Report from Dagstuhl Seminar 25351

Computational Proteomics

Rebekah Gundry*[1], Magnus Palmblad*[2], and Mathias Wilhelm*[3]

1 **University of Nebraska – Omaha, US.** rebekah.gundry@unmc.edu
2 **Leiden University Medical Center, NL.** n.m.palmblad@lumc.nl
3 **TU München – Freising, DE.** mathias.wilhelm@tum.de

Abstract

In 2025 the Dagstuhl Seminar "Computational Proteomics" (25351), part of a series of Dagstuhl Seminars with the same name, brought together experts from proteomics, glycomics and machine learning to address key challenges in the field. Discussions emphasized the need for scalable and interoperable data infrastructures, a new initiative to generate large, AI-ready proteomics datasets, and community standards for reproducible and interpretable machine learning and harmonized glycomics workflows. Participants identified several barriers in clinical translation, multi-omics integration, and quantitative glyco-proteomics, highlighting limited data interoperability, heterogeneous experimental designs, and insufficient statistical and reporting frameworks. The seminar concluded with concrete action plans toward new standards, best practices, and collaborative initiatives to advance reproducible, sustainable and clinically relevant proteomics.

Seminar August 24–29, 2025 – https://www.dagstuhl.de/25351
2012 ACM Subject Classification Theory of computation → Theory and algorithms for application domains; Computing methodologies → Machine learning; Applied computing → Life and medical sciences
Keywords and phrases proteomics, glycomics, glycoproteomics, machine learning, mass spectrometry
Digital Object Identifier 10.4230/DagRep.15.8.46

1 Executive Summary

Rebekah Gundry (University of Nebraska – Omaha, US)
Magnus Palmblad (Leiden University Medical Center, NL)
Mathias Wilhelm (TU München – Freising, DE)

In 2025 the Dagstuhl Seminar "Computational Proteomics" (25351), part of a series of Dagstuhl Seminars with the same name, brought together researchers in proteomics, glycomics, computational biology, translational biomarker research, mass spectrometry, statistics, and machine learning for a week of intense discussions and collaboration. Building on the Dagstuhl Seminar "Computational Proteomics" (23301) in 2023, we extended the agenda in four directions that reflect where our field is now pushing hardest:

Translational Proteomics

The translational proteomics group defined translational proteomics as a continuum from discovery to clinical implementation, spanning basic model systems (cell lines, mouse), human biospecimens, clinical decision support, and ultimately population health. The group

* Editor / Organizer

emphasized that translation is not just "applying proteomics in the clinic," but instead structuring the entire value chain: standardized sample handling, acquisition, annotation, processing, interpretation, and delivery of actionable outputs (e.g. patient stratification, tumor board support). Major barriers identified include a lack of interoperable and well-annotated datasets, underpowered cohorts (especially in rare diseases), weak incentives for repetitive but clinically necessary assays, and difficulty converting molecular readouts into clinical recommendations. The group proposed ENIGMA, a staged, global-scale effort to generate and harmonize >100,000 proteomics datasets, starting in controlled mouse models and extending to human samples, as an AI-ready foundation for translation.

Machine Learning (in Proteomics and Glycomics)

The machine learning (ML) group concluded that the current culture of "incremental performance improvements" is unsustainable and often scientifically marginal. Instead, the group argued for community standards around software quality, reproducibility, interpretability, and dataset (ML/AI) readiness. Discussions focused on updating and extending existing recommendations (e.g. DOME, FAIR4RS) to address maintainability, testability and bias, and on defining what actually constitutes a publishable ML contribution in proteomics or glycomics. The group also highlighted the need for well-annotated, uncertainty-aware training and benchmarking datasets, including glycopeptide data, and began drafting two manuscripts – one on software quality and reporting expectations for ML in proteomics, and one on explainable and interpretable AI in MS-based proteomics.

Glycomics and Glycoproteomics

The "glyco" group focused on two tightly linked goals: improving confidence and comparability in glycan/glycopeptide identification and quantification, and lowering the barrier of entry for new researchers. First, the group outlined a plan for harmonizing glycan search spaces and reporting. A key recommendation is that outputs should carry standardized GlyTouCan identifiers and clearly encode the level of structural specificity (composition-only, topology, full linkage) so that results from different software tools can be compared on a common specificity level. The group also emphasized the need for explicit false-discovery rate (FDR) frameworks for glycan assignments, including topology- and isomer-sensitive scoring. Second, the working group substantially advanced two manuscripts: a best-practices/tutorial document for new glycosylation researchers (terminology, pitfalls, reporting standards), and a focused manuscript on how glycan structure affects glycopeptide signal intensity and the downstream challenges for quantification and biological interpretation. Writing responsibilities, timelines, and revision plans were agreed, and a first integrated draft was produced on-site.

Cross-cutting Topics: Federated Learning, Data Sharing and Credit, Multi-Omics Integration, and Quantitative Glyco-Proteomics

In the second half of the seminar week, people rotated between groups to discuss a number of cross-cutting topics and common challenges: Federated learning and controlled-access clinical data: While federated learning is still rarely used in proteomics due to widespread centralized data deposition, this is expected to change as clinical data are increasingly held locally for regulatory and privacy reasons. The group concluded that now is the time to define incentives, governance, and credit mechanisms for data generators so that high-value but unpublished datasets can still drive model development without leaving institutional boundaries. Multi-omics integration: Participants discussed how to integrate proteomics with transcriptomics,

phosphoproteomics, glycoproteomics, metabolomics/lipidomics, immunopeptidomics, and mass spectrometry imaging (and other spatially resolved data). The consensus was that most current "integration" is actually late-stage comparison of separate analyses. True multi-omics fusion is blocked by non-synchronous sampling, heterogeneous sample preparation, lack of common quality control (standards), differing biological timescales, and underdeveloped statistical control across layers. Guidance is needed on realistic experimental design and on levels of integration (early, mid-level latent, late/pathway). Quantitative glyco-proteomics statistics: The group outlined a plan to benchmark methods for glycoform quantification, normalization, missing value handling, and site occupancy estimation, leveraging tools such as MSstats/MSstats-PTM and experimentally perturbed datasets.

The 2025 Dagstuhl Seminar crystallized a shift in the field: from chasing marginal analytical improvements toward building scalable, interpretable, clinically relevant, and socially sustainable proteomics. The seminar concluded with concrete action items: manuscripts on ML standards and explainability; a best-practices/tutorial manuscript and a quantitative glycoform analysis manuscript from the glyco group; an ENIGMA proposal for large-scale, AI-ready translational proteomics data; a position statement on data sharing, incentive structures, and federated learning; and guidance on realistic multi-omics integration and QC. Together, these efforts define a forward-looking agenda for computational and translational proteomics in the coming years.

The discussions on trustworthiness and quality control in proteomics also fed directly into the planning of a Lorentz Center Workshop on "Trustworthiness in Proteomics", successfully co-organized by Mathias Wilhelm and Magnus Palmblad in Leiden, in February 2026.

2 Table of Contents

3 Working groups

3.1 Working Group Report: Translational Proteomics

Isabell Bludau (Unviversitätsklinikum Heidelberg, DE), Tine Claeys (Ghent University, BE), Stephanie Cologna (University of Illinois – Chicago, US), Melanie Föll (Universitätsklinikum Freiburg, DE), Wassim Gabriel (TU München – Freising, DE), Paula González Menéndez (University of Adelaide, AU), Zhiwei Liu (Westlake University – Hangzhou, CN), Tobias Schmidt (MSAID – Garching, DE), Nicola Ternette (University of Dundee, GB), Hans Wessels (Radboud University Nijmegen, NL), Mathias Wilhelm (TU München – Freising, DE), and Bernd Wollscheid (ETH Zürich, CH)

3.1.1 Scope and Motivation

The translational proteomics working group defined translational proteomics as a stepwise process linking basic proteomics research to clinical and population-level impact. This includes:

- Experiments in model systems (cell lines, mouse models).
- Studies on human biospecimens (tissues, fluids).
- Clinical application (diagnostics, patient stratification, therapy guidance).
- Integration into routine decision-making (e.g. tumor boards).
- Extension to population health and precision medicine at scale.

The group agreed that translational proteomics is not only about "measuring patient samples," but about structuring and standardizing the entire value chain: sample acquisition, metadata capture, MS acquisition, data processing, integration with other omics, interpretation, and delivery of clinically actionable output.

3.1.2 Key Barriers

Data quality and interoperability: Most proteomics datasets are not captured with standardized metadata, ontologies, or controlled vocabularies suitable for clinical use. Heterogeneous assay formats, missing metadata, and inconsistent annotation impede reuse and prevent robust comparisons across sites.

Underpowered cohorts: Many studies operate at N too small for reliable biomarker development, particularly in rare diseases. Cohort fragmentation across institutions and7 slow/complex material transfer agreements impede the assembly of sufficiently powered datasets.

Workflow robustness and reproducibility: Clinical deployment requires durable, validated, auditable workflows. Current research pipelines are often bespoke, sensitive to batch effects, and not easily transferable to regulated environments.

Motivation and incentives: Clinically relevant work often requires repetitive, "boring" targeted assays and rigorous QC. Such work is rarely rewarded academically and is often unfunded. As a result, essential confirmatory and longitudinal measurements are underperformed.

Interpretability and clinical usability: Even when high-quality measurements exist (e.g. phosphoproteomics, glycoproteomics, immunopeptidomics), delivering a clear, defensible recommendation to a tumor board remains difficult. Clinicians want interpretable, validated

evidence (what pathway is active? which drug class is likely to work?), not just feature lists. Regulatory and legal constraints: Clinical proteomics data raise privacy, compliance, and regulatory concerns analogous to those in genomics. Institutions are cautious about releasing raw data, which complicates model development and validation.

3.1.3 Computational / Infrastructural Priorities

Standardized ontologies and metadata schemas (including disease context, acquisition parameters, and sample processing information) are required to make datasets reusable across institutions. Scalable, transparent pipelines are needed to analyze imperfect, real-world study designs (non-ideal controls, heterogeneous sampling times, varying platforms). Multi-omics integration must move beyond naive overlap of "differential features" and instead support mechanistic interpretation and phenotype prediction. Cloud-based or federated frameworks may be required to enable collaborative analysis when data cannot leave clinical boundaries. Quantitative modeling should address proteoforms, phosphorylation, glycosylation, immunopeptidomics, etc., rather than collapsing information to a single "protein ID."

3.1.4 ENIGMA Initiative

To address the chronic shortage of large, harmonized, clinically relevant datasets, the group proposed ENIGMA: an international, multi-site initiative to create an AI-ready proteomics resource on the order of >100,000 samples.
The initiative is structured in phases:
Phase 1 (Pilot): A few thousand well-controlled samples (initially including defined mouse lines and select human material) are processed to establish baseline protocols, metadata standards, benchmarking pipelines, and QC criteria.
Phase 2 (Scale-up): Tens of thousands of samples are acquired across multiple mass spectrometry laboratories worldwide under harmonized DIA-style acquisition. Heterogeneity (instrument type, sample preparation differences) is explicitly captured rather than avoided, to allow development of normalization, batch correction, and cross-lab integration strategies.
Phase 3 (Translation): The trained models and pipelines are applied to human cohorts, including rare disease and clinically complex cases, with the goal of informing stratification, therapeutic targeting, and mechanism-of-action hypotheses.
Critical design features:

- Use of leftover tissues and archived material to reduce ethical/consent burden and accelerate scale.
- Systematic metadata capture (including tissue, condition, preparation, instrument, acquisition settings), with machine learning support for harmonization and outlier detection.
- Community governance around data ownership, licensing, authorship, and credit.
- Defined benchmark subsets for public release, similar in spirit to CPTAC and AlphaFold-style "challenge" datasets.

3.1.5 Outcomes and Next Steps

The working group concluded that translational proteomics requires a coordinated village: clinicians to articulate clinical questions; experimentalists to generate high-quality material; computational groups to build robust, interpretable pipelines; and funding/industry partners to sustain longitudinal measurement. The ENIGMA plan will serve both as a rallying point for funding proposals and as an anchor for future community standards in translational proteomics.

3.2 Working Group Report: Cross-cutting Topics

Frédérique Lisacek (Swiss Institute of Bioinformatics – Geneva, CH), Charlotte Adams (University of Antwerp, BE), Kiyoko Aoki-Kinoshita (Soka University – Tokyo, JP), Gad Armony (Bruker Nederland – Leiderdorp, NL), Wout Bittremieux (University of Antwerp, BE), Isabell Bludau (Unviversitätsklinikum Heidelberg, DE), Robert Chalkley (University of California – San Francisco, US), Tine Claeys (Ghent University, BE), Stephanie Cologna (University of Illinois – Chicago, US), Eric Deutsch (Institute for Systems Biology – Seattle, US), Patrick Emery (Matrix Science Ltd. – London, GB), Melanie Föll (Universitätsklinikum Freiburg, DE), Wassim Gabriel (TU München – Freising, DE), Paula González Menéndez (University of Adelaide, AU), Rebekah Gundry (University of Nebraska – Omaha, US), Devon Kohler (Northeastern University – Boston, US), Lev Levitskiy (University of Southern Denmark – Odense, DK), Klaus Lindpaintner (Bruker – Concord, US), Zhiwei Liu (Westlake University – Hangzhou, CN), Sriram Neelamegham (University at Buffalo – SUNY, US), Magnus Palmblad (Leiden University Medical Center, NL), Daniel Polasky (University of Michigan – Ann Arbor, US), Rene Ranzinger (University of Georgia, US), Tobias Schmidt (MSAID – Garching, DE), Nicola Ternette (University of Dundee, GB), Sergey Vakhrushev (University of Copenhagen, DK), Hans Wessels (Radboud University Nijmegen, NL), Mathias Wilhelm (TU München – Freising, DE), and Dirk Winkelhardt (Ruhr-Universität-Bochum, DE)

The Thursday cross-over sessions brought together participants from translational proteomics, ML, glycomics, clinical proteomics, and multi-omics integration to address challenges that cut across all domains.

3.2.1 Federated Learning, Privacy, and Incentives for Data Sharing

Federated learning (FL) is widely discussed in biomedical AI but is not yet common in proteomics practice. The group identified why: For most discovery proteomics, data are still deposited centrally (e.g. PRIDE, MassIVE). Central pooling is simpler than setting up FL. FL infrastructure is non-trivial: institutions must maintain local compute, synchronize software/hardware, address batch effects and fairness in distributed model updating, and handle versioning and auditing. Repositories themselves would bear significant cost to orchestrate FL at scale. However, participants agreed that this will change as proteomics becomes clinically embedded. Clinical and translational datasets (especially targeted assays, patient-derived longitudinal data, immunopeptidomics, and phospho-signaling panels) are increasingly held locally for regulatory and privacy reasons. In that emerging setting, FL – or at least controlled-access, privacy-aware model training – becomes essential.

The group emphasized that the blockers are not just technical but social:

- Privacy/compliance anxiety: When rules are unclear, many investigators choose not to share data at all.
- Fear of being scooped: High-value clinical datasets are expensive to generate, and groups are reluctant to expose them prior to publication without formal recognition.
- Fear of policing: Investigators worry about reputational risk if preliminary or messy data are scrutinized out of context.

Participants noted that vast quantities of clinically interesting proteomics data remain unpublished and effectively invisible. Releasing even partial access to these datasets (or enabling them to participate in FL-like analysis) would massively accelerate method development and translational validation.
Proposed actions:

- Draft a community position / opinion piece that (a) documents current barriers to sharing unpublished proteomics data, (b) argues for explicit dataset credit and citation, (c) proposes governance and recognition mechanisms for dataset contributors, and (d) frames FL as one of several approaches (not the only one) for responsibly leveraging sensitive data.
- Promote dataset-level DOIs, ORCID linkage, citation tracking, and usage metrics as first-class research outputs.
- Encourage journals and funders to recognize data notes / dataset publications and not penalize manuscripts that build on previously described datasets.

3.2.2 Multi-Omics Integration

The group critically assessed the state of "multi-omics integration" in proteomics-driven biology and translation. The conclusion was that true integration is still rare. Core challenges:

- Different omics layers (genomics, transcriptomics, proteomics, phosphoproteomics, glycoproteomics, metabolomics, lipidomics, immunopeptidomics, spatial MS imaging) are often collected on different material, at different time points, using incompatible extraction chemistries.
- Biological timescales differ: DNA is static, RNA is dynamic on hours, protein abundance and PTMs reflect turnover and signaling, metabolites respond on minutes. Naively correlating steady-state measurements across these layers can be misleading.
- Statistical control is underdeveloped. Integrating multiple high-dimensional datasets multiplies the number of hypotheses tested, but FDR control is often applied separately per layer, if at all.
- QC practices are inconsistent. Many studies still lack standardized spike-ins, technical controls, and batch assessment across all omics layers.

The group distinguished three levels of integration:

- Late integration: Interpret each omics layer independently, then compare or combine conclusions (e.g. pathway enrichment or overlapping differentially regulated features). This is the most common today.
- Mid-level / latent integration: Learn joint low-dimensional representations or network structures across layers (e.g. linking phosphoproteomics to transcriptomics through signaling pathway models).
- Early fusion: Combine raw or minimally processed quantitative data across layers into a single model. This is the most ambitious but also most fragile with respect to batch effects, missingness, and sampling asynchrony.

Use cases discussed:

- Proteogenomics to detect neoantigens, fusion proteins, and tumor-specific sequence variants for immunopeptidomics.
- Phosphoproteomics + total proteomics to distinguish signaling changes from protein abundance changes.
- Glycoproteomics to refine clinically relevant biomarkers (e.g. glycoform-specific PSA outperforms total PSA).
- Spatial MS (laser capture microdissection, MSI-guided microproteomics/metabolomics/glycomics) for tumor microenvironment profiling.

Actionable needs:

- Clear guidance on experimental design for clinically realistic studies, acknowledging that perfect co-sampling is often impossible in the hospital setting.
- Shared QC expectations across layers (technical controls, spike-ins).
- Agreement on how to claim "integration": studies should report whether they did late, mid-level, or early fusion, rather than labeling any multi-assay project as "multi-omics integration."
- Availability of paired, public benchmark datasets suitable for method development.

3.2.3 Quantitative Glyco-Proteomics Statistics

The cross-over discussions also addressed quantification and statistical analysis for glycomics and glycoproteomics, linking glyco specialists, statisticians, and ML practitioners.
Key points:

- Glycopeptide intensities are not directly comparable across glycoforms, because different glycans alter ionization and fragmentation efficiency.
- Missing values should not always be blindly imputed. If an entire glycoform is absent in one condition, treating that as "low abundance" can generate false positives.
- Normalization strategies must reflect biology. Global scaling can erase true global glycosylation shifts; site-focused normalization may be more appropriate.
- Site occupancy and glycoform distribution are biologically meaningful readouts but require modeling analogous to PTM-centric statistical frameworks.

Planned work:

- Evaluate MSstats/MSstats-PTM–style approaches where glycoforms are treated as modified analytes, using controlled datasets (e.g. exoglycosidase-treated samples, known congenital glycosylation defects) to benchmark differential analysis, normalization, and missing value handling.
- Define minimal reporting requirements for glyco-quantitative studies, to make them interpretable and reusable in translational pipelines.

3.3 Working Group Report: Glycomics and Glycoproteomics

Rene Ranzinger (University of Georgia, US), Kiyoko Aoki-Kinoshita (Soka University – Tokyo, JP), Gad Armony (Bruker Nederland – Leiderdorp, NL), Robert Chalkley (University of California – San Francisco, US), Wassim Gabriel (TU München – Freising, DE), Rebekah Gundry (University of Nebraska – Omaha, US), Klaus Lindpaintner (Bruker – Concord, US), Frédérique Lisacek (Swiss Institute of Bioinformatics – Geneva, CH), Sriram Neelamegham (University at Buffalo – SUNY, US), Daniel Polasky (University of Michigan – Ann Arbor, US), Sergey Vakhrushev (University of Copenhagen, DK), and Hans Wessels (Radboud University Nijmegen, NL)

3.3.1 Motivation

Glycomics and glycoproteomics remain computationally and analytically challenging due to structural complexity (branching, linkage, isomerism), diverse acquisition strategies, non-uniform software support, and historically limited standardization. The glyco working group

concentrated on two urgent needs: (i) harmonizing identification and reporting so that results are comparable across software and labs, and (ii) establishing reliable quantitative and statistical practices, especially for glycopeptide-level measurements.

3.3.2 Harmonizing Glycan/Glycopeptide Identification

A core outcome was agreement that glycan and glycopeptide search results must become machine-readable, comparable, and traceable across tools. The group recommends:

- Use of GlyTouCan IDs in all reported results, at minimum at the composition level, with increasing specificity (topology, linkage) where supported by evidence.
- Explicit declaration of structural specificity: Results should indicate whether the assignment is composition-only, topology (branching pattern without full linkage certainty), or fully linkage-resolved.
- Subsumption / hierarchical comparison: Since different tools resolve glycans to different specificity levels, there must be a way to "collapse" structures to a shared lower-specificity representation for comparison and benchmarking.
- Shared glycan reference sets: The group began assembling a curated "human adult reference glycan list," analogous to a reviewed FASTA for proteins. Each participating lab/software group will contribute the glycan lists they currently search (e.g. as used in Byonic, MSFragger-glyco, pGlyco, Protein Prospector, GlycanFinder, etc.). These will be merged, deduplicated, and annotated with GlyTouCan ID, species/context, linkage class (N- vs O-linked), biological source (plasma, tissue, cell line), and whether the glycan is free or peptide-conjugated.
- The aim is to generate a community-supported glycan search space that is biologically grounded, not "everything in GlyTouCan," and thus suitable for routine glycoproteomics searches and cross-tool benchmarking.

3.3.3 Confidence Scoring and FDR for Glycans/Glycopeptides

The group addressed statistical confidence, which currently lags behind peptide-centric proteomics:

- Glycan/glycopeptide identification often lacks robust FDR control, especially at the level of glycan topology or isomer resolution.
- Scoring schemes must consider diagnostic fragment ions, instrument- and method-specific fragmentation behavior (e.g. stepped HCD vs ETD/EThcD), and the ability to distinguish core vs antenna fucosylation, high-mannose vs complex glycans, etc.
- Structural-level FDR (e.g. topology-level "localization" of glycan features on the peptide) was viewed as analogous to PTM localization scoring in phosphoproteomics.

Participants presented ongoing work on topology-aware scoring frameworks that progressively reduce candidate lists using diagnostic ions and instrument-aware fragmentation rules, then apply multi-level FDR (PSM-level, composition-level, topology-level). The consensus was pragmatic: "some FDR is better than none," provided the confidence level is clearly stated. Work remains to prevent overconfidence in cases where only one plausible glycan remains in the database but the MS/MS evidence is insufficient to distinguish isomers.

3.3.4 Quantification, Normalization, and Biological Interpretation

A second major focus was quantification. Glycopeptide signals are strongly influenced by the attached glycan, so intensity is not a straightforward proxy for occupancy or abundance.

The group discussed:

- Handling missing values: imputing partially missing precursors may be acceptable, but imputing entire missing glycoforms is risky and can create false biological conclusions.
- Normalization: global median-centering or total-signal normalization can be misleading if overall glycosylation shifts biologically. Alternative strategies include normalizing within a site to the dominant glycoform, or modeling site occupancy relative to the unmodified protein level.
- Statistical modeling: Treating glycoforms as site-specific PTMs suggests that tools such as MSstats/MSstats-PTM could be adapted by treating each glycoform as the "modified species."

The group proposed benchmarking known perturbation datasets (e.g. exoglycosidase-treated samples, congenital disorders of glycosylation, cell-surface enrichment studies) to evaluate differential analysis, missing value handling, and normalization strategies. This work connects directly to translational aims (e.g. leveraging glycosylation patterns in diagnostics and tumor board discussions) and to ML aims (e.g. training glyco-aware spectral/RT predictors, or "glyco-Prosit" models).

3.3.5 Best Practices and Onboarding for New Researchers

The working group advanced two manuscript efforts that originated in a previous Dagstuhl meeting:

- Best-practices/tutorial manuscript
- Introduces newcomers to protein glycosylation analysis, including released glycans and intact glycopeptides.
- Defines key terminology (composition, topology, linkage-defined structure).
- Recommends that reported results always include GlyTouCan IDs and explicitly state confidence and specificity level.
- Summarizes common pitfalls in identification, quantification, and interpretation.
- Aligns with MIRAGE-style reporting expectations.

During the seminar, section leads were assigned, timelines were agreed upon, prior contributors were re-engaged, and a coordinated writing sprint produced 21 pages of draft text.

Glycoform quantification manuscript

- Examines how glycan structure affects glycopeptide fragmentation and intensity.
- Explains why naïve fold-change analysis across glycoforms can mislead biological interpretation.
- Outlines statistical approaches for site-specific glycoform analysis and occupancy estimation.

3.3.6 Summary

The glycomics/glycoproteomics group delivered concrete progress toward harmonized reporting, FDR-aware scoring, and quantitative interpretation, and established a clear writing and dissemination plan. The group also linked their agenda to ML (glyco-aware prediction models, benchmark datasets) and translational proteomics (inclusion of glycan biology in clinically oriented pipelines).

3.4 Working Group Report: Machine Learning in Proteomics

Tobias Schmidt (MSAID – Garching, DE), Charlotte Adams (University of Antwerp, BE), Wout Bittremieux (University of Antwerp, BE), Tine Claeys (Ghent University, BE), Eric Deutsch (Institute for Systems Biology – Seattle, US), Patrick Emery (Matrix Science Ltd. – London, GB), Wassim Gabriel (TU München – Freising, DE), Devon Kohler (Northeastern University – Boston, US), Lev Levitskiy (University of Southern Denmark – Odense, DK), Zhiwei Liu (Westlake University – Hangzhou, CN), Magnus Palmblad (Leiden University Medical Center, NL), and Dirk Winkelhardt (Ruhr-Universität-Bochum, DE)

3.4.1 Motivation and Problem Statement

The ML working group took a critical view: the field currently rewards incremental improvements (e.g. 2% gain in spectral prediction accuracy) without necessarily delivering interpretability, robustness, reproducibility, or biological/clinical insight. The group argued for a rebalancing of values toward sustainability and impact.

Four core themes structured the discussions:

(i) Software Quality, Standards, and Publication Expectations

Participants assessed the adequacy of existing frameworks such as the DOME recommendations for reporting ML in proteomics/metabolomics and FAIR4RS for FAIR research software. They concluded that these frameworks are necessary but incomplete. Key gaps:

- Maintainability and testability: ML code should be modular, documented, versioned, licensed, and accompanied by retraining instructions and tests. These expectations are rarely enforced.
- Adherence to community standards: ML tools must input and output established HUPO-PSI formats (mzML, mzIdentML, mzTab, SDRF, ProForma, etc.) where applicable, to ensure interoperability.
- Bias and applicability domain: Authors should explicitly analyze dataset composition (organism, tissue, modification class, charge states, instrument type), identify biases (e.g. human and HLA bias in immunopeptidomics, charge state bias, peptide length bias), and articulate where the model should and should not be used.
- Novelty and publication worthiness: DOME specifies how to describe an ML method, but not whether the method represents a meaningful advance. The group argued that journals and reviewers should demand either conceptual novelty, new interpretability, improved robustness/maintainability, or concrete biological/clinical utility – not just a small numeric gain.

Outcome: The group began drafting a manuscript that extends previous Dagstuhl-driven work (including "Interpretation of the DOME Recommendations for Machine Learning in Proteomics and Metabolomics" and recent guidance on FAIR/open-source research software in proteomics) by incorporating the EVERSE research software quality dimensions (modularity, reusability, analysability, modifiability, testability, sustainability). The goal is a reviewer/author checklist that goes beyond reporting and into quality, longevity, and responsible reuse.

(ii) Training Data, Metadata, Uncertainty, and Benchmarking

Robust ML depends on training data that is complete, well-annotated, and accompanied by uncertainty estimates. The group emphasized that current proteomics and glycoproteomics datasets often lack even basic metadata (sample type, organism, sex, preparation method, instrument settings), which blocks fair benchmarking and leads to hidden biases. Priorities:

- Centralized metadata capture: Multiple ongoing efforts (curated SDRF files, reprocessing pipelines, LLM-based annotation of manuscripts and raw mzML headers, large-scale reanalyses such as MassIVE, PRIDE, PeptideAtlas, and MassNet) should not remain siloed. A shared portal under an existing infrastructure (e.g. ProteomeXchange) could collect experimental design metadata, derived identifications/quantifications, QC metrics, and completeness scores.
- Gold standard datasets: For each ML task (fragmentation prediction, retention time prediction, identification, quantification, modification localization, glycopeptide assignment, etc.), the community should define benchmark datasets that include raw data, identifications, quantitative outputs, and experimental design metadata in standard formats.
- These benchmarks must reflect biological and technical diversity (different tissues, instruments, acquisition modes) rather than a single "easy" mixture.
- Uncertainty propagation: Both PSM-level confidence and metadata confidence should be carried through into training. Models should not be trained exclusively on idealized "perfect" identifications; probabilistic or Bayesian treatment of noisy/ambiguous cases is encouraged.
- Synthetic data: Simulated/synthetic data can support benchmarking, experimental design, rare-event modeling, and privacy-preserving analysis. But the group raised integrity and governance concerns: synthetic vendor-format raw files must be clearly watermarked and traceable, to prevent fraud and to allow repositories and journals to distinguish simulated from measured data. Synthetic data are not a substitute for high-quality real data in final model evaluation.

(iii) Interpretability and Explainable AI (XAI)
The group emphasized that interpretable ML is essential if models are to influence biological discovery, clinical triage, or mechanistic reasoning. Two levels were distinguished:

- Explainability: Tools such as SHAP values, saliency maps, causal/graphical models, and feature attribution methods can reveal which features drive predictions in a given model or for a given sample.
- Interpretability: The biological and biochemical meaning of those features must then be contextualized. For example, a classifier that distinguishes tissue-of-origin should identify pathways and protein/glycoform signatures plausibly linked to that tissue, rather than artifacts such as systematic missingness or batch effects.

Planned output: The group outlined a manuscript arguing for explainable and interpretable AI in MS-based proteomics and glycoproteomics. This manuscript will catalog use cases (e.g. tissue-of-origin prediction, phospho-signaling interpretation, glycoform-driven biomarker hypotheses, failure analysis in de novo sequencing), pitfalls (spurious correlations, hidden batch effects), and recommendations for reporting.
(iv) Education and Community Infrastructure
Interest in ML for proteomics and glycomics is accelerating, but training materials remain fragmented. Rather than generating entirely new courses, the group agreed to curate and maintain a community-driven "Awesome Proteomics/ML" style resource, linked to ProteomicsML and ELIXIR TeSS. This resource will collect:

- Introductory material for ML researchers entering proteomics (instrument basics, data structure, pitfalls).
- Practical material for experimentalists entering ML (basic statistics/ML literacy, bias analysis, model evaluation).

- Glyco-specific onboarding resources, which are currently underrepresented compared to proteomics.

The group discussed adding vetted content to TeSS and ProteomicsML, tagging material by audience and task, and exploring lightweight retrieval-augmented assistants based on this curated corpus.

Summary

The ML working group shifted attention from "How do we get slightly better predictions?" to "How do we build models, datasets, software, and training practices that the community can trust, reuse, and interpret – and that actually matter biologically and clinically?" Deliverables in progress include:

- A standards/checklist manuscript integrating DOME, FAIR4RS, and EVERSE for ML in proteomics and glycomics.
- A perspective on explainable and interpretable AI in MS-based proteomics.
- A shared educational and onboarding resource, to be maintained in community infrastructure.

4 Open problems

4.1 Outlook

Rebekah Gundry (University of Nebraska – Omaha, US), Magnus Palmblad (Leiden University Medical Center, NL), and Mathias Wilhelm (TU München – Freising, DE)

The 2025 Dagstuhl Seminar crystallized a shift in the field: from chasing marginal analytical improvements toward building scalable, interpretable, clinically relevant, and socially sustainable proteomics.

Across all working groups, several unifying priorities emerged:

- Standards and reproducibility before novelty.
- Participants called for enforceable expectations around metadata, software quality, model explainability, and statistical rigor – in translational workflows, ML models, glycan/glycopeptide assignment, and multi-omics integration.
- AI-ready, creditable datasets at scale.
- Whether through ENIGMA (to generate >100,000 harmonized datasets across organisms and clinical samples) or through curated benchmark datasets with explicit uncertainty and metadata, the community is moving toward data infrastructure as a collective asset. This shift demands new incentive structures for data contributors, including dataset DOIs, citation tracking, dataset papers, and recognition in funding and hiring.
- Interpretability and clinical relevance.
- The community emphasized not only accurate predictions, but predictions that can be explained, trusted, and acted upon in biological and clinical contexts – from pathway-level interpretation to tumor board recommendations.
- Inclusion of glycosylation, PTMs, and spatial context.

- Participants repeatedly stressed that clinically useful proteomics must incorporate proteoforms, phosphorylation, glycosylation, immunopeptidomics, and spatial heterogeneity, rather than collapsing biology to "protein ID + abundance."
- Bridging to clinical reality.

Translational proteomics was framed as a pipeline problem, not a single breakthrough: standardized acquisition, continuous data ingestion, interpretable computation, and clinically grounded reporting. Federated and privacy-aware learning will likely become central as proteomics enters regulated clinical environments. The seminar concluded with concrete action items: manuscripts on ML standards and explainability; a best-practices/tutorial manuscript and a quantitative glycoform analysis manuscript from the glyco group; an ENIGMA proposal for large-scale, AI-ready translational proteomics data; a position statement on data sharing, incentive structures, and federated learning; and guidance on realistic multi-omics integration and QC. Together, these efforts define a forward-looking agenda for computational and translational proteomics in the coming years.

Last but not least, the participants discussed potential future topics for a next Dagstuhl meeting on computational proteomics. Such a meeting could be structured around a coherent progression from data generation to biological insight. It could begin with the current state of proteomics technologies, focusing on challenges and open gaps in state-of-the-art mass spectrometry, including ion mobility MS, alongside the role of non-MS approaches such as affinity proteomics and single-molecule protein sequencing. From there, the meeting could address fundamental questions of protein identification and characterization, including the status and future of de novo sequencing methods and the ongoing debate around the existence and relevance of the dark proteome. Building on these measurements, structural proteomics – encompassing cross-linking, HDX, FPOP, PTM analysis (including glycans), AlphaFold-based models, and integrated structural modeling – connects molecular structure to function and modification. This in turn motivates deeper consideration of data integration challenges, including N- and P-integration and links between proteomics, cryo-EM, and other structural modalities. At a higher level of abstraction, the integration of multi-omics data with machine learning, foundational models, and the concept of virtual or digital cells offers a unifying computational framework. Finally, the meeting could broaden its scope to challenging and underrepresented application areas such as meta-omics, plant proteomics, paleoproteomics, non-model organisms, and non-human proteomics relevant to food and nutrition, with data integration serving as a cross-cutting theme throughout all topics.

Participants

- Charlotte Adams
University of Antwerp, BE
- Kiyoko Aoki-Kinoshita
Soka University – Tokyo, JP
- Gad Armony
Bruker Nederland – Leiderdorp, NL
- Wout Bittremieux
University of Antwerp, BE
- Isabell Bludau
Unviversitätsklinikum Heidelberg, DE
- Robert Chalkley
University of California – San Francisco, US
- Tine Claeys
Ghent University, BE
- Stephanie Cologna
University of Illinois – Chicago, US
- Eric Deutsch
Institute for Systems Biology – Seattle, US
- Patrick Emery
Matrix Science Ltd. – London, GB
- Melanie Föll
Universitätsklinikum Freiburg, DE
- Wassim Gabriel
TU München – Freising, DE
- Paula González Menéndez
University of Adelaide, AU
- Rebekah Gundry
University of Nebraska – Omaha, US
- Devon Kohler
Northeastern University – Boston, US
- Lev Levitskiy
University of Southern Denmark – Odense, DK
- Klaus Lindpaintner
Bruker – Concord, US
- Frédérique Lisacek
Swiss Institute of Bioinformatics – Geneva, CH
- Zhiwei Liu
Westlake University – Hangzhou, CN
- Sriram Neelamegham
University at Buffalo – SUNY, US
- Magnus Palmblad
Leiden University Medical Center, NL
- Daniel Polasky
University of Michigan – Ann Arbor, US
- Rene Ranzinger
University of Georgia – Athens, US
- Tobias Schmidt
MSAID – Garching, DE
- Nicola Ternette
University of Dundee, GB
- Sergey Vakhrushev
University of Copenhagen, DK
- Hans Wessels
Radboud University Nijmegen, NL
- Mathias Wilhelm
TU München – Freising, DE
- Dirk Winkelhardt
Ruhr-Universität-Bochum, DE
- Bernd Wollscheid
ETH Zürich, CH