Supplementary information

for

Community Evaluation of Glycoproteomics Informatics Solutions Reveals High-Performance Search Strategies of Glycopeptide Data

Rebeca Kawahara[1], Kathirvel Alagesan[2], Marshall Bern[3], Weiqian Cao[4], Robert J Chalkley[5], Kai Cheng[6], Matthew S. Choo[7], Nathan Edwards[8,9], Radoslav Goldman[8,9,10], Marcus Hoffmann[11], Yingwei Hu[12], Yifan Huang[13], Jin Young Kim[14], Doron Kletter[3], Benoit Liquet-Weiland[15,16], Mingqi Liu[4], Yehia Mechref[13], Bo Meng[17], Sriram Neelamegham[6], Terry Nguyen-Khuong[7], Jonas Nilsson[18], Adam Pap[19,20], Gun Wook Park[14], Benjamin L. Parker[21], Cassandra L. Pegg[22], Josef M. Penninger[23,24], Toan K. Phung[22], Markus Pioch[11], Erdmann Rapp[11,25], Enes Sakalli[23], Miloslav Sanda[8,10], Benjamin L. Schulz[22], Nichollas E. Scott[26], Georgy Sofronov[15], Johannes Stadlmann[23], Sergey Y. Vakhrushev[27], Christina M. Woo[28], Hung-Yi Wu[28], Pengyuan Yang[4], Wantao Ying[17], Hui Zhang[12], Yong Zhang[17], Jingfu Zhao[14], Joseph Zaia[29], Stuart M. Haslam[30], Giuseppe Palmisano[31], Jong Shin Yoo[14,32], Göran Larson[33], Kai-Hooi Khoo[34], Katalin F. Medzihradszky[5,19], Daniel Kolarich[2], Nicolle H. Packer[1,2,35], and Morten Thaysen-Andersen[1,35]*

[1]Department of Molecular Sciences, Macquarie University, Sydney, NSW, Australia

[2]Institute for Glycomics, Griffith University Gold Coast Campus, QLD, Australia

[3]Protein Metrics Inc., Cupertino, CA, USA

[4]Institutes of Biomedical Sciences, and the NHC Key Laboratory of Glycoconjugates Research, Fudan University, Shanghai, China

[5]UCSF, School of Pharmacy, Department of Pharmaceutical Chemistry, San Francisco, CA, United States of America

[6]State University of New York, Buffalo, NY, United States of America

[7]Analytics Group, Bioprocessing Technology Institute, Singapore

[8]Clinical and Translational Glycoscience Research Center (CTGRC), Georgetown University, Washington, DC, United States of America

[9]Department of Biochemistry and Molecular & Cellular Biology, Georgetown University, Washington, DC, United States of America

[10]Department of Oncology, Georgetown University, Washington, DC, United States of America

[11]Max Planck Institute for Dynamics of Complex Technical Systems, Bioprocess Engineering, Magdeburg, Germany

[12]Department of Pathology, The Johns Hopkins University, Baltimore, MD, United States of America

[13]Department of Chemistry and Biochemistry, Texas Tech University, TX, United States of America

[14]Research Center of Bioconvergence Analysis, Korea Basic Science Institute, Republic of Korea

[15]Department of Mathematics and Statistics, Macquarie University, Sydney, NSW, Australia

[16]CNRS, Laboratoire de Mathématiques et de leurs Applications de PAU, E2S-UPPA, Pau, France

[17]State Key Laboratory of Proteomics, Beijing Institute of Lifeomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing, China

[18]Proteomics Core Facility, Sahlgrenska academy, University of Gothenburg, Gothenburg, Sweden

[19]BRC, Laboratory of Proteomics Research, Szeged, Hungary

[20]Doctoral School in Biology, Faculty of Science and Informatics, University of Szeged, Szeged, Hungary

[21]Department of Anatomy and Physiology, University of Melbourne, Melbourne, VIC, Australia

[22]School of Chemistry and Molecular Biosciences, University of Queensland, QLD, Australia

[23]IMBA, Institute of Molecular Biotechnology of the Austrian Academy of Sciences, Vienna, Austria

[24]Department of Medical Genetics, Life Sciences Institute, University of British Columbia, Vancouver, BC, Canada

[25]glyXera GmbH, Magdeburg, Germany

[26]Deparment of Microbiology and Immunology, University of Melbourne, Melbourne, VIC, Australia

[27]Copenhagen Center for Glycomics, Department of Cellular and Molecular Medicine, University of Copenhagen, Copenhagen, Denmark

[28]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, United States of America

[29]Department of Biochemistry, Boston University Medical Campus, Boston, MA, United States of America

[30]Department of Life Sciences, Imperial College London, London, UK

[31]Instituto de Ciências Biomédicas, Departamento de Parasitologia, Universidade de São Paulo, São Paulo, SP, Brazil

[32]Graduate School of Analytical Science and Technology, Chungnam National University, Daejeon, Republic of Korea

[33]Department of Laboratory Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

[34]Institute of Biological Chemistry, Academia Sinica, Taipei, Taiwan

[35]Biomolecular Discovery Research Centre, Macquarie University, Sydney, NSW, Australia

Running title: High-performance search strategies for glycoproteomics data analysis

Keywords: Glycoproteomics, glycopeptide, informatics, software, mass spectrometry

***Corresponding author**:

Dr Morten Thaysen-Andersen, PhD

Department of Molecular Sciences

Biomolecular Discovery Research Centre

Faculty of Science & Engineering

Macquarie University - Sydney

NSW-2109, Australia

Content

Extended methods

Study sample

Human serum from a commercial source was used for this study (product number #31876, Thermo Fisher Scientific). As a positive control, 52 fmol of a synthetic *N*-glycopeptide from human vitamin K-dependent protein C (UniProtKB, P04070, EVFVHPNYSK, $Hex_5HexNAc_4NeuAc_2$)[1], was spiked into 5 µg human serum prior to digestion. Proteins were cysteine reduced and alkylated prior to protein digestion using 1:100 (w/w, enzyme:protein substrate) sequence-grade trypsin for 16 h, 37°C in 20 mM aqueous ammonium bicarbonate, pH 8.0. Undigested protein material and large peptides were removed by filtration using a 30 kDa molecular weight cut off membrane (#88502, Thermo Fisher Scientific). The membrane was washed using 30% (v/v) methanol in 0.1% (v/v) aqueous trifluoroacetic acid (TFA). The flow through fraction was collected, evaporated using a SpeedVac, and then resuspended in 200 µL 50% (v/v) acetonitrile (ACN) in 0.1% (v/v) aqueous TFA. Glycopeptide enrichment was performed using Hypersep Retain AX columns (#60107-403, Thermo Fisher Scientific). The columns were prepared according to the manufacturer's instructions and were additionally washed with 100 mM aqueous triethylammonium acetate before equilibration with 95% (v/v) ACN in 1% (v/v) aqueous TFA. The sample was diluted in 3 mL 95% (v/v) ACN in 1% (v/v) aqueous TFA, applied to the columns, and then washed with an additional 3 mL 95% (v/v) ACN in 1% (v/v) aqueous TFA before the glycopeptides were eluted with 1 mL 50% (v/v) ACN in 0.5% (v/v) aqueous TFA. The enriched glycopeptide mixtures were dried using a SpeedVac and resuspended in 0.1% (v/v) aqueous TFA for LC-MS/MS analysis.

Mass spectrometry

The glycopeptides were separated by reversed phase nanoLC using a Thermo Scientific EASY-nLC™ 1200 UPLC system connected to a C_{18} LC column (50 cm length × 75 µm inner

diameter, Thermo Scientific™ EASY-Spray™). Separation was achieved using a 75 min 6-45% (v/v) and 3 min 45-95 % (v/v) gradient of solvent B consisting of 80% (v/v) ACN in 0.1% (v/v) aqueous formic acid in solvent A consisting of 0.1% (v/v) aqueous formic acid at a 300 nL/min flow rate. The separated glycopeptides were detected using a Thermo Scientific™ Orbitrap Fusion™ Lumos™ Tribrid™ mass spectrometer connected directed to the LC. Approximately 1 μg of peptide material was injected on the LC column per run. The same glycopeptide sample was analysed twice using two slightly different acquisition methods producing two related data files (File A and B).

For both methods, MS1 scans were acquired from *m/z* 350 –1,800 in the Orbitrap at a resolution of 120,000 and with an automatic gain control (AGC) of 4 x 10^5 and an injection time of 50 ms. Data-dependent HCD-MS/MS was performed for the 10 most intense precursor ions selecting the highest charge state and the lowest *m/z* in each MS1 full scan. The HCD-MS/MS fragment ions were recorded in the Orbitrap at a resolution of 30,000 and with an AGC of 5 x 10^4, injection time of 60 ms, normalised collision energy (NCE) of 28 and a quadrupole isolation width of 2 Th. Already selected precursors were dynamically excluded for 45 s. Product-dependent (pd) ion triggered re-isolation and fragmentation of precursor ions were enabled upon detection of selected glycan oxonium ions (*m/z* 138.0545, 204.0867 and 366.1396) if these diagnostic ions were amongst the top 20 fragment ions within each HCD-MS/MS spectrum. For File A, pd-triggered ETciD- and CID-MS/MS events were scheduled. The ETciD-MS/MS fragments were detected in the Orbitrap at a resolution of 60,000 with an AGC of 4 x 10^5, injection time of 250 ms, CID NCE of 15, and a quadrupole isolation width of 1.6 Th. Charge-dependent ETD calibration was enabled. The CID-MS/MS fragments were detected in the Orbitrap at a resolution of 30,000 with an AGC of 5 x 10^4, NCE of 30%, injection time of 54 ms, and a quadrupole isolation width of 1.6 Th. For File B, pd-triggered EThcD- and CID-MS/MS events were scheduled. The EThcD-MS/MS fragments were

detected in the Orbitrap at a resolution of 60,000 with an AGC of 4 x 10^5, injection time of 250 ms, HCD NCE of 15, and a quadrupole isolation width of 1.6 Th. Charge-dependent ETD calibration was enabled. The CID-MS/MS fragments were detected in the ion trap at unit resolution using a rapid scan method with an AGC of 1 x 10^4, injection time of 70 ms, NCE of 30, and a quadrupole isolation width of 1.6 Th. File A and B were provided to all participants as raw data files (File A: 684 MB, File B: 811 MB) or as three separate .mgf files containing peak lists of the fragment spectra from the three different fragmentation modes used for File A and B (23.9 MB – 65.6 MB). Conversion to .mgf was performed using ProteoWizard.

Search instructions and reporting template

The participants were requested to use a provided protein search space comprising the entire human proteome (20,231 UniProtKB reviewed sequences, downloaded in January 2018) for their search. In contrast to the fixed protein search space, the participants were free to choose the *N*- and *O*-glycan search space. To limit the number of study variables, participants were asked not to include xylose and any glycan substitutions (e.g. phosphate, sulphate and acetylation) in the glycan search space. The participants were requested to report their team details, identification strategy and the identified glycopeptides in a common reporting template organised as five separate sheets in an Excel file comprising the following categories of information: 1. Team and contact details, 2. Identification strategy and other study information, 3. *N*- and *O*-glycan search space, 4. List of identified *N*- and *O*-glycopeptides, 5. Summary of identified peptides. The returned reports were carefully checked for compliance to the study guideline. See PXD024101via the PRIDE repository[2] for the common reporting template and all deidentified participant reports forming the foundation of this study.

Search engines and pre- and post-processing tools used for the glycopeptide identification

A total of 13 search engines were used for glycopeptide identification: IQ-GPA v2.5[3], Protein Prospector v5.20.23[4], glyXtool^{MS} v0.1.4[5], Byonic v2.16.16[6], Sugar Qb[7], Glycopeptide Search v2.0alpha [8], GlycopeptideGraphMS v1.0/Byonic[9], GlycoPAT v2.0[10] and GPQuest v2.0[11], Mascot v2.5.1[12] or v2.2.07, MS Amanda v1.4.14.8243[13], Sequest-HT (in Proteome Discoverer v 2.2) (**Supplementary Figure S1h**). These tools were used as stand-alone tools or in combinations with other search engines while others were applied with pre- or post-processing tools, including OMSSA v2.1.8, Preview v2.13.2, Protein Prospector MS-filter, MS-GF+/PepArML and pParse v.2.0 (**Supplementary Figure S1i**).

Compilation and comparison of the participant's reports

Information of the participating teams were compiled from the returned reports (**Supplementary Table S1-S2**). The list of intact *N*- and *O*-glycopeptides reported by the 22 teams were compiled into a single table with an unique header (**Supplementary Table S3**). Additional columns were manually added to the compiled table with the purpose of standardising some of the reported text variables and generating unique identifiers (IDs) for the reported glycopeptides and their glycan compositions and source glycoproteins. The glycan composition ID was written as the generic monosaccharide composition as Hex*HexNAc*Fuc*NeuAc*, where * represents the number of the individual monosaccharide residues. Glycopeptides adducted with Na^+ and K^+ were considered and reported by some teams. The adducted glycopeptides were combined with the corresponding non-adducted monosaccharide compositions. UniProtKB identifiers were used as the source protein IDs. The glycopeptide IDs were written as the peptide sequence followed by the generic glycan composition.

The comparisons between the generic glycan compositions, source proteins and glycopeptide IDs reported by the 22 teams were performed using the "pivot table" tool available in Excel,

where the identifier type was placed in “rows”, and the team identifier in “columns”. The variables from each identifier type were compared as summed counts across the 22 teams.

Performance testing

The relative team performance was assessed using a scoring system composed of multiple independent tests designed to score the accuracy (specificity) and coverage (sensitivity) of the reported *N*- and *O*-glycopeptides in orthogonal ways. The raw scores from the individual tests (N1-N6 and O1-O5, described below) were normalised within the range 0-1. These normalised scores were used to establish an overall performance score (range 0-1) measuring the ability to perform accurate and comprehensive *N*- and *O*-glycopeptide analysis. The overall performance score was utilised to separately rank the developer and expert user teams.

a) The synthetic *N*-glycopeptide test (N1): All MS/MS spectra corresponding to the synthetic *N*-glycopeptide from human vitamin K-dependent protein C (peptide sequence: EVFVHPNYSK, glycan composition: $HexNAc_4Hex_5NeuAc_2$) were manually retrieved and annotated from File B. In total, six MS/MS spectra corresponded to the non-adducted synthetic *N*-glycopeptide in charge state 3+ and 4+ spanning the three applied fragmentation modes (HCD-, EThcD- and CID-MS/MS) (**Supplementary Figure S5a-b**). A further three MS/MS spectra (HCD-, EThcD- and CID-MS/MS) corresponded to the K^+-adducted synthetic *N*-glycopeptide in charge state 5+. The sensitivity of the test was determined as the proportion of the 9 MS/MS spectra mapping to the synthetic *N*-glycopeptide that was reported by each team adjusting for the type of fragmentation mode(s) included in their search strategies. The specificity was calculated by the proportion of correctly reported glyco-PSMs corresponding to the synthetic glycopeptide that matched the 9 annotated MS/MS spectra again adjusting for the type of fragmentation mode(s) included in the search

strategies. The score of the test was calculated by averaging the sensitivity and specificity (**Supplementary Figure S5c** and **Supplementary Table S5**).

b) The glycan composition test (N2 and O1). The *N*-glycan composition score was calculated based on the Pearson correlation (R^2) between the expected distribution of *N*-glycans carried by human serum glycoproteins as reported by Clerc et al.[14] and the observed *N*-glycan distribution reported by each team. The *O*-glycan composition score was calculated based on the Pearson correlation (R^2) between the expected distribution of *O*-glycans carried by human serum glycoproteins as reported by Yabu et al.[15] and the observed *O*-glycan distribution reported by each team. The distribution of the *N*- and *O*-glycan compositions was calculated based on the glyco-PSM count of each unique glycan ID relative to the total glyco-PSM count reported by each team.

c) The source glycoprotein test (N3 and O2). The source glycoprotein score was determined from the accuracy (specificity) and coverage (sensitivity) of the reported source glycoproteins relative to the glycoproteins expected in human serum. Reported *N*-glycoproteins previously identified in human serum by both Clerc et al.[14] and Sun et al.[16]. received a score of 2, whereas *N*-glycoproteins only identified by Sun et al. received a score of 1. Source glycoproteins not identified by the two references received no score. Further, reported *O*-glycoproteins previously identified in human serum by Darula et al.[17], Yang et al.[18] and Ye et al.[19] received a score of 3, 2 or 1 according to the number of papers identifying the specific *O*-glycoprotein. The source glycoproteins not reported by any of the references received no score. For both the serum *N*- and *O*-glycoproteins, the number of glyco-PSMs reported by each team were multiplied by the respective source glycoprotein score for each unique glycoprotein ID. The sensitivity of the test was calculated based on the summed glycoprotein score divided by the highest possible total score (number of unique glycoproteins reported by each

team multiplied by the highest theoretical glycoprotein score). The specificity of the test was calculated based on the summed glycoprotein score divided by the number of unique source glycoproteins reported in the selected literature.

d) The glycoproteome coverage test (N4 and O3): The *N*- and *O*-glycoproteome coverage were calculated based on the number of unique glycopeptide IDs (unique peptide sequence and glycan composition) reported by each team.

e) The commonly reported ('consensus') glycopeptide test (N5 and O4): The consensus *N*-glycopeptide score was calculated based on the number of glycopeptide ID commonly reported by 50% of the 22 teams returning *N*-glycopeptide data. The consensus *O*-glycopeptide score was calculated based on the number of glycopeptide ID commonly reported by 30% of the 20 teams returning *O*-glycopeptide data.

f) The NeuGc and multi-Fuc glycopeptide test (N6 and O6). The number of reported *N*- and *O*-glyco-PSMs corresponding to NeuGc and multi-Fuc (Fuc $\geq$ 2) containing glycopeptides was normalised to the total glyco-PSMs reported by each team. Separate *N*- and *O*-glycopeptide scores were then calculated based on the average of non-NeuGc and non-Fuc $\geq$ 2 containing glyco-PSMs for teams that included NeuGc and Fuc $\geq$ 2 containing glycan compositions in their glycan search space.

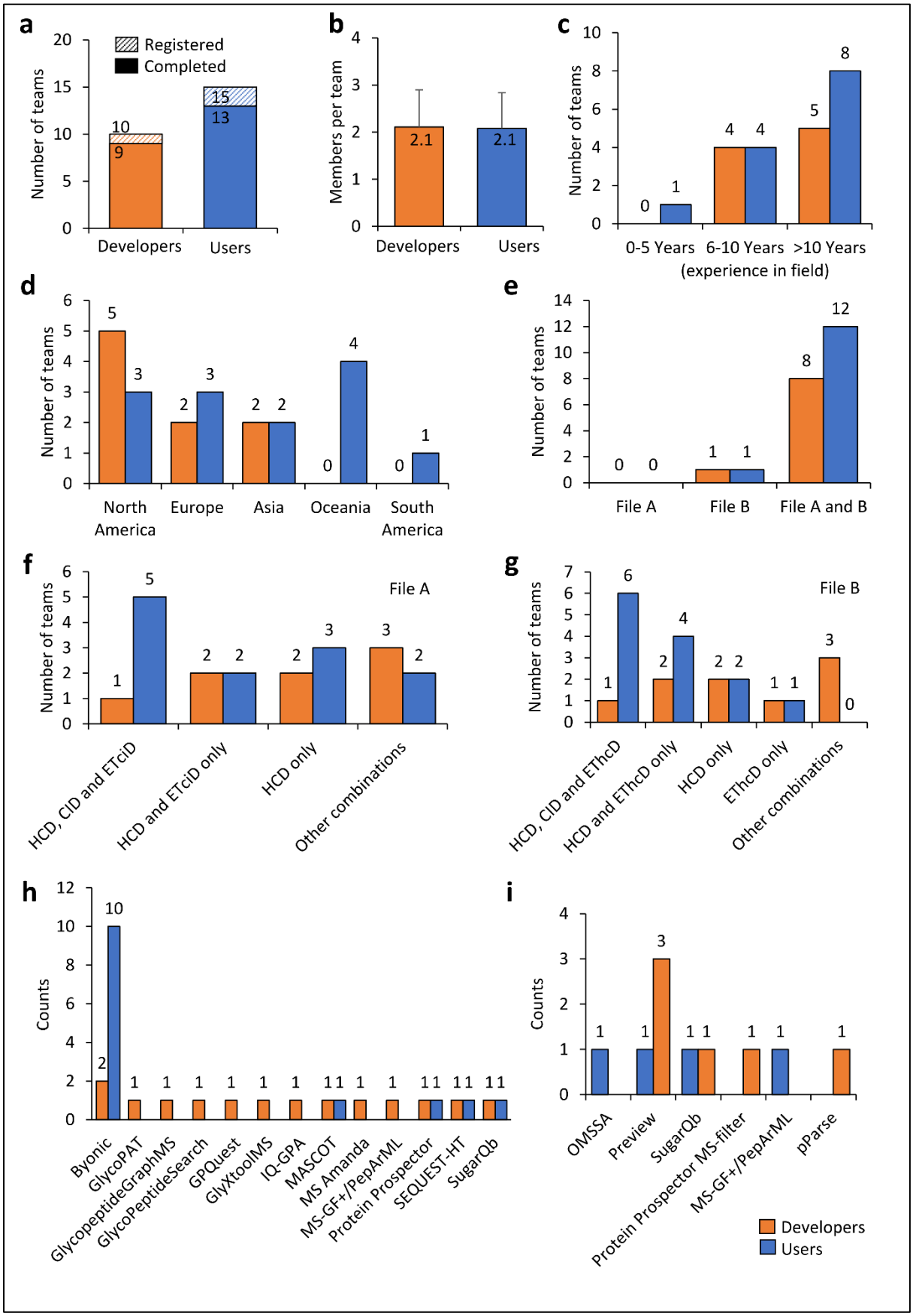
The overall performance scores for *N*- and *O*-glycopeptide analysis were established separately by averaging the scores of the individual performance tests (N1-N6 and O1-O5, respectively).

Statistical analysis

The normalised performance scores from each performance test were compiled with the search parameters and search outputs (average of selected variables). Seven data analysis methods were applied to identify search settings and search output characteristics that were associated with high performance scores including 1) a multiple linear regression model applied with a significance threshold of $p < 0.05$ to identify association between search variables (predictors)

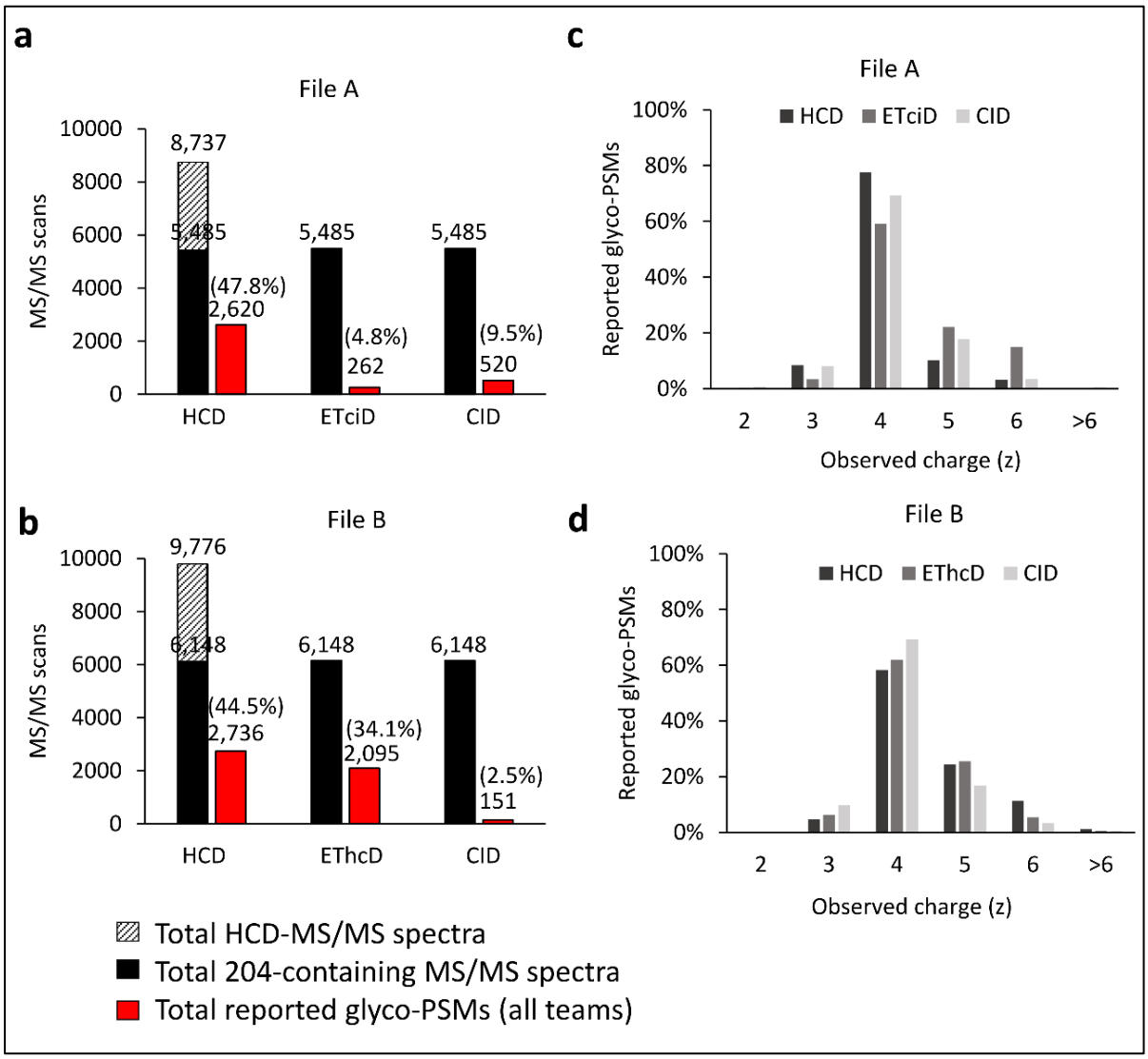
and performances scores (response variable), 2) a ridge linear regression model applied using an induced smoothing paradigm for hypothesis testing[20, 21], 3) a Lasso linear model for variable selection[22], 4) a least angle regression exploiting exact post-selection inference to identify associations[23, 24], 5) a forward stepwise linear regression applied using selective inference to identify association[25], 6) a Random Forest algorithm (an ensemble learning model for regression) applied using a variable of importance score to identify association[26] (a permutation strategy on augmented set of noise variable variables was exploited to define the variable importance cut-off), and 7) a gradient boosting tree algorithm (an ensemble of decision trees for prediction) applied using a similar strategy as the Random Forest algorithm to select important associations[27, 28]. Only associations commonly observed across a minimum of three different statistical methods were considered in this study.

Supplementary Figure S1



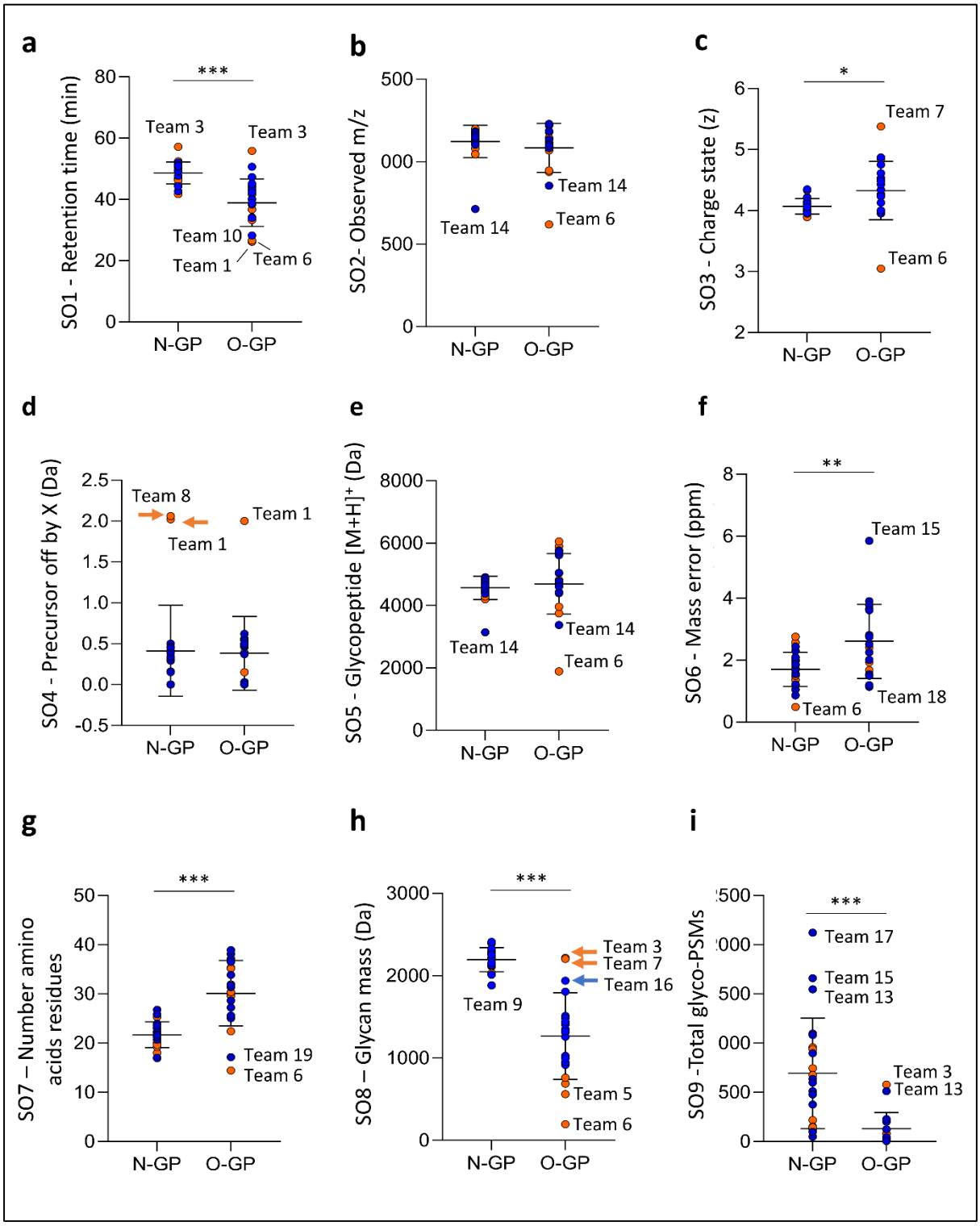
Supplementary Figure S1. Overview of the participating teams and their search strategies grouped according to their status as either developers (orange) or expert users (blue) of glycoproteomics software. **a.** Number and type of teams that registered for and completed the study. Note that a few registered teams did not complete the study; individuals within these non-completing teams and their data (if any) were not included in the study. **b.** Average number of members in each of the completing teams. Data is represented by mean ± SD (n = 9, developers and n = 13, expert users). **c.** The self-reported experience in glycoproteomics of each team. **d.** Team origin by continent. **e.** Data files (File A and/or B) handled by the teams. **f**-**g.** Type of fragmentation spectra used by teams to identify glycopeptides. **h**. Search engine(s) and **i.** pre- and post-processing tools used for the glycopeptide identification.

Supplementary Figure S2



Supplementary Figure S2. Overview of the MS/MS data and charge state distribution of the reported glycopeptides. **a-b.** The total number of all recorded HCD-MS/MS scans within File A and B (striped bars), the total number of *m/z* 204-containing MS/MS scans (glycopeptide MS/MS spectra, black bars) and the total number of glyco-PSMs collectively reported from all teams (red bars) over the different fragmentation methods. **c-d.** Charge state distribution of the reported glyco-PSMs from File A and B (data are plotted as the mean calculated from all teams).

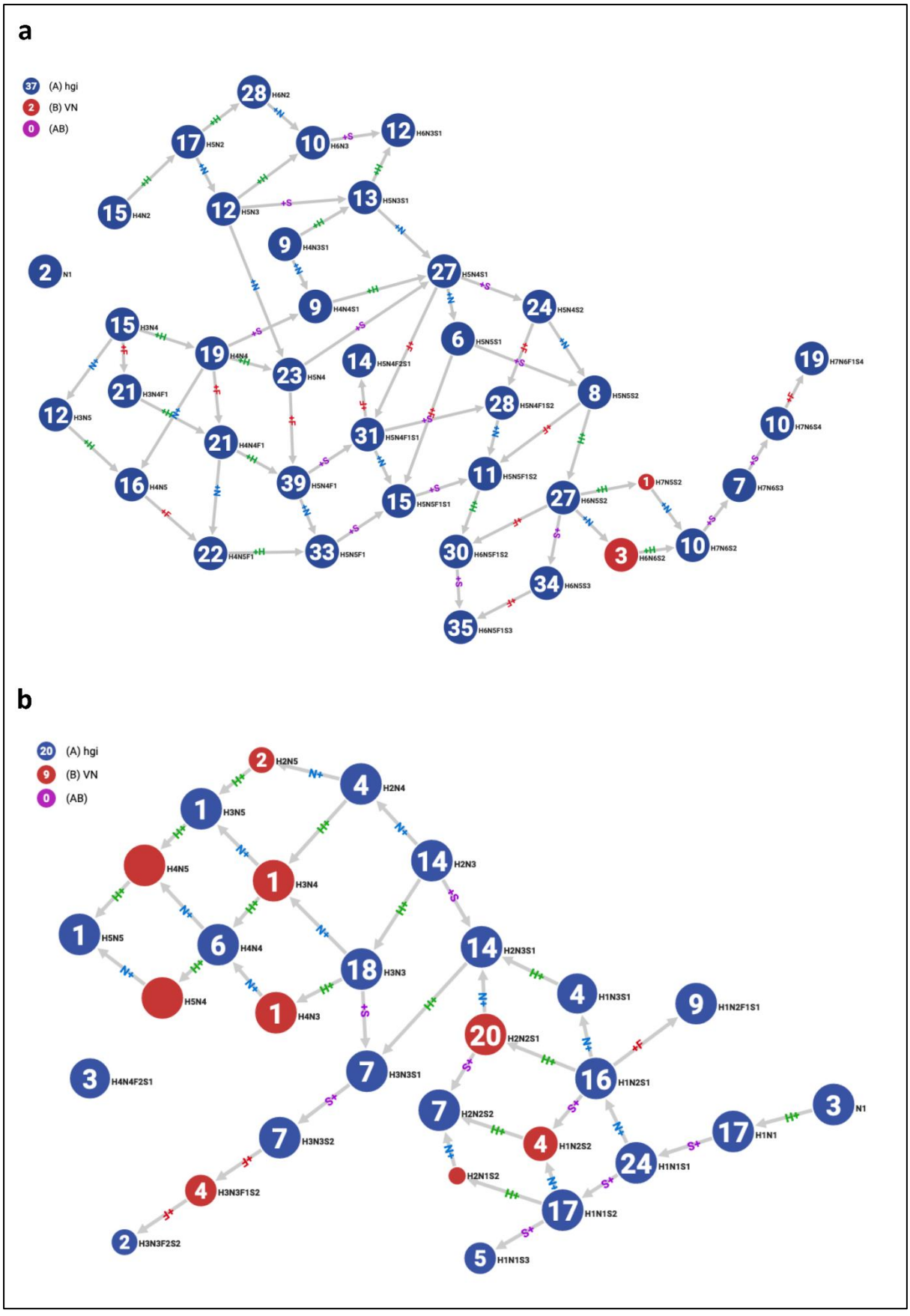
Supplementary Figure S3



Supplementary Figure S3. Team-centric overview of the search output data from the glycopeptide identification process (SO1-SO9). Distribution of the **a.** LC retention time (min), **b.** observed glycopeptide *m/z*, **c.** observed charge state (z), **d.** observed precursor selection off

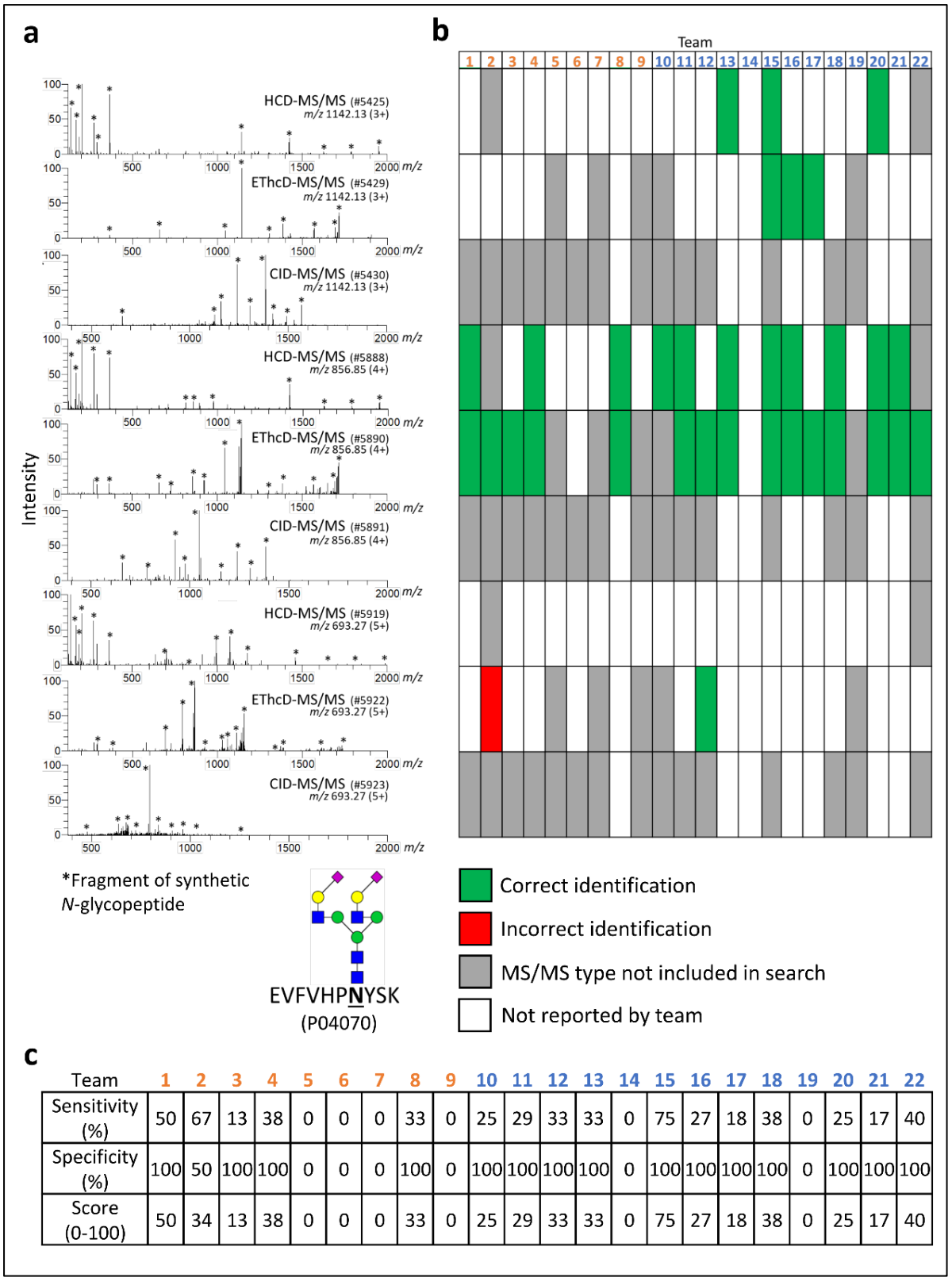
by X (Da, positive values only), (**e**) observed glycopeptide mass $[M+H]^+$ (Da), **f.** actual mass error of observed glycopeptides (ppm, positive values only), **g.** length of observed glycopeptides, **h.** glycan mass of observed glycopeptides (M, Da), **i.** total *N*- and *O*-glyco-PSMs reported by the participants. The mean and SDs of data from all teams are also indicated for each graph. Developer data are plotted in orange and expert user data are plotted in blue. Teams reporting data outside the SDs have been labelled. The *N*-glycopeptide (N-GP) data were statistically compared to the *O*-glycopeptide (O-GP) data using t-tests where $*p < 0.05$, $**p < 0.01$ and $***p < 0.001$.

Supplementary Figure S4



Supplementary Figure S4. Biosynthesis-centric network analysis of the *N*- and *O*-glycan compositions carried by the **a.** 163 consensus *N*-glycopeptides and **b.** 23 consensus *N*-glycopeptides using Glyconnect Compozitor v1.0.0. Each node corresponds to a glycan composition either reported within in the consensus list of glycopeptides arising from this study (blue circles) or manually added to biosynthetically connect the glycan compositions by a single glycan processing step (red circles). Both networks showed close biosynthetic relationship between the consensus *N*- and *O*-glycan structures reported in this study supporting the correctness of their identification.

Supplementary Figure S5

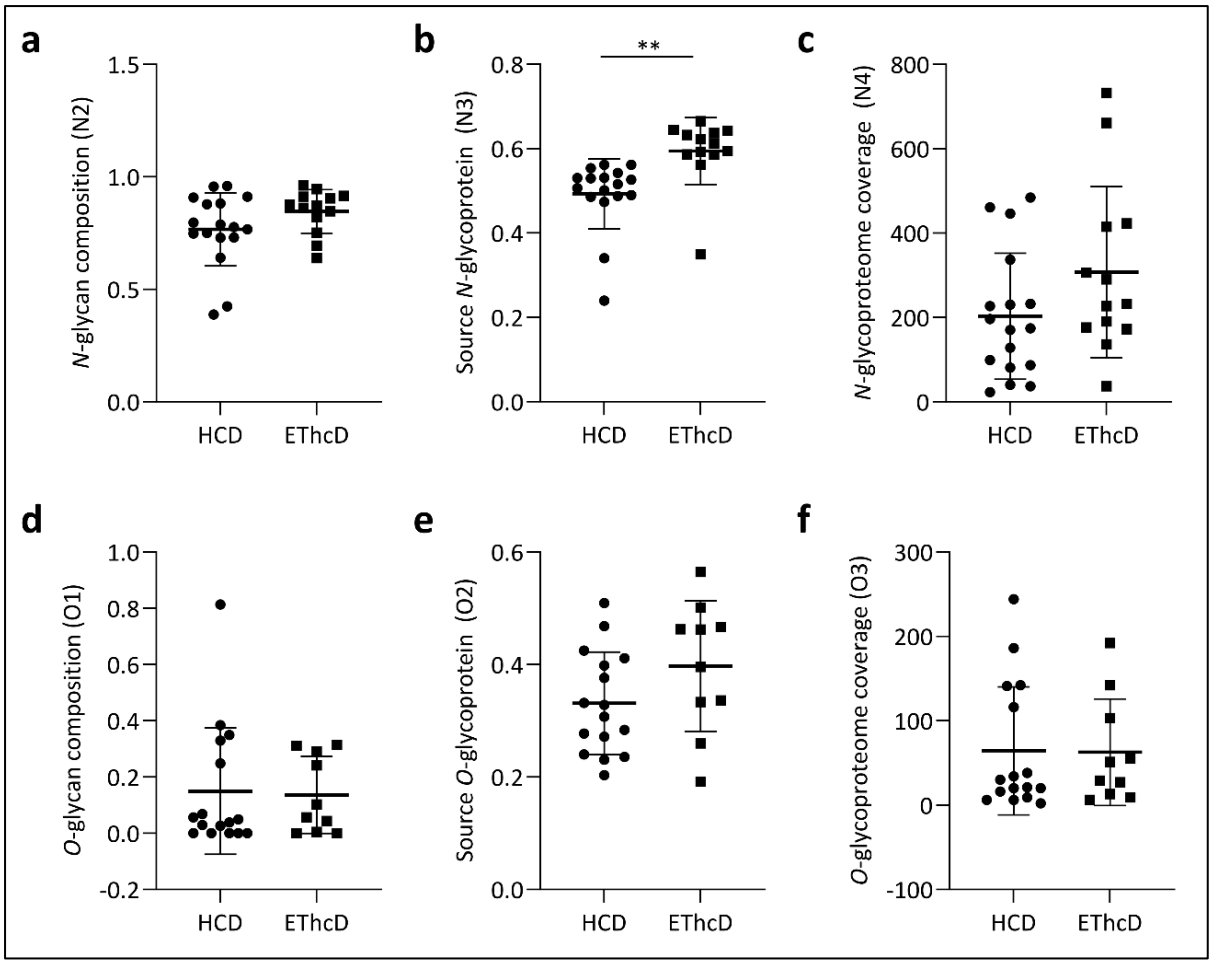


Team	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Sensitivity (%)	50	67	13	38	0	0	0	33	0	25	29	33	33	0	75	27	18	38	0	25	17	40
Specificity (%)	100	50	100	100	0	0	0	100	0	100	100	100	100	0	100	100	100	100	0	100	100	100
Score (0-100)	50	34	13	38	0	0	0	33	0	25	29	33	33	0	75	27	18	38	0	25	17	40

Supplementary Figure S5. Data underpinning the synthetic glycopeptide performance test (N1). **a.** MS/MS spectra corresponding to the non-adducted synthetic *N*-glycopeptide

(EVFVHPNYSK, $Hex_5HexNAc_4NeuAc_2$, UniProtKB, P04070, see insert for schematics) in charge state 3+ and 4+ (six top spectra) and the K^+-adducted synthetic *N*-glycopeptide in charge state 5+ (three bottom spectra) arising from the three fragmentation modes (HCD-, EThcD- and CID-MS/MS) used to generate File B. **b.** Overview of the 9 MS/MS spectra of the synthetic *N*-glycopeptide (from panel a) that were either correctly identified (green), incorrectly identified (red), or not reported by each team (white). Spectra arising from fragmentation mode(s) not included in the search strategy chosen by each team were not included in the assessment (indicated in grey). **c.** Performance scores arising from the test determined for each team based on the sensitivity and specificity of the identification of the 9 MS/MS spectra corresponding to the synthetic *N*-glycopeptide.

Supplementary Figure S6



Supplementary Figure S6. Comparison of the raw (non-normalised) performance scores arising from the glycopeptide identifications based on HCD- or EThcD-MS/MS data. Only glycopeptides unambiguously reported by either HCD- or EThcD-MS/MS data were included in this analysis. **a.** *N*-glycan composition (N2), **b.** source *N*-glycoprotein (N3), and **c.** *N*-glycoproteome coverage (N4) were calculated using HCD-MS/MS glyco-PSMs reported by 17 teams and EThcD-MS/MS glyco-PSMs reported by 13 teams. Significance was tested using unpaired student's t-test, ** $p < 0.01$). **d.** *O*-glycan composition (O2), **e.** source *O*-glycoprotein (O2) and **f.** *O*-glycoproteome coverage (O3) were calculated using HCD-MS/MS glyco-PSMs reported by 16 teams and EThcD-MS/MS glyco-PSMs reported by 10 teams.

References used in the supplementary information

1. Stavenhagen, K. et al. Quantitative mapping of glycoprotein micro-heterogeneity and macro-heterogeneity: an evaluation of mass spectrometry signal strengths using synthetic peptides and glycopeptides. *J Mass Spectrom* **48**, 627-639 (2013).
2. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* **47**, D442-D450 (2019).
3. Park, G.W. et al. Integrated GlycoProteome Analyzer (I-GPA) for Automated Identification and Quantitation of Site-Specific N-Glycosylation. *Sci Rep* **6**, 21175 (2016).
4. Baker, P.R., Trinidad, J.C. & Chalkley, R.J. Modification site localization scoring integrated into a search engine. *Mol Cell Proteomics* **10**, M111 008078 (2011).
5. Pioch, M., Hoffmann, M., Pralow, A., Reichl, U. & Rapp, E. glyXtool(MS): An Open-Source Pipeline for Semiautomated Analysis of Glycopeptide Mass Spectrometry Data. *Anal Chem* **90**, 11908-11916 (2018).
6. Bern, M., Kil, Y.J. & Becker, C. Byonic: advanced peptide and protein identification software. *Curr Protoc Bioinformatics* **Chapter 13**, Unit13 20 (2012).
7. Stadlmann, J., Hoi, D.M., Taubenschmid, J., Mechtler, K. & Penninger, J.M. Analysis of PNGase F-Resistant N-Glycopeptides Using SugarQb for Proteome Discoverer 2.1 Reveals Cryptic Substrate Specificities. *Proteomics* **18**, e1700436 (2018).
8. Pompach, P., Chandler, K.B., Lan, R., Edwards, N. & Goldman, R. Semi-automated identification of N-Glycopeptides by hydrophilic interaction chromatography, nano-reverse-phase LC-MS/MS, and glycan database search. *J Proteome Res* **11**, 1728-1740 (2012).

9. Choo, M.S., Wan, C., Rudd, P.M. & Nguyen-Khuong, T. GlycopeptideGraphMS: Improved Glycopeptide Detection and Identification by Exploiting Graph Theoretical Patterns in Mass and Retention Time. *Anal Chem* **91**, 7236-7244 (2019).

10. Liu, G. et al. A Comprehensive, Open-source Platform for Mass Spectrometry-based Glycoproteomics Data Analysis. *Mol Cell Proteomics* **16**, 2032-2047 (2017).

11. Toghi Eshghi, S., Shah, P., Yang, W., Li, X. & Zhang, H. GPQuest: A Spectral Library Matching Algorithm for Site-Specific Assignment of Tandem Mass Spectra to Intact N-glycopeptides. *Anal Chem* **87**, 5181-5188 (2015).

12. Bollineni, R.C., Koehler, C.J., Gislefoss, R.E., Anonsen, J.H. & Thiede, B. Large-scale intact glycopeptide identification by Mascot database search. *Sci Rep* **8**, 2117 (2018).

13. Dorfer, V. et al. MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *J Proteome Res* **13**, 3679-3684 (2014).

14. Clerc, F. et al. Human plasma protein N-glycosylation. *Glycoconj J* **33**, 309-343 (2016).

15. Yabu, M., Korekane, H. & Miyamoto, Y. Precise structural analysis of O-linked oligosaccharides in human serum. *Glycobiology* **24**, 542-553 (2014).

16. Sun, S. et al. Site-Specific Profiling of Serum Glycoproteins Using N-Linked Glycan and Glycosite Analysis Revealing Atypical N-Glycosylation Sites on Albumin and alpha-1B-Glycoprotein. *Anal Chem* **90**, 6292-6299 (2018).

17. Darula, Z., Sarnyai, F. & Medzihradszky, K.F. O-glycosylation sites identified from mucin core-1 type glycopeptides from human serum. *Glycoconj J* **33**, 435-445 (2016).

18. Yang, W., Ao, M., Hu, Y., Li, Q.K. & Zhang, H. Mapping the O-glycoproteome using site-specific extraction of O-linked glycopeptides (EXoO). *Mol Syst Biol* **14**, e8486 (2018).

19. Ye, Z., Mao, Y., Clausen, H. & Vakhrushev, S.Y. Glyco-DIA: a method for quantitative O-glycoproteomics with in silico-boosted glycopeptide libraries. *Nat Methods* **16**, 902-910 (2019).

20. Sottile, G., Cilluffo, G. & Muggeo, V.M.R. (2019). The R package islasso: estimation and hypothesis testing in lasso regression.

21. Cilluffo, G., Sottile, G., La Grutta, S. & Muggeo, V.M. The Induced Smoothed lasso: A practical framework for hypothesis testing in high dimensional regression. *Stat Methods Med Res* **29**, 765-777 (2020).

22. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33**, 1-22 (2010).

23. Hastie, T. & Efron, B. (2013). Lars: Least Angle Regression, Lasso and Forward Stagewise. R package version 1.2.

24. Tibshirani, R.J., Jonathan, T., Lockhart, R. & Tibshirani, R. (2014). Exact Post-Selection Inference for Sequential Regression Procedures.

25. Tibshirani, R. et al. (2019). SelectiveInference: Tools for Post-Selection Inference. R package version 1.2.5.

26. Breiman, L. Bagging Predictors. *Machine Learning* **24**, 123–140 (1996).

27. Efron, B. & Hastie, T. Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. (Cambridge University Press, 2016).

28. Greenwell, B., Boehmke, B. & Cunningham, J. Generalized Boosted Regression Models. (R package version 2.1.8; 2020).