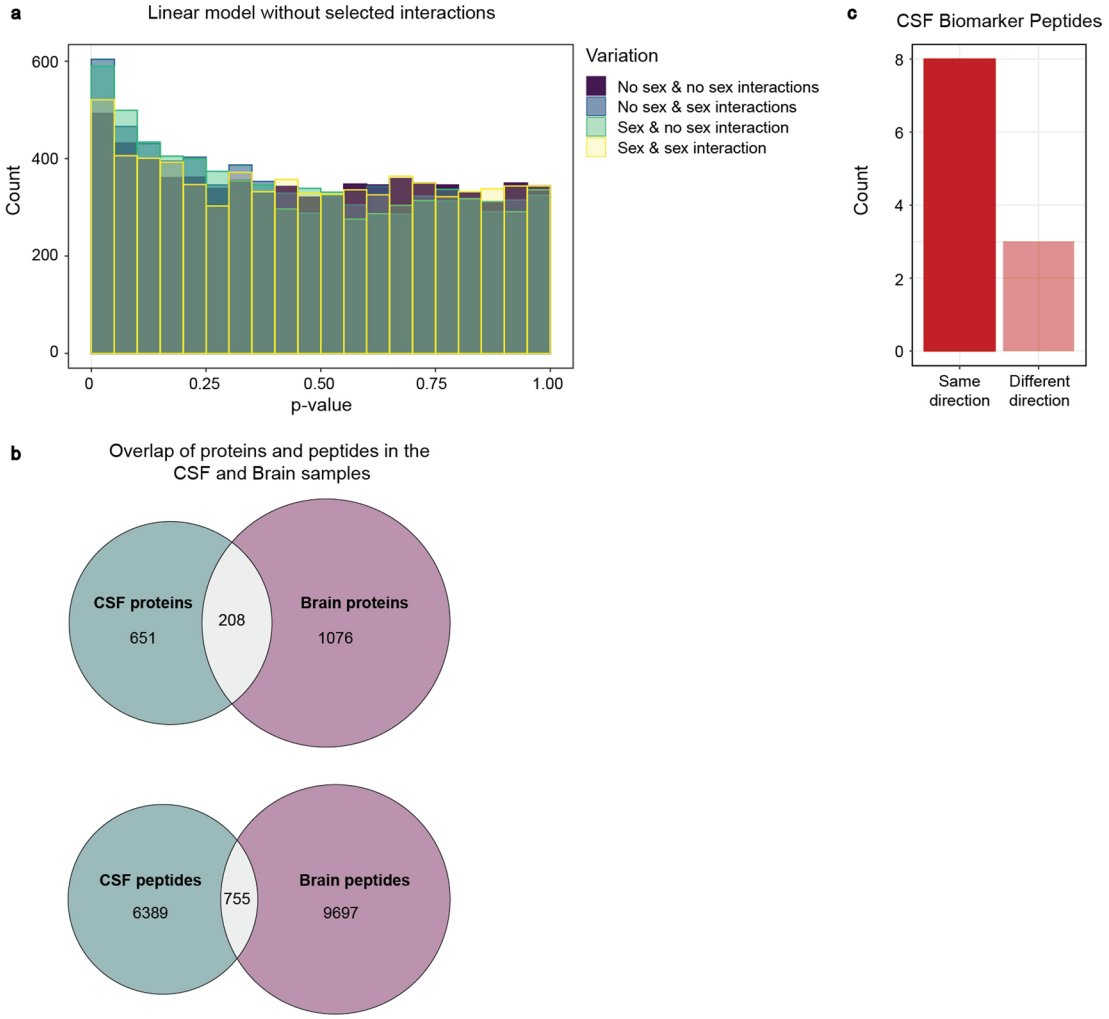
Extended Data Fig. 1 | Global characteristics of study population. **a**, Characteristics of the study cohort. P values were estimated via Wilcoxon rank sum test. **b**, Age distribution within the healthy (HG) and PD (PDG) cohort groups, separated by sex. Boxplots: median, center; first and third quantile, lower and upper hinges; largest/smallest value no further than 1.5 * inter-quantile range of the hinge, whiskers; data points beyond are defined as outliers and plotted individually. P values are indicated (Wilcoxon rank sum test, n = 51 subjects in HG, n = 52 subjects PDG). **c**, GO enrichment of all identified proteins in the CSF proteome (trypsin-only control data) using the human proteome (UniProt FASTA, July 2019) as the background. Only the 10 terms with the highest enrichment per GO domain are shown. Numerical data for graphs in b and c are available as source data.

https://doi.org/10.1038/s41594-022-00837-0



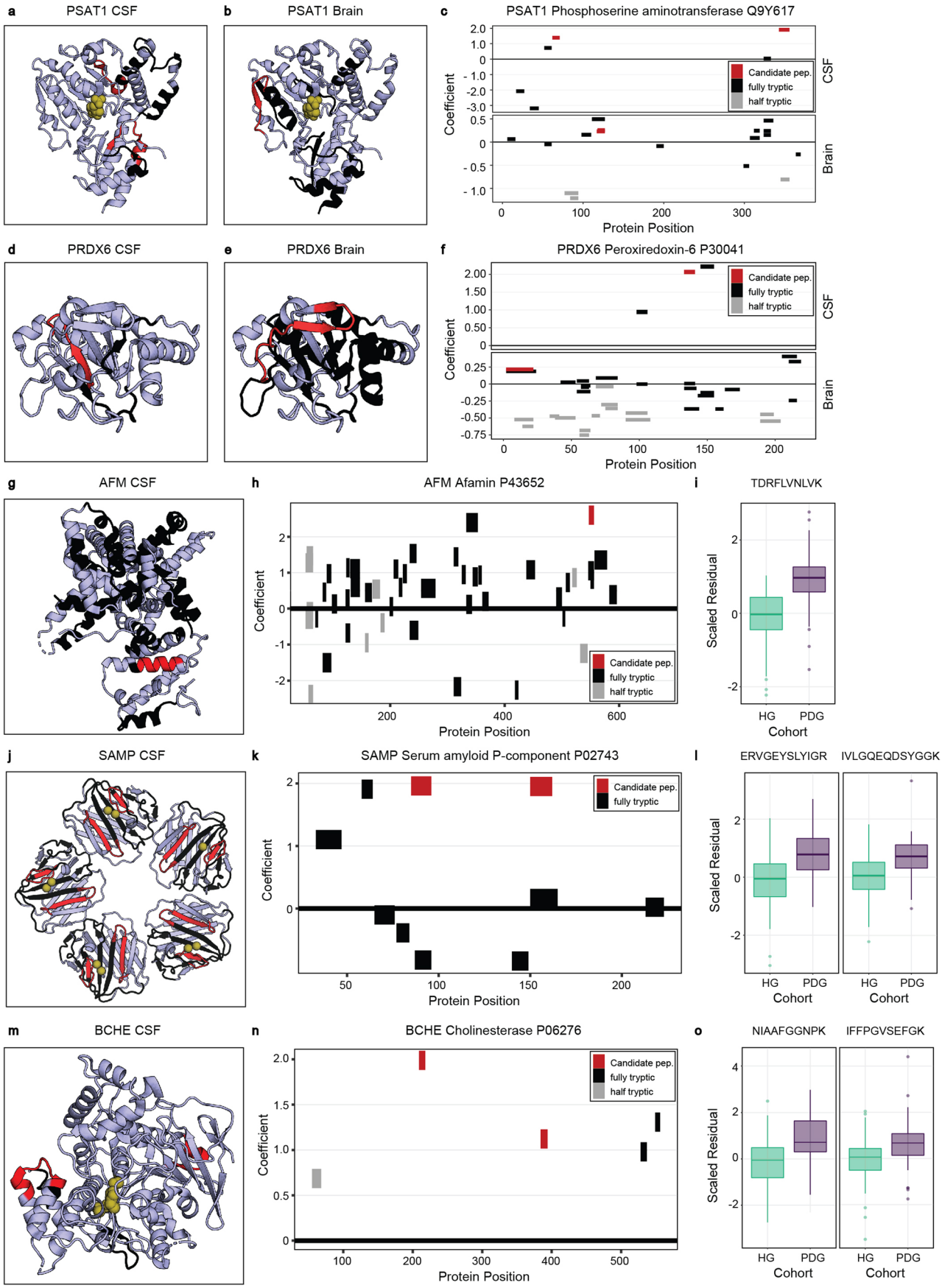
Extended Data Fig. 2 | Comparison of structural features of variable and non-variable peptides, and the proteins containing these peptides, in CSF. **a**, Distribution of secondary structures as loops, alpha-helix and beta-strands for variable/non-variable peptides, as predicted using PSIPRED. Boxplots for all panels: median, center; first and third quantile, lower and upper hinges; largest/smallest value no further than 1.5 * inter-quantile range of the hinge, whiskers; data points beyond are defined as outliers and plotted individually. P values are indicated (Wilcoxon rank sum test; ns, not significant; 9385 non-variable, 386 medium-variable, 117 high-variable peptides, from 51 subjects) **b**, **c**, Predicted propensity of peptides to bind DNA or RNA. **d**, Predicted solvent accessible surface area of variable/non-variable peptides. **e**, Number of high confidence interaction in STRING for proteins with at least one highly variable peptide (red), as compared to all other proteins (gray). **f**, **g**, Sequence length and number of domains as annotated in the PFAM database for proteins with at least one highly variable peptide (red), as compared to all other proteins (gray). **h**, Side-by-side view of affected peptides from in situ (left, reproduced from Fig. 2g for comparison) and in vitro experiments (right). Structure of human brain fructose bisphosphate aldolase (PDB entry 1XFB). The enzyme is a homotetramer, one subunit is shown as light blue cartoon and the other 3 subunits are shown as gray surface. The substrate is represented as yellow spheres, based on an alignment of PDB entry 1XFB with the ligand bound structure of the muscle isoform (PDB entry 4ALD). For the *in situ* data (left), the highly variable peptides are highlighted in dark red (bimodal) and salmon (unimodal). For the in vitro data (right), the significant peptides in the presence and absence of fructose 1,6-bisphosphate are highlighted in red ($\log_2$ fold change < −1 or > 1). The bimodal and non-bimodal peptide identified in situ are encircled. Numerical data for graphs in a-g are available as source data.

https://doi.org/10.1038/s41594-022-00837-0



Extended Data Fig. 3 | Effects of the sex variable on the linear model and overlap between the brain and CSF data sets. a, The histogram visualizes the P values (calculated via t-statistics) of the cohort variable estimated from the linear model describing effects of structural variation, with the indicated combinations of the sex variable and interactions with sex taken into account. For all models, the first bar (extreme left) indicates significant (<0.05) P values. **b**, Number of proteins and peptides of the CSF and brain samples used for estimating the linear models, visualizing the overlapping peptides and proteins between the two tissues. **c**, Number of candidate peptide from the CSF where the coefficients change in the same or different direction in the brain samples. Numerical data for graphs in a and c are available as source data.

https://doi.org/10.1038/s41594-022-00837-0

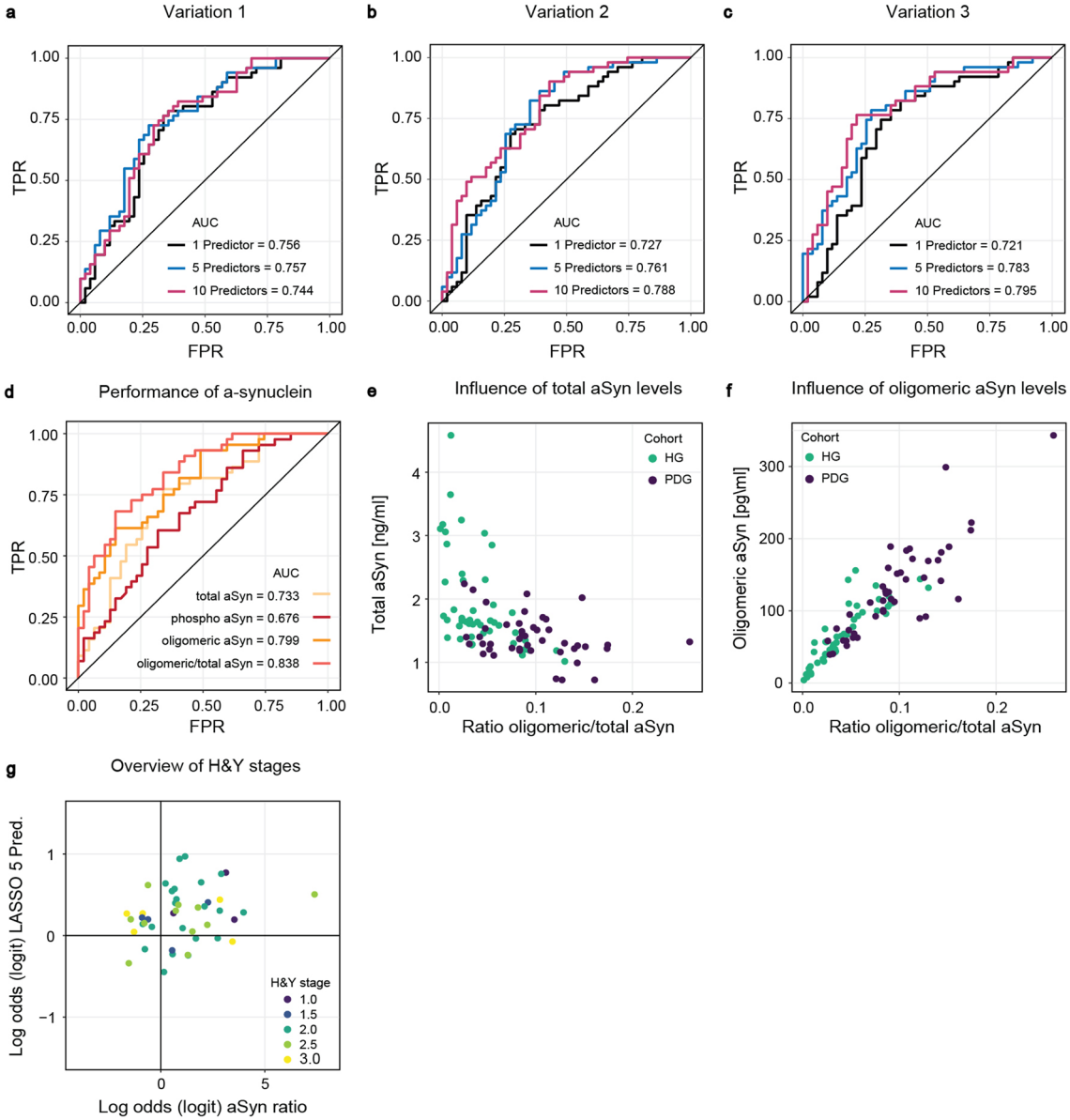


Extended Data Fig. 4 | See next page for caption.

https://doi.org/10.1038/s41594-022-00837-0

Extended Data Fig. 4 | Structural changes in selected CSF and brain proteins. **a**, **b**, Structure of PSAT1 (PDB 3E77) colored according to peptides in CSF (**a**) and brain (**b**) data. PSAT catalyzes an important step in serine biosynthesis. Black indicates all analyzed peptides; red indicates the candidate peptide (a) or peptides with a significant P value (b). One candidate CSF peptide is only 6.0 Å away from the active site. **c**, Coverage plot for all analyzed PSAT1 peptides in CSF (top) and brain (bottom). Black represents fully tryptic; gray represents half tryptic peptides and red bars the candidate/significant peptide in the CSF/brain. **d**, **e**, Structure of PRDX6 (PDB PRX1) colored according to peptides in CSF (**d**) and brain (**e**) data. Colors as in a, b. **f**, Coverage plot for all analyzed PRDX6 peptides in CSF (top) and brain (bottom). Colors as in c. **g**, Structure of AFM (PDB 5OKL). **h**, Coverage plot for AFM. AFM level in serum exosomes was linked to PD progression[4]. **i**, Scaled residual plots for AFM. Boxplots are as in Extended Data Fig. 2 (n = 51 subjects in each HG and PDG). **j**, Structure of SAMP (PDB entry 1GYK) as pentamer with bound calcium (yellow spheres), an abundance-based plasma biomarker candidate for PD[112]. **k**, Coverage plot of SAMP. **l**, Scaled residual plots for SAMP (ERVGEYSLYIGR: n = 47 subjects in HG and n = 50 subjects in PDG; IVLGQEQDSYGGK: n = 51 subjects in each HG and PDG). **m**, Structure of BCHE (PDB entry 1P0I) in complex with butanoic acid (yellow spheres). Activity of this enzyme is decreased in PD with dementia (PDD)[113], note that BCH inhibition was not used in our cohort. **n**, Coverage plot for BCHE. **o**, Scaled residual plot for BCHE (NIAAFGGNPK: n = 51 subjects in each HG and PDG; IFFPQVSEFGK: n = 43 subjects in HG and n = 48 subjects in PDG). Peptide colors in panels G, J, M are as in D; colors in panels H, K, N are as in C. Numerical data for graphs in c, f, h, i, k, l, n and o are available as source data.

https://doi.org/10.1038/s41594-022-00837-0



Extended Data Fig. 5 | Classification of Parkinson's Disease in CSF data. **a**, ROC curves for classification of PD based on LiP peptide variation. In this case, LiP peptide intensities were neither corrected for trypsin-only peptide intensities nor for protein abundance. **b**, ROC curves for classification of PD based on LiP peptide variation. In this case, LiP peptide intensities were not corrected for protein abundance. **c**, ROC curves for classification of PD based on LiP peptide variation. In this case, LiP peptide intensities were not corrected for trypsin-only peptide intensities. **d**, ROC curves for classification of PD based on ELISA measurement of different a-synuclein species from. **e**, Total α-synuclein levels compared to the ratio of the oligomeric/total α-synuclein level across the cohort. **f**, Oligomeric α-synuclein levels compared to the ratio of the oligomeric/total α-synuclein level across the cohort. For (**e**) and (**f**), each dot represents a single individual. **g**, Comparison of classification of the PDG using the ratio of oligomeric to total a-synuclein (log odds plotted on x axis) and using a combination of five LiP peptide levels (log odds plotted on y axis). Each point represents an individual and the HY-stage is indicated by color. Numerical data for graphs in a-g are available as source data.

nature research

Corresponding author(s): Andreas Beyer, Paola Picotti

Last updated by author(s): Jun 22, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Xcalibur (4.1) / Tune (2.9-290402/2.9.4.2954)
Data analysis	SpectroMine (2.0.190613.43665), human UniProt FASTA (July, 2019), Spectronaut 13 (13.5.190812), Python (3.7), DISOPRED (v3.1), Psipred (v3.5), PyMOL (v2.4.1), R (4.0.3), topGO (2.36.0), glmnet (4.1-1), lawstat (3.4), multimode(1.5), protti (0.2.0)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

PRIDE accession number for all mass spectrometry raw files: PXD034120; PDB entry: 1XFB, 4ALD, 3NVQ, 5KC6, 1J78, 3E77, PRX1, 5OKL, 1GYK, 1P0I; AlphaFold: AF-Q9GZP4-F1 ; human UniProt FASTA (July, 2019). GWAS information was retrieved from https://www.ebi.ac.uk/gwas/ accessed on 2020626. All other data needed to evaluate the conclusions in the paper are present in the supplementary materials. Source data are provided with this paper.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Calculation of sample size was not performed. We obtained the biofluid samples through the outpatient clinic for movement disorder of the VU University Medical Center Amsterdam, Netherlands. The number of samples per cohort group was restricted by the availability of suitable biofluid samples. The postmortem brain tissue samples were collected by the Netherlands Brain Bank (www.brainbank.nl) in Amsterdam, Netherlands .
Data exclusions	No data was excluded for the analysis of the variability of the healthy human proteome . One sample was excluded from the analysis of the proteomic changes of PD due to missing total protein measurement. 11 samples were excluded in the LASSO classification section when combining LiP and a-synuclein measurements due to missing a-synuclein measurements.
Replication	Measurements of the CSF samples were not replicated due to limited sample availability. For each brain tissue sample a trypsin-only and LiP sample, in triplicate, was prepared.
Randomization	Sample preparation of CSF was performed in batches of 7 with an optimal distribution of cohort group, age and sex. The brain tissue samples were all prepared together. The measurement of the samples on the mass spectrometer were carried out in an randomized fashion.
Blinding	The experiments were not carried out blinded. This approach was chosen to allow to correct for biases during data analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	CSF Healthy Group: 52 individuals (Age: 51-82, Male: 65.4%) CSF Parkinson Disease Group: 51 individuals (Age: 41-84, Male: 39.2%) Brain Healthy Group: 5 individuals (Age: 81-84, all female) Brain Parkinson Disease Group: 5 individuals (Age: 81-83, all female)
Recruitment	Patients, donating CSF, were reimbursed for travel costs and offered a lunch. No further compensation. The source population comprised of PD patients who were referred to the participating academic hospital, VU University Medical Center Amsterdam (VUmc), and fulfilled acknowledged diagnostic criteria. By including also patients from non-university nearby peripheral hospitals, the PROGRESS-PD-cohort were more representative of the general PD population. The VUmc excluded secondary and tertiary referrals of unusual cases to avoid selection bias. The healthy controls were recruited through an advertisement on the website http://www.parkinson-vereniging.nl and in the magazine 'Papaver' of the Dutch Parkinson Foundation (Parkinson Vereniging) and from spouses and acquaintances of the patients that visit the outpatient clinic for movement disorders. The controls were matched with the parkinsonian patients for age and gender. The brain tissue samples were collected by the Netherlands Brain bank (NBB) and the Normal Aging Brain Collection Amsterdam (NABCA). By signing the Informed Consent form, donors gave permission for post-mortem brain autopsy and use

	of their brain material and medical records for research purposes. Donors or relatives did not receive any compensation.
Ethics oversight	The study was approved by the local ethics committee of the VU University Medical Center, and all participants (or next of kin) gave written informed consent.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

3 Chapter II: Separating protein structural changes from changes in protein abundance, alternative splicing and post-translational modifications at high resolution and on a proteome-wide scale

Contribution statement

I implemented the LiPAnalyzeR R package, performed all analysis and created all figures in the manuscript. I was actively engaged in conceptualizing the project with input from Jan Grossbach and Andreas Beyer. Sample preparation, limited proteolysis and LC-MS/MS data acquisition was performed by Valentina Cappelletti and Christian Doerig, I performed library generation and data independent acquisition using Spectronaut with help from Christian Doerig. Valentina Cappelletti wrote the methods on sample preparation, limited proteolysis and LC-MS/MS data acquisition. I wrote the rest of the manuscript with Andreas Beyer and feedback from all co-authors.

Data and code availability

All data and code can be assessed over:
https://uni-koeln.sciebo.de/s/8lQhk9Bg0VwbLp5

The LiPAnalyzeR package has been deposited on GitHub:
https://github.com/LuiseNagel/LiPAnalyzer

Separating protein structural changes from changes in protein abundance, alternative splicing and post-translational modifications at high resolution and on a proteome-wide scale

Nagel, L.[1], Grossbach, J.[1], Cappelletti, V.[2], Doerig, C.[2], Picotti, P.[2], Beyer, A.[1,3,4]

[1] Cluster of Excellence Cellular Stress Responses in Aging-associated Diseases (CECAD), University of Cologne, Cologne, Germany.
[2] Institute of Molecular Systems Biology, Department of Biology, ETH Zurich, Zurich, Switzerland.
[3] Faculty of Medicine and University Hospital of Cologne, and Center for Molecular Medicine Cologne, University of Cologne, Cologne, Germany.
[4] Institute for Genetics, Faculty of Mathematics and Natural Sciences, University of Cologne, Cologne, Germany.

Abstract

Combining limited proteolysis with mass spectrometry (LiP-MS) is an emerging method for quantifying proteome-wide structural changes. Differences in proteinase K (PK) accessibility of native proteins are utilized to concurrently assess alterations in the structural accessibility of thousands of proteins *in situ*. However, deconvoluting the different signal contributions, such as changes in protein abundance, to the LiP-MS signal remains a challenging issue and hinders full exploitation of the data. Here, we propose a two-step approach that first removes unwanted variations from the LiP signal that are not caused by kinase (e.g., PK) accessibility variations and then infers the effects of variables of interest on the remaining signal using a contrast model. Using LiP-MS data from two yeast species and humans we demonstrate that combining peptide-level and protein-level variation from control measurements in a constrained model to remove unwanted variation outperforms simpler approaches previously employed. Our framework provides a uniquely powerful approach for deconvolving LiP-MS signals and deriving contributions of structural changes, protein abundance, post-translational modifications and alternative splicing.

Introduction

The structural state of protein is closely linked to its function and as a result to corresponding phenotypes[1]. Protein structural changes are important mediators of cell extrinsic- and intrinsic signaling, metabolic adaptation, molecular stress and genetic variability[2,3]. Recent technological advances now enable the mapping of protein structural changes on a proteomic scale[4–7]. One of those methods, limited proteolysis coupled to Mass Spectrometry (LiP-MS), utilizes differences in the structural accessibility of protein regions to an unspecified kinas, such as Proteinase K (PK), to simultaneously quantify alterations in the structural state of thousands of proteins *in situ*[4,8,9]. LiP-MS has facilitated research on various biological topics including metabolic adaptation[10], thermostability[11], cancer-related conformational changes[12,13], protein-protein interactions[14–16] including proteome-specific protein-protein interactions[17], aging related changes in yeast[18], *Caenorhabditis elegans*[19] and mice[20] as well as utilizing multi-dimensional protein-structural changes as a new class of disease biomarkers[21].

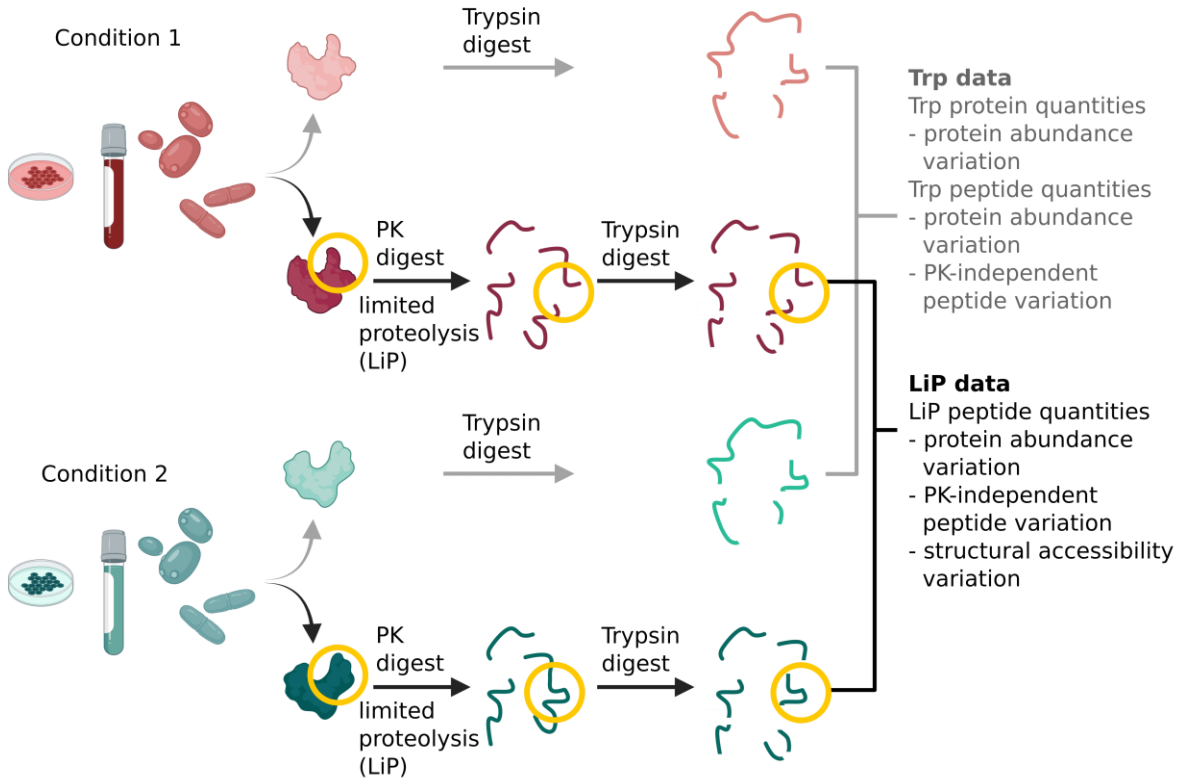


Figure 1: Experimental LiP-MS workflow. *The proteome from samples with different conditions are extracted in nondenaturing, native conditions and digested with (a) trypsin-only (b) PK (short time, limited proteolysis) followed by trypsin. Regions with a difference in the three-dimensional conformation (yellow) show a different digest pattern in the limited proteolysis step. Subsequently, LiP and Trp intensities are quantified via mass spectrometry. (Created with BioRender.com)*

To quantify structural protein accessibility changes between conditions, lysates from both conditions are first subjected to a short (i.e. limited) proteinase K (PK) digest that will preferentially digest the most accessible and flexible regions of proteins, followed by a conventional complete trypsin digest (LiP) (Figure 1). Protein accessibility differences between the two conditions will likely result in

different PK digestion patterns (Figure 1, yellow circled region). For example if a certain part of a protein is more accessible in condition 1 compared to condition 2, the respective protein region will be more digested by the unspecific PK in the first condition compared to the second. As a consequence, peptides from that protein region will be differentially digested between the conditions during the LIP step. Accessibility changes detected this way may result from protein structural rearrangements or from variable shielding of surface areas, e.g., through ligand binding. Mass-spectrometry (MS) is used to quantify those variations in peptide abundance on a proteome-wide scale.

The analysis of LiP-MS data faces a number of challenges, some of which are common to many types of molecular high-throughput data (e.g. correcting for technical variation such as batch effects, multiple hypothesis correction), while others are new and specific to LiP-MS data. First, changes in LiP peptide signals (i.e. the MS signal of a peptide subjected to the double digest with PK and trypsin) may result from factors other than PK accessibility effects, like variation in protein abundance, alternative splicing or post-translational modifications (PTMs). In order to account for such confounding effects, LiP-MS studies typically also perform a second control digest without the limited PK digest and only digesting with trypsin (Figure 1). The underlying notion is that PK-independent effects (such as protein abundance changes or alternative splicing) would equally affect both the LiP- as well as the trypsin-only (Trp) digest. Hence, MS signal changes that are common to both are likely not due to PK accessibility effects. Analyses schemes are required that appropriately combine the data from LiP-treated and Trp control samples to carve out PK-specific effects. Second, the PK digest generates many half-tryptic peptides due to the unspecific cleavage behavior of PK. Half-tryptic peptides are usually not present when only performing a trypsin digest. Half-tryptic peptides from a LiP-treated sample may carry additional information about the specific location of protein structural changes. However, the analysis of half-tryptic peptides is hampered by the fact that their sequences are more stochastic than fully tryptic peptides and that they do not exist in the trypsin-only control samples. To address these challenges, a computational approach is required that allows to distinguish the origin of variation in the LiP signal and subsequently identify the signal caused by differences in PK accessibility. Additionally, batch effects and biological variability can introduce further variation in the LiP-signal that requires methods for the removal of this unwanted variation.

Existing analysis frameworks for proteomics data do not fully address all of these challenges. Approaches, typically used in the analysis of MS data that are not specifically designed for LiP-MS data[22–24], are capable of correcting for covariates such as batch effects, but do not account for LiP-specific challenges in the data. Some methods designed for or previously applied to LiP-MS data do not fully correct for all of the unwanted variation. For example MSstatsLiP corrects LiP signals for variation caused by protein abundance changes, but it does not correct for peptide-level changes e.g. due to alternative splicing[9]. We are not aware of any tool appropriately correcting LiP peptide signals for all PK-independent variations.

Hence, we have developed a comprehensive method which enables the removal of unwanted variation (RUV) from LiP-MS data with different experimental setups and made available as the R package LiPAnalyzeR. Our analysis reveals the importance of removing unwanted variation at the protein- and peptide level, thereby creating a reference framework for the analysis of LiP-MS data.

Results

LiPAnalyzeR – A tailored bioinformatic pipeline for inferring structural variation from LiP-MS data

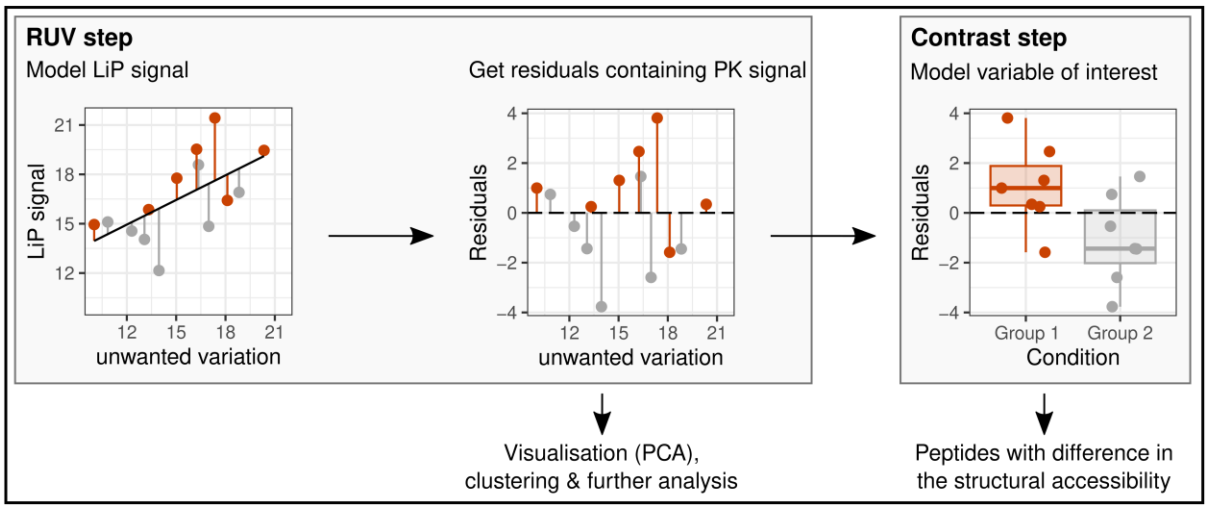


***Figure 2**: **Overview LiPAnalyzeR pipeline.** In the RUV step, unwanted variation such as PK-independent effects are regressed out from the LiP signal, resulting in residuals in which only the variation of interest is contained. These residuals can then be used for data visualization (PCA), clustering and further downstream analysis. In the contrast step, the effects of the variable of interest are then modeled for the residuals, leading to the identification of peptides which show differences in the structural accessibility linked to the modeled contrast.*

The measured signal of a peptide from the LiP-MS assay ('LiP peptide') is affected by changes in PK accessibility – which are of interest – , and other (usually unwanted) covariates like protein abundance (Figure 3a, visualized with four fission yeast strains, see the following section), changes of the proteoform (e.g. alternative splicing and post-translational modifications, PTMs; Figure 3b) and technical factors such as batch effects. LiPAnalyzeR assumes that actual structural effects (changes of PK accessibility) only affect the signal of LiP peptides, while protein levels ('Trp proteins') and signals of peptides in the trypsin-only control measurements ('Trp peptides') can only report on effects other than PK accessibility (Figure 3c). Therefore, signals of protein abundances (usually estimated from several peptides as single peptides are poor estimators of protein abundance[25]) and Trp peptides can be used to correct for biological signal variation of the LiP peptides that is not due to PK accessibility changes. LiPAnalyzeR removes unwanted variation (i.e. corrects for PK-unrelated covariates) by modeling the measured signal of a LiP peptide as a function of its covariates, regressing the measured signals on those unwanted covariates, and retaining residuals containing the variation of interest. The residuals resulting from this model may subsequently be used for further downstream analysis, such as dimensional reduction e.g. PCA or clustering, and are utilized to estimate effects of conditions or treatments of interest on protein accessibility in a second regression model. Thus, LiPAnalyzeR infers structural accessibility variation in two steps (Figure 2, Figure 3): first correcting for all unwanted covariates (RUV model) and second modeling the effects of variables of interest on the remaining signal variation (contrast model). The result of the first step is the corrected LiP signal (residual), from which the contribution of unwanted effects has been removed, hence are independent of protein abundance variation or variation of peptide levels that are not related to PK-accessibility changes. The RUV model may also be used to correct for additional biological covariates, e.g. age, if their effect on structural accessibility variation is not of interest. The contrast model is used to estimate effect sizes,

comparable to a fold change, for variables of interest and for obtaining estimates of statistical significance (p-values, see methods for details).

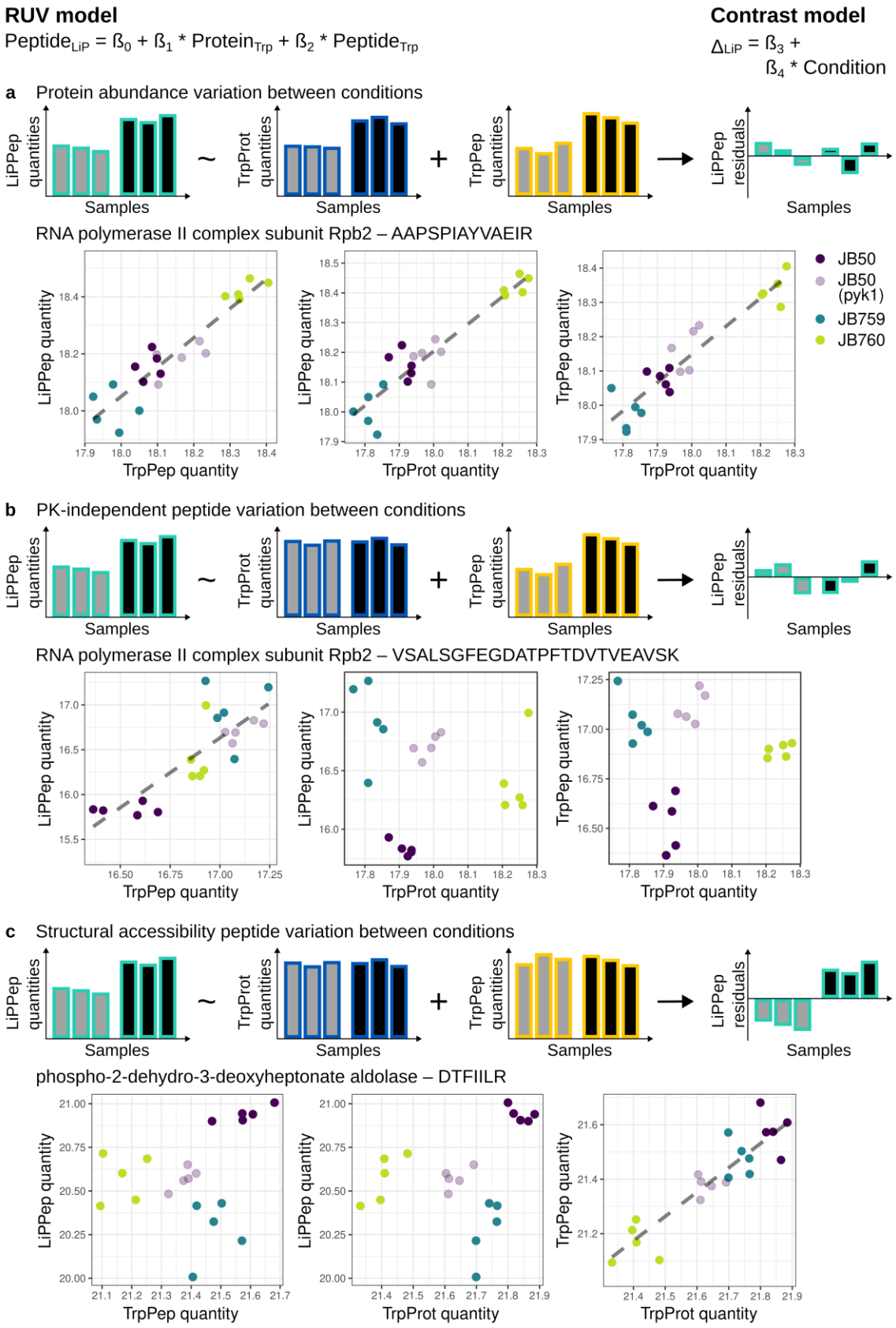


***Figure 3**: **Computational approach for inferring accessibility changes between two or more conditions from LiP data. a) Protein abundance variation.** Top: Schematic overview of the LiPAnalyzeR steps with a peptide with a protein abundance effect reflected in all LiP and Trp quantities; Bottom: LiP peptide (LiPPep), Trp peptide (TrpPep) and Trp protein (TrpProt) of a peptide showing a protein abundance difference between the JB50 and JB759 strain in the fission yeast data (gray dashed line visualizes linear regression). The RUV model removes protein abundance variation from the LiP peptides and no structural variation shows in the subsequent contrast model. Strains are plotted in different colors (JB50: purple, PYK1 mutant: light purple, JB759: blue, JB760: green). **b) PK-independent peptide variation.** Top: Schematic overview of the LiPAnalyzeR steps with a peptide with a PK-independent peptide variation reflected in LiP and Trp peptides. Bottom: LiP peptide, Trp peptide and Trp protein of a peptide showing a PK-independent peptide variation between the JB50 and JB759 strain in the fission yeast data. The RUV model removes PK-independent peptide variation from the LiP peptides and no structural variation shows in the subsequent contrast model. **c) Structural accessibility variation.** Top: Schematic overview of the LiPAnalyzeR steps with a peptide with a structural accessibility variation reflected only in the LiP peptides. Bottom: LiP peptide, Trp peptide and Trp protein of a peptide showing a structural accessibility variation between the JB50 and JB759 strain in the fission yeast data. The RUV model does not account for the structural accessibility variation, hence the signal is still reflected in the resulting residuals and detected by the contrast model.*

Datasets used for benchmarking LiPAnalyzeR

Multiple datasets were utilized for exploring the analysis of structural accessibility changes from LiP-MS data and benchmarking our approach implemented in LiPAnalyzeR. These datasets encompass various species and experimental setups, including multiple yeast species and human data, reflecting distinct organisms.

We used datasets from two distantly related yeast species – fission yeast (*Schizosaccharomyces pombe*) and budding yeast (*Saccharomyces cerevisiae*). The fission yeast data consists of three different isolates (JB50, JB759, JB760) and a mutant of one of these isolates (JB50 PYK1, see methods)[26]. For each of the four strains, five biological replicates were measured in matching LiP and Trp samples. This dataset provides a more diverse biological variation compared to other yeast datasets used in this study, since it contains four distinct yeast strains, but does not have a high number of biological replicates and no technical replicates measured. It is used for visualizing general properties of LiP-MS yeast data and basic explorations of analyzing LiP-MS data when aiming to infer structural accessibility variation. The budding yeast dataset consists of eleven biological replicates each of two well characterized isolates which were measured in matching LiP and Trp samples[27]. Two technical replicates were measured for every individual sample (see methods for detail). This dataset has a smaller amount of biological variation, since it only contains two different strains, but provides a higher number of replicates and technical replicates, therefore serving as a dataset to investigate consistency and reproducibility of results from inferring structurally accessibility variation using different approaches.

We additionally utilized a human cerebrospinal fluid (CSF) dataset containing 52 samples from healthy individuals each measured with LiP and Trp digest[21]. This human data set differs from the yeast data, in that the samples differ more in study variables factors (such as age, sex or environment) and also have a much more complex and varied genetic background. For instance the budding yeast data contains only two fairly diverse genotypes, driving strong abundance effects for many protein levels. The human CSF data contains a greater complexity than yeast data and has many more PK-independent peptide effects, specifically alternative splicing as 95% of multi-exon genes produce at least one alternatively spliced isoform[28]. Therefore the human CSF dataset especially allows

investigating approaches to differentiate PK-independent peptide effects from variation in the structural accessibility. In addition to the fission yeast dataset, we use the human CSF for the general exploration of LiP-MS data properties and to benchmark our LiPAnalyzeR approach for inferring structural accessibility variation.

LiP peptide intensities are also affected by non-structural variation

Signals of almost all LiP peptides are positively correlated with their cognate Trp peptides and Trp proteins both in human (Figure 4a) and fission yeast (Extended Data Figure 1a), suggesting a strong contribution of PK-independent signals on LiP-peptide signals. The variation in the abundance of a protein affecting the LiP peptide intensity can be estimated more robustly by summarizing the available levels of all Trp peptides as compared to using only the matched Trp peptide of the LiP peptide of interest[25]. Human proteomes are characterized by greater complexity compared to yeast proteomes, where especially alternative splicing substantially increases the diversity of proteoforms originating from the same gene. We therefore hypothesized that accounting for peptide-specific variation in addition to protein abundance variation would be even more relevant for human samples compared to yeast samples. Indeed, Trp peptide signals and protein abundance were about equally correlated with LiP peptide signals in the human data, with the Trp peptides showing a slightly greater average correlation (Figure 4b). As opposed to that, in case of the fission yeast data LiP peptides signals were more closely correlated with the protein abundances, than with the matching Trp peptide signals (Extended Data Figure 1b).

One example of PK-independent peptide variation (i.e. nonrandom intensity changes in a portion but not all of the Trp peptides of a protein) is alternative splicing in the haptoglobin-related protein (HPR, Uniprot ID: P00739). HPR has an alternative isoform (P00739-2) and additionally a very high sequence similarity with haptoglobin (HP, Uniprot ID: P00738) and its alternative isoform (P00738-2). Here, all peptides present in the main isoform of HPR are used for visualizing a peptide-specific signal caused by alternative splicing in a human CSF cohort (Figure 4c, top). The vast majority of HPR peptides (R2-R5) is present in all four isoforms of HPR and HP, but four peptides of Region 1 (R1, residues 58-72) are not present in isoform 2 of HP (Figure 4c). Pearson correlation coefficients estimated for every peptide between LiP/Trp peptide and Trp protein signals across all samples showed a strong positive correlation for all peptides from R2-R5, but not for peptides from R1. Trp and LiP peptides located in R1 show the same deviation from the estimated protein abundances, while being highly correlated with each other, which is consistent with the notion that they may be commonly affected by alternative splicing. Investigating the behavior of the individual samples identifies samples from five donors that apparently expressed higher levels of the alternative isoform of HPR (P00739-2), while all other donors predominantly expressed the main isoform (P00739). A representative peptide from region R1 had LiP and Trp peptide signals that were lower in those five samples than expected based on the estimated protein abundance (Figure 4d, blue samples show an altered splicing pattern in R1), as opposed to an example peptide from the unaffected region R3. Protein abundance was dominated by peptides that are common to all isoforms and hence not accounting for peptide-specific PK-independent signal variation would have falsely identified a PK effect for peptides from region R1.

These examples from fission yeast (Figure 3b) and human (Figure 4) LiP-MS datasets underline the importance of correcting for both protein- and peptide-specific – but PK-independent – variation in

the LiP-MS signals. Further, the extent to which those corrections influence conclusions about specific PK effects depends on the dataset at hand.

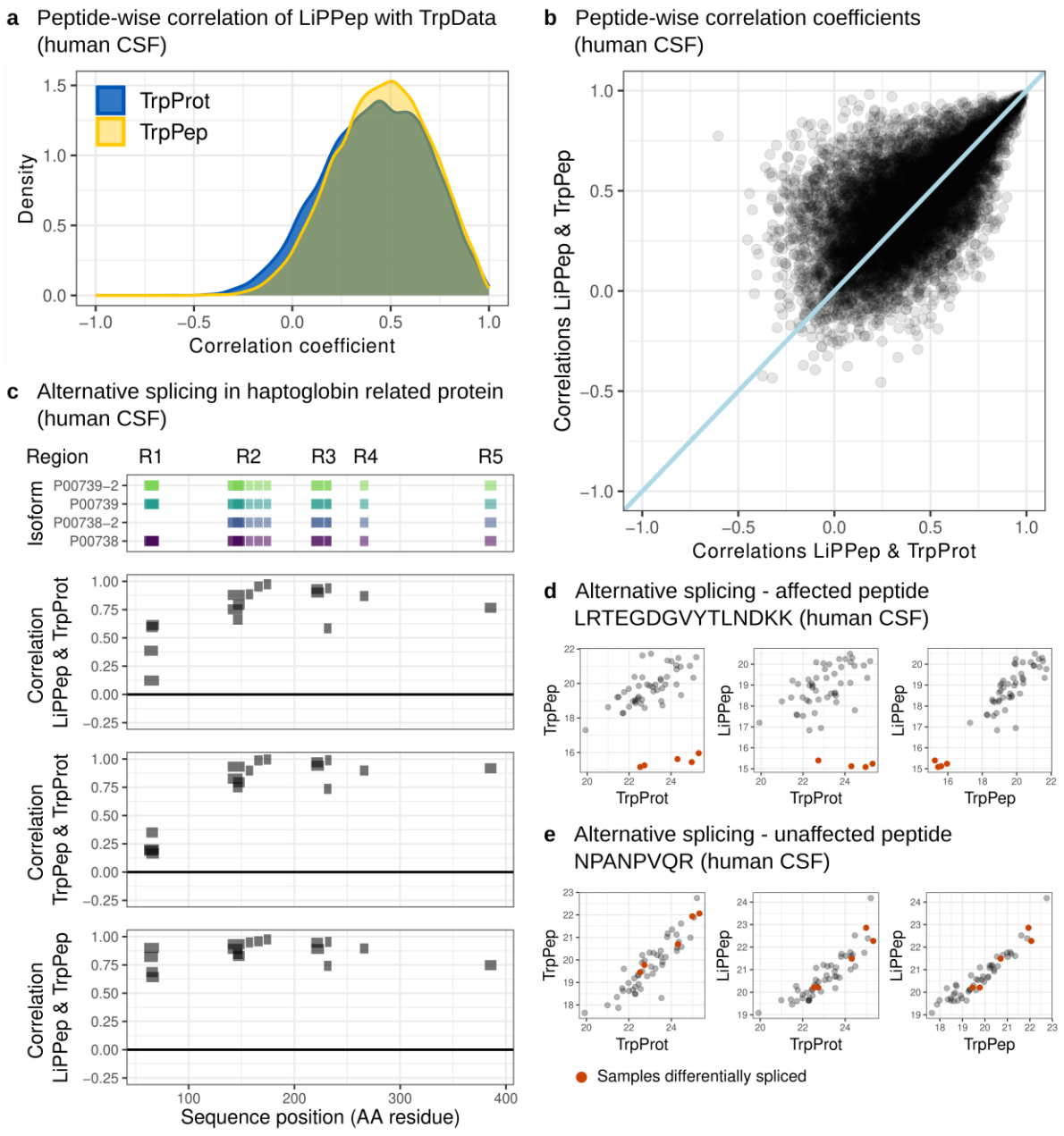


Figure 4: Investigation of effects across LiP peptides, Trp peptides and Trp protein quantities. a) *Peptide-wise Pearson correlation coefficients of LiPPep quantities with the TrpPep (yellow) and TrpProt (blue) quantities from the human CSF data.* ***b)*** *Correlation coefficients of a) plotted against each other. A line going through the origin with a slope of 1 is added (light blue).* ***c)*** *Alternatively spliced haptoglobin-related protein (P00739) visualized across LiP and Trp data. The occurrence of all peptides found in the LiP and Trp data contained in the main isoform P00739 are visualized along the protein position for the present isoforms of haptoglobin and the haptoglobin-related protein (top). Peptide-wise Pearson correlation coefficients between LiPPep & TrpProt (second from top), TrpPep & TrpProt (third from top) and LiPPep &TrpPep (bottom) are also visualized along the protein residuals. The protein is divided into five regions (R1-R5), where R1 is affected by alternative splicing.* ***d)*** *Example peptide of the haptoglobin-related protein affected by alternative splicing (R1). Samples with alternative splicing in R1*

are red. ***e)*** *Example peptide of the haptoglobin-related protein showing no effect by alternative splicing (R3). Samples with alternative splicing in R1 are red.*

Strategies for removing unwanted variation from LiP-MS data

Based on biochemical principles, the PK-independent part of the LiP signal variation should be perfectly correlated with the Trp signal variation. Based on this, one should be able to correct for the PK-independent contribution by computing ratios of LiP and Trp signals (LiP/Trp ratios). In reality however, both the LiP and Trp signals are subject to measurement noise, which reduces the actual correlation between them. Simply computing LiP/Trp ratios would be suboptimal in that noise from the Trp measurements would be added to the LiP signal. The regression-based approach used by LiPAnalyzeR addresses this problem by computing the empirical association between the two measures (LiP peptide signal and Trp peptide signal) for each peptide individually, thereby accounting for peptide-specific noise. To demonstrate this limitation of LiP/Trp ratios we have correlated LiP/Trp ratios with the respective Trp signals (i.e. with the matching Trp peptide and the protein abundance estimated from the Trp peptides (TrpProt); Figure 5). If the LiP/Trp ratios were fully corrected for PK-independent signal variation included in both the LiP peptides and Trp measurements, the LiP/Trp ratios should show little correlation with the Trp signals. This however is not the case. Instead, the LiP/Trp ratios show strong negative correlations with the Trp peptide and Trp protein signals (Figure 5a, b). This bias is particularly strong for peptides where a significant structural accessibility effect is inferred based on LiP/Trp ratios, but not based on the RUV approach of LiPAnalyzeR (Extended Data Figure 1c,d; Supplementary text). Hence, computing ratios 'corrects' for variation in the Trp data that is uncorrelated with the LiP measurements, resulting in false signals.

The RUV model of LiPAnalyzeR corrects for Trp peptide and Trp protein signals, even though the Trp peptide signal alone should, in principle, be affected by all relevant PK-independent signal variation, i.e., including variation of protein abundance and peptide effects. However, LiPAnalyzeR includes protein abundance (usually estimated from several peptides) as an additional covariate, because single peptides are poor estimators of protein abundance[25]. Integrating the information of several peptides from the same protein reduces noise in the data. Thus, PK-independent signal variation in the LiP measurements that is mainly due to protein abundance variation is generally better corrected by using aggregate measures of protein abundance based on multiple peptides. Only correcting for Trp protein but not for Trp peptide signal would not be sufficient to remove PK-independent peptide variation from the LiP signal. To demonstrate these points, LiP peptides were corrected by running the RUV regression of LiPAnalyzeR including either Trp peptide or Trp protein signals, but not both. Pearson's correlation coefficients were then computed between the residuals from these reduced RUV models and the respective Trp data type not used for the correction. If correcting only for Trp peptides or proteins alone would be sufficient to correct for all PK-independent effects, the resulting residuals should be independent of the Trp data type not used in the RUV step. This, however, is not the case. Instead, LiP signals corrected using the reduced RUV models still retain substantial correlation with Trp protein or Trp peptide signal (Figure 5c, d). In the case of fission yeast (Figure 5c) correcting only for protein abundance but not for Trp peptides induces a smaller bias compared to the human data (Figure 5d), which is in line with the earlier observation that the human data is more strongly affected by PK-independent peptide effects such as alternative splicing.

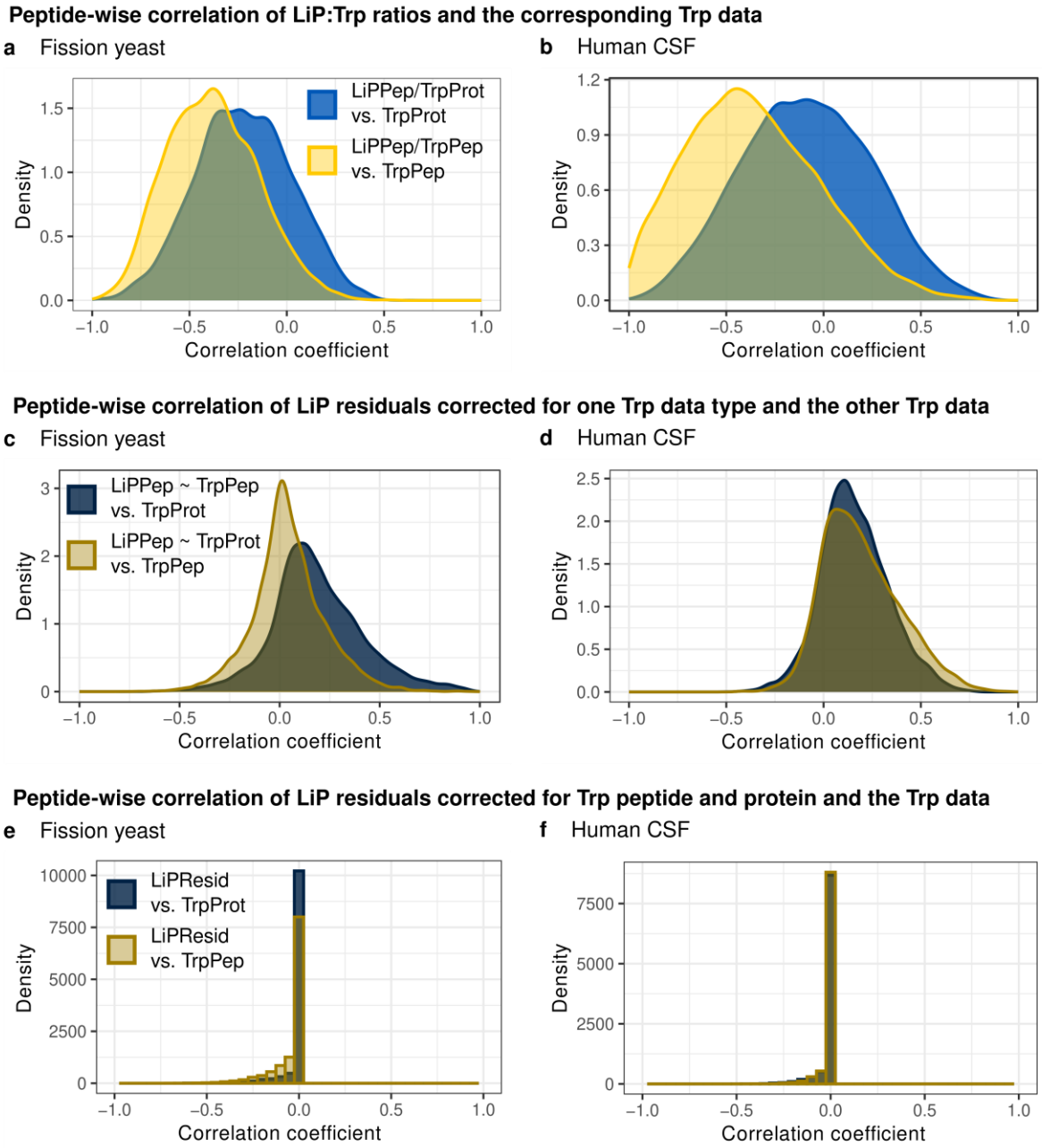


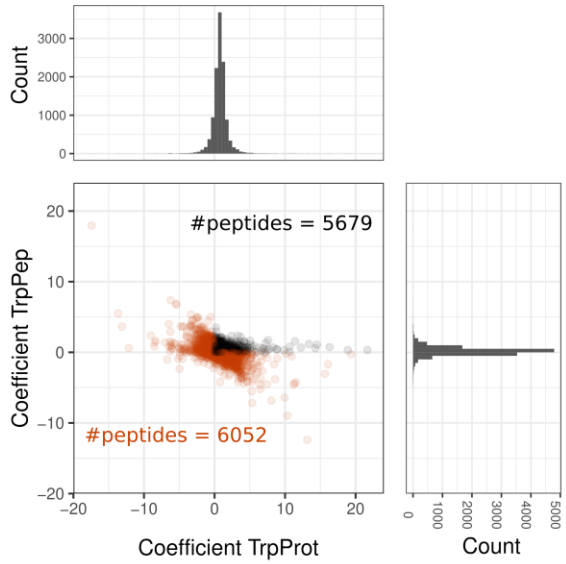
Figure 5: Correcting LiP peptides for variation also present in corresponding Trp data. a), b) *Peptide-wise Pearson's correlation coefficients between the ratio of LiP peptides to Trp peptides and Trp peptide quantities (yellow) as well as between the ratio of LiP peptides to Trp proteins and the Trp protein quantities (blue) in* ***a)*** *fission yeast and* ***b)*** *human CSF data.* ***c), d)*** *Peptide-wise Pearson's correlation coefficients between the residuals of the RUV step run with only Trp peptides or only Trp proteins from LiP peptide quantities. Residuals estimated from models using Trp peptides as a variable are correlated against Trp protein quantities (dark blue) and residuals of those models built with Trp proteins are correlated against the Trp peptide quantities (dark yellow) in both* ***c)*** *fission yeast and* ***d)*** *human CSF data.* ***e), f)*** *Residuals of complete RUV step using both Trp data as variables correlated against Trp peptide (dark yellow) and Trp protein (dark blue) in* ***e)*** *fission yeast and* ***f)*** *human CSF data.*

Applying the complete RUV model – including Trp peptide and protein quantities as covariates – to the LiP peptide completely removes any bias for the vast majority of peptides (Figure 5e, f). There is a small negative bias remaining, which results from the fact that LiPAnalyzeR constrains the coefficients for the Trp peptide and protein correction to be non-negative. Taken together, these results demonstrate that (1) a simple ratio approach leads to artifactual signals and (2) correcting for both Trp peptide signals and protein abundance is necessary.

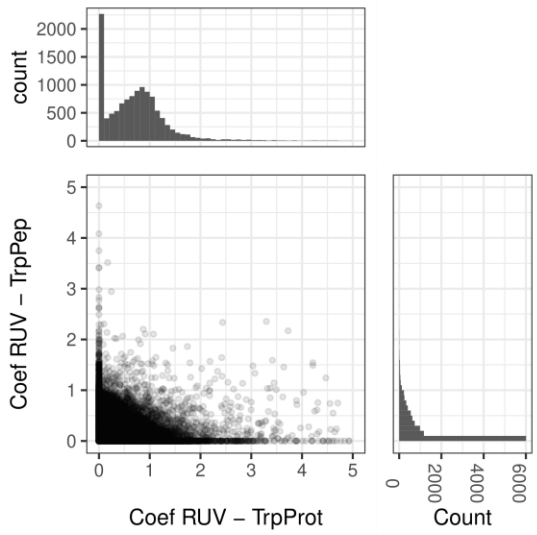
Superior performance of applying a constrained RUV model prior to the contrast model

To reduce overfitting, LiPAnalyzeR constrains the coefficients or the Trp peptide and protein signal to be non-negative. Negative coefficients would be biochemically implausible, since it can be expected that PK-independent effects affect the Trp and LiP data in the same direction, which is supported by the strong positive correlation of those data (Figure 3a, Extended Data Figure 1a). For instance, it would not be plausible that an increase in the abundance of a protein would result in a decrease in the signal of one of its LiP peptides. When LiP peptide signals are modeled without constraints, many of the Trp peptide and protein coefficients become negative (Figure 6a, Extended Data Figure 2a). Importantly, in the vast majority of cases, only one of the two coefficients is negative, which strongly suggests overfitting, where a too extreme positive coefficient estimated for one of the Trp data is 'balanced' by a negative coefficient for the other. The fission yeast peptide RALIDDSPCSEFPR from the 60S ribosomal protein L14 is an example of a peptide where applying an unconstrained RUV model results in over-fitting (Figure 6d). In this case, almost all variation in the LiP peptide signal can be explained by the Trp peptide signal, whereas the protein abundance only poorly correlates with the LiP peptide. The constrained RUV model sets the coefficient for the protein to zero and estimates the correction coefficient for the Trp peptide very close to 1, which is biochemically plausible. An unconstrained RUV model on the other hand results in a larger positive coefficient for the Trp peptide (1.172) and a negative coefficient for the protein (-1.463). Interestingly, the constrained RUV model mostly sets one of the two coefficients to or close to zero (Figure 6b, Extended Data Figure 2b, c, d). Thus, for most peptides either the Trp peptide or the protein abundance is selected, but rarely both. Noticeably, more LiP peptides are Trp peptide driven in the human CSF data than in the fission yeast data (Figure 6b,E Extended Data Figure 2b).

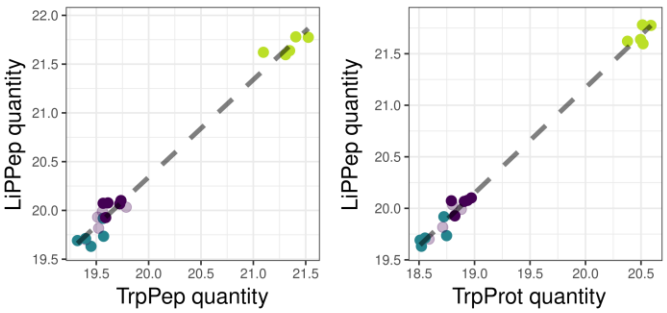
a RUV step without contraints (fission yeast)
LiPPep ~ TrpProt + TrpPep + Batch



b RUV step with constraints (fission yeast)
LiPPep ~ TrpProt + TrpPep + Batch



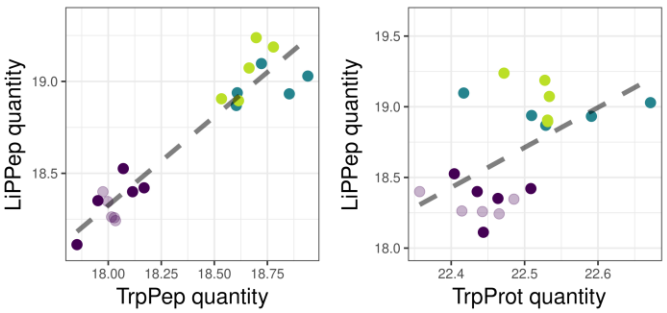
c Example peptide showing over-fitting if RUV and contrast model are combined (fission yeast)



Coefficients from RUV + contrast model combined model (c.m.):
LiPPep ~ TrpProt + TrpPep + Strain + Batch
individual model (i.m.):
LiPPep ~ TrpProt + TrpPep + Batch
LiPResid ~ Strain

	TrpProt	TrpPep	JB759	JB760
c.m.	0.073	0.152	-0.240	1.281
s.m.	0.856	0.165	0.026	-0.046

d Example peptide showing over-fitting if RUV is run without constraints (fission yeast)



JB50 JB759
JB50 (pyk1) JB760

Coefficients from RUV model
no constraints (n.c)/ with constraints (w.c.)
LiPPep ~ TrpProt + TrpPep + Batch

	Intercept	TrpProt	TrpPep	Batch
n.c.	30.099	-1.463	1.172	-0.106
w.c.	0.973	0	0.966	-0.085

e Strain coefficients from models estimated on technical replicates (budding yeast)

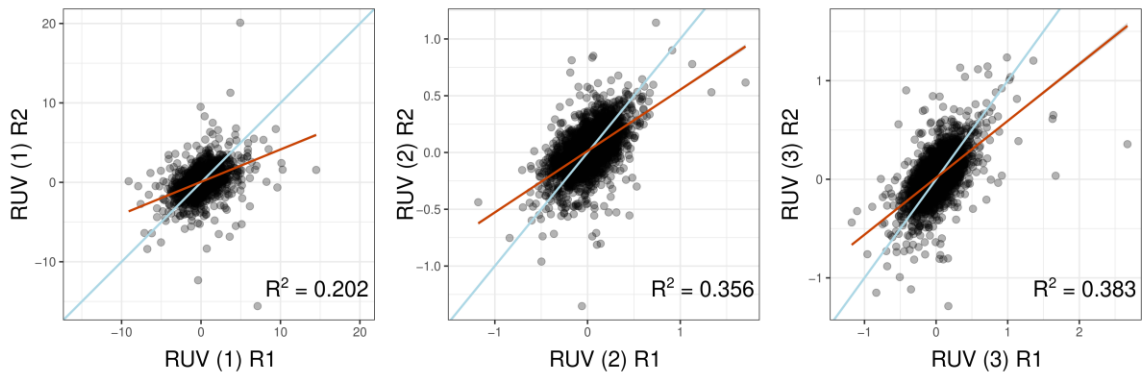


Figure 6: Comparing different regression approaches for removing Trp signal from LiP peptide quantities. a) *Peptide-specific coefficients for TrpPep and TrpProt from RUV models without constraints removing Trp variation from LiP peptide quantities in fission yeast. Peptides with at least one negative coefficient are displayed in orange.* ***b)*** *Peptide-specific coefficients for TrpPep and TrpProt from RUV models with constraints (coefficients of TrpPep and TrpProt >=0) in fission yeast.* ***c)*** *LiP peptide quantities plotted against Trp peptide (left) and Trp protein (middle) quantities for the peptide GLPLEAVTTIAK from dihydroxyacetone kinase Dak1. Different fission yeast strains are displayed in different colors (JB50: purple, PYK1 mutant: light purple, JB759: blue, JB760: green). Coefficients from modeling the LiP peptides in a combined model with RUV and contrast in one (c.m.) and modeling LiP peptides in separated RUV and contrast models (i.m.) are displayed (right). Coefficients with a significant p-value are red.* ***d)*** *LiP peptide quantities plotted against Trp peptide (left) and Trp protein (middle) quantities for the peptide RALIDSPCSEFPR from 60s ribosomal protein L14. Coefficients from a model without constraints (n.c.) and with constraints (w.c., coefficients of TrpPep and TrpProt >=0) are displayed (right). Colors as in c).* ***e)*** *Peptide-wise coefficients for strain effects estimated on the technical replicates of budding yeast data using different modeling approaches: (1) combining the RUV and contrast step into one model (Equation 5, no constraints to the model), (2) running RUV without constraints and subsequently the contrast model on the resulting residuals (Equations 1-4, no constraints in model 1) and (3) running RUV with constraints and subsequently the contrast model on the resulting residuals (Equations 1-4, constraints in model 1 as described in method section). A regression line (red) and a line going through the origin with a slope of 1 is added (light blue) is shown in each plot.*

LiPAnalyzeR first removes unwanted variation (RUV step) and subsequently estimates effects for variables of interest (contrast step). In principle it would be possible to estimate the effects of unwanted covariates (such as PK-independent variation) together with the effects of interest in a single model, which would correspond to combining the RUV model and the contrast model. While technically feasible, this approach results in an increase in the number of false positive structural accessibility changes. For example, the average protein abundance often depends on the genotype[29], hence the variable of interest (e.g. genotype) is confounded with a variable that is not of interest (e.g. protein abundance). As a consequence, LiP peptides from this protein correlated with both genotype and protein abundance – even if the folding of the protein remains unaffected by the genotype. Combining both variables in a single model can result in the protein abundance signal being modeled in the genotype (here: strain) variable, since this also correlates with the LiP peptide signal. Thus, a combined model falsely assumes an effect of the genotype on the LiP peptide (Figure 6c). The two-step approach first completely removes the protein abundance effect and subsequently no genotype effect is inferred in the contrast model, resulting in a more conservative approach.

To demonstrate that the two-step constraint regression approach chosen for LiPAnalyzeR leads to more reproducible results, we systematically explored three different modeling options: (1) combining the RUV and contrast step into one model (Equation 5, no constraints to the model), (2) running RUV without constraints and subsequently the contrast model on the resulting residuals (Equations 1-4, no constraints in model 1) and (3) running RUV with constraints and subsequently the contrast model on the resulting residuals (Equations 1-4, constraints in model 1 as described in method section). We then used the consistency of strain coefficients estimated in the contrast models from different sets of replicates as a quality measure for the models: if the coefficients are more similar between replicates, the modeling approach is more robust and less prone to overfit. These approaches were applied to technical replicates as well as biological replicates of a budding yeast dataset to compare the stability of the inferred structural variation signals (see methods for details). The strain effects estimated using the combined model were an order of magnitude larger than those resulting from the two-step

approach and less consistent between replicates (Figure 6e, Extended Data Figure 3a, b, c). Also the correction coefficients for Trp peptide and protein abundance were less consistent when using a combined model (Figure Extended Data Figure 3a, c). Further, constraining the Trp peptide and protein coefficients in the RUV model results in higher similarities between replicates compared to an unconstrained RUV model, which further supports the notion from above that an unconstrained model results in overfitting. Comparing results of the contrast modeling with the three approaches on the fission yeast data led to similar results (Extended Data Figure 4a-4f).

Taken together, these results show that first removing unwanted variation using a constrained RUV model before applying a contrasts model results in more conservative effect estimates and reduces overfitting.

Including half-tryptic peptides can improve coverage and sensitivity

The PK digest of the LiP protocol produces many half-tryptic peptides, i.e. peptides with a trypsin digestion site at only one end, while the other end of the peptide was digested by PK. Hence, for the vast majority of proteins both fully-tryptic and half-tryptic peptides can be measured, resulting in a complex pattern of many peptides spanning the protein residues (Figure 7a). There are protein regions where no matching peptides are found, other regions are only covered by fully or half-tryptic peptides, while some regions are covered by both fully and half-tryptic peptides. Often it is not only a single half-tryptic peptide which matches a fully tryptic peptide, but multiple half-tryptic peptides matching one or multiple (overlapping) fully-tryptic peptides or not matching a fully-tryptic peptide at all. Since these PK induced half-tryptic peptides do not exist in the trypsin-only control samples, they cannot be directly matched to a Trp peptide as a control like the fully tryptic LiP peptides

One strategy to deal with this ambiguity is to exclude any half-tryptic peptides from the analysis, which is the default setting in LiPAnalyzeR, but leaves a part of the data unused. Hence, LiPAnalyzeR also implements the following strategy to remove unwanted variation and then estimate coefficients of interest for half-tryptic peptides specifically. First, LiPAnalyzeR correlates the signal of a half-tryptic LiP peptide with all fully-tryptic peptides from the Trp data matching the same protein(s). Subsequently the Trp peptide with the largest positive correlation is used in the RUV model. All other aspects (e.g. correction for protein abundance) are performed in the same way as for fully tryptic peptides. Using the most strongly correlated peptide for this correction creates a sampling bias and may result in over-correcting the signal of the half-tryptic peptide. When many fully tryptic peptides were measured for a protein in the Trp condition, one of them may correlate well with a half tryptic LiP peptide by chance alone. The procedure shrinks the residuals towards zero and therefore takes a conservative approach. This shows when comparing results from this approach to those derived from the alternative approach, where only the Trp protein and not peptide quantities are given into the RUV model. Adding the most correlated Trp peptide to the RUV model results in less structural alterations being detected in half-tryptic peptides, although the overall directionality of the signal remains the same (Extended Data Figure 5d, e, f). It is also possible that a PK-independent effect included in the half-tryptic peptide, such as a PTM, is not reflected in any fully-tryptic peptide and hence cannot be accounted for. Despite this limitation half-tryptic peptides may still add useful information especially in combination with fully tryptic peptides from the same protein region.

The half-tryptic LiP peptides show a higher average correlation to the matched fully-tryptic peptides from the Trp data, than to the corresponding protein abundance (Extended Data Figure 5a, b). Hence, the Trp peptide is more frequently picked by the RUV model than the Trp protein, although the overall coefficients estimated for the Trp data by the RUV models are comparable with those estimated for the fully-tryptic peptides (Figure 7b, Extended Data Figure 5c). The signal of the half-tryptic peptides can add to the confidence one has in structural rearrangements detected in the fully-tryptic peptides before. E.g. if only a single fully-tryptic peptide in a protein shows a change in the structural accessibility, a half-tryptic peptide in the same region also having a structural signal can corroborate the finding (Figure 8c). Since the fully-tryptic peptides are individually chosen by correlation, it is unlikely that the structural effects in the Trp peptides of that region are simply induced into the LiP residuals by the model. In case of the aldolase in the fission yeast data, a strong structural effect spanning large regions of the protein can be detected in the fully-tryptic peptides between JB50 and JB759, but not between JB50 and JB760 (Figure 7d). The same pattern can be observed just as strongly in the half-tryptic peptides, with there being numerous significantly changing ones in JB759, but none in JB760. This shows how stable half-tryptic peptides can represent a structural effect also detected in the fully-tryptic peptides.

Even though the more stochastic digestion patterns of half-tryptic peptides increase their variability between replicates (Extended Data Figure 5g, h), they can still be used to gain confidence in structural alterations found by fully-tryptic peptides or detect structural variability in regions exclusively covered by half-tryptic peptides. Analysis on the fully-tryptic peptides should still be run with the matching peptides of the Trp control samples in the RUV model prior to the contrast model. Half-tryptic peptides are analyzed in an independent LiPAnalyzeR run as described above.

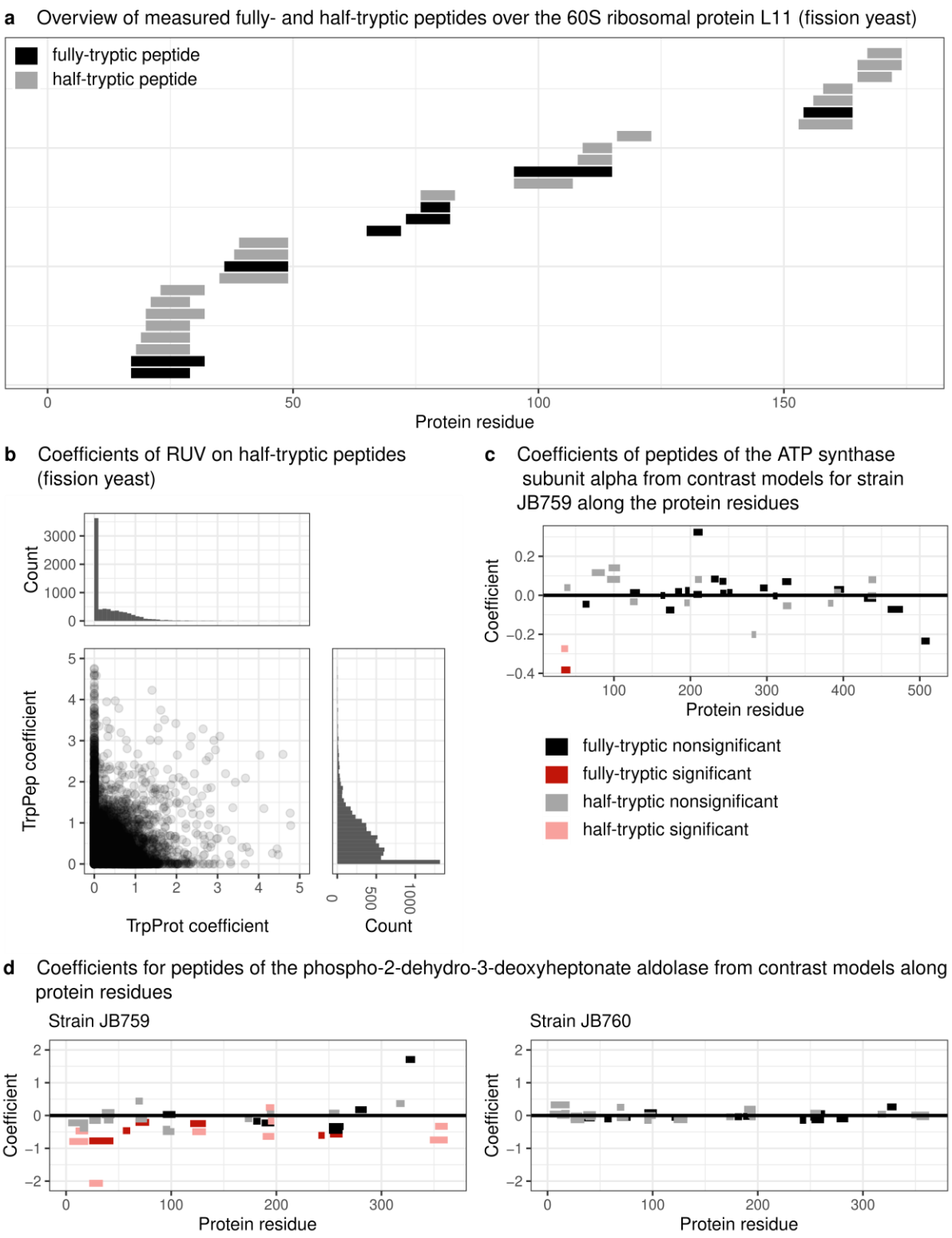


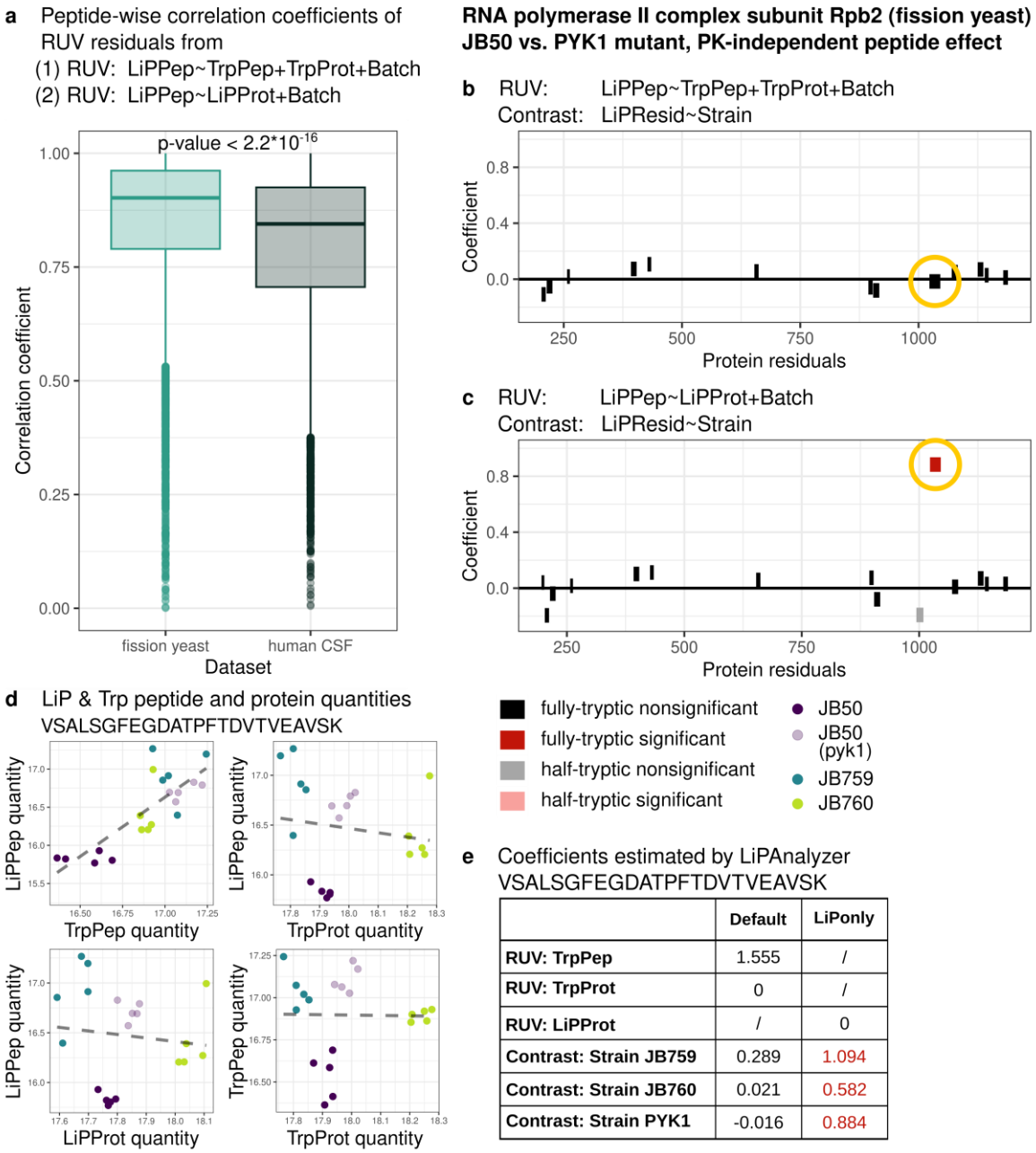
Figure 7: a) *Fully-tryptic (black) and half-tryptic (gray) peptides detected in the 60S ribosomal protein L11 in fission yeast along the protein residues.* ***b)*** *Peptide-specific coefficients for TrpPep and TrpProt from RUV models with constraints (coefficients of TrpPep and TrpProt >=0) in fission yeast.* ***c)*** *Coefficients from contrast models for fully-tryptic and half-tryptic peptides of the ATP synthase subunit alpha in the JB759 strain using JB50 as reference strain. RUV for the fully-tryptic peptides (dark) was performed with the default LiPAnalyzeR pipeline, while half-tryptic peptides (light) were analyzed in the HT-only modus. Peptides with a significant corresponding p-value are displayed in red.* ***d)*** *Coefficients from contrast models for fully-tryptic and half-tryptic peptides of the*

phospho-2-dehydro-3-deoxyheptonate aldolase in the JB759 (left) and JB760 strain using JB50 as reference strain. RUV for the fully-tryptic peptides (dark) was performed with the default LiPAnalyzeR pipeline, while half-tryptic peptides (light) were analyzed in the HT-only modus. Peptides with a significant corresponding p-value are displayed in red.

Working without trypsin-only control samples

Trypsin-only control measurements are essential for correcting for peptide-specific PK-independent effects. However, without such measurements it is still possible to correct for protein-level effects. Without trypsin-only control measurements, protein abundances may be estimated from the respective LiP measurements, assuming that the majority of peptides of a protein are not affected by structural changes. The correction for protein abundance variation in the RUV step is then performed using protein abundances estimated from the LiP data (LiPProt), while neglecting the correction of PK-independent peptide variation.

We first ran the RUV step on the fission yeast and human CSF data (1) with Trp peptide and protein quantities and (2) with LiP proteins only. Comparing the residuals from these two models showed a significantly higher correlation between the residuals in the fission yeast data compared to the human CSF data (Figure 7a). This was expected, since we observed PK-independent peptide variation to be more prominent in human CSF as compared to yeast samples before. We subsequently estimated strain effects with the contrast model, using the residuals estimated from RUV alternatives described above and obtained a high similarity between the them (Pearson's correlation coefficients strain coefficients: JB759 = 0.899, JB760 = 0.851, JB50 PYK1 mutant = 0.926).



e Coefficients estimated by LiPAnalyzer
VSALSGFEGDATPFTDVTVEAVSK

	Default	LiPonly
RUV: TrpPep	1.555	/
RUV: TrpProt	0	/
RUV: LiPProt	/	0
Contrast: Strain JB759	0.289	1.094
Contrast: Strain JB760	0.021	0.582
Contrast: Strain PYK1	-0.016	0.884

Figure 8: Example results of LiPAnalyzeR run in different modes on the peptides of the RNA polymerase II complex subunit Rpb on fission yeast strains. a) *Peptide-wise Pearson's correlation coefficients between the residuals from the RUV step run with (1) Trp peptide and protein quantities and (2) LiP protein quantities in fission yeast and human CSF data (Student's t-test: p-value < 2.2*10^{-16}). (Median, center; first and third quartile, lower and upper hinges; largest/smallest value no further than 1.5 × interquartile range of the hinge, whiskers; data points beyond are defined as outliers and plotted individually.)* ***b)*** *Peptide-specific coefficients of the PYK1 mutant estimated in the OLS model using JB50 as the reference strain plotted along the protein sequence. The OLS model was run in the LiPAnalyzeR default mode. Full tryptic peptides with no significant p-value are visualized in black, if the p-value is significant they are plotted in red. Half tryptic peptides are depicted in the same colors but in a lighter shade. The peptide VSALSGFEGDATPFTDVTVEAVSK with a PK-independent peptide effect is marked in yellow.* ***c)*** *As a), but the OLS models were run in the LiP only mode.* ***d)*** *LiP and Trp peptide and protein quantities of VSALSGFEGDATPFTDVTVEAVSK plotted against each other. Colors indicate different strains (JB50: purple, PYK1 mutant: light purple, JB759: blue, JB760: green).* ***e)*** *Table with coefficients estimated by LiPAnalyzeR for the*

fission yeast peptide VSALSGFEGDATPFTDVTVEAVSK with the RUV step using(1) Trp peptide and protein quantities and (2) LiP protein quantities. Coefficients with a significant p-value are depicted in red.

Running the RUV model utilizing the Trp measurements on the peptides of the RNA polymerase II complex subunit Rpb2 (RPB2) in the fission yeast data leads to no strain-dependent difference in structural accessibility being inferred by the subsequent contrast model (Figure 6b, 6e). On the contrary, if the RUV step is run using only the LiP protein estimates the contrast model infers a structurally accessibility variation between the strains for the peptide VSALSGFEGDATPFTDVTVEAVSK (Figure 6v, 6e). This is caused by the PK-independent peptide effect which is reflected in both the LiP and Trp peptide measurements but not in protein abundance signals (Figure 6d). Therefore, not including the Trp peptide into the model prevents the RUV model from accounting for this effect, resulting in the contrast model inferring a false positive structural accessibility signal.

Utilizing LiP protein abundances instead of Trp measurements for correcting the LiP peptides is a valid option if no Trp data is at hand. However, this approach is limited by (1) the inability of correcting for PK-independent effects on peptides and (2) the presence of large scale effects, affecting a large proportion of LiP-peptides, such as protein aggregation, thus affecting the estimation of protein levels from LiP-data. How important the removal of PK-independent peptide variation is, depends on both the research question and the individual dataset/origin species.

Discussion

LiPAnalyzeR implements a robust approach for removing unwanted variation from LiP-MS signal and subsequently identifying peptides with accessibility changes. This framework consists of a RUV step employing constrained regression, followed by a contrast model in which effect sizes and corresponding p-values representing PK-dependent signal changes are estimated. Our work has demonstrated the need to separate these two steps, because otherwise effects of confounding variables may incorrectly be split between a LiP effect and a confounding effect (e.g. protein abundance change). Such 'effect bleeding' only occurs if two covariates are confounded, such as batch confounded with treatment, or protein abundance confounded with LiP effects. Whereas the former can mostly be avoided by proper experimental design, it is practically impossible to prevent the latter. If a treatment has an effect only on the abundance of a protein, it would also alter the intensity of peptides measured in the LiP assay, even in complete absence of any folding change. A model jointly modeling the LiP signal as a function of protein abundance and treatment runs the risk to incorrectly split the effect between the two, even though there is no effect of the treatment on PK accessibility.

As shown in this work, Trp peptides and proteins mostly positively correlate with the LiP peptides. Protein abundances are generally less noisy than a single peptide measurement, since they are estimated and averaged across multiple peptides. Protein abundance is therefore often a better estimator of the true LiP peptide level than the observed matching Trp peptide level and hence, often preferentially selected by the RUV model for the correction step. If however, peptide-specific PK-independent effects dominate (such as alternative splicing or PTMs), our modeling approach will select the matching trypsin-only peptide to correct the LiP-MS signal. Consistent with this notion we found that in the comparably complex human CSF data compared to the simpler fission yeast data, Trp peptides were selected much more frequently. Hence, while accounting for Trp peptides is always beneficial to minimize false positive effects, it is even more important in data where many PK-independent peptide effects are expected.

If both Trp peptide and protein signal are removed from the LiP signal in the RUV step, it is important to prevent the corresponding coefficients from becoming negative, since negative coefficients would be biochemically implausible. Constraining coefficients within plausible ranges reduced overfitting and increased reproducibility of the effect estimation. Running the RUV step without constraints resulted in negative coefficients for over half of all peptides in budding yeast (6052 out of 11731 peptides), while in the human data less than a third of the peptides were affected (2956 out of 9863 peptides; Figure 6a, Extended Data Figure 2a). This is consistent with the previous observation: the yeast data seems to be much less dominated by peptide-specific effects. If however, there is only a protein level effect, modeling it as a function of peptide- and protein level increases the risk for overfitting. Overfitting may lead to 'overcorrection' when attempting to remove the influence of a confounding variable. Thus, the constrained approach presented here reduces to some extent the risk of 'overcorrection'. We note however that using two variables (protein abundance and protein level) to correct for PK-independent effects remains a conservative approach with respect to detecting PK-specific effects. Therefore, LiPAnalyzeR can also be run using either protein abundance or peptide level alone in the RUV step if this is deemed necessary for the question at hand.

The trypsin-only control experiment helps correct for variation in the LiP-MS data that is not due to PK-accessibility changes. Thus, these control experiments can be omitted whenever no PK-

independent variation in protein- and peptide levels is expected, e.g. when mapping drug-protein binding using cell lysates[30,31]. In all other cases it is advantageous to have such control data. If for any reason Trp data is not available, protein abundances estimated from the LiP data may also be used to correct for PK-independent signal variation. The underlying notion is that PK-accessibility changes will rarely affect the whole protein. Hence, the average peptide signal may serve as a proxy for the PK-independent signal of any specific peptide in the same protein. This approach has two disadvantages compared to having a Trp control: first, it is impossible to correct for peptide-specific PK-independent effects (such as splicing), because no independent control measurement for individual peptides exists. Second, estimating protein abundance from the LiP data may be flawed since there is no guarantee that the PK-digest does not change the average protein signal. Especially in the case of proteins with only few detected peptides, the average signal may be biased by LiP effects on one or two of those peptides. Thus, the importance of the Trp control depends on both the present dataset and the research question. The more PK-independent peptide effects, such as splicing, are expected in a dataset, the more important it is, to also add trypsin-control samples to the experimental setup. Our comparison of the fission yeast and human data sets confirmed that notion: not using the Trp controls had a much larger impact on the LiP-effect estimates in the human compared to the fission yeast data.

The RUV model generates residuals that are corrected for confounding factors such as protein abundance or PK-independent effects, thus informing on the structural state of the peptide.Therefore, beyond estimating effect sizes in a contrast model, these residuals should be used for all downstream analysis focusing on LiP effects, such as, clustering of peptides based on their structural effects across conditions or for visualizing structural changes.

LiPAnalyzeR is a versatile tool that can be applied beyond detecting protein structural effects. LiPAnalyzeR can also be used to detect PK-independent effects on the signal of Trp peptides by modeling Trp peptide signals as a function of protein levels. These effects might be indicative for the presence of alternative splicing or PTMs. As in the analysis of structural effects in LiP peptides, further variables such as batch membership may be included in the RUV model. LiPAnalyzeR additionally allows applying the models for inferring differences in protein abundance between conditions. These features enable querying a single LiP-MS dataset for a great diversity of scientific questions, and associating protein structural effects with protein abundance variation and changes in proteoforms.

Our work provides a unified, robust statistical framework for the analysis of LiP-MS data; it addresses several challenges specific to this type of data. It clearly demonstrates the need to separate the effect estimation from correcting for confounding variables and opens up new possibilities to disentangle protein structural effects from changes in protein abundance and proteoforms, thus maximizing biological insight from this rich data.

Methods

Limited proteolysis and liquid chromatography-mass spectrometry of fission yeast

Sample preparation and limited proteolysis

Five biological replicates of the Schizosaccharomyces pombe JB50, JB759, JB760 and JB50-PYK1 mutant strains were cultivated and harvested as described in Kamrad *et al.*, 2020[26]. Frozen cell pellets were resuspended in 400 µl cold LB, mixed with the same volume of acid-washed glass beads (Sigma-Aldrich), transferred to a FastPrep-24TM 5G Instrument (MP Biomedicals), and disrupted at 4°C by 8 rounds of bead-beating at 30 s with 200 s pauses between the runs. Samples were centrifuged (2 min, 1,000 g, 4°C), supernatants collected, and protein concentrations determined with the bicinchoninic acid assay (Thermo Fisher Scientific). Proteome extracts were divided into two samples: a control sample (tryptic control, TrP) undergoing only tryptic digestion to measure protein abundance changes, and a LiP sample subjected to a double-protease digestion with an unspecific protease followed by trypsin digestion to provide information on protein structural changes. Proteinase K (from Tritirachium album, Sigma Aldrich) was added to 100 µg of the LiP samples at an E:S ratio of 1:100 (w/w) and incubated for 5 min at 25°C. The same volume of water was added to 100 µg of the control samples. Digestion reactions were stopped by heating LiP samples at 99°C for 5 min, followed by the addition of sodium deoxycholate (Sigma Aldrich) to a final concentration of 5%. Control samples underwent the same procedure. Both LiP and TrP samples were then subjected to complete tryptic digestion under denaturing conditions. Peptides where reduced by incubation of samples with tris(2-carboxyethyl)phosphine (Thermo Fisher Scientific) to a final concentration of 5 mM for 30 min at 37 °C. Next, the alkylation of free cysteine residues was achieved by adding iodoacetamide (Sigma Aldrich) to a final concentration of 40 mM for 30 min at 25°C in the dark. Samples were diluted with freshly prepared 0.1 M ammonium bicarbonate to a final concentration of 1% sodium deoxycholate. Samples were predigested with lysyl endopeptidase LysC (Wako Chemicals) at an enzyme/substrate ratio of 1:100. After 2 hours at 37°C, sequencing-grade porcine trypsin (Promega) was added to a final enzyme/substrate ratio of 1:100, and samples were incubated for 16 h at 37°C under shaking at 800 rpm. Protease digestion was quenched by lowering the reaction pH (< 3) The peptide mixtures were loaded onto Sep-Pak tC18 cartridges or 96 wells elution plates (Waters), desalted, and eluted with 80% acetonitrile, 0.1% formic acid. After elution from the cartridges, peptides were dried in a vacuum centrifuge, resolubilized in 0.1% formic acid, and analyzed by mass spectrometry.

LC-MS/MS data acquisition

Peptide digests were analyzed on an Orbitrap Q Exactive Plus mass spectrometer (Thermo Fisher) equipped with a nanoelectrospray ion source and a nano-flow LC system (Easy-nLC 1000, Thermo Fisher). For shotgun LC-MS/MS data dependent acquisition (DDA), 1 µl peptide digest from each biological replicate was injected at a concentration of 1 mg/ml. 1 µl of the same samples were also measured in data-independent acquisition (DIA) mode. Peptides were separated on a 40 cm x 0.75 µm i.d. column packed in-house with 1.9 µm C18 beads (Dr. Maisch Reprosil-Pur 120). For LC fractionation, buffer A was 0.1% formic acid and buffer B was 0.1% formic acid in 100% acetonitrile using a linear LC gradient from 5% to 25% or 5% to 35% acetonitrile, respectively, over 120 min and a flowrate of 300 nL/min and the column was heated to 50°C.

For DDA measurement on the Orbitrap Q Exactive Plus, MS1 scans were acquired over a mass range of 350-1500 m/z with a resolution of 70,000. The 20 most intense precursors that exceeded 1300 ion counts were selected for collision induced dissociation and the corresponding MS2 spectra were acquired at a resolution of 35000, collected for maximally 55 ms. All multiply charged ions were used to trigger MS-MS scans followed by a dynamic exclusion for 30 s. Singly charged precursor ions and ions of undefinable charged states were excluded from fragmentation.

For DIA measurements, 20 variable-width DIA isolation windows were recursively acquired. The DIA isolation setup included a 1 m/z overlap between windows, as described in (Piazza et al., 2018). DIA-MS2 spectra were acquired at a resolution of 17500 with a fixed first mass of 150 m/z and an AGC target of 1 x 106. To mimic DDA fragmentation, normalized collision energy was 25, calculated based on the doubly charged center m/z of the DIA window. Maximum injection times were automatically chosen to maximize parallelization resulting in a total duty cycle of approximately 3 s. A survey MS1 scan from 350 to 1500 m/z at a resolution of 70,000, with AGC target of 3 x 106 or 120 ms injection time was acquired in between the acquisitions of the full DIA isolation window sets.

Data analysis: library generation and data-independent acquisition

The data were searched in Spectronaut version 15.7.220308.50606 (Biognosys). Hybrid libraries for the tryptic control and the LiP samples consisting of the corresponding DDA and DIA runs were created based on a Pulsar search using the default settings, with the exception of digest type, which was set to “semi-specific” for the LiP samples only, and the minimal peptide length, which was set to 6. The data were searched against. The data were searched against customized fasta files[32]. The targeted data extraction was performed in Spectronaut with default settings except for the machine learning, which was set to “across experiment”, imputation which was deactivated and the data filtering, which was set to “Qvalue”. The FDR was set to 1% on peptide and protein levels. The LiP and tryptic control samples were searched separately. Results were exported using the LiPAnalyzeR_SpectroScheme format.

Data preparation

LiP and Trp quantities were extracted from the exported data from Spectronaut using the columns ‘*PEP.Quantity*’ for peptide and ‘*PG.Quantity*’ for protein quantities. Peptides missing a measurement in one of the samples were removed from any downstream analysis. Peptide and protein quantities were log2-transformed. Batch correction was performed on the data using the *removebatch()* function from the r-package *limma*[33]. Analysis on the data without RUV models were performed on the batch corrected data. If RUV was performed, it was run on the not batch corrected data and batch was instead added into the model as a coefficient.

Limited proteolysis and liquid chromatography-mass spectrometry of budding yeast

Sample preparation, limited proteolysis and LC-MS/MS data acquisition

For this study we used 11 biological replicates of the laboratory BY4716 S. cerevisiae strain, an S288C derivative (MATα lys2Δ0) and 11 biological replicates of the wild isolate RM11-1a (MATa leu2Δ0 ura3Δ0 ho::KAN)[27]. Single colonies of the 2 strains were picked from frash plates and inoculated in synthetic complete medium (SC,Formedium) and grown for 6 hours at 30°C under shaking at 150 rpm. The pre-cultures were inoculated into fresh SC medium cultures to a final OD600 of 0.0004 and grown overnight at 30°C under constant shaking at 150 rpm. When cultures reached OD600 = 0.8±0.1 the liquid medium was removed by 5 min centrifugation at 800 x g, RT. The pellets were washed with 20 ml of PBS 1X and centrifuged at 800 x g, RT. Cell pellets were finally resuspended in RT lysis buffer (100 mM HEPES, 1 mM MgCl2, 150 mM KCl, pH 7.5), flash-frozen and stored at −80°C. Cell lysis was performed as described above, using a FastPrep-24TM 5G Instrument (MP Biomedicals), with the following settings: speed = 5.5 m/s, time = 30 sec, 8 cycles, rest time = 200 sec. Proteome extracts were then processed as described above and peptides analyzed on an Orbitrap Q Exactive Plus mass spectrometer (Thermo Fisher), see „LC-MS/MS data acquisition“ section above.

Data analysis: library generation and data-independent acquisition

The data were searched in Spectronaut version 15.7.220308.50606 (Biognosys). Libraries for the tryptic control and the LiP samples consisting of the DIA runs were created based on a Pulsar search using the default settings, with the exception of digest type, which was set to “semi-specific” for the LiP samples only, and the minimal peptide length, which was set to 6. The data were searched against customized fasta files[34]. The targeted data extraction was performed in Spectronaut with default settings except for the machine learning, which was set to “across experiment”, imputation which was deactivated and the data filtering, which was set to “Qvalue”. The FDR was set to 1% on peptide and protein levels. The LiP and tryptic control samples were searched separately. Results were exported using the LiPAnalyzeR_SpectroScheme format.

Data preparation

LiP and Trp quantities were extracted from the exported data from Spectronaut using the columns ‘*PEP.Quantity*’ for peptide and ‘*PG.Quantity*’ for protein quantities. Technical replicates were combined taking the mean of both measurements while omitting NAs. Peptides missing a measurement in one of the samples were removed from any downstream analysis. Peptide and protein quantities were log2-transformed. Batch correction was performed on the data using the *removebatch()* function from the r-package *limma*[33].

For comparing the results of different groups with each other, the data was split into technical and biological replicate groups. For the technical replicates, the two replicates per strain were each assigned to a different group, resulting in two technical replicates groups with eleven biological replicates of each strain in each of the groups (R1, R2). For creating the biological replicates groups (G1, G2), the summarizing technical replicates were divided by batch with G1 consisting of batch 1-5 and G2 batch 6-10.

Limited proteolysis and liquid chromatography-mass spectrometry of human CSF

Human CSF samples were prepared and analyzed as described by Mackmull *et al.*[21]. The results from the performed Spectronaut search were exported using the LiPAnalyzeR_SpectroScheme format.

Data preparation

LiP and Trp quantities were extracted from the exported data from Spectronaut using the columns '*PEP.Quantity*' for peptide and '*PG.Quantity*' for protein quantities. Peptides missing a measurement in more than 20 samples in either the healthy or PD samples were removed from any downstream analysis. Peptide and protein quantities were log2-transformed. Batch correction was performed on the complete dataset using the *removebatch()* function from the r-package *limma*[33]. Subsequently, only the 52 healthy CSF samples were exported and used for analysis in this work.

Computational method of LiPAnalyzeR

The RUV step removes unwanted variation from the LiP peptide matrix $Y_{LiP} = (y_{LiP,ij})$ with peptide i in sample j using a model that describes the contribution of unwanted covariables X to the measured peptide intensity:

$$Y_{LiP} = \beta_0 + \beta_1 X_{Prot} + \beta_2 X_{Trp} + \underline{\beta_m}\,\underline{X} + \epsilon \quad (1)$$

This model estimates the contribution of (whole) protein levels $X_{Prot} = (x_{Prot,ij})$, Trp peptide intensities $X_{Trp} = (x_{Trp\,ij})$ and other unwanted covariates, such as batch effects $\underline{X} = (x_{ijm})$ to the LiP signal variation of each peptide i independently. The matrices X_{Prot}, X_{Trp} and unobserved errors $\epsilon = (\epsilon_{ij})$ have the dimensionality $a \times b$; β_0, $\beta_1$and β_2 are vectors of the length a with $\beta_1$and β_2 being defined as ≥ 0; the matrix $\underline{X} = (x_{ijm})$ which models further variation factors for RUV m is $a \times b \times c$ and $\underline{\beta_m}$ is a matrix of $a \times c$. Here, a = number of peptides, b = number of samples and c = number of further covariables for RUV.

We then estimate all the variation which can be explained by X_{Prot} , X_{Trp} and $\underline{X}$:

$$\tilde{Y}_{LiP} = \beta_0 + \beta_1 X_{Prot} + \beta_2 X_{Trp} + \underline{\beta_m}\,\underline{X} \quad (2)$$

defining the $a \times b$ matrix $\tilde{Y}_{LiP} = (\tilde{y}_{LiP,ij})$ which is subsequently used to estimate the residual matrix $\hat{Y}_{LiP} = (\hat{y}_{LiP,ij})$, which is also $a \times b$.

$$\hat{Y}_{LiP} = Y_{LiP} - \tilde{Y}_{LiP} \quad (3)$$

$\hat{Y}_{LiP}$ now contains structural accessibility variation of interest, since all PK-dependent variation contained in trypsin-only data has been removed, as well as further variation caused by factors such as batch. Accessibility variation for the covariable(s) of interest n can now be inferred in the contrast model

$$\hat{Y}_{LiP} = \beta_{0'} + \underline{\beta_n}\,\underline{Z} \quad (4)$$

where $\underline{Z} = (z_{jin})$ is a $a \times b \times d$ matrix and $\underline{\beta_n}$ is a matrix of $a \times d$, with d = number of variables of interest. $\underline{\beta_n}$ is then evaluated via t statistics, with the degrees of freedom used in the RUV model (Equation 1) being removed from the available degrees of freedom prior to estimating the p-value. Subsequently, we correct for multiple testing effects by computing corrected p-values using the Benjamini–Hochberg approach over all peptides from the same protein ('protein-wise FDR').

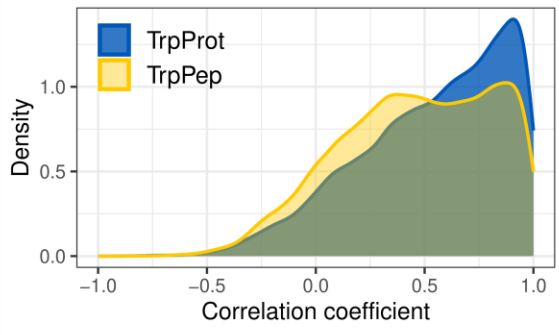
Combining the RUV (Equation 1) and contrast (Equation 2) model into one regression (section *'Superior performance of applying a constrained RUV model prior to the contrast model'*) results in

$$Y_{LiP} = \beta_0 + \beta_1 X_{Prot} + \beta_2 X_{Trp} + \underline{\beta_p}\,\underline{W} + \epsilon \quad (5)$$

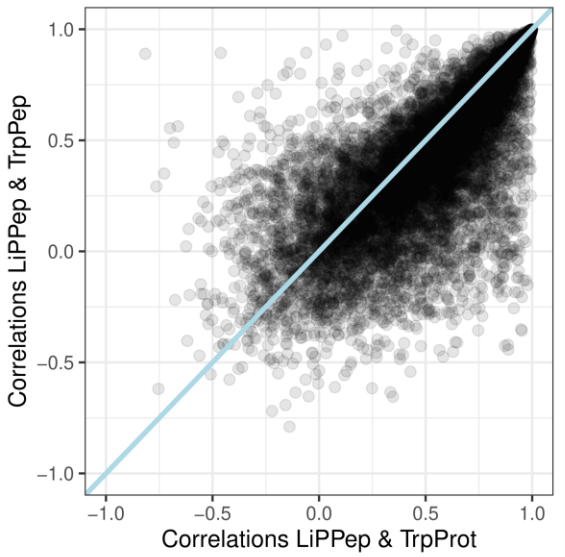
were $\underline{Z} = (z_{jip})$ is a $a \times b \times e$ matrix and $\underline{\beta_p}$ is a matrix of $a \times e$, with $e = c + d$ combining covariates/covariables for RUV and the covariates/covariables of interest into one matrix.

Extended Data Figures

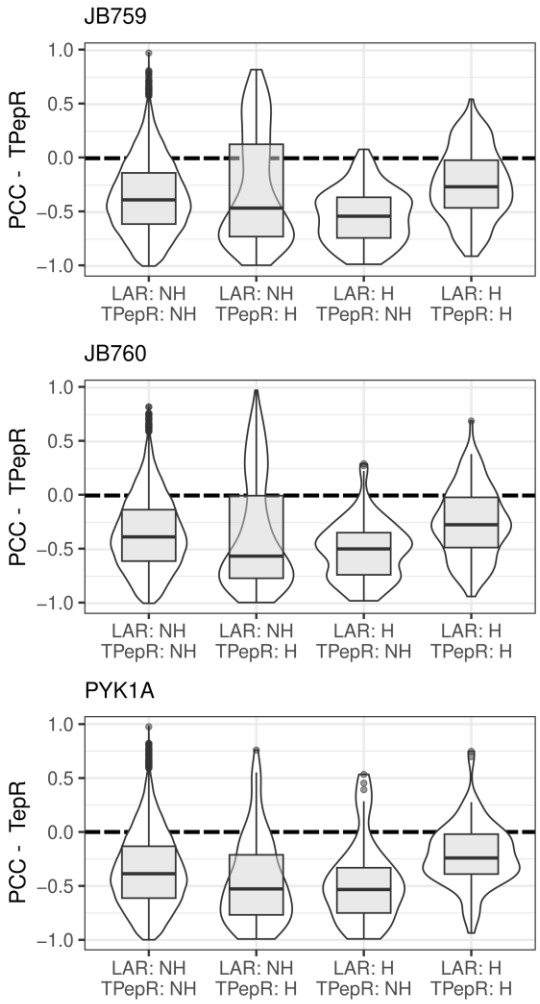
a Peptide-wise correlation of LiPPep with Trp Data (fission yeast)



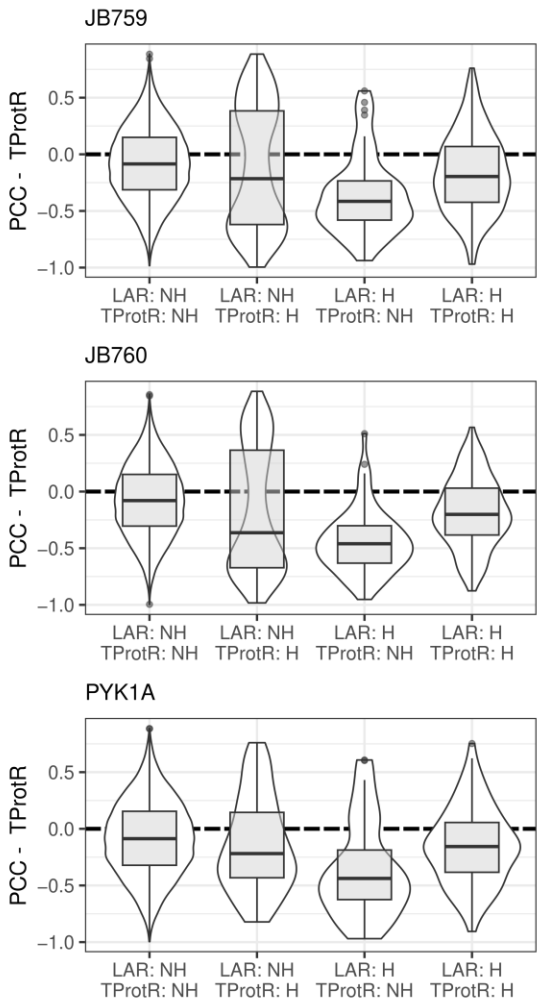
b Peptide-wise correlation coefficients (fission yeast)



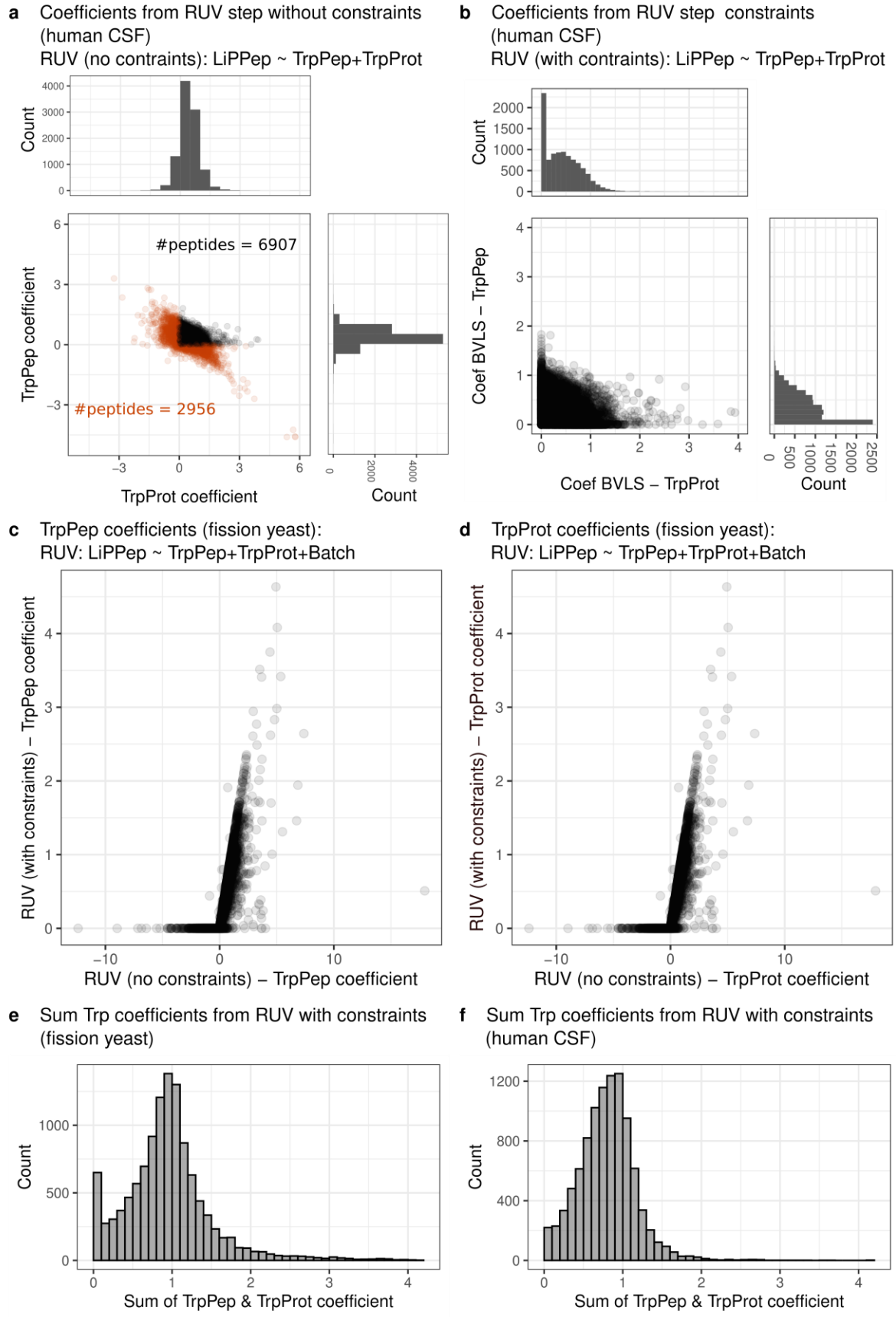
c PPC of TPepR and TrpPep quantity split by Classification of each peptide after TPepR or RUV correction (fission yeast)



d PPC of TProtR and TrpProt quantity split by Classification of each peptide after TProtR or RUV correction (fission yeast)

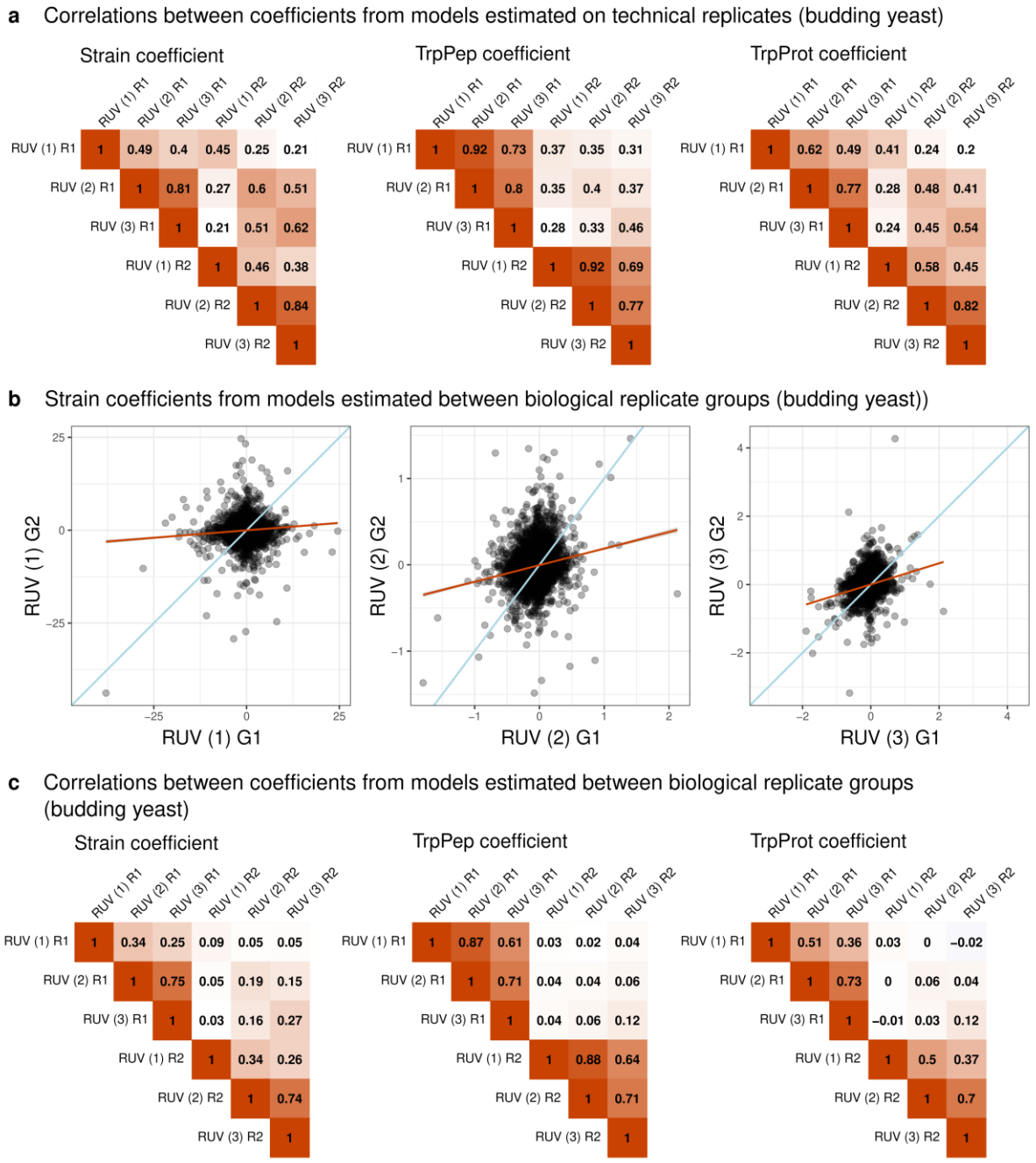


Extended Data Figure 1: Investigation of effects across LiPPep, TrpPep and TrpProt data. a) *Peptide-wise Pearson correlation coefficients of LiPPep quantities with the TrpPep (yellow) and TrpProt (blue) quantities from the fission yeast data.* ***b)*** *Correlation coefficients of a) plotted against each other for each peptide. A line going through the origin with a slope of 1 is added (light blue).* ***c)*** *Pearson's correlation coefficient of the ratio of LiP to Trp peptides (TPepR) to Trp peptide quantities split by their classifications (NH = no hit, H = hit) using the ratio approach (TPepR) or RUV by LiPAnalyzeR (LAR) to correct for Trp quantities in the fission yeast data. Classification was performed by the contrast model of LiPAnalyzeR using JB50 as the reference strain, hence results for all other strains are shown. (For details and interpretation see supplementary text.)* ***d)*** *Same as in c but using the ratio of LiP peptides to Trp proteins (TProtR) instead. (For details and interpretation see supplementary text.)*



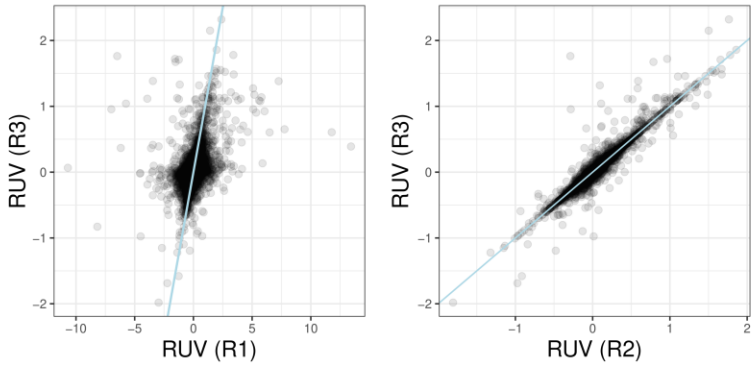
Extended Data Figure 2: Utilizing OLS or BVLS to regress out Trp data effects from LiP peptide quantities. a) *Peptide-specific coefficients for Trp peptide and Trp protein quantities from the RUV step applied without setting constraints to the coefficients to the fission yeast data. Peptides with at least one negative coefficient are*

displayed in orange. ***b)*** *Peptide-specific coefficients for TrpPep and TrpProt from RUV models with constraints (coefficients of TrpPep and TrpProt >=0) in human CSF.* ***c)*** *Peptide-specific coefficients for Trp peptide quantities regressed from the Lip peptide data applying the RUV step without constraints (x-axis) or with constraints (y-axis, Trp coefficients >=0) to the fission yeast data. Coefficients estimated for TrpPep and TrpProt are set to be non-negative in the BVLS models, while no constrictions are set in the OLS models.* ***d)*** *Same as in c) but peptide-specific coefficients for Trp protein are visualized.* ***e)*** *Sum of the Trp peptide and protein coefficients from the RUV step with constraints (Trp coefficients >=0) in the fission yeast data.* ***f)*** *Sum of the Trp peptide and protein coefficients from the RUV step with constraints (Trp coefficients >=0) in the human CSF data.*



Extended Data Figure 3: Analyzing structural accessibility variation with different RUV approaches in budding yeast replicates. a) *Pearson's correlation coefficients of peptide-wise coefficients estimated on the technical replicates of the budding yeast data using different modeling approaches: (1) combining the RUV and contrast step into one model (Equation 5, no constraints to the model), (2) running RUV without constraints and subsequently the contrast model on the resulting residuals (Equations 1-4, no constraints in model 1) and (3) running RUV with constraints and subsequently the contrast model on the resulting residuals (Equations 1-4, constraints in model 1 as described in method section).* ***b)*** *Peptide-wise coefficients for strain effects estimated on the biological replicate groups of budding yeast data using modeling approaches as in a).* ***c)*** *Pearson's correlation coefficients of peptide-wise coefficients estimated on the biological replicate groups of the budding yeast data. Modeling approaches as in a).*

a JB759 strain coefficients from different modelling approaches

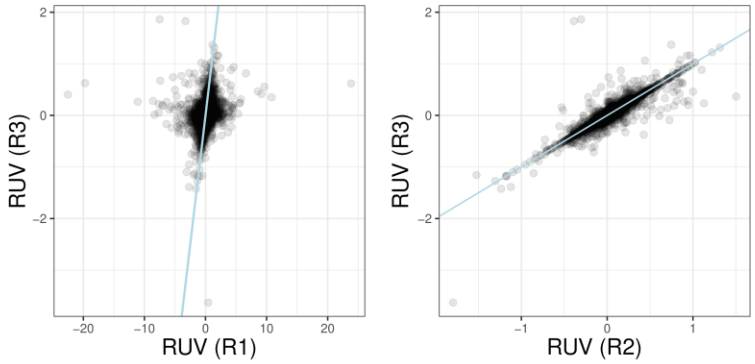


b JB759 strain hits

	RUV (1) Hit	**RUV (1)** No Hit
RUV (3) Hit	117	128
RUV (3) No Hit	448	11037

	RUV (2) Hit	**RUV (2)** No Hit
RUV (3) Hit	211	34
RUV (3) No Hit	22	11463

c JB760 strain coefficients from different modelling approaches

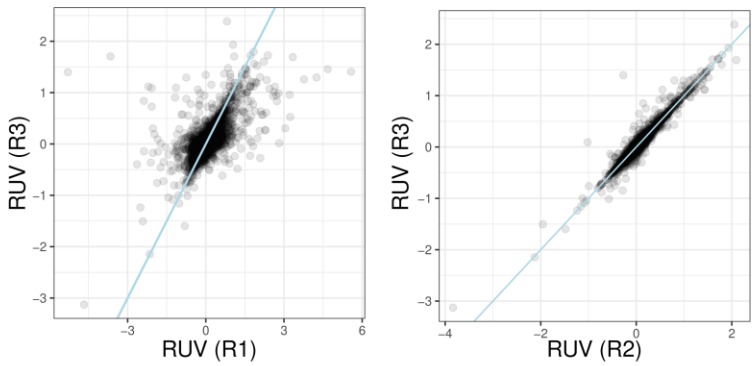


d JB760 strain hits

	RUV (1) Hit	**RUV (1)** No Hit
RUV (3) Hit	145	99
RUV (3) No Hit	516	10970

	RUV (2) Hit	**RUV (2)** No Hit
RUV (3) Hit	217	27
RUV (3) No Hit	20	11466

e PYK1A strain coefficients from different modelling approaches

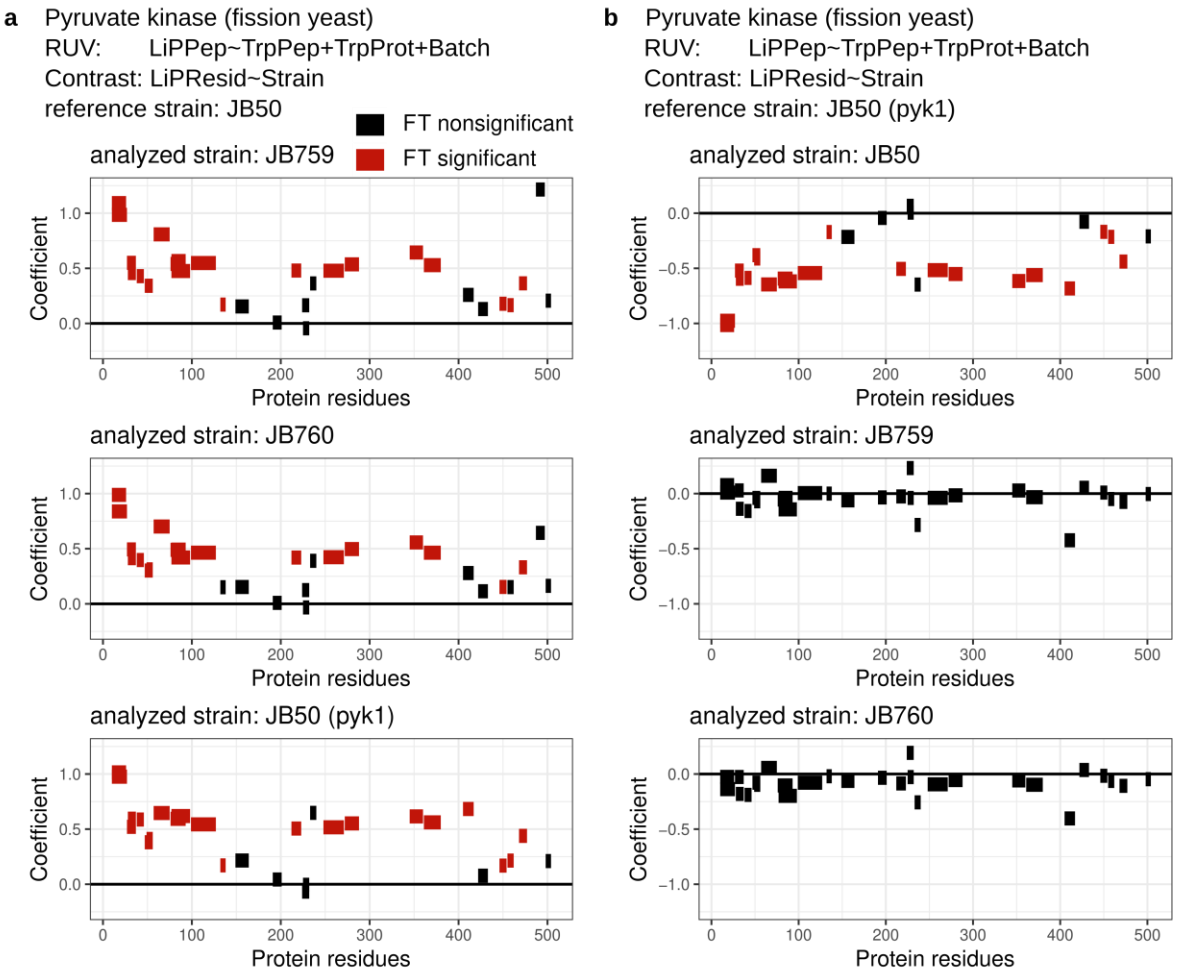


f PYK1A strain hits

	RUV (1) Hit	**RUV (1)** No Hit
RUV (3) Hit	87	49
RUV (3) No Hit	170	11424

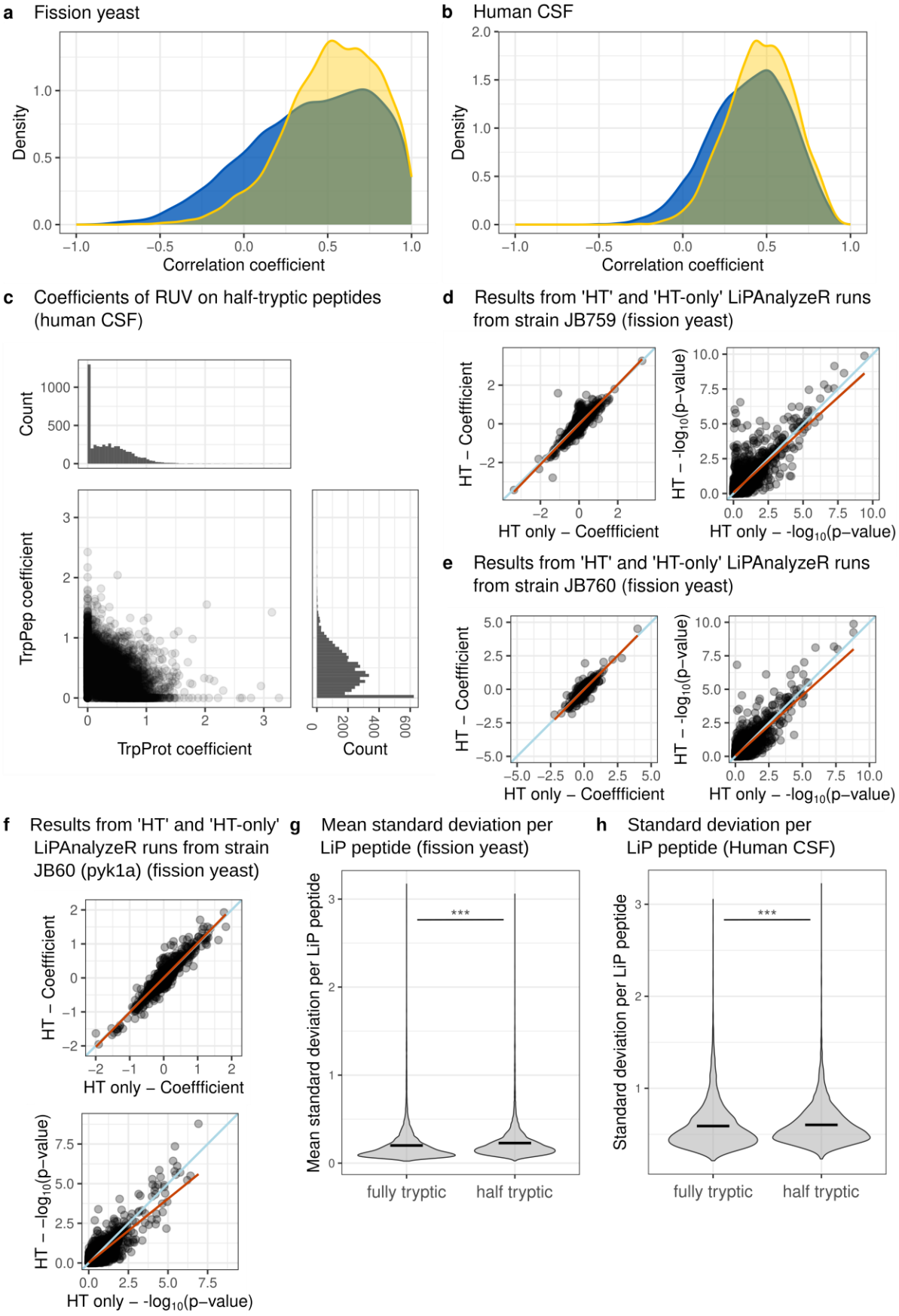
	RUV (2) Hit	**RUV (2)** No Hit
RUV (3) Hit	116	20
RUV (3) No Hit	19	11575

Extended Data Figure 4: Analyzing structural accessibility variation with different RUV approaches in fission yeast replicates a) *Peptide-wise coefficients for the JB759 strain, with JB50 as the reference strain, using different modeling approaches: (1) combining the RUV and contrast step into one model (Equation 5, no constraints to the model), (2) running RUV without constraints and subsequently the contrast model on the resulting residuals (Equations 1-4, no constraints in model 1) and (3) running RUV with constraints and subsequently the contrast model on the resulting residuals (Equations 1-4, constraints in model 1 as described in method section). A line going through the origin with a slope of 1 is added (light blue).* ***b)*** *Number of peptides with a significant p-value for the coefficients representing the JB759 strain after protein-wise FDR when modeling the data with the same models as in a).* ***c)*** *As c) but for the JB760 strain.* ***d)*** *As c) but for the JB760 strain.* ***e)*** *As c) but for the PYK1 mutant.* ***f)*** *As c) but for the PYK1 mutant.*



Extended Data Figure 5: LiPAnalyzeR example results on the protein pyk1 in fission yeast. a) *Peptide-specific coefficients of all peptides from the pyruvate kinase estimated with LiPAnalyzeR plotted along the protein sequence for the JB759 strain (top), JB760 strain (middle) and PYK1 mutant (bottom) in the fission yeast data. LiPAnalyzeR in the default model and JB50 was set as the reference strain. Full tryptic peptides with no significant p-value are visualized in black, if their p-value is significant they are plotted in red. (For details and interpretation see supplementary text.)* ***b)*** *As a) but, for the JB50 strain (top), JB769 strain (middle) and JB760 strain (bottom) in the fission yeast data, setting the PYK1 mutant as the reference strain. (For details and interpretation see supplementary text.)*

Peptide-wise correlation of LiP peptides with TrpData



Extended Data Figure 6: Running LiPAnalyzeR on half-tryptic peptides only. a), b) *Peptide-wise Pearson's correlation coefficients between the LiP peptides and and the highest correlating Trp peptide quantities (yellow) as well as between LiP peptides and the Trp protein quantities (blue) in* ***a)*** *fission yeast and* ***b)*** *human CSF data.* ***c)*** *Peptide-specific coefficients for TrpPep and TrpProt from RUV models with constraints (coefficients of TrpPep and TrpProt >=0) in human CSF.* ***d)*** *Coefficients (left) and corresponding p-values (right) for half-tryptic peptides comparing strain JB759 to JB50 in the contrast model after applying the RUV model using the 'HT' (correcting half-tryptic LiP peptides using only the Trp protein levels) or 'HT-only' (correcting half-tryptic LiP peptides using the most correlating Trp peptide and the Trp protein levels). Linear regressions (orange) and lines going through the origin with a slope of 1 (light blue) are added.* ***e)*** *Same as d) but comparing the JB760 strain to JB50.* ***f)*** *Same as d) but comparing the PYK1 mutate of JB50 to JB50.* ***g)*** *Mean standard deviation of fully and half tryptic peptides in the fission yeast data (Wilcoxon rank sum test: p-value <0.0001 (***)) . The mean standard deviation is depicted as a black line. P-****h)*** *Mean standard deviation of fully and half tryptic peptides in the human CSF data (Wilcoxon rank sum test: p-value <0.0001 (***)). The mean standard deviation is depicted as a black line.*

Supplementary text

Ratio correction approach increases number of false classification of structural changes

To quantify the impact the use of a ratio approach to correct for Trp effects in the LiP data has on the peptides identified as being structurally altered, the ratio approach was applied to the fission yeast data. Instead of applying the RUV models, ratios of the LiP peptides to either the Trp peptides (TPepR) or Trp proteins (TProtR) were estimated and subsequently used as input for the contrast model (Equation 4), setting JB50 as the reference strain for estimating strain effects. The data was also analyzed using the standard LiPAnalyzeR pipeline (Equation 1-4). P-values estimated in the contrast model were FDR corrected protein-wise and a peptide was defined as being structurally altered if it had a corrected p- value below 0.05.

Peptides were classified based on if they were defined as structurally altered (hit) or not structurally altered (no hit) by utilizing a ratio approach (TR) before applying the contrast model or using the default LiPAnalyzeR (LAR) pipeline. This sorts the peptides into four groups for every investigated strain: 1) Hit in TR & LAR, 2) hit in TR & no hit in LAR, 3) no hit in TR & and it LAR and 4) no it in TR & LAR. Pearson's correlation coefficients of the ratio score with the respective Trp data (Figure 4a) were plotted for each of the four groups for all strains in the Trp peptide ratio (Extended Data Figure 1c) and Trp protein ratio (Extended Data Figure 1d) version. The more negative the correlation coefficient is, the more Trp signal was transferred into the ratios and the more we expect our results to be biased by the fact that a ratio approach was applied instead of the RUV models. It can be observed that for peptides where both approaches agree (group 1 & 4) the average correlation of the ratio to the respective Trp data is much closer to zero then for peptides where there is no agreement between the approaches (group 2 & 3). Here, many more peptides have a strong negative correlation coefficient.

Hence, the more Trp signal is added into the ratio by choosing this approach, the more the final results diverge from the results we get when using our RUV model. This analysis clearly shows that using a ratio approach will increase the number of false positive and false negative peptides identified to be altered in the structural PK accessibility in further downstream analysis.

Structural pattern of pyk1 mutation extremely stable across different strains

For the pyruvate kinase (pyk1) we know that it has a single nucleotide polymorphism and associated metabolic changes in JB759, JB760 and the JB60 PYK1 mutant compared to JB50[26].

Analyzing pyk1 with LiPAnalyzeR using JB50 as the reference strain, leads to a large number of peptides with a change in PK accessibility in all of the other strains as we would expect based on prior knowledge (Extended Data Figure 4a). Additionally, the results of this analysis also reveal how conserved these changes are between JB759, JB760 and the PYK1 mutant. A very specific pattern arises when plotting the coefficients of LiPAnalyzeR for every strain for every peptide, with the same peptides being effected into the same directionality across the complete protein. When using the PYK1 mutant as the reference strain, there are no peptides with a significant change in PK accessibility to the JB759 and JB760 strain. In JB50 there are many peptides with a significant change in the structural accessibility with the opposite coefficient pattern estimated for PYK1 using JB50 as the reference. The

example of pyk1 demonstrates the robustness of structural accessibility changes in LiP data inferred by LiPAnalyzeR.

Bibliography

1. Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* **450**, 964–972 (2007).

2. Nussinov, R., Tsai, C.-J. & Jang, H. Protein ensembles link genotype to phenotype. *PLOS Comput. Biol.* **15**, e1006648 (2019).

3. Tzeng, S.-R. & Kalodimos, C. G. Protein activity regulation by conformational entropy. *Nature* **488**, 236–240 (2012).

4. Feng, Y. *et al.* Global analysis of protein structural changes in complex proteomes. *Nat. Biotechnol.* **32**, 1036–1044 (2014).

5. Savitski, M. M. *et al.* Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science* **346**, 1255784 (2014).

6. De Souza, N. & Picotti, P. Mass spectrometry analysis of the structural proteome. *Curr. Opin. Struct. Biol.* **60**, 57–65 (2020).

7. Kaur, U. *et al.* Proteome-Wide Structural Biology: An Emerging Field for the Structural Analysis of Proteins on the Proteomic Scale. *J. Proteome Res.* **17**, 3614–3627 (2018).

8. Schopper, S. *et al.* Measuring protein structural changes on a proteome-wide scale using limited proteolysis-coupled mass spectrometry. *Nat. Protoc.* **12**, 2391–2410 (2017).

9. Malinovska, L. *et al.* Proteome-wide structural changes measured with limited proteolysis-mass spectrometry: an advanced protocol for high-throughput applications. *Nat. Protoc.* **18**, 659–682 (2023).

10. Cappelletti, V. *et al.* Dynamic 3D proteomes reveal protein functional alterations at high resolution in situ. *Cell* **184**, 545-559.e22 (2021).

11. Leuenberger, P. *et al.* Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. *Science* **355**, eaai7825 (2017).

12. Granato, D. C. *et al. Conformational changes in saliva proteome guides discovery of cancer aggressiveness related markers*. http://biorxiv.org/lookup/doi/10.1101/2023.08.04.552034 (2023) doi:10.1101/2023.08.04.552034.

13. Liu, F. & Fitzgerald, M. C. Large-Scale Analysis of Breast Cancer-Related Conformational Changes in Proteins Using Limited Proteolysis. *J. Proteome Res.* **15**, 4666–4674 (2016).

14. Sztacho, M. *et al.* Limited Proteolysis-Coupled Mass Spectrometry Identifies Phosphatidylinositol 4,5-Bisphosphate Effectors in Human Nuclear Proteome. *Cells* **10**, 68 (2021).

15. Backe, S. J. *et al.* PhosY-secretome profiling combined with kinase-substrate interaction screening defines active c-Src-driven extracellular signaling. *Cell Rep.* **42**, 112539 (2023).

16. Khanppnavar, B. *et al. Regulatory sites of CaM-sensitive adenylyl cyclase AC8 revealed by cryo-EM and structural proteomics*. http://biorxiv.org/lookup/doi/10.1101/2023.03.03.531047 (2023) doi:10.1101/2023.03.03.531047.

17. Holfeld, A. *et al.* Systematic identification of structure-specific protein–protein interactions. 2023.02.01.522707 Preprint at https://doi.org/10.1101/2023.02.01.522707 (2023).

18. Paukštytė, J. *et al.* Global analysis of aging-related protein structural changes uncovers enzyme-polymerization-based control of longevity. *Mol. Cell* **83**, 3360-3376.e11 (2023).

19. Sui, X. *et al.* Global proteome metastability response in isogenic animals to missense mutations and polyglutamine expansions in aging. 2022.09.28.509812 Preprint at https://doi.org/10.1101/2022.09.28.509812 (2022).

20. Shuken, S. R. *et al.* Limited proteolysis–mass spectrometry reveals aging-associated changes in cerebrospinal fluid protein abundances and structures. *Nat. Aging* **2**, 379–388 (2022).

21. Mackmull, M.-T. *et al.* Global, in situ analysis of the structural proteome in individuals with Parkinson's disease to identify a new class of biomarker. *Nat. Struct. Mol. Biol.* **29**, 978–989 (2022).

22. Choi, M. *et al.* MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* **30**, 2524–2526 (2014).

23. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of

(prote)omics data. *Nat. Methods* **13**, 731–740 (2016).

24. Röst, H. L. *et al.* OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods* **13**, 741–748 (2016).

25. Carrillo, B., Yanofsky, C., Laboissiere, S., Nadon, R. & Kearney, R. E. Methods for combining peptide intensities to estimate relative protein abundance. *Bioinformatics* **26**, 98–103 (2010).

26. Kamrad, S. *et al.* Pyruvate kinase variant of fission yeast tunes carbon metabolism, cell regulation, growth and stress resistance. *Mol. Syst. Biol.* **16**, e9270 (2020).

27. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science* **296**, 752–755 (2002).

28. Bush, S. J., Chen, L., Tovar-Corona, J. M. & Urrutia, A. O. Alternative splicing and the evolution of phenotypic novelty. *Philos. Trans. R. Soc. B Biol. Sci.* **372**, 20150474 (2017).

29. Foss, E. J. *et al.* Genetic basis of proteome variation in yeast. *Nat. Genet.* **39**, 1369–1375 (2007).

30. Piazza, I. *et al.* A machine learning-based chemoproteomic approach to identify drug targets and binding sites in complex proteomes. *Nat. Commun.* **11**, 4200 (2020).

31. Morretta, E. *et al.* Novel insights on the molecular mechanism of action of the anti-angiogenic pyrazolyl-urea GeGe-3 by functional proteomics. *Bioorganic Chem.* **115**, 105168 (2021).

32. Weith, M. *et al.* Genetic effects on molecular network states explain complex traits. *Mol. Syst. Biol.* **19**, e11493 (2023).

33. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

34. Grossbach, J. *et al.* The impact of genomic variation on protein phosphorylation states and regulatory networks. *Mol. Syst. Biol.* **18**, e10712 (2022).

4 Discussion and Conclusions

LiP-MS presents a novel approach for identifying structural changes under near-native conditions. It enables the detection of conformational changes, protein misfolding, protein aggregation and protein binding, including ligand binding and PPIs, on a proteome-wide scale. LiP-MS data provides information not only about structural changes in the proteome but also about changes in protein abundance, splicing, and PTMs. This allows for multiple proteome measurements to be obtained from a single experiment.

This thesis presents a novel bioinformatic method for inferring structural protein alterations from the first human LiP-MS data set. This approach distinguished structural alterations from PK-independent protein changes, such as protein abundance changes, PTMs or alternative splicing. We applied our pipeline to a large PD cohort and demonstrated, to our knowledge for the first time, that global *in situ* analysis of a body fluid with LiP-MS provides a new type of molecular readout of disease. Multivariate feature selection enabled the construction of models that classify individuals with PD and healthy donors based on signals of structural changes of CSF proteins. These models outperformed those based solely on protein abundance information from the same cohort. A general computational pipeline for the analysis of LiP-MS data was then developed based on this work. After investigating and comparing various approaches that address LiP specific challenges, we implemented a two-step approach that first removes unwanted variations from the LiP signal and then infers the effects of variables on the structural accessibility of proteins. We are presenting a new framework to the scientific community that can be applied to a broad range of LiP-MS experiments. This framework enables for the separation of LiP-MS signals by distinguishing the contributions of structural changes, protein abundance, post-translational modifications, and alternative splicing.

4.1 Future directions of using LiP-MS in PD-related research

LiP-MS was used to analyze the CSF of both healthy individuals and those with PD. The results demonstrate that global *in situ* analysis of a body fluid can provide a new type of molecular readout of a human disease state. Using a novel bioinformatic pipeline, we identified 76 structurally altered proteins in the CSF and showed that protein structural information is more effective than protein abundance information in discriminating between healthy participants and those with PD. Our study shows that LiP-MS has the potential to identify novel structural biomarker candidates for disease on a proteome-wide scale. This can facilitate the generation of hypotheses about underlying disease processes.

4.1.1 Perspectives on biomarker discovery for early diagnosis of Parkinson's Disease

To advance the discovery of proteomic structural biomarkers for early PD diagnosis, it is necessary to validate the identified structural changes. This requires a second CSF cohort with a wider range of PD disease states to ensure sensitivity to different progression levels. The application of LiP-MS to the whole proteome, as performed in this study, can be used to potentially identify additional structural biomarker candidates. However, combining LiP with a more targeted MS application, such as Selected Reaction Monitoring (SRM) or Parallel Reaction Monitoring (PRM), may be more effective in validating the structural alterations identified in our PD study. Both targeted applications select, fragment, and quantify specific precursors[117,118]. In the case of SRM, only specific fragments of these precursors are quantified. These approaches restrict the number of peptides that can be measured in a single run (typically tens to hundreds), but enable absolute quantification[119]. Using these methods, it may be

possible to measure all peptides that cover a region of suspected structural change, even if only a limited number of peptides were detectable in a proteome-wide LiP-MS experiment. This may result in higher sensitivity and resolution, allowing to potentially restrict biomarkers candidates to those that can be used to develop structure-specific antibodies for reliable quantification of the structure in routine diagnostics[120].

In addition, it would be valuable to investigate the molecular cause of a change in PK accessibility in a protein region in PD, determining which differences are due to conformational changes, misfolding, or shielding of surface areas, i.e., caused by PPI or binding of a small molecule. This investigation would require additional experimental approaches to examine the specific structural changes detected by LiP-MS. To investigate differences in PPIs, the first step is to filter those PK accessibility changes that occur in or near known binding interfaces. AlphaFold2 could be used to model a three-dimensional structure for a suspected PPI or protein complex[39]. Then, LiP patterns could be projected onto the predicted protein complex structure. Binding-induced changes in structural conformation may occur in protein regions beyond the binding interface, altering PK-accessibility in peptides distant from the binding interface. A three-dimensional structural representation of altered conformation upon binding can help interpret differences in the PK accessibility observed in all affected protein regions. Pull-down experiments could further confirm an increase in specific protein-protein interactions or protein complexes under specific conditions. Further, cross-linking combined with mass spectrometry (XL-MS) could be used to investigate the source of variation in PK accessibility[121]. To obtain structural information about proteins using MS, a cross-linker, which is a chemical reagent, is inserted between two functional groups in a protein or protein complex. This allows for the modeling of a three-dimensional structure of the protein or protein complex based on the position and spacing of cross-linked amino acids[73]. Standard experimental approaches for direct visualization of altered conformations, misfolding or interactions, such as Cryo-EM[32] or NMR[29], could also be used. However, these methods may currently not be feasible for all structural changes, particularly those that depend on the near-native state and concentration of the protein. Cryo-EM and NMR can be used to study structural rearrangements such as amyloid structure, protein misfolding, and protein aggregation[122].

An important step in biomarker development is to investigate the specificity of the candidate biomarkers. The initial search for structural biomarker candidates of PD was performed on a cohort consisting of healthy individuals and those with PD. Future studies should aim to assess the specificity of these biomarkers. In the case of PD, there are several neurodegenerative diseases that share symptoms and molecular pathological processes with PD, such as dementia with Lewy bodies (DLB), AD or multiple system atrophy (MSA)[123–126]. All of these diseases share neuropathological markers with PD. As our novel candidate biomarkers for PD are defined by quantifying differences between healthy and PD diseased individuals, it is likely, that some of these markers will also be present in one of these similar diseases[127,128]. To ensure the specificity of new biomarkers for PD, it is necessary to include additional cohorts from other diseases.

In complex, multifactorial disorders, such as PD, it is unlikely that a single measurement is unlikely to diagnose or monitor disease progression and response to treatment[129]. Therefore, the overall goal should be to combine multiple data, including protein abundance, structural changes, as well as genetic, imaging and clinical data, for a more diverse, but also more sensitive and specific classification[130–132]. Our study demonstrates that combining the oligomeric-to-total α-Syn ratio with PD-related structural changes significantly improves the accuracy of classifying individuals as healthy

or diseased, compared to using either measure alone. In future work, this should be further expanded to include other data, such as imaging and clinical data.

4.1.2 Predicting progression of Parkinson's Disease through the structural proteome

Structural changes in the proteome contain information that complements other layers of the proteome, such as protein abundance. This provides novel insights into the functional and metabolic state, as well as the interactome, of cells and organisms[75]. This thesis provides evidence that LiP-MS is more effective than traditional molecular approaches, such as analyzing protein abundance alone, in investigating pathological processes in complex disorders, such as PD. LiP-MS not only provides information about the structural changes in the proteome, as well as changes in protein abundance, alternative splicing, and PTMs. This allows for multiple measurements of the proteome to be derived from a single experiment, making it an attractive tool for simultaneously measuring different types of disease-related changes in the proteome.

PD poses many challenges in terms of disease diagnosis, progression and treatment monitoring and LiP-MS is a promising, novel tool to provide complementary information to existing data in this field[133,134]. One of the challenges in PD is measuring and ideally predicting its progression in an individual[135]. The cohort used in this thesis represented early stages of the disease, and therefore, limited variation within disease progression was available. LiP-MS could be applied to a cohort with a wider range of PD progression to infer structural changes in the proteome that correlate with the clinical progression of the disease. Including PD individuals with more motor and dementia-driven subtypes of PD would be of great interest, as it would allow the study of structural changes in relation to different clinical symptoms. These data could also be used to identify previously unknown proteins and pathways associated with PD, as well as to discover new molecular drivers of various clinical phenotypes. In addition, some existing cohorts also include longitudinal samples from the same individuals[136]. Such cohorts allow for tracking of disease progression in the same individual, which is of great interest for understanding the pathological processes of PD over time. The application of LiP-MS in these cohorts could provide novel insights into how molecular processes differ between individuals and develop within an individual. These projects could result in the identification of progression and clinical phenotype markers. Based on such studies, patient-specific monitoring of disease progression could be greatly improved. This would enable clinicians to more accurately assess treatment efficacy, providing individuals with PD the opportunity to benefit from personalized treatment options.

Identifying biomarkers that can predict the progression of Parkinson's disease, including the onset of symptoms and the rate of progression, would significantly enhance the treatment and management of the disease, and improve the quality of life for those affected[133]. To achieve this, a cohort of individuals with early-stage Parkinson's disease, whose future progression is known, would be required. Detecting signals of structural accessibility variations in proteins that are linked with future disease progression could facilitate individualized treatment starting at an early disease stage. Therefore, LiP-MS is a promising tool for conducting such a study. Improving the prediction of the expected progression of cognitive and physical abilities in individuals with PD would provide crucial information for those affected by Parkinson's and their social environment. Anticipating potential increases in health and social care needs and planning for individual requirements would be a valuable achievement in creating an optimal environment for individuals with the disease.

4.1.3 Evaluate the capability of the structural proteome to stratify subtypes of Parkinson's Disease

PD is a disorder that presents with significant heterogeneity in its symptoms, progression, and underlying biological mechanisms among affected individuals[137,138]. As a result, PD is typically classified into subtypes, and there is a great incentive to identify differences between these subtypes to facilitate better management, personalized treatment options, and improved individual well-being.

PD subtyping traditionally relies on the observed phenotype, including specific motor symptoms (e.g., tremor-dominant versus non-tremor PD), axial symptoms, dementia, and genetic causes[139]. However, many studies apply these classifications exclusively, limiting individuals to only one subtype[140–142]. This can be problematic due to the fact that some classifications are based on symptoms, such as specific motor symptoms, while others may be based on the cause of the disease, for example, if a genetic mutation is observed in the individual and their family, and even other are based on the time of onset of disease[143,144]. Therefore, this classification system does not consider that, for example, individuals with early-onset of PD may also exhibit a motor-driven subtype and, as a result, share pathological drivers with those that have a motor-driven subtype with later onset[144,145]. Defining molecular and clinical differences between PD subtypes can be confounded by the fact that the subtype criteria are not mutually exclusive, which may make the subtypes indistinguishable at the level of pathophysiological processes. A subtype can also be based on the presence of a known genetic mutation that can cause PD or increase the likelihood of developing PD[146]. It is unlikely that affected individuals share pathophysiological traits that differentiate them from other individuals with PD, as their clinical phenotypes can also vary significantly[147]. Instead, individuals with a known genetic driver of PD still have PD phenotypes with different motor symptoms, with or without dementia, and thus overlapping neuropathologic drivers with other subtypes[148].

These traditional subtypes usually exhibit poor longitudinal stability as an individual's subtype classification often changes as they experience different or more symptoms over time[149,150]. It is important to note that the underlying neuropathological processes are expected to occur as early as 20 years before the clinical onset of symptoms[151,152]. Therefore, if a person with PD does not show signs of dementia at the time of classification but develops them over time, it can be expected that the pathological processes that cause dementia were present at the time of subtyping[153]. In a study aimed at identifying molecular drivers of dementia in PD, this individual would, however, be classified as part of the PD-without-dementia group. This would dilute the signals caused by dementia-related pathological processes, making it more challenging to identify these processes. There is a increasing incentive in the field to improve subtyping of PD, moving towards data-driven subtyping instead of relying solely on an individual's current symptoms[139,154–157]. Novel approaches to subtyping typically involve following individuals over time and classifying them based on their progression. This may include the rate at which they progress and/or the order in which symptoms occure[155,157]. These classifications, which rely on the progression of the disease, have been shown to be more stable in individuals with the disease[139] and more accurate in predicting clinically relevant milestones, such as dementia, care placement, or death[158,159]. This subtyping may include measures of different α-Syn pathologies and commonly known gene mutations, but it typically does not include other omics measures. However, subtyping approaches require more time as the progression of an individual must be observed over time, making direct personalized treatment based on subtype impossible.

Neuropathologic processes that are not yet reflected in the disease phenotype may be important in defining and understanding different subtypes of PD. To assess this, better quantification of the pathological processes in PD is needed, which may also allow subtyping based on these molecular

changes. LiP-MS is a great technology for measuring these pathological changes on a proteome-wide scale – allowing for the simultaneous quantification of changes that occur at different levels of the proteome. The fact that LiP-MS also detects changes in the interactome is advantageous for studying subtle changes that may not be reflected at the abundance level of the proteome[75].

There is a growing recognition in the scientific community that PD may not be a single disease entity, but should rather be viewed as a spectrum of neurodegenerative disorders with distinct subtypes or variations[160,161]. In addition to the heterogeneity within PD, it has been suggested that other diseases associated with Lewy bodies should also be included in the same classification system, and that the goal should not be to clearly distinguish them from PD. For instance, the DLB Consortium has recently suggested that DLB and Parkinson's disease dementia (PDD) should be viewed as two extremes on the same continuum, rather than as two different diseases[123,126]. There is a growing trend towards conducting molecular investigations of PD and PD-related diseases simultaneously[127]. Cohorts compromising multiple Lewy body diseases are analyzed using omics and imaging techniques, with the goal of identifying both overlapping and divergent molecular mechanisms[162,163]. The use of LiP-MS in these cohorts can provide complementary information by identifying differences in the structural proteome that other omics approaches do not measure. LiP-MS is a promising tool for revealing minor functional changes, e.g., activity changes in specific pathways or protein interactions, that may be critical for disease. Investigating whether LiP-MS can differentiate between different Lewy body diseases or if the structural heterogeneity within PD is greater than its distance to DBL is an intriguing research question.

Investigation of the pathological processes that cause disease and drive disease progression in PD and similar diseases using LiP-MS may provide new insights, revealing overlapping and divergent molecular mechanisms with the potential to positively impact disease progression prediction and disease treatment.

4.2 Perspectives on computational analysis of LiP-MS data

We have investigated different approaches for analyzing LiP-MS data, considering the impact of different biological and technical variables on the overall analysis and results. As a result, we have developed LiPAnalyzeR, a computational pipeline designed to distinguish between structural accessibility changes and PK-independent effects in LiP-MS data.

LiPAnalyzeR is a versatile framework that can be used to analyze a wide range of LiP-MS experiments. The user can define RUV and contrast models according to their data set. For instance, in experiments where different conditions are derived from the same cell lysate sample, additional correction of the LiP signal for protein abundance variation and potential PK-independent effects is not necessary. To determine the structural accessibility variations in fully-tryptic and half-tryptic peptides, LiP intensities can be analyzed directly. For this, the contrast model shapplied on the LiP peptide intensities directly instead of LiP residuals from the RUV models. To remove experimental batch effects or other technical variation, an RUV step can be added. In addition, multiple conditions can be included in the contrast model, e.g., if the goal is to study disease state and age in the same cohort. The framework also allows for modeling interactions between conditions and examining complex relationships among variables to capture their impact on outcomes, providing a more refined understanding of the data.

4.2.1 Expanding the inclusion of half-tryptic peptides in the analysis of structural alterations

Inferring and interpreting changes in PK accessibility from half-tryptic peptides is challenging because these peptides do not exist in the Trp data. This makes it difficult to distinguish between signal caused by structural accessibility variation and signal due to PK-independent variation. LiPAnalyzeR currently uses a conservative approach to analyze half-tryptic peptides. LiPAnalyzeR identifies the fully-tryptic Trp peptide, which originates from the same protein as the analyzed half-tryptic peptide, and has the highest correlation to the half-tryptic LiP peptide. This fully-tryptic Trp peptide is then used in the RUV step along with the Trp protein abundance. This approach represents an improvement over an analysis that only corrects for changes in protein abundance. However, the analysis of half-tryptic peptides still has a higher chance of false classifications compared to the analysis of fully-tryptic peptides.

The current implementation of LiPAnalyzeR for half-tryptic peptides may miss structural changes indicated by the signal of half-tryptic peptides resulting from PK-independent effects, such as a conformational shift driven by a distant PTM in the same protein. If the PTM signal is reflected in one of the fully-tryptic peptides in the Trp data, that fully-tryptic peptide may be the highest correlating fully-tryptic peptide with the half-tryptic peptide. The RUV model would include this fully-tryptic peptide to correct the intensity of the half-tryptic LiP peptide for the PTM effect. The signal in the half-tryptic peptides induced by the structural shift due to the PTM would, however, be consistent with the PTM signal in the fully tryptic peptide and therefore be removed by the RUV model. The true structural signal in the half-tryptic peptide will then not be present in the LiP residuals and cannot be detected in the contrast model. A false negative finding would result from missing this structural change. Furthermore, a half-tryptic peptide may be affected by a PK-independent peptide effect that cannot be measured in any of the fully-tryptic Trp peptides. For example, a half-tryptic peptide may contain a post-translational modification effect, but if no corresponding fully-tryptic peptide was measured, the RUV model cannot account for this variation, leading to a falsely inferred structural change.

Restricting the fully-tryptic peptides, which are candidates for being included in the RUV step, to those that overlap with the analyzed half-tryptic peptides could potentially improve the analysis of half-tryptic peptides. This would remove half-tryptic peptides from the analysis for which no matching fully tryptic peptides were identified, but could optimize the peptide matching for the remaining half-tryptic peptides. Thus, only local PK-independent peptide effects would be removed from the signal of the half-tryptic peptide. This would prevent scenarios such as the one described above, where a distant PTM effect could be used to remove the structural signal from a half-tryptic peptide during the RUV step. This alternative approach could also reduce the probability of PTM effects being exclusive to a half-tryptic peptide, since the half-tryptic peptide would need to overlap at least partially with the fully-tryptic peptide. The half-tryptic peptide can contain missed trypsin cleavage sites, which means it may not completely overlap with the fully tryptic peptide, leaving a small chance that a PTM located in a region of the half-tryptic peptide is not covered by the fully-tryptic peptide. To address this issue, one potential solution is to require that the candidate fully-tryptic peptide covers the entire sequence of the half-tryptic peptide. However, this approach may result in a reduction of the number of half-tryptic peptides. It is recommended to compare these alternative approaches to the current implementation with careful evaluation.

In the future, it may be beneficial to consider alternative methods for analyzing half-tryptic peptides within the same data set. It could be useful to apply different approaches within the same data set, depending on the availability of matching fully-tryptic peptides. Annotations should be added to the

results of each half-tryptic peptide. These annotations would inform the user whether LiPAnalyzeR has removed PK-independent peptide effects and is confident that only changes due to changes in PK accessibility are quantified. If LiPAnalyzeR cannot distinguish between PK-independent and PK-dependent peptide effects in a particular half-tryptic peptide, it will be noted. The user can then interpret the results of LiPAnanlyzeR considering the larger biological picture.

4.2.2 Characterizing changes in intrinsic disordered protein regions with LiP-MS

The majority of the proteome exists in stable three-dimensional structures, but a significant portion is structurally heterogeneous, known as intrinsically disordered proteins (IDPs) and intrinsically disordered protein regions (IDRs)[164–166]. These regions are found across all kingdoms of life and are particularly prevalent in eukaryotes, where they occur in approximately 33% of the proteins[165]. Despite, or rather because of, the lack of a defined structural conformation, IDRs are critical for many cellular processes, including signal transduction, protein biosynthesis, cell division, circadian biology and cellular homeostasis[166–171], as they can adapt many different structures[172]. They significantly contribute to macromolecular interactions through their conformational plasticity and adaptability, typically interacting with other biomolecules, serving as, for example, flexible linkers, binding interfaces for simultaneous interactions with multiple partners or cellular sensors[166,173–175]. Therefore, the study of IDRs is of great interest. However, their lack of a distinct three-dimensional conformation poses challenges for many approaches such as X-ray crystallography[164] or AlphaFold[37,41].

LiP-MS data can be used to evaluate changes in interaction for IDRs depending on their binding mechanism. There are three primary mechanisms by which IDRs bind to other biomolecules: (1) coupled folding-upon-binding, in which IDRs adopt a specific structural conformation upon binding to a specific biomolecule[176], (2) fuzzy binding or fuzzy complexes, in which IDRs exist in a finite number of multiple bound-state conformations within a protein complex[177], and (3) disordered complex formation, in which IDRs exist in a completely disordered bound-state complex and both binding partners involved remain disordered[166]. If an IDR is easily accessible in its unbound state, such as when it is located in a loop region on the surface of a protein, and becomes buried upon binding, making it less accessible for PK, this change will be detectable in the LiP-MS data using the standard LiP-AnalyzeR pipeline. However, if there is no overall difference in the accessibility when an IDR binds, new evaluation approaches are needed. In these cases, it is reasonable to hypothesize that binding events associated with type (1) can be detected in LiP data, potentially this is also possible for bindings of type (2). However, for type bindings (3), since the disordered state of the IDR is independent of its binding state, it is more likely that no difference in PK digest occurs, making it undetectable through LiP-MS.

If there is no overall difference in accessibility of an IDR between conditions, one approach to further investigate this IDR is to compare the half-tryptic digest patterns of these conditions. Assuming an example where the IDR of interest is mostly unbound and disordered in condition 1, the IDR would have a slightly different structural conformation at the moment of PK digest in each protein molecule in a sample. Therefore, the PK digest of the IDR would be very heterogeneous, resulting in a large variety of half-tryptic peptides present in every individual sample. If the same IDR is mostly folded upon binding in condition 2, PK cleavages would be much more conserved between different molecules of the same protein in each condition due to the identical three-dimensional conformation. Therefore, it is expected that half-tryptic peptides resulting from the PK digest will be less diverse, with certain cleavages occurring more frequently than others. These differences in the occurrence of

half-tryptic peptides in an IDR or IDP could be used to indicate whether it is largely disordered or bound upon binding in a given condition. This analysis may also be successful in fuzzy binding states, where a finite number of structural states exist. However, a higher number of structural alternating states may be indistinguishable from a disordered state.

An alternative approach is based on the idea that some IDRs may be primarily involved in specific binding events that differ between samples in the absence of a strong external signal. For example, in an experimental setup designed to study structural aspects of the stress response in yeast, an external stress stimulus is applied to condition 2. For this example, we assume that the IDR of interest functions as a stress sensor, a known phenomenon[175,178]. The IDR of interest responds uniformly to the same stressor and stress level in all condition 2 samples, resulting in a consistent structural conformation and a conserved pattern of PK accessibility. In contrast, samples from condition 1 are not exposed to the external stressor, but rather experience a low level of stress specific to each colony (i.e., samples). This stress is not conserved between samples, but is unique to each one. Therefore, the same IDR may have different three-dimensional shapes not within but between samples of condition 1. Each sample in condition 1 has its individual PK accessibility pattern in this specific region. However, the mean accessibility of the IDR in condition 1 may be similar to that of the other condition, rendering the standard analysis of LiP-MS data. To detect these alterations in the accessibility of the IDR in these samples, the variation of the PK-accessibility within each condition should be estimated and subsequently compared in between conditions. This can be achieved by using Levene's test[179] to determine whether the variance of a structural signal differs between conditions. To ensure accurate results, it is crucial to differentiate between variance caused by structural signal and variance caused by PK-independent effects. Therefore, it is recommended to conduct such an analysis on LiP residuals obtained from the RUV model of LiPAnalyzeR.

4.2.3 Combining individual signals to reflect intra and interprotein structural changes

LiP-MS data analysis, such as with LiPAnalyzeR, is typically conducted on peptide level, resulting in peptide-specific LiP residuals (resulting from the RUV models) or structural scores (resulting from the contrast models). However, changes in protein structure may not be limited to a single peptide, as they may affect multiple peptides of the structurally altered protein. Additionally, changes in PPIs and protein complexes can affect the accessibility of peptides derived from multiple involved proteins simultaneously. Integrating peptide-specific PK accessibility scores from the same protein or protein complex and drawing conclusions about structural changes affecting multiple peptides is currently the responsibility of the user. Implementing an approach that identifies such structural events affecting multiple peptides and reports them would increase the interpretability of the results.

Identifying major structural rearrangements in proteins cannot be accomplished by solely examining neighboring peptides in the protein sequence, as peptides that are in close proximity to each other in three-dimensional space may not be adjacent in the AA sequence[180]. One promising approach is to cluster LiP signals from the same protein or protein complex based on whether they behave similarly. This can help identify structural changes that occur simultaneously, as they may reflect the same structural event. Similar pipelines have been successfully applied to identify subcomplexes in proteins or to detect functional proteoform groups[113,181,182]. When attempting to combine structural signals, it is crucial to use the LiP signal that reflects variations in PK accessibility. To ensure that the estimated clusters are not driven by protein abundance, PK-independent peptide variation, or technical artifacts,

the residuals resulting from the RUV step of LiPAnalyzeR are the optimal choice for this type of analysis.

To identify peptides in a protein that are affected by the same structural event or to identify structural events that occur simultaneously, the first step is to quantify which peptides change their accessibility at the same time. This can be achieved by performing pairwise correlation of peptides belonging to the same protein across all samples. In this step, it is crucial that the data contains a sufficient amount of variation for these correlations to be meaningful. Performing this step on data with binary structural variation, e.g., comparing yeast cultures exposed to only two different perturbations, may not yield promising results. Instead, data with a wide range of changes, ideally continuous variation, such as a large cohort of PD patients at different disease stages, would be optimal for obtaining these correlation values. To ensure accurate clustering, it is recommended to use absolute correlation values when assessing PK-accessibility. This is because the same structural event may increase PK-accessibility in one peptide while decreasing it in another. Peptides with high correlation coefficients can then be combined into a cluster for further analysis. These individual clusters contain different peptides from the same protein that reflect either a single or multiple structural events that occur simultaneously. An example of multiple, simultaneous events is the binding of a small molecule that induces a conformational switch in another region of the protein. This can be viewed as two structural events, where one is dependent on the other. To visualize these structural events occurring in different peptides, they could be mapped onto the three-dimensional structure of the protein, highlighting the regions of change in three-dimensional space. This approach could also be used to summarize structural changes in PPIs and protein complexes by analyzing all peptides from proteins involved in the interaction together, rather than just those from a single protein.

This approach would allow the integration of the structural LiP signal from fully-tryptic and half-tryptic peptides by clustering LiP signals that arise from the same structural event into a single signal, regardless of the peptide type they were measured in. The reduction in feature space also increases statistical power by minimizing the number of events in multiple testing correction, potentially resulting in the detection of more distinct structural events.

4.2.4 Estimating indices of protein structural events to understand dynamic functional states

It would be of great interest to summarize the structural state of each protein within a single index. This would enable tracking changes in the structure of a protein as a whole, for example, under continuous conditions, such as aging, without inferring structural rearrangements for each individual peptide or clustered peptide group. A simple example of this is estimating an index for a protein that aggregates during aging. This aggregation can be quantified from individual peptides or groups of peptides that lose or gain PK accessibility due to this shift. However, an index that represents the overall level of aggregation in each individual would be useful to visualize and monitor the aggregation progression in all individuals of a cohort.

This could be addressed by performing a dimensional reduction on the structural LiP signal, i.e., the LiP residual estimated in LiPAnalyzeR, excluding PK-independent variations. Since the goal is to estimate a structural index specific to each protein, the dimensional reduction would be applied to all peptides of the same protein simultaneously. To accomplish this, a matrix containing the LiP residuals of all peptides of a specific protein (i.e., features) over all samples should be created and used as the input for the dimensional reduction algorithm. The dimensional reduction would result in a one-dimensional or multidimensional index (e.g., the loadings for one or several principal components) for

each sample, describing the structural state of the protein of interest. In the example described above, where the target signal is the increase in aggregation over time, a one-dimensional index should be sufficient to describe this process and is easy to interpret. However, for more complex structural changes in a protein, two or more dimensional indices may be necessary to accurately represent them.

An example of such a scenario is an enzyme that occurs in both an inactive and an active state. These states differ significantly in their three-dimensional conformation, and can therefore be quantified using LiP-MS. Additionally, the enzyme has a phosphorylation site, and the phosphorylation causes a small structural shift in the protein. Thus, there are two possible conformational changes that can occur in the same protein, resulting in four combinations (1) inactive without a phosphorylation-induced shift, (2) inactive with a phosphorylation-induced shift, (3) active without a phosphorylation-induced shift, and (4) active with a phosphorylation-induced shift. If the conformational shift induced by the phosphorylation additionally leads to a higher probability that the enzyme is present in the active conformation, although a protein can also be phosphorylated and inactive, as observed in kinases[183], these two types of structural conformations (inactive vs. active and phosphorylated vs. non-phosphorylated) are correlated but not entirely dependent on each other. Representing them in a one-dimensional index would obscure this effect. However, a two-dimensional index that shows the shift from active to inactive state in the first dimension and the phospho-induced shift in the second dimension would allow observing this interplay. In cases with further conformational states, the dimensions of the index may need to be even higher.

Estimating one-dimensional indices is straightforward because it does not require any protein-specific decisions about the number of dimensions of the index, and the interpretation is simple. However, they may not be the optimal choice for more complex structural changes that are frequently present in the proteome. Multidimensional indices may be more effective in representing various structural events that occur in a protein, but are more challenging to interpret and may require further processing, such as grouping samples based on this structural protein index. This approach could also be used to study protein complexes by performing dimensional reduction on all peptides corresponding to proteins involved in the complex instead of just peptides from a single protein. Furthermore, additional dimensions representing changes in protein abundance or in PK-independent peptide effects could be added to the index. This would create a single, multidimensional index that reports the overall changes in a specific protein.

4.3 Outlook

LiP-MS facilitates the characterization of structural alterations in proteins under near-native conditions on a proteome-wide scale, additionally providing information on changes in protein abundance, PTMs, and alternative splicing. This work introduces a versatile computational framework, designed to distinguish between structural accessibility changes and PK-independent effects in LiP-MS data. This innovative framework establishes a standardized and reliable approach for the comprehensive analysis of LiP-MS data. It leverages the full potential of the data, thus significantly contributing to the accessibility and usability of this technique within the scientific community.

Additional extensions to LiPAnalyzeR will further increase the interpretability of results on a protein and functional level, as well as enable the detection of structural events that are currently not reliably detected, such as changes in IDRs. Summarizing individual changes into structural events or creating protein-specific indices will increase the signal-to-noise ratio, and decrease the size of the feature space. This will allow for a higher sensitivity and an increase in statistical power as the number of events in multiple testing corrections are minimized.

Bibliography

1. Harper, J. W. & Bennett, E. J. Proteome complexity and the forces that drive proteome imbalance. *Nature* **537**, 328–338 (2016).
2. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13**, 227–232 (2012).
3. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535–550 (2016).
4. Bludau, I. & Aebersold, R. Proteomic and interactomic insights into the molecular basis of cell functional diversity. *Nat. Rev. Mol. Cell Biol.* **21**, 327–340 (2020).
5. Pandey, A. & Mann, M. Proteomics to study genes and genomes. *Nature* **405**, 837–846 (2000).
6. Hartl, F. U., Bracher, A. & Hayer-Hartl, M. Molecular chaperones in protein folding and proteostasis. *Nature* **475**, 324–332 (2011).
7. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
8. Li, J. J., Bickel, P. J. & Biggin, M. D. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* **2**, e270 (2014).
9. Signor, S. A. & Nuzhdin, S. V. The Evolution of Gene Expression in cis and trans. *Trends Genet.* **34**, 532–544 (2018).
10. Novoa, E. M. & Ribas De Pouplana, L. Speeding with control: codon usage, tRNAs, and ribosomes. *Trends Genet.* **28**, 574–581 (2012).
11. Hershko, A. & Ciechanover, A. THE UBIQUITIN SYSTEM. *Annu. Rev. Biochem.* **67**, 425–479 (1998).
12. Wilhelm, B. T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239–1243 (2008).
13. Tian, B. & Manley, J. L. Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.* **18**, 18–30 (2017).
14. Lopez, A. J. ALTERNATIVE SPLICING OF PRE-mRNA: Developmental Consequences and Mechanisms of Regulation. *Annu. Rev. Genet.* **32**, 279–305 (1998).
15. Buljan, M. *et al.* Tissue-Specific Splicing of Disordered Segments that Embed Binding Motifs Rewires Protein Interaction Networks. *Mol. Cell* **46**, 871–883 (2012).
16. Brar, G. A. Beyond the Triplet Code: Context Cues Transform Translation. *Cell* **167**, 1681–1692 (2016).
17. Touriol, C. *et al.* Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons. *Biol. Cell* **95**, 169–178 (2003).
18. Walsh, C. T., Garneau-Tsodikova, S. & Gatto, G. J. Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew. Chem. Int. Ed Engl.* **44**, 7342–7372 (2005).
19. Rogers, L. D. & Overall, C. M. Proteolytic Post-translational Modification of Proteins: Proteomic Tools and Methodology *. *Mol. Cell. Proteomics* **12**, 3532–3542 (2013).
20. Seo, J.-W. & Lee, K.-J. Post-translational Modifications and Their Biological Functions: Proteomic

Analysis and Systematic Approaches. *BMB Rep.* **37**, 35–44 (2004).

21. Grant, B. J., Gorfe, A. A. & McCammon, J. A. Large conformational changes in proteins: signaling and other functions. *Curr. Opin. Struct. Biol.* **20**, 142–147 (2010).

22. Balchin, D., Hayer-Hartl, M. & Hartl, F. U. In vivo aspects of protein folding and quality control. *Science* **353**, aac4354 (2016).

23. Alberts, B. *et al.* Analyzing Protein Structure and Function. in *Molecular Biology of the Cell. 4th edition* (Garland Science, 2002).

24. Ilari, A. & Savino, C. Protein Structure Determination by X-Ray Crystallography. in *Bioinformatics: Data, Sequence Analysis and Evolution* (ed. Keith, J. M.) 63–87 (Humana Press, 2008). doi:10.1007/978-1-60327-159-2_3.

25. Abola, E., Kuhn, P., Earnest, T. & Stevens, R. C. Automation of X-ray crystallography. *Nat. Struct. Biol.* **7**, 973–977 (2000).

26. Ballone, A., Lau, R. A., Zweipfenning, F. P. A. & Ottmann, C. A new soaking procedure for X-ray crystallographic structural determination of protein–peptide complexes. *Acta Crystallogr. Sect. F Struct. Biol. Commun.* **76**, 501–507 (2020).

27. Kleckner, I. R. & Foster, M. P. An introduction to NMR-based approaches for measuring protein dynamics. *Biochim. Biophys. Acta BBA - Proteins Proteomics* **1814**, 942–968 (2011).

28. Takeuchi, K., Baskaran, K. & Arthanari, H. Structure determination using solution NMR: Is it worth the effort? *J. Magn. Reson.* **306**, 195–201 (2019).

29. Puthenveetil, R. & Vinogradova, O. Solution NMR: A powerful tool for structural and functional studies of membrane proteins in reconstituted environments. *J. Biol. Chem.* **294**, 15914–15931 (2019).

30. Hu, Y. *et al.* NMR-Based Methods for Protein Analysis. *Anal. Chem.* **93**, 1866–1879 (2021).

31. Klukowski, P., Riek, R. & Güntert, P. Rapid protein assignments and structures from raw NMR spectra with the deep learning technique ARTINA. *Nat. Commun.* **13**, 6151 (2022).

32. Bai, X., McMullan, G. & Scheres, S. H. W. How cryo-EM is revolutionizing structural biology. *Trends Biochem. Sci.* **40**, 49–57 (2015).

33. Egelman, E. H. The Current Revolution in Cryo-EM. *Biophys. J.* **110**, 1008–1012 (2016).

34. DiIorio, M. C. & Kulczyk, A. W. Exploring the Structural Variability of Dynamic Biological Complexes by Single-Particle Cryo-Electron Microscopy. *Micromachines* **14**, 118 (2022).

35. Renaud, J.-P. *et al.* Cryo-EM in drug discovery: achievements, limitations and prospects. *Nat. Rev. Drug Discov.* **17**, 471–492 (2018).

36. Kuhlman, B. & Bradley, P. Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.* **20**, 681–697 (2019).

37. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583 589 (2021).

38. Perrakis, A. & Sixma, T. K. AI revolutions in biology. *EMBO Rep.* **22**, e54046 (2021).

39. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. *Nat. Commun.* **13**, 1265 (2022).

40. Bryant, P. *et al.* Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. *Nat. Commun.* **13**, 6028 (2022).

41. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).

42. Jendrusch, M., Korbel, J. O. & Sadiq, S. K. *AlphaDesign: A* de novo *protein design framework based on AlphaFold*. http://biorxiv.org/lookup/doi/10.1101/2021.10.11.463937 (2021) doi:10.1101/2021.10.11.463937.

43. Schopper, S. *et al.* Measuring protein structural changes on a proteome-wide scale using limited proteolysis-coupled mass spectrometry. *Nat. Protoc.* **12**, 2391–2410 (2017).

44. Wrabl, J. O. *et al.* The role of protein conformational fluctuations in allostery, function, and revolution. *Biophys. Chem.* **159**, 129–141 (2011).

45. Feng, Y. *et al.* Global analysis of protein structural changes in complex proteomes. *Nat. Biotechnol.* **32**, 1036–1044 (2014).

46. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).

47. Nørregaard Jensen, O. Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Curr. Opin. Chem. Biol.* **8**, 33–41 (2004).

48. Smits, A. H. & Vermeulen, M. Characterizing Protein–Protein Interactions Using Mass Spectrometry: Challenges and Opportunities. *Trends Biotechnol.* **34**, 825–834 (2016).

49. Malinovska, L. *et al.* Proteome-wide structural changes measured with limited proteolysis-mass spectrometry: an advanced protocol for high-throughput applications. *Nat. Protoc.* **18**, 659–682 (2023).

50. Fontana, A. *et al.* Probing protein structure by limited proteolysis. *Acta Biochim. Pol.* **51**, 299–321 (2004).

51. Carrillo, B., Yanofsky, C., Laboissiere, S., Nadon, R. & Kearney, R. E. Methods for combining peptide intensities to estimate relative protein abundance. *Bioinformatics* **26**, 98–103 (2010).

52. Savitski, M. M. *et al.* Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science* **346**, 1255784 (2014).

53. Mateus, A., Määttä, T. A. & Savitski, M. M. Thermal proteome profiling: unbiased assessment of protein state through heat-induced stability changes. *Proteome Sci.* **15**, 13 (2017).

54. Tan, C. S. H. *et al.* Thermal proximity coaggregation for system-wide profiling of protein complex dynamics in cells. *Science* **359**, 1170–1177 (2018).

55. Mateus, A. *et al.* Thermal proteome profiling in bacteria: probing protein state in vivo. *Mol. Syst. Biol.* **14**, e8242 (2018).

56. Le Sueur, C., Hammarén, H. M., Sridharan, S. & Savitski, M. M. Thermal proteome profiling: Insights into protein modifications, associations, and functions. *Curr. Opin. Chem. Biol.* **71**, 102225 (2022).

57. Mateus, A. *et al.* Thermal proteome profiling for interrogating protein interactions. *Mol. Syst. Biol.* **16**, e9232 (2020).

58. Peck Justice, S. A. *et al.* Mutant thermal proteome profiling for characterization of missense

protein variants and their associated phenotypes within the proteome. *J. Biol. Chem.* **295**, 16219–16238 (2020).

59. Piazza, I. *et al.* A machine learning-based chemoproteomic approach to identify drug targets and binding sites in complex proteomes. *Nat. Commun.* **11**, 4200 (2020).

60. De Souza, N. & Picotti, P. Mass spectrometry analysis of the structural proteome. *Curr. Opin. Struct. Biol.* **60**, 57–65 (2020).

61. Frey, L. *et al. On the pH-dependence of α-synuclein amyloid polymorphism and the role of secondary nucleation in seed-based amyloid propagation*. http://biorxiv.org/lookup/doi/10.1101/2023.06.25.546428 (2023) doi:10.1101/2023.06.25.546428.

62. Basalla, J. L. *et al.* Dissecting the phase separation and oligomerization activities of the carboxysome positioning protein McdB. *eLife* **12**, e81362 (2023).

63. Zhang, Z. & Shah, B. Limited Proteolysis Coupled with Mass Spectrometry for Simultaneous Evaluation of a Large Number of Protein Variants for Their Impact on Conformational Stability. *Anal. Chem.* **93**, 14263–14271 (2021).

64. Meng, H. *et al.* Discovery of a cooperative mode of inhibiting RIPK1 kinase. *Cell Discov.* **7**, 1–18 (2021).

65. Sarker, H. *et al.* Glucocorticoids Bind to SARS-CoV-2 S1 at Multiple Sites Causing Cooperative Inhibition of SARS-CoV-2 S1 Interaction With ACE2. *Front. Immunol.* **13**, 906687 (2022).

66. Sun, W. *et al.* Small molecule activators of TAK1 promotes its activity-dependent ubiquitination and TRAIL-mediated tumor cell death. *Proc. Natl. Acad. Sci.* **120**, e2308079120 (2023).

67. Khanppnavar, B. *et al. Regulatory sites of CaM-sensitive adenylyl cyclase AC8 revealed by cryo-EM and structural proteomics*. http://biorxiv.org/lookup/doi/10.1101/2023.03.03.531047 (2023) doi:10.1101/2023.03.03.531047.

68. Samanta, A., Kiselar, J., Pumroy, R. A., Han, S. & Moiseenkova-Bell, V. Y. Structural insights into the molecular mechanism of mouse TRPA1 activation and inhibition. *J. Gen. Physiol.* **150**, 751–762 (2018).

69. Sztacho, M. *et al.* Limited Proteolysis-Coupled Mass Spectrometry Identifies Phosphatidylinositol 4,5-Bisphosphate Effectors in Human Nuclear Proteome. *Cells* **10**, 68 (2021).

70. Holfeld, A. *et al. Systematic identification of structure-specific protein–protein interactions*. http://biorxiv.org/lookup/doi/10.1101/2023.02.01.522707 (2023) doi:10.1101/2023.02.01.522707.

71. Morretta, E. *et al.* Novel insights on the molecular mechanism of action of the anti-angiogenic pyrazolyl-urea GeGe-3 by functional proteomics. *Bioorganic Chem.* **115**, 105168 (2021).

72. Piazza, I. *et al.* A Map of Protein-Metabolite Interactions Reveals Principles of Chemical Communication. *Cell* **172**, 358-372.e23 (2018).

73. Piersimoni, L., Kastritis, P. L., Arlt, C. & Sinz, A. Cross-Linking Mass Spectrometry for Investigating Protein Conformations and Protein–Protein Interactions–A Method for All Seasons. *Chem. Rev.* **122**, 7500–7531 (2022).

74. Zampieri, M. *et al.* High-throughput metabolomic analysis predicts mode of action of uncharacterized antimicrobial compounds. *Sci. Transl. Med.* **10**, eaal3973 (2018).

75. Cappelletti, V. *et al.* Dynamic 3D proteomes reveal protein functional alterations at high resolution in situ. *Cell* **184**, 545-559.e22 (2021).

76. Mehta, V. *et al.* Structure of Mycobacterium tuberculosis Cya, an evolutionary ancestor of the mammalian membrane adenylyl cyclases. *eLife* **11**, e77032 (2022).

77. Hartl, F. U. Protein Misfolding Diseases. (2017).

78. Liu, F. & Fitzgerald, M. C. Large-Scale Analysis of Breast Cancer-Related Conformational Changes in Proteins Using Limited Proteolysis. *J. Proteome Res.* **15**, 4666–4674 (2016).

79. Li, J. *et al.* Ab0135 Protein Conformational Changes and Functional Alterations in Dermal Fibroblasts from Patients with Systemic Sclerosis. *Ann. Rheum. Dis.* **81**, 1197–1197 (2022).

80. Schürch, P. M. *et al.* Calreticulin mutations affect its chaperone function and perturb the glycoproteome. *Cell Rep.* **41**, 111689 (2022).

81. Kikis, E. A., Gidalevitz, T. & Morimoto, R. I. Protein homeostasis in models of aging and age-related conformational disease. *Adv. Exp. Med. Biol.* **694**, 138–159 (2010).

82. Paukštytė, J. *et al.* Global analysis of aging-related protein structural changes uncovers enzyme-polymerization-based control of longevity. *Mol. Cell* **83**, 3360-3376.e11 (2023).

83. Sui, X. *et al. Global proteome metastability response in isogenic animals to missense mutations and polyglutamine expansions in aging*. http://biorxiv.org/lookup/doi/10.1101/2022.09.28.509812 (2022) doi:10.1101/2022.09.28.509812.

84. Shuken, S. R. *et al.* Limited proteolysis–mass spectrometry reveals aging-associated changes in cerebrospinal fluid protein abundances and structures. *Nat. Aging* **2**, 379–388 (2022).

85. Lu, H. *et al.* DiLeu Isobaric Labeling Coupled with Limited Proteolysis Mass Spectrometry for High-Throughput Profiling of Protein Structural Changes in Alzheimer's Disease. *Anal. Chem.* **95**, 9746–9753 (2023).

86. Granato, D. C. *et al. Conformational changes in saliva proteome guides discovery of cancer aggressiveness related markers*. http://biorxiv.org/lookup/doi/10.1101/2023.08.04.552034 (2023) doi:10.1101/2023.08.04.552034.

87. Strimbu, K. & Tavel, J. A. What are Biomarkers? *Curr. Opin. HIV AIDS* **5**, 463–466 (2010).

88. Veenstra, T. D. *et al.* Biomarkers: Mining the Biofluid Proteome *. *Mol. Cell. Proteomics* **4**, 409–418 (2005).

89. Rifai, N., Gillette, M. A. & Carr, S. A. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat. Biotechnol.* **24**, 971–983 (2006).

90. Nakayasu, E. S. *et al.* Tutorial: best practices and considerations for mass-spectrometry-based protein biomarker discovery and validation. *Nat. Protoc.* **16**, 3737–3760 (2021).

91. Babuin, L. & Jaffe, A. S. Troponin: the biomarker of choice for the detection of cardiac injury. *CMAJ* **173**, 1191–1202 (2005).

92. Aldous, S. J. Cardiac biomarkers in acute myocardial infarction. *Int. J. Cardiol.* **164**, 282–294

(2013).

93. Du Clos, T. W. Function of C-reactive protein. *Ann. Med.* **32**, 274–278 (2000).

94. C-Reactive Protein (Blood) - Health Encyclopedia - University of Rochester Medical Center. https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=c_reactive_protein_serum.

95. Schilbach, K. *et al.* Biomarkers of GH action in children and adults. *Growth Horm. IGF Res.* **40**, 1–8 (2018).

96. Gaucher Disease Testing. *National Gaucher Foundation* https://www.gaucherdisease.org/gaucher-diagnosis-treatment/testing/.

97. Lyons, T. J. & Basu, A. Biomarkers in diabetes: hemoglobin A1c, vascular and tissue markers. *Transl. Res.* **159**, 303–312 (2012).

98. Schrag, A., Ben-Shlomo, Y. & Quinn, N. How valid is the clinical diagnosis of Parkinson's disease in the community? *J. Neurol. Neurosurg. Psychiatry* **73**, 529–534 (2002).

99. Green, A. J. E. RT-QuIC: a new test for sporadic CJD. *Pract. Neurol.* **19**, 49–55 (2019).

100. Parkinson disease. https://www.who.int/news-room/fact-sheets/detail/parkinson-disease.

101. Parkinson's Disease: Challenges, Progress, and Promise | National Institute of Neurological Disorders and Stroke. https://www.ninds.nih.gov/current-research/focus-disorders/parkinsons-disease-research/parkinsons-disease-challenges-progress-and-promise.

102. Yilmaz, R., Hopfner, F., van Eimeren, T. & Berg, D. Biomarkers of Parkinson's disease: 20 years later. *J. Neural Transm.* **126**, 803–813 (2019).

103. Spillantini, M. G. *et al.* α-Synuclein in Lewy bodies. *Nature* **388**, 839–840 (1997).

104. Bader, J. M. *et al.* Proteome profiling in cerebrospinal fluid reveals novel biomarkers of Alzheimer's disease. *Mol. Syst. Biol.* **16**, e9356 (2020).

105. Mosleth, E. F. *et al.* Cerebrospinal fluid proteome shows disrupted neuronal development in multiple sclerosis. *Sci. Rep.* **11**, 4087 (2021).

106. Majbour, N. K. *et al.* Oligomeric and phosphorylated alpha-synuclein as potential CSF biomarkers for Parkinson's disease. *Mol. Neurodegener.* **11**, 7 (2016).

107. van Dijk, K. D. *et al.* Changes in endolysosomal enzyme activities in cerebrospinal fluid of patients with Parkinson's disease. *Mov. Disord.* **28**, 747–754 (2013).

108. van Steenoven, I. *et al.* α-Synuclein species as potential cerebrospinal fluid biomarkers for dementia with lewy bodies. *Mov. Disord.* **33**, 1724–1733 (2018).

109. van Dijk, K. D. *et al.* Cerebrospinal fluid and plasma clusterin levels in Parkinson's disease. *Parkinsonism Relat. Disord.* **19**, 1079–1083 (2013).

110. van Dijk, K. D. *et al.* Reduced α-synuclein levels in cerebrospinal fluid in Parkinson's disease are unrelated to clinical and imaging measures of disease severity. *Eur. J. Neurol.* **21**, 388–394 (2014).

111. Ong, S.-E. & Mann, M. Mass spectrometry–based proteomics turns quantitative. *Nat. Chem. Biol.* **1**, 252–262 (2005).

112. Xiao, Z., Prieto, D., Conrads, T. P., Veenstra, T. D. & Issaq, H. J. Proteomic patterns: their potential for disease diagnosis. *Mol. Cell. Endocrinol.* **230**, 95–106 (2005).

113. Bludau, I. *et al.* Systematic detection of functional proteoform groups from bottom-up proteomic datasets. *Nat. Commun.* **12**, 3810 (2021).

114. Tazi, J., Bakkour, N. & Stamm, S. Alternative splicing and disease. *Biochim. Biophys. Acta* **1792**, 14–26 (2009).

115. Xu, H. *et al.* PTMD: A Database of Human Disease-associated Post-translational Modifications. *Genomics Proteomics Bioinformatics* **16**, 244–251 (2018).

116. Paukštytė, J. *et al.* Global analysis of aging-related protein structural changes uncovers enzyme-polymerization-based control of longevity. *Mol. Cell* **83**, 3360-3376.e11 (2023).

117. Meyer, J. G. & Schilling, B. Clinical applications of quantitative proteomics using targeted and untargeted data-independent acquisition techniques. *Expert Rev. Proteomics* **14**, 419–429 (2017).

118. Schmidt, A. & Schreiner, D. Quantitative Detection of Protein Splice Variants by Selected Reaction Monitoring (SRM) Mass Spectrometry. in *Alternative Splicing: Methods and Protocols* (eds. Scheiffele, P. & Mauger, O.) 231–246 (Springer US, 2022). doi:10.1007/978-1-0716-2521-7_14.

119. Mendes, M. L. & Dittmar, G. Targeted proteomics on its way to discovery. *PROTEOMICS* **22**, 2100330 (2022).

120. Ofek, G. *et al.* Elicitation of structure-specific antibodies by epitope scaffolds. *Proc. Natl. Acad. Sci.* **107**, 17880–17887 (2010).

121. Yu, C. & Huang, L. Cross-Linking Mass Spectrometry (XL-MS): an Emerging Technology for Interactomics and Structural Biology. *Anal. Chem.* **90**, 144–165 (2018).

122. Fändrich, M. & Schmidt, M. Methods to study the structure of misfolded protein states in systemic amyloidosis. *Biochem. Soc. Trans.* **49**, 977–985 (2021).

123. Jellinger, K. A. & Korczyn, A. D. Are dementia with Lewy bodies and Parkinson's disease dementia the same disease? *BMC Med.* **16**, 34 (2018).

124. Tateno, F., Sakakibara, R., Kawai, T., Kishi, M. & Murano, T. Alpha-synuclein in the Cerebrospinal Fluid Differentiates Synucleinopathies (Parkinson Disease, Dementia With Lewy Bodies, Multiple System Atrophy) From Alzheimer Disease. *Alzheimer Dis. Assoc. Disord.* **26**, 213 (2012).

125. Jellinger, K. A. Neuropathological Aspects of Alzheimer Disease, Parkinson Disease and Frontotemporal Dementia. *Neurodegener. Dis.* **5**, 118–121 (2008).

126. Walker, L., Stefanis, L. & Attems, J. Clinical and neuropathological differences between Parkinson's disease, Parkinson's disease dementia and dementia with Lewy bodies – current issues and future directions. *J. Neurochem.* **150**, 467–474 (2019).

127. Menšíková, K. *et al.* Lewy body disease or diseases with Lewy bodies? *Npj Park. Dis.* **8**, 1–11 (2022).

128. Compta, Y. & Revesz, T. Neuropathological and Biomarker Findings in Parkinson's Disease and Alzheimer's Disease: From Protein Aggregates to Synaptic Dysfunction. *J. Park. Dis.* **11**, 107–121.

129. Delenclos, M., Jones, D. R., McLean, P. J. & Uitti, R. J. Biomarkers in Parkinson's disease: Advances and strategies. *Parkinsonism Relat. Disord.* **22**, S106–S110 (2016).

130. Kohannim, O. *et al.* Boosting power for clinical trials using classifiers based on multiple biomarkers. *Neurobiol. Aging* **31**, 1429–1442 (2010).

131. Liu, R., Wang, X., Aihara, K. & Chen, L. Early Diagnosis of Complex Diseases by Molecular Biomarkers, Network Biomarkers, and Dynamical Network Biomarkers. *Med. Res. Rev.* **34**, 455–478 (2014).

132. Sloan, A. *et al.* Design and analysis considerations for combining data from multiple biomarker studies. *Stat. Med.* **38**, 1303–1320 (2019).

133. Evers, L. J. W., Peeters, J. M., Bloem, B. R. & Meinders, M. J. Need for personalized monitoring of Parkinson's disease: the perspectives of patients and specialized healthcare providers. *Front. Neurol.* **14**, (2023).

134. Tolosa, E., Garrido, A., Scholz, S. W. & Poewe, W. Challenges in the diagnosis of Parkinson's disease. *Lancet Neurol.* **20**, 385–397 (2021).

135. Leaver, K. & Poston, K. L. Do CSF Biomarkers Predict Progression to Cognitive Impairment in Parkinson's disease patients? A Systematic Review. *Neuropsychol. Rev.* **25**, 411–423 (2015).

136. Discover | EPDN. https://discover.epnd.org/.

137. Foltynie, T., Brayne, C. & Barker, R. A. The heterogeneity of idiopathic Parkinson's disease. *J. Neurol.* **249**, 138–145 (2002).

138. Greenland, J. C., Williams-Gray, C. H. & Barker, R. A. The clinical heterogeneity of Parkinson's disease and its therapeutic implications. *Eur. J. Neurosci.* **49**, 328–338 (2019).

139. Johansson, M. E., van Lier, N. M., Kessels, R. P. C., Bloem, B. R. & Helmich, R. C. Two-year clinical progression in focal and diffuse subtypes of Parkinson's disease. *Npj Park. Dis.* **9**, 1–11 (2023).

140. Fereshtehnejad, S.-M. & Postuma, R. B. Subtypes of Parkinson's Disease: What Do They Tell Us About Disease Progression? *Curr. Neurol. Neurosci. Rep.* **17**, 34 (2017).

141. Eisinger, R. S. *et al.* Motor subtype changes in early Parkinson's disease. *Parkinsonism Relat. Disord.* **43**, 67–72 (2017).

142. Marras, C. & Chaudhuri, K. R. Nonmotor features of Parkinson's disease subtypes. *Mov. Disord.* **31**, 1095–1102 (2016).

143. Marras, C. & Lang, A. Parkinson's disease subtypes: lost in translation? *J. Neurol. Neurosurg. Psychiatry* **84**, 409–415 (2013).

144. Klein, C. & Westenberger, A. Genetics of Parkinson's Disease. *Cold Spring Harb. Perspect. Med.* **2**, a008888 (2012).

145. Inamdar, N., Arulmozhi, D., Tandon, A. & Bodhankar, S. Parkinson's Disease: Genetics and Beyond. *Curr. Neuropharmacol.* **5**, 99–113 (2007).

146. Blauwendraat, C., Nalls, M. A. & Singleton, A. B. The genetic architecture of Parkinson's disease. *Lancet Neurol.* **19**, 170–178 (2020).

147. Lunati, A., Lesage, S. & Brice, A. The genetic landscape of Parkinson's disease. *Rev. Neurol. (Paris)* **174**, 628–643 (2018).

148. Ferreira, M. & Massano, J. An updated review of Parkinson's disease genetics and clinicopathological correlations. *Acta Neurol. Scand.* **135**, 273–284 (2017).

149. von Coelln, R. *et al.* The inconsistency and instability of Parkinson's disease motor subtypes. *Parkinsonism Relat. Disord.* **88**, 13–18 (2021).

150. Simuni, T. *et al.* How stable are Parkinson's disease subtypes in de novo patients: Analysis of the PPMI cohort? *Parkinsonism Relat. Disord.* **28**, 62–67 (2016).

151. Chiang, H.-L. & Lin, C.-H. Altered Gut Microbiome and Intestinal Pathology in Parkinson's Disease. *J. Mov. Disord.* **12**, 67–83 (2019).

152. Gaig, C. & Tolosa, E. When does Parkinson's disease begin? *Mov. Disord.* **24**, S656–S664 (2009).

153. Aarsland, D. *et al.* Parkinson disease-associated cognitive impairment. *Nat. Rev. Dis. Primer* **7**, 1–21 (2021).

154. Chen, Q. *et al.* Data-driven subtyping of Parkinson's disease: comparison of current methodologies and application to the Bochum PNS cohort. *J. Neural Transm. Vienna Austria 1996* **130**, 763–776 (2023).

155. Fereshtehnejad, S.-M., Zeighami, Y., Dagher, A. & Postuma, R. B. Clinical criteria for subtyping Parkinson's disease: biomarkers and longitudinal progression. *Brain* **140**, 1959–1976 (2017).

156. D'Sa, K. *et al.* Prediction of mechanistic subtypes of Parkinson's using patient-derived stem cell models. *Nat. Mach. Intell.* **5**, 933–946 (2023).

157. Zhang, X. *et al.* Data-Driven Subtyping of Parkinson's Disease Using Longitudinal Clinical Records: A Cohort Study. *Sci. Rep.* **9**, 797 (2019).

158. Ygland Rödström, E. & Puschmann, A. Clinical classification systems and long-term outcome in mid- and late-stage Parkinson's disease. *Npj Park. Dis.* **7**, 1–9 (2021).

159. De Pablo-Fernández, E., Lees, A. J., Holton, J. L. & Warner, T. T. Prognosis and Neuropathologic Correlation of Clinical Subtypes of Parkinson Disease. *JAMA Neurol.* **76**, 470–479 (2019).

160. Farrow, S. L., Cooper, A. A. & O'Sullivan, J. M. Redefining the hypotheses driving Parkinson's diseases research. *Npj Park. Dis.* **8**, 1–7 (2022).

161. Thenganatt, M. A. & Jankovic, J. Parkinson Disease Subtypes. *JAMA Neurol.* **71**, 499–504 (2014).

162. Abdelmoaty, M. M. *et al.* Clinical biomarkers for Lewy body diseases. *Cell Biosci.* **13**, 209 (2023).

163. Chan, Y. H., Wang, C., Soh, W. K. & Rajapakse, J. C. Combining Neuroimaging and Omics Datasets for Disease Classification Using Graph Neural Networks. *Front. Neurosci.* **16**, (2022).

164. Dunker, A. K. *et al.* Intrinsically disordered protein. *J. Mol. Graph. Model.* **19**, 26–59 (2001).

165. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *J. Mol. Biol.* **337**, 635–645 (2004).

166. Holehouse, A. S. & Kragelund, B. B. The molecular basis for cellular function of intrinsically disordered protein regions. *Nat. Rev. Mol. Cell Biol.* 1–25 (2023) doi:10.1038/s41580-023-00673-0.

167. Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **16**, 18–29 (2015).

168. Brodsky, S., Jana, T. & Barkai, N. Order through disorder: The role of intrinsically disordered regions in transcription factor binding specificity. *Curr. Opin. Struct. Biol.* **71**, 110–115 (2021).

169. Cuylen, S. *et al.* Ki-67 acts as a biological surfactant to disperse mitotic chromosomes. *Nature* **535**, 308–312 (2016).

170. Pelham, J. F., Dunlap, J. C. & Hurley, J. M. Intrinsic disorder is an essential characteristic of components in the conserved circadian circuit. *Cell Commun. Signal.* **18**, 181 (2020).

171. Witus, S. R. *et al.* BRCA1/BARD1 intrinsically disordered regions facilitate chromatin recruitment and ubiquitylation. *EMBO J.* **42**, e113565 (2023).

172. Oldfield, C. J. & Dunker, A. K. Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. *Annu. Rev. Biochem.* **83**, 553–584 (2014).

173. van der Lee, R. *et al.* Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **114**, 6589–6631 (2014).

174. Borcherds, W. *et al.* Disorder and residual helicity alter p53-Mdm2 binding affinity and signaling in cells. *Nat. Chem. Biol.* **10**, 1000–1002 (2014).

175. Moses, D., Ginell, G. M., Holehouse, A. S. & Sukenik, S. Intrinsically disordered regions are poised to act as sensors of cellular chemistry. *Trends Biochem. Sci.* **48**, 1019–1034 (2023).

176. Robustelli, P., Piana, S. & Shaw, D. E. Mechanism of Coupled Folding-upon-Binding of an Intrinsically Disordered Protein. *J. Am. Chem. Soc.* (2020) doi:10.1021/jacs.0c03217.

177. Tompa, P. & Fuxreiter, M. Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. *Trends Biochem. Sci.* **33**, 2–8 (2008).

178. Cuevas-Velazquez, C. L. *et al.* Intrinsically disordered protein biosensor tracks the physical-chemical effects of osmotic stress on cells. *Nat. Commun.* **12**, 5438 (2021).

179. Schultz, B. B. Levene's Test for Relative Variation. *Syst. Biol.* **34**, 449–456 (1985).

180. Clore, G. M. & Gronenborn, A. M. Determination of Three-Dimensional Structures of Proteins and Nucleic Acids in Solution by Nuclear Magnetic Resonance Spectroscop. *Crit. Rev. Biochem. Mol. Biol.* **24**, 479–564 (1989).

181. Hollunder, J., Beyer, A. & Wilhelm, T. Identification and characterization of protein subcomplexes in yeast. *PROTEOMICS* **5**, 2082–2089 (2005).

182. Dermit, M., Peters-Clarke, T. M., Shishkova, E. & Meyer, J. G. Peptide Correlation Analysis (PeCorA) Reveals Differential Proteoform Regulation. *J. Proteome Res.* **20**, 1972–1980 (2021).

183. Roskoski, R. Src kinase regulation by phosphorylation and dephosphorylation. *Biochem. Biophys. Res. Commun.* **331**, 1–14 (2005).