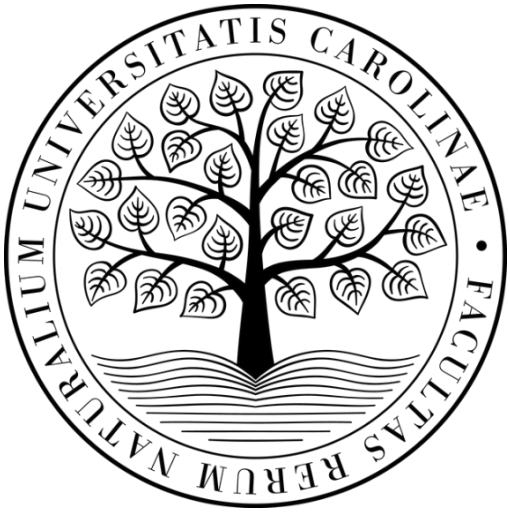
CHARLES UNIVERSITY

Faculty of Science

Department of Biochemistry



Mgr. Marek Zákopčaník

Bioinformatic evaluation of mass spectrometric data obtained by analysis of proteins modified by reactive oxygen species

Bioinformatické vyhodnocení hmotnostně-spektrometrických dat získaných analýzou proteinů modifikovaných reaktivními formami kyslíku

DOCTORAL THESIS

Supervisor: RNDr. Petr Novák, Ph.D.

Prague, 2024

Declaration

I declare that I have worked on this thesis independently and that I have listed all the sources of information, and the literature used. This work, or a substantial part of it, has not been submitted to obtain another or the same academic degree.

Prague _____ _____

Mgr. Marek Zákopčaník

Declaration of authorship

I declare Mgr. Marek Zákopčaník contributed significantly to the experiments and to all scientific publications contained in this Ph.D. thesis. He performed most of the evaluations, substantially contributed to the planning of the experiments, and took a significant part in the primary data interpretation and their preparation for publication.

Prague _____ _____

RNDr. Petr Novák, Ph.D.

Acknowledgements

I would like to thank my supervisor, Dr. Petr Novák, for guiding me through all the years of my work and for his time and willingness to help me. My thanks also go to Dr. Dmitry Loginov for productive cooperation and discussions. I would also like to thank the whole group, the Laboratory of Structural Biology and Cell Signalling at the Institute of Microbiology, CAS, for creating a great environment and for all the acts of camaraderie.

I sincerely thank my family for their unconditional support throughout all the years of my studies.

I would also like to acknowledge our funding sources, namely the Technology Agency of the Czech Republic grant (ODEEP-EU TH86010001), and the Ministry of Education, Youth and Sports of the Czech Republic grant "PHOTOMACHINES - Photosynthetic cell redesign for high yields of therapeutic peptides" (CZ.02.01.01/00/22_008/0004624).

Abstract

The field of structural mass spectrometry (MS) is a rapidly developing area comprising of many versatile techniques for the characterisation of biomolecules under native or near-native conditions. Among them, reactive radical labelling represents a promising approach, delivering insight into the conformational adjustments and biomolecular dynamics. Specifically, reactive oxygen species are capable of providing snapshots of the solvent-accessible regions of the studied system. Fast Photochemical Oxidation of Proteins (FPOP) represents a powerful and accessible platform for protein footprinting, but the complexity of FPOP data presents significant challenges in bioinformatic analysis. This thesis aims to contribute to overcoming these difficulties by investigating the important aspects of FPOP data evaluation and proposing a novel FPOP workflow.

In the first part of this work, different identification frameworks for the analysis of the FPOP data were evaluated. A detailed comparison between two widely used search engines, Mascot and PEAKS, was conducted to assess their performance in identifying modified peptides. The search engines were compared in their consistency across replicates and the confidence in reliability of the modification site assignment.

Next, the thesis explored the potential of data-independent acquisition (DIA) as an improvement for the FPOP analysis. DIA offers the advantage of comprehensive and unbiased data collection, capturing a more complete set of fragmentation spectra. The use of manually validated spectral library matched with the DIA data improved the reliability of identification of modified peptides and the spatial coverage of the analysis, when compared to the workflow using routine filtration steps. Furthermore, the reproducibility of quantification was evaluated for data-dependent, data-independent, and MS1-only acquisitions.

The last part showcases an identification of suspected modification in the structure of blue laccase B76 from the termite species *Neocapritermes taracua*. This work demonstrated the utilisation of an open search to characterise reactive oxygen species-induced modification.

Keywords:

proteins, protein footprinting, reactive oxygen species, mass spectrometry, liquid chromatography, bioinformatics

Abstrakt

Strukturní hmotnostní spektrometrie je dynamicky se rozvíjející oblastí, která zahrnuje mnoho všestranných technik pro charakterizaci biomolekul v podmínkách blízkým jejich přirozenému prostředí. Jednou z těchto technik je značení pomocí reaktivních radikálů, jenž umožňuje studovat konformační a interakční dynamiku proteinů. Reaktivní formy kyslíku jsou díky své rychlosti značení schopné poskytnout „snímek" oblastí dostupných rozpouštědlu. Rychlá fotochemická oxidace proteinů (z anglického „Fast Photochemical Oxidation of Proteins – FPOP) je výkonná a dostupná metoda pro studium struktury proteinů, ale složitost FPOP dat ztěžuje následnou analýzu a interpretaci. Cílem této práce je přispět k překonání těchto potíží, a to zaměřením se na aspekty důležité pro vyhodnocení FPOP dat a navržením nového postupu pro zpracování FPOP analýz.

V první části této práce byl zkoumán vliv použitého počítačového programu pro prohledávání hmotnostně-spektrometrických dat na výsledky FPOP analýz. Dva široce používané programy, Mascot a PEAKS, byly porovnány z hlediska jejich schopnosti identifikovat modifikované peptidy. Byla posouzena jejich konzistence napříč replikáty a jejich spolehlivost přiřazení modifikace ke konkrétní pozici v sekvenci peptidu.

Dále se práce zaměřila na využití datově-nezávislého sběru dat pro vylepšení FPOP analýzy. Datově-nezávislý sběr dat přináší komplexní záznam dat, který není ovlivněn měřeným vzorkem, a poskytuje úplnější sadu fragmentačních spekter. Ručně validovaná spektrální knihovna, porovnaná s datově-nezávislými daty poskytla spolehlivější identifikace modifikovaných peptidů a vyšší prostorové pokrytí proteinů, než postup s rutinními filtračními kroky. Byla porovnána reprodukovatelnost kvantifikace FPOP dat při sběru dat v datově-závislém a datově-nezávislém režimu, a režimu bez fragmentace.

Poslední část této doktorské práce demonstruje identifikaci spekulované modifikace ve struktuře modré lacasy B76 z termita druhu *Neocapritermes taracua*. Za využití otevřeného prohledávání hmotnostně-spektrometrických dat byla odhalena modifikace vytvořená reaktivními formami kyslíku.

Klíčová slova:

bílkoviny, pulsní značení bílkovin, reaktivní formy kyslíku, hmotnostní spektrometrie, kapalinová chromatografie, bioinformatika

[ANGLICKY]

Table of contents

Abbreviations 9
1. Introduction 10
1.1. High-resolution methods for studying protein structure 10
1.2. Structural mass spectrometry 12
1.2.1. Native MS and ion mobility 12
1.2.2. Chemical cross-linking 13
1.2.3. Mass spectrometry-based protein labelling 14
1.2.3.1. Hydrogen/deuterium exchange 14
1.2.3.2. Amino acid-targeted labelling 16
1.2.3.3. Reactive radical (untargeted) labelling 18
1.3. Identification of proteomic mass spectrometry data 30
1.3.1. Database search engines 31
1.3.1.1. Historical development of database search engines 31
1.3.1.2. Mechanisms of database search engines 32
1.3.1.3. Application of database search in covalent labelling MS 33
1.3.2. Spectral library search 34
1.3.2.1. Historical development of library search engines 34
1.3.2.2. Mechanisms of spectral library search engines 34
1.3.2.3. Application of spectral library search in covalent labelling MS 35
2. Aims of the thesis 37
3. Methods 38
3.1. List of experimental methods used 38
4. Results and Discussion 39
4.1. Publication I 39
4.2. Publication II 47
4.3. Publication III 53
5. Summary 57
6. List of publications 58
7. References 59
Attached publications 74

Abbreviations

CV	coefficient of variation
cryo-EM	cryogenic electron microscopy
DDA	data-dependent acquisition
DIA	data-independent acquisition
ESI	electrospray ionisation
FDR	false discovery rate
FFAP	Fast Fluoroalkylation of Proteins
FPOP	Fast Photochemical Oxidation of Proteins
Hb	haemoglobin
HbHp	haemoglobin-haptoglobin complex
Hp	haptoglobin
HOS	higher order structure
HDX	hydrogen/deuterium exchange
HRF	hydroxyl radical footprinting
ID	identification
IM	ion mobility
$[1/K_0] = V \cdot s \cdot cm^{-2}$	ion mobility (quantity)
LC	liquid chromatography
MS	mass spectrometry
MALDI	matrix-assisted laser desorption/ionisation
nanoESI	nanoflow electrospray ionisation
NMR	nuclear magnetic resonance
PSM	peptide-spectrum match
PTM	post-translational modification
PCA	Principal Component Analysis
RT	retention time
SASA	solvent-accessible surface area
SSM	spectrum-spectrum match

1. Introduction

Proteins represent the ultimate manifestation of genetic information in all living organisms. Their utility ranges from structural support, signalling, and immunological response to catalytical activity, transport, and replication of genetic information. Although nature has seemingly 'only' 20 building blocks available, the combinatorics together with an array of posttranslational modifications allow for immense variety among individual proteoforms. The function of a protein molecule is inherently dependent on its structure, which originates from the order of the amino acids in the polypeptide chain – primary structure. Certain regions of the polypeptide chain can allow for the proximity of carbonyl and amino groups of the protein backbone, forming α-helices, β-sheets and other secondary structures. The coordination of these local folds with the help of van der Waals forces, hydrophobic interactions, disulphide bridges, hydrogen and ionic bonds determines the overall shape of a single polypeptide chain – tertiary structure. Some proteins, such as haemoglobin, nucleosomes, or antibodies, achieve their full potential only in the form of several polypeptide chains arranged mainly via noncovalent interactions into the quaternary structure. All these levels of assembly come together to form a unique functional unit with the ability to attract, repulse, reshape itself, or induce changes based on its surroundings and interaction partners. Therefore, the study of protein structure aids in the investigation of molecular machinery of the living organisms.

1.1. High-resolution methods for studying protein structure

Throughout history, molecular biology has gained several techniques for studying protein structure. These approaches differ in the physical or chemical principles which they employ, and consequently in the information they can provide about the system in question.

Starting with the first structure of myoglobin solved by X-ray crystallography in 1958[1], this method has been the main source of high-resolution snapshots of protein structures. It employs analysis of the diffraction patterns obtained by illuminating clusters of protein molecules regularly arranged in crystals[2]. Because the X-ray wavelengths are near the distances between atoms (acting as resonators), the electromagnetic radiation interferes with their electrons, which results in diffraction. This is amplified by the repeating of the array, leading to diffraction patterns that can be translated into electron density maps. These maps are then used to fit the atom

coordinates, which then form the structure as a whole. Over time, this technique has been refined to provide structures with resolution close to 1 Å. However, to utilise this technique, a crystal of sufficient size (exceeding 0.1 mm) is required. This may not be achievable with certain proteins that have, for example, a significant proportion of unstructured regions[3]. It is also worth noting that the conformation in which a certain protein crystalizes may substantially differ from its *in vivo* form.

Nuclear magnetic resonance (NMR) represents a fundamentally different approach, allowing the study of proteins in solution and thus granting access to the full dynamics of the molecule. It is based on the ability of atoms with non-zero nuclear spin (e.g. ^{1}H, ^{13}C, ^{15}N and ^{31}P) to change their orientation in an external magnetic field upon absorption of electromagnetic radiation in the radio range[4]. This causes nuclei occupying lower energy spin states (aligned with the external magnetic field) to transition into higher energy spin states (aligned against the field). The exact energy difference between those states is then measured upon system relaxation. Since the absorbed frequency is dependent on the type of the atom and on the chemical surroundings, NMR measurements obtain information about spatial constraints, leading to the reconstruction of the protein structure. Because the method captures the protein in solution, the resulting structure is not a single snapshot but rather a superposition of all observed states. Although it brings unquestionable advantages, NMR is not without limits. This includes protein size, because proteins larger than 40 kDa produce NMR spectra that are too convoluted to be interpreted unambiguously. Secondly, the demand for isotopically labelled protein variants increases the costs. Furthermore, there is a need for protein stability, with experiment duration reaching up to days and requiring concentrations around 1 mM, these conditions may not be feasible for some proteins.

Due to recent advancements in direct electron detectors, cryogenic electron microscopy (cryo-EM) has emerged as a promising competitor to high-resolution structural techniques[5,6]. Its principle relies on transmission electron microscopy of samples that underwent *vitrification*. This shock-freezing in liquid alkane solution captures proteins in their current state, encapsulating them in vitreous ice, while preventing the formation of crystalline ice[7]. In this way the protein molecules can be irradiated by the electron beam and by acquiring thousands of two-dimensional scans with different orientations of the protein, a three-dimensional structure can be reconstructed[8]. With the ongoing development, cryo-EM is achieving resolution similar to that of X-ray crystallography and NMR spectroscopy. Cryo-EM is proving

useful in the investigation of large complexes[9], highly symmetric structures[10], and studies of therapeutic antibodies[11,12].

1.2. Structural mass spectrometry

In addition to the traditional high-resolution techniques mass spectrometry (MS) has become widely used for the study of proteins. The wide adaptation was powered by the introduction of novel 'soft' ionisation methods in the 1980s, namely electrospray ionisation (ESI) and matrix-assisted laser desorption/ionisation (MALDI)[13]. This allowed for gentle transfer of biomolecules to the gas phase in the form of charged ions, which are detectable by mass analysers. In the case of proteins, mass spectrometry is often coupled with reverse phase liquid chromatography (LC) as the separation method of choice, leading to the predominant use of ESI. This was further improved by the introduction of nanoflow electrospray ionisation (nanoESI)[14]. NanoESI not only minimises the amount of sample consumption, but also deals better with salt contamination, allows for detection of non-surface active analytes, (such as glycosides and glycoproteins), and produces more stable electrospray for aqueous solutions[15].

The methods of structural mass spectrometry comprise several techniques that are complementary to the high-resolution approaches mentioned above. Despite not being able to provide information on the precise location of atoms, structural MS can provide insight indicative of the spatial arrangement of the molecules and their post-translational modifications (PTMs). Mass spectrometry techniques benefit from high sensitivity and the availability of coupled separation techniques. Due to the extensive range of methods, structural MS is almost unlimited in terms of the protein size, complexity or presence of unstructured regions, obstacles that are often insurmountable for other approaches. Moreover, MS techniques can investigate the changes induced by diverse conditions and interaction partners, making them invaluable in the study of protein-protein dynamics or protein-ligand screening. All this is possible while requiring significantly less protein amount per analysis[16–20].

1.2.1. Native MS and ion mobility

Native mass spectrometry brings the possibility of gently enough transfer of biomolecules from the solution to the gas phase so that the compounds remain in their intact state, even preserving their PTMs and noncovalent interactions[21]. By leveraging the 'soft' electrospray ionisation, native MS is capable of delivering information on

proteins with superior biological relevance. Native mass spectrometry has been an invaluable tool to study protein complexes, metalloenzymes, nucleic acid binding factors, or antibody selectivity[22–25]. Since the charge state of the protein after electrospray ionisation is dependent on the structure (number of accessible ionisable groups), the method is suitable for the detection of protein unfolding, whether the stability is probed by solution, interaction partners, or thermal conditions[26,27].

Ion mobility (IM) adds another layer of information by separating the analytes according to their shape and size-to-charge ratio[28]. In the ion mobility analyser molecules are propelled by an electric field, while being slowed by an inert gas. The more compact a molecule is, the fewer collisions with the inert gas it will undergo, travelling through the analyser faster. The ion mobility is then derived from the time of travel. Coupling of IM to mass spectrometry enables the determination of the collision cross-section, as a projection of the compound shape on its average orientation[21]. The application of ion mobility ranges from individual protein folding[29], through viral interactions[30], to DNA nanostructure assembly[31].

1.2.2. Chemical cross-linking

Cross-linking reagents represent a bifunctional footprinting compounds, that can target amino acid side chains in a specific or semi-specific manner, to provide information on protein distance constraints and protein-protein interactions. The reaction occurs under near-native conditions and is not limited by the size of the protein or the complexity of the system[32]. In the classical bottom-up approach the protein sample is subsequently digested and analysed by LC-MS to reveal which regions are proximal. Two residues can be linked only if their distance is within the length of the cross-linker, an information which can be utilised for structure determination.

The properties of the cross-linking reagents can be tailored by selecting the reactive functional group or by the length of the linker region. The spectrum ranges from the 'zero-length' cross-linkers, that use the functional groups of the polypeptide chain without incorporating any linker region, which are used to report on sites of direct contact within the salt bridge distance[33]; through ~11 Å, typical for mapping the protein interface; up to 60 Å of linker length, employed for in cell cross-linking studies[34].

The linker region allows for incorporation of labile bonds to produce a cross-linker, that is cleavable upon fragmentation. This provides characteristic fragment ions, which aids the process of identification[35]. Another approach to reduce the data complexity is to use isotopically labelled reagents. This not only allows identification of cross-linking products with higher confidence but can also be utilised for quantitative studies of protein in two conformational states[36].

1.2.3. Mass spectrometry-based protein labelling

While mass spectrometry allows for the study of proteins in their near-native state, the greatest strength in the pursuit of higher order structure (HOS) elucidation lies in the utilisation of covalent and radical labelling. Whether it is reversible or irreversible, covalent labelling results in a mass shift of the targeted areas, which can be detected by mass spectrometry. Since the terminology has not been unified, the terms 'labelling' and 'footprinting' will be used interchangeably throughout the text.

Hydrogen/deuterium exchange (HDX) is a representative of reversible labelling technique, because the deuterium readily undergoes back exchange with available hydrogen (e. g. from solvent). In the most commonly used setup, only the backbone accessibility information is utilised, making the HDX also untargeted. On the other hand, irreversible protein footprinting uses probes that label the solvent-accessible surface area (SASA) in a stable manner. The technique is designed on the premises that the functional groups of the amino acid side chains can be altered by specific chemical reagents, the reactivity depends on the solvent accessibility of the residue, and the initial phase of the labelling does not disturb the native state of the protein. The irreversible footprinting reagents can either be targeted towards specific side chain motifs, or they can react with a broad range of amino acids, functioning as untargeted probes.

1.2.3.1. Hydrogen/deuterium exchange

The history of hydrogen/deuterium exchange started in the 1950s, with pork insulin as the first protein studied with HDX in 1954[37]. At first, this technique was coupled with NMR spectroscopy, which offered sufficient spatial resolution on amino acid level[38]. Coupling with mass spectrometry came several decades later in 1991, when it was used to observe conformational changes in bovine ubiquitin induced by methanol[39]. Since then, HDX-MS has developed into a versatile technique with well-established methodology.

To briefly introduce the principle, HDX employs the ability of O-H, N-H and S-H hydrogens in the protein to exchange rapidly for deuterium in the D_2O solution (Figure 1)[40]. The deuterium uptake can then be measured as a mass shift using mass spectrometry. Although the basis for labelling is covalent, it is fully reversible. Despite the exchange occurring on both protein backbone and amino acid side chains, the interchangeable side chain hydrogens exchange at such rapid rate that they revert to their original state upon the briefest encounter of H_2O environment (e.g., proteolysis, separation). The H→D exchange rate is influenced by several factors including the pH, temperature, the solvent accessibility, and intramolecular hydrogen bonding[41–44]. The first two parameters play an important role in the experimental design, allowing to limit the exchange by lowering the pH to approximately 2.5 and flash-freezing the sample. The latter reflect the protein conformation and thus allow HDX to provide information on the protein higher order structure. The isotope exchange is fastest for unstructured regions, where the backbone amides do not participate in hydrogen bonding. However, structured regions involve a whole assembly of hydrogen bonds, combined with limited solvent access of such areas, hindering the exchange.

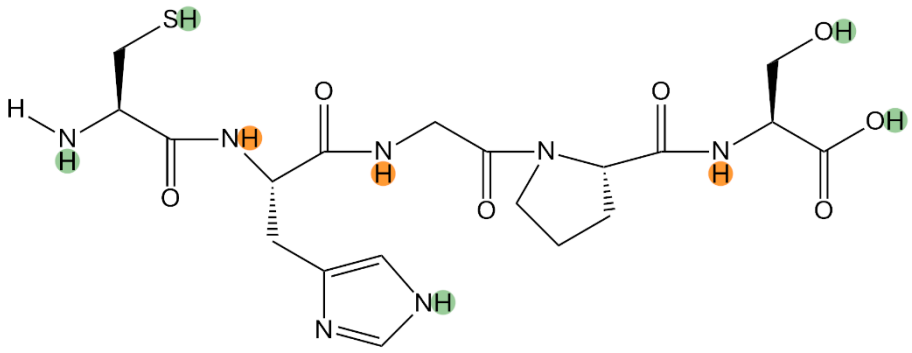


Figure 1 | Exchangeable hydrogens in the polypeptide chain. Hydrogens labelled green are exchanging too fast to be readily detectable. The exchange of backbone amide hydrogens (labelled orange) is detectable by common HDX setup.

Hydrogen/deuterium exchange serves as an invaluable tool for characterising protein binding sites[45], study of protein folding[46], investigating protein complex formation dynamics[47], effective biotherapeutics discovery[48] and examination of protein-membrane interactions[49]. The method benefits from years of development, that resulted in an established workflow[50] and copious amount of HDX data processing software[51]. While the standard bottom-up HDX approach generally provides information on the peptide level, options to increase the spatial resolution of the method have been explored. One approach deconvolutes the single amino acid resolution from many sequentially overlapping peptides[52], another employs electron

transfer dissociation to obtain fragmentation spectra without the intramolecular gas-phase hydrogen scrambling, which would erase the relevance of the labelling[53].

1.2.3.2. Amino acid-targeted labelling

Selective covalent labelling employs chemical reagents that modify specific functional groups of amino acid side chains, making them complementary to HDX. This technique assumes that the amino acid side chains will be modified by the reagent, depending on their solvent accessibility and their reactivity, and that the side chains involved in conformational changes or protein interactions will demonstrate different extent of modification. The following is an overview of amino acids and their available labelling approaches.

- Arginine

The positively charged end of the arginine side chain allows for nucleophilic attack, typically by reagents with neighbouring carbonyl groups such as phenylglyoxal[54], cyclohexadione[55] or 2,3-butanedione[56]. Since polar residues are often involved in protein interactions and catalysis[57], targeting arginine can provide information about the influence of specific conditions on protein reactivity.

- Carboxylic acids

Due to their charged nature, aspartic and glutamic acid residues are often located on the protein surface or taking part in electrostatic interactions, making them amendable to labelling but limiting their relevancy toward information about protein higher-order structure. Labelling of carboxyl groups in the protein structure is performed using 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride (EDC), glycinamide-HCl, or glycine ethyl ester. Modification of carboxylic acids was used to probe differences in the tertiary structure of Ras protein between its active (GTP-bound) and inactive (GDP-bound) forms[58], in the study of the membrane topology of the guinea pig cytochrome P450[59], or to characterise the structure of monoclonal antibody[60].

- Cysteine

The highly reactive nucleophilic thiol group of cysteine makes it one of the most reactive amino acids. The majority of cysteines in proteins form disulphide bonds, stabilising the tertiary structure. Free cysteine residues can be modified with iodoacetamide, iodoacetic acid and related derivatives[61,62]; chloroacetamide; N-alkylmaleimide[63] or vinylpyridine[64]. The high reactivity allows for investigation of

which cysteines participate in the formation of disulphide bonds, by first labelling the intact protein with one reagent and utilising an alternative reagent to target cysteines, which become accessible after reduction [usually by dithiothreitol, *tris*(2-carboxyethyl)phosphine, or 2-mercaptoethanol][65,66].

- Histidine

The ability to release and gain proton in its imidazole ring in near-physiological pH conditions designs this amino acid to play a key role in the catalytic mechanism of many emzymes[67]. Thus, histidine attracts attention despite its low abundance in proteins. The most specific and commonly used histidine labelling reagent is diethylpyrocarbonate, first characterised as a trypsin inactivator[68,69].

- Lysine

The ε-NH_2 moieties of lysine exemplify (same as the N-terminal) primary amino groups, which are most commonly modified by acetylation. This is usually performed via reaction with organic acid anhydrides, such as acetic anhydride[70], succinic anhydride[71], and *N*-hydroxysuccinimide esters[72]. Since the lysines are often located on the protein surface, they represent an alluring target for covalent labelling. Due to its positive charge at physiological pH, lysine is typically the target of study of interactions with the surrounding polar environment[73], or with negatively charged ligands, such as DNA[74].

- Tryptophan

The electron rich heterocyclic aromatic side chain is the basis for the hydrophobicity of tryptophan, which causes its rare occurrence on the protein surface. Therefore, the number of labelling studies targeting tryptophan is limited. However, beside the usual labelling reagents including 2-hydroxy-5-nitrobenzyl bromide[75], *N*-bromosuccinimide[76], and 2-nitrophenylsulphenyl chloride[77], novel labelling reagents, such as sulfonyl-triazoles[78] or *N*-sulfonyl oxaziridine[79], are emerging.

- Tyrosine

Tyrosine can be the carrier of several post-translational modifications[80–82]; hence, it is amendable to covalent labelling. In addition, it can either be exposed or buried, making it an interesting footprinting target. Chemical modifications target either the aromatic ring: nitration, usually by tetranitromethane[83], and iodination[84]; or the hydroxyl group: derivatization by *N*-acetylimidazole[85].

- Other amino acids

The amino acids with aliphatic side chains (alanine, glycine, isoleucine, leucine, proline, and valine) are generally inert to targeted chemical labelling. Although some studies explored the possibilities of C-H derivatization[86–88], they are not applicable to whole proteins and require conditions that are far from physiological. The carboxamide side chains of asparagine and glutamine do not allow for virtually any targeted labelling[89]. While methionine readily undergoes oxidation to the corresponding sulfoxide or sulfone, its reactivity to other means of chemical modification is feasible only under acidic conditions[90–92], diminishing the biological relevance of such experiment on proteins. Serine and threonine both contain hydroxyl groups, which can carry PTMs like phosphorylation and *O*-glycosylation[93], however, the stronger nucleophiles present in the protein sequence, such as cysteine, lysine and tyrosine, usually siphon the derivative reagents addressed to these amino acids[51,89].

Although the use of amino acid-specific labelling reagents offers straightforward and often precisely targeted structure footprinting, elucidating information on the residue solvent accessibility, their practical employment is limited. Targeting of one amino acid inherently leads to low spatial coverage. Moreover, in contrast to the HDX already mentioned, chemical reagents often introduce relatively large functional groups, which could alter the chemical properties of the side chains, such as charge, hydrophobicity, and the ability to participate in bonds. This, combined with generally longer incubation periods (ranging from minutes to hours), increase the risk of perturbation of the protein structure. Together with reaction conditions, which sometimes diverge notably from the protein native conditions, the biological relevance of such obtained data should be evaluated with caution. Taken together, structural biology would undoubtedly benefit from covalent labelling reagent that offers greater spatial coverage through wider reactivity; provides more relevant information on the solvent accessibility of the labelled residue; and perturbs the protein native structure only to a limited extent.

1.2.3.3. Reactive radical (untargeted) labelling

The need to overcome these limitations kindled the development of untargeted probes. These non-specific labelling reagents carry the advantage of broad structural coverage because of their wider reactivity. The reagents are still modifying the amino acid side chains, providing information complementary to HDX experiments. Moreover, the labelling reactions using radical probes occur on a microsecond

timescale[94–96], allowing to capture snapshots of the protein folding steps or other transient states.

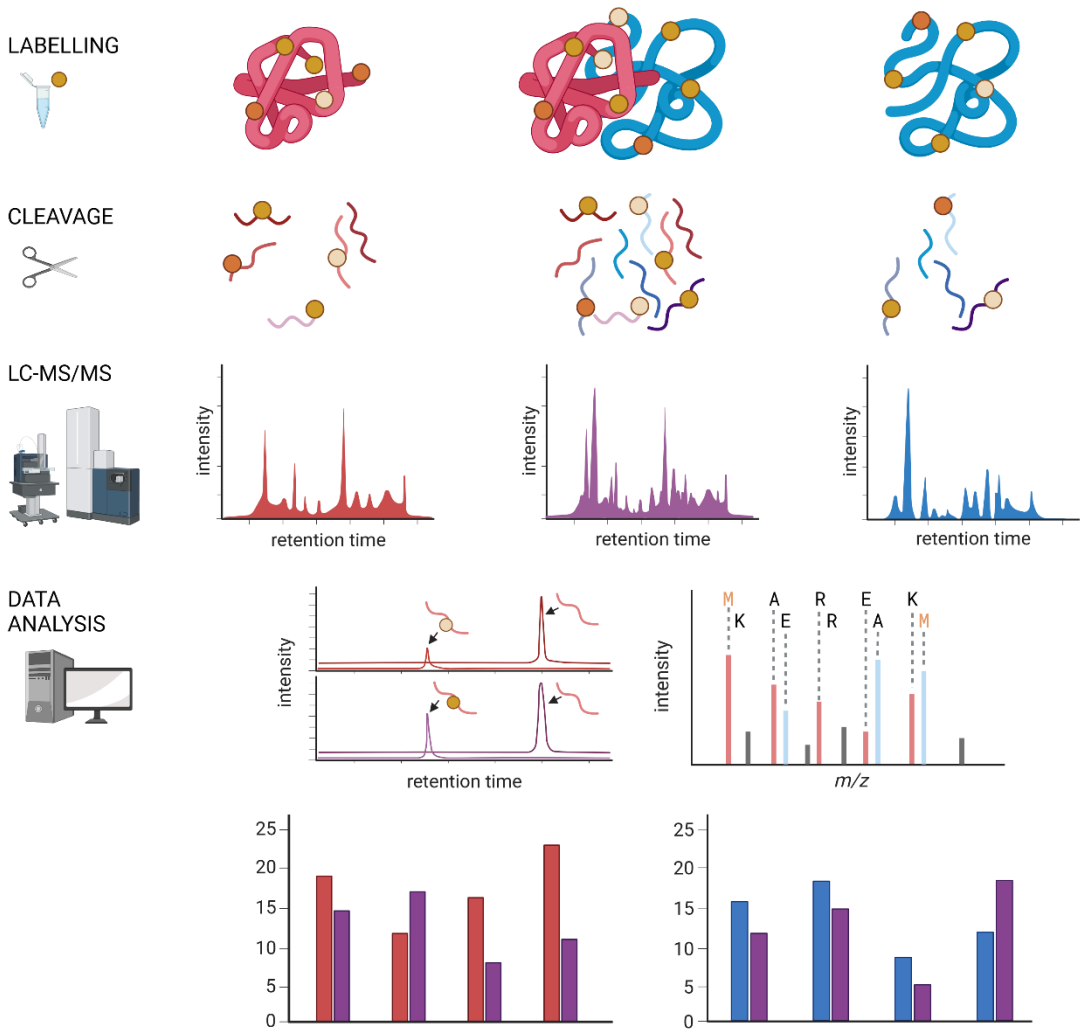


Figure 2 | Scheme of the differential protein labelling experiment. Proteins are labelled both individually and in a complex. After cleavage, the peptides are analysed using LC-MS/MS. The exact site of modification is determined from fragmentation spectra, and the extent of modification is quantified from MS1 chromatogram trace. The effect of protein complex formation is then observed as a difference between the unbound and bound state. Created with BioRender.com

Generally, the protein labelling workflow uses a differential setup, comparing the extent of modification in the unbound versus bound state (Figure 2). The protein (or proteins in the case of study of complexes) is labelled individually and with its interaction partner(s). On the basis of the footprinting probe, one or multiple amino acid side chains of the modified peptide are altered, resulting in mass shift detectable by MS. After protein digestion, the individual peptides are separated using liquid chromatography and analysed with MS/MS. Fragmentation spectra are used to identify the modified peptide and the exact position of the modification within the

peptide sequence. The precise determination of the site is crucial for evaluation of the extent of modification in detail necessary to obtain comprehensive structural information. In principle, the radical labelling approach may reach single amino acid resolution.

HYDROXYL RADICALS

The first application of the hydroxyl radical (·OH) was in nucleic acid footprinting[97]. After the hydroxyl radical is generated, it abstracts solvent accessible hydrogen in the DNA/RNA backbone, which subsequently breaks down the phosphodiester bond. The formed fragments are then analysed using electrophoresis. Solvent-accessible regions are cleaved preferentially; thus, the method provides information on nucleic acid higher order structure. Later, the concept was employed for the study of protein structure. While the first adaptations used hydroxyl radicals to cleave the protein backbone, it was later found that the ·OH acts predominantly as a labelling reagent of the amino acid side chains[98,99]. Protein hydroxyl radical footprinting (HRF) was first demonstrated by a group around Chance in 1999, with the identification of the footprinting products and the proposed reactivity order[100].

The major advantage of ·OH stems from its unparalleled resemblance to water. The hydroxyl radical is comparable to the H_2O molecule in both van der Waals area and hydrophobicity[101], leading to superior biological relevance of this labelling. Furthermore, it exhibits broad reactivity across amino acids and the individual reaction mechanisms are well explored[102]. There are two main starting mechanisms for ·OH oxidation – either by hydrogen abstraction from saturated carbon site, or hydroxyl addition to unsaturated carbon-carbon double bonds or aromatic rings. Hence, radical intermediates created are primed to follow subsequent reactions, based on structure (strength of the C-H bond, substituent electronegativity, and organoradical stability) and proximal environment (steric effects, protein structure-induced microenvironment). The hydroxyl radical preferentially reacts with the side chains of sulphur-containing and aromatic amino acids, while glycine, alanine, aspartate, and asparagine rank as the least reactive to ·OH, yielding them impractical as hydroxyl radical labelling targets (Table 1, page 21). Modifications of serine and threonine are also considered marginal due to their low abundance and low detectability. Practically speaking, the side chains of 14 out of 20 amino acids are exploitable for ·OH-based labelling. While the hydroxyl radical footprinting produces many different modifications, depending on the amino acids, in most cases it results

in incorporation of one or two oxygen atoms. Interestingly, the amino acids differ, based on the source of oxygen atom, which is incorporated during the modification process[103]. They can be divided into 3 classes, depending on whether they preferentially uptake oxygen from hydrogen peroxide, dissolved oxygen, or both. An overview of possible ·OH-induced modifications is presented in Table 1.

Table 1 | Rate constants for reactions of free amino acids with ·OH and possible mass shifts of primary oxidation products[102,104].

	rate (M^{-1} s^{-1})	mass shifts
cysteine	3.5×10^{10}	-15.9772 +31.9898 +47.9847
tryptophan	1.3×10^{10}	+3.9949 +15.9949 +31.9898 +47.9847
tyrosine	1.3×10^{10}	+15.9949 +31.9898 +47.9847
methionine	8.5×10^{9}	-32.0085 +15.9949 +31.9898
phenylalanine	6.9×10^{9}	+15.9949 +31.9898 +47.9847
histidine	4.8×10^{9}	-23.0160 -22.0320 -10.0320 +4.9789 +15.9949
arginine	3.5×10^{9}	-43.0534 +13.9793 +15.9949
cystine	2.1×10^{9}	+31.9898 +47.9847
isoleucine	1.8×10^{9}	+13.9793 +15.9949
leucine	1.7×10^{9}	+13.9793 +15.9949
valine	8.5×10^{8}	+13.9793 +15.9949
proline	6.5×10^{8}	+13.9793 +15.9949
glutamine	5.4×10^{8}	+13.9793 +15.9949
threonine	5.1×10^{8}	-2.0157 +15.9949
lysine	3.5×10^{8}	+13.9793 +15.9949
serine	3.2×10^{8}	-2.0157 +15.9949
glutamate	2.3×10^{8}	-30.0106 +13.9793 +15.9949
alanine	7.7×10^{7}	+15.9949
aspartate	7.5×10^{7}	-30.0106 +15.9949
asparagine	4.9×10^{7}	+15.9949
glycine	1.7×10^{7}	–

It is worth noting that aromatic and some aliphatic amino acids exhibit the presence of FPOP structural isomers[105,106]. These amino acids can be labelled in different positions of their side chain, leading to isobaric modified peptides. These isobars exhibit different chemical properties and can be separated chromatographically (Figure 3, page 22). Although these cases bring additional insight into the solvent accessibility at the sub-residual level, they significantly

complicate the evaluation of FPOP data. To correctly assess the extent of modification of the individual residues, all isobaric forms of the modified residue should be identified and quantified.

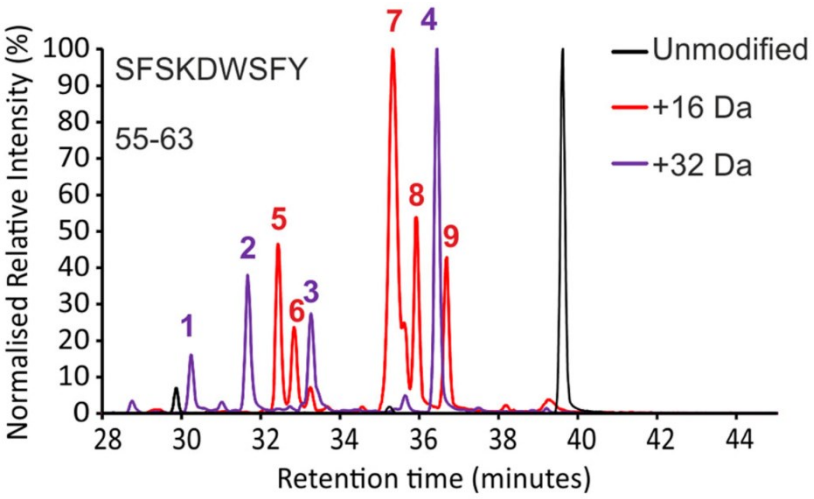


Figure 3 | Chromatographically resolved isobaric FPOP modifications of Trp60. Reprinted from[105].

Since ·OH radicals were previously used for nucleic acid footprinting, many techniques to generate the radicals directly in solution have been developed. Their fundamentals, advantages, and shortcomings are discussed in the following sections.

- Fenton chemistry

In 1894 Fenton reported on the oxidation of tartaric acid in the presence of hydrogen peroxide and Fe^{2+} salt[107], a chemistry also later described for other transition metal ions in lower oxidation states[108]. But it was not until 1932, when Haber and Weiss proposed the reaction mechanism[109]:

$$Fe^{2+} + H_2O_2 \rightarrow Fe^{3+} + OH^- + \cdot OH \quad (1)$$

$$Fe^{3+} + H_2O_2 \rightarrow Fe^{2+} + HO_2^{\bullet} + H^+ \quad (2)$$

$$H_2O_2 + \cdot OH \rightarrow HO_2^{\bullet} + H_2O \quad (3)$$

$$HO_2^{\bullet} \rightarrow O_2^{\bullet -} + H^+ \quad (4)$$

$$Fe^{3+} + HO_2^{\bullet} \rightarrow Fe^{2+} + O_2 + H^+ \quad (5)$$

$$Fe^{3+} + O_2^{\bullet -} \rightarrow Fe^{2+} + O_2 \quad (6)$$

$$Fe^{2+} + \cdot OH + H^+ \rightarrow Fe^{3+} + H_2O \quad (7)$$

However, the pH optimum of the reaction lies between 3 and 4, making it hardly compatible with the majority of biomolecules. The method was improved by the

introduction of the chelating agent ethylenediaminetetraacetic acid (EDTA), to prevent iron ions from precipitating at physiological pH, and ascorbate, which is capable of reducing Fe^{3+}–EDTA to Fe^{2+}-EDTA, allowing the reaction to start over. This approach was then widely utilised in probing nucleic acid structures[110]. To date, Fenton chemistry remains the most accessible approach to generate hydroxyl radicals, since there are no requirements for specialised instrumentation. In addition to nucleic acid footprinting, it was also used for protein labelling. For example, chelated Fe ions allow for attachment directly on the protein molecule, creating a localised probe, that generates OH radicals *in situ*[111]. While this method opens doors to intriguing applications, the reaction time exposes the protein to prolonged attacks of ·OH radicals, which could lead to structure perturbation.

- Electron pulse radiolysis

Electron pulse radiolysis employs the phenomena of ionising radiation, when high-energy electrons pass through the aqueous environment, ionising the molecules of water. This is achieved by accelerating the electrons in linear or van de Graff accelerators to energies ranging from 1 to 30 MeV. This energy is sufficient to begin the following reactions[112,113]:

$$H_2O \overset{h\nu}{\rightarrow} H_2O^{\bullet+} + e^- \quad (8)$$

$$H_2O^{\bullet+} + H_2O \rightarrow \cdot OH + H_3O^+ \quad (9)$$

$$e^- + nH_2O \rightarrow e^-_{aq} + H_2O^* \quad (10)$$

$$H_2O^* \rightarrow \cdot OH + H^+ \quad (11)$$

This method of generating hydroxyl radicals does not require hydrogen peroxide, which can react with metal-binding proteins, invoking Fenton-like chemistry, or cause alterations of the protein structure independently of oxidation[114]. Nevertheless, the need for an electron accelerator, together with safety protocols necessary for work with electrons in the γ-ray energy range, prevent mass adaptation of this technique and synchrotron water radiolysis has taken its place.

- Synchrotron water radiolysis

Synchrotrons are capable of accelerating photons to energies of 1 to 100 keV at a rate around 10^{15} photons per second[115]. This X-ray beam launches the photoelectric effect, when the photons transfer their energy to electrons upon collision. Excited electrons are escaping the valence layer of water molecules, generating radicals. The

reaction follows the same mechanism as depicted in equations 8–11, differing only in the 'energy-carriers'.

Chance and his team reported the first adaptation of protein footprinting by hydroxyl radicals generated by synchrotron water radiolysis in 2002[116], introducing the HRF on a millisecond time scale. Due to its fast reaction rate and precise control over the photon beam, this technique allows for highly reproducible results. The limiting factor is, of course, the access to the synchrotron radiation facility.

The methods to generate hydroxyl radicals discussed so far are all burdened by disadvantages. In Fenton-like chemistry the activity of radicals cannot be easily controlled, and water radiolysis by electron pulse or synchrotron require specialised facilities. Thus, a novel technical solution was needed, which would combine the simplicity of Fenton reagents with the ability to control the radical dosage, as in water radiolysis techniques.

- Laser photolysis of hydrogen peroxide

Exposing hydrogen peroxide to UV light around 253 nm causes homolytic cleavage of the molecule to two hydroxyl radicals (equation 12). The ·OH can then react with H_2O_2 molecules, that did not undergo photolysis, in subsequent chain reactions (equations 13 and 14). The competing self-quenching reaction is denoted as equation 15[(ref. 117,118)].

$$H_2O_2 \xrightarrow{h\nu} 2 \cdot OH \quad (12)$$

$$\cdot OH + H_2O_2 \rightarrow H_2O + HO_2^{\bullet} \quad (13)$$

$$HO_2^{\bullet} + H_2O_2 \rightarrow H_2O + O_2 + \cdot OH \quad (14)$$

$$2 \cdot OH \rightarrow H_2O_2 \quad (15)$$

The hydroxyl radical footprinting of proteins by UV photolysis of hydrogen peroxide was first introduced by Sharp in 2004[(ref. 119)]. The method was demonstrated on lysozyme and β-lactoglobulin A, as representatives of different secondary and tertiary structures. The extents of modification were compared to solvent-accessible surface area derived from NMR structures to a good agreement. However, this protocol operated with an H_2O_2 concentration of 15% and UV irradiation for up to 5 minutes.

In order to minimise the labelling duration and hydrogen peroxide concentration, the workflow proposed by Aye, Low and Sze in 2006[(ref. 120)] swapped the UV lamp for

Nd:YAG laser with 266 nm wavelength. Hydrogen peroxide was added from 0.3% to 1% concentrations only shortly before irradiation. This nanosecond pulse labelling delivered up to 2 mJ of energy per pulse to solutions of ubiquitin and apomyoglobin, as model proteins. The labelling reaction was stopped by flash freezing and excess H_2O_2 was removed during subsequent lyophilisation. Upon solubilisation and trypsin digestion, the samples were analysed by LC-FT-ICR MS and LC-MS/MS. The study showed that a single laser pulse suffices to introduce oxidations in solvent-accessible regions.

Concurrently, Hambly and Gross have developed similar approach[94], employing a KrF exciplex laser tuned to 248 nm. This wavelength not only is closer to ideal for H_2O_2 cleavage, but also minimises the laser absorption by proteins. Using 50 mJ per pulse, the technique allowed even lower concentrations of hydrogen peroxide (~0.04%). The capillary system, featuring a syringe flow regulation and a narrow transparent window, allowed for fine synchronisation with the laser, to achieve a single pulse per unit of volume (Figure 4). Immediately after sample collection, agarose-immobilised catalase was added to decompose any excess H_2O_2. Together with the presence of glutamine as a radical scavenger, this workflow allowed for a controlled and tuneable labelling process. This platform was later coined as Fast Photochemical Oxidation of Proteins (FPOP).

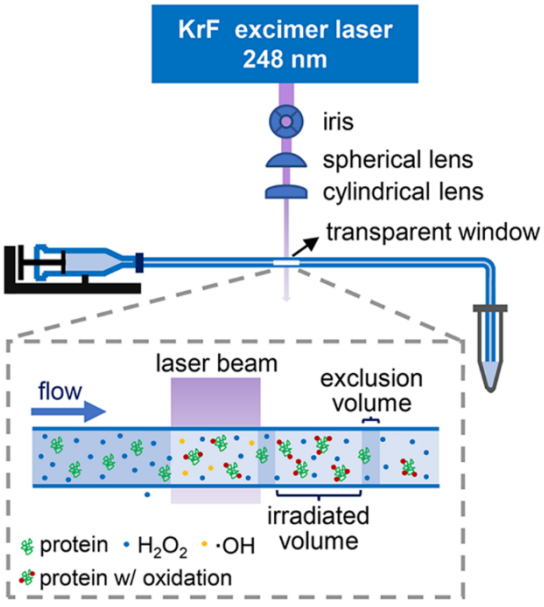
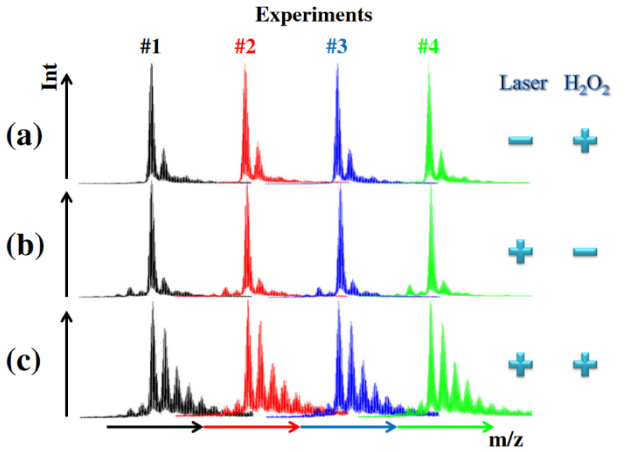


Figure 4 | Schematic representation of the FPOP platform. Reprinted from[121].

Despite the use of H_2O_2 as a precursor to hydroxyl radicals, the FPOP platform minimises the amount needed to efficiently label the biomolecules analysed, lowering

the possibility of hydrogen peroxide influence on the native structure of the protein. Later improvements introduced mixing of the protein sample with H_2O_2 solution by syringe system only seconds before laser irradiation, together with a more precise description of the experimental protocol, increasing the reproducibility[122]. The updated platform design was demonstrated on cytochrome *c*, as a case of redox-active protein, a species that can interfere with hydrogen peroxide on its own. Comparison of the FPOP analysis with two control conditions (non-laser and non-H_2O_2) proved significant differences in the extent of modification (Figure 5), and a substantial decrease of H_2O_2-induced oxidation was observed, when compared to the previous platform arrangement.



% unmodified protein for standard experiments					
Condition	Exp. #1	Exp. #2	Exp. #3	Exp. #4	Average
Non-laser (a)	61.6	61.9	60.6	59.7	61.0 ± 1.0
Non-H_2O_2 (b)	63.1	63.7	64.5	64.0	63.8 ± 0.6
FPOP (c)	26.8	25.4	26.3	26.6	26.3 ± 0.6

Figure 5 | Mass spectra of FPOP-labelled cytochrome *c* (+15 charge) obtained by using a modified FPOP platform in quadruplicate for (A) the 'non-laser' control, (B) the 'non-H_2O_2' control, and (C) standard FPOP conditions. The unmodified protein fraction (%) for each experiment is included in the table below the figure. Reprinted from[122].

Overall, FPOP retains the main advantages of tuneability, such as in synchrotron water radiolysis, while requiring less expensive and more accessible equipment. The timescale of this technique is short enough to prevent the outcome to be affected by structural perturbations caused by the ·OH radicals[102]. Furthermore, FPOP can be performed under physiological conditions and even *in vivo*[123]. It is worth noting that hydrogen peroxide can produce additional reactions that lead to overoxidation from secondary radicals[124]. Thus, adherence to strict experimental conditions is necessary to obtain reproducible and reliable results[125].

Several other marginal techniques for hydroxyl radical generation were explored, but they never reached a wider adaptation. These include high voltage electrical discharge[126], peroxonitrous acid disproportionation[127], X-ray rotating anode radiolysis[128], γ-radiation[129] and water sonolysis[130].

OTHER RADICAL LABELLING

While hydroxyl radicals are considered the most versatile and broadly applicable radical labelling probe, the untargeted footprinting toolkit offers alternative reagents, with different mechanisms, reactivity, and specificity (Figure 6), improving the versatility of reactive radical labelling.

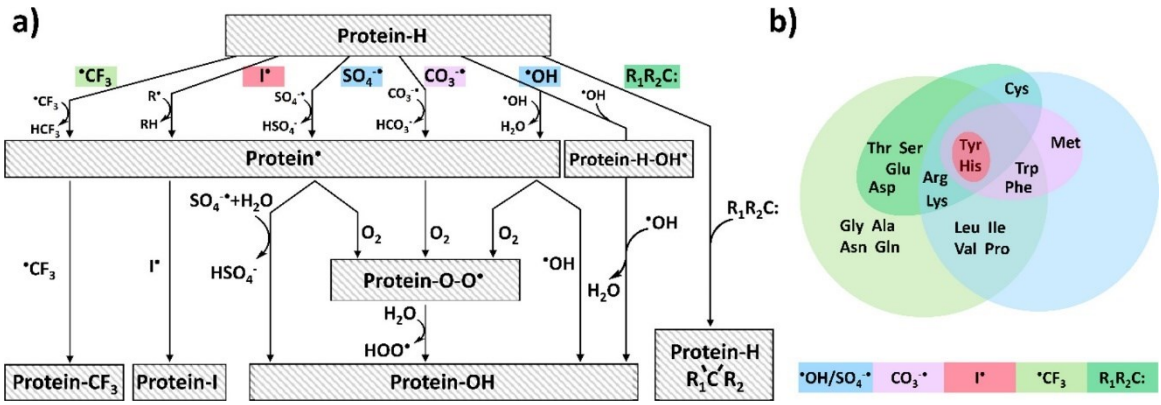


Figure 6 | Overview of untargeted radical footprinting reagents. (A) Schematic of the proposed pathways. (B) Specificity towards amino acid residues. Reprinted from[51].

- Sulphate radical anion

The sulphate radical anion ($SO_4^{-\bullet}$) is a strong oxidising agent, which (same as ·OH) has the ability to abstract hydrogen from biomolecules, inducing modification of the amino acid side chain, either by fragmentation or hydratation[95]. Furthermore, the sulphate radical anion reacts with lipid membranes and nucleic acids[131]. The first use of $SO_4^{-\bullet}$ was demonstrated in 2005 with the radical generation by microwave-assisted metal-catalysed oxidation from persulphate[132], and five years later it was introduced on the FPOP platform[95]. Although the exact reaction mechanism has not yet been fully investigated, the results imply that $SO_4^{-\bullet}$ exhibits different reactivity to amino acids than hydroxyl radical[95], and also differs in the proteins labelled when employed as in cell footprinting reagent[133], presumably due to its negative charge. These findings suggest that protein footprinting experiments may benefit from the combination of these two radical reagents.

- Carbonate radical anion

Carbonate radical anion ($CO_3^{-\bullet}$) as a footprinting agent was first demonstrated on the FPOP platform by Zhang et al.[134]. The workflow uses hydroxyl radicals from the photodissociation of hydrogen peroxide to generate carbonate radical anions from carbonate/bicarbonate buffer. $CO_3^{-\bullet}$ exhibits selectivity toward electron-rich residues with the highest reactivity with methionine and tryptophane, followed by tyrosine and moderate reactivity with histidine and phenylalanine. Granting the carbonate radical anion labels only certain amino acids, it can serve as a valuable footprinting probe for studies of protein dynamics and interactions, which involve this subset of residues.

- Carbene radicals

Carbene radicals represent highly reactive, non-specific labelling probes with very short lifetime, with methylene as the first reagent representative described in 2000(ref. 135). The radical precursors are inert to proteins, but once generated, their reactivity ensures rapid labelling, without the need for a scavenger. Mechanism-wise, carbene radicals can undergo insertion into X-H bonds and addition to C=C bonds[136]. The labelling yield of the technique was greatly improved by the employment of 'photoleucine', a diazirine-modified amino acid, activated by a 355 nm exciplex laser pulse[137]. In the study, carbene labelling proved viable for the investigation of protein-peptide interaction, with calmodulin and myosin light chain kinase peptide as the example system, and apo-/holo-calmodulin conformational changes.

- Iodine radical

Iodine is known to modify tyrosine residues in thyroid hormone production via enzymatical pathway[138], or even non-enzymatically[84], but its application in the form of radicals is rare. In 2008, photoactivated iodo-tyrosine was introduced, enabling protein backbone cleavage in the gas phase mediated by the iodine radical[139]. The cleavage was almost completely site-selective, with marginal histidine involvement observed. The direct footprint of iodine radicals was demonstrated by group around Gross in 2012, with the use of FPOP platform[140]. Iodine radicals generated by photodissociation of 4-iodobenzoic acid were used to label myoglobin, carbonic anhydrase II and insulin proteoforms. The data analysis showed extensive modifications of tyrosine and relatively minor extents of histidine iodination.

- Trifluoromethyl radical

While fluor rarely occurs in biomolecules[141], its high electronegativity allows for interesting chemistry. The trifluoromethyl radical (·CF_3) labelling was first introduced by Cheng et al.[142], inspired by the C–H fluoroalkyl functionalisation of the heterocycles reported earlier[143]. Laser-initiated labelling starts with generation of hydroxyl radical from H_2O_2, which react with $NaSO_2CF_3$, producing trifluoromethyl radical. The ·CF_3 then readily reacts with 18 of 20 amino acids, omitting cysteine and methionine. The signature +68 Da mass shift of the trifluoromethyl probe is also accompanied by minor ·OH labelling byproducts (Figure 7).

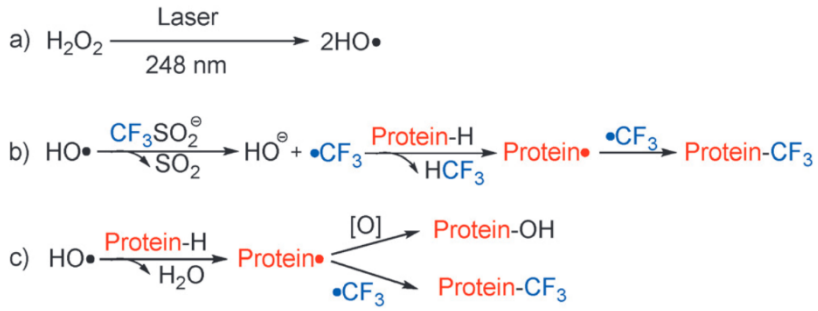


Figure 7 | Reaction mechanism of the trifluoromethyl labelling. (A) First, the hydroxyl radicals are produced by laser photodissociation of hydrogen peroxide. (B) The ·OH then generates a trifluoromethyl radical, which labels the protein. (C) Since hydroxyl radicals are present in solution, they compete with trifluoromethyl radicals. Reprinted from[142].

Interestingly, a dual labelling workflow has emerged, utilising the footprinting information of both hydroxyl and trifluoromethyl radicals[145]. This was done by carefully optimising the doses of both radicals. The main advantage of the concurrent labelling is that the covalent probes used exhibit mutually complementary amino acid reactivity. Furthermore, peptides labelled with ·CF_3 and ·OH have opposite shift in hydrophobicity, which aids in the chromatographic separation.

As a follow-up to trifluoromethyl radical labelling, the groups around Novák and Beier reported on the potential of Togni reagents as a source of radical fluoroalkylation probes[146]. The study demonstrated the preference of fluoroalkyl radicals towards aromatic amino acids (tryptophane, histidine, tyrosine, phenylalanine) and cysteine. The workflow, later coined as Fast Fluoroalkylation of Proteins (FFAP)[144], generates fluoroalkyl radicals from Togni reagents by reductive decomposition with ascorbic acid (Figure 8, page 30).

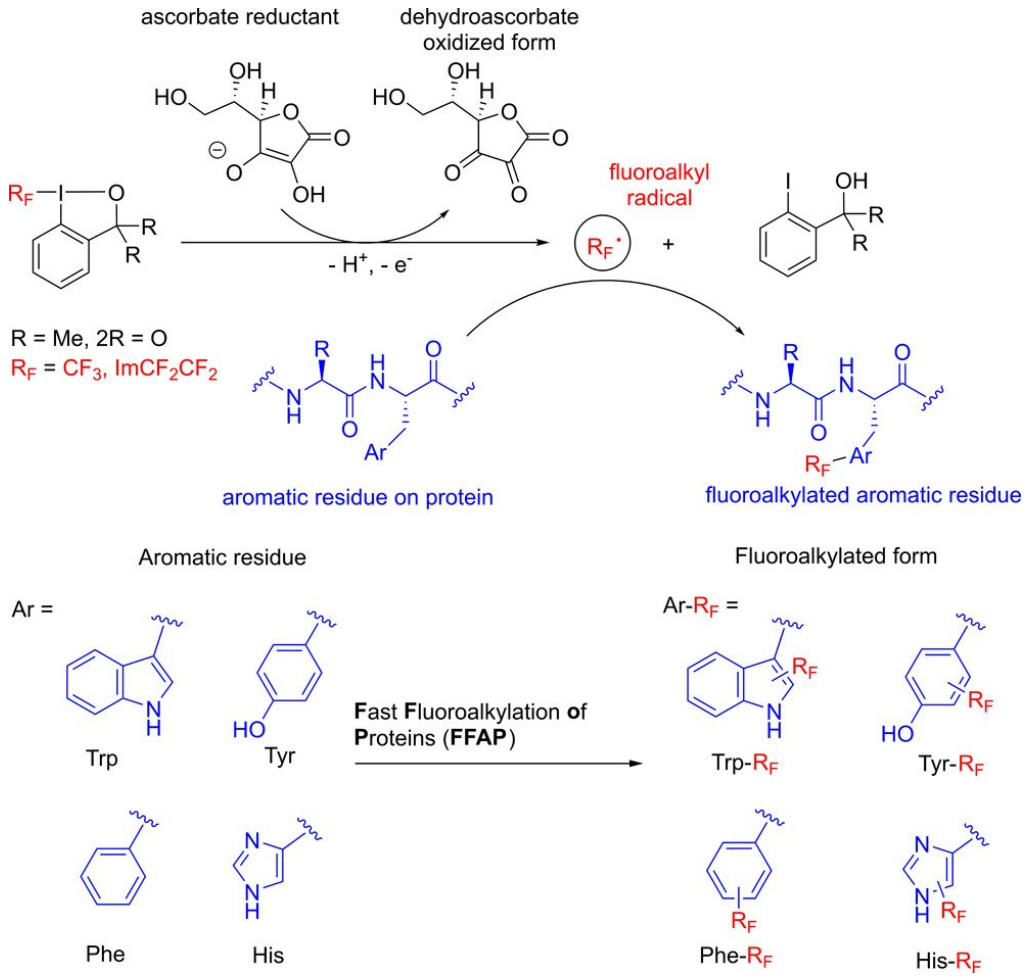


Figure 8 | Reaction scheme of Fast Fluoroalkylation of Proteins (FFAP). Reprinted from[144].

Labelling proteins with untargeted radical probes often produces complex proteomic samples, which, upon digestion, give rise to plethora of peptidoforms. At the same time, labelling experiments require sensitive and reliable identifications of the modified peptides and, furthermore, precise localisation of the modified site within the peptide. The complexity of FPOP data is additionally increased by the existence of isobaric modifications – peptides bearing the same modification, on the same residue, but on different C-atom. So far, there is no evaluation software that would address all the challenges inherent to FPOP data analysis. Possible search strategies and applicable identification workflows are discussed in the following chapter.

1.3. Identification of proteomic mass spectrometry data

Mass spectrometry has earned the flagship status among proteomics techniques, covering the entire range of analysis complexity. While the measurement of intact proteins and their subsequent fragmentation (top-down approach) allows for a more

direct assessment of proteomic samples, the analytical difficulty increases immensely with the complexity of the sample. For that reason complex protein samples are usually subjected to protein cleavage and the resulting peptide mixture is introduced to the mass spectrometer gradually (bottom-up approach). This separation is generally done by liquid chromatography. After ionisation, the *m/z* signals and intensities of the peptides are first acquired (MS1). The subjecting of the peptides to fragmentation yields tandem mass spectra (MS2 or MS/MS), from which the amino sequence and possible modification are determined. The *m/z* signals destined for fragmentation can be targeted and narrowly-isolated based on the intensity or selection made prior to analysis, an approach called data-dependent acquisition (DDA). Alternatively, a larger subset of the MS1 trace mass range can be selected, maximising the number of compounds fragmented, denoted data-independent acquisition (DIA). DDA provides clear relation between the precursor ion and its fragmentation spectrum but can miss out on important but low intensity signals, while DIA covers more of the MS1 signals, but its tandem mass spectra are typically very convoluted, which presents a great challenge in the data interpretation.

To interpret the information obtained by bottom-up mass spectrometry analysis, two main approaches – database search and library search – have been invented. Their brief history and major characteristics are discussed in the following sections.

1.3.1. Database search engines

1.3.1.1. Historical development of database search engines

The rise of database search engines in the 1990s brought a paradigm shift in the proteomic mass spectrometry data processing. Before, the MS spectra interpretation was performed fully manually or with the aid of early computer programs that still required user evaluation of the output[147,148]. In 1994, SEQUEST introduced an algorithm capable of uninterpreted peptide identification[149]. The algorithm would identify amino acid sequences from the database that match the measured precursor mass and then correlate the experimental fragmentation spectrum with the predicted MS/MS spectra of the candidates (using cross-correlation score). Following was the Mascot search engine released in 1999, which brought web-based accessibility, thread scalability, and probabilistic scoring that further advanced the peptide identification[150]. The scoring based on the MOWSE algorithm considered the peptide

mass, length, and frequency of the sequencies in the database to derive the probability that the peptide was assigned by chance.

Since then, numerous database search engines have emerged, including open-source TANDEM in 2004, as a modular platform[151], or Andromeda in 2011, with an emphasis on scalability[152], and commercial PEAKS DB in 2012, incorporating *de novo* sequencing[153], or Byonic in 2012[154], just to list a few of the most popular.

1.3.1.2. Mechanisms of database search engines

Database search engines operate by generating database from the provided protein sequencies, performing *in silico* digestion and fragmentation. The experimentally obtained MS/MS spectra are then scored against the database queries (Figure 9). The candidate peptides are selected based on the mass tolerance, specified enzyme parameters, and defined post-translational modifications. The results are typically in form of a list of peptide sequences matched to fragment spectra (peptide-spectrum match, PSM). Individual search engines differ in the steps of preprocessing of experimental spectra, such as charge deconvolution, deisotoping, centroiding, removing the assumed background noise, selecting subsets of the most intense peaks, or spectrum binning[150,152,155]. The crucial distinction lies in the scoring algorithm, which can be based on spectral correlation functions, probability, count of matched fragments, expectation value, or others[156,157].

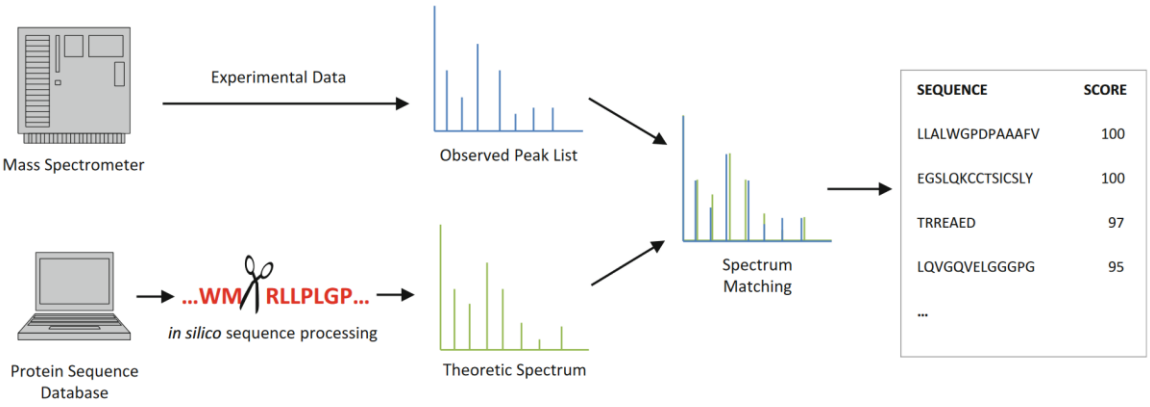


Figure 9 | Schematic illustration of the sequence database search. The experimental spectrum is compared to the generated theoretical one and the match is scored, resulting in a list of peptide-spectrum matches. Reprinted from[158].

In an attempt to control the amount of truly correct and incorrect matches (the false discovery rate, FDR), the array of results is subjected to statistical estimation of confidence and error rates[159]. This can be done by either target-decoy approach or the empirical Bayes approach. Target-decoy searching assumes that target and decoy

databases are mutually exclusive, so that decoy hits are not excluding true hits, and that matches to decoy and peptide false positives are equally likely[160]. The list of PSMs is then subjected to a score cut-off, so that the FDR is below the desired threshold. While being easy to implement, this approach prolongs the search time and raises questions as to whether the reversed or randomised sequences acting as decoys truly represent false peptide matches. The empirical Bayes approach uses the *q*-value, as a measure of significance in terms of FDR, as opposed to *p*-value, which is a measure of significance in terms of the false positive ratio[161]. The algorithm models the most likely distributions of correct and incorrect identifications, fits them to the experimental data, and derives the score cut-off. However, the use of parametric assumptions and set coefficients in the modelling can sometimes fail to provide the correct population estimation.

1.3.1.3. Application of database search in covalent labelling MS

The main advantage of the database search engine is its adaptability to different use cases via parametrisation. The option to modulate the search discriminant factors, such as mass tolerance, together with defining variable modifications and their specificity allows for precise control over the search space. This is particularly beneficial in analyses of untargeted labelling experiments. Furthermore, the ability to identify unpredicted peptides by relaxing search parameters ('open search') enables the discovery of unknown peptidoforms, possibly leading to an improved informational yield of such experiments.

However, a large search space directly increases the need for computational power, due to the vast possibilities that need to be evaluated. These possibilities grow with each defined modification-residue combination, allowed enzyme misscleavage, and relaxed mass tolerance. Additionally, with a more extensive list of candidate peptides, the probability of high scoring incorrect match increases. This can lead to a higher incidence of false identifications, impairing the subsequent reliability assessment in the form of FDR filtering. Therefore, the evaluation of the database search results of untargeted footprinting analyses should be performed with caution, considering the diversity of protein forms in the sample. It is also necessary to note that because of the nature of data-independent acquisition, the fragmentation spectra measured in this mode do not have direct relation to specific *m/z* values and thus are not amenable to conventional database search. Several algorithmic solutions have been developed to transform the DIA data into pseudo-DDA form, which enables the

utilization of database search engines (e. g., DIA-Umpire[162], Group-DIA[163]), or attempt direct identification (e. g. PECAN[164], PEAKS DeepNovo-DIA[165]). However, these solutions are outperformed by library-based approaches[166–168]. Considering the search space complexity inherent to untargeted labelling, compound identification directly from DIA data presents a challenge that is immensely demanding on computational resources.

1.3.2. Spectral library search

1.3.2.1. Historical development of library search engines

The development of spectral library search engines was driven by the need for more efficient and sensitive peptide identification compared to database search. As a concept, spectral library search appeared in the 1970s with the development of mass spectra reference system, allowing for peak and intensity search among 8124 reference spectra[169]. Breakthrough came later on with the leap advancement in mass spectrometry and computational methods. In the 2000s, several research groups worked on algorithms for spectral library search of mass spectrometry data, leading to the introduction of multiple search engines in 2006. That includes Bibliospec, reporting spectrum comparability across instruments and laboratories[170], X!Hunter, as a counterpart to the database search engine X!Tandem[171], and SpectraST, which gained a broad user base and was implemented in the Trans Proteomic Pipeline[172]. During the same time, the initial collaborative spectral libraries were being built, providing the first reference databases[173–175].

1.3.2.2. Mechanisms of spectral library search engines

In contrast to database search engines, which generate theoretical spectra from protein sequence databases, library search engines utilise curated real-case spectra as a reference. The first step is the library construction, which still relies on the database search identification of the compounds. Ideally, the spectra originate from a well-characterised sample, and the peptide-spectrum matches are thoroughly validated. Spectral library entry then consists of the reference fragmentation spectrum and the corresponding annotation, including the peptide sequence, precursor ion charge, modifications, protein, retention time, or ion mobility. Second, the experimental spectra are matched against the reference in the library, assessing the similarity of fragmentation peaks alignment and their intensity, so called

spectrum-spectrum matches (SSMs). This similarity is often calculated through the dot product[172] or later introduced probabilistic scoring[176].

The reliability of the assignments is validated with methods similar to those used in the database search. In the case of the target-decoy approach, the decoy spectra need to resemble the real data, so the search engine views them as equal. The first implementation used shuffling of real target spectra so that they correspond to fragments expected from the shuffled sequence[177]. Alternatively, the decoys can be generated by simple swapping of the precursors, without the need for ion-type control, unannotated peaks interference, or additional need for computational resources[178].

Due to the immense variability of the real-case fragmentation spectra (both diverse peak position and intensity), the modelling of probabilistic distributions from the result data using the empirical Bayes approach poses as a challenging task. Therefore, the spectral profile must first be standardised, allowing for parametric estimation of the two populations - 'top-scoring incorrect' matches following uniform distribution and 'top-scoring correct' matches acting as a probabilistic outliers[179]. This approach allows for spectral library search results validation even in cases where the target-decoy approach cannot be employed.

1.3.2.3. Application of spectral library search in covalent labelling MS

Since the spectral library search utilises validated real-case spectra, they are superior in the accuracy of peptide identification[180]. While the theoretical fragmentation patterns generated by database search engines work only with the fragment positions, spectral library queries capture the full complexity of individual peptides and their unique fragmentation patterns. This is especially significant in the case of labelling experiments, since the modification that is introduced into the peptide sequence may alter the distribution of ions upon fragmentation. Spectral library search further benefits from the significantly improved speed owning to simple indexation of candidate peptides, without the need for theoretical spectra generation, and to the reduced search space, containing only previously observed candidates (and optionally decoys). Since the spectral library formation is controlled, it allows for inclusion of low abundance compounds that could be overlooked by the conventional database search approach. This is particularly important for labelling analyses, where the detection of modified peptides with low abundance can still prove relevant to the investigation of the protein structure and dynamic. Needless to say, once the spectral

library is completed, it can deliver reproducible performance when employed in analogous use cases. Moreover, as mentioned above, spectral libraries proved valuable in identifying compounds from the DIA proteomic data, with software tools like OpenSWATH[181], Skyline[182], EncyclopeDIA[183], or Spectronaut[184].

While the reduced search space improves search speed and sensitivity, it also reduces the ability to gain broader insight. The most limiting factor of spectral libraries is the dependence on the reference data set completeness when the missing compounds are simply not detected. Furthermore, since the approach uses real-case spectra, the quality of the identification framework stands and falls with the quality of the reference queries. Individual fragmentation patterns need to be clear, contaminant-free, and acquired with sufficient resolution. Additionally, the validation process during the building of the spectral library necessitates thorough quality control, as oversights can ultimately thwart the search results.

Both database and spectral library search engines play a vital role in the identification of proteomic mass spectrometry data, each with their own advantages and limitations. Database search engines are adaptable to different conditions and capable of identifying novel peptidoforms, making them appropriate for analysing MS data in discovery mode. However, their algorithms based on theoretical spectra and result processing sometimes lead to a lowered sensitivity toward low abundant peptidoforms. Spectral library search engines excel in identification accuracy and confidence while requiring less computational time. Furthermore, the set of reference spectra can be tailored to encompass all the relevant peptidoforms. Nevertheless, these benefits come at the expense of possibly limited scope and versatility, two factors which are often needed for the full utilisation of labelling experiments.

Ideally, a combined approach would fuse both approaches into a suite tailored for covalent labelling data sets. Such method then leverages the database search to explore the full landscape of potential peptidoforms derived from labelling, while utilising the rapid and highly confident identification of the spectral library search later on. While a complementary approach has been reported, demonstrating improvements in both identification and quantification reproducibility[185], the design has not yet seen widespread adaptation.

2. Aims of the thesis

The aims of this thesis were to contribute to the development of an evaluation workflow for mass spectrometric data obtained by analysis of proteins modified by reactive oxygen species. The objective was to improve the versatility of the technique and to aid in its widespread utilisation.

Specific aims were to:

- validate and compare different identification algorithms;
- investigate and assess modes of acquisition of the FPOP data;
- design and test a novel approach to acquire and evaluate FPOP data.

3. Methods

The publications included in this thesis contain a complete description of all the methods and experimental procedures used. The description is detailed enough to allow their reproduction. Therefore, this chapter consists only of a list of experimental techniques that were used.

3.1. List of experimental methods used

- Fast Photochemical Oxidation of Proteins and Fast Fluoroalkylation of Proteins
- bottom-up mass spectrometry analysis of modified proteins in data-dependent, data-independent, and MS1-only acquisition modes (timsTOF SCP)
- evaluation of FPOP data by Mascot, PEAKS, and FragPipe search engines
- comparison of search results using Principal Component Analysis, one-ANOVA, Tukey Honest Significant Difference post hoc test
- processing and evaluation of MS data using Python and R scripts

4. Results and Discussion

4.1. Publication I

With FPOP's ability to label majority of amino acid side chains (14 out of 20), the exact position of the modification in the peptide must be derived not only from its specificity but mainly from experimental evidence – fragmentation coverage of the modification site. Since the identification of modified peptides is significantly influenced by the search engine employed[186], the first part of the work was aimed at the evaluation of different identification workflows. Two search engines – Mascot and PEAKS – were chosen for this comparison, since they provide evaluation of modification assignment confidence (MD-score and AScore, respectively), which can be used for further evaluation of the identification reliability.

The search of FPOP analyses of haemoglobin (Hb), haptoglobin (Hp), and HbHp complex were conducted for each of three replicates separately and the ratios of mutual identifications (IDs) were established for both search engines (Figure 10, page 40). From the point of identifications consistently occurring in all three replicates, Mascot outperformed PEAKS with an average of 26% of mutual IDs, compared to an average of 22% in case of PEAKS. This ratio was similar between the Hb, Hp and HbHp analyses. Operating with only mutual identifications, Mascot's search yielded more IDs (in terms of absolute count) in the case of Hb and HbHp samples. The intersection of consistently identified peptides between Mascot and PEAKS revealed that only around 30% of the IDs were shared. Since Mascot is limited during its search in the number of defined variable modifications to 9, Cys trioxidation, His → Asn and His → Asp were only defined for the PEAKS search. Although these modifications represented around a third of the exclusive identifications, removing them from the list of defined variable modifications did not result in more mutual identifications with Mascot. This means that the comparison was not affected by different "starting positions", but variability is caused by differences in the identification logic and algorithms of search engines. Especially in the case of ambiguous peptides, each search engine could propose a different site of modification. Next, the distribution of modification types was compared between Mascot and PEAKS for the HbHp sample (Table 2, page 40). Oxidations, carbonyls, and His +5 Da modifications were more abundant in the Mascot results.

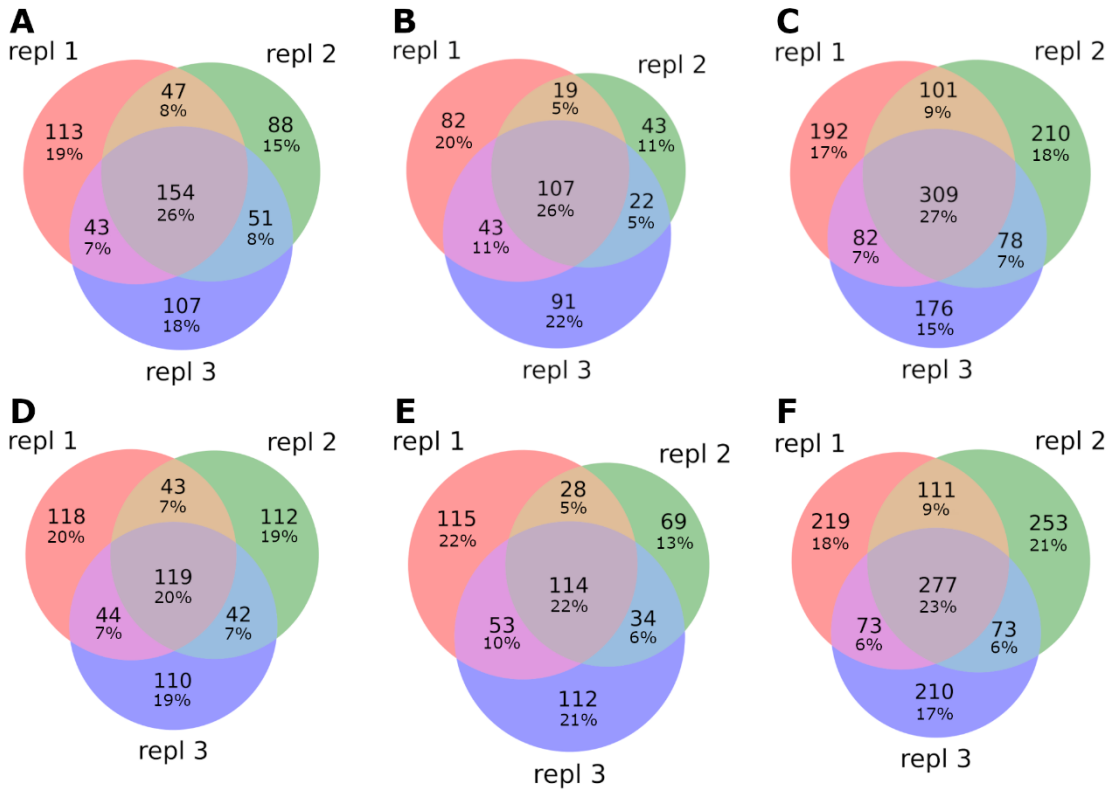


Figure 10 | Intersections of IDs between replicates for Mascot searches of (A) haemoglobin; (B) haptoglobin; (C) HbHp complex samples; and for PEAKS searches of (D) haemoglobin; (E) haptoglobin; (F) HbHp complex samples. Overall, Mascot exhibits higher consistency within replicates.

Table 2 | Types of modifications detected by Mascot and PEAKS in the HbHp complex.

modification	**Mascot**		**PEAKS**	
	count	percentage	count	percentage
oxidation	144	47%	112	40%
dioxidation	36	12%	30	11%
carbonyl	61	20%	42	15%
decarboxylation	22	7%	21	8%
Met -32Da	7	2%	4	1%
His +5Da	19	6%	9	3%
His -10Da	16	5%	10	4%
Arg deguanidination	4	1%	5	2%
His->Asp*	NA		13	5%
His->Asn*	NA		14	5%
trioxidation*	NA		17	6%

* only defined for PEAKS search

In order to introduce a straightforward FPOP data evaluation pipeline, we looked for a metric to estimate the identification reliability from available measures. We opted

out for Principal Component Analysis (PCA) with PSM score, probability of correct modification site assignment (MD-score for Mascot, AScore for PEAKS), and precursor intensity as variables (further referred to as Score, Site/logAScore, and Intensity, respectively). The PCA carried out using R packages yielded a three-dimensional visualisation, where the modified peptides distributed into discrete clusters (Figure 11). This separation was observed in all samples (Hb, Hp, and HbHp) for both search engines.

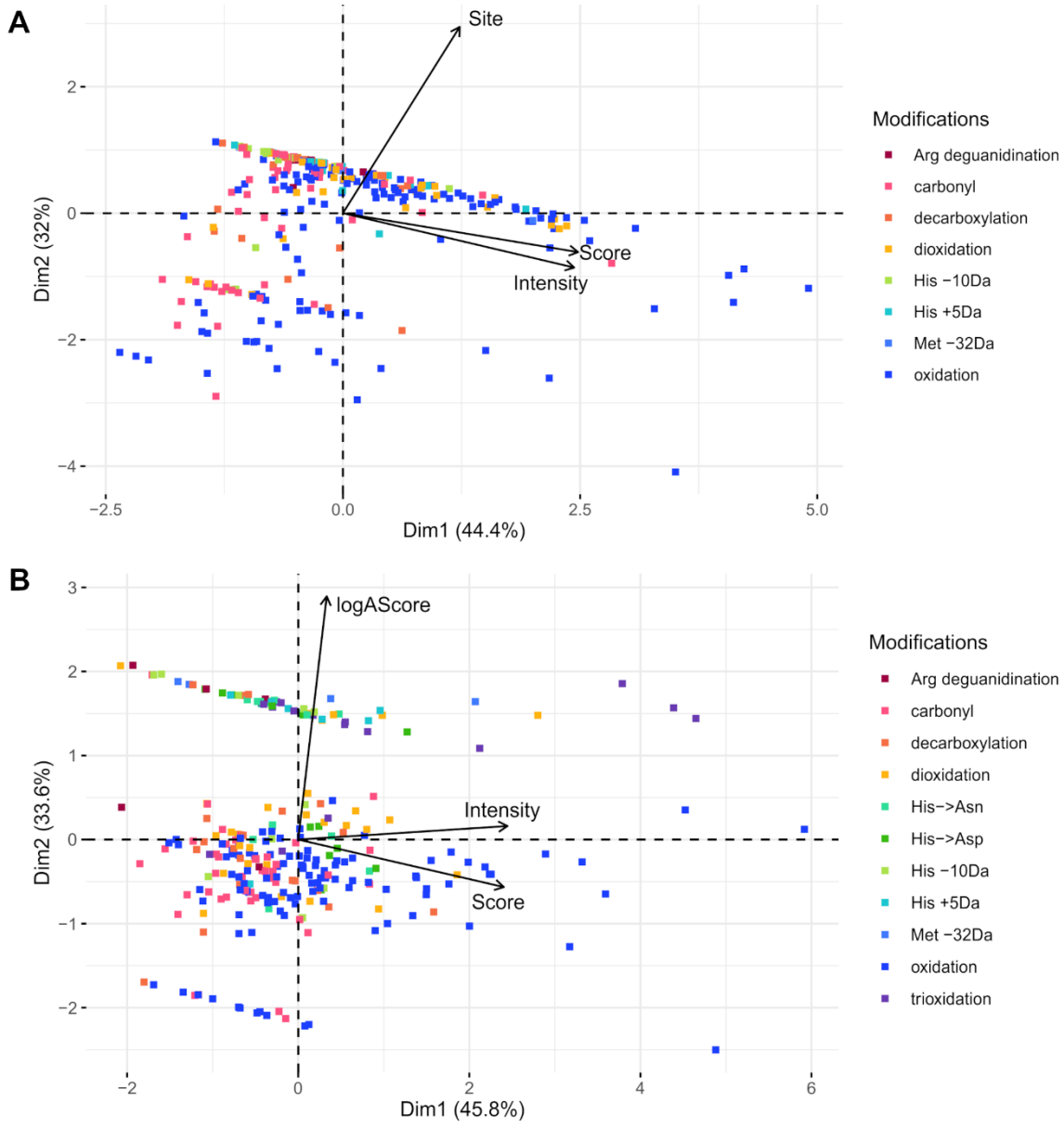


Figure 11 | Principal component analysis of the modified peptides identified in the HbHp complex sample by (A) Mascot and (B) PEAKS. The plot of colour-coded modifications shows their distribution within dimensions defined by the PCA. The vectors show the correlation of the variables with the PCA dimensions. The logAScore variable represents the PEAKS' probability of Site determination on logarithmic scale for linearisation. In both cases, the IDs are distributed into three clusters. For Mascot, most of the IDs are in the top cluster, whereas for PEAKS, the majority of the IDs are in the middle cluster.

The input variables are shown in the form of vectors to visualise their correlation with the PCA dimensions. For both Mascot and PEAKS, the Score and Intensity were positively correlated with the first dimension and showed a mutually opposing trend in the third dimension. Since the first dimension of the PCA explains the most variability, the analyses show that Score and Intensity correlate with each other, but not with the confidence in the determination of the modification site (Site/logAScore). However, it was the Site/logAScore vector along which the clusters of identifications were separated.

In the PCA of the HbHp complex, most of the Mascot assigned modified peptides were located in the top of three clusters, suggesting that Mascot identifies the modifications with high confidence. Manual validation confirmed that the oxidations, carbonyls, His -10 Da, and His +5 Da were supported by sufficient fragmentation coverage, confirming the assessment provided by Mascot. The accumulation of IDs in the middle area indicates lower certainty in Site determination, arising from incomplete fragmentation coverage, leaving two or three possible options for the modification localisation. Finally, few data points collocating at the bottom of the graph represent IDs that lack fragmentation coverage needed for at least approximate Site localisation. Altogether, the reliability of Mascot identifications highly corresponds to the distribution within PCA.

The top cluster of identifications obtained by PEAKS consisted exclusively of modifications with the highest possible confidence. Detailed inspection revealed that the modifications were assigned solely by their definition, which means that the peptides contained only one possible position for the modification. Thus, the assignment of the ultimate Site score did not correspond to the quality of the fragmentation spectra. The majority of the PEAKS IDs grouped around the middle area of the graph without providing a clear correlation between the fragmentation coverage and the Site estimation (AScore). The third cluster at the bottom comprised of identifications with AScore = 0, despite the fragmentation coverage. Overall, the separation of PEAKS IDs into cluster provided no clear indication on the reliability of the assignment of FPOP modifications.

To compare the quality of the assessment of Mascot and PEAKS, four IDs that occur in the results of both search engines were selected, based on their PSM scores. First, Met15 oxidation in the VADALTNAVAHVDDMPNALSALSDLHAHK haemoglobin peptide in its triply charged form at m/z 1004.84 and retention time (RT) 16.61 min.

Since methionine oxidation is very common and extensive modification, both search engines should assign this modification with high confidence. Mascot identification was based on almost complete sequence coverage of the modified peptide by fragment ions (Figure 12A), which lead to a very high PSM Score of 153.34 and 100% confidence in the Site determination. While PEAKS also assigned numerous fragment ions (Figure 12B), it missed y14 and b18 ions, which are among the most intense ions in the spectrum. Although this peptide-spectrum match was evaluated at Score 105.68, the confidence in the modification position was low (AScore 11.12), locating this identification in the lower area of the middle PCA cluster.

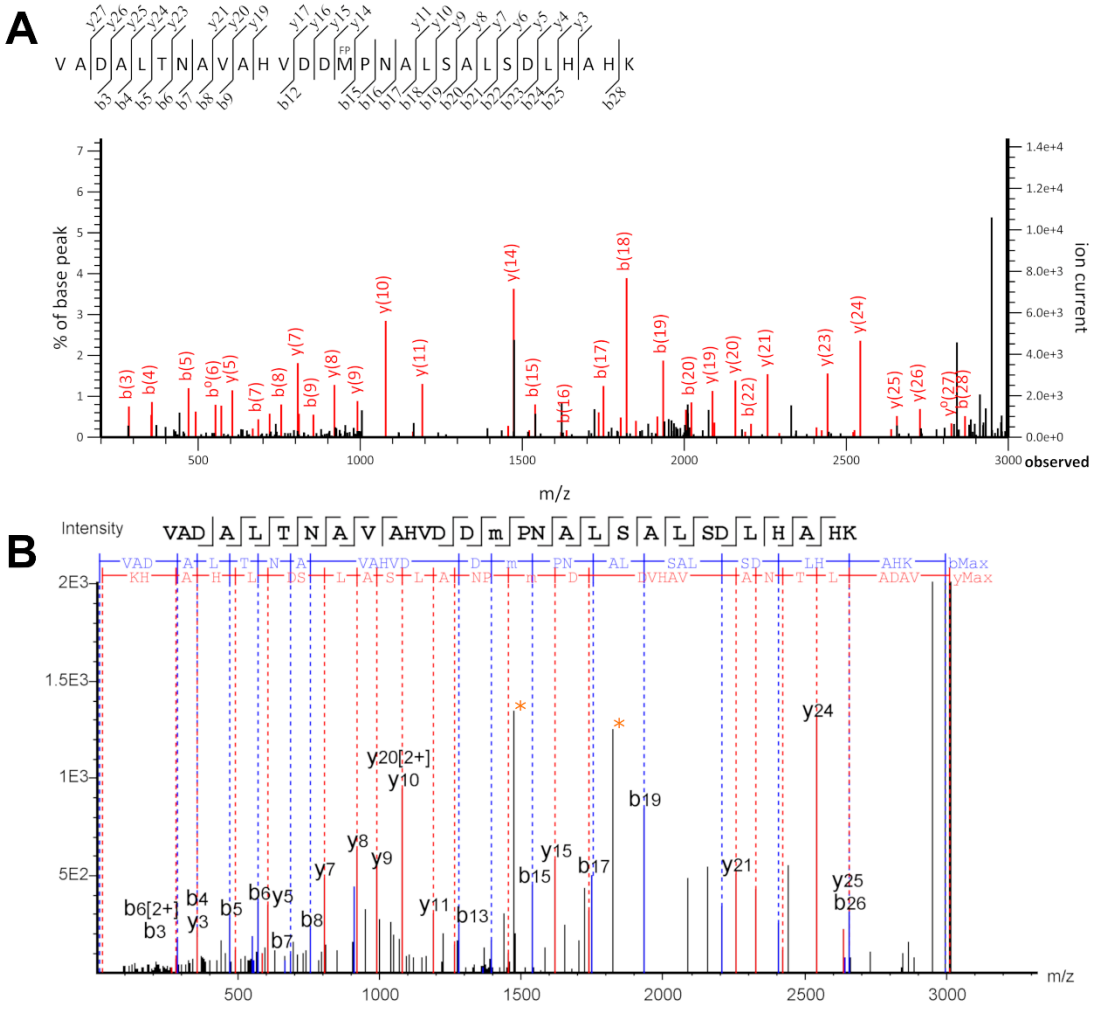


Figure 12 | Fragmentation spectra of Met15 oxidation (+16 Da) in the haemoglobin peptide VADALTNAVAHVDDMPNALSALSDLHAHK in its triply charged form at *m/z* 1004.84 and RT 16.61 min from the sample of the HbHp complex identified by (A) Mascot, which localised the modification with 100% confidence, and (B) PEAKS, which did not assign the y14 and b18 ions (labelled *) and evaluated the site confidence low (AScore 11.12).

The second selected modification is His11 carbonyl in the haemoglobin peptide VVAGVANALAHK in its doubly charged form at *m/z* 582.33 and RT 9.44 min.

Carbonyls represent low-abundance modifications, usually with a lower intensity. The assignment of modification in Mascot result (Score 32.1) is based on fragmentation ions b4, y5 and y7–y10 (Figure 13A), which do not suffice for exact Site determination, only narrow the possible positions to Leu9 or His11. Mascot leans towards the latter with 49.82% certainty, positioning this ID in the middle cluster. The identification of PEAKS is based on y-series fragment ions, with b-ions nearing the noise level (Figure 13B). Even with the assignment of almost all possible fragment ions and a medium-level Score of 71.05, the modification localisation confidence remains at AScore 14.04 and location of this ID in the lower part of the middle cluster, same as in the previous example.

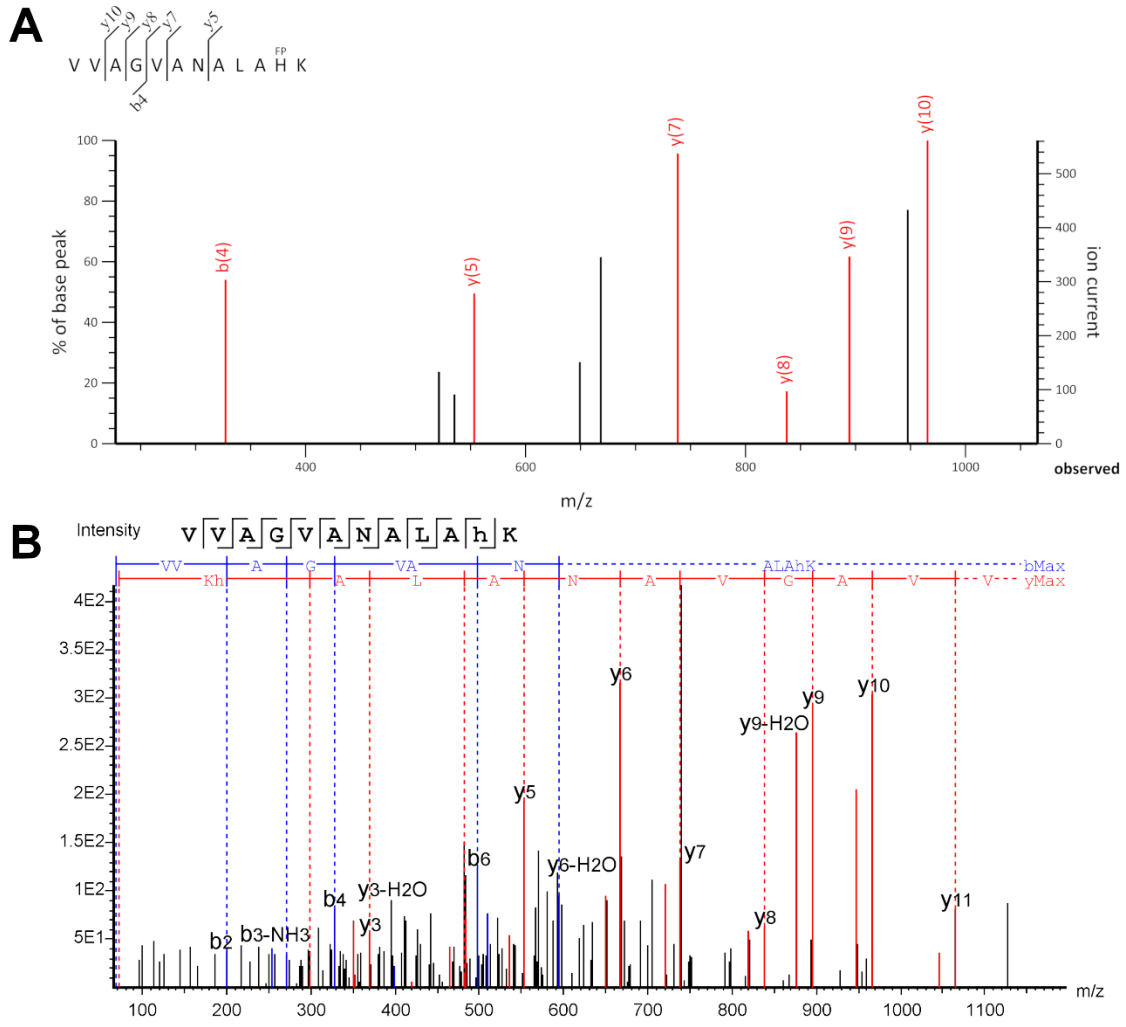


Figure 13 | Fragmentation spectra of His11 carbonyl (+14 Da) from the haemoglobin peptide VVAGVANALAHK in its doubly charged form at *m/z* 582.33 and RT 9.44 min from the sample of the HbHp complex identified by (A) Mascot (Site 49.82%), and (B) PEAKS (AScore 14.04). PEAKS shows better fragmentation coverage than Mascot, yet the certainty of site determination is low.

Third, Val5 oxidation in the VVAGVANALAHK peptide in its doubly charged form at *m/z* 583.34 at RT 5.9 min. This modified peptide represents identification which was assigned high Mascot Score (102.79), and lower PSM score in the case of PEAKS (60.44). The fragmentation spectrum for Mascot identification consists solely of y-ion fragments, missing only y1 and y8 (Figure 14A). Since FPOP oxidation is not defined for glycine, Mascot assigned the modification to valine with 99.99% confidence, resulting in ID localisation in the top cluster of PCA. The fragmentation spectrum of PEAKS identification (Figure 14B) is missing only ions y1, b1, b7 and b10, which leads to a confident localisation of the oxidation (AScore 26.57), and middle PCA cluster. Thus, this ID provides a reliable determination of the modification site, regardless of the lower PSM score.

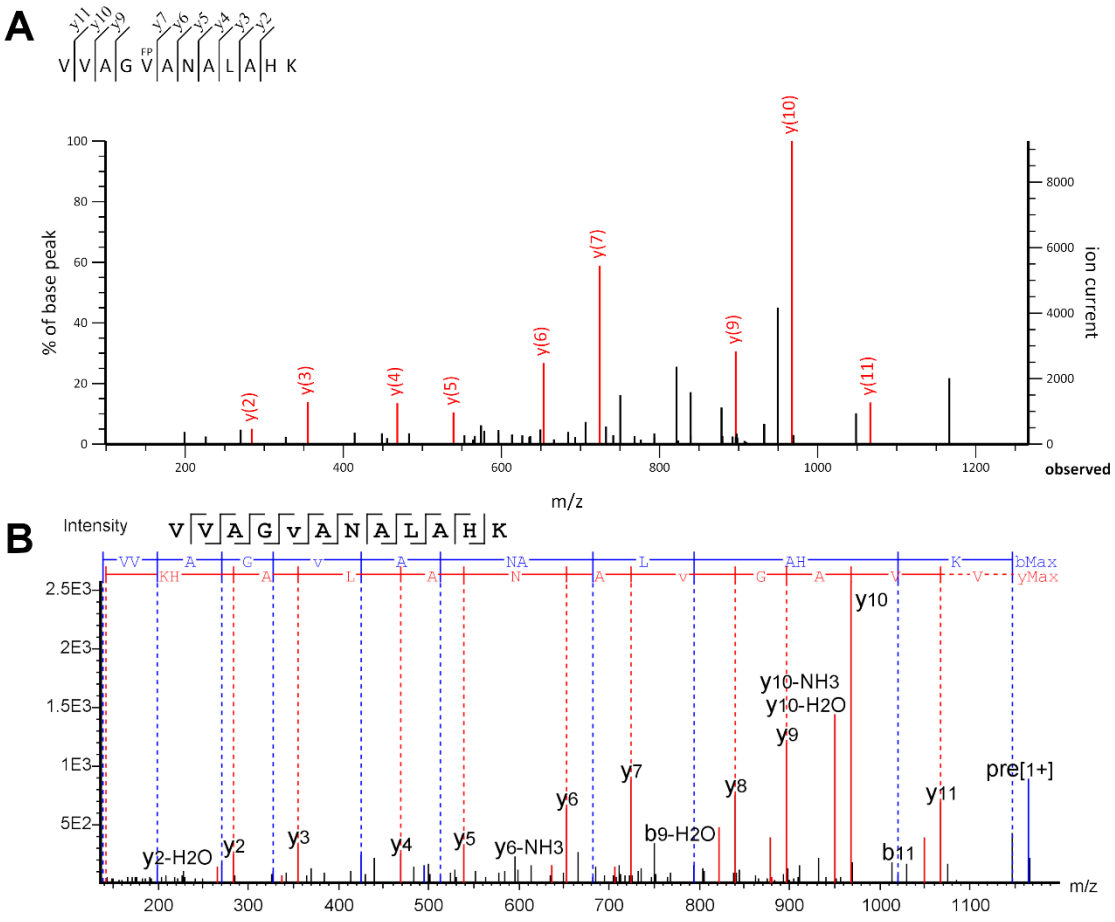


Figure 14 | Fragmentation spectra of Val5 oxidation (+16 Da) in the haemoglobin peptide VVAGVANALAHK in its doubly charged form at *m/z* 583.34 at RT 5.9 min from the sample of the HbHp complex identified by (A) Mascot (Site: 99.99%), and (B) PEAKS (AScore: 26.57).

Lastly, the Glu5 decarboxylation in the haptoglobin peptide KQLVEIEK in its mono charged form at *m/z* 956.58 at RT 5.43 min is an ID with a low Score from both Mascot and PEAKS (23.72 and 50.06, respectively). Mascot's determination of the

modification site relies on low intensity b6-H_20 and b6-NH_3 ions (Figure 15A), leading to estimation of 68.98% confidence and localisation on the lower border of the top PCA cluster. The PEAKS fragmentation spectrum consists of 11 out of 14 possible fragments, but several of them have a very low intensity (Figure 15B). However, the resulting AScore is 24.32, signalling reliable confidence and the middle PCA cluster.

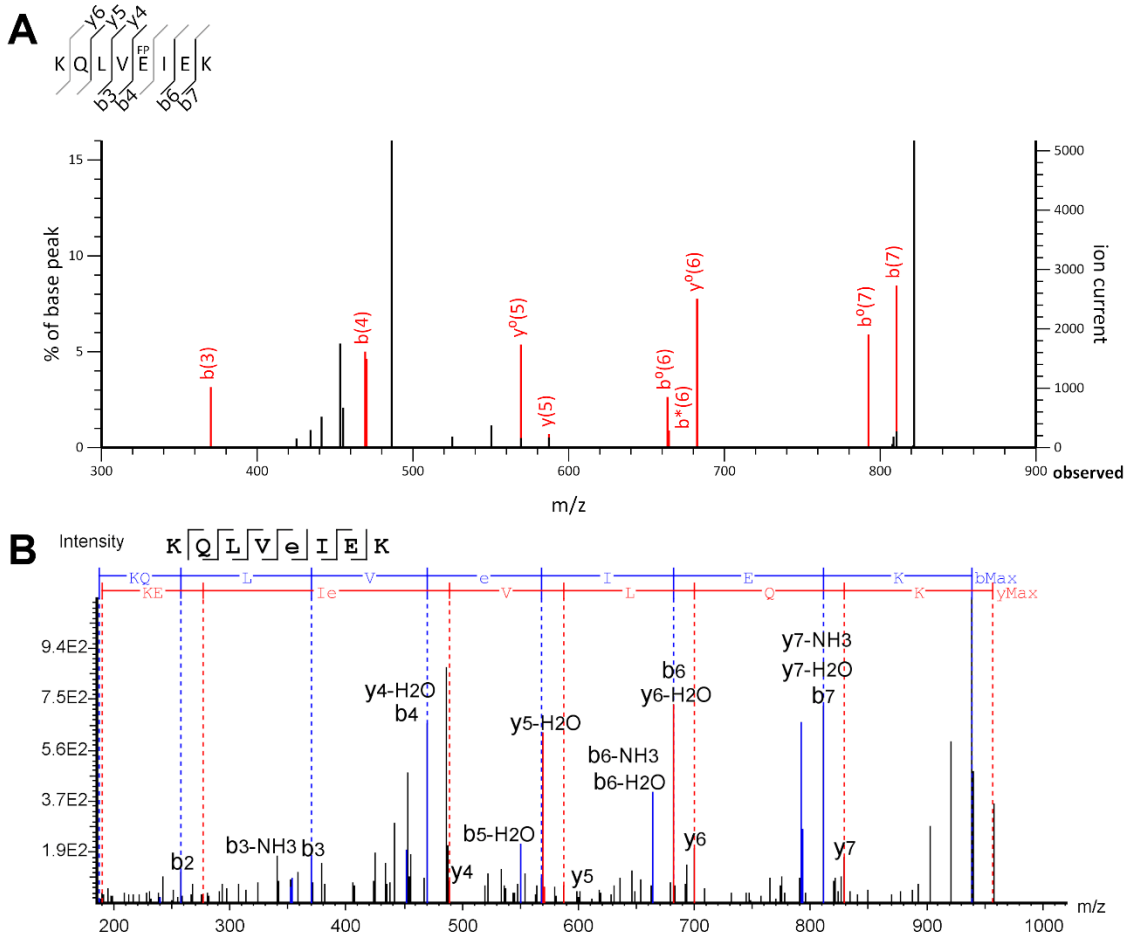


Figure 15 | Fragmentation spectra of the Glu5 decarboxylation (-30 Da) in the haptoglobin peptide KQLVEIEK in its monocharged form at m/z 956.58 at RT 5.43 min from the HbHp complex sample as identified by (A) Mascot (Site: 68.98%), and (B) PEAKS (AScore: 24.3). Even though only low intensity ions differentiate between the Glu5 and Glu7 positions, both search engines are assigning confident site determination.

This work demonstrated that the results obtained from the FPOP analysis are highly dependent on the identification framework selected, with the identifications provided by Mascot and PEAKS only partially overlapping. Interestingly, Mascot exhibited higher consistency of identifications among replicates, compared to PEAKS. The use of PCA as a tool for determining the reliability of modification localisation proved viable in the case of Mascot. The comparison of mutual IDs for Mascot and PEAKS showed that the site localisation confidence estimations provided by Mascot

exhibit a clearer relation to their respective fragmentation spectra. PEAKS, however, does not provide a direct link between the fragmentation coverage of the modification site and the confidence estimation. Visualisation by PCA could be used in dedicated FPOP evaluation software for quick estimation of the quality of the identifications. On the contrary, the clustering of PEAKS identifications does not clearly correspond to identification reliability, necessitating manual validation for most of the results.

4.2. Publication II

While the automated approach in FPOP data evaluation is robust and provides fast assessment, the results obtained by this technique are so intricate that it would undoubtedly benefit from the complex picture provided by data independent acquisition. However, since the isolation window of the DIA duty cycle allows for multiple peptides to be co-fragmented, the interpretation of the fragmentation spectra of such heavily modified compounds is not yet feasible. The proposed workflow utilises the additional information (that DIA brings over DDA) for validation purposes. Moreover, quantification of the extent of modification from the MS1 trace of DIA data exhibited the lowest variability.

Using the same model of HbHp complex, an FPOP experiment was measured in DDA, DIA, and MS1-only mode on the timsTOF SCP. Two open-source software were selected for this workflow – FragPipe[187] for the purpose of DDA data search, and Skyline[182] as a tool for spectral library building and validation. Skyline conveniently allows to import both timsTOF raw data and FragPipe *interact.pep.xml* result files directly, extracting information about defined modifications and score cut-off corresponding to defined false discovery rate. During this workflow the latter is disregarded, because the variability among FPOP modifications results in several distinct categories, each of which would require its own FDR estimation[188]. Skyline allows to export a list of individual fragments contained in the library in tabular form, enabling automated processing with, for example, Python. During processing of this data set, a prototype Python script was used to evaluate the fragmentation coverage of the modification site of individual identifications. Thus, in case of multiple identifications originating from the same chromatographic peak, the IDs with insufficient fragmentation coverage of the modification site could be disregarded.

Skyline further calculates the similarity of the isotopic profile of the precursor peptide, which proves useful for detecting cases such as when the carbonyl (+14 Da) modification is incorrectly assigned to a peptide that, in fact, bears an oxidation

(+16 Da). For instance, the haptoglobin peptide AVGDKLPECEAVCGKPK, detected at 7.8 min in its triple charged form at both *m/z* 624.6397 and 625.3113, indicates the presence of both +14 and +16 Da forms (Figure 16), with both of these modifications assigned to Leu6. Consonant with the experience with FPOP data, peptides modified by both carbonyl and oxidation on the same amino acid residue are usually chromatographically resolved.

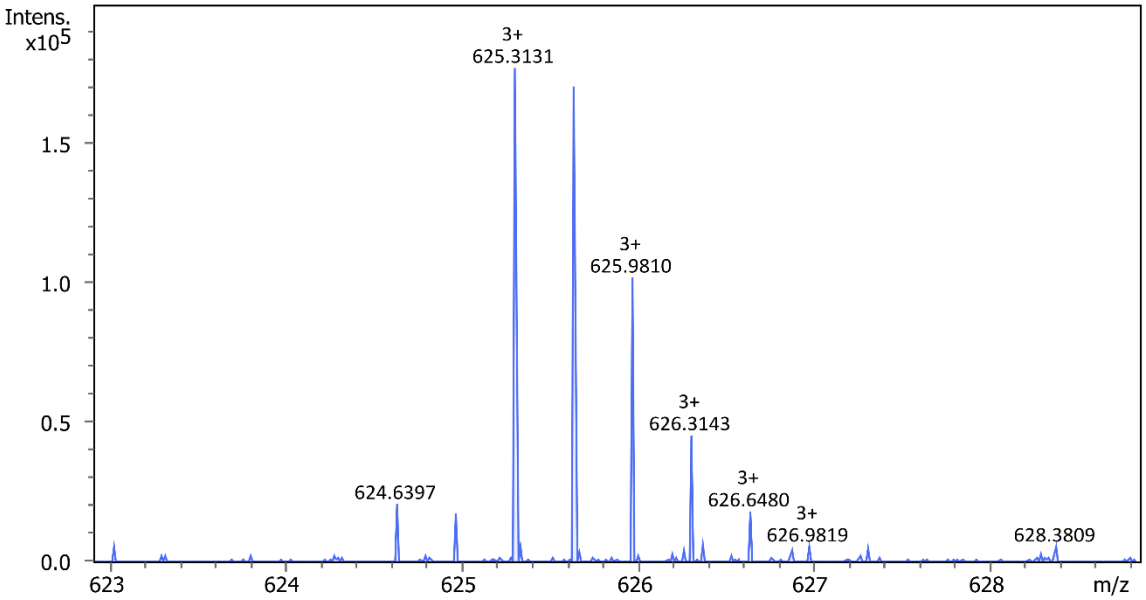


Figure 16 | An example of seemingly coeluting carbonyl and oxidation modifications in the haptoglobin peptide AVGDKLPECEAVCGKPK in its triply charged form at RT 7.8 min. The carbonyl modification (+14 Da) corresponds to *m/z* 624.6397, and the oxidation (+16 Da) corresponds to *m/z* 625.3131.

The extracted ion mobilogram (Figure 17, page 49) shows two peptide forms resolved by ion mobility ($1/K_0$). The first at $1/K_0$=0.8904 corresponding to the +16 Da modification, and the second peak at $1/K_0$=0.9237 composing of a mixture of +14 and +16 Da modifications. This means that the hypothetical carbonyl modification does not exhibit a different collision cross-section than the oxidation. Upon closer examination of the fragmentation spectra (Figure 18, page 50), it was revealed that both of these forms bear a +16 Da modification on the same residue (Leu6). Both the type and site of modification are unambiguously assigned by b5 ion at *m/z* 471.257±0.001, b6 ion at *m/z* 600.337±0.001 and the doubly charged y11 ion at *m/z* 637.797±0.001 (Figure 18A, B and C, respectively, page 50). This rules out the presence of carbonyl modified form.

It is worth noting the presence of a triply charged precursor ion with neutral loss of water at *m/z* 619.309±0.002 (Figure 18C, page 50). While it was detected in both

fragmentation spectra, it is more abundant in the spectrum of the second compound, supporting the evidence of two different peptide forms. Furthermore, this lability of the second peptidoform corresponds to the mixed signal of +14 and +16 Da precursors in the MS1 trace resolved by ion mobility. This indicates that different positions of oxidative modification on the aliphatic chain exhibit different chemical behaviour. Although the existence of isobaric FPOP modifications has already been observed, so far, they have been resolved only chromatographically[105,106]. This example shows that with the help of ion mobility-enabled instruments, FPOP could bring a level of structural information details not yet seen. While information about multiple forms of the same residue modification can be obtained, the current state-of-the-art does not allow for accurate individual quantification of these forms. The tailored FPOP evaluation suite should ideally consider the employment of ion mobility trace to benefit from additional resolving dimension.

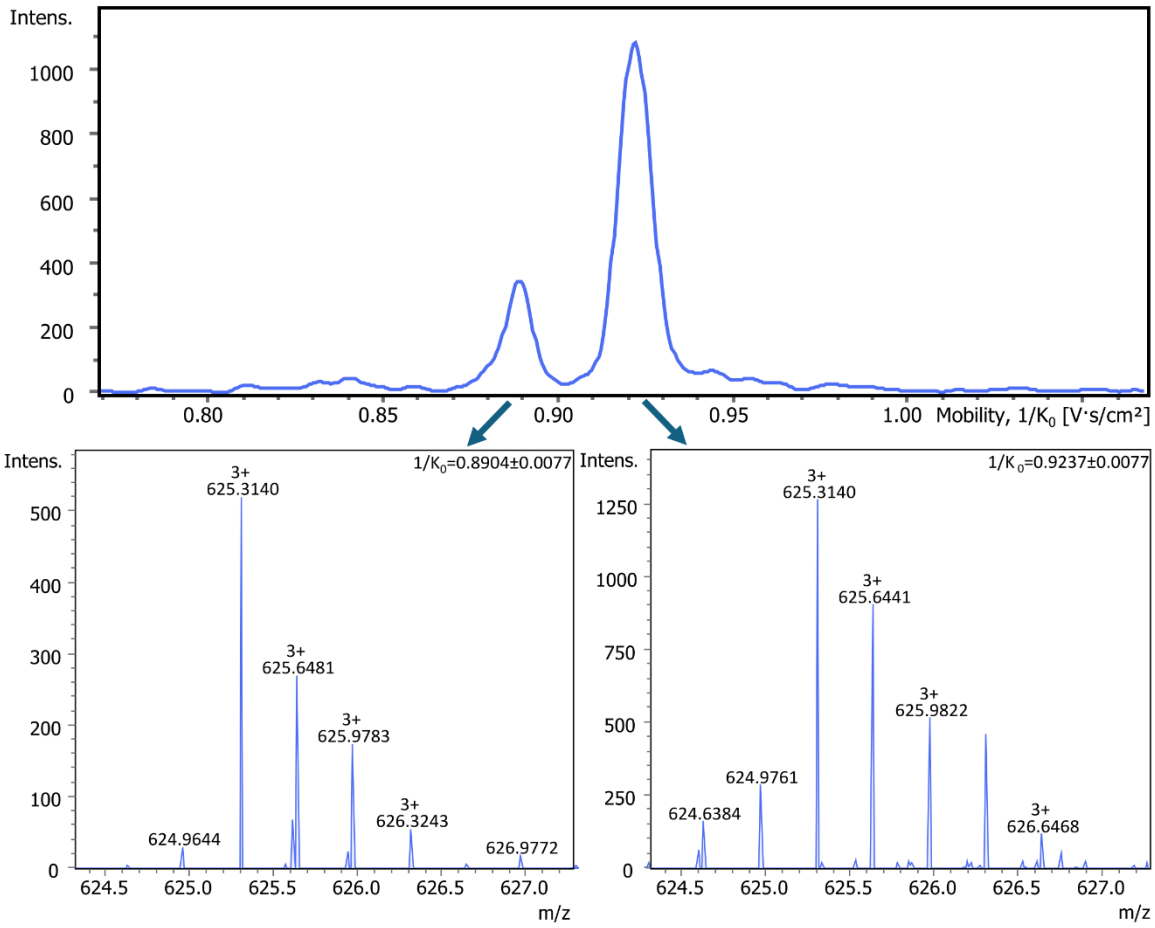


Figure 17 | Extracted ion mobilogram of the haptoglobin peptide AVGDKLPECEAVCGKPK in its triply charged form at RT 7.8 min. The mobilogram shows two resolved peaks corresponding to +16 Da modification (bottom left) and mixture of +14 and +16 Da signal (bottom right).

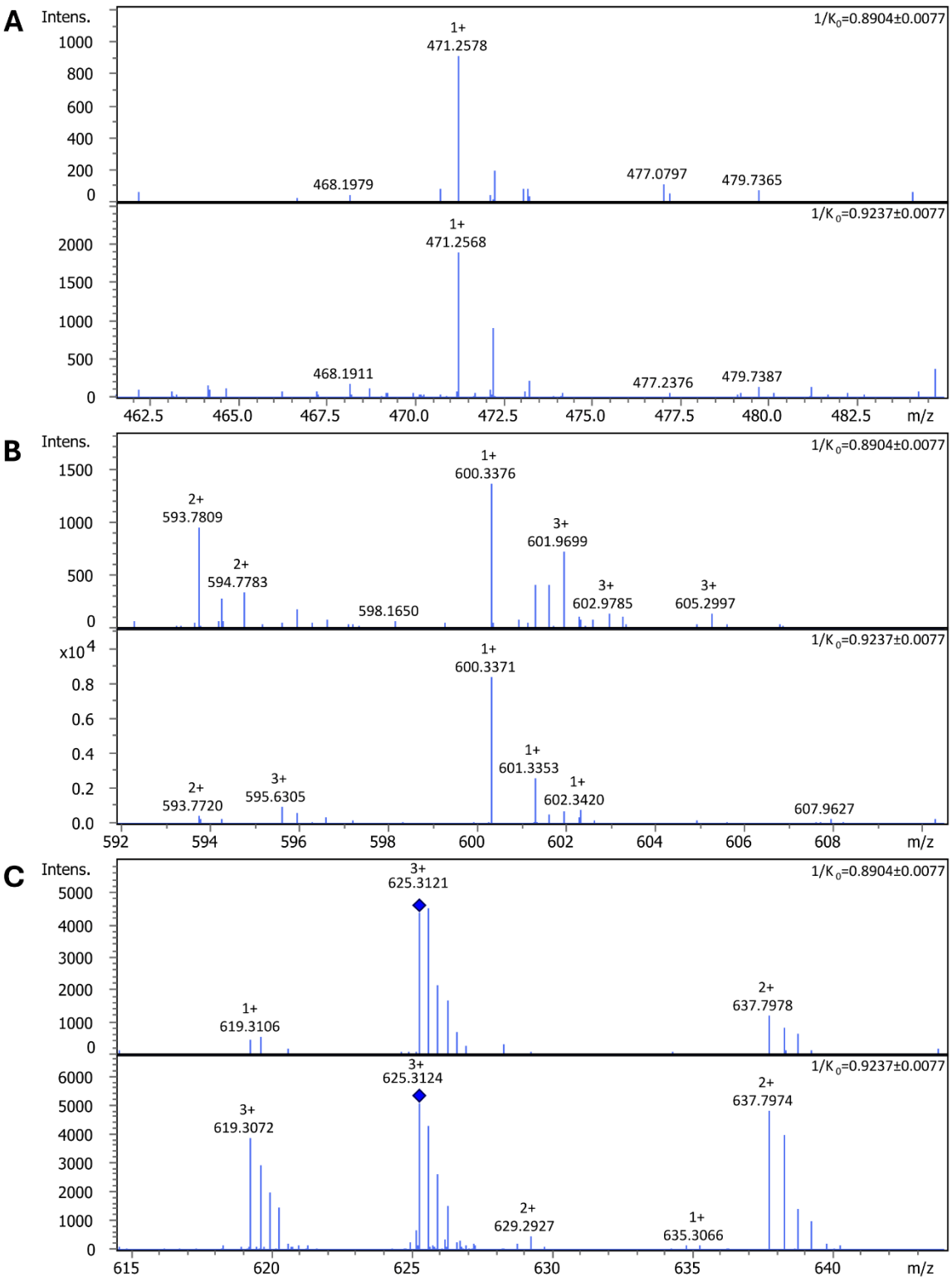


Figure 18 | Selected details from fragmentation spectra of two coeluted peptidoforms of the haptoglobin peptide AVGDKLPECEAVCGKPK in its triply charged form at *m/z* 625.312±0.001 at RT 7.8 min separated by ion mobility. Both forms bear an oxidation (+16 Da) on the Leu6, assigned by (A) b5 ion at *m/z* 471.257±0.001, (B) b6 ion at *m/z* 600.337±0.001, and (C) doubly charged y11 ion at *m/z* 637.797±0.001. The ◆ denotes the precursor. The ion at *m/z* 619.309±0.002 corresponds to the precursor with a neutral loss of water.

The first round of validation steps led to the rejection of 70% of the unfiltered identifications. Next, the fragmentation spectra of the library entries were matched against the DIA data to calculate the spectral similarity. Most of the identifications were matched with a dot product score of 0.8 or higher, showing a high level of similarity. A score of 0.7 was selected as a threshold for identifications that were manually validated. This led to rejection of additional 28 IDs due to low intensity or inconsistent presence among replicates, further improving the quality of the spectral library. Although this approach results in a more laborious validation process, the resulting spectral library contains more identifications, which are also more reliable.

The validated spectral library was compared with the workflow using Mascot and automated results filtering as described in Publication I. In order to preserve comparability, identifications with unambiguous single FPOP modification from the library were used. As illustrated in Table 3, the ratios of individual FPOP modifications remained mostly similar, with carbonyls and decarboxylations less abundant in the validated library. Nevertheless, the newly proposed approach yielded more modified identifications (333) than the unique IDs provided by Mascot (200).

Table 3 | Representation of modifications in the validated spectral library and in the Mascot search results.

modification	**validated library**		**Mascot***	
	count	percentage	count	percentage
Arg deguanidination	4	1%	2	1%
carbonyl	24	7%	20	10%
decarboxylation	23	7%	18	9%
dioxidation	36	11%	17	9%
His +5Da	23	7%	10	5%
His -10Da	16	5%	9	5%
Met -32Da	5	2%	3	2%
oxidation	200	60%	121	61%
trioxidation	2	1%	NA	

* only unique modified peptides

The manually validated library outperformed the automated Mascot workflow on both ends of the amino acid side chains reactivity scale, containing more IDs of less reactive aliphatic amino acids (Val, Leu, and Pro) as well as amino acids with more reactive side chains (His, Met, Phe, Trp, and Tyr). This set of identifications also provided a better mapping of the HbHp interaction surface (Figure 19, page 52),

enhancing the structural information provided by this method. Leveraging the full potential of FPOP labelling is crucial for its widespread utilisation.

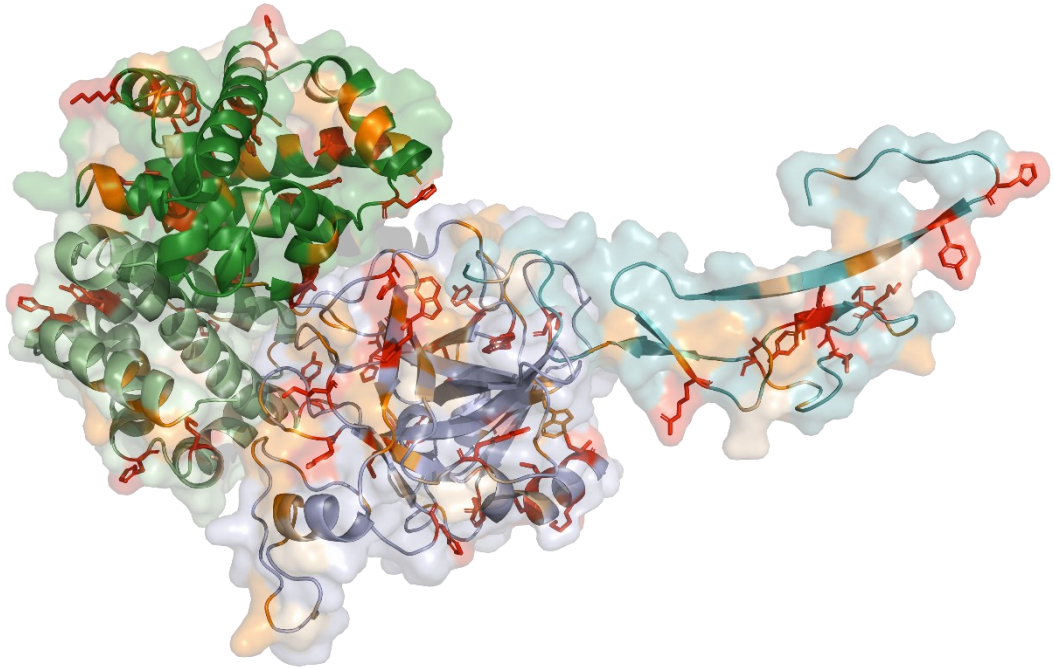


Figure 19 | Structure of the HbHp complex with quantified FPOP modifications. Hb alpha is shown in pale green, Hb beta in dark green, Hp alpha in cyan, and Hp beta in dark blue. The extent of modification is separated into three tiers: 1-3 % (beige); 3-10 % (orange); 10% and more (red). The most modified residues are also shown as sticks. Structure adapted from[189].

One limitation of this workflow is the neglect of isobaric modifications, which should be taken into account to assess the full extent of modification[105,106]. Although the incorporation of isobaric peptides is not currently available in Skyline, tailored software solution for FPOP evaluation should allow for this.

The accepted identifications were quantified on the level of precursor (MS1 trace) for DDA, DIA, and MS1-only modes. To assess the variability of individual acquisition modes, coefficients of variation (CVs) were calculated (Figure 20, page 53). Quantification obtained from DIA data demonstrated the least variability with median CV 0.1233, MS1-only mode performed similarly with median CV 0.1494 and slightly broader interquartile range. Finally, the DDA mode exhibited the most variability, with a median CV of 0.2104 and the widest interquartile range. The one-way ANOVA test followed by the Tukey Honest Significant Difference post hoc test confirmed that the differences between all test pairs are significant with *p*-values for the DIA–DDA and DDA–MS1-only pairs lower than 1.8×10^{-10}, and the DIA–MS1-only pair having a *p*-value of 0.00338.

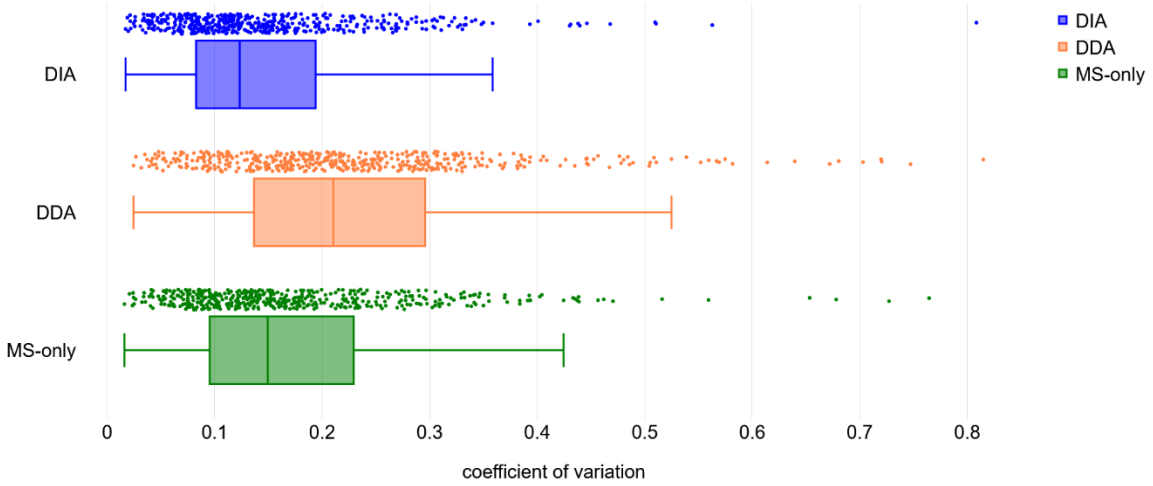


Figure 20 | Variability of quantification. The coefficients of variation were calculated from the quantification of the modifications identified (five replicates) for each acquisition mode. The line represents the median.

Furthermore, this approach was utilised for evaluation of Fast Fluoralkylation of Proteins analysis on the same HbHp system with analogous results. At first, the library contained 74 Togni-modified identifications. The mapping of DIA data revealed 11 unreliable IDs. Similarly to the first experiment, DIA exhibited the least variability. This shows that the proposed workflow is adaptable and effective for various radical probes used in protein footprinting.

This work demonstrated the viability of the DIA workflow for FPOP by combining the validated DDA library with the DIA data for further validation of the identifications. Although the validation step is time-consuming, it can be aided by a suitable level of automatization. Furthermore, quantification of DIA data exhibits the lowest variability on the timsTOF SCP. While the identification of extensively modified peptides (such as the products of FPOP analysis) from DIA data is not yet computationally feasible, the superior degree of informational complexity over that of DDA can be leveraged as an additional degree of validation.

4.3. Publication III

The previous publications were representative of closed search approach, working with the FPOP modifications as well-defined alterations of the amino acid residues. While defining them as variable modifications allows to control the size of the search space, some cases could call for a different approach. Such as the study of the structure of laccase BP76 of the termite species *Neocapritermes taracua*, which plays a crucial role in their self-sacrifice defence strategy. This oxidoreductase catalyses conversion

of hydroquinones stored in the termite body to highly toxic benzoquinones (Figure 21).

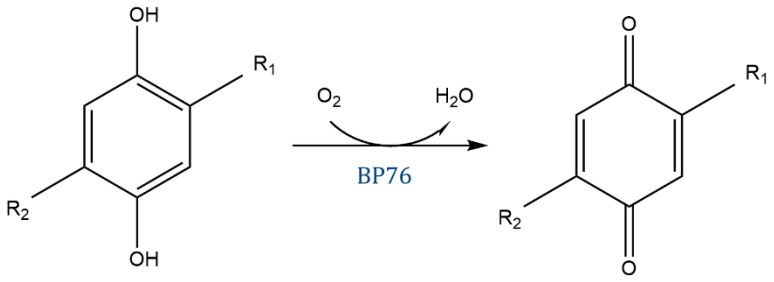


Figure 21 | Scheme of the reaction catalysed by laccase BP76 (R_1: -H / -CH_3; R_2: -H / -OH).

The structure of the laccase BP76 was determined by X-ray diffraction at 1.8 Å, utilising the signal from its copper atoms (Figure 22A, page 55). The study revealed that the enzyme retains its catalytic activity even in the solid state, owing to the combined effects of compact folding, glycosylation, and the stabilisation of the unstructured regions either by disulphide bridges or tight dimer interface. However, one of the structural elements required more investigation. Two cysteine residues in the BP76 sequence are not participating in disulphide bridges – Cys614, which coordinates one of the copper atoms; and Cys292, located at the active site of the enzyme, which binds to ε-carbon of Lys189 via metastable cross-link (Figure 22B, page 55). This covalent interaction was modelled based on the electron density of the X-ray diffraction but demanded further confirmation.

The LC-MS data were searched using FragPipe 21.1 in open search mode. Although the region containing Cys292 was not covered by the analysis, peptides 169–189 and 169–200 containing the Lys189 were detected. The fragmentation spectrum of the ion signal at *m/z* 920.95 at RT 46.30 min pointed out quadruply charged peptide ^{169}YDFFANNEGTLYWHSHDGLQKLDGIQGSIVVR^{200} with a mass shift of +0.9840 Da on Lys189 (Figure 23, page 55) corresponding to lysine oxidation to α-aminoadipic semialdehyde, also known as allysine. This was further supported by the detection of the triply charged ^{169}YDFFANNEGTLYWHSHDGLQK^{189} peptide at *m/z* 848.38 at retention time 36.71 min bearing the same modification (Figure 24, page 56). This peptide exhibited approximately 4-times lower intensity, indicating that the cleavage after allysine-modified Lys189 is hindered.

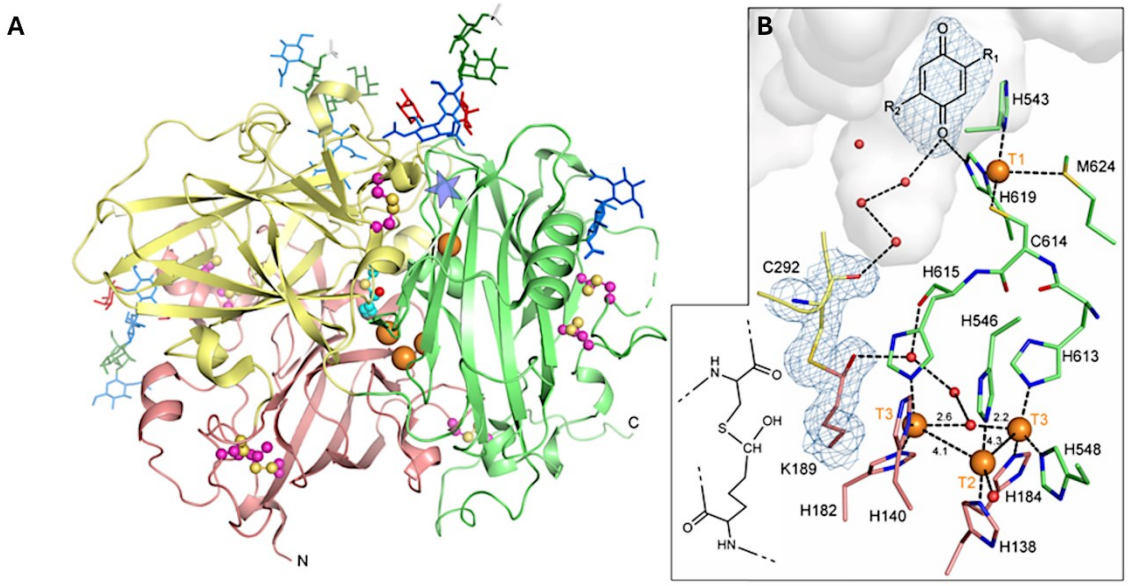


Figure 22 | Structure of the termite laccase BP76. (A) BP76 monomer with laccase domains coloured red, yellow, and green, and the active site marked with a blue star. The disulfide bridges and the Lys189–Cys292 cross-link are highlighted as magenta and cyan balls and sticks, respectively. Sulphur atoms are yellow, oxygen red, and copper orange. Glycans are shown as sticks coloured according to the Symbol Nomenclature for Glycans. (B) BP76 active site featuring the Lys189–Cys292 cross-link, water molecules linking it to the active site, the protein's Connolly surface, a 2mFo-DFc electron density map at 1σ, chemical formulas of the hemithioacetal cross-link and the reaction product (R1: -H or -CH_3, R2: -H or -OH), and distances in Å between the three copper types (T1–3). Adapted from[190].

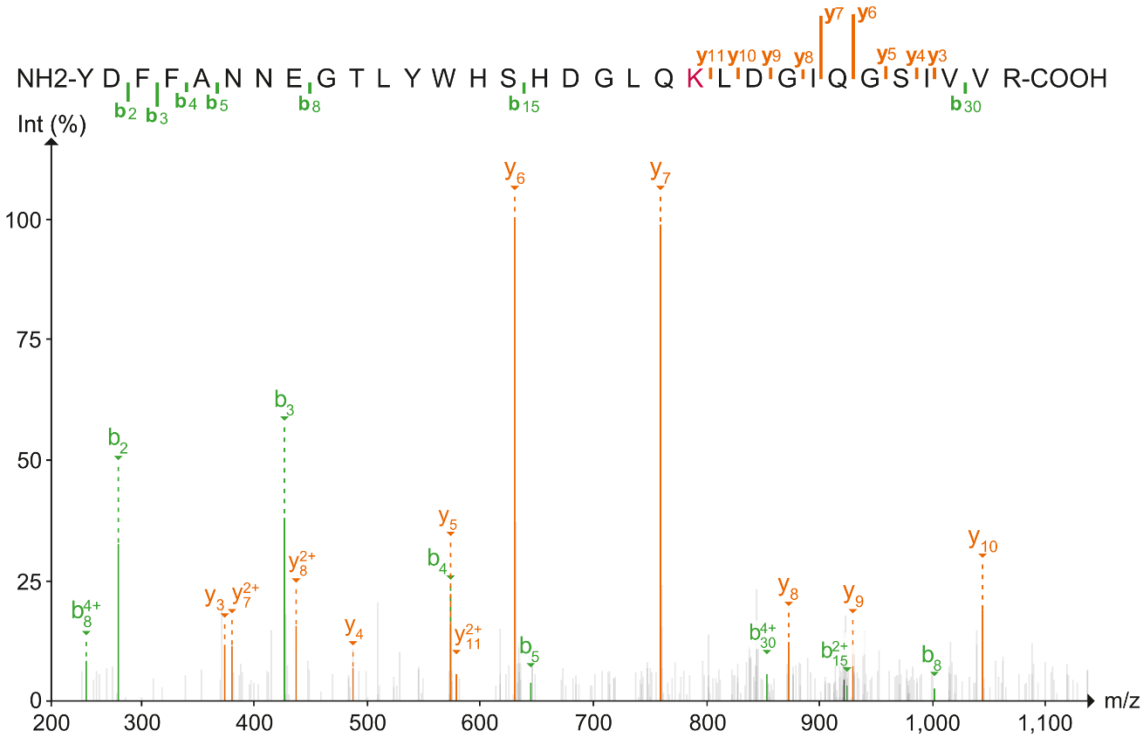


Figure 23 | Fragmentation spectrum of the YDFFANNEGTLYWHSHDGLQKLDGIQGSIVVR peptide in its quadruply charged form at *m/z* 920.95 and RT 46.30 min bearing +0.9840 Da modification of Lys189 (highlighted in red). This mass shift corresponds to aldehyde moiety replacing the amino moiety.

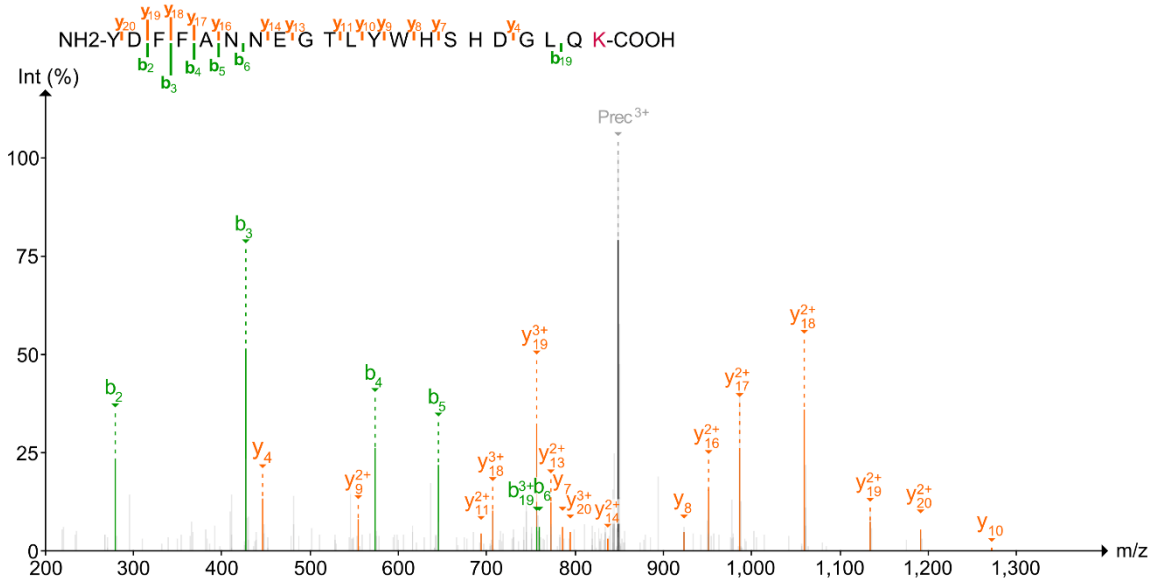


Figure 24 | Fragmentation spectrum of the YDFFANNEGTLYWHSHDGLQK peptide in its triply charged form at *m/z* 848.38 and RT 36.71 min bearing a +0.9840 Da modification of Lys189 (highlighted in red).

This post-translational modification was previously observed in collagen and elastin as a product of lysyl oxidase, which is also a copper metalloenzyme[191]. Furthermore, it is assumed that this modification is formed in cyt *c* in the presence of peroxide-induced radicals. This could lead to structural changes in the cyt *c* molecule, enabling its peroxidase activity during apoptosis[192,193]. The formation of allysine in BP76 could follow a similar mechanism that likely necessitates reactive oxygen species. The gained aldehyde moiety enables the Lys189 to react with the sulfhydryl of the proximal Cys292, forming the reversible hemithioacetal cross-link, detected by the X-ray crystallography (Figure 22B).

The first two publications demonstrated workflows with well-defined and characterised FPOP modifications. However, the work done in Publication III has shown that reactive oxygen species can give rise to hardly predictable modifications, which can also serve as a source of structural information. Although such modifications probably occur with low abundance, including them in footprinting experiments, data evaluation could enhance the information provided by the technique. To incorporate such modifications, a specialised software solution would ideally allow for open search workflow, allowing to combine the results with known modifications. This could be beneficial not only for existing experimental designs, but also for new radical labelling-based structure study techniques developed in the future.

5. Summary

This thesis aimed to contribute to the development of an evaluation workflow for mass spectrometric data obtained by the analysis of proteins modified by reactive oxygen species. Following are the obtained results, published in the three attached publications:

- The impact of peptide identification framework selection on the outcomes of Fast Photochemical Oxidation of Proteins (FPOP) analyses was assessed, comparing two search engines with different algorithmic approaches – Mascot and PEAKS.
- Only 31% of identifications (IDs) were mutual between the search engines, Mascot exhibited higher consistency between replicates (around 26% of mutual IDs), compared to PEAKS (around 22% of mutual IDs).
- Mascot was able to reliably assign the modification site even at lower peptide-spectrum match (PSM) scores, and its site confidence evaluation was more strongly related to the fragmentation coverage. PEAKS did not show a clear relationship of the modification site determination with the PSM scores or fragmentation coverage.
- The reproducibility of quantification across DDA, DIA, and MS1-only acquisition modes was compared on FPOP analysis of the HbHp complex using coefficients of variation. DIA and MS1-only acquisition modes demonstrated significantly better reproducibility compared to DDA, even for low-intensity identifications.
- A novel approach of a manually validated spectral library enabled matching of the identifications obtained by database search against the DIA data, improving the identification confidence.
- The manual validation process resulted in an increase in the spatial resolution of the method compared to conventional approach with automated filtering steps.
- The proposed workflow was also performed for Fast Fluoroalkylation of Proteins analysis, demonstrating its adaptability across various protein footprinting techniques.
- An open search was employed to elucidate the formation of a reversible Lys-Cys cross-link in the structure of the blue laccase BP76 from *Neocapritermes taracua* termite. The stabilising effect involves the modification of Lys to α-amino-adipic semialdehyde, presumably through a radical mechanism of reactive oxygen species.

6. List of publications

PUBLICATIONS SUPPORTING THIS THESIS

1. **Zakopcanik, M.;** Kavan, D.; Novak, P. & Loginov, D. S. Quantifying the Impact of the Peptide Identification Framework on the Results of Fast Photochemical Oxidation of Protein Analysis. *J. Proteome Res.* **23**, 609–617 (2024).

2. **Zakopcanik, M.;** Kavan, D.; Kukacka, Z.; Novak, P. & Loginov, D. S. Data-Independent Acquisition Represents a Promising Alternative for Fast Photochemical Oxidation of Proteins (FPOP) Samples Analysis. *Anal. Chem.* **96**, 11273–11279 (2024).
3. Škerlová, J.; Brynda, J.; Šobotník, J.; **Zákopčaník, M.**; Novák, P.; Bourguignon, T.; Sillam-Dussès, D. & Řezáčová, P. Crystal Structure of Blue Laccase BP76, a Unique Termite Suicidal Defense Weapon. *Structure* (2024).

OTHER PUBLICATIONS BY AUTHOR

Charras, Q.; Gupta, S.; Tichý, M.; Al-Amoudi, A.; Lukeš, M.; Koník, P.; Sobotka, R.; **Zákopčaník, M.**; Novák, P.; Komenda, J.; Naschberger, A.: 1.8 Å Resolution Structure Of The Cyanobacterial Photosystem I Assembly Intermediate Lacking Cytoplasmic Subunits. (Nature Plants, under review), preprint available at: https://doi.org/10.21203/rs.3.rs-4640615/v1.

7. References

1. Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R. G.; Wyckoff, H.; Phillips, D. C.: A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* **181**:4610, 662–666 (1958).
2. Wlodawer, A.; Minor, W.; Dauter, Z.; Jaskolski, M.: Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS J.* **275**:1, 1–21 (2008).
3. Dobson, C. M.: Biophysical Techniques in Structural Biology. *Annu. Rev. Biochem.* **88**, 25–33 (2019).
4. Marion, D.: An Introduction to Biological NMR Spectroscopy. *Mol. Cell. Proteomics* **12**:11, 3006–3025 (2013).
5. Li, X.; Mooney, P.; Zheng, S.; Booth, C.; Braunfeld, M. B.; Gubbens, S.; Agard, D. A.; Cheng, Y.: Electron counting and beam-induced motion correction enable near atomic resolution single particle cryoEM. *Nat. Methods* **10**:6, 584–590 (2013).
6. Milazzo, A.-C.; Cheng, A.; Moeller, A.; Lyumkis, D.; Jacovetty, E.; Polukas, J.; Ellisman, M. H.; Xuong, N.-H.; Carragher, B.; Potter, C. S.: Initial evaluation of a direct detection device detector for single particle cryo-electron microscopy. *J. Struct. Biol.* **176**:3, 404–408 (2011).
7. Tivol, W. F.; Briegel, A.; Jensen, G. J.: An Improved Cryogen for Plunge Freezing. *Microsc. Microanal.* **14**:5, 375–379 (2008).
8. Thompson, R. F.; Walker, M.; Siebert, C. A.; Muench, S. P.; Ranson, N. A.: An introduction to sample preparation and imaging by cryo-electron microscopy for structural biology. *Methods* **100**, 3–15 (2016).
9. Meynier, V.; Hardwick, S. W.; Catala, M.; Roske, J. J.; Oerum, S.; Chirgadze, D. Y.; Barraud, P.; Yue, W. W.; Luisi, B. F.; Tisné, C.: Structural basis for human mitochondrial tRNA maturation. *Nat. Commun.* **15**:1, 4683 (2024).
10. Passchier, T. C.; White, J. B. R.; Maskell, D. P.; Byrne, M. J.; Ranson, N. A.; Edwards, T. A.; Barr, J. N.: The cryoEM structure of the Hendra henipavirus nucleoprotein reveals insights into paramyxoviral nucleocapsid architectures. *Sci. Rep.* **14**:1, 14099 (2024).
11. Sheward, D. J. *et al.*: Structural basis of broad SARS-CoV-2 cross-neutralization by affinity-matured public antibodies. *Cell Rep. Med.* **5**:6, 101577 (2024).
12. Zhao, G.-X.; Fang, X.-Y.; Bu, G.-L.; Chen, S.-J.-B.; Sun, C.; Li, T.; Xie, C.; Wang, Y.; Li, S.-X.; Meng, N.; Feng, G.-K.; Zhong, Q.; Kong, X.-W.; Liu, Z.; Zeng, M.-S.: Potent human monoclonal antibodies targeting Epstein-Barr virus gp42 reveal vulnerable sites for virus infection. *Cell Rep. Med.* **5**:5, 101573 (2024).
13. Fenn, J.: Electrospray ionization mass spectrometry: How it all began. *J. Biomol. Tech.* **13**:3, 101–118 (2002).
14. Wilm, M.; Mann, M.: Analytical properties of the nanoelectrospray ion source. *Anal. Chem.* **68**:1, 1–8 (1996).

15. Karas, M.; Bahr, U.; Dülcks, T.: Nano-electrospray ionization mass spectrometry: addressing analytical problems beyond routine. *Fresenius' J. Anal. Chem.* **366**:6, 669–676 (2000).

16. Benesch, J. L. P.; Ruotolo, B. T.: Mass Spectrometry: an Approach Come-of-Age for Structural and Dynamical Biology. *Curr. Opin. Struct. Biol.* **21**:5, 641–649 (2011).

17. Sharon, M.: How far can we go with structural mass spectrometry of protein complexes? *J. Am. Soc. Mass Spectrom.* **21**:4, 487–500 (2010).

18. Heck, A. J. R.: Native mass spectrometry: a bridge between interactomics and structural biology. *Nat. Methods* **5**:11, 927–933 (2008).

19. Beveridge, R.; Calabrese, A. N.: Structural Proteomics Methods to Interrogate the Conformations and Dynamics of Intrinsically Disordered Proteins. *Front. Chem.* **9,** 603639 (2021).

20. Struwe, W. B.; Robinson, C. V.: Relating glycoprotein structural heterogeneity to function – insights from native mass spectrometry. *Curr. Opin. Struct. Biol.* **58**, 241–248 (2019).

21. Konijnenberg, A.; Butterer, A.; Sobott, F.: Native ion mobility-mass spectrometry and related methods in structural biology. *Biochim. Biophys. Acta, Proteins Proteomics* **1834**:6, 1239–1256 (2013).

22. Loo, J. A.: Studying noncovalent protein complexes by electrospray ionization mass spectrometry. *Mass Spectrom. Rev.* **16**:1, 1–23 (1997).

23. Han, Y.; Desai, A. A.; Zupancic, J. M.; Smith, M. D.; Tessier, P. M.; Ruotolo, B. T.: Native ion mobility-mass spectrometry reveals the binding mechanisms of anti-amyloid therapeutic antibodies. *Protein Sci.* **33**:6, e5008 (2024).

24. Schneeberger, E.-M.; Breuker, K.: Native Top-Down Mass Spectrometry of TAR RNA in Complexes with a Wild-Type tat Peptide for Binding Site Mapping. *Angew. Chem., Int. Ed.* **56**:5, 1254–1258 (2017).

25. Cubrilovic, D.; Biela, A.; Sielaff, F.; Steinmetzer, T.; Klebe, G.; Zenobi, R.: Quantifying Protein-Ligand Binding Constants using Electrospray Ionization Mass Spectrometry: A Systematic Binding Affinity Study of a Series of Hydrophobically Modified Trypsin Inhibitors. *J. Am. Soc. Mass Spectrom.* **23**:10, 1768–1777 (2012).

26. Vahidi, S.; Stocks, B. B.; Konermann, L.: Partially Disordered Proteins Studied by Ion Mobility-Mass Spectrometry: Implications for the Preservation of Solution Phase Structure in the Gas Phase. *Anal. Chem.* **85**:21, 10471–10478 (2013).

27. Moghadamchargari, Z.; Huddleston, J.; Shirzadeh, M.; Zheng, X.; Clemmer, D. E.; Raushel, F. M.; Russell, D. H.; Laganowsky, A.: Intrinsic GTPase Activity of K-RAS Monitored by Native Mass Spectrometry. *Biochemistry* **58**:31, 3396–3405 (2019).

28. Kanu, A. B.; Dwivedi, P.; Tam, M.; Matz, L.; Hill Jr., H. H.: Ion mobility–mass spectrometry. *J. Mass Spectrom.* **43**:1, 1–22 (2008).

29. Smith, D. P.; Giles, K.; Bateman, R. H.; Radford, S. E.; Ashcroft, A. E.: Monitoring Copopulated Conformational States During Protein Folding Events Using Electrospray Ionization-Ion Mobility Spectrometry-Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **18**:12–9, 2180–2190 (2007).

30. Jacobs, A.; Hoover, H.; Smith, E.; Clemmer, D. E.; Kim, C.-H.; Kao, C. C.: The intrinsically disordered N-terminal arm of the brome mosaic virus coat protein specifically recognizes the RNA motif that directs the initiation of viral RNA replication. *Nucleic Acids Res.* **46**:1, 324–335 (2018).

31. van Dyck, J. F.; Burns, J. R.; Le Huray, K. I. P.; Konijnenberg, A.; Howorka, S.; Sobott, F.: Sizing up DNA nanostructure assembly with native mass spectrometry and ion mobility. *Nat. Commun.* **13**:1, 3610 (2022).

32. O'Reilly, F. J.; Rappsilber, J.: Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology. *Nat. Struct. Mol. Biol.* **25**:11, 1000–1008 (2018).

33. Rivera-Santiago, R. F.; Sriswasdi, S.; Harper, S. L.; Speicher, D. W.: Probing structures of large protein complexes using zero-length cross-linking. *Methods* **89**, 99–111 (2015).

34. Graziadei, A.; Rappsilber, J.: Leveraging crosslinking mass spectrometry in structural and cell biology. *Structure* **30**:1, 37–54 (2022).

35. 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**:8, 7500–7531 (2022).

36. Kukacka, Z.; Rosulek, M.; Strohalm, M.; Kavan, D.; Novak, P.: Mapping protein structural changes by quantitative cross-linking. *Methods* **89**, 112–120 (2015).

37. Hvidt, A.; Linderstrøm-Lang, K.: Exchange of hydrogen atoms in insulin with deuterium atoms in aqueous solutions. *Biochim. Biophys. Acta* **14**, 574–575 (1954).

38. Glickson, J. D.; McDonald, C. C.; Phillips, W. D.: Assignment of tryptophan indole NH proton resonances of lysozyme. *Biochem. Biophys. Res. Commun.* **35**:4, 492–498 (1969).

39. Katta, V.; Chait, B. T.; Carr, S.: Conformational changes in proteins probed by hydrogen-exchange electrospray-ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* **5**:4, 214–217 (1991).

40. Englander, S. W.; Kallenbach, N. R.: Hydrogen exchange and structural dynamics of proteins and nucleic acids. *Q. Rev. Biophys.* **16**:4, 521–655 (1983).

41. Englander, S. W.; Downer, N. W.; Teitelbaum, H.: Hydrogen Exchange. *Annu. Rev. Biochem.* **41**, 1972, 903–924 (1972).

42. Bai, Y.; Milne, J. S.; Mayne, L.; Englander, S. W.: Primary Structure Effects on Peptide Group Hydrogen Exchange. *Proteins* **17**:1, 75–86 (1993).

43. Smith, D. L.; Deng, Y.; Zhang, Z.: Probing the Non-covalent Structure of Proteins by Amide Hydrogen Exchange and Mass Spectrometry. *J. Mass Spectrom.* **32**:2, 135–146 (1997).

44. Nguyen, D.; Mayne, L.; Phillips, M. C.; Walter Englander, S.: Reference Parameters for Protein Hydrogen Exchange Rates. *J. Am. Soc. Mass Spectrom.* **29**:9, 1936–1939 (2018).

45. Liao, W.-L.; Dodder, N. G.; Mast, N.; Pikuleva, I. A.; Turko, I. V.: Steroid and Protein Ligand Binding to Cytochrome P450 46A1 as Assessed by Hydrogen−Deuterium Exchange and Mass Spectrometry. *Biochemistry* **48**:19, 4150–4158 (2009).

46. Krishna, M. M. G.; Maity, H.; Rumbley, J. N.; Lin, Y.; Englander, S. W.: Order of Steps in the Cytochrome *c* Folding Pathway: Evidence for a Sequential Stabilization Mechanism. *J. Mol. Biol.* **359**:5, 1410–1419 (2006).

47. Marcsisin, S. R.; Engen, J. R.: Molecular insight into the conformational dynamics of the Elongin BC complex and its interaction with HIV-1 Vif. *J. Mol. Biol.* **402**:5, 892–904 (2010).

48. Nakazawa, S.; Hashii, N.; Harazono, A.; Kawasaki, N.: Analysis of oligomeric stability of insulin analogs using hydrogen/deuterium exchange mass spectrometry. *Anal. Biochem.* **420**:1, 61–67 (2012).

49. Cao, J.; Burke, J. E.; Dennis, E. A.: Using Hydrogen/Deuterium Exchange Mass Spectrometry to Define the Specific Interactions of the Phospholipase A2 Superfamily with Lipid Substrates, Inhibitors, and Membranes. *J. Biol. Chem.* **288**:3, 1806–1813 (2013).

50. Masson, G. R. *et al.*: Recommendations for performing, interpreting and reporting hydrogen deuterium exchange mass spectrometry (HDX-MS) experiments. *Nat. Methods* **16**:7, 595–602 (2019).

51. Liu, X. R.; Zhang, M. M.; Gross, M. L.: Mass Spectrometry-based protein footprinting for higher-order structure analysis: Fundamentals and applications. *Chem. Rev.* **120**:10, 4355–4454 (2020).

52. Kan, Z.-Y.; Walters, B. T.; Mayne, L.; Englander, S. W.: Protein hydrogen exchange at residue resolution by proteolytic fragmentation mass spectrometry analysis. *Proc. Natl. Acad. Sci. U. S. A.* **110**:41, 16438–16443 (2013).

53. Rand, K. D.; Zehl, M.; Jensen, O. N.; Jørgensen, T. J. D.: Protein Hydrogen Exchange Measured at Single-Residue Resolution by Electron Transfer Dissociation Mass Spectrometry. *Anal. Chem.* **81**:14, 5577–5584 (2009).

54. Krell, T.; Pitt, A. R.; Coggins, J. R.: The use of electrospray mass spectrometry to identify an essential arginine residue in type II dehydroquinases. *FEBS Lett.* **360**:1, 93–96 (1995).

55. Wanigasekara, M. S. K.; Huang, X.; Chakrabarty, J. K.; Bugarin, A.; Chowdhury, S. M.: Arginine-Selective Chemical Labeling Approach for Identification and Enrichment of Reactive Arginine Residues in Proteins. *ACS Omega* **3**:10, 14229–14235 (2018).

56. Schepens, I.; Ruelland, E.; Miginiac-Maslow, M.; Le Maréchal, P.; Decottignies, P.: The Role of Active Site Arginines of Sorghum NADP-Malate Dehydrogenase in Thioredoxin-dependent Activation and Activity. *J. Biol. Chem.* **275**:46, 35792–35798 (2000).

57. Bogan, A. A.; Thorn, K. S.: Anatomy of hot spots in protein interfaces1. *J. Mol. Biol.* **280**:1, 1–9 (1998).

58. Akashi, S.; Shirouzu, M.; Terada, T.; Ito, Y.; Yokoyama, S.; Takio, K.: Characterization of the Structural Difference between Active and Inactive Forms of the Ras Protein by Chemical Modification Followed by Mass Spectrometric Peptide Mapping. *Anal. Biochem.* **248**:1, 15–25 (1997).

59. Izumi, S.; Kaneko, H.; Yamazaki, T.; Hirata, T.; Kominami, S.: Membrane Topology of Guinea Pig Cytochrome P450 17α Revealed by a Combination of Chemical Modifications and Mass Spectrometry. *Biochemistry* **42**:49, 14663–14669 (2003).

60. Kaur, P.; Tomechko, S. E.; Kiselar, J.; Shi, W.; Deperalta, G.; Wecksler, A. T.; Gokulrangan, G.; Ling, V.; Chance, M. R.: Characterizing monoclonal antibody structure by carboxyl group footprinting. *mAbs* **7**:3, 540–552 (2015).

61. Anson, M. L.: The Reactions of Iodine and Iodoacetamide with Native Egg Albumin. *J. Gen. Physiol.* **23**:3, 321–331 (1940).

62. Maithal, K.; Ravindra, G.; Balaram, H.; Balaram, P.: Inhibition of *Plasmodium falciparum* Triose-phosphate Isomerase by Chemical Modification of an Interface Cysteine. *J. Biol. Chem.* **277**:28, 25106–25114 (2002).

63. Britto, P. J.; Knipling, L.; Wolff, J.: The Local Electrostatic Environment Determines Cysteine Reactivity of Tubulin. *J. Biol. Chem.* **277**:32, 29018–29027 (2002).

64. Friedman, M.: Application of the S-Pyridylethylation Reaction to the Elucidation of the Structures and Functions of Proteins. *J. Protein Chem.* **20**:6, 431–453 (2001).

65. Codina, A.; Vilaseca, M.; Tarragó, T.; Fernández, I.; Ludevid, D.; Giralt, E.: Location of Disulfide bonds in mature α-L;-fucosidase from pea. *J. Pept. Sci.* **7**:6, 305–315 (2001).

66. Shetterly, S.; Tom, I.; Yen, T.-Y.; Joshi, R.; Lee, L.; Wang, P. G.; Macher, B. A.: α1,3 galactosyltransferase: new sequences and characterization of conserved cysteine residues. *Glycobiology* **11**:8, 645–653 (2001).

67. Schneider, F.: Histidine in Enzyme Active Centers. *Angew. Chem., Int. Ed. Engl.* **17**:8, 583–592 (1978).

68. Hullán, L.; Szontagh, T.; Turtóczky, I.; Fedorcsák, I.: The inactivation of trypsin by diethyl pyrocarbonate. *Acta Chem. Scand.* **19**:10, 2440–2441 (1965).

69. Foti, S.; Marletta, D.; Saletti, R.; Petrone, G.; Daolio, S.: Fast-atom bombardment mass spectrometry of peptide derivatives with diethylpyrocarbonate. *Rapid Commun. Mass Spectrom.* **5**:7, 336–339 (1991).

70. Steiner, R. F.; Albaugh, S.; Fenselau, C.; Murphy, C.; Vestling, M.: A mass spectrometry method for mapping the interface topography of interacting proteins, illustrated by the melittin-calmodulin system. *Anal. Biochem.* **196**:1, 120–125 (1991).

71. Przybylski, M.; Glocker, M. O.; Nestel, U.; Schnaible, V.; Blüggel, M.; Diederichs, K.; Weckesser, J.; Schad, M.; Schmid, A.; Welte, W.; Benz, R.: X-ray crystallographic and mass spectrometric structure determination and functional characterization of succinylated porin from *Rhodobacter capsulatus*: Implications for ion selectivity and single-channel conductance. *Protein Sci.* **5**:8, 1477–1489 (1996).

72. Blumberg, S.; Vallee, B. L.: Superactivation of thermolysin by acylation with amino acid N-hydroxysuccinimide esters. *Biochemistry* **14**:11, 2410–2419 (1975).

73. Suckau, D.; Mak, M.; Przybylski, M.: Protein surface topology-probing by selective chemical modification and mass spectrometric peptide mapping. *Proc. Natl. Acad. Sci. U. S. A.* **89**:12, 5630–5634 (1992).

74. Scaloni, A.; Monti, M.; Acquaviva, R.; Tell, G.; Damante, G.; Formisano, S.; Pucci, P.: Topology of the Thyroid Transcription Factor 1 Homeodomain−DNA Complex. *Biochemistry* **38**:1, 64–72 (1999).

75. Koshland, D. E.; Karkhanis, Y. D.; Latham, H. G.: An Environmentally-Sensitive Reagent with Selectivity for the Tryptophan Residue in Proteins. *J. Am. Chem. Soc.* **86**:7, 1448–1450 (1964).

76. Patchornik, A.; Lawson, W. B.; Witkop, B.: Selective Cleavage of Peptide Bonds. II. The Tryptophyl Peptide Bond and The Cleavage Of Glucagon1. *J. Am. Chem. Soc.* **80**:17, 4747–4748 (1958).

77. Scoffone, Ernesto.; Fontana, Angelo.; Rocchi, Raniero.: Sulfenyl halides as modifying reagents for polypeptides and proteins. I. Modification of tryptophan residues. *Biochemistry* **7**:3, 971–979 (1968).

78. Hahm, H. S.; Toroitich, E. K.; Borne, A. L.; Brulet, J. W.; Libby, A. H.; Yuan, K.; Ware, T. B.; McCloud, R. L.; Ciancone, A. M.; Hsu, K.-L.: Global targeting of functional tyrosines using sulfur-triazole exchange chemistry. *Nat. Chem. Biol.* **16**:2, 150–159 (2020).

79. Xie, X.; Moon, P. J.; Crossley, S. W. M.; Bischoff, A. J.; He, D.; Li, G.; Dao, N.; Gonzalez-Valero, A.; Reeves, A. G.; McKenna, J. M.; Elledge, S. K.; Wells, J. A.; Toste, F. D.; Chang, C. J.: Oxidative cyclization reagents reveal tryptophan cation–π interactions. *Nature* **627**:8004, 680–687 (2024).

80. Abello, N.; Kerstjens, H. A. M.; Postma, D. S.; Bischoff, R.: Protein Tyrosine Nitration: Selectivity, Physicochemical and Biological Consequences, Denitration, and Proteomics Methods for the Identification of Tyrosine-Nitrated Proteins. *J. Proteome Res.* **8**:7, 3222–3238 (2009).

81. Stone, M. J.; Chuang, S.; Hou, X.; Shoham, M.; Zhu, J. Z.: Tyrosine sulfation: an increasingly recognised post-translational modification of secreted proteins. *N. Biotechnol.* **25**:5, 299–317 (2009).

82. Hunter, T.: Why nature chose phosphate to modify proteins. *Philos. Trans. R. Soc., B* **367**:1602, 2513–2516 (2012).

83. Riordan, J. F.; Sokolovsky, M.; Vallee, B. L.: Tetranitromethane. A Reagent for the Nitration of Tyrosine and Tyrosyl Residues of Proteins1. *J. Am. Chem. Soc.* **88**:17, 4104–4105 (1966).

84. Hobba, G. D.; Forbes, B. E.; Parkinson, E. J.; Francis, G. L.; Wallace, J. C.: The insulin-like growth factor (IGF) binding site of bovine insulin-like growth factor binding protein-2 (bIGFBP-2) probed by iodination. *J. Biol. Chem.* **271**:48, 30529–30536 (1996).

85. Simpson, R. T.; Riordan, J. F.; Vallee, B. L.: Functional Tyrosyl Residues in the Active Center of Bovine Pancreatic Carboxypeptidase A. *Biochemistry* **2**:3, 616–622 (1963).

86. Gong, W.; Zhang, G.; Liu, T.; Giri, R.; Yu, J.-Q.: Site-Selective C(sp3)–H Functionalization of Di-, Tri-, and Tetrapeptides at the N-Terminus. *J. Am. Chem. Soc.* **136**:48, 16940–16946 (2014).

87. Rodríguez, N.; Romero-Revilla, J. A.; Fernández-Ibáñez, M. Á.; Carretero, J. C.: Palladium-catalyzed N-(2-pyridyl)sulfonyl-directed C(sp3)–H γ-arylation of amino acid derivatives. *Chem. Sci.* **4**:1, 175–179 (2012).

88. Osberger, T. J.; Rogness, D. C.; Kohrt, J. T.; Stepan, A. F.; White, M. C.: Oxidative diversification of amino acids and peptides by small-molecule iron catalysis. *Nature* **537**:7619, 214–219 (2016).

89. deGruyter, J. N.; Malins, L. R.; Baran, P. S.: Residue-Specific Peptide Modification: A Chemist's Guide. *Biochemistry* **56**:30, 3863–3873 (2017).

90. Gharakhanian, E. G.; Deming, T. J.: Versatile Synthesis of Stable, Functional Polypeptides via Reaction with Epoxides. *Biomacromolecules* **16**:6, 1802–1806 (2015).

91. Gharakhanian, E. G.; Deming, T. J.: Chemoselective synthesis of functional homocysteine residues in polypeptides and peptides. *Chem. Commun.* **52**:30, 5336–5339 (2016).

92. Kramer, J. R.; Deming, T. J.: Reversible chemoselective tagging and functionalization of methionine containing peptides. *Chem. Commun.* **49**:45, 5144–5146 (2013).

93. Bagwan, N.; El Ali, H. H.; Lundby, A.: Proteome-wide profiling and mapping of post translational modifications in human hearts. *Sci. Rep.* **11**:1, 2184 (2021).

94. Hambly, D. M.; Gross, M. L.: Laser Flash Photolysis of Hydrogen Peroxide to Oxidize Protein Solvent-Accessible Residues on the Microsecond Timescale. *J. Am. Soc. Mass Spectrom.* **16**:12, 2057–2063 (2005).

95. Gau, B. C.; Chen, H.; Zhang, Y.; Gross, M. L.: Sulfate Radical Anion as a New Reagent for Fast Photochemical Oxidation of Proteins. *Anal. Chem.* **82**:18, 7821–7827 (2010).

96. Zhang, B.; Rempel, D. L.; Gross, M. L.: Protein Footprinting by Carbenes on a Fast Photochemical Oxidation of Proteins (FPOP) Platform. *J. Am. Soc. Mass Spectrom.* **27**:3, 552–555 (2016).

97. Tullius, T. D.; Dombroski, B. A.: Hydroxyl radical 'footprinting': high-resolution information about DNA-protein contacts and application to lambda repressor and Cro protein. *Proc. Natl. Acad. Sci. U. S. A.* **83**:15, 5469–5473 (1986).

98. Hawkins, C. L.; Davies, M. J.: Generation and propagation of radical reactions on proteins. *Biochim. Biophys. Acta, Bioenerg.* **1504**:2, 196–219 (2001).

99. Garrison, W. M.: Reaction mechanisms in the radiolysis of peptides, polypeptides, and proteins. *Chem. Rev.* **87**:2, 381–398 (1987).

100. Maleknia, S. D.; Brenowitz, M.; Chance, M. R.: Millisecond Radiolytic Modification of Peptides by Synchrotron X-rays Identified by Mass Spectrometry. *Anal. Chem.* **71**:18, 3965–3973 (1999).

101. Takamoto, K.; Chance, M. R.: Radiolytic protein footprinting with mass spectrometry to probe the structure of macromolecular complexes. *Annu. Rev. Biophys. Biomol. Struct.* **35**, 251–276 (2006).

102. Xu, G.; Chance, M. R.: Hydroxyl radical-mediated modification of proteins as probes for structural proteomics. *Chem. Rev.* **107**:8, 3514–3543 (2007).

103. Liu, X. R.; Zhang, M. M.; Zhang, B.; Rempel, D. L.; Gross, M. L.: Hydroxyl-Radical Reaction Pathways for the Fast Photochemical Oxidation of Proteins Platform As Revealed by 18O Isotopic Labeling. *Anal. Chem.* **91**:14, 9238–9245 (2019).

104. Xu, G.; Chance, M. R.: Radiolytic Modification and Reactivity of Amino Acid Residues Serving as Structural Probes for Protein Footprinting. *Anal. Chem.* **77**:14, 4549–4555 (2005).

105. Cornwell, O.; Radford, S. E.; Ashcroft, A. E.; Ault, J. R.: Comparing Hydrogen Deuterium Exchange and Fast Photochemical Oxidation of Proteins: a Structural Characterisation of Wild-Type and ΔN6 β 2-Microglobulin. *J. Am. Soc. Mass Spectrom.* **29**, 2413–2426 (2018).

106. Polák, M.; Yassaghi, G.; Kavan, D.; Filandr, F.; Fiala, J.; Kukačka, Z.; Halada, P.; Loginov, D. S.; Novák, P.: Utilization of Fast Photochemical Oxidation of Proteins and Both Bottom-up and Top-down Mass Spectrometry for Structural Characterization of a Transcription Factor-dsDNA Complex. *Anal. Chem.* **94**:7, 3203–3210 (2022).

107. Fenton, H. J. H.: LXXIII.—Oxidation of tartaric acid in presence of iron. *J. Chem. Soc., Trans.* **65**:0, 899–910 (1894).

108. Goldstein, S.; Meyerstein, D.; Czapski, G.: The Fenton reagents. *Free Radical Biol. Med.* **15**:4, 435–445 (1993).

109. Haber, F.; Weiss, J.: Über die Katalyse des Hydroperoxydes. *Naturwissenschaften* **20**:51, 948–950 (1932).

110. Tullius, T. D.; Greenbaum, J. A.: Mapping nucleic acid structure by hydroxyl radical cleavage. *Curr. Opin. Chem. Biol.* **9**:2, 127–134 (2005).

111. Cheal, S. M.; Ng, M.; Barrios, B.; Miao, Z.; Kalani, A. K.; Meares, C. F.: Mapping Protein−Protein Interactions by Localized Oxidation: Consequences of the Reach of Hydroxyl Radical. *Biochemistry* **48**:21, 4577–4586 (2009).

112. Wardman, P.: Application of pulse radiolysis methods to study the reactions and structure of biomolecules. *Rep. Prog. Phys.* **41**:2, 259 (1978).

113. Simic, M. G.: [2] Pulse radiolysis in study of oxygen radicals. In: *Methods in Enzymology*. Abelson, J. N.; Simon, M. I.; Sies, H.; Glazer, A. N. (Ed.). New York: Academic Press, 1990 p. 186, 89–100. ISBN 978-0121820879.

114. Xu, G.; Kiselar, J.; He, Q.; Chance, M. R.: Secondary Reactions and Strategies To Improve Quantitative Protein Footprinting. *Anal. Chem.* **77**:10, 3029–3037 (2005).

115. Sclavi, B.; Woodson, S.; Sullivan, M.; Chance, M. R.; Brenowitz, M.: Time-resolved synchrotron X-ray "footprinting", a new approach to the study of nucleic acid structure and function: application to protein-DNA interactions and RNA folding 1. *J. Mol. Biol.* **266**:1, 144–159 (1997).

116. Kiselar, J. G.; Maleknia, S. D.; Sullivan, M.; Downard, K. M.; Chance, M. R.: Hydroxyl radical probe of protein surfaces using synchrotron X-ray radiolysis and mass spectrometry. *Int. J. Radiat. Biol.* **78**:2, 101–114 (2002).

117. Weeks, J. L.; Matheson, M. S.: The Primary Quantum Yield of Hydrogen Peroxide Decomposition1. *J. Am. Chem. Soc.* **78**:7, 1273–1278 (1956).

118. Elliot, A. J.: *Rate Constants and G-Values for the Simulation of the Radiolysis of Light Water over the Range 0-300 (Degrees) C.* Chalk River Laboratories, AECL-11073. (1994).

119. Sharp, J. S.; Becker, J. M.; Hettich, R. L.: Analysis of Protein Solvent Accessible Surfaces by Photochemical Oxidation and Mass Spectrometry. *Anal. Chem.* **76**:3, 672–683 (2004).

120. Aye, T. T.; Low, T. Y.; Sze, S. K.: Nanosecond Laser-Induced Photochemical Oxidation Method for Protein Surface Mapping with Mass Spectrometry. *Anal. Chem.* **77**:18, 5814–5822 (2005).

121. Li, K. S.; Shi, L.; Gross, M. L.: Mass Spectrometry-Based Fast Photochemical Oxidation of Proteins (FPOP) for Higher Order Structure Characterization. *Acc. Chem. Res.* **51**:3, 736 (2018).

122. Zhang, Y.; Rempel, D. L.; Zhang, H.; Gross, M. L.: An Improved Fast Photochemical Oxidation of Proteins (FPOP) Platform for Protein Therapeutics. *J. Am. Soc. Mass Spectrom.* **26**:3, 526–529 (2015).

123. Espino, J. A.; King, C. D.; Jones, L. M.; Robinson, R. A. S.: In Vivo Fast Photochemical Oxidation of Proteins Using Enhanced Multiplexing Proteomics. *Anal. Chem.* **92**:11, 7596–7603 (2020).

124. Vahidi, S.; Konermann, L.: Probing the Time Scale of FPOP (Fast Photochemical Oxidation of Proteins): Radical Reactions Extend Over Tens of Milliseconds. *J. Am. Soc. Mass Spectrom.* **27**:7, 1156–1164 (2016).

125. Wagner, W. J.; Moyle, A. B.; Wagner, N. D.; Rempel, D. L.; Gross, M. L.: Evaluating Chemical Footprinting-Induced Perturbation of Protein Higher Order Structure. *Anal. Chem.* **96**:23, 9693–9703 (2024).

126. Maleknia, S. D.; Chance, M. R.; Downard, K. M.: Electrospray-assisted modification of proteins: a radical probe of protein structure. *Rapid Commun. Mass Spectrom.* **13**:23, 2352–2358 (1999).

127. King, P. A.; Jamison, E.; Strahs, D.; Anderson, V. E.; Brenowitz, M.: 'Footprinting' proteins on DNA with peroxonitrous acid. *Nucleic Acids Res.* **21**:10, 2473–2478 (1993).

128. Henn, A.; Halfon, J.; Kela, I.; Orion, I.; Sagi, I.: Nucleic acid fragmentation on the millisecond timescale using a conventional X-ray rotating anode source: application to protein–DNA footprinting. *Nucleic Acids Res.* **29**:24, e122 (2001).

129. Ottinger, L. M.; Tullius, T. D.: High-Resolution in Vivo Footprinting of a Protein−DNA Complex Using γ-Radiation. *J. Am. Chem. Soc.* **122**:24, 5901–5902 (2000).

130. Makino, K.; Mossoba, M. M.; Riesz, P.: Formation of ·OH and ·H in Aqueous Solutions by Ultrasound Using Clinical Equipment. *Radiat. Res.* **96**:2, 416–421 (1983).

131. Xiao, R.; Liu, K.; Bai, L.; Minakata, D.; Seo, Y.; Kaya Göktaş, R.; Dionysiou, D. D.; Tang, C.-J.; Wei, Z.; Spinney, R.: Inactivation of pathogenic microorganisms by sulfate radical: Present and future. *Chem. Eng. J. (Amsterdam, Neth.)* **371**, 222–232 (2019).

132. Bridgewater, J. D.; Vachet, R. W.: Using Microwave-Assisted Metal-Catalyzed Oxidation Reactions and Mass Spectrometry To Increase the Rate at Which the Copper-Binding Sites of a Protein Are Determined. *Anal. Chem.* **77**:14, 4649–4653 (2005).

133. McKenzie-Coe, A. A.; Johnson, D. T.; Peacock, R. B.; Zhang, Z.; Jones, L. M.: Evaluating the Sulfate Radical Anion as a New Reagent for In-Cell Fast Photochemical Oxidation of Proteins. *J. Am. Soc. Mass Spectrom.* **32**:7, 1644–1647 (2021).

134. Zhang, M. M.; Rempel, D. L.; Gross, M. L.: A Fast Photochemical Oxidation of Proteins (FPOP) platform for free-radical reactions: the carbonate radical anion with peptides and proteins. *Free Radical Biol. Med.* **131**, 126–132 (2019).

135. Richards, F. M.; Lamed, R.; Wynn, R.; Patel, D.; Olack, G.: Methylene as a possible universal footprinting reagent that will include hydrophobic surface areas: Overview and feasibility: Properties of diazirine as a precursor. *Prot. Sci.* **9**:12, 2506–2517 (2000).

136. Blencowe, A.; Hayes, W.: Development and application of diazirines in biological and synthetic macromolecular systems. *Soft Matter* **1**:3, 178–205 (2005).

137. Jumper, C. C.; Schriemer, D. C.: Mass Spectrometry of Laser-Initiated Carbene Reactions for Protein Topographic Analysis. *Anal. Chem.* **83**:8, 2913–2920 (2011).

138. Cavalieri, R. R.: Iodine Metabolism and Thyroid Physiology: Current Concepts. *Thyroid* **7**:2, 177–181 (1997).

139. Ly, T.; Julian, R. R.: Residue-Specific Radical-Directed Dissociation of Whole Proteins in the Gas Phase. *J. Am. Chem. Soc.* **130**:1, 351–358 (2008).

140. Chen, J.; Cui, W.; Giblin, D.; Gross, M. L.: New Protein Footprinting: Fast Photochemical Iodination Combined with Top-Down and Bottom-Up Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **23**:8, 1306–1318 (2012).

141. Chan, K. K. J.; O'Hagan, D.: Chapter Eleven - The Rare Fluorinated Natural Products and Biotechnological Prospects for Fluorine Enzymology. In: *Methods in Enzymology*. Hopwood, D. A. (Ed.). New York: Academic Press, 2012, p. 219–235. ISBN 978-0123942913.

142. Cheng, M.; Zhang, B.; Cui, W.; Gross, M. L.: Laser-initiated Radical Trifluoromethylation of Peptides and Proteins and Its Application to Mass Spectrometry-Based Protein Footprinting HHS Public Access. *Angew. Chem., Int. Ed. Engl.* **56**:45, 14007–14010 (2017).

143. Fujiwara, Y.; Dixon, J. A.; O'Hara, F.; Funder, E. D.; Dixon, D. D.; Rodriguez, R. A.; Baxter, R. D.; Herlé, B.; Sach, N.; Collins, M. R.; Ishihara, Y.; Baran, P. S.: Practical and innate carbon–hydrogen functionalization of heterocycles. *Nature* **492**:7427, 95–99 (2012).

144. Fojtík, L.; Fiala, J.; Pompach, P.; Chmelík, J.; Matoušek, V.; Beier, P.; Kukačka, Z.; Novák, P.: Fast Fluoroalkylation of Proteins Uncovers the Structure and Dynamics of Biological Macromolecules. *J. Am. Chem. Soc.* **143**:49, 20670–20679 (2021).

145. Jain, R.; Dhillon, N. S.; Farquhar, E. R.; Wang, B.; Li, X.; Kiselar, J.; Chance, M. R.: Multiplex chemical labeling of amino acids for protein footprinting structure assessment. *Anal. Chem.* **94**:27, 9819–9825 (2022).

146. Rahimidashaghoul, K.; Klimánková, I.; Hubálek, M.; Korecký, M.; Chvojka, M.; Pokorný, D.; Matoušek, V.; Fojtík, L.; Kavan, D.; Kukačka, Z.; Novák, P.; Beier, P.: Reductant-Induced Free Radical Fluoroalkylation of Nitrogen Heterocycles and Innate Aromatic Amino Acid Residues in Peptides and Proteins. *Chemistry* **25**:69, 15779–15785 (2019).

147. Johnson, R. S.; Biemann, K.: Computer program (SEQPEP) to aid in the interpretation of high-energy collision tandem mass spectra of peptides. *Biomed. Environ. Mass Spectrom.* **18**:11, 945–957 (1989).

148. Hines, W. M.; Falick, A. M.; Burlingame, A. L.; Gibson, B. W.: Pattern-based algorithm for peptide sequencing from tandem high energy collision-induced dissociation mass spectra. *J. Am. Soc. Mass Spectrom.* **3**:4, 326–336 (1992).

149. Eng, J. K.; McCormack, A. L.; Yates, J. R.: An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**:11, 976–989 (1994).

150. Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S.: Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**:18, 3551–3567 (1999).

151. Craig, R.; Beavis, R. C.: TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**:9, 1466–1467 (2004).

152. Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M.: Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J. Proteome Res.* **10**:4, 1794–1805 (2011).

153. Zhang, J.; Xin, L.; Shan, B.; Chen, W.; Xie, M.; Yuen, D.; Zhang, W.; Zhang, Z.; Lajoie, G. A.; Ma, B.: PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. *Mol. Cell. Proteom.* **11**:4, (2012).

154. Bern, M.; Kil, Y. J.; Becker, C.: Byonic: Advanced Peptide and Protein Identification Software. *Curr. Protoc. Bioinf.* **40**:1, 13.20.1-13.20.14 (2012).

155. Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H.: Open Mass Spectrometry Search Algorithm. *J. Proteome Res.* **3**:5, 958–964 (2004).

156. Amir, S. H.; Yuswan, M. H.; Aizat, W. M.; Mansor, M. K.; Desa, M. N. M.; Yusof, Y. A.; Song, L. K.; Mustafa, S.: Comparative database search engine analysis on massive tandem mass spectra of pork-based food products for halal proteomics. J. *Proteomics* **241**, 104240 (2021).

157. Nesvizhskii, A. I.; Vitek, O.; Aebersold, R.: Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **4**:10, 787–797 (2007).

158. Verheggen, K.; Martens, L.; Berven, F. S.; Barsnes, H.; Vaudel, M.: Database Search Engines: Paradigms, Challenges and Solutions. In: *Modern Proteomics – Sample Preparation, Analysis and Practical Applications.* Mirzaei, H.; Carrasco, M. (Ed.). Switzerland: Springer International Publishing, 2016, p. 147–156, ISBN 978-3-319-41446-1.

159. Benjamini, Y.; Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**:1, 289–300 (1995).

160. Elias, J. E.; Gygi, S. P.: Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**:3, 207–214 (2007).

161. Storey, J. D.; Tibshirani, R.: Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* **100**:16, 9440–9445 (2003).

162. Tsou, C.-C.; Avtonomov, D.; Larsen, B.; Tucholska, M.; Choi, H.; Gingras, A.-C.; Nesvizhskii, A. I.: DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **12**:3, 258–264 (2015).

163. Li, Y.; Zhong, C.-Q.; Xu, X.; Cai, S.; Wu, X.; Zhang, Y.; Chen, J.; Shi, J.; Lin, S.; Han, J.: Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files. *Nat. Methods* **12**:12, 1105–1106 (2015).

164. Ting, Y. S.; Egertson, J. D.; Bollinger, J. G.; Searle, B. C.; Payne, S. H.; Noble, W. S.; MacCoss, M. J.: PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nat. Methods* **14**:9, 903–908 (2017).

165. Tran, N. H.; Qiao, R.; Xin, L.; Chen, X.; Liu, C.; Zhang, X.; Shan, B.; Ghodsi, A.; Li, M.: Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nat. Methods* **16**:1, 63–66 (2019).

166. Rosenberger, G.; Bludau, I.; Schmitt, U.; Heusel, M.; Hunter, C. L.; Liu, Y.; MacCoss, M. J.; MacLean, B. X.; Nesvizhskii, A. I.; Pedrioli, P. G. A.; Reiter, L.; Röst, H. L.; Tate, S.; Ting, Y. S.; Collins, B. C.; Aebersold, R.: Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat. Methods* **14**:9, 921–927 (2017).

167. Zhang, F.; Ge, W.; Huang, L.; Li, D.; Liu, L.; Dong, Z.; Xu, L.; Ding, X.; Zhang, C.; Sun, Y.; A, J.; Gao, J.; Guo, T.: A Comparative Analysis of Data Analysis Tools for Data-Independent Acquisition Mass Spectrometry. *Mol. Cell. Proteomics* **22**:9, (2023).

168. Gotti, C.; Roux-Dalvai, F.; Joly-Beauparlant, C.; Mangnier, L.; Leclercq, M.; Droit, A.: Extensive and Accurate Benchmarking of DIA Acquisition Methods and Software Tools Using a Complex Proteomic Standard. *J. Proteome Res.* **20**:10, 4801–4814 (2021).

169. Heller, S. R.: Conversational mass spectral retrieval system and its use as an aid in structure determination. *Anal. Chem.* **44**:12, 1951–1961 (1972).

170. Frewen, B. E.; Merrihew, G. E.; Wu, C. C.; Noble, W. S.; MacCoss, M. J.: Analysis of Peptide MS/MS Spectra from Large-Scale Proteomics Experiments Using Spectrum Libraries. *Anal. Chem.* **78**:16, 5678–5684 (2006).

171. Craig, R.; Cortens, J. C.; Fenyo, D.; Beavis, R. C.: Using Annotated Peptide Mass Spectrum Libraries for Protein Identification. *J. Proteome Res.* **5**:8, 1843–1849 (2006).

172. Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R.: Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**:5, 655–667 (2007).

173. Craig, R.; Cortens, J. P.; Beavis, R. C.: Open Source System for Analyzing, Validating, and Storing Protein Identification Data. *J. Proteome Res.* **3**:6, 1234–1242 (2004).

174. Desiere, F.; Deutsch, E. W.; King, N. L.; Nesvizhskii, A. I.; Mallick, P.; Eng, J.; Chen, S.; Eddes, J.; Loevenich, S. N.; Aebersold, R.: The PeptideAtlas project. *Nucleic Acids Res.* **34**, D655–D658 (2006).

175. Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; Stein, S. E.; Aebersold, R.: Building consensus spectral libraries for peptide identification in proteomics. *Nat. Methods* **5**:10, 873–875 (2008).

176. Dasari, S.; Chambers, M. C.; Martinez, M. A.; Carpenter, K. L.; Ham, A.-J. L.; Vega-Montoto, L. J.; Tabb, D. L.: Pepitome: evaluating improved spectral library search for identification complementarity and quality assessment. *J. Proteome Res.* **11**:3, 1686–1695 (2012).

177. Lam, H.; Deutsch, E. W.; Aebersold, R.: Artificial Decoy Spectral Libraries for False Discovery Rate Estimation in Spectral Library Searching in Proteomics. *J. Proteome Res.* **9**:1, 605–610 (2010).

178. Cheng, C.-Y.; Tsai, C.-F.; Chen, Y.-J.; Sung, T.-Y.; Hsu, W.-L.: Spectrum-based Method to Generate Good Decoy Libraries for Spectral Library Searching in Peptide Identifications. *J. Proteome Res.* **12**:5, 2305–2310 (2013).

179. Shao, W.; Zhu, K.; Lam, H.: Refining similarity scoring to enable decoy-free validation in spectral library searching. *Proteomics* **13**:22, 3273–3283 (2013).

180. Deutsch, E. W.: Tandem Mass Spectrometry Spectral Libraries and Library Searching. In: *Data Mining in Proteomics*. Hamacher, M.; Eisenacher, M.; Stephan, C. (Ed.). Totowa: Humana Press, 2011, p. 225–232. ISBN 978-1-60761-987-1.

181. Röst, H. L.; Rosenberger, G.; Navarro, P.; Gillet, L.; Miladinović, S. M.; Schubert, O. T.; Wolski, W.; Collins, B. C.; Malmström, J.; Malmström, L.; Aebersold, R.: OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **32**:3, 219–223 (2014).

182. MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J.: Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**:7, 966–968 (2010).

183. Searle, B. C.; Pino, L. K.; Egertson, J. D.; Ting, Y. S.; Lawrence, R. T.; MacLean, B. X.; Villén, J.; MacCoss, M. J.: Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nat. Commun.* **9**:1, 5128 (2018).

184. Bruderer, R.; Bernhardt, O. M.; Gandhi, T.; Miladinović, S. M.; Cheng, L.-Y.; Messner, S.; Ehrenberger, T.; Zanotelli, V.; Butscheid, Y.; Escher, C.; Vitek, O.; Rinner, O.; Reiter, L.: Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues. *Mol. Cell. Proteomics* **14**:5, 1400–1410 (2015).

185. Fernández-Costa, C.; Martínez-Bartolomé, S.; McClatchy, D.; Yates, J. R. I.: Improving Proteomics Data Reproducibility with a Dual-Search Strategy. *Anal. Chem.* **92**:2, 1697–1701 (2020).

186. Yuan, Z. F.; Lin, S.; Molden, R. C.; Garcia, B. A.: Evaluation of proteomic search engines for the analysis of histone modifications. *J. Proteome Res.* **13**:10, 4470–4478 (2014).

187. Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I.: MSFragger: ultrafast and comprehensive peptide identification in shotgun proteomics. *Nat. Methods* **14**:5, 513–520 (2017).

188. Rojas Ramírez, C.; Espino, J. A.; Jones, L. M.; Polasky, D. A.; Nesvizhskii, A. I.: Efficient Analysis of Proteome-Wide FPOP Data by FragPipe. *Anal. Chem.* **95**:44, 16131–16137 (2023).

189. Stødkilde, K.; Torvund-Jensen, M.; Moestrup, S. K.; Andersen, C. B. F.: Structural basis for trypanosomal haem acquisition and susceptibility to the host innate immune system. *Nat. Commun* **5**:1, 5487 (2014).

190. Škerlová, J.; Brynda, J.; Šobotník, J.; Zákopčaník, M.; Novák, P.; Bourguignon, T.; Sillam-Dussès, D.; Řezáčová, P.: Crystal structure of blue laccase BP76, a unique termite suicidal defense weapon. *Structure* (2024) doi:10.1016/j.str.2024.07.015.

191. Eyre, D. R.; Paz, M. A.; Gallop, P. M.: Cross-Linking in Collagen and Elastin. *Annu. Rev. Biochem.* **53**, 717–748 (1984).

192. Tomášková, N.; Novák, P.; Kožár, T.; Petrenčáková, M.; Jancura, D.; Yassaghi, G.; Man, P.; Sedlák, E.: Early modification of cytochrome *c* by hydrogen peroxide triggers its fast degradation. *Int. J. Biol. Macromol.* **174**, 413–423 (2021).

193. Yin, V.; Shaw, G. S.; Konermann, L.: Cytochrome c as a Peroxidase: Activation of the Precatalytic Native State by H2O2-Induced Covalent Modifications. *J. Am. Chem. Soc.* **139**:44, 15701–15709 (2017).

DECLARATION OF AUTHORSHIP

1. **Zakopcanik, M.;** Kavan, D.; Novak, P. & Loginov, D. S. Quantifying the Impact of the Peptide Identification Framework on the Results of Fast Photochemical Oxidation of Protein Analysis. *J. Proteome Res.* **23**, 609–617 (2024).

I declare that Marek Zákopčaník contributed around 90% to Publication I. He processed and analysed the data, chose and implemented the metrics for comparison, and wrote and prepared the manuscript.

2. **Zakopcanik, M.;** Kavan, D.; Kukacka, Z.; Novak, P. & Loginov, D. S. Data-Independent Acquisition Represents a Promising Alternative for Fast Photochemical Oxidation of Proteins (FPOP) Samples Analysis. *Anal. Chem.* **96**, 11273–11279 (2024).

I declare that Marek Zákopčaník contributed around 80% to Publication II. He composed and optimised the processing workflow, analysed the data, evaluated the tested approaches, and wrote and prepared majority of the manuscript.

3. Škerlová, J.; Brynda, J.; Šobotník, J.; **Zákopčaník, M.**; Novák, P.; Bourguignon, T.; Sillam-Dussès, D. & Řezáčová, P. Crystal Structure of Blue Laccase BP76, a Unique Termite Suicidal Defense Weapon. *Structure* (2024).

I declare that Marek Zákopčaník contributed around 10% to Publication III. He optimised and conducted the open search of the MS data, interpreted the spectra and prepared respective parts of the manuscript (methods, results and figures).

In Vestec 25.9.2024

RNDr. Petr Novák, Ph.D.

Attached publications

Publication I

Zakopcanik, M.; Kavan, D.; Novak, P. & Loginov, D. S.

Quantifying the Impact of the Peptide Identification Framework on the Results of Fast Photochemical Oxidation of Protein Analysis.

Journal of Proteome Research, 2024

My contribution: processing and analysis of the data, implementation of the comparison metrics, writing and preparation of the manuscript

Journal of proteome research

pubs.acs.org/jpr

Article

Open Access

This article is licensed under CC-BY 4.0

Quantifying the Impact of the Peptide Identification Framework on the Results of Fast Photochemical Oxidation of Protein Analysis

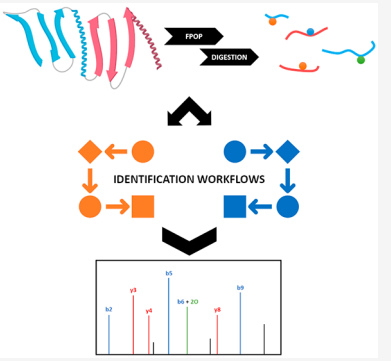
Marek Zakopcanik, Daniel Kavan, Petr Novak, and Dmitry S. Loginov*

Cite This: *J. Proteome Res.* 2024, 23, 609−617

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Fast Photochemical Oxidation of Proteins (FPOP) is a promising technique for studying protein structure and dynamics. The quality of insight provided by FPOP depends on the reliability of the determination of the modification site. This study investigates the performance of two search engines, Mascot and PEAKS, for the data processing of FPOP analyses. Comparison of Mascot and PEAKS of the hemoglobin−-haptoglobin Bruker timsTOF data set (PXD021621) revealed greater consistency in the Mascot identification of modified peptides, with around 26% of the IDs being mutual for all three replicates, compared to approximately 22% for PEAKS. The intersection between Mascot and PEAKS results revealed a limited number (31%) of shared modified peptides. Principal Component Analysis (PCA) using the peptide-spectrum match (PSM) score, site probability, and peptide intensity was applied to evaluate the results, and the analyses revealed distinct clusters of modified peptides. Mascot showed the ability to assess confident site determination, even with lower PSM scores. However, high PSM scores from PEAKS did not guarantee a reliable determination of the modification site. Fragmentation coverage of the modification position played a crucial role in Mascot assignments, while the AScore localizations from PEAKS often become ambiguous because the software employs MS/MS merging.



KEYWORDS: *FPOP, search engine, structural proteomics*

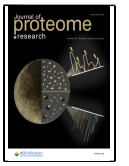
Downloaded via Marek Zakopcanik on June 19, 2024 at 11:03:35 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

INTRODUCTION

Information about the higher-order structure of proteins is vital for our understanding of their function. Mass spectrometry-based methods offer a wide range of approaches to study protein structure,[1] interactions,[2] and dynamics.[3] One of these approaches is the covalent labeling of proteins. Covalent labeling might be implemented by different means. For example, the common technique hydrogen/deuterium exchange (HDX) relies on the exchange of amide hydrogens on the backbone of the protein.[4] The efficiency of this method of labeling is based on solvent accessibility and the local hydrogen bonding network. HDX already has a well-established workflow and numerous data analysis software packages.[5−8] Other methods of protein footprinting use reactive radical species such as fluoroalkyl radicals,[9,10] carbonate radical,[11] carbene diradical,[12] and hydroxyl radical.[13]

Fast Photochemical Oxidation of Proteins (FPOP) employs labeling of amino acid side chains by hydroxyl radicals.[13] The covalent modification of the residues is irreversible. Moreover, the •OH properties are similar to those of the H_2O molecules, offering a biologically relevant image of the protein state.[14] The nonspecific nature of this reaction offers theoretically no limit in coverage, but in practice, the difference in reactivity between the amino acid side chains and the •OH varies by 10^3,[14,15] resulting in the predominance of more reactive residue modifications (Cys, Met, Trp, Tyr, Phe, His) over less reactive (aliphatic and other polar) amino acids.[14] Therefore, the extent of modification is not proportional to just the solvent accessibility but depends on the side chain reactivity as well. Furthermore, high confidence in the modification position is necessary for reliable and relevant results. While FPOP continues to develop, most recently via a top-down proteomic approach,[16,17] the bottom-up approach is most commonly used.[18,19] Currently, FPOP analysis is held back by tedious data processing, and the workflow has not yet been unified like in the case of HDX. A recent protocol by Liu et al.[20] introduces guidelines to establish an FPOP platform but is limited in the number of modifications used and relies on Byos (Protein Metrics, CA, USA), which includes the FPOP workflow as part of the suite and cannot be obtained separately. One of the other proprietary programs used for FPOP data analysis is Sequest, but it is restricted in the number of possible modifications, thus requiring separate consecutive searches.[21] The recently introduced algorithm update for cloud-based search engine Bolt promises an

Received: June 30, 2023
Revised: October 17, 2023
Accepted: December 6, 2023
Published: December 29, 2023



ACS Publications

© 2023 The Authors. Published by American Chemical Society

https://doi.org/10.1021/acs.jproteome.3c00390
J. Proteome Res. 2024, 23, 609−617

increased number of modifications while decreasing the processing time[22] but so far lacks adaptation. There are two open source programs that might be used for FPOP data processing: MSS-clean[23] and PepFoot.[24] Their primary designation is footprinting experiments with a specific covalent probe, thus being able to run analysis with only one variable mass shift at a time. MSFragger,[25] Comet,[26] and MS-GF+[27] are standard open source proteomic search engines that allow enough variable modifications to be suitable for FPOP, but they require employment of further processing modules to provide certainty assessment for the modification site.

The steps of FPOP data processing involve a database search, identification of the compounds from the MS/MS spectra, assignment of chromatographic peaks, and quantification of the signal intensity from the LC chromatogram. As previously published,[28] the choice of search engine can significantly influence the identification of modified peptides in a data set. We decided to explore this aspect of data processing for FPOP analyses. Two representatives of different algorithmic approaches were chosen for this experiment: Mascot,[29] as a purely database search-based software, and PEAKS,[30] which employs *de novo* sequencing as a workflow improvement. Mascot searches the MS/MS data provided in the form of a peak list against a database of choice. Each peptide-spectrum match (PSM) is assigned a score based on the probability that the observed match is a random event. Confidence in the localization of the peptide modification is expressed as a percentage value derived from the difference in score between the two most probable sites.[31] The number of variable modifications that can be defined for a search is limited to 9. PEAKS accepts raw analysis files from a variety of vendors (Agilent, Bruker, Thermo, and Waters) and runs its own data refinement. In pursuit of more reliable identification, PEAKS performs additive merging of spectra based on the precursor m/z and time window. Furthermore, it combines *de novo* and database search results together with additional features as an input for the internal Linear Discrimination Function to derive the p-value and ultimately score.[30] Modification localization is reported as an AScore, reflecting the probability of the reported position compared to all other possible positions. There is no limit to the number of variable modifications defined for the PEAKS search.

For the evaluation of search engine performance, an FPOP hemoglobin−haptoglobin data set from an experiment previously performed in our lab[32,33] was selected for the study. The data set is available on ProteomeXchange[34] (PXD021621). It represents a complex proteomic sample with features such as disulfide bonds and glycosylations, which must be taken into account during the sample treatment.[35] Steps such as carbamidomethylation of Cys and deglycosylation alter the peptide features and increase the complexity of the sample, raising the bar for search engines.

METHODS

Data Analysis

Data were searched against a database consisting of sequences of Hb (Uniprot IDs P69905 and P68871) and Hp (Uniprot ID P00738) supplemented with a deglycosylated form of β subunit, together with 246 sequences of common contaminants (including trypsin and catalase), using Mascot Server 2.7.0 software (Matrix Science Inc., Boston, MA, USA) and PEAKS Studio X+ software (Bioinformatics Solutions Inc., Waterloo, ON, Canada). The MGF files for Mascot were generated in Compass DataAnalysis v.5.3 (Bruker Daltonics, Billerice, MA, USA). The database was infused with reverse sequence decoys. Only tryptic peptides with up to two misscleavages were allowed. The tolerance of the precursor ions was set at 12 ppm, and the mass tolerance for the MS/MS fragment ions was set at 0.05 Da. The significance threshold was set to $p \leq 0.05$. The considered variable modifications are listed in Table S1. Only peptides with a single FPOP modification were considered for the analysis. This condition was applied because the FPOP modifications might alter the protein's structure, which could make solvent accessible areas that are not present in the native state of the protein. Peptide-spectrum matches were cut off by default values: $-10\log_{10}(P) \geq 2$ for Mascot and $-10\log_{10}(P) \geq 15$ for PEAKS. The intensities of the peptides were determined using Compass DataAnalysis v.5.3.

Due to the fact that Mascot's results were obtained as a list of all identifications, the modified peptides with multiple IDs were clustered in a 30 s retention time window, from which the best scoring ID was selected.

The intersections of the modifications detected within the replicates and between the Mascot and PEAKS results were visualized using Venn diagrams using the "matplotlib_venn"[36] Python package.

The results were subjected to Principal Component Analysis (PCA), where the parameters evaluated were score, site probability (MD-score for Mascot and AScore for PEAKS), and precursor intensity. For the purpose of linearization, $\log_{10}$ (AScore) was used for the PEAKS results. The PCA and subsequent visualization of the results have been carried out in R[37] using the "FactoMineR",[38] "factoextra",[39] and "corrplot"[40] packages.

RESULTS AND DISCUSSION

Modifications Detected within Replicates

Using Mascot and PEAKS, a peptide search was done separately for each replicate. The resulting lists of IDs only partially overlapped (Figure S1). In the hemoglobin (Hb) sample, Mascot identified a total of 603 modified peptides with 154 IDs being mutual for all three replicates (Table 1). In the haptoglobin (Hp) sample, a total of 407 modified peptides were identified, of which 107 were mutual. In the hemoglobin−haptoglobin (HbHp) complex sample, 1148 modified peptides were identified, of which 309 were mutual. As for PEAKS, the search of the Hb sample yielded a total of 588 modified peptides, of which 119 were mutual for all three replicates. For the Hp sample, 525 modified peptides were identified, of which 114 were mutual. In the sample of the HbHp complex, a total of 1216 modified peptides were identified, with 277 of them being mutual.

Table 1. Modified Peptides in Hb, Hp, and HbHp Complex Samples Identified by Mascot and PEAKS

	Mascot			PEAKS		
	all	mutual		all	mutual	
Hb	603	154	25.5%	588	119	20.2%
Hp	407	107	26.3%	525	114	21.7%
HbHp	1148	309	26.9%	1216	277	22.8%

https://doi.org/10.1021/acs.jproteome.3c00390
J. Proteome Res. 2024, 23, 609−617

The portion of mutual IDs to the total amount of IDs was higher in the case of Mascot, around 26%, in contrast to PEAKS, which averaged around 22%. This shows a higher consistency in the identification of modified peptides by Mascot. The Mascot results include occurrences of the same modified peptide (meaning type and position) from different retention times. This proves to be interesting, especially for encompassing all possible variants of modified peptides, which could reside on different atoms of the amino acid side chain, resulting in elution of the same peptide at multiple retention times (isobaric forms of modifications).[41]

For further evaluation, only IDs detected in all three replicates were accepted.

Intersection between Mascot and PEAKS

Considering only modified peptides identified in all 3 replicates, a comparison between Mascot and PEAKS was made. Matching IDs possessed the same type, site of modification, and retention time. Only a limited number of identified modified peptides were shared among search engines (Figure 1). It is worth noting that the proportion of mutual

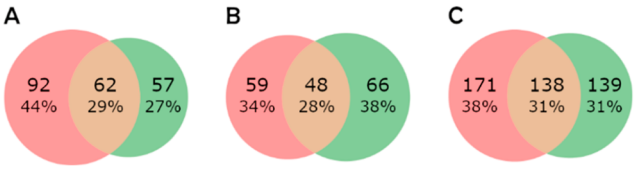


Figure 1. Intersections (orange) of IDs by Mascot (red) and PEAKS (green) in the (A) hemoglobin sample, (B) haptoglobin sample, and (C) HbHp complex sample.

IDs remains the same for analyses of single peptides as well as for the HbHp complex. This suggests that the IDs done by search engines are not related to the complexity of the data set but to the logic of identification.

The observed mutual exclusivity might be explained to some extent due to the Mascot search engine being limited in the number of possible defined variable modifications to 9. Based on our prior evaluation of the current data set,[32,33] together with previously published data processing setups[20] and described rates of reactivity,[14,15] the His → Asn, His → Asp, and trioxidation modifications were defined for the PEAKS search only (Table S1). His already has 4 different modifications defined, and we were more interested in Met −32 Da, as a modification of the most reactive residue. Arg deguanidination was included for both search engines to widen the spatial resolution of the method. Trioxidation was not used in Mascot, because it is an extensive modification occurring mainly on Cys that requires the reaction of 3 radicals,[18] reducing the relevance for structural information. Moreover, the inclusion of positional isomers in the Mascot results (mentioned above) further increases the number of unique IDs. The modifications, which were defined only in PEAKS searches, make up around one-third of the exclusive IDs. The setup of a PEAKS search with the same set of modifications as Mascot was tested, but the number of mutual IDs did not increase. Another reason for the differences in IDs could originate from fragmentation spectra, where the ions do not cover the exact position of the modification, and the precise position cannot be determined. In these cases, each of the search engines could tend to a different localization of the modification. Lastly, the results suggest that the search engines differ in the matter of identified modifications, with Mascot generally being able to detect more oxidations, carbonyls, and His +5 Da over PEAKS (Table 2).

Table 2. Types of Modifications Detected by Mascot and PEAKS in the HbHp Complex

	Mascot		PEAKS	
oxidation	144	47%	112	40%
dioxidation	36	12%	30	11%
carbonyl	61	20%	42	15%
decarboxylation	22	7%	21	8%
Met −32 Da	7	2%	4	1%
His +5 Da	19	6%	9	3%
His −10 Da	16	5%	10	4%
Arg deguanidination	4	1%	5	2%
His → Asp[a]	NA		13	5%
His → Asn[a]	NA		14	5%
trioxidation[a]	NA		17	6%

[a]Only defined for search in PEAKS.

Principal Component Analysis

Principal Component Analysis (PCA) was employed to further analyze the results of the selected search engines using the PSM score, the probability of the modification site, and the intensity of the peptide as criteria. The PCA revealed the grouping of IDs into distinct clusters and relations among the selected criteria. For each condition (Hb, Hp, and HbHp), the PCA yielded 3 dimensions, with the representation of supplied variables (Score, Site, and Intensity) being similar across conditions.

The PCA statistical models for hemoglobin-only and haptoglobin-only samples (Figures S2−S7) show a similar behavior (in terms of the ID's distribution and the representation of variables) as for the HbHp complex sample. For this reason, the HbHp complex sample was selected for further examination.

In the case of the HbHp complex, a statistical evaluation was performed for modified peptides of both proteins listed together. The results obtained by Mascot contained 175 modified peptides of Hb and 134 modified peptides of Hp. The first two dimensions yielded by PCA explained 44.4% and 32% of the data set variability, respectively. Dimension 1 represents mainly the Score and Intensity (both by roughly 60%), and the Site is represented by 15% (Figure S8A). The Score and Intensity exhibit a 75% positive correlation with Dimension 1 (Figure S9A). This alignment indicates a very high cross-correlation between Score and Intensity. Dimension 2 represents almost 85% of Site variability, with more than 92% positive correlation of Site with Dimension 2.

The distribution of the modified peptides in the PCA plot (Figure 2A) shows that most of the IDs are accumulated in a horizontal cluster at the top of the plot. A closer look shows that oxidations predominate toward the right side (alongside the Score and Site vector), in contrast to the left side, which is populated by a more diverse spectrum of modifications, including carbonyls, His −10 Da, and His +5 Da. On the basis of manual validation, the modified peptides from this cluster are identified with a high certainty, which corresponds to precise coverage of the modified residue in the MS/MS spectrum. An example is Trp oxidation in a peptide with the sequence SAVTALWGK in its singly charged form at m/z

https://doi.org/10.1021/acs.jproteome.3c00390
J. Proteome Res. 2024, 23, 609−617

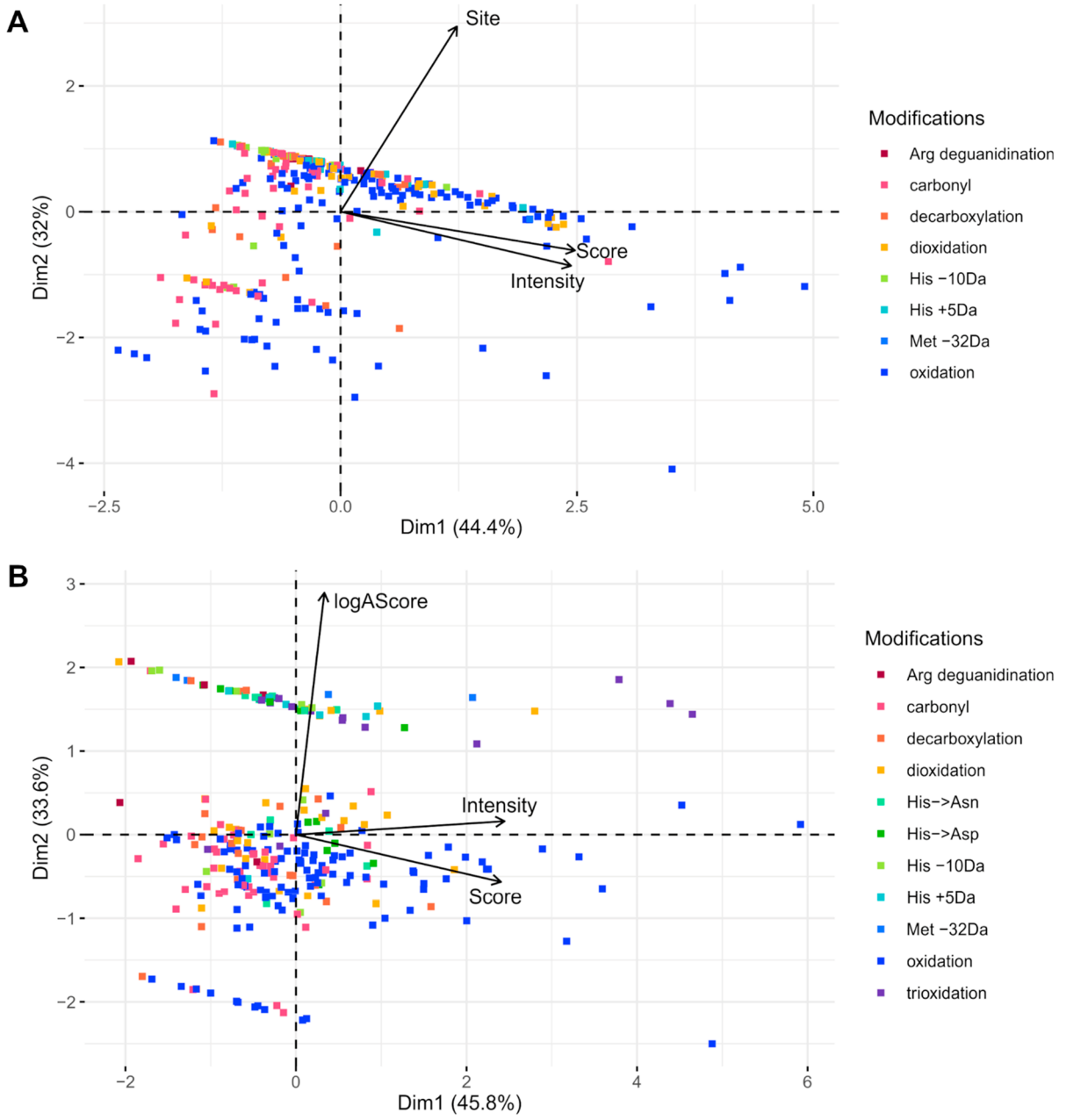


Figure 2. Principal component analysis of modified peptides identified in the HbHp complex sample by (A) Mascot and (B) PEAKS. The plot of color-coded modifications shows their distribution within dimensions defined by the PCA. The vectors show the correlation of the variables with the PCA dimensions. The logAScore variable represents the probability of site determination. In both cases, the IDs are distributed into 3 clusters. For Mascot, most of the IDs are in the top cluster, while for PEAKS, the majority of the IDs are in the middle cluster.

948.51 and RT 9.2 min (Figure 3A). The spectrum shows 10 fragment ions out of 16, including a b3−b8 ion series with b6 and b7 ions covering exactly the site of modification. The second linear cluster of IDs spanning the middle of the distribution shows once again that oxidations are mainly on the right side of the cluster, whereas carbonyls and dioxidations are located on the left side. The exact site of modifications falling into this group could not always be precisely determined due to an incomplete or unconvincing fragmentation spectrum. Leu carbonyl in the peptide VLSPADKTNVK in its doubly charged form at m/z 593.33 at RT 12.5 min (Figure 3B) is such a case. The spectrum contains only y-ions, with y10 missing, leaving both Val1 and Leu2 as possible sites of modification. The IDs scattered in the bottom area of the PCA plot (mainly oxidations) are subject to great uncertainty with insufficient peptide fragment coverage, for example, Tyr10 oxidation in a peptide with the sequence DYAEVGRVGY-VSGWGR in its triply charged form at m/z 596.29 and RT 9.8 min (Figure 3C).

Because there is no fragmentation coverage between the y6 and y15 ions, there are 5 possible sites of modification, making this ID unusable. Given these points, the PCA model for Mascot shows a good separation of reliable, questionable, and completely unreliable IDs.

As for the PEAKS search of the HbHp complex, the input list contained 150 Hb modified peptides and 124 Hp modified

https://doi.org/10.1021/acs.jproteome.3c00390
J. Proteome Res. 2024, 23, 609−617

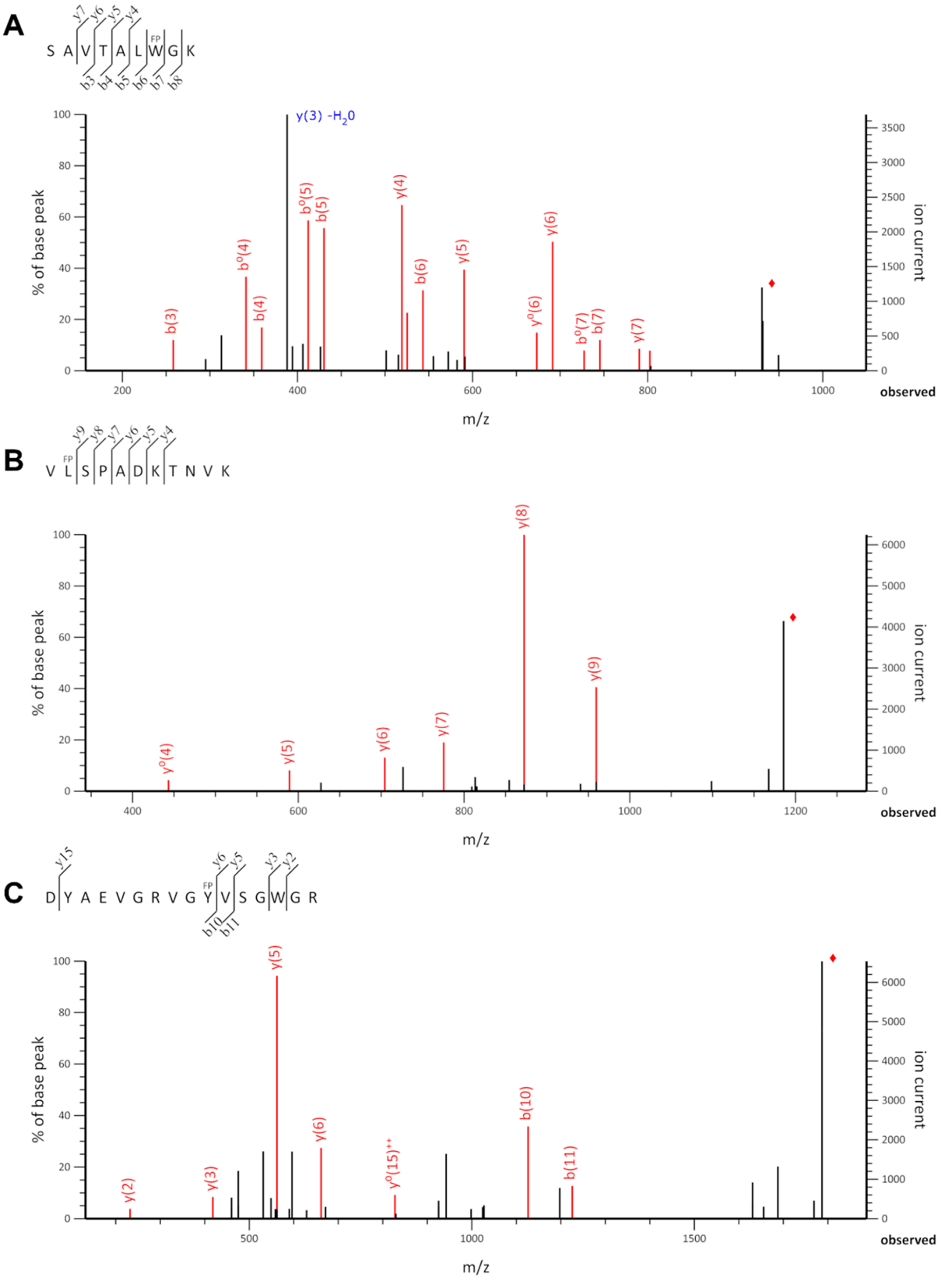


Figure 3. Fragmentation spectra of selected IDs from the PCA model of the Mascot search results for the HbHp complex. The ♦ sign labels the precursor ion. (A) Trp7 oxidation (+15.995) in the peptide SAVTALWGK in its singly charged form at m/z 948.51 and RT 9.2 min from the top cluster, y3 $-H_2O$ undetected by Mascot was annotated manually; (B) Leu2 carbonyl (+13.979) in the peptide VLSPADKTNVK in its doubly charged form at m/z 593.33 and RT 12.5 min from the middle cluster; (C) Tyr10 oxidation (+15.995) in the peptide DYAEVGRVGYVSGWGR in its triply charged form at m/z 596.29 and RT 9.8 min from the bottom cluster. The representative IDs show that the PCA is able to separate reliable, questionable, and unreliable modified peptides.

https://doi.org/10.1021/acs.jproteome.3c00390
J. Proteome Res. 2024, 23, 609−617

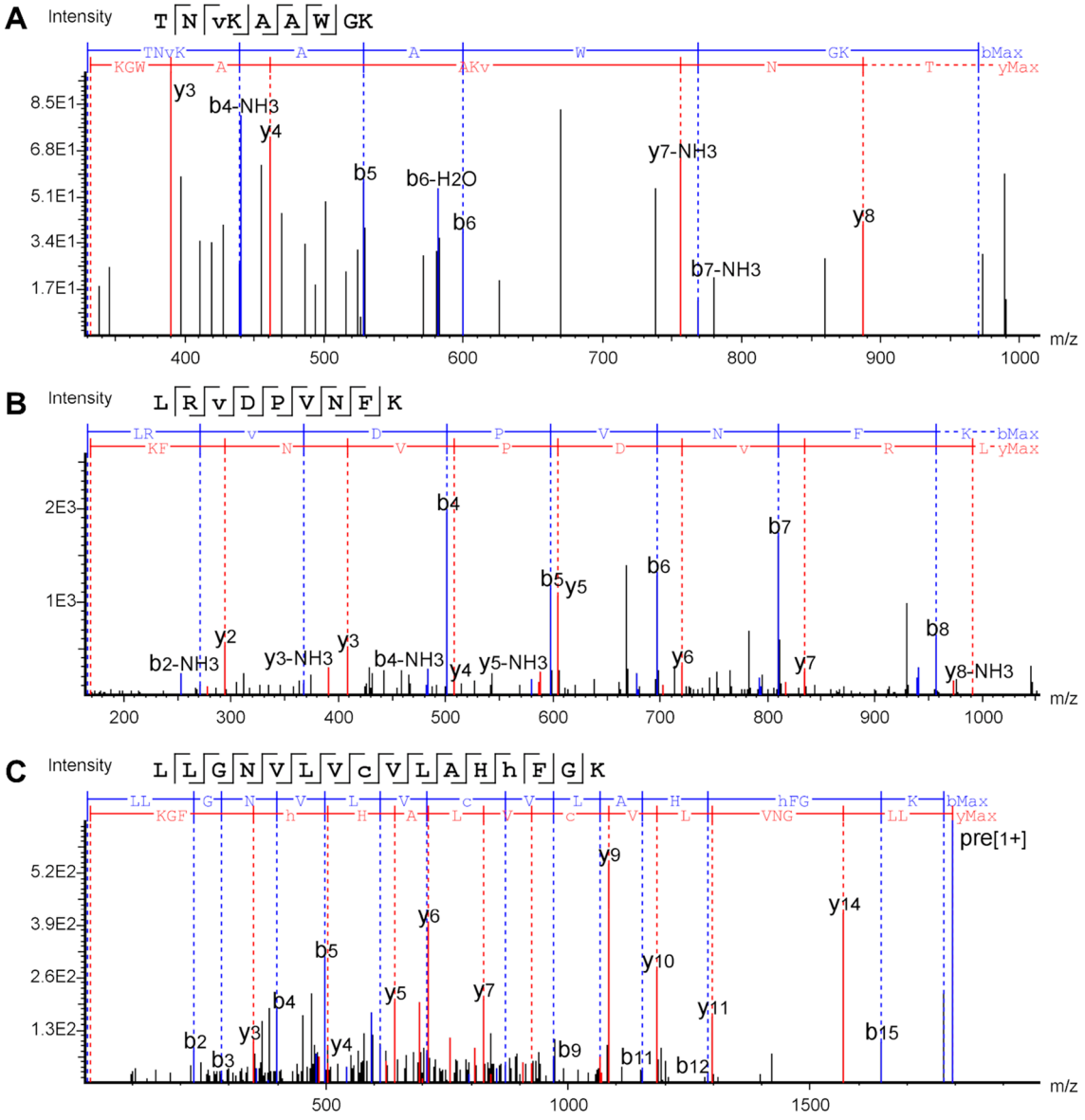


Figure 4. Fragmentation spectra of selected IDs from the PCA model of the PEAKS search results for the HbHp complex: (A) Val3 carbonyl (+13.979) in the peptide TNVKAAWGK in its singly charged form at m/z 988.52 at RT 7.27 min from the top cluster; (B) Val3 oxidation (+15.995) in the peptide with the sequence LRVDPVNFK in its doubly charged form at m/z 552.31 at RT 8.19 min from the middle cluster; (C) His13 oxidation (+15.995) in the peptide LLGNVLVCVLAHHFGK in its doubly charged form at m/z 897.00 at RT 21.01 min from the bottom cluster. The selected IDs show that the fragmentation coverage of the site of modification and signal intensity do not have a clear relation to the AScore and position within the PCA clusters.

peptides. Dimension 1 accounts for 45.4% of the variability, and Dimension 2 accounts for 33.3%. Dimension 1 represents Score and Intensity by 68% (Figure S8B), with both variables being positively correlated with Dimension 1 by over 82%, but the Score and Intensity variables show only partial cross-correlation. One possible explanation is the additive merging of MS/MS spectra. While fragments from several scans could contribute to the PSM score, the intensity of the compound in the chromatogram might be low. Dimension 2 represented the entirety of the logAScore variability (99.97%) and was fully positively correlated (99.99%). The PCA model (Figure 2B) shows separation into 3 clusters. Manual validation has shown that the top horizontal cluster consists predominantly of IDs that are unambiguous by definition. This means that there is only one residue in the peptide sequence that could carry out the modification. For illustration, the fragmentation spectrum of Val carbonyl in the TNVKAAWGK peptide in its singly charged form at m/z 988.52 at RT 7.27 min (Figure 4A) consists of very low intensity signals. Nevertheless, this ID has been assigned the highest certainty possible. The second cluster spans around the center of origin and includes most of the IDs. This cluster predominantly consists of oxidations, carbonyls, decarboxylations, and dioxidations. Closer inspection revealed that most of these IDs have a low AScore, although the fragment coverage seems sufficient.

For example, the fragmentation spectrum of Val3 oxidation in a peptide with the sequence LRVDPVNFK in its doubly charged form at m/z 552.31 at RT 8.19 min (Figure 4B) contains 14 out of 16 fragment ions; yet, the assigned AScore is 20.41. The third, linear cluster at the bottom consists of IDs

https://doi.org/10.1021/acs.jproteome.3c00390
J. Proteome Res. 2024, 23, 609−617

with absolute uncertainty of position (with AScore = 0), although the fragmentation spectra might cover the entire peptide sequence. For instance, His13 oxidation in the LLGNVLVCVLAHHFGK peptide in its doubly charged form at m/z 897.00 at RT 21.01 min (Figure 4C), whose fragmentation spectrum contains almost all possible fragments. Our explanation is that the merged spectrum contains a number of low-intensity signals, and the fragment ions determining the oxidation position to be His13 are among those low-intensity signals (y3, y4, b11, b12). Additionally, in the sequence scheme on the top, PEAKS indicates that there is supposedly a b14 fragment, but such a signal is not present in the spectrum. Furthermore, after manual validation of the spectrum, we were able to annotate ions 367.2, which corresponds to y3 carrying oxidation, and 1081.44, which corresponds to y9 carrying carbonyl. Thus, sometimes the AScore cannot be confident. These results show that, even though the PEAKS IDs are separated by PCA into distinct clusters, the relation to certainty in site determination is inconclusive. From the modified peptides identified by both Mascot and PEAKS, 4 representatives have been selected for comparison. First, Met15 oxidation in a peptide with the sequence VADALTNAVAHVDDMPNALSALSDLHAHK in its triply charged form at m/z 1004.84 and RT 16.61 min is a highly probable and extensive modification. Both search engines were able to identify numerous fragmentation ions (Figure S10) and thus assign a very high score (Mascot: 153.34; PEAKS: 105.68), but PEAKS did not account for relatively strong y14 and b18 ion signals assigned by Mascot. Furthermore, Mascot evaluated the certainty of oxidation localization as almost 100%, determining it to be located in the top cluster of the PCA plot. PEAKS, on the other hand, evaluated its certainty only as AScore 11.12, which ranks among the lower ones and means localization to the lower part of the middle cluster.

Second, the His11 carbonyl from the peptide with the sequence VVAGVANALAHK in its doubly charged form at m/z 582.33 and RT 9.44 min represents a low abundance modification. The identification done by Mascot comes from ions b4, y5, and y7−y10, which do not cover the site of modification (Figure S11A). The assigned score of 32.1 is among the lower ones. The coverage is not sufficient to differentiate between positions Leu9 and His11, yet Mascot sides with the latter, signaling 49.82% certainty, ranking within the middle cluster of the PCA plot. His11 carbonyl identified by PEAKS relies on the fragmentation spectrum with numerous clear y-ion signals, together with several b-ions on the level of noise (Figure S11B), resulting in a middle tier score of 71.05. Although PEAKS shows almost AA-level fragmentation sequencing, AScore 14.04 indicates not very high certainty of the position and localization to the lower part of the middle cluster.

Val5 oxidation in the VVAGVANALAHK peptide in its doubly charged form at m/z 583.34 at RT 5.9 min is the third example. In this case, Mascot assigned a high score of 102.79, while PEAKS assessed the identification at a score of 60.44, which represents the low end of the evaluation scale for PEAKS. The fragmentation spectrum used for identification by Mascot contains y-ions only, with just y1 and y8 ions missing (Figure S12A). Since the y7 and y9 ions border the Gly-Val residues and the oxidation is not defined for Gly, Mascot gives a site determination confidence of 99.99%, ranking as the top cluster ID. The identification by PEAKS relies on 10 (out of 11) y-ions and 8 b-ions (Figure S12B). Although the PSM score assigned by PEAKS is among the lower ones, the fragmentation coverage is sufficient to give the AScore of 26.57, meaning middle cluster in the PCA.

Finally, the Glu5 decarboxylation in the peptide KQLVEIEK in its monocharged form at m/z 956.58 at RT 5.43 min represents identification, which has a low score from both search engines (Mascot: 23.72; PEAKS: 50.06). The fragmentation spectrum for identification done by Mascot (Figure S13A) lacks precise coverage of the site of modification, with only low intensity b6-H_2O and b6-NH_3 ions differentiating between the Glu5 and Glu7 positions. This results in 68.98% confidence in site determination and localization to the lower part of the top cluster. The fragmentation spectrum for PEAKS identification (Figure S13B) contains 11 out of 14 possible fragments, but the signals for b2, b5, b6, y4, and y5 have a borderline noise-level intensity. The site determination is evaluated with a fairly confident AScore 24.32, meaning middle cluster.

These results demonstrate that Mascot is able to assess confident site determination for modifications, even in cases of a lower PSM score. On the other hand, a high PSM score of PEAKS IDs does not guarantee a reliable determination of the modification site. Similarly, IDs with lower PSM score can still result in a reliable localization of the modification. From a user point of view, Mascot assignments have a clear relation to the fragmentation coverage of the modification position, while the AScore value does not provide reliable guidance on the certainty of modification localization. The AScore evaluation appears to be low, even in cases with a high fragmentation coverage. Additionally, the highest possible confidence (AScore of 1000) is reachable only for unambiguously defined modifications. Furthermore, the additional merging performed by PEAKS yields its identifications and classification inconclusive, as there is little to no option to confirm PEAKS identifications against the raw data.

CONCLUSIONS

The results of this study show that the selection of the search engine for FPOP data evaluation has a significant impact on the detected modifications. We compared the consistency of selected search engines within replicates and compared results of the search engines with each other, where Mascot shows greater consistency over PEAKS. The number of identifications detected by both search engines was limited (31%). Principal Component Analysis revealed distinct clusters of modified peptides based on their scores, site probabilities, and intensities. Mascot identifications show a more confident site determination and a clear relationship between the quality of the fragmentation spectrum and the assigned site confidence. From the point of view of data evaluation, Mascot requires less manual validation. Most of the IDs belong to the top cluster and are reliable. Manual validation of the IDs from the middle cluster is advisable, with oxidations generally scoring better than carbonyls. Modifications below the middle cluster should be discarded. Clustering of PEAKS IDs does not relate clearly to the reliability of identification, which leads to the need for manual validation of the majority of the results. Furthermore, data refinement performed by PEAKS prevents backchecking with raw data.

Overall, this study provides insights into the performance and differences between Mascot and PEAKS in the analysis of FPOP data. We conclude that using Mascot for FPOP data

https://doi.org/10.1021/acs.jproteome.3c00390
J. Proteome Res. 2024, 23, 609−617

leads to more reliable and understandable IDs and requires less manual validation. Still, there is the need for further development and automatization of the data processing workflow for FPOP analysis.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.3c00390.

Defined variable modifications (Table S1), intersections of IDs for Hb and Hp samples (Figure S1), quality of variable representation for PCA of Hb sample (Figure S2), variable correlations for PCA of Hb sample (Figure S3), PCA for Hb sample (Figure S4), quality of variable representation for PCA of Hp sample (Figure S5), variable correlations for PCA of Hp sample (Figure S6), PCA for Hp sample (Figure S7), quality of variable representation for PCA of HbHp sample (Figure S8), variable correlations for PCA of HbHp (Figure S9), and fragmentation spectra of example modified peptides (Figures S10−S13) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Dmitry S. Loginov − *Institute of Microbiology, The Czech Academy of Sciences, 14220 Prague, Czech Republic;* orcid.org/0000-0002-4164-6145; Email: dmitry.loginov@biomed.cas.cz

Authors

Marek Zakopcanik − *Institute of Microbiology, The Czech Academy of Sciences, 14220 Prague, Czech Republic; Department of Biochemistry, Faculty of Science, Charles University, 12820 Prague, Czech Republic;* orcid.org/0009-0003-9039-7151

Daniel Kavan − *Institute of Microbiology, The Czech Academy of Sciences, 14220 Prague, Czech Republic*

Petr Novak − *Institute of Microbiology, The Czech Academy of Sciences, 14220 Prague, Czech Republic;* orcid.org/0000-0001-8688-529X

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.3c00390

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by Czech Scientific Foundation (22-27695S) and the EU H2020 program project EPIC-XS (grant agreement ID: 823839).

■ ABBREVIATIONS

FPOP, Fast Photochemical Oxidation of Proteins; Hb, hemoglobin; Hp, haptoglobin; HbHp, hemoglobin−haptoglobin; PSM, peptide-spectrum match; RT, retention time (min)

■ REFERENCES

(1) Pan, J.; Zhang, S.; Borchers, C. H. Comparative Higher-Order Structure Analysis of Antibody Biosimilars Using Combined Bottom-up and Top-down Hydrogen-Deuterium Exchange Mass Spectrometry. *Biochim. Biophys. Acta* **2016**, *1864* (12), 1801−1808.

(2) Liu, F.; Rijkers, D. T. S.; Post, H.; Heck, A. J. R. Proteome-Wide Profiling of Protein Assemblies by Cross-Linking Mass Spectrometry. *Nat. Methods* **2015**, *12*, 1179−1184.

(3) Leney, A. C.; Rafie, K.; Van Aalten, D. M. F.; Heck, A. J. R. Direct Monitoring of Protein O-GlcNAcylation by High-Resolution Native Mass Spectrometry. *ACS Chem. Biol.* **2017**, *12* (8), 2078−2084.

(4) Zhang, Z.; Smith, D. L. Determination of Amide Hydrogen Exchange by Mass Spectrometry: A New Tool for Protein Structure Elucidation. *Protein Sci.* **1993**, *2* (4), 522−531.

(5) Brown, K. A.; Wilson, D. J. Bottom-up Hydrogen Deuterium Exchange Mass Spectrometry: Data Analysis and Interpretation. *Analyst* **2017**, *142* (16), 2874−2886.

(6) Lau, A. M.; Claesen, J.; Hansen, K.; Politis, A. Deuteros 2.0: Peptide-Level Significance Testing of Data from Hydrogen Deuterium Exchange Mass Spectrometry. *Bioinformatics* **2021**, *37* (2), 270−272.

(7) Narang, D.; Lento, C.; Wilson, D. J. HDX-MS: An Analytical Tool to Capture Protein Motion in Action. *Biomedicines* **2020**, *8* (7), 224.

(8) Vávra, J.; Sergunin, A.; Stráňava, M.; Kádek, A.; Shimizu, T.; Man, P.; Martínková, M. Hydrogen/Deuterium Exchange Mass Spectrometry of Heme-Based Oxygen Sensor Proteins. *Methods Mol. Biol. Clifton NJ.* **2023**, *2648*, 99−122.

(9) Cheng, M.; Zhang, B.; Cui, W.; Gross, M. L. Laser-Initiated Radical Trifluoromethylation of Peptides and Proteins and Its Application to Mass Spectrometry-Based Protein Footprinting HHS Public Access. *Angew. Chem., Int. Ed. Engl.* **2017**, *56* (45), 14007−14010.

(10) Fojtík, L.; Fiala, J.; Pompach, P.; Chmelík, J.; Matoušek, V.; Beier, P.; Kukačka, Z.; Novák, P. Fast Fluoroalkylation of Proteins Uncovers the Structure and Dynamics of Biological Macromolecules. *J. Am. Chem. Soc.* **2021**, *143* (49), 20670−20679.

(11) Zhang, M. M.; Rempel, D. L.; Gross, M. L. A Fast Photochemical Oxidation of Proteins (FPOP) Platform for Free-Radical Reactions: The Carbonate Radical Anion with Peptides and Proteins. *Free Radic. Biol. Med.* **2019**, *131*, 126−132.

(12) Smith, R. A. G.; Knowles, J. R. Letter: Aryldiazirines. Potential Reagents for Photolabeling of Biological Receptor Sites. *J. Am. Chem. Soc.* **1973**, *95* (15), 5072−5073.

(13) Hambly, D. M.; Gross, M. L. Laser Flash Photolysis of Hydrogen Peroxide to Oxidize Protein Solvent-Accessible Residues on the Microsecond Timescale. *J. Am. Soc. Mass Spectrom.* **2005**, *16* (12), 2057−2063.

(14) Liu, X. R.; Zhang, M. M.; Gross, M. L. Mass Spectrometry-Based Protein Footprinting for Higher-Order Structure Analysis: Fundamentals and Applications. *Chem. Rev.* **2020**, *120* (10), 4355−4454.

(15) Xu, G.; Chance, M. R. Hydroxyl Radical-Mediated Modification of Proteins as Probes for Structural Proteomics. *Chem. Rev.* **2007**, *107* (8), 3514−3543.

(16) Yassaghi, G.; Kukačka, Z.; Fiala, J.; Kavan, D.; Halada, P.; Volný, M.; Novák, P. Top-Down Detection of Oxidative Protein Footprinting by Collision-Induced Dissociation, Electron-Transfer Dissociation, and Electron-Capture Dissociation. *Anal. Chem.* **2022**, *94* (28), 9993−10002.

(17) Polák, M.; Yassaghi, G.; Kavan, D.; Filandr, F.; Fiala, J.; Kukačka, Z.; Halada, P.; Loginov, D. S.; Novák, P. Utilization of Fast Photochemical Oxidation of Proteins and Both Bottom-up and Top-down Mass Spectrometry for Structural Characterization of a Transcription Factor-dsDNA Complex. *Anal. Chem.* **2022**, *94* (7), 3203−3210.

(18) Lu, Y.; Zhang, H.; Niedzwiedzki, D. M.; Jiang, J.; Blankenship, R. E.; Gross, M. L. Fast Photochemical Oxidation of Proteins Maps the Topology of Intrinsic Membrane Proteins: Light-Harvesting Complex 2 in a Nanodisc. *Anal. Chem.* **2016**, *88*, 8827.

(19) Charvátová, O.; Foley, B. L.; Bern, M. W.; Sharp, J. S.; Orlando, R.; Woods, R. J. Quantifying Protein Interface Footprinting by Hydroxyl Radical Oxidation and Molecular Dynamics Simulation:

Application to Galectin-1. *J. Am. Soc. Mass Spectrom.* **2008**, *19* (11), 1692−1705.
(20) Liu, X. R.; Rempel, D. L.; Gross, M. L. Protein Higher-Order-Structure Determination by Fast Photochemical Oxidation of Proteins and Mass Spectrometry Analysis. *Nat. Protoc.* **2020**, *15* (12), 3942−3970.
(21) Rinas, A.; Espino, J. A.; Jones, L. M. An Efficient Quantitation Strategy for Hydroxyl Radical-Mediated Protein Footprinting Using Proteome Discoverer. *Anal. Bioanal. Chem.* **2016**, *408* (11), 3021−3031.
(22) Chea, E. E.; Prakash, A.; Jones, L. M. The Utilization of the Search Engine, Bolt, to Decrease Search Time and Increase Peptide Identifications in Hydroxyl Radical Protein Footprinting-Based Workflows. *Proteomics* **2021**, *21* (21−22), 2000295.
(23) Ziemianowicz, D. S.; Sarpe, V.; Schriemer, D. C. Quantitative Analysis of Protein Covalent Labeling Mass Spectrometry Data in the Mass Spec Studio. *Anal. Chem.* **2019**, *91* (13), 8492−8499.
(24) Bellamy-Carter, J.; Oldham, N. J. PepFoot: A Software Package for Semiautomated Processing of Protein Footprinting Data. *J. Proteome Res.* **2019**, *18* (7), 2925−2930.
(25) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: Ultrafast and Comprehensive Peptide Identification in Shotgun Proteomics. *Nat. Methods* **2017**, *14* (5), 513−520.
(26) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: An Open-Source MS/MS Sequence Database Search Tool. *PROTEOMICS* **2013**, *13* (1), 22−24.
(27) Kim, S.; Pevzner, P. A. MS-GF+ Makes Progress towards a Universal Database Search Tool for Proteomics. *Nat. Commun.* **2014**, *5*, 5277.
(28) Yuan, Z. F.; Lin, S.; Molden, R. C.; Garcia, B. A. Evaluation of Proteomic Search Engines for the Analysis of Histone Modifications. *J. Proteome Res.* **2014**, *13* (10), 4470−4478.
(29) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data. *ELECTROPHORESIS* **1999**, *20* (18), 3551−3567.
(30) Zhang, J.; Xin, L.; Shan, B.; Chen, W.; Xie, M.; Yuen, D.; Zhang, W.; Zhang, Z.; Lajoie, G. A.; Ma, B. PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. *Mol. Cell. Proteomics* **2012**, *11* (4), M111.010587.
(31) Savitski, M. M.; Lemeer, S.; Boesche, M.; Lang, M.; Mathieson, T.; Bantscheff, M.; Kuster, B. Confident Phosphorylation Site Localization Using the Mascot Delta Score. *Mol. Cell. Proteomics* **2011**, *10* (2), S1−S12.
(32) Loginov, D. S.; Fiala, J.; Chmelik, J.; Brechlin, P.; Kruppa, G.; Novak, P. Benefits of Ion Mobility Separation and Parallel Accumulation-Serial Fragmentation Technology on timsTOF Pro for the Needs of Fast Photochemical Oxidation of Protein Analysis. *ACS Omega* **2021**, *6* (15), 10352−10361.
(33) Loginov, D. S.; Fiala, J.; Brechlin, P.; Kruppa, G.; Novak, P. Hydroxyl Radical Footprinting Analysis of a Human Haptoglobin-Hemoglobin Complex. *Biochim. Biophys. Acta - Proteins Proteomics* **2022**, *1870* (2), 140735.
(34) Deutsch, E. W.; Bandeira, N.; Perez-Riverol, Y.; Sharma, V.; Carver, J. J.; Mendoza, L.; Kundu, D. J.; Wang, S.; Bandla, C.; Kamatchinathan, S.; Hewapathirana, S.; Pullman, B. S.; Wertz, J.; Sun, Z.; Kawano, S.; Okuda, S.; Watanabe, Y.; MacLean, B.; MacCoss, M. J.; Zhu, Y.; Ishihama, Y.; Vizcaíno, J. A. The ProteomeXchange Consortium at 10 Years: 2023 Update. *Nucleic Acids Res.* **2023**, *51* (D1), D1539−D1548.
(35) Pompach, P.; Brnakova, Z.; Sanda, M.; Wu, J.; Edwards, N.; Goldman, R. Site-Specific Glycoforms of Haptoglobin in Liver Cirrhosis and Hepatocellular Carcinoma. *Mol. Cell. Proteomics* **2013**, *12* (5), 1281−1293.
(36) Tretyakov, K. Area-Weighted Venn-Diagrams for Python/Matplotlib; https://github.com/konstantint/matplotlib-venn (accessed 2023-03-03).
(37) R Core Team. R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing, 2022; https://www.R-project.org/.
(38) Lê, S.; Josse, J.; Husson, F. FactoMineR: An R Package for Multivariate Analysis. *J. Stat. Softw.* **2008**, *25* (1), 1−18.
(39) Kassambara, A.; Mundt, F. Factoextra: Extract and Visualize the Results of Multivariate Data Analyses; 2020; https://CRAN.R-project.org/package=factoextra (accessed 2023-03-03).
(40) Wei, T.; Simko, V. R Package "Corrplot": Visualization of a Correlation Matrix; 2021; https://github.com/taiyun/corrplot (accessed 2023-03-03).
(41) Cornwell, O.; Radford, S. E.; Ashcroft, A. E.; Ault, J. R. Comparing Hydrogen Deuterium Exchange and Fast Photochemical Oxidation of Proteins: A Structural Characterisation of Wild-Type and ΔN6 β 2-Microglobulin. *J. Am. Soc. Mass Spectrom.* **2018**, *29*, 2413−2426.

https://doi.org/10.1021/acs.jproteome.3c00390
J. Proteome Res. 2024, 23, 609−617

Supplementary data

Quantifying the Impact of the Peptide Identification Framework on the Results of Fast Photochemical Oxidation of Proteins Analysis

Marek Zakopcanik[1,2], Daniel Kavan[1], Petr Novak[1], Dmitry S. Loginov[*,1]

[1]Institute of Microbiology, The Czech Academy of Sciences, 14220 Prague, Czech Republic

[2]Faculty of Science, Charles University, 12843 Prague, Czech Republic

*Email: dmitry.loginov@biomed.cas.cz

List of Supplementary data

Defined variable modifications (Table S1), intersections of IDs for Hb and Hp samples (Figure S1), quality of variables representation for PCA of Hb sample (Figure S2), variables correlation for PCA of Hb sample (Figure S3), PCA for Hb sample (Figure S4), quality of variables representation for PCA of Hp sample (Figure S5), variables correlation for PCA of Hp sample (Figure S6), PCA for Hp sample (Figure S7), quality of variables representation for PCA of HbHp sample (Figure S8), variables correlation for PCA of HbHp (Figure S9), fragmentation spectra of example modified peptides (Figures S10 – S13).

Table S1. Overview of defined variable modifications.

modification name	**mass difference (Da)**	**amino acid**
carbamidomethylation	+57.02146	C
FPOP oxidation	+15.994915	F, H, I, K, L, M, P, R, V, W, Y
FPOP dioxidation	+31.989828	C, F, M, W, Y
FPOP carbonyl	+13.979265	E, H, I, L, P, Q, R, V
FPOP decarboxylation	-30.010565	E, D
FPOP Met-32Da	-32.008457	M
FPOP His+5Da	+4.9789	H
FPOP His-10Da	-10.031968	H
FPOP Arg deguanidination	-43.053432	R
FPOP trioxidation*	+47.984745	C, W
FPOP His->Asp*	-22.03197	H
FPOP His->Asn*	-23.01598	H

* only defined for search in PEAKS

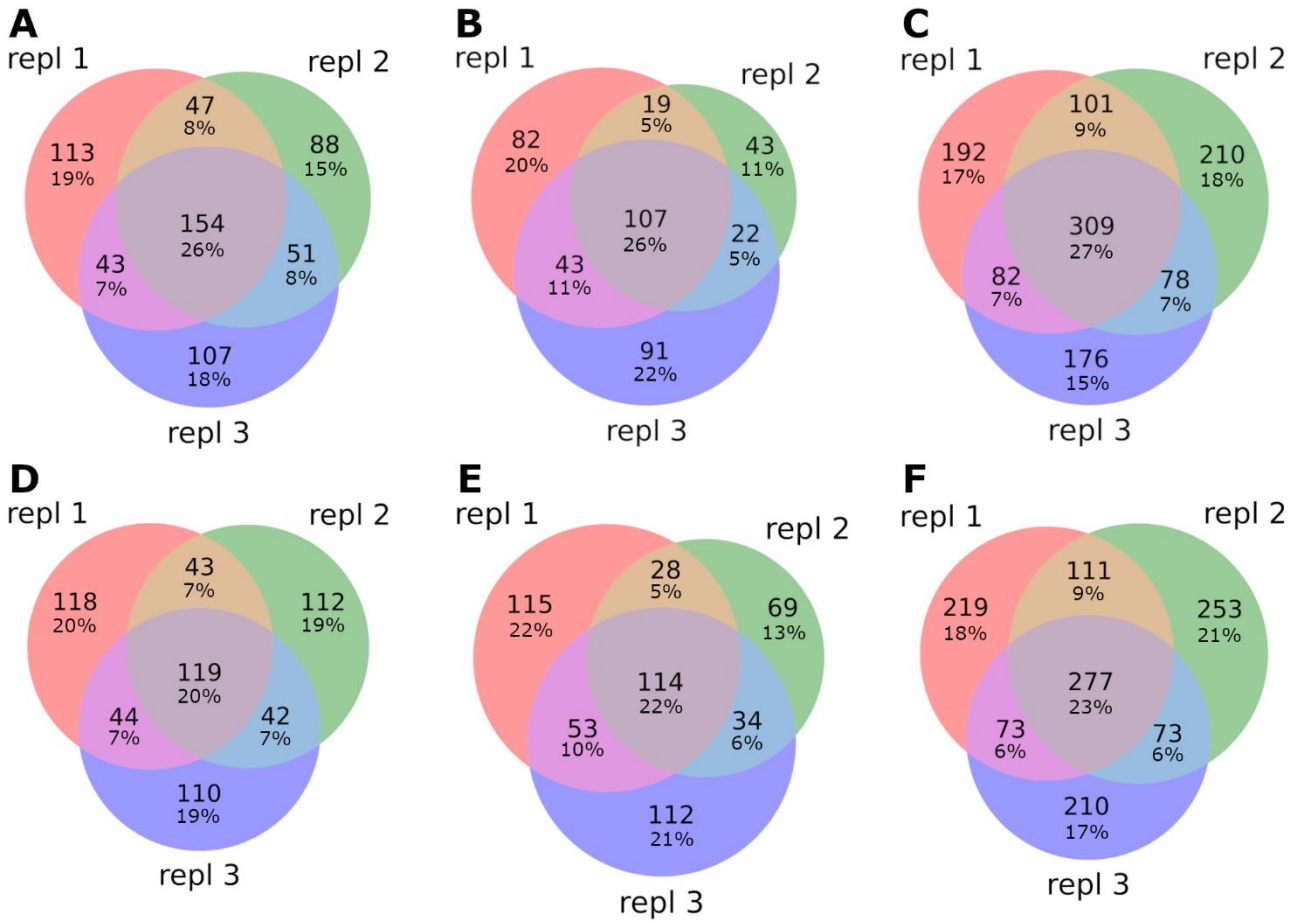


Figure S1. Intersections of IDs between replicates for Mascot searches of (A) Hb; (B) Hp; (C) HbHp complex samples; and for PEAKS searches of (D) Hb; (E) Hp; (F) HbHp complex samples. Mascot exhibits higher consistency within replicates.

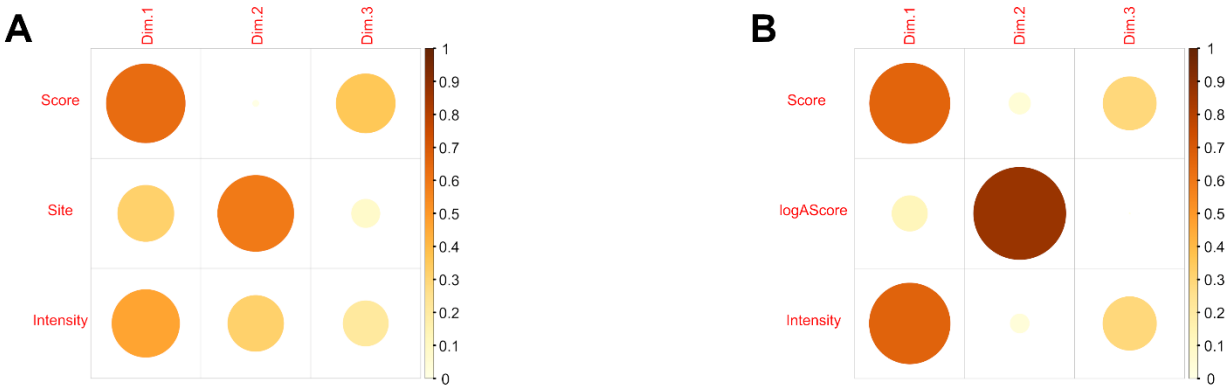


Figure S2. Quality of representation of the selected variables by the dimensions of the PCA for Hb sample searched by (A) Mascot, and (B) PEAKS. The extent of representation is visualized by both size and colour.

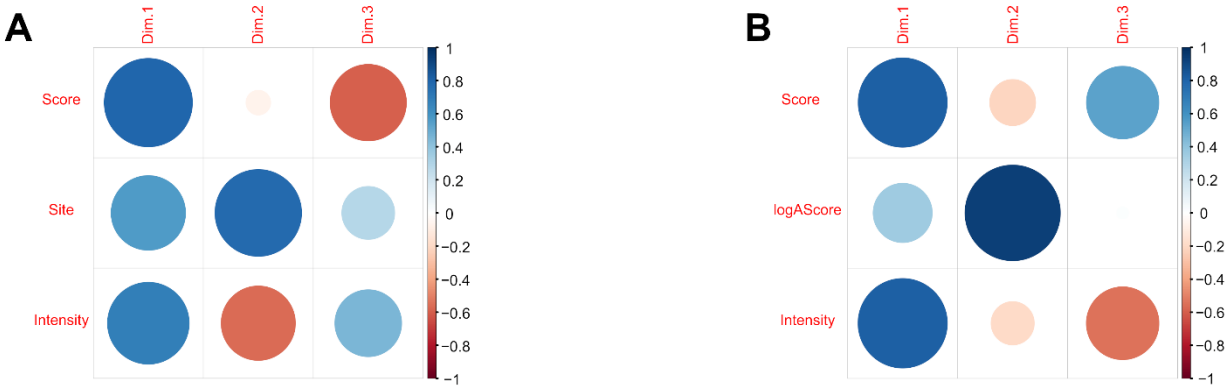


Figure S3. Correlation of the selected variables and the dimensions of the PCA for Hb sample searched by (A) Mascot, and (B) PEAKS. The extent of correlation is visualized by size and the trend of correlation as colour.

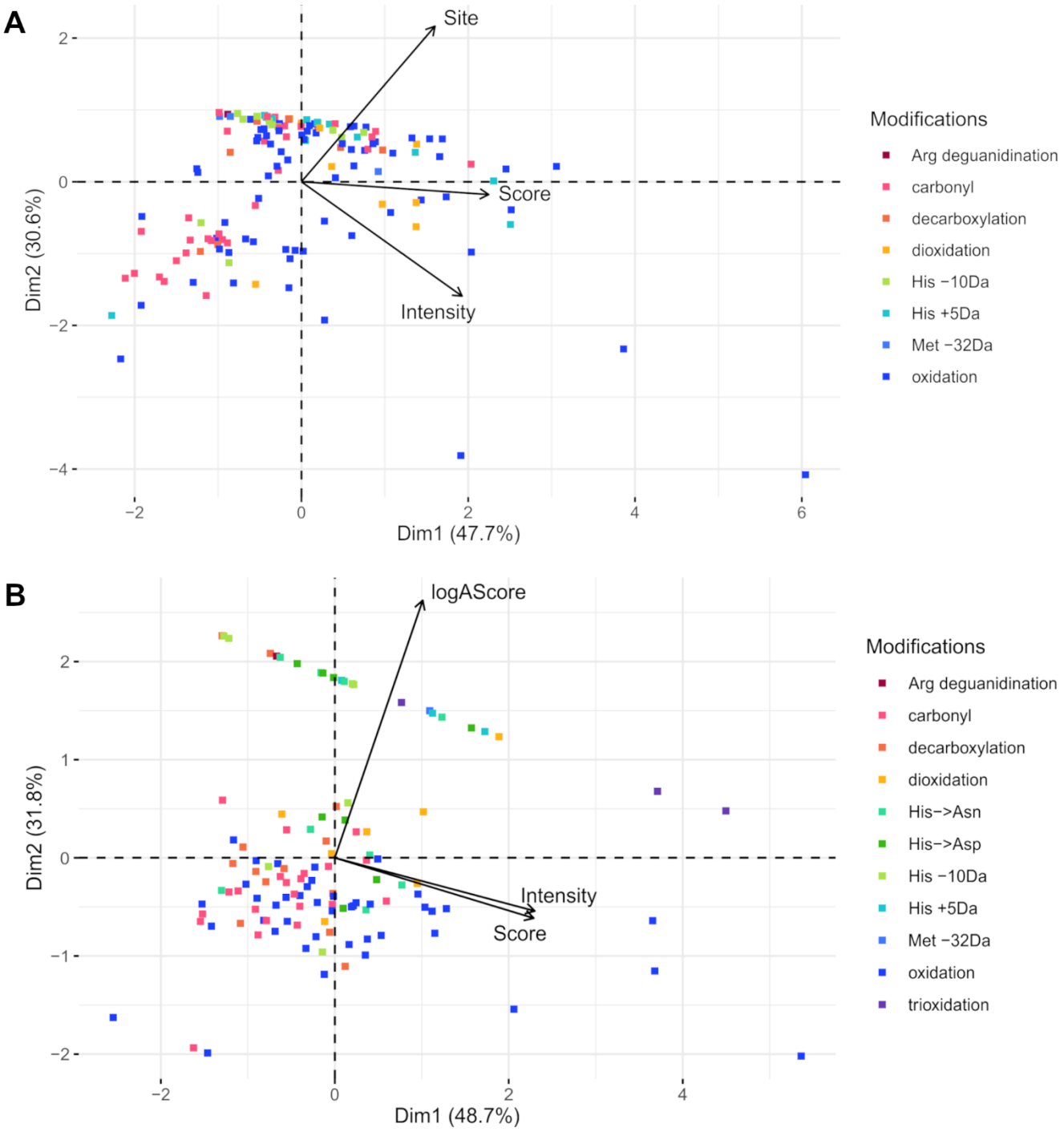


Figure S4. Principal component analysis of modifications identified in Hb sample by (A) Mascot and (B) PEAKS. Plot of color-coded modifications shows their distribution within dimensions defined by PCA. The vectors show correlation of variables with the PCA dimensions. The logAScore variable represents probability of site determination.

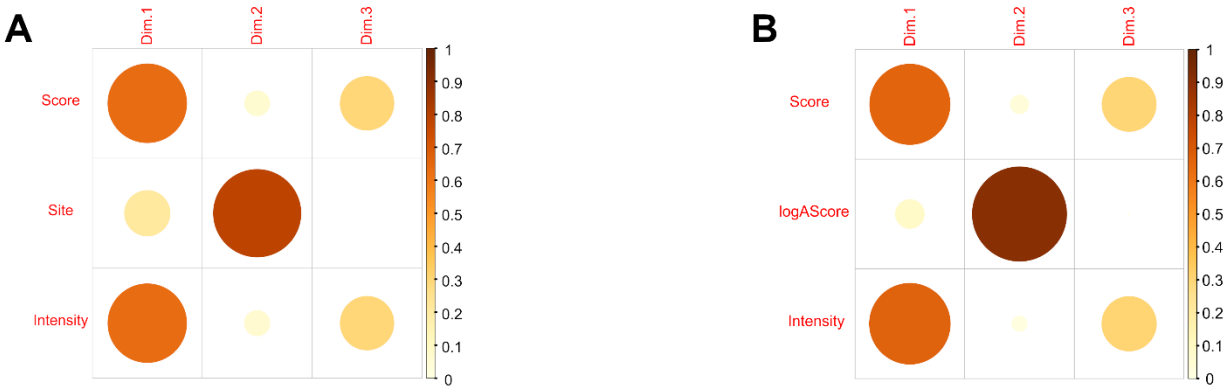


Figure S5. Quality of representation of the selected variables by the dimensions of the PCA for Hp sample searched by (A) Mascot, and (B) PEAKS. The extent of representation is visualized by both size and colour.

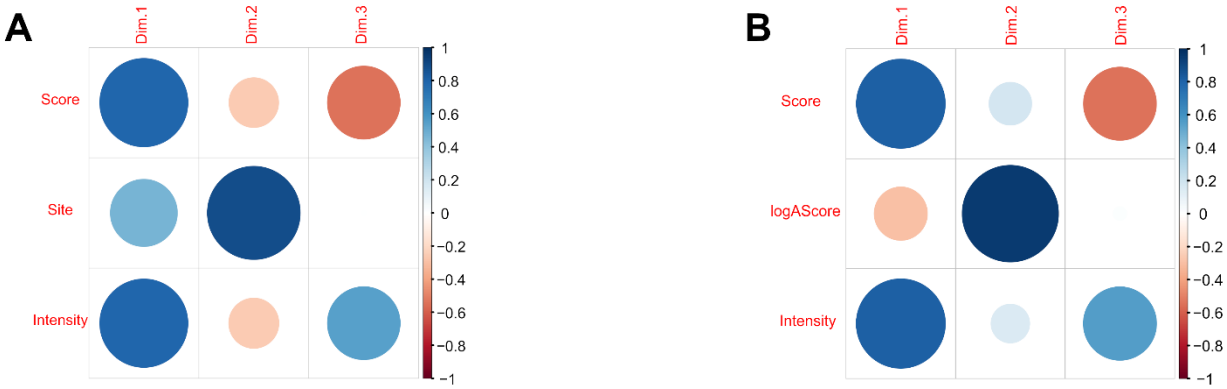


Figure S6. Correlation of the selected variables and the dimensions of the PCA for Hp sample searched by (A) Mascot, and (B) PEAKS. The extent of correlation is visualized by size and the trend of correlation as colour.

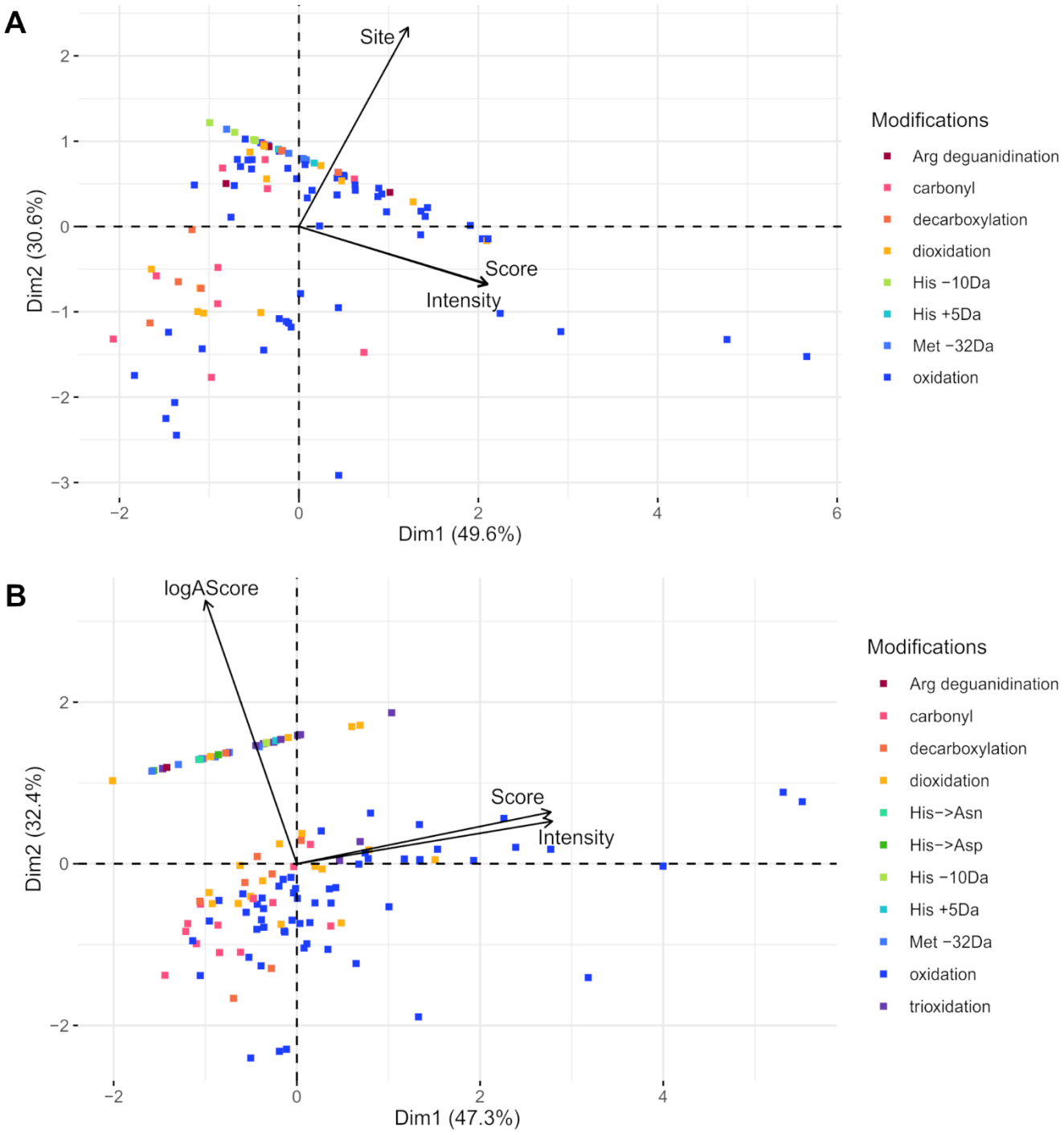


Figure S7. Principal component analysis of modifications identified in Hp sample by (A) Mascot and (B) PEAKS. Plot of color-coded modifications shows their distribution within dimensions defined by PCA. The vectors show correlation of variables with the PCA dimensions. In (A) the Score and Intensity vectors overlap. The logAScore variable represents probability of site determination.

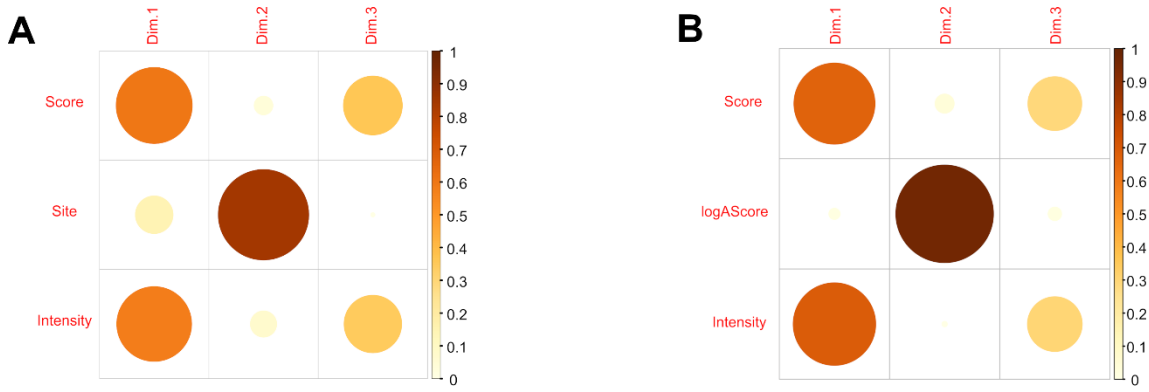


Figure S8. Quality of representation of the selected variables by the dimensions of the PCA for (A) Mascot search and (B) PEAKS search of the HbHp complex sample. The extent of representation is visualized by both size and colour.

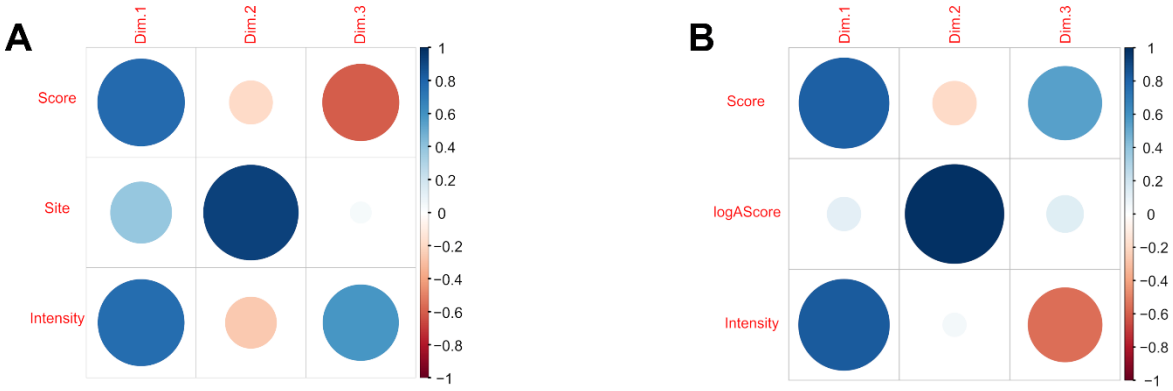


Figure S9. Correlation of the selected variables and the dimensions of the PCA for HbHp complex sample for (A) Mascot and (B) PEAKS. The extent of correlation is visualized by size and the trend of correlation as colour.

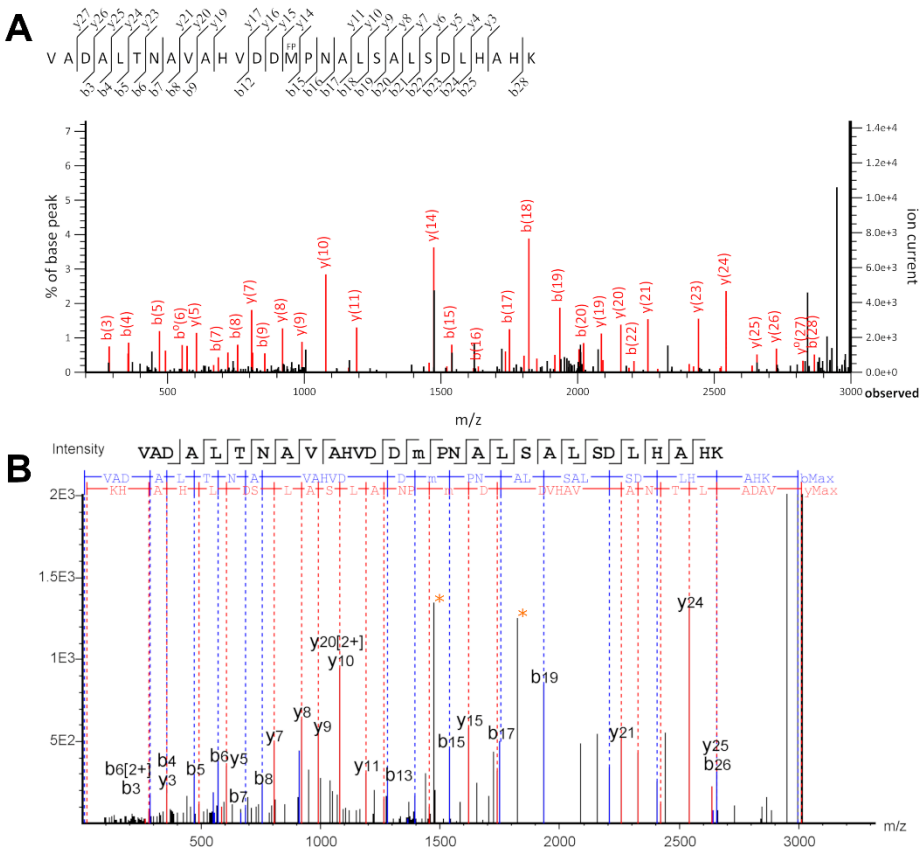


Figure S10. Fragmentation spectra of Met15 oxidation (+15.994915) in the peptide VADALTNAVAHVDDMPNALSALSDLHAHK in its triply charged form at m/z 1004.84 and RT 16.61 from the HbHp complex sample as identified by (A) Mascot (Site: 98.26%), and (B) PEAKS (AScore: 11.12). PEAKS did not account for relatively strong y14 and b18 ions (labelled *). Although the fragmentation coverage is almost complete for both search engines, Mascot shows high confidence, while PEAKS gives rather low AScore.

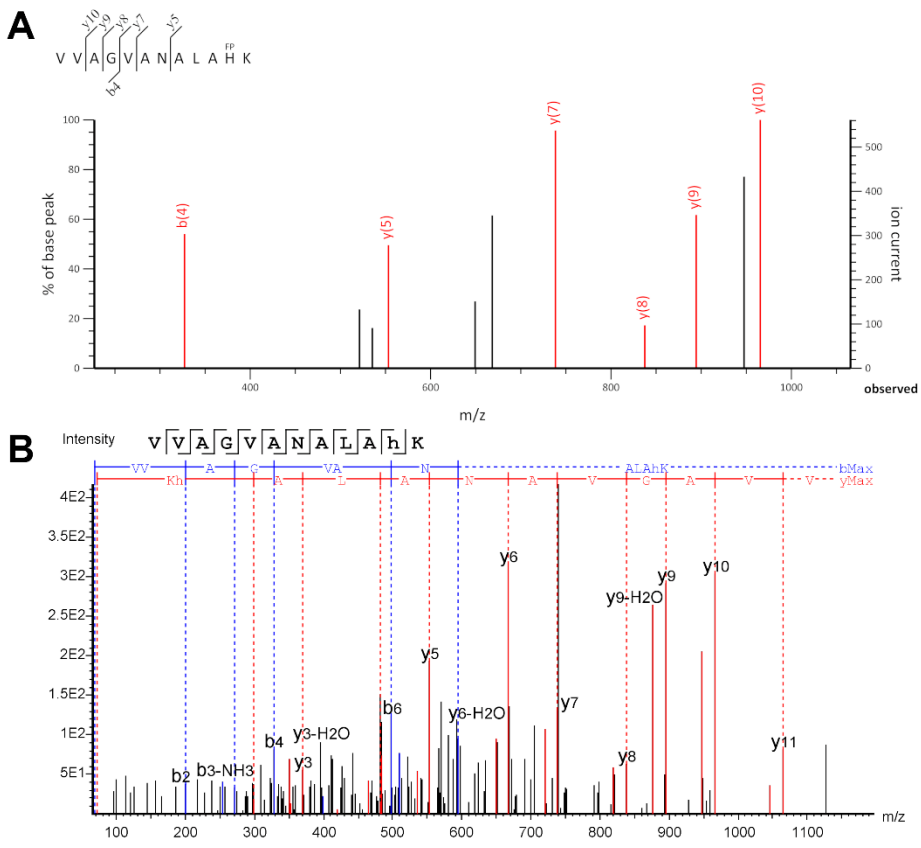


Figure S11. Fragmentation spectra of His11 carbonyl (+13.979265) from peptide VVAGVANALAHK in its doubly charged form at m/z 582.33 and RT 9.44 from the HbHp complex sample as identified by (A) Mascot (Site: 49.82%), and (B) PEAKS (AScore: 14.04). PEAKS shows better fragmentation coverage than Mascot, yet the certainty of site determination is low.

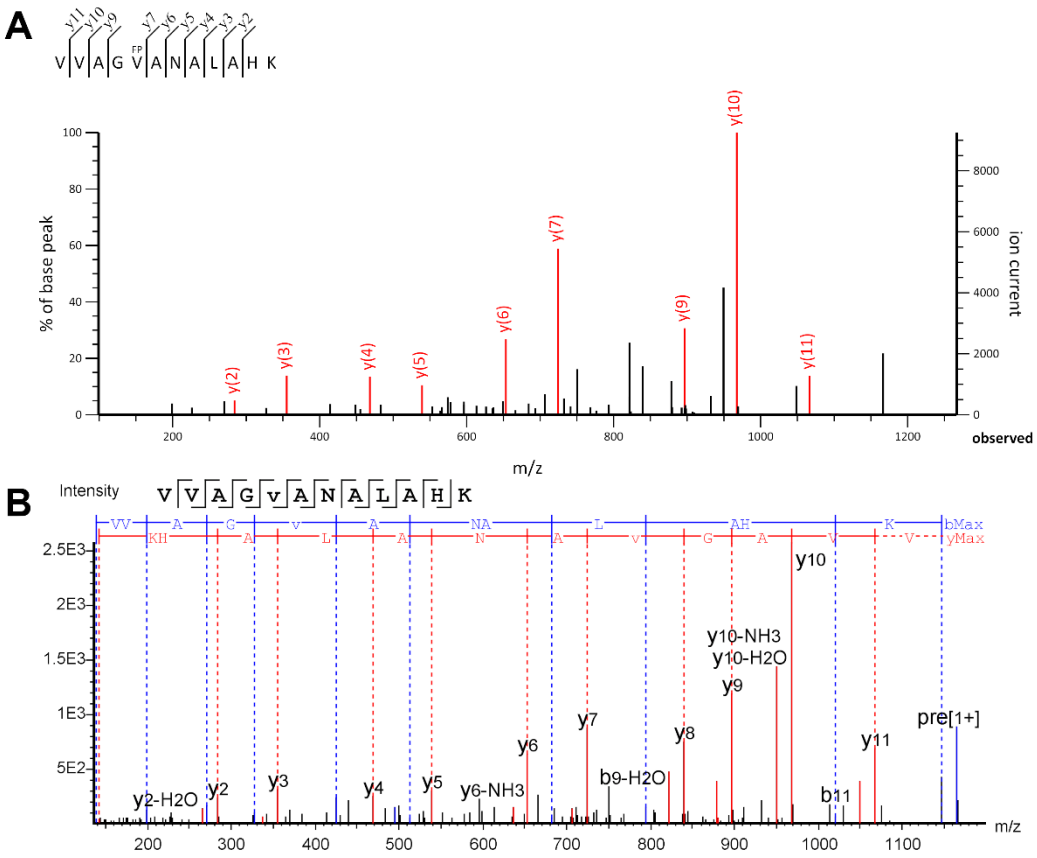


Figure S12. Fragmentation spectra of Val5 oxidation (+15.994915) in the peptide VVAGVANALAHK in its doubly charged form at m/z 583.34 at RT 5.9 from the HbHp complex sample as identified by (A) Mascot (Site: 99.99%), and (B) PEAKS (AScore: 26.57). For high intensity fragmentation spectra, both Mascot and PEAKS are able to determine the site of modification with high confidence.

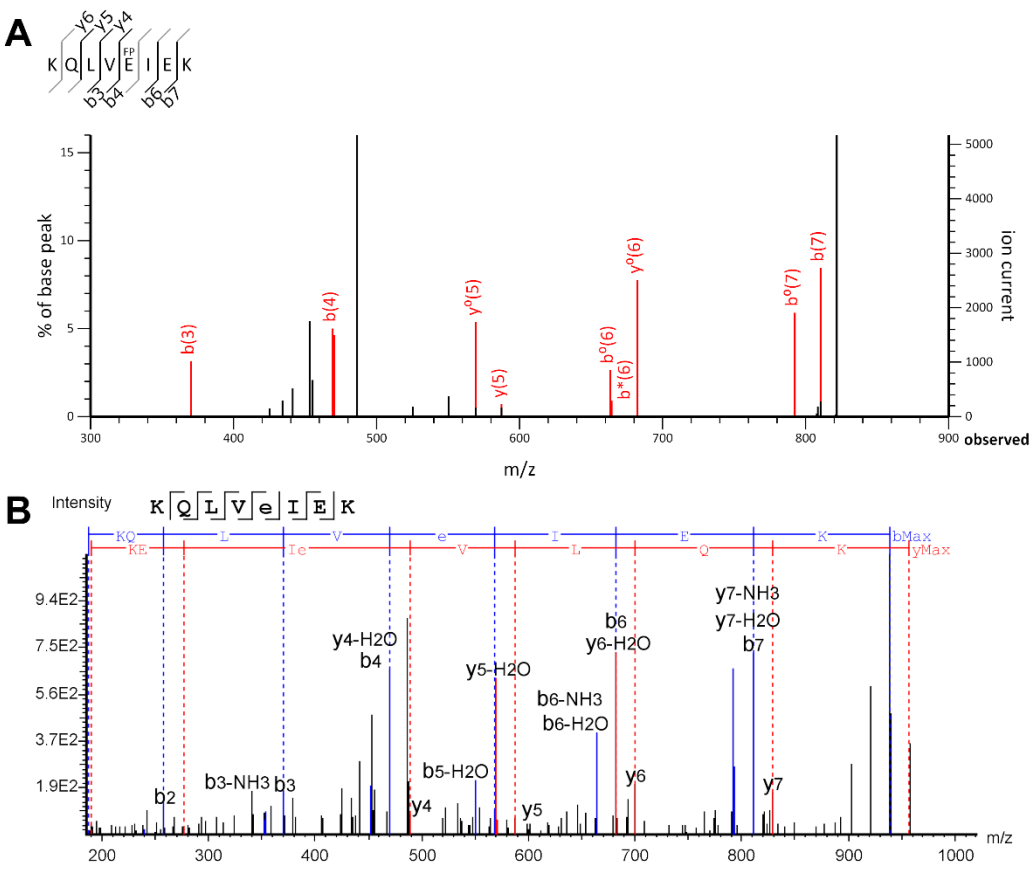


Figure S13. Fragmentation spectra of the Glu5 decarboxylation (-30.010565) in peptide with the sequence KQLVEIEK in its mono charged form at m/z 956.58 at RT 5.43 from the HbHp complex sample as identified by (A) Mascot (Site: 68.98%), and (B) PEAKS (AScore: 24.32). Even though only low intensity ions differentiate between the Glu5 and Glu7 positions, both search engines are assigning confident site determination.

Publication II

Zakopcanik, M.; Kavan, D.; Kukacka, Z.; Novak, P. & Loginov, D. S.

Data Independent Acquisition Represents a Promising Alternative for Fast Photochemical Oxidation of Proteins (FPOP) Samples Analysis.

Analytical Chemistry, 2024

My contribution: composing and optimisation of the processing workflow, data analysis, evaluation of acquisition approaches, writing and preparation of the manuscript

analytical chemistry

Open Access
This article is licensed under CC-BY 4.0

pubs.acs.org/ac
Article

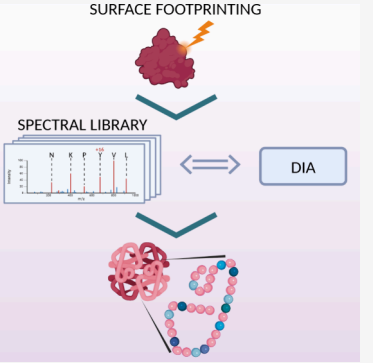
Data-Independent Acquisition Represents a Promising Alternative for Fast Photochemical Oxidation of Proteins (FPOP) Samples Analysis

Marek Zakopcanik, Daniel Kavan, Zdenek Kukacka, Petr Novak, and Dmitry S. Loginov*

Cite This: https://doi.org/10.1021/acs.analchem.4c01084
Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Fast Photochemical Oxidation of Proteins (FPOP) is a protein footprinting method utilizing hydroxyl radicals to provide valuable information on the solvent-accessible surface area. The extensive number of oxidative modifications that are created by FPOP is both advantageous, leading to great spatial resolution, and challenging, increasing the complexity of data processing. The precise localization of the modification together with the appropriate reproducibility is crucial to obtain relevant structural information. In this paper, we propose a novel approach combining validated spectral libraries together with utilizing DIA data. First, the DDA data searched by FragPipe are subsequently validated using Skyline software to form a spectral library. This library is then matched against the DIA data to filter out nonrepresentative IDs. In comparison with FPOP data processing using only a search engine followed by generally applied filtration steps, the manually validated spectral library offers higher confidence in identifications and increased spatial resolution. Furthermore, the reproducibility of quantification was compared for DIA, DDA, and MS-only acquisition modes on timsTOF SCP. Comparison of coefficients of variation (CV) showed that the DIA and MS acquisition modes exhibit significantly better reproducibility in quantification (CV medians 0.1233 and 0.1494, respectively) compared to the DDA mode (CV median 0.2104).



Downloaded via 147.231.236.14 on July 6, 2024 at 13:24:46 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

INTRODUCTION

Nowadays, a comprehensive structural study of proteins requires not only a snapshot of one state but also dynamic information which can provide some insight into various peculiarities of protein complex formation,[1−3] protein−DNA interactions,[4,5] protein folding,[6,7] etc. From this point of view, radical footprinting methods represent a valuable tool.

Protein surface footprinting provides information related to the surface accessible area of residue side chains, resulting in biologically relevant data concerning interaction or folding dynamics. Footprinting might be implemented by means of different radical probes.[8−10] Among them, the Fast Photochemical Oxidation of Proteins (FPOP) introduced in 2005[11] is the most popular. It is a covalent labeling-based method that utilizes hydroxyl radicals. Due to ·OH properties and their wide reactivity toward amino acids, this method has the potential to provide valuable information even on minor conformational changes of protein residue side chains.[12] On the other hand, broad reactivity results in an extreme number of variable modifications, which should be accounted for during data processing. The precise localization of the oxidative modification in the peptide sequence is a keystone of FPOP analysis. Thus, data processing is a bottleneck in FPOP analysis. One way to overcome this issue is to restrict the spatial resolution of the method by reducing the number of modifications allowed per search.[13]

The challenge of FPOP data processing was addressed in our recent study.[14] We showed that the results of the FPOP experiment are highly dependent on a selected peptide identification framework, with only 30% of the identified modifications mutual for the Mascot and PEAKS search engines. Applied statistical analysis in combination with manual validation of assigned modifications revealed that Mascot provided more reliable FPOP data over PEAKS. This approach allows us to reduce the time for manual validation of assigned FPOP modifications but is still quite laborious.

One of the possibilities to avoid the need for repeated ID validation for each experiment on the same set of proteins might be implementation of a reference spectral library. Linking the peptide parameters (sequence, m/z, charge, retention time, and possibly ion mobility) together with its characteristic fragmentation pattern is the underlying principle of spectral libraries.[15] Compared to the older approach of database search, spectral library search is limited to identifications already present in the library. On the other

Received: February 27, 2024
Revised: June 27, 2024
Accepted: June 28, 2024

ACS Publications
© XXXX The Authors. Published by American Chemical Society

hand, spectral library search offers greater sensitivity, is less time-consuming, provides a higher number of peptide-spectrum matches at the same error rate, and is capable of detecting less obvious matches.[16,17] The employment of proficient control mechanisms can result in a library consisting of highly confident identifications and precisely assigned PTMs, or oxidative modifications in the case of FPOP analysis. Additionally, the implementation of library search in the FPOP data processing workflow allows the utilization of data-independent acquisition (DIA).

Data collection in mass spectrometry-based proteomics can take place in either full-scan mode or tandem MS, further divided into data-dependent acquisition (DDA) and DIA. In DDA mode, a limited number of precursor ions (topN) are selected from survey scans based on their parameters (typically the intensity and charge state) and then subjected to fragmentation, generating fragmentation spectra. In DIA, the mass spectrometer sequentially isolates and fragments predefined precursor windows across the desired mass range. This approach allows for a comprehensive coverage of precursors across various abundance levels in the defined mass range. DIA acquisition provides data with significantly less missing values, resulting in better reproducibility of the analysis over DDA mode.

Due to the extensive number of modifications that need to be considered during the evaluation of FPOP data, the formation of spectral libraries from DIA data is not yet viable. To obtain the spectral library for the HbHp complex, we opted for the combination of FragPipe for the search of the DDA data and Skyline for the validation part. The FragPipe[18] is an open-source pipeline that offers fast and precise identification of proteomic data. Skyline[19] is an open-source application for targeted proteomics. The software offers to generate a spectral library from FragPipe search results and view annotated fragmentation spectra associated with individual library entries. Together with dot product scoring and the option to filter on the basis of the number of matched ions, this allows for a convenient way to validate each identification. For FPOP to bear relevant structural information, the certainty of the modification position is crucial. Thus, only spectra with sufficient fragmentation coverage should be included in the spectral library.

METHODS

Chemicals and Materials. Human haptoglobin phenotype 1−1 (Hp), human hemoglobin (Hb), and catalase were purchased from Sigma-Aldrich. PNgase F was from New England BioLabs Inc. Mass spectrometry-grade trypsin was from Promega Corporation. All solvents were purchased from Merck (Germany) in LC-MS grade. Other chemicals reported in this article were purchased from Sigma-Aldrich.

FPOP of the HbHp Complex. Protein solutions at 0.3 mg/mL concentrations were prepared in 150 mM ammonium acetate (pH 7.0), Hp and Hb were mixed in an equimolar ratio to form a complex. FPOP analysis was carried out as described previously[2] with minor modifications. Prior to FPOP analysis, histidine was added to the samples at 1 mM concentration. The protein complex was mixed with 20 mM H_2O_2 using a continuous capillary flow system (1:1 v/v) and irradiated with a laser shot (15 Hz, 20 ns pulse duration, 2.24 mJ/mm^2 radiant exposure) by a 248 nm KrF laser (Coherent, COMPex50). Quenching of hydroxyl radicals was done with 75 mM L-methionine immediately after irradiation. The samples were collected in a plastic tube with catalase to remove H_2O_2 excess. The FPOP experiment was conducted in pentaplicate.

FFAP of the HbHp Complex. The FFAP experiment was performed as described previously.[8] Briefly, lyophilized Hp and Hb were dissolved in 50 mM ammonium carbonate buffer pH 7.4 and mixed together in 1:1 ratio. The equimolar mixture of Hp and Hb (0.3 mg/mL) was labeled by acetic Togni reagent dissolved in dimethylsulfoxide (50 mg/mL) upon activation by L-Ascorbic acid (10 mg/mL). The final concentration of acetic Togni reagent in the reaction was 7.5 mM. The labeling reaction was stopped after 3 s by adding tryptophan (10 mg/mL). Five μg of the labeled complex was desalted using a C4 microtrap column (Optimize Technologies, Oregon City, OR) and dried using a SpeedVac concentrator. Experiment was done in biological triplicate.

Protein Digestion and LC-MS and MS/MS Data Acquisition. The collected FPOP samples were diluted (3:1 ratio) with 150 mM 4-ethylmorpholine buffer (pH 8.5), supplemented with 15% acetonitrile (ACN), and FFAP samples were dissolved in 50 mM 4-ethylmorpholine buffer (pH 8.5), supplemented with 5% ACN. Reduction and alkylation were achieved by adding tris(2-carboxyethyl)-phosphine and 2-chloroacetamide at 10 mM and 30 mM concentrations, respectively. The samples were incubated at 70 °C for 5 min and cooled down on a bench. Deglycosylation was carried out by adding PNGase F in 1:20 ratio followed by overnight incubation at 37 °C. Subsequently, trypsin was added to the samples in a 1:20 ratio and the samples were digested for 8 h at 37 °C. The digestion was terminated by adding trifluoroacetic acid (TFA) to a final concentration of 0.2%. Samples were desalted using a C18 microtrap column (Optimize Technologies, Oregon City, OR) and dried using a SpeedVac concentrator.

The samples were resuspended in 20 μL of 2% ACN, 0.1% TFA. In addition, a pooled sample was prepared by mixing an equal volume of all replicates in triplicate. All samples were analyzed using the Vanquish Neo UHPLC liquid chromatography system (Thermo Scientific) coupled with a timsTOF SCP mass spectrometer equipped with Captive spray (Bruker Daltonics). LC conditions were the same for all acquisition modes (DDA, MS1, and DIA). The peptides were trapped in the C18 trap column (PepMap Neo C18 5 μm, 300 μm × 5 mm, Thermo Scientific). After 3 min of trapping, the peptides were eluted from the trap column and separated on a C18 column (DNV PepMap Neo 75 μm × 150 mm, 2 μm, Thermo Scientific) by a 55 min long linear 0.1% formic acid (FA)−ACN gradient from 5% (v/v) to 35% (v/v) ACN at a flow rate of 350 nL/min. Both columns were heated to 50 °C. The DDA acquisition was done using parameters from the standard proteomic PASEF method. Spectra were recorded within the range of 100−1700 m/z, while the ion mobility was scanned from 0.7 to 1.3 Vs/cm^2. The method consisted of a TIMS survey scan lasting 166 ms, followed by 5 PASEF MS/MS scans. The total cycle time for this process was 1.04 s. The target intensity was set at 14000 with an intensity threshold of 500. The precursors for data-dependent acquisition were fragmented with an ion mobility-dependent collision energy that increased linearly from 20 to 59 eV. Active exclusion was enabled for 0.4 min with the mass width 0.015 m/z and 1/K0 width 0.015 V × s/cm^2 to prevent repeated selection of the same precursor ions. The same method with switched off PASEF mode was used to collect only MS1 scans.

https://doi.org/10.1021/acs.analchem.4c01084
Anal. Chem. XXXX, XXX, XXX−XXX

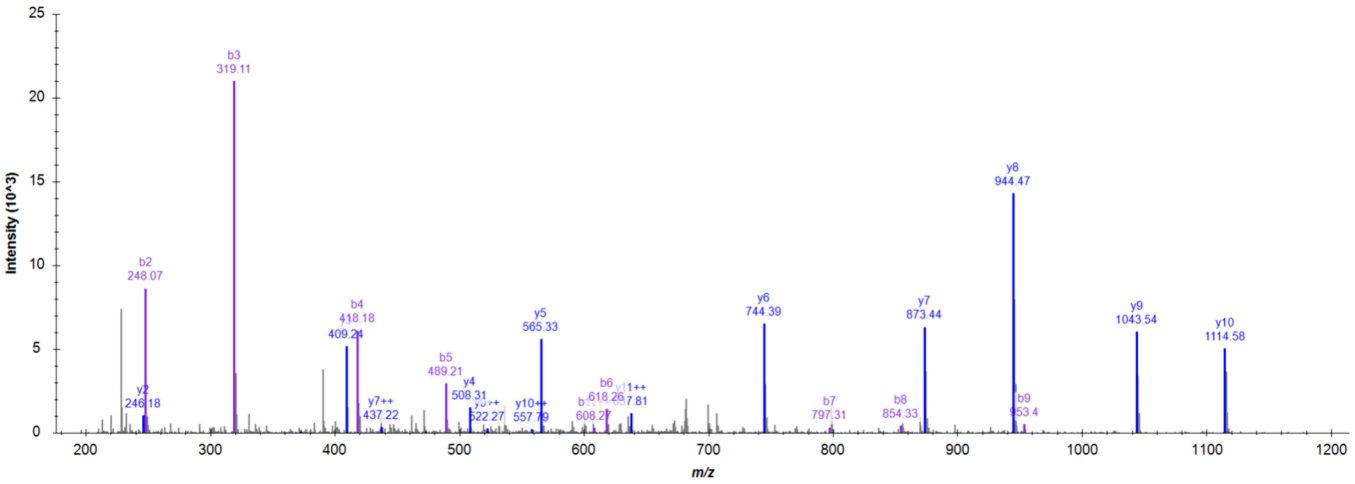


Figure 1. Fragmentation spectrum of Tyr7 oxidation in the peptide SCAVAEYGVYVK in its doubly charged form at m/z 681.32 and RT 16.70 min. The spectrum contains 19 out of 22 possible fragments, with y5 and y6 providing confidence for the modification site, representing optimal entry for the spectral library.

DIA acquisition was done using the standard dia-PASEF method. Spectra were recorded within the range of 100−1700 m/z, while the ion mobility was scanned from 0.64 to 1.45 Vs/cm^2. The number of dia-PASEF scans was 24, separated into three ion mobility windows per scan covering an m/z range from 400 to 1000 by 25 Th windows and an ion mobility range from 0.64 to 1.37 Vs/cm^2 with an estimated cycle time of 0.96 s. The accumulation and the ramp time was 100 ms, and collision energy settings were the same as in the DDA mode. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium[20] via the PRIDE[21] partner repository with the data set identifier PXD050132 and DOI 10.6019/PXD050132.

Data Analysis. Data obtained in DDA mode were searched using FragPipe 20.0 (Nesvizhskii Lab, University of Michigan, MI, USA against a database containing sequences of Hb (Uniprot IDs P69905 and P68871) and Hp (Uniprot ID P00738) supplemented with a deglycosylated form of β subunit, together with 246 sequences of common contaminants (including trypsin, PNGase F, and catalase). For each database entry, a reverse decoy sequence was appended. The precursor ion tolerance was set to 12 ppm, and the fragment mass tolerance was set to 0.05 Da. Only tryptic peptides (without cleavage before P) with up to two misscleavages were considered. Cysteine carboxymethylation was defined as fixed modification and FPOP modifications were denoted as variable modifications (Table S1). A maximum of 2 variable modifications per peptide was allowed. This limitation was chosen in order to operate within reasonable search space. On behalf of validation, the MSBooster rescoring was employed together with Percolator (version 3.06) and ProteinProphet (part of Philosopher 5.0).

Data collected from pooled samples was processed using Mascot as described in Zakopcanik et al.[14] Briefly, raw data were converted to MGF files in Compass DataAnalysis v.5.3 (Bruker Daltonics, Billerice, MA, USA). The files obtained were then searched using Mascot 2.7.0 (Matrix Science Inc., Boston, MA, USA). The search parameters were set the same as for FragPipe search (12 ppm precursor tolerance, 0.05 Da fragment tolerance, same cleavage rules and same database). Because the number of variable modifications being limited in Mascot, trioxidation was not included. The significance threshold was set at $p \leq 0.05$. Only identifications present in all replicates were accepted.

The results of the FragPipe search were processed in Skyline 23.1.0.268 (MacCoss Lab, University of Washington, Seattle, WA, USA). First, the five *interact.pep.xml* files were loaded together with five uncalibrated *mgf* files (generated by FragPipe). The cutoff score calculated in the *filter.log* file was not used, since the FPOP results represent a heterogeneous data set and would require separate thresholds for each group of identifications.[22] The precursor charges were set to 2−4 with a mass accuracy of 12 ppm. Fragment ions were restricted to *b* and *y*, with charges to 1 and 2 and library ion match tolerance of 0.05 m/z. The ion mobility resolving power was kept at the default value of 30. Trypsin [KR|P] with a maximum of 2 misscleavages was set for the digestion of the Hb and Hp sequences and up to 2 FPOP modifications, defined the same way as for FragPipe search. Each entry of the resulting library was manually inspected. Several aspects were considered during the validation. First, the overall sequence fragmentation coverage, while ensuring that the annotated fragments represent the majority of the tandem spectrum. Second, the Skyline-calculated dot product score, comparing the theoretical profile of the library entry and the actual isotopic distribution of the precursor in each of the replicates. All IDs that scored less than 0.9 of the isotopic dot product were closely inspected, which helped to detect misassignments, mainly carbonyl IDs that were wrongly assigned to peptides modified with single oxidation. Finally, special attention was paid to the presence of fragment ions that determine the exact position of the modification. In case that the modification could not be precisely localized, it was included as an "unlocalized mass shift". Identifications with low precursor intensity (below 2500 in Skyline notation) were excluded from the library. As a next step of validation, the spectral library was used as a reference to match against the DIA data in Skyline. A dot product score was used as a measure of spectral similarity. All spectra with a dot product score less than 0.7 were selected for closer examination.

Next, DIA and MS-only acquired data were mapped onto a respective library match using m/z, RT, and dot product score as criteria in Skyline software. Confident identifications were quantified in each of the acquisition modes at MS1 level using

https://doi.org/10.1021/acs.analchem.4c01084
Anal. Chem. XXXX, XXX, XXX−XXX

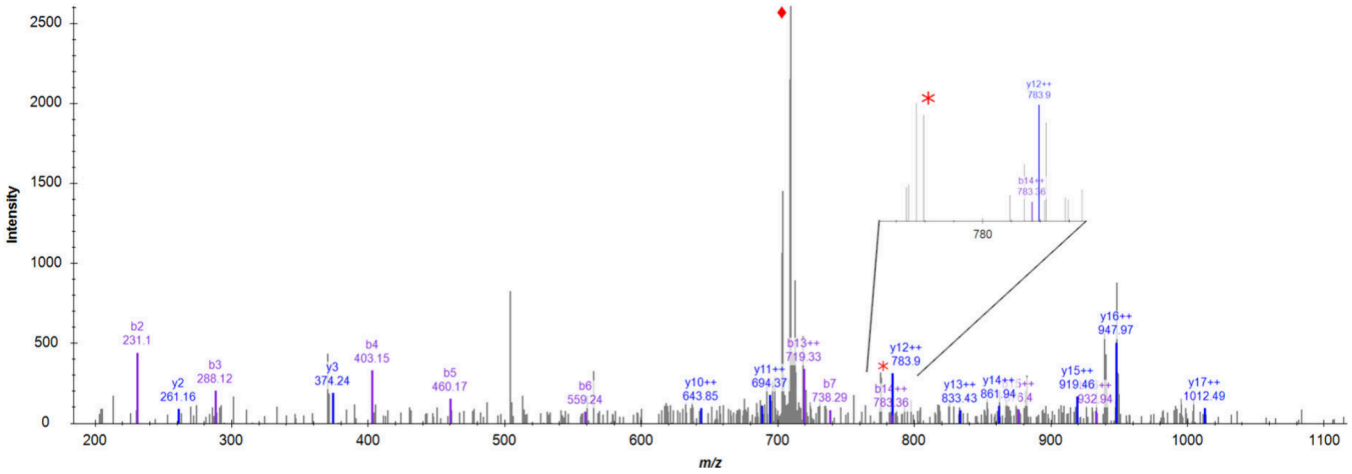


Figure 2. Fragmentation spectrum of Tyr7 oxidation in a peptide TEGDGVYTLNNEKQWINK in its triply charged form at m/z 709.01 and RT 12.22 min. Most of the assigned fragment ions have low intensity. Additionally, the doubly charged y12 fragment, which should support the localization of the modification, has an unmodified counterpart with the same intensity (labeled *). Analysis of raw data revealed at least two coeluted peptide species separated by ion mobility. The ♦ symbol denotes the precursor. Because of the uncertainty of the modification site, together with the low intensity of the assigned fragment ions, this ID was excluded from the spectral library.

Compass DataAnalysis 5.3 (Bruker Daltonics, Billerice, MA, USA). Based on our previous experience with FPOP data, identifications were quantified in charge states 1−5, which is necessary to encompass complete information about the extent of modification. The extent of modification (EoM) itself was calculated for each identification individually as the intensity of the modified peptide divided by the sum of intensities of the unmodified peptide and all of its modified variants. The proportion of unmodified peptide was calculated using the same formula.

$$EoM_i = \frac{I_i}{I_{unmod} + \sum I_{mod}}$$

RESULTS AND DISCUSSION

The optimization of the DIA workflow was performed on the FPOP oxidized HbHp protein complex, which represents a complex protein sample containing disulfide bonds and carbohydrate part. The HbHp complex was measured in pentaplicates on timsTOF SCP in two MS/MS acquisition modes—DDA, DIA—and once in MS-only mode. The DDA analyses were searched using FragPipe, and the results were loaded into Skyline to form a spectral library. The parameters considered during manual validation were the isotopic profile of the precursor peptide, the dot product of the fragment ion spectra, the number of assigned fragment ions, and the fragmentation coverage of the site of modification. For example, the fragmentation spectrum of the peptide SCAVAE-YGVYVK at m/z 681.32 (doubly charged) and RT 16.70 min bearing oxidation on Tyr7 (Figure 1) shows almost complete fragmentation coverage (19 out of 22 possible fragments). Furthermore, the site of modification is unambiguously determined by fragments y5 and y6. Together with the average dot product score of 0.96, this spectrum represents the ideal library entry. On the other hand, the fragmentation spectrum of the peptide TEGDGVYTLNNEKQWINK at m/z 709.01 (triply charged) and RT 12.22 min bearing oxidation on Tyr7 (Figure 2) consists of 21 fragments (of 36 possible), many of them at a borderline noise level. The modification site was assigned based on the $y12^{+2}$ fragment, but the spectrum contains a contradictory signal of the unmodified $y12^{+2}$ ion, indicating the possible presence of other coeluted peptides. This was further supported by examining the ion mobility information where the presence of two forms was revealed (Figure S1). Since the modification site cannot be determined confidently, the intensity of the assigned fragments is low, and the dot product score is low (0.62), this identification was rejected from the spectral library. Of 1780 identified peptides by the FragPipe search, a total of 522 were accepted, of which 447 bear the FPOP modification. This validation step produced a spectral library that included the retention time of the identifications. Although manual validation is time-consuming, the created spectral library can now be used for all future analyses of the same set of proteins (when employing the same chromatography conditions).

Next, the created spectral library was matched against the DIA data to calculate the dot product score (Figure 3). Most IDs exhibit a dot product score greater than 0.8, indicating high similarity between the spectra. Identifications with a dot

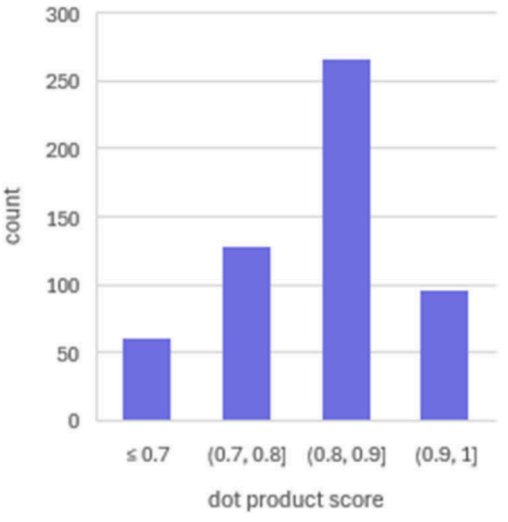


Figure 3. Distribution of the dot product score for DIA data matched against the manually validated spectral library.

https://doi.org/10.1021/acs.analchem.4c01084
Anal. Chem. XXXX, XXX, XXX−XXX

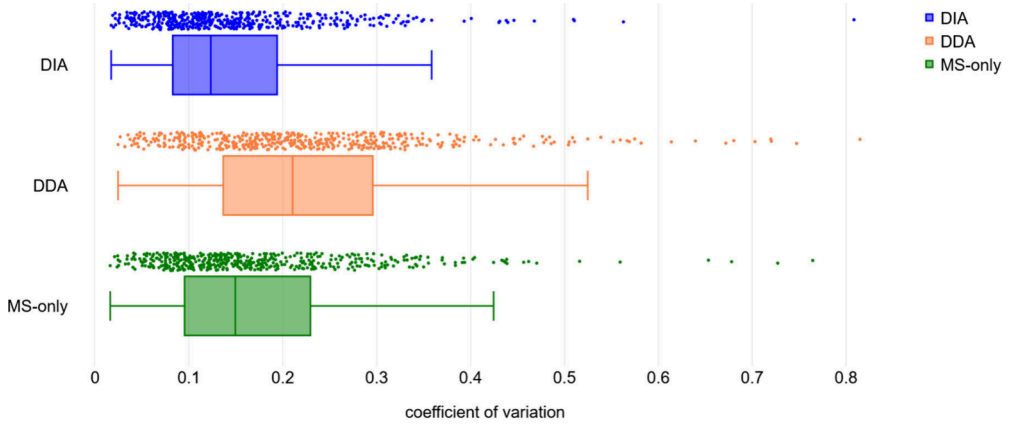


Figure 4. Variability of quantification. Coefficients of variation were calculated from the quantification of identified modifications (five replicates) for each acquisition mode. The line represents the median. Created with box and whisker plot maker.[25]

product score lower than 0.7 were selected for closer examination. The low dot product score could be an indication that the modification does not occur in all replicates or is of very low abundance. Therefore, even if the identification comes from a good quality fragmentation spectrum, the modified species is not representative of the state of the protein. Using raw data for validation of these low-scoring modified peptides, an additional 28 IDs were excluded from the data set. The step of matching the identifications against DIA data represents an important additional layer of validation.

The identifications obtained by this novel approach were compared with a routine workflow employing Mascot and automated filtering steps, as described previously.[14] Due to the nature of the previously published approach, only IDs with a single FPOP modification and a precise localization of the modification site were used for this comparison. The set of manually validated modifications consists mainly of oxidations (60%), with dioxidations being the second most common modification (11%), followed by carbonylation, decarboxylation, and His +5 Da (each 7%). The complete overview of the modifications is available in Table S2. The ratio of the types of modifications is similar to the results obtained by employing the Mascot search engine, but the number of modified IDs from Mascot search (256) is lower than in the manually validated library (333). To put this comparison in more context, the following aspects of each approach should be considered. Apart from utilizing information from five replicates in the case of library approach, versus three pooled replicates in the case of the Mascot search approach, the main difference lies in the validation mechanism. The approach using Mascot represents automated filtering in the form of accepting only IDs which were detected in all of the replicates. On the other hand, for the construction of the manually validated library, only one identification from the FragPipe database search was sufficient (in the case of no conflicting IDs from other replicates). Although manual validation of the data set requires more time, once it is finished, the library can be used repeatedly for every future experiment on this set of proteins. Furthermore, it is worth noting that the Mascot results contain IDs of isobaric peptides from different retention times as separate entries. When comparing the number of IDs without isobaric forms, the difference becomes even more apparent (library: 333 IDs, Mascot: 200 IDs). The set of additional IDs present in the manually validated library included modifications of less reactive aliphatic amino acids (Val, Leu, Pro) as well as more reactive ones (His, Met, Phe, Trp, and Tyr). Several of the IDs present exclusively in the manually validated data set were from the interaction surface of the Hb and Hp subunits of the complex. This demonstrates that the additional effort results in better coverage globally, as well as in the biologically relevant areas of the protein complex. Taken together, the approach using manual validation of the identifications leads to a wider population of unique modifications and increased spatial resolution, compared to routine filtering steps, requiring the ID to be present in all replicates as a validation mechanism. However, the lack of isobaric peptides could result in an incomplete assessment of the extent of modification in the case of aromatic amino acids.[23,24] To incorporate isobaric peptides into the proposed workflow, manual intervention is required during the validation step using Skyline. The user interface allows multiple entries of the same peptide (and the same modification) with different retention times. Such occurrences could be discovered using the extracted ion chromatogram, which Skyline reconstructs from the peak list data. DDA or DIA data would be needed for further confirmation of such IDs.

To validate the novel approach of using the DIA acquisition for FPOP, we compared the quantification of the accepted identifications between DIA, DDA, and MS-only modes. For each ID, the extent of modification was calculated. In order to evaluate the reproducibility of each acquisition method, a coefficient of variation (CV) was determined for each identification. Figure 4 shows the distribution of CVs for DIA, DDA, and MS-only using boxplots. More detailed parameters of the CV distribution are shown in Table 1. The DIA acquisition mode exhibits the lowest median (0.1233), as well as the lowest interquartile range, indicating the lowest variability. The characteristics of the MS-only data set are close

Table 1. Parameters of CVs for DIA, DDA, and MS-Only Acquisition Methods

	DIA	DDA	MS-only
Minimum:	0.0172	0.0245	0.0161
Q1:	0.0828	0.1366	0.0954
Median:	0.1233	0.2104	0.1494
Q3:	0.1939	0.2958	0.2294
Maximum:	0.8085	0.8148	0.7645
Mean ($\bar{x}$):	0.1501	0.2301	0.1748

https://doi.org/10.1021/acs.analchem.4c01084
Anal. Chem. XXXX, XXX, XXX−XXX

to those of the DIA data set, with a slightly higher median (0.1494) and a broader interquartile range. However, the DDA data set varies the most, with the median of 0.2104 and the widest interquartile range of the three acquisition modes. The extents of modification of selected peptides representing low, medium, and high intensity species are plotted in the Figure S2. These results suggest that the DIA mode, together with the MS-only mode, brings better reproducibility in the quantification of the FPOP data.

To assess whether there are significant differences in variability (relative to means) among the three data sets, a one-way ANOVA test was performed.[26] The ratio of variation between the sample means and the variation within the samples (F-statistic) is 58.067, which exceeds the threshold of H_0 acceptance threshold (3.0024). This, together with the p-value converging to 0, indicates that there is a significant difference in variability between at least two of the data sets. Posthoc analysis using the Tukey Honestly Significant Difference test showed that the differences between all of the test pairs are significant ($p < 0.01$). While the CV values for the DDA mode differ substantially, with the differences of the DIA-DDA and DDA-MS-only pairs with a p-value lower than 1.8×10^{-10}, the p-value of the DIA-MS-only pair is equal to 0.00338. The complete results for the ANOVA and the Tukey Honestly Significant Difference test are available in Tables S3 and S4.

Furthermore, low-intensity identifications were inspected separately, as a suspected source of variability. The distribution of the lowest intensity tertile for each acquisition mode (Figure 5) retains the similarity between the DIA and MS modes.

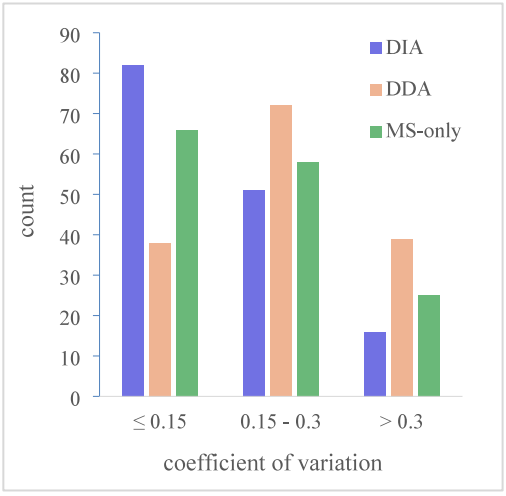


Figure 5. Coefficients of variation for the lowest-intensity tertile of identifications.

Compared to DDA, quantification from DIA or MS-only data offers higher reproducibility even for low-intensity identifications.

The proposed workflow is a universal approach that can be very effective regardless of the radical probes used for protein footprinting. Fast fluoroalkylation of proteins (FFAP) was selected as an alternative labeling technique to demonstrate the effectivity of the DIA workflow. FFAP utilizes fluoroalkyl radicals, generated by the reductive decomposition of Togni reagents with ascorbic acid, to label proteins within seconds at aromatic residues.[8] The manually validated FFAP spectral library of the HbHp complex contained 74 Togni modified identifications. The mapping of the DIA data hinted at 15 IDs with the dot product score ≤0.7 (Figure S3), and 11 of them were removed from the validated library due to either low precursor intensity or nonconsistent presence in the replicates. Overall, the FFAP data set of the HbHp complex contained 57 reliable modifications after removing isobaric and low intensive peptides (after quantification in DataAnalysis). The extent of modification and coefficients of variation were calculated in the same way as those for the FPOP analysis. Similarly to FPOP results, quantification of FFAP data acquired in DIA mode has the lowest variability compared to DDA and MS-only acquisition modes (Figure S4). DIA data exhibit the lowest median variability (0.1382) compared to DDA and MS-only, with medians 0.1689 and 0.1781, respectively (Table S5).

CONCLUSIONS

In this paper, we have demonstrated the potential of DIA for the protein footprinting data evaluation. The proposed workflow combines DDA data to create a validated spectral library, which is then matched against DIA data for further improvement of the IDs reliability.

Furthermore, the reproducibility of DIA, DDA, and MS-only acquisition modes for FPOP and FFAP analysis on timsTOF SCP was compared. The variability of quantification was assessed by coefficients of variation. The results reported here show that quantification from DIA data offers significantly better reproducibility compared to DDA and MS-only data.

The proposed workflow would undoubtedly benefit from automation, specifically in the step of validation of the identified spectra. Such automation could include evaluation of the fragmentation coverage of the modified peptide, and/or a carefully selected intensity threshold to cut off collision spectra originating from low abundant precursors, which often lead to spectra of insufficient quality.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.4c01084.

Overview of defined variable modifications (Table S1); Extracted ion mobilogram of the peptide TEGDGVYTLNNEKQWINK (Figure S1); Representation of modifications in the validated spectral library and the Mascot search results (Table S2); Extent of modifications of selected peptides (Figure S2); Results of one-way ANOVA test (Table S3); Results of Tukey Honestly Significant Difference test (Table S4); Dot product score for DIA data matched against the spectral library for the Togni labeled data set (Figure S3); Variability of quantification of FFAP experiment (Figure S4); CV distribution of quantification for FFAP experiment (Table S5). (PDF)

AUTHOR INFORMATION

Corresponding Author

Dmitry S. Loginov − *Institute of Microbiology, The Czech Academy of Sciences, 14220 Prague, Czech Republic;* orcid.org/0000-0002-4164-6145; Email: dmitry.loginov@biomed.cas.cz

https://doi.org/10.1021/acs.analchem.4c01084
Anal. Chem. XXXX, XXX, XXX−XXX

Authors

Marek Zakopcanik − *Institute of Microbiology, The Czech Academy of Sciences, 14220 Prague, Czech Republic; Department of Biochemistry, Faculty of Science, Charles University, 12820 Prague, Czech Republic;* orcid.org/0009-0003-9039-7151

Daniel Kavan − *Institute of Microbiology, The Czech Academy of Sciences, 14220 Prague, Czech Republic*

Zdenek Kukacka − *Institute of Microbiology, The Czech Academy of Sciences, 14220 Prague, Czech Republic;* orcid.org/0000-0001-7569-843X

Petr Novak − *Institute of Microbiology, The Czech Academy of Sciences, 14220 Prague, Czech Republic;* orcid.org/0000-0001-8688-529X

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.4c01084

Funding

This work was mainly financially supported by the Czech Science Foundation (22-27695S), the Technology Agency of the Czech Republic (ODEEP-EU TH86010001), the Ministry of Education, Youth and Sports of the Czech Republic grant PHOTOMACHINES - Photosynthetic cell redesign for high yields of therapeutic peptides (CZ.02.01.01/00/22_008/0004624) and the Academy of Sciences of the Czech Republic (RVO: 61388971).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge the Centre of molecular structure Core Facility at BIOCEV, a facility funded by European Regional Development Funds (CZ.1.05/1.1.00/02.0109 BIOCEV) and supported by the Czech Infrastructure for Integrative Structural Biology (Structural mass spectrometry CF - LM2018127 CIISB for CMS BIOCEV funded by MEYS CR).

■ ABBREVIATIONS

ACN, acetonitrile; CV, coefficient of variation; DDA, data-dependent acquisition; DIA, data-independent acquisition; FA, formic acid; FFAP, Fast Fluoroalkylation of Proteins; FPOP, Fast Photochemical Oxidation of Proteins; Hb, hemoglobin; Hp, haptoglobin; HbHp, hemoglobin−haptoglobin; PSM, peptide-spectrum match; RT, retention time (min); TFA, trifluoroacetic acid

■ REFERENCES

(1) Heming, J. D.; Huffman, J. B.; Jones, L. M.; Homa, F. L. *Journal of Virology* **2014**, *88* (1), 225−236.

(2) Loginov, D. S.; Fiala, J.; Brechlin, P.; Kruppa, G.; Novak, P. *Biochim. Biophys. Acta - Proteins Proteom.* **2022**, *1870* (2), 140735.

(3) Du, Y.; Duc, N. M.; Rasmussen, S. G. F.; Hilger, D.; Kubiak, X.; Wang, L.; Bohon, J.; Kim, H. R.; Wegrecki, M.; Asuru, A.; Jeong, K. M.; Lee, J.; Chance, M. R.; Lodowski, D. T.; Kobilka, B. K.; Chung, K. Y. *Cell* **2019**, *177* (5), 1232−1242.e11.

(4) Sperry, J. B.; Shi, X.; Rempel, D. L.; Nishimura, Y.; Akashi, S.; Gross, M. L. *Biochemistry* **2008**, *47* (6), 1797−1807.

(5) Polák, M.; Černý, J.; Novák, P. *Anal. Chem.* **2024**, *96* (4), 1478−1487.

(6) Chen, X.; Shao, Z.; Marinkovic, N. S.; Miller, L. M.; Zhou, P.; Chance, M. R. *Biophys. Chem.* **2001**, *89* (1), 25−34.

(7) Walters, B. T.; Mayne, L.; Hinshaw, J. R.; Sosnick, T. R.; Englander, S. W. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (47), 18898−18903.

(8) Fojtík, L.; Fiala, J.; Pompach, P.; Chmelík, J.; Matoušek, V.; Beier, P.; Kukačka, Z.; Novák, P. *J. Am. Chem. Soc.* **2021**, *143* (49), 20670−20679.

(9) Cheng, M.; Zhang, B.; Cui, W.; Gross, M. L. *Angew. Chem., Int. Ed. Engl.* **2017**, *56* (45), 14007−14010.

(10) Zhang, M. M.; Rempel, D. L.; Gross, M. L. *Free radical biology & medicine* **2019**, *131*, 126−132.

(11) Hambly, D. M.; Gross, M. L. *J. Am. Soc. Mass Spectrom.* **2005**, *16* (12), 2057−2063.

(12) Loginov, D. S.; Fiala, J.; Chmelik, J.; Brechlin, P.; Kruppa, G.; Novak, P. *ACS Omega* **2021**, *6* (15), 10352−10361.

(13) Liu, X. R.; Rempel, D. L.; Gross, M. L. *Nat. Protoc.* **2020**, *15* (12), 3942−3970.

(14) Zakopcanik, M.; Kavan, D.; Novak, P.; Loginov, D. S. *J. Proteome Res.* **2023**, *23* (2), 609−617.

(15) Griss, J. *PROTEOMICS* **2016**, *16* (5), 729−740.

(16) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. *PROTEOMICS* **2007**, *7* (5), 655−667.

(17) Shao, W.; Lam, H. *Mass Spectrom. Rev.* **2017**, *36* (5), 634−648.

(18) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. *Nat. Methods* **2017**, *14* (5), 513−520.

(19) MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J. *Bioinformatics* **2010**, *26* (7), 966−968.

(20) Deutsch, E. W.; Bandeira, N.; Perez-Riverol, Y.; Sharma, V.; Carver, J. J.; Mendoza, L.; Kundu, D. J.; Wang, S.; Bandla, C.; Kamatchinathan, S.; Hewapathirana, S.; Pullman, B. S.; Wertz, J.; Sun, Z.; Kawano, S.; Okuda, S.; Watanabe, Y.; MacLean, B.; MacCoss, M. J.; Zhu, Y.; Ishihama, Y.; Vizcaíno, J. A. *Nucleic Acids Res.* **2023**, *51* (D1), D1539−D1548.

(21) Perez-Riverol, Y.; Bai, J.; Bandla, C.; García-Seisdedos, D.; Hewapathirana, S.; Kamatchinathan, S.; Kundu, D. J.; Prakash, A.; Frericks-Zipper, A.; Eisenacher, M.; Walzer, M.; Wang, S.; Brazma, A.; Vizcaíno, J. A. *Nucleic Acids Res.* **2022**, *50* (D1), D543−D552.

(22) Rojas Ramírez, C.; Espino, J. A.; Jones, L. M.; Polasky, D. A.; Nesvizhskii, A. I. *Anal. Chem.* **2023**, *95* (44), 16131−16137.

(23) Polák, M.; Yassaghi, G.; Kavan, D.; Filandr, F.; Fiala, J.; Kukačka, Z.; Halada, P.; Loginov, D. S.; Novák, P. *Anal. Chem.* **2022**, *94* (7), 3203−3210.

(24) Cornwell, O.; Radford, S. E.; Ashcroft, A. E.; Ault, J. R. *J. Am. Soc. Mass Spectrom.* **2018**, *29*, 2413−2426.

(25) *Box and whisker plot maker*. Statistics Kingdom. https://www.statskingdom.com/advanced-boxplot-maker.html (accessed 2023-12-12).

(26) *ANOVA Calculator - One Way ANOVA and Tukey HSD test*. Statistics Kingdom. https://www.statskingdom.com/180Anova1way.html (accessed 2023-12-12).

SUPPLEMENTARY DATA

Data-independent acquisition represents a promising alternative for FPOP samples analysis

Marek Zakopcanik[1,2], Daniel Kavan[1], Zdenek Kukacka[1], Petr Novak[1], Dmitry S. Loginov[*,1]

[1]Institute of Microbiology, The Czech Academy of Sciences, 14220 Prague, Czech Republic

[2]Faculty of Science, Charles University, 12843 Prague, Czech Republic

*Email: dmitry.loginov@biomed.cas.cz

List of Supplementary data

Overview of defined variable modifications (Table S1); Extracted ion mobilogram of the peptide TEGDGVYTLNNEKQWINK (Figure S1); Representation of modifications in the validated spectral library and Mascot search results (Table S2); Extent of modifications of selected peptides (Figure S2); Results of one-way ANOVA test (Table S3); Results of Tukey Honestly Significant Difference test (Table S4); Dot product score for DIA data matched against the spectral library for the Togni labeled data set (Figure S3); Variability of quantification of FFAP experiment (Figure S4); CV distribution of quantification for FFAP experiment (Table S5).

Table S1. Overview of defined variable modifications.

modification name	mass difference (Da)	amino acid
FPOP oxidation	+15.994915	F, H, I, K, L, M, P, R, V, W, Y
FPOP dioxidation	+31.989828	F, M, W, Y
FPOP carbonyl	+13.979265	E, H, I, L, P, Q, R, V
FPOP decarboxylation	-30.010565	E, D
FPOP Met-32Da	-32.008457	M
FPOP His+5Da	+4.9789	H
FPOP His-10Da	-10.031968	H
FPOP Arg deguanidination	-43.053432	R
FPOP trioxidation*	+47.984745	W

* not defined for Mascot

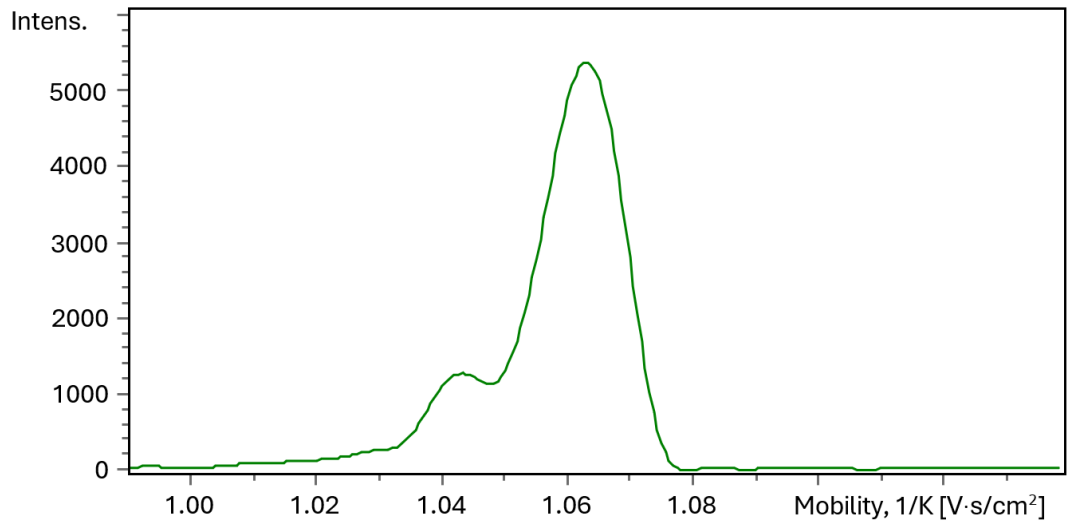


Figure S1. Extracted ion mobilogram of the peptide TEGDGVYTLNNEKQWINK at m/z 709.01 (triply charged) and RT 12.22 min bearing oxidation presumably on Tyr7. At least two coeluted subspecies separated by ion mobility explain the ambiguity of the localization of the oxidation from the MS/MS spectrum.

Table S2. Representation of modifications in the validated spectral library and in the Mascot search results.

modification	**validated library**		**Mascot***	
	count	percentage	count	percentage
Arg deguanidination	4	1%	2	1%
carbonyl	24	7%	20	10%
decarboxylation	23	7%	18	9%
dioxidation	36	11%	17	9%
His +5Da	23	7%	10	5%
His -10Da	16	5%	9	5%
Met -32Da	5	2%	3	2%
oxidation	200	60%	121	61%
trioxidation	2	1%	N/A	N/A

* only unique modified peptides

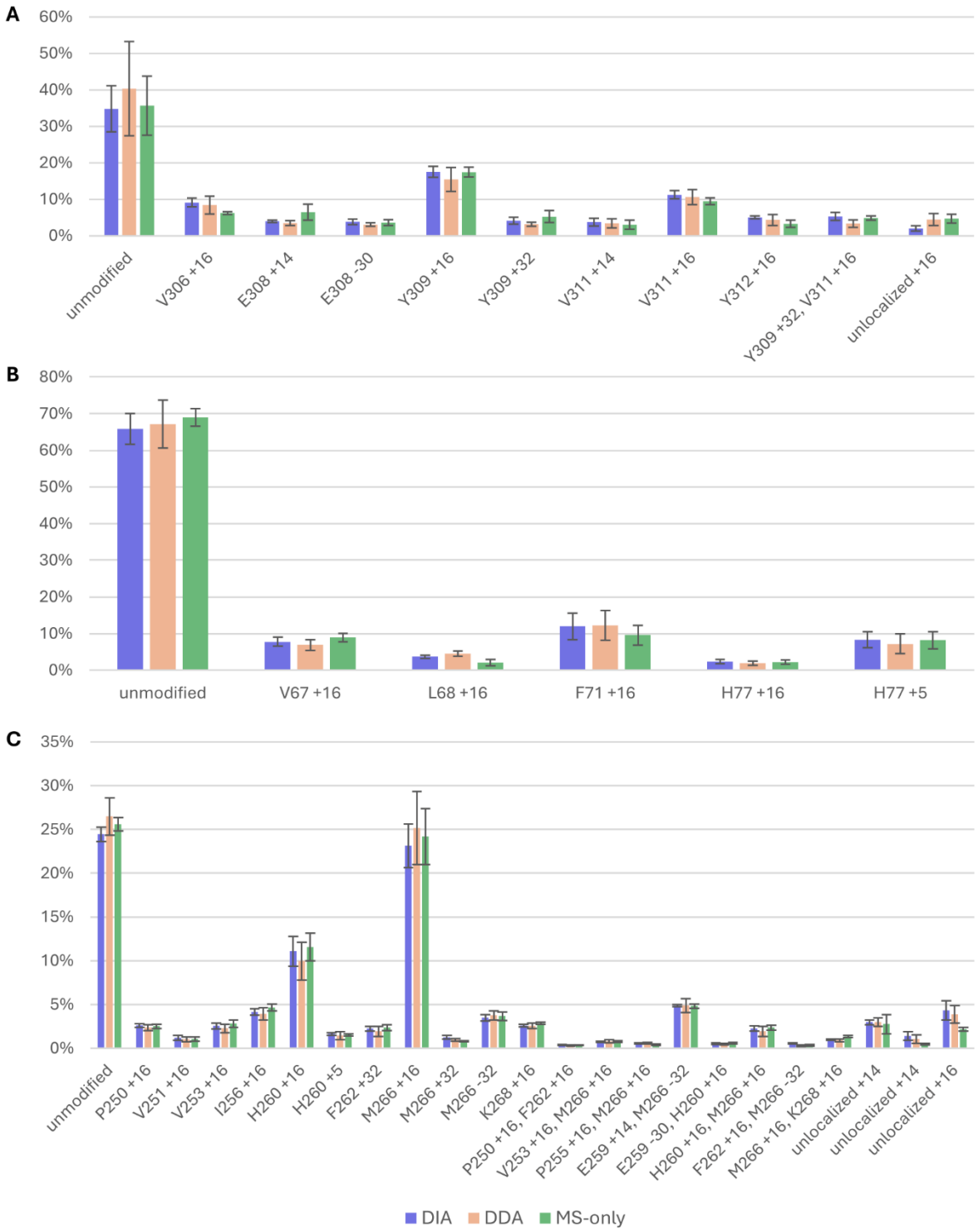


Figure S2. The extents of modification and the unmodified fraction of selected peptides. (A) Low intensity peptide SCAVAEYGVYVK; (B) medium intensity peptide KVLGAFSDGLAHLDNLK; (C) high intensity peptide SPVGVQPILNEHTFCAGMSK. The standard deviation shows that the DDA data exhibit greater variability over the DIA and MS data.

Table S3. Results of one-way ANOVA test for CVs of quantification from DIA, DDA and MS acquisition modes.

Source	DF	Sum of Square	Mean Square	F Statistic	P-value
Groups (between groups)	2	1.5001	0.75	58.0667	0
Error (within groups)	1338	17.2827	0.01292		
Total	1340	18.7827	0.01402		

Table S4. Results of Tukey Honestly Significant Difference test for CVs of quantifications from DIA, DDA and MS acquisition modes.

Pair	Difference	SE	Q	Lower CI	Upper CI	Critical Mean	p-value
DIA-DDA	0.08	0.005376	14.8824	0.06216	0.09784	0.01784	1.732e-10
DIA-MS	0.0247	0.005376	4.5975	0.00688	0.04255	0.01784	0.003377
DDA-MS	0.0553	0.005376	10.2849	0.03745	0.07312	0.01784	1.75e-10

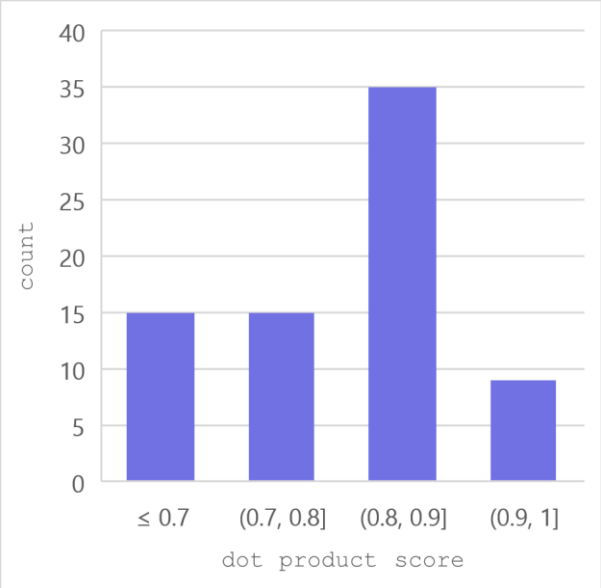


Figure S3. Distribution of the dot product score for DIA data matched against the manually validated spectral library for the Togni labeled data set.

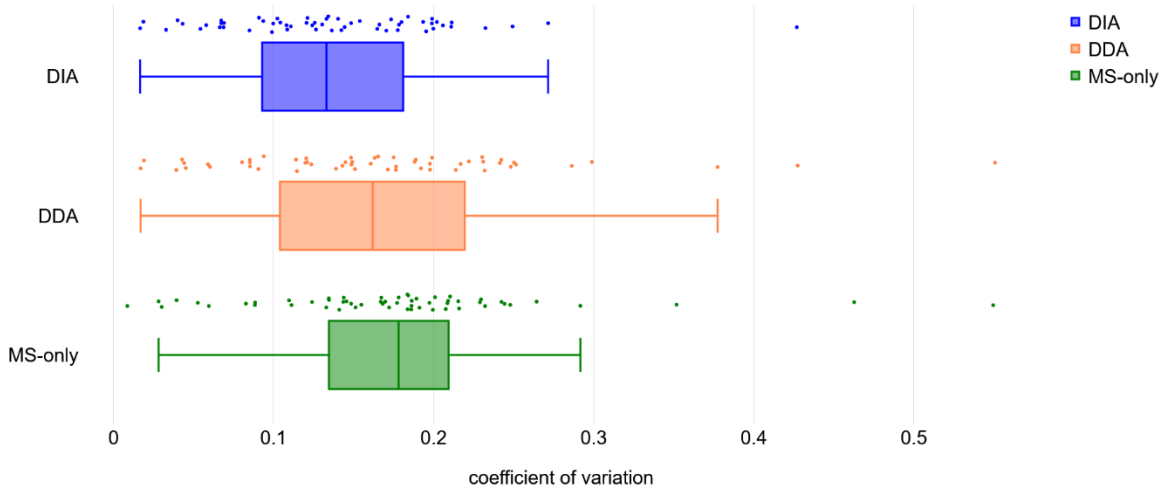


Figure S4. Variability of quantification of the Togni labeled data set. Coefficients of variation were calculated from the quantification of identifications (three replicates) for each acquisition mode.

Table S5. Parameters of CVs for DIA, DDA and MS-only acquisition modes for the Togni labeled data set.

	DIA	**DDA**	**MS-only**
Minimum:	0.01665	0.01692	0.008634
Q1:	0.09288	0.104	0.1346
Median:	0.133	0.1619	0.1781
Q3:	0.1809	0.2194	0.2093
Maximum:	0.4269	0.5507	0.5495
Mean ($\bar{x}$):	0.1382	0.1689	0.1781

Publication III

Škerlová, J.; Brynda, J.; Šobotník, J.; **Zákopčaník, M.**; Novák, P.; Bourguignon, T.; Sillam-Dussès, D. & Řezáčová, P.

Crystal Structure of Blue Laccase BP76, a Unique Termite Suicidal Defense Weapon.

Structure, 2024

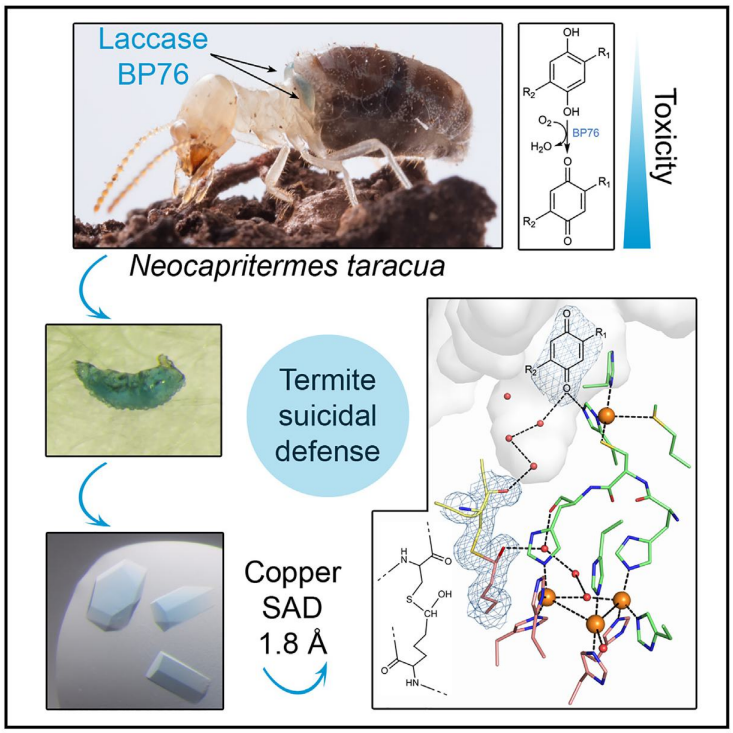
My contribution: optimisation and conducting of the open search of the MS data, spectra interpretation, writing and preparation of the respective manuscript sections

Short Article

Structure

Crystal structure of blue laccase BP76, a unique termite suicidal defense weapon

Graphical abstract



Authors

Jana Škerlová, Jiří Brynda, Jan Šobotník, ..., Thomas Bourguignon, David Sillam-Dussès, Pavlína Řezáčová

Correspondence

rezacova@uochb.cas.cz

In brief

Aging workers of the termite *Neocapritermes taracua* defend their colony by suicidal body rupture, mixing deposits of blue laccase BP76 with hydroquinones to produce a toxic sticky liquid. Škerlová et al. determined the structure of native BP76 that reveals stabilization strategies including a unique Lys–Cys crosslink, ensuring catalytic activity of this enzyme in solid state.

Highlights

- BP76 is a blue laccase involved in termite suicidal defense
- Crystal structure of native BP76 isolated from natural source solved by copper SAD
- BP76 is stabilized by compact folding, disulfides, dimerization, and glycosylation
- K189 is oxidized to allysine and forms a hemithioacetal cross-link with C292

Škerlová et al., 2024, Structure *32*, 1–5
October 3, 2024 © 2024 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.
https://doi.org/10.1016/j.str.2024.07.015

CellPress

Please cite this article in press as: Škerlová et al., Crystal structure of blue laccase BP76, a unique termite suicidal defense weapon, Structure (2024), https://doi.org/10.1016/j.str.2024.07.015

Structure

CellPress

Short Article

Crystal structure of blue laccase BP76, a unique termite suicidal defense weapon

Jana Škerlová,[1] Jiří Brynda,[1] Jan Šobotník,[2,3] Marek Zákopčaník,[4] Petr Novák,[4] Thomas Bourguignon,[5] David Sillam-Dussès,[6] and Pavlína Řezáčová[1,7,*]

[1]Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, 160 00 Prague, Czech Republic
[2]Faculty of Tropical AgriSciences, Czech University of Life Sciences, 165 00 Prague, Czech Republic
[3]Institute of Entomology, Biology Centre, Czech Academy of Sciences, 370 05 České Budějovice, Czech Republic
[4]Institute of Microbiology, Czech Academy of Sciences, 142 20 Prague, Czech Republic
[5]Okinawa Institute of Science and Technology Graduate University, Okinawa 904-0495, Japan
[6]Laboratory of Experimental and Comparative Ethology, UR 4443, Université Sorbonne Paris Nord, 93430 Villetaneuse, France
[7]Lead contact
*Correspondence: rezacova@uochb.cas.cz
https://doi.org/10.1016/j.str.2024.07.015

SUMMARY

Aging workers of the termite *Neocapritermes taracua* can defend their colony by sacrificing themselves by body rupture, mixing the externally stored blue laccase BP76 with hydroquinones to produce a sticky liquid rich in toxic benzoquinones. Here, we describe the crystal structure of BP76 isolated from *N. taracua* in its native form. The structure reveals several stabilization strategies, including compact folding, glycosylation, and flexible loops with disulfide bridges and tight dimer interface. The remarkable stability of BP76 maintains its catalytic activity in solid state during the lifespan of *N. taracua* workers, providing old workers with an efficient defensive weapon to protect their colony.

INTRODUCTION

Workers of the termite species *Neocapritermes taracua* accumulate defensive compounds throughout their lives. As their foraging abilities decline with age, they readily use these compounds to protect their nestmates. In an act of self-sacrifice by body rupture known as autothysis, hydroquinone substrates contained in the salivary glands are mixed with blue solid protein deposits or "blue crystals," which are stored in cuticular pouches (Figure 1A). These deposits contain a single protein of approximately 76 kDa called BP76. A laccase, BP76 converts the relatively harmless hydroquinones into toxic benzoquinones (Figure 1B), which are contained in a sticky liquid that immobilizes and poisons the opponent.[1,2] This defense mechanism, comprising BP76 and the associated secretory glands, is unique among insects.[3]

RESULTS AND DISCUSSION

We isolated solid BP76 deposits from aging workers of the termite species *N. taracua* collected in French Guiana. The deposits contain pure, dimeric, thermally stable BP76 (see Figure S1), exhibiting laccase activity in the native solid state (Figure 1C). Although the BP76 deposits have been previously referred to as "crystals," we did not observe any X-ray diffraction, indicating their amorphous character. After dissolving and recrystallizing BP76 (see Figure S1), we observed that the crystals retained the enzymatic activity (Figure 1D).

We determined the structure of native, enzymatically active BP76 from a natural source at 1.8 Å by single-wavelength anomalous diffraction using a signal from copper atoms naturally occurring in the protein. BP76 is a typical blue laccase comprising three cupredoxin domains (Figure 2A).[4] The active site (Figure 2B) has a resting state configuration captured in a partially reduced state.[4,5] It contains a type 1 copper atom in domain 3, where one-electron oxidation of substrates occurs. The axial ligand of this copper is a methionine residue (M624), similar to prokaryotic and plant multicopper oxidases, but unlike the closest laccase homologs from higher fungi, where a leucine or a phenylalanine[4] serves as the axial ligand. During catalysis, electrons from the substrate are transferred to a trinuclear copper cluster located between domains 1 and 3, which contains one type 2 copper atom and two type 3 copper atoms (Figure 2B). At the trinuclear copper site, dioxygen is reduced to water by four electrons.

BP76 retains catalytic activity in the solid state, both in the deposits on the termite's back and in the crystals used for structural analysis. This is achieved by combining several stabilization strategies, including compact folding,[6] glycosylation,[7] and the stabilization of flexible loops by disulfide bridges[8] and/or in a tight dimer interface.[6,9] BP76 contains eight disulfide bridges located in flexible loops, which differ markedly from the two to three disulfide bridges typically found in the closest structural homologs from higher fungi and plants.[4] Moreover, BP76 exhibits native insect *N*-glycosylation at four sites, two of which undergo difucosylation, a pattern also observed in other insects.[10–12]

All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Please cite this article in press as: Škerlová et al., Crystal structure of blue laccase BP76, a unique termite suicidal defense weapon, Structure (2024), https://doi.org/10.1016/j.str.2024.07.015

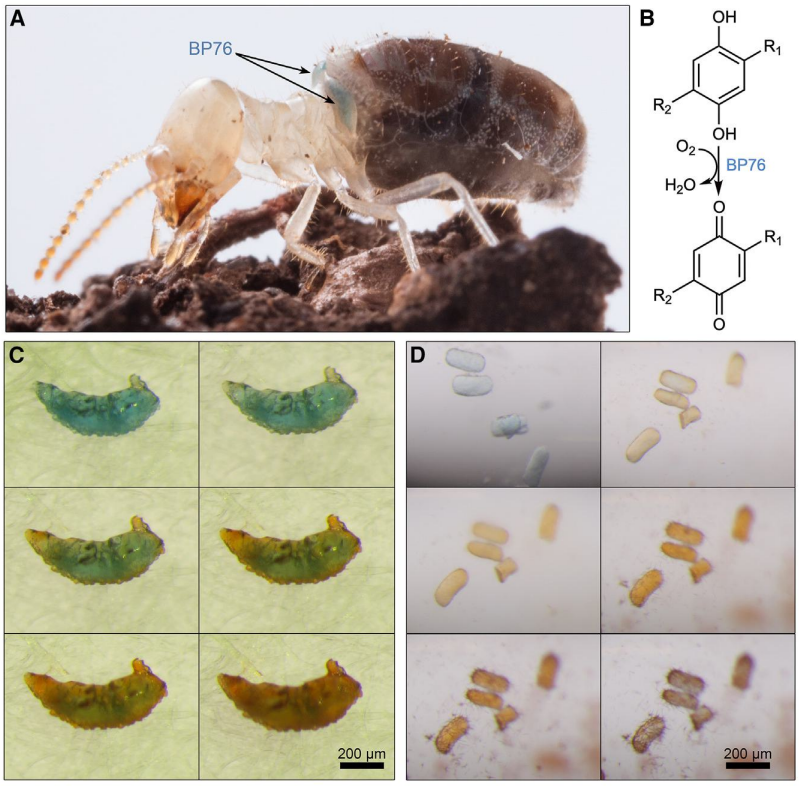


Figure 1. BP76 is enzymatically active in native solid deposits and in the recrystallized state

(A) Aging worker of *Neocapritermes taracua* with pouches containing BP76 deposits.

(B) Reaction catalyzed by BP76 (R1: -H or -CH_3, R2: -H or -OH).

(C) Laccase reaction resulting in a brown product in BP76 solid deposits at 0, 10, 20, 30, 40, and 60 s and (D) in BP76 crystals at 0, 15, and 30 s, and 5, 15, and 60 min (left to right, top to bottom).

Interestingly, some of the glycans are sulfated at the O3 oxygen of the α(1→3)-branch mannose. While Man-α(1→3) sulfation has been documented in echinoderms,[13] *N*-glycans in insects are sulfated preferentially at the α(1→6)-branch mannose[11,14,15] or disulfated at Man-α(1→3) and Man-α(1→6).[11,15]

BP76 forms a tight, chain-swapped dimer with an interface area of 3,349 Å^2 (14% of total monomer surface) and a gain in solvation free energy of −28.7 kcal/mol (Figure 2C).[16] Many of the flexible loops are involved in the interface, as is the C terminus, which folds in an extended conformation onto the surface of the second laccase molecule near the N terminus. None of the numerous structurally homologous laccases form a dimer with a similar interface; only ascorbate oxidase in zucchini has been found to have such an interface, albeit much less tight (1,851 Å^2, −16.6 kcal/mol).[17]

BP76 contains two cysteine residues that are not involved in disulfide bridges: C614, which coordinates type 1 copper; and C292 (domain 2), which forms a covalent bond with the ε-carbon of residue K189 (domain 1). This bond, located in the region of electron transfer, is linked through a chain of water molecules to the trinuclear cluster (see Figures 2A and 2B). Mass spectrometry analysis of BP76 revealed the oxidation of K189 to α-aminoadipic semialdehyde, also known as allysine (see Figure S3). Allysine is a post-translational modification generated by copper-dependent lysyl oxidase, forming cross-links in collagen and elastin.[18] It is also autogenerated in cytochrome *c* by peroxide-induced radicals, triggering its peroxidase activity in apoptosis.[19,20] Similarly, BP76 might autogenerate allysine through a radical mechanism, possibly involving reactive oxygen species. The aldehyde moiety of allysine K189 then readily reacts with the sulfhydryl of proximal C292 to form a reversible hemithioacetal cross-link (Figure 2B).

Lys–Cys cross-links have been reported,[21–23] where the amine and thiol groups are linked through an oxygen atom and can function as a redox switch and regulate enzyme function.[24,25] However, the type of Lys–Cys cross-link reported here has not been observed before. Unusual covalent connections between amino acid side chains have been identified in other metalloproteins, such as a Cys–Tyr bond in catalase,[26] and notably a His–Trp bond in multicopper bilirubin oxidase.[27,28] Some of these bonds are autogenerated by the metalloproteins and play a role in protein stabilization.[29,30] The K189–C292 cross-link overlaps with a disulfide bond in zucchini ascorbate oxidase,[17] which indicates its importance in the structure and stability of BP76.

Termite defensive fluids contain quinones,[31] and the substrate-binding pocket[32] of BP76 contains a mixture of naturally occurring BP76 reaction products such as 2-methyl-1,4-benzoquinone and 1,4-benzoquinone, which are found in the body rupture liquid of *N. taracua* workers.[1] The same quinones are found in the labial gland secretion of soldiers in several termite species.[33] A structurally similar compound, 2-hydroxy-5-methyl-1,4-benzoquinone, found in *Odontotermes magdalenae* soldiers,[34] also seems to be present in the BP76 active site. Quinones form numerous hydrophobic interactions in the hydrophobic pocket (Figure 2D). The first ketone contacts polar residues at the pocket opening, while the second ketone interacts with the catalytic histidine H619, which transfers electrons to T1 copper[32] and connects through a chain of water molecules to the K189–C292 cross-link. This interaction network alternatively involves also the 5-hydroxy group of the quinone (Figure 2B). The methyl group points toward a hydrophobic region that can, upon a slight rearrangement of the F438 and M465 residues, accommodate 2-ethyl-1,4-hydroquinone, also found in the liquid produced from the *N. taracua* defensive body rupture.[1]

Our structural study reveals the molecular architecture of a unique insect laccase that functions as a highly peculiar suicidal defense weapon to protect the termite colony. This insect laccase structure reveals native insect post-translational

Please cite this article in press as: Škerlová et al., Crystal structure of blue laccase BP76, a unique termite suicidal defense weapon, Structure (2024), https://doi.org/10.1016/j.str.2024.07.015

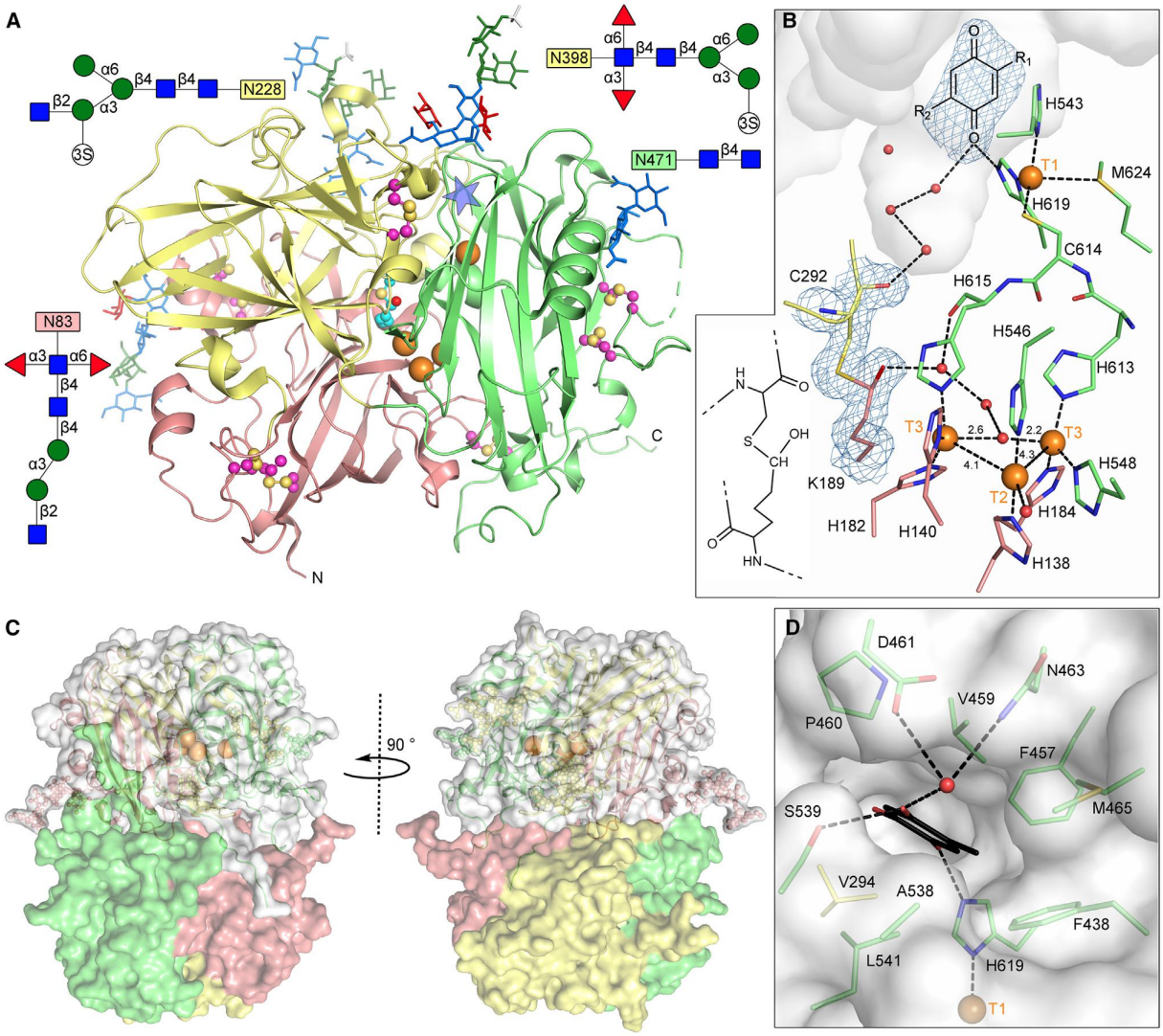


Figure 2. Structure of BP76

(A) BP76 monomer with laccase domains colored in red, yellow, and green, respectively; the active site is marked with a star. Disulfide bridges and the K189–C292 cross-link are highlighted as magenta and cyan balls and sticks, respectively. Sulfur atoms are yellow, oxygen red, and copper orange. Glycans are shown as sticks colored according to the Symbol Nomenclature for Glycans (SNFG); SNFG diagrams are also shown.

(B) BP76 active site featuring the K189–C292 cross-link, water molecules linking it to the active site, the protein's Connolly surface, a $2mF_o$-DF_c electron density map at 1σ, chemical formulas of the hemithioacetal cross-link and the reaction product (R1: -H or -CH_3, R2: -H or -OH), and distances in Å between the three copper types (T1–3).

(C) BP76 dimer shown as a surface representation with colored domains; in one monomer, a cartoon with glycans (balls and sticks) is shown below a white semitransparent surface.

(D) Active site pocket with 2-hydroxy-5-methyl-1,4-benzoquinone. Also see Figure S2 for further details.

modification. This study explains the structural basis for the remarkable stability of BP76 and its ability to maintain catalytic activity in the solid state for several months, the expected life-span of the *N. taracua* worker.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - *Neocapritermes taracua*
- METHOD DETAILS
  - BP76 deposits isolation
  - BP76 characterization

Please cite this article in press as: Škerlová et al., Crystal structure of blue laccase BP76, a unique termite suicidal defense weapon, Structure (2024), https://doi.org/10.1016/j.str.2024.07.015

- Laccase activity tests
- Crystallization
- X-ray structure determination
- Mass spectrometry data acquisition
- Mass spectrometry data analysis

- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.str.2024.07.015.

ACKNOWLEDGMENTS

We would like to thank our colleagues from IOCB, Prague – Dr. Zdeněk Voburka for the N-terminal sequencing, Dr. Pavel Majer for discussions about chemistry, and Dr. Petr Pachl for help with X-ray data analysis. Diffraction data were collected at beamline MX14.1 at the BESSY II electron-storage ring operated by the Helmholtz-Zentrum Berlin, and we would like to thank Dr. Frank Lennartz and Dr. Manfred Weiss for their support. Jan Šobotník was supported by the Internal Grant Agency of the Faculty of Tropical AgriSciences, Czech University of Life Sciences, Prague, Czech Republic (grant no. 20243114). Petr Novák acknowledges the Ministry of Education, Youth and Sports of the Czech Republic grant PHOTOMACHINES – Photosynthetic cell redesign for high yields of therapeutic peptides (CZ.02.01.01/00/22_008/0004624) and the Structural Mass Spectrometry Core Facility of CIISB, Instruct-CZ Centre, supported by MEYS CR (LM2018127) and the European Regional Development Fund Project, UP "CIISB" (no. CZ.02.1.01/0.0/0.0/18_046/0015974).

AUTHOR CONTRIBUTIONS

Conceptualization, P.Ř.; methodology, J. Škerlová, J.B., and P.N.; formal analysis, J. Škerlová, J.B., M.Z., and P.N.; investigation, J. Škerlová, J.B., M.Z., and P.N.; resources, D.S.-D., J. Šobotník, P.N., and P.Ř.; data curation, J.B. and J. Škerlová; writing – original draft, J. Škerlová; writing – review and editing, J. Škerlová, J.B., P.N., J. Šobotník, T.B., D.S.-D., and P.Ř.; visualization, J. Škerlová and M.Z.; supervision, J. Šobotník, T.B., and P.R.; funding acquisition, J. Šobotník and P.N.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 12, 2024
Revised: June 21, 2024
Accepted: July 22, 2024
Published: August 15, 2024

REFERENCES

1. Bourguignon, T., Šobotník, J., Brabcová, J., Sillam-Dussès, D., Buček, A., Krasulová, J., Vytisková, B., Demianová, Z., Mareš, M., Roisin, Y., and Vogel, H. (2016). Molecular Mechanism of the Two-Component Suicidal Weapon of Neocapritermes taracua Old Workers. Mol. Biol. Evol. *33*, 809–819. https://doi.org/10.1093/molbev/msv273.
2. Sobotnik, J., Bourguignon, T., Hanus, R., Demianova, Z., Pytelkova, J., Mares, M., Foltynova, P., Preisler, J., Cvacka, J., Krasulova, J., and Roisin, Y. (2012). Explosive backpacks in old termite workers. Science *337*, 436. https://doi.org/10.1126/science.1219129.
3. Sobotnik, J., Kutalova, K., Vytiskova, B., Roisin, Y., and Bourguignon, T. (2014). Age-dependent changes in ultrastructure of the defensive glands of Neocapritermes taracua workers (Isoptera, Termitidae). Arthropod Struct. Dev. *43*, 205–210. https://doi.org/10.1016/j.asd.2014.02.003.
4. Hakulinen, N., and Rouvinen, J. (2015). Three-dimensional structures of laccases. Cell. Mol. Life Sci. *72*, 857–868. https://doi.org/10.1007/s00018-014-1827-5.
5. Rulisek, L., and Ryde, U. (2013). Theoretical studies of the active-site structure, spectroscopic and thermodynamic properties, and reaction mechanism of multicopper oxidases. Coord. Chem. Rev. *257*, 445–458.
6. Russell, R.J., Ferguson, J.M., Hough, D.W., Danson, M.J., and Taylor, G.L. (1997). The crystal structure of citrate synthase from the hyperthermophilic archaeon pyrococcus furiosus at 1.9 A resolution. Biochemistry *36*, 9983–9994. https://doi.org/10.1021/bi9705321.
7. Shental-Bechor, D., and Levy, Y. (2009). Folding of glycoproteins: toward understanding the biophysics of the glycosylation code. Curr. Opin. Struct. Biol. *19*, 524–533. https://doi.org/10.1016/j.sbi.2009.07.002.
8. Matsumura, M., Becktel, W.J., Levitt, M., and Matthews, B.W. (1989). Stabilization of phage T4 lysozyme by engineered disulfide bonds. Proc. Natl. Acad. Sci. USA *86*, 6562–6566. https://doi.org/10.1073/pnas.86.17.6562.
9. Salminen, T., Teplyakov, A., Kankare, J., Cooperman, B.S., Lahti, R., and Goldman, A. (1996). An unusual route to thermostability disclosed by the comparison of Thermus thermophilus and Escherichia coli inorganic pyrophosphatases. Protein Sci. *5*, 1014–1025. https://doi.org/10.1002/pro.5560050604.
10. Rendić, D., Wilson, I.B., and Paschinger, K. (2008). The glycosylation capacity of insect cells. Croat. Chem. Acta *81*, 7–21.
11. Stanton, R., Hykollari, A., Eckmair, B., Malzl, D., Dragosits, M., Palmberger, D., Wang, P., Wilson, I.B.H., and Paschinger, K. (2017). The underestimated N-glycomes of lepidopteran species. Biochim. Biophys. Acta. Gen. Subj. *1861*, 699–714. https://doi.org/10.1016/j.bbagen.2017.01.009.
12. Staudacher, E., Altmann, F., März, L., Hård, K., Kamerling, J.P., and Vliegenthart, J.F. (1992). Alpha 1-6(alpha 1-3)-difucosylation of the asparagine-bound N-acetylglucosamine in honeybee venom phospholipase A2. Glycoconj. J. *9*, 82–85. https://doi.org/10.1007/BF00731703.
13. Vanbeselaere, J., Jin, C., Eckmair, B., Wilson, I.B.H., and Paschinger, K. (2020). Sulfated and sialylated N-glycans in the echinoderm Holothuria atra reflect its marine habitat and phylogeny. J. Biol. Chem. *295*, 3159–3172. https://doi.org/10.1074/jbc.RA119.011701.
14. Cabrera, G., Salazar, V., Montesino, R., Támbara, Y., Struwe, W.B., Leon, E., Harvey, D.J., Lesur, A., Rincón, M., Domon, B., et al. (2016). Structural characterization and biological implications of sulfated N-glycans in a serine protease from the neotropical moth Hylesia metabus (Cramer [1775]) (Lepidoptera: Saturniidae). Glycobiology *26*, 230–250. https://doi.org/10.1093/glycob/cwv096.
15. Kurz, S., Aoki, K., Jin, C., Karlsson, N.G., Tiemeyer, M., Wilson, I.B.H., and Paschinger, K. (2015). Targeted release and fractionation reveal glucuronylated and sulphated N- and O-glycans in larvae of dipteran insects. J. Proteomics *126*, 172–188. https://doi.org/10.1016/j.jprot.2015.05.030.
16. Krissinel, E., and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. J. Mol. Biol. *372*, 774–797. https://doi.org/10.1016/j.jmb.2007.05.022.
17. Messerschmidt, A., Luecke, H., and Huber, R. (1993). X-ray structures and mechanistic implications of three functional derivatives of ascorbate oxidase from zucchini. Reduced, peroxide and azide forms. J. Mol. Biol. *230*, 997–1014. https://doi.org/10.1006/jmbi.1993.1215.
18. Eyre, D.R., Paz, M.A., and Gallop, P.M. (1984). Cross-linking in collagen and elastin. Annu. Rev. Biochem. *53*, 717–748. https://doi.org/10.1146/annurev.bi.53.070184.003441.
19. Tomaskova, N., Novak, P., Kozar, T., Petrencakova, M., Jancura, D., Yassaghi, G., Man, P., and Sedlak, E. (2021). Early modification of cytochrome c by hydrogen peroxide triggers its fast degradation. Int. J. Biol. Macromol. *174*, 413–423. https://doi.org/10.1016/j.ijbiomac.2021.01.189.
20. Yin, V., Shaw, G.S., and Konermann, L. (2017). Cytochrome c as a Peroxidase: Activation of the Precatalytic Native State by H(2)O(2)-Induced Covalent Modifications. J. Am. Chem. Soc. *139*, 15701–15709. https://doi.org/10.1021/jacs.7b07106.
21. Matthews, B.W. (2021). Recognizing lysine-cysteine crosslinks in proteins. Protein Sci. *30*, 1491–1492. https://doi.org/10.1002/pro.4149.

Please cite this article in press as: Škerlová et al., Crystal structure of blue laccase BP76, a unique termite suicidal defense weapon, Structure (2024), https://doi.org/10.1016/j.str.2024.07.015

22. Ruszkowski, M., and Dauter, Z. (2016). On methylene-bridged cysteine and lysine residues in proteins. Protein Sci. *25*, 1734–1736. https://doi.org/10.1002/pro.2958.

23. Wang, J. (2019). Crystallographic identification of spontaneous oxidation intermediates and products of protein sulfhydryl groups. Protein Sci. *28*, 472–477. https://doi.org/10.1002/pro.3568.

24. Wensien, M., von Pappenheim, F.R., Funk, L.M., Kloskowski, P., Curth, U., Diederichsen, U., Uranga, J., Ye, J., Fang, P., Pan, K.T., et al. (2021). A lysine-cysteine redox switch with an NOS bridge regulates enzyme function. Nature *593*, 460–464. https://doi.org/10.1038/s41586-021-03513-3.

25. Rabe von Pappenheim, F., Wensien, M., Ye, J., Uranga, J., Irisarri, I., de Vries, J., Funk, L.M., Mata, R.A., and Tittmann, K. (2022). Widespread occurrence of covalent lysine-cysteine redox switches in proteins. Nat. Chem. Biol. *18*, 368–375. https://doi.org/10.1038/s41589-021-00966-5.

26. Diaz, A., Horjales, E., Rudino-Pinera, E., Arreola, R., and Hansberg, W. (2004). Unusual Cys-Tyr covalent bond in a large catalase. J. Mol. Biol. *342*, 971–985. https://doi.org/10.1016/j.jmb.2004.07.027.

27. Akter, M., Tokiwa, T., Shoji, M., Nishikawa, K., Shigeta, Y., Sakurai, T., Higuchi, Y., Kataoka, K., and Shibata, N. (2018). Redox Potential-Dependent Formation of an Unusual His-Trp Bond in Bilirubin Oxidase. Chemistry *24*, 18052–18058. https://doi.org/10.1002/chem.201803798.

28. Koval, T., Svecova, L., Ostergaard, L.H., Skalova, T., Duskova, J., Hasek, J., Kolenko, P., Fejfarova, K., Stransky, J., Trundova, M., and Dohnalek, J. (2019). Trp-His covalent adduct in bilirubin oxidase is crucial for effective bilirubin binding but has a minor role in electron transfer. Sci. Rep. *9*, 13700. https://doi.org/10.1038/s41598-019-50105-3.

29. Cooley, R.B., Rhoads, T.W., Arp, D.J., and Karplus, P.A. (2011). A diiron protein autogenerates a valine-phenylalanine cross-link. Science *332*, 929. https://doi.org/10.1126/science.1205687.

30. Griese, J.J., Roos, K., Cox, N., Shafaat, H.S., Branca, R.M.M., Lehtiö, J., Gräslund, A., Lubitz, W., Siegbahn, P.E.M., and Högbom, M. (2013). Direct observation of structurally encoded metal discrimination and ether bond formation in a heterodinuclear metalloprotein. Proc. Natl. Acad. Sci. USA *110*, 17189–17194. https://doi.org/10.1073/pnas.1304368110.

31. Prestwich, G.D. (1984). Defense Mechanisms of Termites. Annu. Rev. Entomol. *29*, 201–232. https://doi.org/10.1146/annurev.en.29.010184.001221.

32. Bertrand, T., Jolivalt, C., Briozzo, P., Caminade, E., Joly, N., Madzak, C., and Mougin, C. (2002). Crystal structure of a four-copper laccase complexed with an arylamine: insights into substrate recognition and correlation with kinetics. Biochemistry *41*, 7325–7333. https://doi.org/10.1021/bi0201318.

33. Sobotnik, J., Jirosova, A., and Hanus, R. (2010). Chemical warfare in termites. J. Insect Physiol. *56*, 1012–1021. https://doi.org/10.1016/j.jinsphys.2010.02.012.

34. Olagbemiro, T.O., Lajide, L., Sani, K.M., and Staddon, B.W. (1988). 2-Hydroxy-5-methyl-1,4-benzoquinone from the salivary gland of the soldier termitesOdontotermes magdalenae. Experientia *44*, 1022–1024. https://doi.org/10.1007/BF01939911.

35. Kabsch, W. (2010). Xds. Acta Crystallogr. D Biol. Crystallogr. *66*, 125–132. https://doi.org/10.1107/S0907444909047337.

36. Bury, C.S., Brooks-Bartlett, J.C., Walsh, S.P., and Garman, E.F. (2018). Estimate your dose: RADDOSE-3D. Protein Sci. *27*, 217–228. https://doi.org/10.1002/pro.3302.

37. Sheldrick, G.M. (2010). Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. Acta Crystallogr. D Biol. Crystallogr. *66*, 479–485. https://doi.org/10.1107/s0907444909038360.

38. Cowtan, K. (2006). The Buccaneer software for automated model building. 1. Tracing protein chains. Acta Crystallogr. D Biol. Crystallogr. *62*, 1002–1011. https://doi.org/10.1107/S0907444906022116.

39. Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G.W., McCoy, A., et al. (2011). Overview of the CCP4 suite and current developments. Acta Crystallogr. D Biol. Crystallogr. *67*, 235–242. https://doi.org/10.1107/s0907444910045749.

40. Kovalevskiy, O., Nicholls, R.A., Long, F., Carlon, A., and Murshudov, G.N. (2018). Overview of refinement procedures within REFMAC5: utilizing data from different sources. Acta Crystallogr. D Struct. Biol. *74*, 215–227. https://doi.org/10.1107/s2059798318000979.

41. Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. *60*, 2126–2132. https://doi.org/10.1107/S0907444904019158.

42. Moriarty, N.W., Grosse-Kunstleve, R.W., and Adams, P.D. (2009). electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation. Acta Crystallogr. D Biol. Crystallogr. *65*, 1074–1080. https://doi.org/10.1107/S0907444909029436.

43. Liebschner, D., Afonine, P.V., Baker, M.L., Bunkóczi, G., Chen, V.B., Croll, T.I., Hintze, B., Hung, L.W., Jain, S., McCoy, A.J., et al. (2019). Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. Acta Crystallogr. D Struct. Biol. *75*, 861–877. https://doi.org/10.1107/S2059798319011471.

44. Liebschner, D., Afonine, P.V., Moriarty, N.W., Poon, B.K., Sobolev, O.V., Terwilliger, T.C., and Adams, P.D. (2017). Polder maps: improving OMIT maps by excluding bulk solvent. Acta Crystallogr. D Struct. Biol. *73*, 148–157. https://doi.org/10.1107/S2059798316018210.

45. Williams, C.J., Headd, J.J., Moriarty, N.W., Prisant, M.G., Videau, L.L., Deis, L.N., Verma, V., Keedy, D.A., Hintze, B.J., Chen, V.B., et al. (2018). MolProbity: More and better reference data for improved all-atom structure validation. Protein Sci. *27*, 293–315. https://doi.org/10.1002/pro.3330.

46. Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. Nat. Struct. Biol. *10*, 980. https://doi.org/10.1038/nsb1203-980.

47. Holm, L., and Rosenström, P. (2010). Dali server: conservation mapping in 3D. Nucleic Acids Res. *38*, W545–W549. https://doi.org/10.1093/nar/gkq366.

48. Jubb, H.C., Higueruelo, A.P., Ochoa-Montaño, B., Pitt, W.R., Ascher, D.B., and Blundell, T.L. (2017). Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. J. Mol. Biol. *429*, 365–371. https://doi.org/10.1016/j.jmb.2016.12.004.

49. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 3.0.

50. Ericsson, U.B., Hallberg, B.M., Detitta, G.T., Dekker, N., and Nordlund, P. (2006). Thermofluor-based high-throughput stability optimization of proteins for structural studies. Anal. Biochem. *357*, 289–298. https://doi.org/10.1016/j.ab.2006.07.027.

51. Písačková, J., Procházková, K., Fábry, M., and Řezáčová, P. (2013). Crystallization of the Effector-Binding Domain of Repressor DeoR from Bacillus subtilis. Cryst. Growth Des. *13*, 844–848. https://doi.org/10.1021/cg301551v.

52. Lennartz, F., Jeoung, J.H., Ruenger, S., Dobbek, H., and Weiss, M.S. (2022). Determining the oxidation state of elements by X-ray crystallography. Acta Crystallogr. D Struct. Biol. *78*, 238–247. https://doi.org/10.1107/S2059798321013048.

Please cite this article in press as: Škerlová et al., Crystal structure of blue laccase BP76, a unique termite suicidal defense weapon, Structure (2024), https://doi.org/10.1016/j.str.2024.07.015

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
Perfluoropolyether Cryo Oil	Hampton Research	Cat#HR2-814
Gel Filtration Markers Kit for Protein Molecular Weights 6,500-66,000 Da	Sigma-Aldrich	Cat#MWGF70
Gel Filtration Markers Kit for Protein Molecular Weights 12,000-200,000 Da	Sigma-Aldrich	Cat#MWGF200
Sypro Orange	Invitrogen	S6650
4-ethylmorpholine	Sigma-Aldrich	Cat#04499
Tris(2-carboxyethyl)phosphine hydrochloride	Sigma-Aldrich	Cat# 68957
2-Chloroacetamide	Sigma-Aldrich	Cat# 22790
Sequencing Grade Modified Trypsin	Promega	Cat#V5111
Lys-C, Mass Spec Grade	Promega	Cat#VA1170
Water for chromatography (LC-MS Grade) LiChrosolv®	Supelco	Cat#1.15333
Acetonitrile hypergrade for LC-MS LiChrosolv®	Supelco	Cat#1.00029
Acetic acid 100% for LC-MS LiChropur™	Supelco	Cat#5.33001
Formic acid 98% - 100% for LC-MS LiChropur™	Supelco	Cat#5.33002
Trifluoroacetic acid eluent additive for LC-MS, LiChropur™	Supelco	Cat#80457
Critical commercial assays		
JCSG+ Screen	Qiagen	Cat#130720
Index screen	Hampton Research	Cat#HR2-144
Deposited data		
X-ray structure of BP76	This paper	PDB: 8R7I
Mass spectrometric analysis of BP76	This paper	ProteomeXchange: PXD053642 http://www.ebi.ac.uk/pride
Software and algorithms		
XDS Versions Jun 1, 2017 and Jun 30, 2023	Kabsch[35]	https://xds.mr.mpg.de/
RADDOSE-3D V3.0.803	Bury et al.[36]	https://raddo.se/
SHELXC v2016/1, SHELXD v2013/2, SHELXE v2016/3	Sheldrick[37]	https://shelx.uni-goettingen.de/
Buccaneer v1.5	Cowtan[38]	https://www.ccp4.ac.uk/html/cbuccaneer.html
CCP4 8.0.017	Winn et al.[39]	https://www.ccp4.ac.uk/
Refmac 5.8.0425	Kovalevskiy et al.[40]	https://www2.mrc-lmb.cam.ac.uk/groups/murshudov/content/refmac/refmac.html
Coot 0.9.8.92	Emsley and Cowtan[41]	https://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot/
eLBOW	Moriarty et al.[42]	https://phenix-online.org/documentation/reference/elbow.html
Phenix 1.20.1-4487	Liebschner et al.[43]	https://phenix-online.org/
Polder	Liebschner et al.[44]	https://phenix-online.org/version_docs/1.18rc2-3788/reference/polder.html
MolProbity server	Williams et al.[45]	http://molprobity.biochem.duke.edu/
PDB validation server	Berman et al.[46]	https://validate-rcsb-2.wwpdb.org/
Dali server	Holm and Rosenstrom[47]	http://ekhidna2.biocenter.helsinki.fi/dali/
PISA server	Krissinel and Henrick[16]	https://www.ebi.ac.uk/pdbe/pisa/
Arpeggio web server	Jubb et al.[48]	https://biosig.lab.uq.edu.au/arpeggioweb/
Pymol 3.0.0	Schrodinger, LLC[49]	https://pymol.org/

(*Continued on next page*)

Please cite this article in press as: Škerlová et al., Crystal structure of blue laccase BP76, a unique termite suicidal defense weapon, Structure (2024), https://doi.org/10.1016/j.str.2024.07.015

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
FragPipe 21.1	Nesvizhskii Lab, University of Michigan, MI, USA	https://fragpipe.nesvilab.org/
Other		
Äkta Basic FPLC system	Amersham Biosciences	N/A
LightCycler 480 II	Roche	https://sequencing.roche.com/us/en/products/group/lightcycler-480-ii.html
Procise 491 HT Protein Sequencer	Applied Biosystems	N/A
MRC 3 lense crystallization plates	Swissci	Cat#3W96T-UVP
Crystal Gryphon	Art Robbins Instruments	https://www.artrobbins.com/gryphon
Oryx8	Douglas Instruments	https://www.douglas.co.uk/oryx8.htm
Minstrel Desktop Crystal Imaging System	Rigaku	N/A
timsTOF SCP	Bruker Daltonics	https://www.bruker.com/en/products-and-solutions/mass-spectrometry/timstof/timstof-scp.html
PepMap™ Neo Trap Cartridge	Thermo Fisher Scientific	Cat#174500
Double nanoViper™ PepMap™ Neo UHPLC Columns	Thermo Fisher Scientific	Cat#DNV75150PN
Vanquish™ Neo UHPLC System	Thermo Fisher Scientific	Cat#VN-S10-A-01

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Pavlina Rezacova (rezacova@uochb.cas.cz).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Coordinates and structure factors have been deposited at PDB and are publicly available as of the date of publication. Mass spectrometry data have been deposited at ProteomeXchange and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Neocapritermes taracua

Workers of *Neocapritermes taracua* were isolated from several nest fragments collected on December the 5th, 2014, in the forest near Petit Saut dam (N 5.06656°, W 52.98297°), French Guiana.

METHOD DETAILS

BP76 deposits isolation

Living *N. taracua* workers were gently extracted from all nest fragments and immobilized on ice, and older individuals with visible BP76 deposits were manually separated and stored at -20°C until further use. The BP76 deposits, each weighing about 5 to 30 μg per individual, were carefully dissected from several thousands of old workers, yielding a total amount of approximately 35 mg of BP76 deposits. The dissected protein deposits were stored at -20°C until further use.

BP76 characterization

The solid deposits were dissolved in 20 mM Hepes pH 7.0, 100 mM NaCl and 40 μg of the dissolved protein were analyzed by silver-stained SDS-PAGE under reducing conditions using standard protocols.

Analytical size-exclusion chromatography was performed with 30 μl of BP76 at 75 mg/ml in 20 mM Hepes pH 7.0, 100 mM NaCl on Superdex200 10/300 GL using Äkta Basic FPLC system (Amersham Biosciences, UK), at a flow rate of 0.5 ml/min. The molecular weight calibration standards (Gel Filtration Markers Kit for Protein Molecular Weights 6,500-66,000 Da, Sigma-Aldrich, USA)

Please cite this article in press as: Škerlová et al., Crystal structure of blue laccase BP76, a unique termite suicidal defense weapon, Structure (2024), https://doi.org/10.1016/j.str.2024.07.015

contained aprotinin (6.5 kDa), cytochrome C (12.4 kDa), carbonic anhydrase (29 kDa), bovine serum albumin (66 kDa), and alcohol dehydrogenase (150 kDa) was added from the Gel Filtration Markers Kit for Protein Molecular Weights 12,000-200,000 Da, Sigma-Aldrich, USA).

Protein stability was tested by Thermofluor assay[50] at the final protein concentration of 0.9 mg/ml in the presence of 8× concentrated Sypro Orange (Invitrogen, USA) using a real-time PCR LightCycler 480 II (Roche, Switzerland) as described before.[51] A temperature gradient from 20 to 95°C with a rate of 0.83°C/min was applied. The excitation wavelength was 465 nm, and fluorescence intensity was recorded at 580 nm.

The N-terminal amino acid sequence was determined using standard Edman degradation protocol and the Procise Protein Sequencing System (Procise 491 HT Protein Sequencer, Applied Biosystems, USA).

Laccase activity tests

Enzymatic activity was tested in the solid BP76 deposits and in BP76 crystals. In the case of the solid deposits, laccase reaction was followed by placing a native solid laccase deposit on filter paper soaked with a saturated solution of the colorigenic substrate 2,6-dimethoxyphenol. Photographs were taken after 0, 10, 20, 30, 40, and 60 seconds. The deposits turned orange upon production of the colored product. In the case of BP76 crystals, laccase reaction was followed by placing the crystals into a saturated solution of the colorigenic substrate 2,6-dimethoxyphenol in the crystallization solution. Pictures were taken after 0, 15, and 30 seconds, and 5, 15, and 60 minutes. The crystals turned orange upon production of the colored product, which later formed brown crystals on the surface of the BP76 crystals that turned blue again after substrate depletion.

Crystallization

Based on the results of the Thermofluor buffer screen the deposits were dissolved for crystallization experiments in 20 mM Bicine pH 9.0, 100 mM NaCl. Crystallization trials were performed at 18°C by the sitting-drop vapor diffusion technique in MRC 3 lense crystallization plates (Swissci, UK) with a reservoir volume of 40 μl using the Crystal Gryphon (Art Robbins Instruments, USA) and Oryx8 (Douglas Instruments, UK) crystallization robots as well as the Minstrel Desktop Crystal Imaging System (Rigaku, Japan). First lentil-shaped microcrystals were obtained with 45 mg/ml BP76 and JCSG$^+$ Screen (Qiagen, Netherlands) condition #69 (2.4 M sodium malonate pH 7.0) in sitting drops containing 200 nl of each protein and precipitant solutions after 3 weeks. The microcrystals were harvested into a tube with 50 μl of precipitant and vortexed with a steel seed bead (Hampton Research, USA). The resulting seed stock was matrix micro-seeded into a fresh plate under identical experimental conditions, but using the Index screen (Hampton Research, USA) and replacing 40 nl of reservoir solution in each drop with 40 nl of the seed stock. Hexagonal rod-shaped crystals of good diffraction quality grew in 2-3 weeks in Index condition #2 (100 mM sodium acetate pH 4.5, 2 M ammonium sulfate) and were harvested after reaching the final size in about six weeks, washed in Perfluoropolyether Cryo Oil (Hampton Research, USA) for cryoprotection, and flash-cooled in liquid nitrogen for diffraction experiments. Additionally, elongated crystals with blunt edges grew under the same experimental conditions, both with and without using the seed stock, in Index condition #4 (100 mM Bis-Tris pH 6.5, 2 M ammonium sulfate) in 1-3 weeks. These crystals had worse diffraction quality and were used for testing the enzymatic activity.

X-ray structure determination

Two isomorphous diffraction data sets were collected at the wavelengths of 1.362 Å and 0.918 Å at beamline MX14.1 at the BESSY II electron-storage ring operated by the Helmholtz-Zentrum Berlin. The diffraction data were processed in XDS[35] and the resolution limit was based on the R_{meas} values that increased below the resolution of 1.8 Å. Estimated average diffraction weighted dose was calculated by the program Raddose-3D[36] at BESSY II[52]; extrapolations were made using the closest-in-date calibrations from year 2018, and for the anomalous data collected at 9.103 keV, extrapolations were done using calibrations at 13.5 keV. The structure was determined by the single-wavelength anomalous diffraction method using the SHELX suite.[37] Anomalous signal from native copper atoms was analyzed in SHELXC. Initial phases were obtained by SHELXD. Heavy-atom positions for all four expected copper atoms with occupancies between 0.864 and 1.000 were identified by SHELXD, which generated the substructure and initial electron density maps. SHELXE was used for density modification using the native data set and to build an initial polyalanine model. Further model building was done automatically in Buccaneer[38] from the CCP4 package[39] and the model was refined in Refmac[40] and Coot.[41] Ligand geometry definition was created in the program eLBOW[42] from the Phenix[43] suite. Omit map analysis was performed in the program Polder[44] from the Phenix suite.[43] MolProbity[45] and PDB validation[46] servers were used for model validation. For data collection and refinement statistics, see Table S1; the coordinates were deposited into the Protein Data Bank (PDB: 8R7I). Structure similarity search was performed using the Dali[47] server and dimer interfaces were analyzed using the PISA[16] server. Interatomic interactions were analyzed using the program Contact from the CCP4 package[39] and the Arpeggio web server.[48] Structure visualizations were created in Pymol.[49]

The asymmetric unit contained a single BP76 molecule, and we were able to model residues 23-505 and 509-655 of the polypeptide chain. Residues 1-22 represent a propeptide that is not present in the native protein, as confirmed by N-terminal sequencing (see Figure S5); residues 506-510 and 656-657 could not be modeled due to disorder. The electron density indicated several clear discrepancies with the published sequence (Uniprot: A0A0S3AND7) and the following mutations were made in the model: E55G, E233S, and R382S. These mutations were confirmed by mass spectrometry analysis as described below. However, given the natural origin

Please cite this article in press as: Škerlová et al., Crystal structure of blue laccase BP76, a unique termite suicidal defense weapon, Structure (2024), https://doi.org/10.1016/j.str.2024.07.015

of the protein, the structure likely represents a mixture of various protein species differing in sequence and there might be further sites of sequence ambiguity, especially in the solvent-exposed flexible regions. The following glycosylation were built in the model:

GlcNAc-β(1→2)Man-α(1→3)Man-β(1→4)GlcNAc-β(1→4)[Fuc-α(1→6)][Fuc-α(1→3)]GlcNAc-β-N83, Man-α(1→6)[GlcNAc-β(1→2)Man-3-sulfate-α(1→3)]Man-β(1→4)GlcNAc-β(1→4)GlcNAc-β-N228, Man-α(1→6)[Man-3-sulfate-α(1→3)]Man-β(1→4)GlcNAc-β(1→4)[Fuc-α(1→6)][Fuc-α(1→3)]GlcNAc-β-N398, GlcNAc-β(1→4)GlcNAc-β-N471.

The K189–C292 cross-link was clearly apparent in the SHELXE-generated electron density map, as well as in the omit map calculated using the data from the first 60° of collection (see Figure S2A).

The water molecule coordinated in the trinuclear copper cluster was modeled as a hydroxide anion, based on the local inter-atomic distances[4,5] (Figure 2B), in combination with a chloride anion, each with occupancy 0.5 (see Figure S2B).

The substrate-binding pocket contained electron density shaped as a modified phenyl ring (see Figure S2C). We could not satisfactorily explain this density with components of the crystallization solution. We have explained the density by a mixture of known products of the BP76 reaction, 2-methyl-1,4-benzoquinone and 1,4-benzoquinone, and we have also built in the model 2-hydroxy-5-methyl-1,4-benzoquinone, as the electron density map suggested an additional polar modification of the phenyl ring (see Figures S2C and S2E). Each of the three quinone ligands was modeled with an occupancy of 0.33. However, there are still unresolved peaks in the mF_o-DF_c electron density map around the ligands (see Figure S2C), suggesting that the composition of the ligand binding site might be more complicated, considering also the natural source of the preparation. Nevertheless, the Polder map analysis[44] supported the validity of the modeled quinone products (see Figure S2C).

Mass spectrometry data acquisition

One μg of protein was diluted with 50 mM 4-ethylmorpholine/acetate (pH 8.5) to a final volume of 20 μl. The sample was reduced and alkylated by adding 10 mM Tris(2-carboxyethyl)phosphine and 30 mM 2-chloroacetamide, followed by incubation at 70°C for 5 min. A two-step digestion was performed. First, LysC (Promega, USA) was added to the samples in the ratio of 1:20 and incubated at 37°C for 1 h. Second, trypsin (Promega, USA) was added in the ratio of 1:20 and incubated at 37°C overnight. The digestion process was terminated by the addition of trifluoroacetic acid to a final concentration of 0.1% (v/v).

LC-MS/MS analyses were performed by using the Vanquish Neo UHPLC system (Thermo Fisher Scientific, USA) coupled with a timsTOF SCP mass spectrometer equipped with Captive spray (Bruker Daltonics, USA). Spectra were collected in a positive data dependent acquisition mode. Peptides were trapped onto a C18 trap column (PepMap Neo C18 5 μm, 300 μm × 5 mm, Thermo Fisher Scientific, USA) in 0.1% formic acid for 3 min and then eluted to a C18 analytical column (Double nanoViper PepMap Neo 75 μm × 150 mm, 2 μm, Thermo Fisher Scientific, USA). Both columns were heated to 50°C. Peptide separation was carried out using a linear water-acetonitrile gradient of 5%-35% acetonitrile over 35 min at a flow rate of 350 nl/min. The standard proteomics PASEF method was applied for data acquisition. The scan range was set from 0.6 to 1.6 V s/cm^2 and ramp time was 100 ms. A total of 10 PASEF MS/MS scans were performed. Precursor ions selected for fragmentation were in the m/z range of 100-1700 with charge states ≥2+ and ≤6+. Active exclusion was enabled for 0.4 min to prevent repeated selection of the same precursor ions.

Mass spectrometry data analysis

MS data were searched using FragPipe 21.1 (Nesvizhskii Lab, University of Michigan, USA) against a database containing sequence of the BP76 protein (residues 23-657) and an additional sequence of the BP76 protein containing the suspected mutations, together with 243 sequences of common contaminants (including trypsin). For each database entry, a reverse decoy sequence was appended. Only tryptic peptides (without cleavage after P) with up to two miscleavages were considered. An open search was employed with the following settings. The precursor mass tolerance was set to -150 to +500 Da, the fragment mass tolerance was set to 0.05 Da, the "Report mass shift as variable mod" option was set to "Yes, keep delta mass" and the "localize mass shift" option was enabled. Cysteine carboxymethylation was defined as a fixed modification, and N-terminal acetylation and Met oxidation were defined as variable modifications. A maximum of three variable modifications per peptide was allowed. On behalf of validation, PeptideProphet (with default settings for open search) and ProteinProphet were employed (both part of Philosopher 5.1.0), together with sequential FDR filtering to 1% on the protein level.

The open search ion signal at m/z 921.43 4 by FragPipe pointed out peptide sequence ^{169}YDFFANNEGTLYWHSHDG LQKLDGIQGSIVVR^{200} with a mass shift of +0.9840 Da on K189 (see Figure S3) corresponding to lysine oxidation to α-aminoadipic semialdehyde, also known as allysine. This was supported by detecting ^{169}YDFFANNEGTLYWHSHDGLQK^{189} peptide bearing the same modification but with approximately four times lower intensity, indicating hindered cleavage properties of the modified lysine. The hemithioacetal cross-link observed in the crystal structure could not be detected, because it is unstable and fully reversible.

Furthermore, three suspected amino acid mutations were investigated, namely E55G, E233S, and R382S. Presence of all mutations, as well as the native forms was confirmed from the MS/MS data. The fragmentation spectrum of the peptide N^{46}QCARE CIAEEAPK^{59} (See Figure S4A) exhibits similar coverage as the E55G mutated form (See Figure S4B). The spectrum of the E55G mutant contains high-intensity signals of y2 to y7 fragments, providing confidence for the mutation site. In terms of quantity, the mutated peptide has around 60% of the intensity of the wild-type peptide. The E233S mutation was discovered on peptide S^{204}TDLNGDTYDLDVPDHTLLILDWINTTAGER^{234} (See Figure S4C) and the mutated form (See Figure S4D) is supported by b29 and b30 fragments. Due to the high length of the peptide and low abundance of the precursor ion, the fragmentation spectra offer only partial sequence coverage. The wild-type peptide is about two times more prevalent than the mutated form. The R382S mutation was identified through b2 to b14 series (See Figure S4F), covering the site of the mutation. The fragmentation spectrum of the

Please cite this article in press as: Škerlová et al., Crystal structure of blue laccase BP76, a unique termite suicidal defense weapon, Structure (2024), https://doi.org/10.1016/j.str.2024.07.015

wild-type peptide Y^{370}SGTNEERNEVPRYNGGFAR^{389} (See Figure S4E) is impaired by incomplete fragmentation, with the precursor ion being the second-highest signal. The remaining unannotated high-intensity signals could not be assigned. The intensity of the mutated peptide is about 45% of the intensity of the wild-type peptide.

QUANTIFICATION AND STATISTICAL ANALYSIS

X-ray diffraction data collection and refinement statistics are summarized in Table S1.

Structure, Volume *32*

Supplemental Information

Crystal structure of blue laccase BP76, a unique termite suicidal defense weapon

Jana Škerlová, Jiří Brynda, Jan Šobotník, Marek Zákopčaník, Petr Novák, Thomas Bourguignon, David Sillam-Dussès, and Pavlína Řezáčová

Table S1. Data collection and refinement statistics, related to STAR Methods.
Values in parentheses refer to the highest resolution shell.

Data collection statistics		
Dataset	Native (used for refinement)	Anomalous (used for phasing)
Space group	*P*321	*P*321
Cell parameters (Å; °)	a = b = 163.5, c = 89.2; α = β = 90, γ = 120	a = b = 162.7, c = 89.3; α = β = 90, γ = 120
No. of molecules in AU	1	1
Molecular weight (kDa)	76	76
No. of amino acids in AU	635	635
No. of Cu atoms in AU	4	4
Matthews coefficient	4.5	4.5
Solvent content (%)	73	73
Wavelength (Å)	0.918	1.362
Estimated average diffraction weighted dose (MGy)	0.90	1.74
Resolution (Å)	47.20-1.80 (1.91-1.80)	47.00-1.85 (1.96-1.85)
No. of unique reflections	126,421 (20,205)	225,770 (35,710)
Multiplicity	10.0 (10.2)	27.0 (8.6)
Completeness (%)	99.7 (99.5)	99.5 (97.3)
R_{meas}	14.3 (86.4)	9.6 (88.3)
$CC_{(1/2)}$	99.8 (90.0)	99.9 (81.1)
Average I/σ(I)	12.1 (2.4)	17.4 (2.2)
Wilson B ($Å^2$)	26.9	31.6
SHELX SAD phasing statistics		
SHELXC d''/σ(I) for resolution shells (Å)		
inf.-9.91		4.09
9.91-6.10		2.68
6.10-4.59		2.13
4.59-3.76		1.85
3.76-3.21		1.72
3.21-2.83		1.36
2.83-2.54		1.09
2.54-2.31		0.96
2.31-2.13		0.87
2.13-1.98		0.82
1.98-1.85		0.78
SHELXD no. of Cu atoms in substructure		4
SHELXD CC (%)		44.6
SHELXD CC(weak) (%)		25.0
SHELXD FOM (%)		69.6
SHELXE No. of built polyA residues		594
SHELXE pseudo-free CC (%)		80.7
SHELXE FOM (%)		73.1

Refinement statistics		
Resolution range (Å)	47.20-1.80 (1.84-1.80)	
No. of reflections in working set	123,978 (9,003)	
No. of reflections in test set	3,179 (230)	
R value (%)	13.2 (22.8)	
R_{free} value (%)	14.5 (23.6)	
Estimated overall coordinate error based on R/R_{free}	0.062/0.060	
RMSD bond length (Å)	0.015	
RMSD angle (º)	1.545	
No. of modeled residues in AU	630	
No. of atoms in AU	6,513	
No. of protein atoms in AU	5,123	
Mean B value ($Å^2$) protein/glycans/ligands/solvent	27.3/63.9/34.5/43.8	
Ramachandran plot statistics		
Residues in favored regions (%)	96.4	
Residues in allowed regions (%)	3.6	

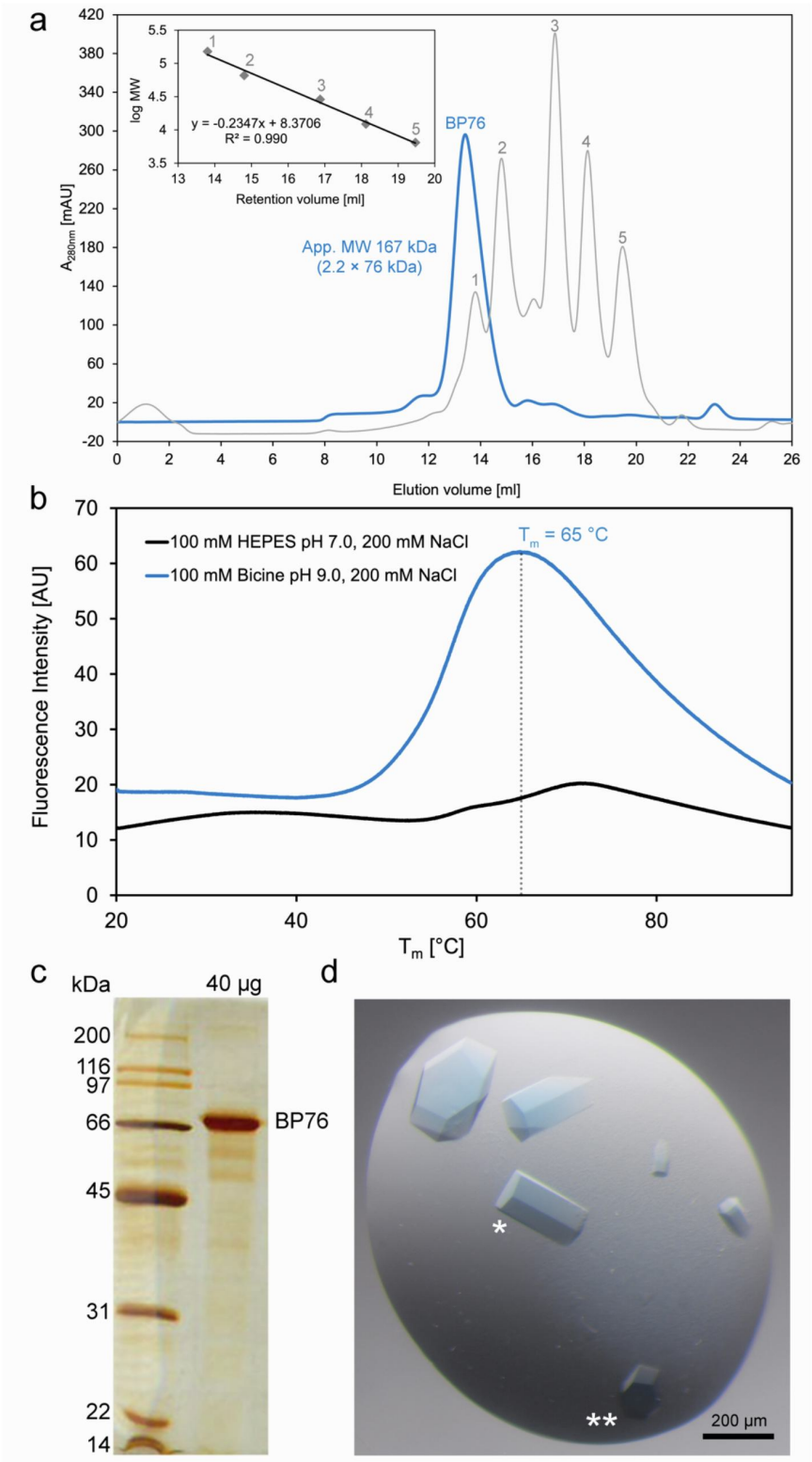


Figure S1. Characterization of BP76, related to STAR Methods. (a) Size-exclusion chromatogram documenting the dimeric state of BP76. The calibration chromatogram and curve are shown in grey for alcohol dehydrogenase (1), bovine serum albumin (2), carbonic anhydrase (3), cytochrome C (4), and aprotinin (5). (b) Thermofluor assay curves for buffer composition representing the original (black) and optimized (blue) solubilization buffers. 20 mM buffer and 100 mM salt were used for protein solubilization. (c) Silver-stained SDS-PAGE analysis of the solubilized BP76 deposits. (d) BP76 crystals

used for experimental phasing (*) and for native X-ray data collection (**) exhibited comparable diffraction quality.

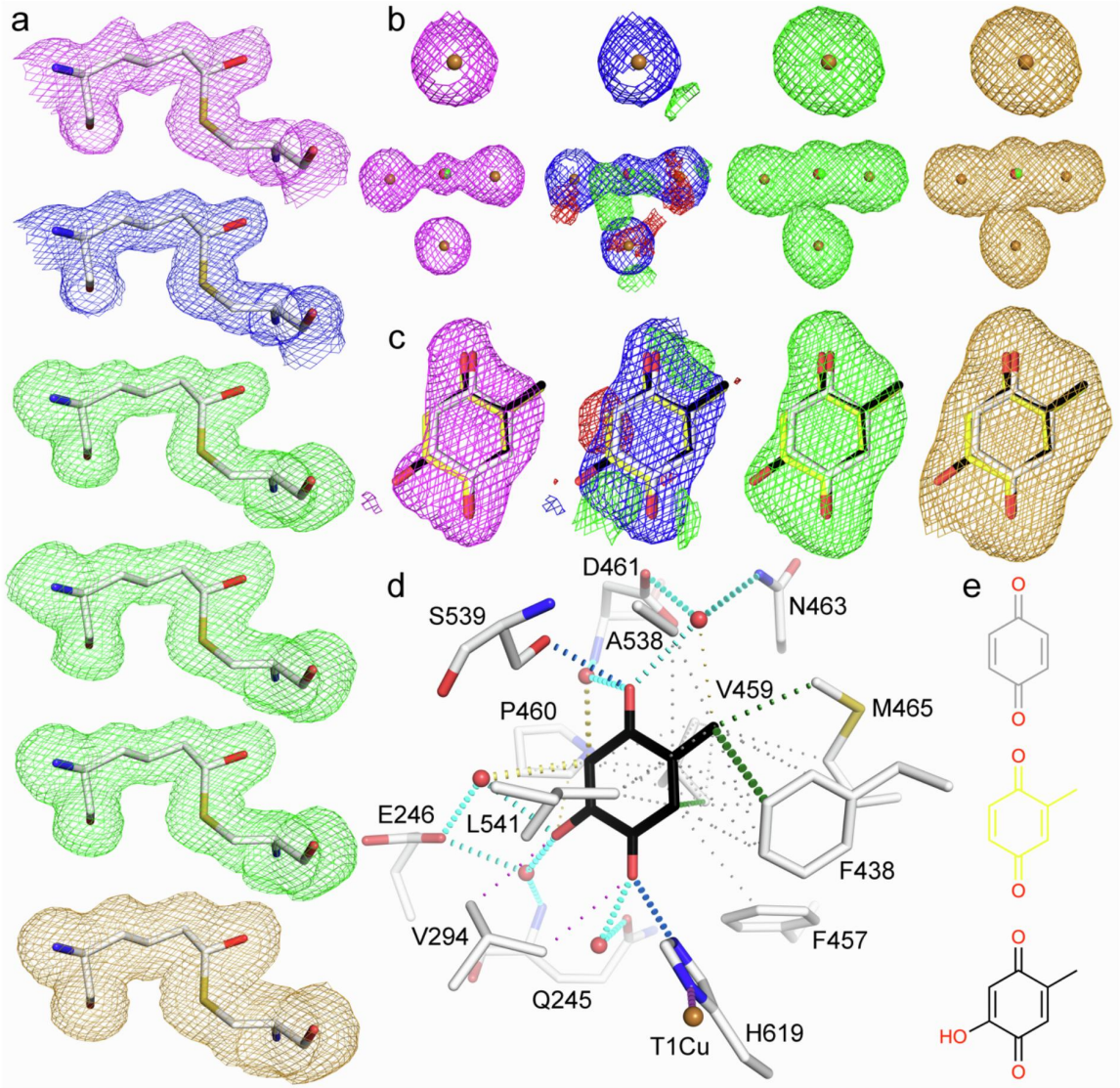


Figure S2. Electron density maps and ligand interaction details, related to STAR Methods and to Figure 2. (a) electron density maps for the K189–C292 cross-link from top to bottom: SHELX experimental map at 1σ; final refinement $2mF_o$-DF_c map (blue, 1σ) and mF_o-DF_c map (green and red, 3σ); omit map using data from the first collected 60 ° (3σ); omit map using data from the last collected 60 ° (3σ); omit map using complete data (3σ); Polder omit map (3σ). (b) electron density maps for the copper and hydroxide/chloride ions from left to right: SHELX experimental map at 3σ; final refinement $2mF_o$-DF_c map (blue, 3σ) and mF_o-DF_c map (green and red, 3σ); omit map using complete data (3σ); Polder omit map (3σ). (c) electron density maps for the ligands (each modeled with occupancy 0.33, see also panel e) from left to right: SHELX experimental map at 1σ; final refinement $2mF_o$-DF_c map (blue, 1σ) and mF_o-DF_c map (green and red, 3σ); omit map using complete data (3σ); Polder omit map (3σ). (d) details of interactions between 2-hydroxy-5-methyl-1,4-benzoquinone and BP76, as determined by Arpeggio web server, are shown as dotted lines (the thicker the dotted line, the shorter the interaction distance). Direct and water-mediated hydrogen bonds are colored blue and cyan, respectively, weak hydrogen bonds are magenta, weak polar interactions are yellow, hydrophobic interactions are green (within van der Waals distance) and grey (within 5 Å), and the coordination bond with type 1 copper (T1Cu) is purple. (e) chemical formulas for the ligands 1,4-benzoquinone (grey), 2-methyl-1,4-benzoquinone (yellow), and 2-hydroxy-5-methyl-1,4-benzoquinone (black). Oxygen atoms are colored red, nitrogen blue, sulfur yellow, copper orange, and chlorine green.

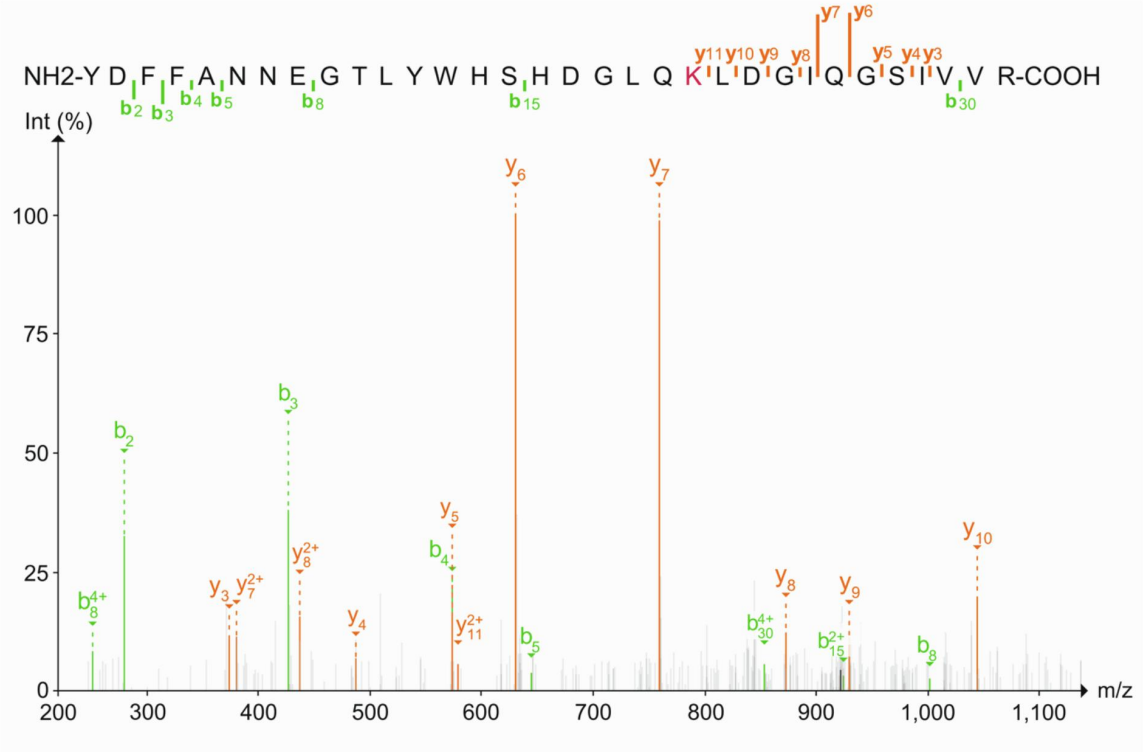


Figure S3. Mass spectrometry analysis of K189 oxidation, related to STAR Methods. Fragmentation spectrum is shown for the peptide YDFFANNEGTLYWHSHDGLQKLDGIQGSIVVR in its quadruply charged form at m/z 920.95 and RT 46.30 min bearing +0.9840 Da modification of K189 (highlighted in red). This mass shift corresponds to allysine form, *i.e.* an aldehyde moiety replacing the amino moiety of lysine.

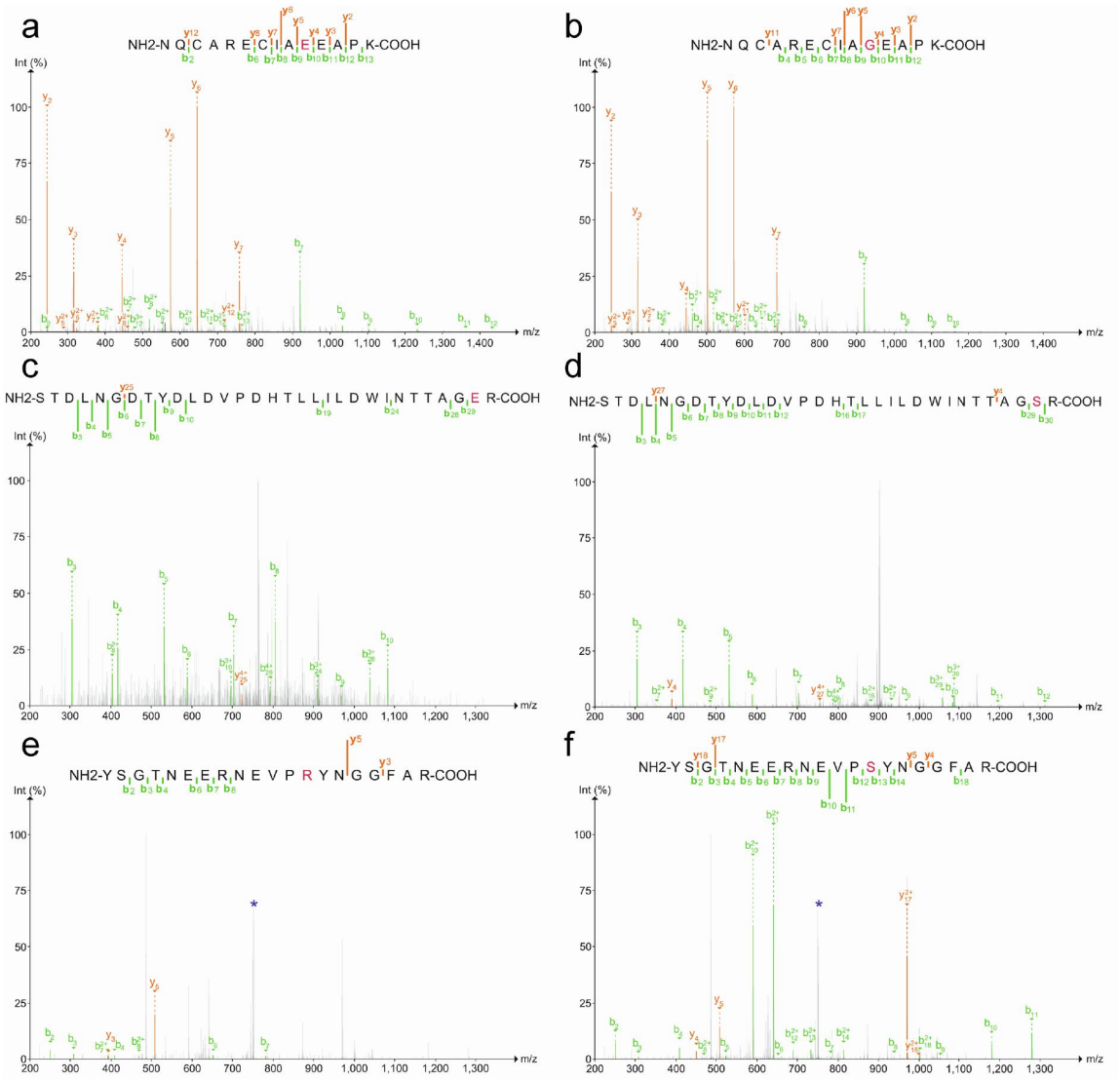


Figure S4. Mass spectrometry identification of suspected mutations in BP76, related to STAR Methods. Mutated and wild-type peptides are shown with mutation sites highlighted in red. Fragmentation spectra are shown for (a) the peptide NQCARECIAEEAPK in its triply charged form at m/z 559.26 and RT 20.06 min, and (b) the E55G mutated peptide NQCARECIAGEAPK; (c) the STDLNGDTYDLDVPDHTLLILDWINTTAGER peptide in its quintuply charged form at m/z 740.14 and RT 41.11 min, and (d) the E233S mutated peptide STDLNGDTYDLDVPDHTLLILDWINTTAGSR; (e) the peptide YSGTNEERNEVPRYNGGFAR in its triply charged form at m/z 750.33 and RT 25.02 min, and (f) the R382S mutated peptide YSGTNEERNEVPSYNGGFAR. Asterisks denote precursor ions.

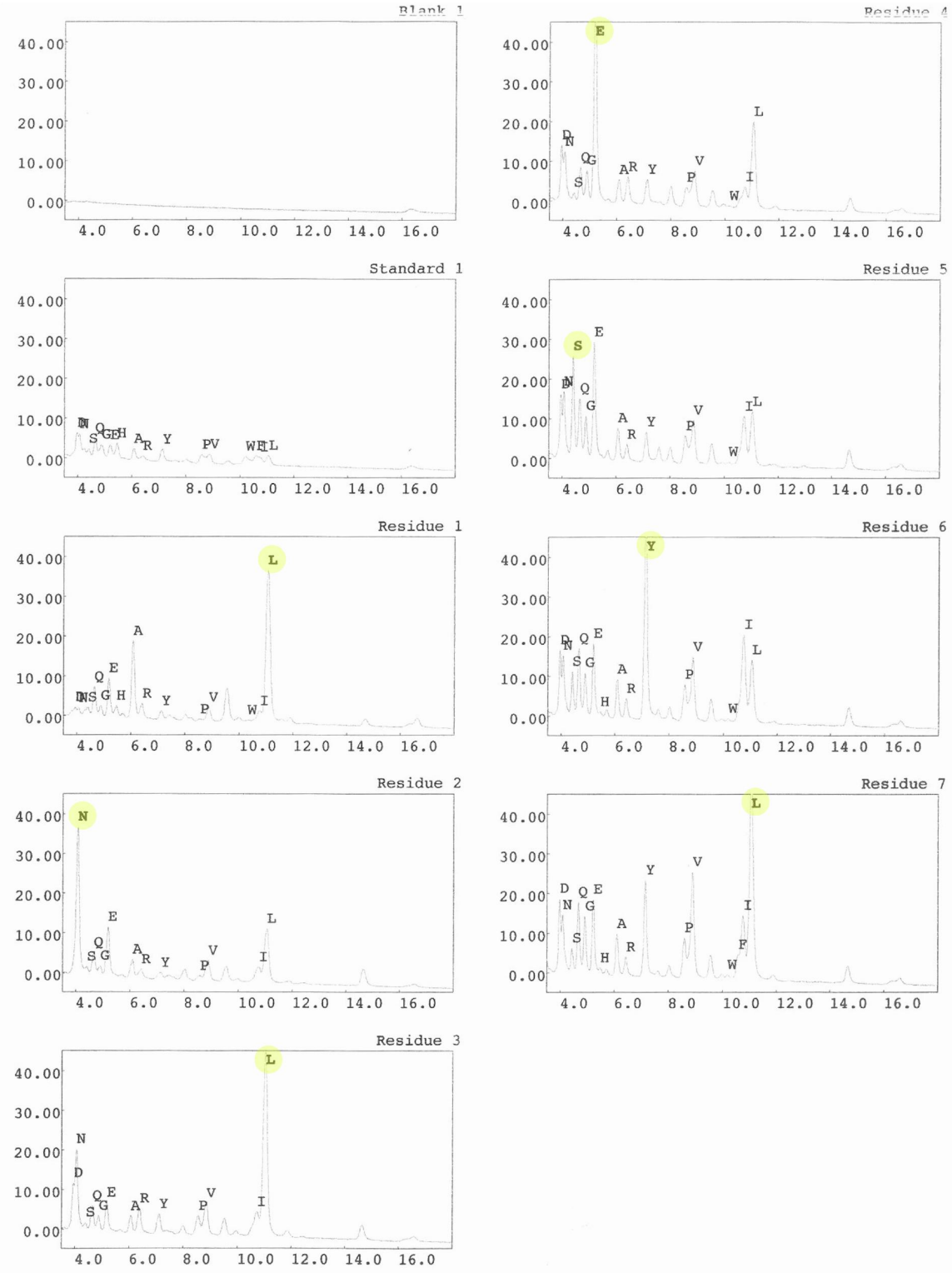


Figure S5. N-terminal sequencing of BP76, related to STAR Methods. Sequencing chromatograms are shown with time (min) on the x-axis and voltage (mV) on the y-axis. Dominant peaks corresponding to the N-terminal sequence LNLESYL are highlighted in yellow.