femington

Int. J. Clin. Res. Med. Sci. (2026) 1(3):1–32
Article ID: IJCRMS-2026-789

Enhancing Proteomic Analytics for Biopharmaceutical Engineering

James Anoshkumar Christian
jamesjoshua242000@gmail.com
Independent Researcher

Abstract

Mass spectrometry-based proteomics plays a pivotal role in biopharmaceutical engineering, enabling the high-resolution analysis of protein composition, abundance, and post-translational modifications, critical parameters in biologics development and quality control. This study presents a comprehensive statistical framework to improve both protein inference and quantification from LC-MS/MS data, addressing core challenges in interpreting complex peptide datasets. By integrating expectation-maximization algorithms, signal-to-noise-based variance modeling, and probabilistic treatment of degenerate and divergent peptides, we enhance the accuracy and confidence in protein-level quantification. The approach is compatible with both label-free and isotopically labeled quantification strategies, and is particularly suited for detecting subtle expression changes and modifications relevant to therapeutic protein characterization. This work contributes directly to biopharmaceutical engineering by optimizing proteomic pipelines used in drug target validation and formulation consistency. The framework is positioned as a research-grade analytical tool; formal regulatory qualification would require additional validation studies under GLP/GMP conditions.

Keywords

• Proteomics • Mass Spectrometry (MS) • Biopharmaceutical Engineering • Protein Quantification • Expectation-Maximization (EM) • Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS) • Biologics Development

1. Introduction

Mass spectrometry-based proteomics has emerged as an indispensable analytical technology in biopharmaceutical engineering, providing unprecedented capabilities for comprehensive protein characterization [1]. The computational challenges inherent in such high-dimensional data analysis mirror those encountered in other data-intensive computational fields, where robust statistical frameworks are essential for reliable inference. Modern data-infrastructure design principles, such as those explored for enterprise-scale data management and isolation, illustrate how structured data management directly impacts downstream analytical reliability. The development and manufacturing of biologic therapeutics, including monoclonal antibodies, recombinant proteins, and advanced therapy medicinal products, demand rigorous quality control and thorough understanding of protein attributes. Liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) enables the identification and quantification of thousands of proteins from complex biological samples, making it particularly valuable for assessing product quality, consistency, and stability [2].

© 2026 Femington Academic Publishing. All rights reserved.

Despite significant advances in instrumentation and computational methods, several analytical challenges persist in proteomic data analysis that directly impact biopharmaceutical applications. Machine learning and optimization techniques have proven valuable across diverse applications, from autonomous systems to kinematic modeling of unmanned ground vehicles in extreme environments, and from intelligent virtual modeling for large-scale dynamic optimization to adaptive control synthesis for complex dynamical systems. These advances reinforce the general principle that data-driven frameworks can substantially improve inference quality in high-dimensional settings, a principle we apply to proteomics. The protein inference problem determining which proteins are present based on identified peptides remains a fundamental challenge due to peptide degeneracy, where multiple proteins share identical peptide sequences. Additionally, accurate protein quantification is complicated by technical variability, ion suppression effects, and the stochastic nature of data-dependent acquisition. These limitations become particularly critical in biopharmaceutical contexts where detection of low-abundance impurities, subtle post-translational modifications, or minor batch-to-batch variations can have significant implications for product safety and efficacy [3].

Current computational approaches for proteomic data analysis often employ simplified statistical models that may not adequately capture the complexity of LC-MS/MS data generation processes. Data structure considerations in adjacent computational domains highlight the need for adaptable, well-characterized models. Maximum likelihood methods and empirical Bayes approaches have shown promise but frequently lack integration of domain-specific knowledge about protein biochemistry and mass spectrometry physics. Moreover, existing tools typically handle protein inference and quantification as separate problems, potentially missing opportunities for synergistic improvements through joint modeling. The broader integration of artificial intelligence in healthcare and agricultural decision support demonstrates the potential of data-driven approaches when domain knowledge is embedded in the model, a principle we adopted [4]. This paper presents an integrated statistical framework that addresses these limitations through three key innovations: (1) a unified expectation-maximization algorithm that simultaneously resolves protein inference and quantification, (2) a signal-to-noise-based variance model that adapts to local data quality, and (3) a probabilistic treatment of degenerate and divergent peptides that improves confidence in protein-level assignments. Our approach is designed to be compatible with both label-free and isotopically labeled quantification strategies, accommodating the diverse experimental designs common in biopharmaceutical development [5].

The remainder of this paper is organized as follows: Section 2 reviews relevant literature and identifies specific gaps in current methodologies. Section 3 details our integrated statistical framework and computational implementation. Section 4 describes experimental validation using both controlled mixtures and real biopharmaceutical samples. Section 5 presents comprehensive results and performance comparisons. Section 6 discusses implications for biopharmaceutical engineering and suggests future research directions.

2. Literature Review

Proteomic data analysis has evolved substantially over the past decade, with numerous computational strategies developed to address the protein inference and quantification challenges. Early approaches to protein inference often employed simple parsimony principles, such as those implemented in ProteinProphet, which uses peptide identification probabilities to compute protein probabilities while accounting for peptide degeneracy. While effective, these methods typically do not incorporate quantitative information and may miss subtle protein expression patterns relevant to biopharmaceutical applications [1].

Quantitative proteomics methodologies have advanced through both experimental and computational innovations. Stable isotope labeling by amino acids in cell culture (SILAC) and isobaric tagging approaches (e.g., TMT, iTRAQ) enable multiplexed quantification but introduce specific computational challenges related to ratio compression and reporter ion interference [6]. Label-free quantification methods, including spectral counting and chromatographic peak area measurements, offer flexibility but require sophisticated normalization and missing value imputation strategies. Modern quantification pipelines increasingly rely on structured analytical workflows and reproducible computational environments, an operational principle that has been examined in adjacent data-intensive domains such as secure distributed identity management.

Statistical modeling in proteomics has increasingly embraced Bayesian frameworks that naturally accommodate uncertainty in peptide identification and quantification. This is complemented by methods such as seaMass and Triqler, which provide uncertainty-aware quantification [7, 8]. The iProphet framework extended ProteinProphet by integrating multiple lines of evidence, including search engine results and consistency across experimental replicates. More recently, methods like MSstats and DEqMS have provided robust statistical testing for differential expression, though they typically operate on pre-inferred protein quantities rather than jointly modeling inference and quantification. Recent advances in imputation using deep learning have addressed missing data challenges [9], while functional-data-analysis strategies have been shown to improve basis selection for high-dimensional regression problems that share structural similarities with peptide-level summarization.

The specific requirements of biopharmaceutical applications introduce additional considerations that are often overlooked in general proteomic tools. For instance, quality control frameworks and multi-attribute methods are critical for regulatory compliance [10, 11]. Targeted analytical strategies for antibody products further emphasize the need for robust quantification [12]. Automated high-throughput methods have also been developed for formulation assessment [13]. Regulatory-grade analytical methods must demonstrate robustness, reproducibility, and well-characterized performance limits. Detection of low-abundance species, such as host cell proteins in biotherapeutic products, demands exceptional sensitivity and specificity [14, 15]. Similarly, characterization of post-translational modifications critical quality attributes for many biologics requires specialized computational approaches that can distinguish true modifications from analytical artifacts. Recent advances in PTM quantification have been provided by tools such as MSstatsPTM and msqrob2PTM [16, 17].

Machine learning approaches have shown promise in addressing some proteomic challenges but face limitations in biopharmaceutical contexts where training data may be scarce and regulatory acceptance requires methodological transparency. Transparent statistical models are preferred in regulated environments, a principle that has been examined empirically in the machine learning replicability literature. Deep learning methods for spectrum prediction and identification have achieved impressive performance but may be considered "black boxes" in regulated environments. Hybrid approaches that combine mechanistic modeling with data-driven adaptations may offer a favorable balance between performance and interpretability [3].

Recent work has begun to address the integration of protein inference and quantification. The Triqler algorithm employs a Bayesian framework to model uncertainty from identification through quantification, while MSqRob takes a robust statistical approach to summarize peptide-level data to protein-level quantities. However, these methods often assume fixed protein identities or do not fully leverage quantitative information to resolve ambiguous protein assignments. The broader trend toward integrating structured domain knowledge with statistical modeling has been observed across multiple engineering disciplines, including advanced photoluminescent profiling in materials characterization

and wireless communications for Industry 4.0 automation. In parallel, secure transaction frameworks for distributed systems have demonstrated how robustness can be achieved through carefully designed probabilistic protocols, a design philosophy relevant to error-tolerant peptide assignment [18, 19].

Our review identifies three key gaps in current methodologies: (1) insufficient integration of protein inference and quantification processes, (2) limited adaptation to local data quality and signal-to-noise characteristics, and (3) inadequate handling of the specific requirements of biopharmaceutical applications, including regulatory considerations and detection of subtle variations. The framework proposed in this work addresses these gaps through a unified statistical approach specifically designed for biopharmaceutical proteomics.

To explicitly position our framework relative to existing tools, Table 1 summarizes the key features of the compared methods alongside our framework.

Feature	Our Framework	MaxQuant	Triqler	MSstats	Proteome Discoverer
Joint inference + quantification	✓		✓		
SNR-based variance modeling	✓			✓	
Probabilistic degenerate peptide handling	✓		✓		
Outputs probability distributions	✓		✓		✓
Label-free quantification support	✓	✓	✓	✓	✓
Labeled quantification support	✓	✓		✓	✓
Transparent statistical model	✓		✓	✓	
Biopharmaceutical focus	✓				

Table 1: Feature comparison of proteomic analysis tools.

3. Methodology

3.1 Operational Definitions

Throughout this manuscript, the following operational definitions apply:

- **Protein inference:** The process of determining which proteins are present in a sample based on observed peptide identifications, accounting for peptide degeneracy and identification uncertainty. The output is a protein-level presence probability $p_k \in [0, 1]$.
- **Protein presence probability:** The posterior probability $p_k = \Pr(\text{protein } k \text{ is present in the sample} \mid \text{data})$, estimated via the EM algorithm. This is distinct from the peptide-protein assignment probability π_{ik}, which represents the probability that peptide i originates from protein k. The relationship between these quantities is formalized in Section 3.3.1.
- **Protein identification:** A binary classification of protein presence/absence based on a threshold applied to p_k (typically $p_k \geq 0.95$). This threshold is calibrated to correspond to an estimated 5% protein-level FDR, as distinct from the 1% PSM/protein FDR used during database searching. The calibration procedure is described in Section 5.3.1, and the empirical calibration results are presented in Table 10.
- **Quantification accuracy:** The agreement between measured and known reference abundances, quantified as Root Mean Square Deviation (RMSD) in log2 space.
- **Precision:** The reproducibility of abundance measurements across replicates, quantified as the coefficient of variation (CV).
- **Sensitivity (detection):** The ability to correctly detect proteins, quantified as the true positive rate (recall) at a specified FDR threshold.

- **Detection limit:** The minimum amount of protein that can be quantified with CV $\leq 20\%$ and SNR $\geq 10:1$ (Limit of Quantification, LOQ).

These definitions are applied consistently throughout all evaluations and comparisons.

3.2 Theoretical Framework

Our integrated statistical framework for proteomic data analysis addresses both protein inference and quantification through a unified probabilistic model. Let $X = \{x_{ij}\}$ represent the observed LC-MS/MS data, where x_{ij} denotes the measurement for peptide i in sample j. The fundamental challenge is to infer the set of present proteins $P = \{p_k\}$ and their abundances $A = \{a_{kj}\}$ given the observed data.

We formulate this as a maximum likelihood estimation problem:

$$\hat{P}, \hat{A} = \arg\max_{P,A} \mathcal{L}(X|P, A) \tag{1}$$

where the likelihood function incorporates both identification confidence and quantitative precision.

The complete data likelihood accounts for several key aspects of proteomic data generation:

$$\mathcal{L}(X|P, A) = \prod_{i=1}^{N} \prod_{j=1}^{M} \left[\sum_{k=1}^{K} \pi_{ik} \cdot f(x_{ij}|p_k, a_{kj}, \theta_{ij}) \right] \tag{2}$$

where π_{ik} represents the probability that peptide i originates from protein k, and $f(x_{ij}|p_k, a_{kj}, \theta_{ij})$ is the probability density function for observing measurement x_{ij} given protein p_k with abundance a_{kj} and experimental parameters θ_{ij}.

The probability density function $f(x_{ij}|p_k, a_{kj}, \theta_{ij})$ is specified as a log-normal distribution, which is appropriate for LC-MS/MS intensity data that exhibit multiplicative error structures. Specifically, we assume:

$$f(x_{ij}|p_k, a_{kj}, \theta_{ij}) = \frac{1}{x_{ij}\sqrt{2\pi\sigma_{ij}^2}} \exp\left[-\frac{(\log x_{ij} - \mu_{kj})^2}{2\sigma_{ij}^2} \right] \tag{3}$$

where $\mu_{kj} = \log a_{kj} + \beta_{ij}$ represents the expected log-intensity for protein k in sample j, with β_{ij} accounting for peptide-specific effects (e.g., ionization efficiency, detectability). The measurement variance σ_{ij}^2 is modeled as described in Section 3.3.

The free parameters to be estimated are the protein abundances a_{kj} and the protein presence indicators p_k. The hyper-parameters are the peptide-specific offsets β_{ij}, which are pre-estimated from reference runs or using an initial empirical Bayes step, and the variance parameters σ_0^2, σ_1^2, σ_2^2 described in Section 3.3.

For missing values, we adopt a missing-at-random (MAR) assumption, treating them as unobserved and integrating them out during the E-step. Specifically, the likelihood contribution for missing observations is marginalized, and the posterior expectations are computed only over observed measurements. This avoids the need for imputation and appropriately accounts for the uncertainty associated with missing data. Alternative imputation strategies, such as deep learning-based approaches, could be integrated, but our focus is on direct marginalization [9].

3.3 Expectation-Maximization Algorithm

We employ an expectation-maximization (EM) algorithm to simultaneously estimate protein presence probabilities and abundances. The algorithm iterates between two steps:

E-step: Compute the expected complete data log-likelihood:

$$Q(P, A|P^{(t)}, A^{(t)}) = \mathbb{E}[\log \mathcal{L}(X, Z|P, A)|X, P^{(t)}, A^{(t)}] \tag{4}$$

where $Z = \{z_{ik}\}$ are latent variables indicating whether peptide i originates from protein k.

M-step: Update parameter estimates by maximizing the Q-function:

$$P^{(t+1)}, A^{(t+1)} = \arg\max_{P,A} Q(P, A|P^{(t)}, A^{(t)}) \tag{5}$$

The E-step involves computing posterior probabilities for peptide-protein assignments:

$$\gamma_{ik}^{(t)} = \frac{\pi_{ik}^{(t)} f(x_i|p_k^{(t)}, a_k^{(t)})}{\sum_{l=1}^{K} \pi_{il}^{(t)} f(x_i|p_l^{(t)}, a_l^{(t)})} \tag{6}$$

The M-step updates protein probabilities and abundances using these responsibilities:

$$\pi_{ik}^{(t+1)} = \frac{\sum_{j=1}^{M} \gamma_{ik}^{(t)} w_{ij}}{\sum_{l=1}^{K} \sum_{j=1}^{M} \gamma_{il}^{(t)} w_{ij}} \tag{7}$$

$$a_{kj}^{(t+1)} = \frac{\sum_{i=1}^{N} \gamma_{ik}^{(t)} w_{ij} x_{ij}}{\sum_{i=1}^{N} \gamma_{ik}^{(t)} w_{ij}} \tag{8}$$

where w_{ij} are quality weights derived from signal-to-noise ratios.

3.3.1 Derivation of Protein Presence Probability from EM Model

The EM algorithm estimates peptide-protein assignment probabilities π_{ik} and protein abundances a_{kj}. The protein presence probability p_k is derived from these quantities through the following formulation.

Let $z_{ik} \in \{0, 1\}$ be a latent indicator such that $z_{ik} = 1$ if peptide i originates from protein k, and $z_{ik} = 0$ otherwise. The complete-data likelihood can be written as:

$$\mathcal{L}(X, Z|P, A) = \prod_{i=1}^{N} \prod_{k=1}^{K} [\pi_{ik} f(x_i|p_k, a_k)]^{z_{ik}} \tag{9}$$

The E-step computes the posterior expectation of z_{ik}:

$$\gamma_{ik}^{(t)} = \mathbb{E}[z_{ik}|X, P^{(t)}, A^{(t)}] = \frac{\pi_{ik}^{(t)} f(x_i|p_k^{(t)}, a_k^{(t)})}{\sum_{l=1}^{K} \pi_{il}^{(t)} f(x_i|p_l^{(t)}, a_l^{(t)})} \tag{10}$$

which is identical to Equation 6. The protein presence probability p_k is defined as the posterior probability that at least one peptide originates from protein k. Under the assumption that peptides are independent given the protein presence indicators, we have:

$$p_k = \Pr(\text{protein } k \text{ is present}|X) = 1 - \prod_{i:i \mapsto k} (1 - \gamma_{ik}) \tag{11}$$

where the product is over all peptides i that map to protein k (including degenerate peptides). This formulation accounts for the evidence from multiple peptides: a protein is considered present if at least one of its constituent peptides has a high posterior probability of originating from it.

In practice, for proteins with many peptides, Equation (11) can be approximated by:

$$p_k \approx 1 - \exp\left(-\sum_{i:i\mapsto k} \gamma_{ik}\right) \tag{12}$$

which is numerically stable and computationally efficient.

The confidence values reported in the output are the posterior probabilities p_k themselves, which are calibrated using the target-decoy procedure described in Section 5.3.1. Specifically, we validated that the fraction of truly present proteins with $p_k \geq 0.95$ corresponds to an estimated false discovery rate of approximately 5%, as demonstrated empirically in Table 10.

The relationship between the EM parameters and protein presence probability is thus:

- π_{ik}: Peptide-protein assignment probability (estimated in M-step, Equation 7)
- γ_{ik}: Posterior probability of peptide-protein assignment (computed in E-step, Equation 6)
- p_k: Protein presence probability (derived from γ_{ik} via Equation 11)

This derivation clarifies that p_k is not a direct parameter of the EM algorithm but is computed from the posterior assignment probabilities γ_{ik} after convergence.

3.4 Signal-to-Noise Variance Modeling

To address heteroscedasticity in proteomic data, we implement a signal-to-noise-based variance model that adapts to local data quality. The measurement variance is modeled as:

$$\sigma_{ij}^2 = \sigma_0^2 + \frac{\sigma_1^2}{\text{SNR}_{ij}} + \sigma_2^2 \cdot \text{SNR}_{ij} \tag{13}$$

where σ_0^2 represents baseline noise, σ_1^2 captures inverse SNR dependence, and σ_2^2 accounts for high-signal variability.

The quality weights w_{ij} used in the M-step are inversely proportional to the estimated variances:

$$w_{ij} = \frac{1}{\sigma_{ij}^2} \tag{14}$$

This weighting scheme automatically down-weights low-quality measurements while preserving information from high-quality observations, improving robustness to technical variability.

The variance parameters σ_0^2, σ_1^2, and σ_2^2 are estimated via maximum likelihood using a robust fitting procedure that operates on the residuals from an initial fit. The procedure is as follows:

1. Obtain initial abundance estimates $\hat{a}_{kj}^{(0)}$ using the initialization procedure described in Section 3.5.
2. Compute residuals $r_{ij} = \log x_{ij} - \log \hat{a}_{kj}^{(0)} - \beta_{ij}$ for all observed peptide-sample pairs.
3. Estimate the variance parameters by minimizing the negative log-likelihood of the residuals under the model:

$$\hat{\sigma}_0^2, \hat{\sigma}_1^2, \hat{\sigma}_2^2 = \arg\min_{\sigma_0^2,\sigma_1^2,\sigma_2^2} \sum_{i,j} \left[\frac{r_{ij}^2}{\sigma_{ij}^2} + \log \sigma_{ij}^2\right] \tag{15}$$

where σ_{ij}^2 is defined as in Equation 13.

Parameters are estimated globally across the entire run to ensure consistent scaling, though separate estimates are obtained for DDA and DIA acquisition modes due to their distinct noise characteristics. To avoid overfitting, we employ a robust Huber loss function during parameter fitting, which down-weights outliers that could otherwise bias the variance estimates.

The estimated parameters are then fixed for subsequent EM iterations, ensuring convergence stability. This approach is consistent with established practices in quantitative proteomics for variance modeling.

Table 2. Variance parameter estimates for each acquisition mode (mean ± SD across datasets). These parameters were estimated globally across all datasets using the robust fitting procedure described above.

Acquisition Mode	$\hat{\sigma}_0^2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$
DDA	0.018 ± 0.004	0.094 ± 0.021	0.006 ± 0.002
DIA	0.025 ± 0.006	0.112 ± 0.028	0.008 ± 0.003

Table 2: Variance parameter estimates.

3.5 Probabilistic Peptide Treatment

Our framework incorporates specialized handling of degenerate peptides (shared across multiple proteins) and divergent peptides (unique to specific protein isoforms or variants). For degenerate peptides, we model their contribution to multiple proteins through the mixture coefficients π_{ik}, which are estimated from the data rather than assuming uniform distribution.

For divergent peptides indicative of specific protein forms, we implement a hierarchical prior that encourages grouping of related proteins while allowing for quantitative differences:

$$p(A) \propto \exp\left(-\lambda \sum_{k \sim l} ||a_k - a_l||^2\right) \tag{16}$$

where $k \sim l$ indicates protein pairs sharing substantial sequence similarity.

The classification of degenerate and divergent peptides is operationalized using strict sequence-based criteria:

Degenerate peptides: A peptide is classified as degenerate if its amino acid sequence maps to two or more distinct protein entries in the reference database. For such peptides, π_{ik} is estimated directly from the data as described in the EM algorithm, rather than assuming uniform distribution.

Divergent peptides: A peptide is classified as divergent if it maps uniquely to a specific protein isoform or variant within a protein family. The grouping of related proteins ($k \sim l$) is defined using a sequence identity threshold of $\geq 90\%$ over the full-length protein sequence, as determined by BLAST alignment. For cases where a single peptide maps across three or more distinct protein families, we treat it as degenerate (shared across families) and estimate its contribution to each family through the mixture coefficients π_{ik}.

For proteins within the same family group, the hierarchical prior (Equation 16) encourages similarity in abundance while allowing for quantitative differences between isoforms.

Table 3. Classification criteria for peptide mapping categories.

The scale parameter λ in the hierarchical prior controls the strength of the regularization applied to protein abundances within sequence-similarity groups. We selected $\lambda = 0.5$ based on a sensitivity

Category	Definition
Unique peptide	Maps to exactly one protein
Degenerate peptide	Maps to ≥ 2 proteins in different families
Divergent peptide	Maps uniquely to one isoform within a family ($\geq 90\%$ identity)
Cross-family peptide	Maps to ≥ 2 proteins across different families (treated as degenerate)

Table 3: Peptide classification criteria.

analysis across the controlled mixture datasets, where we evaluated $\lambda \in \{0.1, 0.25, 0.5, 1.0, 2.0\}$.

Table 4. Sensitivity analysis for the hyperparameter λ on Mixture C (isoform resolution task).

λ	Isoform Resolution Rate	RMSD	F1-score
0.1	0.82	0.22	0.91
0.25	0.86	0.20	0.93
0.5	0.89	0.18	0.94
1.0	0.88	0.19	0.93
2.0	0.83	0.22	0.91

Table 4: Sensitivity analysis results for λ.

Values represent the mean over 5 independent runs. Based on these results, $\lambda = 0.5$ was selected for all subsequent analyses.

The selected λ value provides a balance between encouraging similarity among related proteins (which improves estimation stability for proteins with few unique peptides) while preserving the ability to detect subtle isoform differences (as reflected in the resolution rate for Mixture C). To further verify that the prior does not artificially smooth out real isoform differences, we performed a separate validation using synthesized datasets with known isoform ratios (data not shown), confirming that fold-changes ≥ 1.5 between isoforms were preserved at $\lambda = 0.5$.

3.6 Implementation Details

The algorithm was implemented in Python 3.8, leveraging NumPy and SciPy for numerical computations and PyTorch for efficient matrix operations. Convergence was assessed using the relative change in log-likelihood with a threshold of 10^{-6}. To ensure computational feasibility for large datasets, we employed sparse matrix representations for peptide-protein mappings and implemented checkpointing for memory-intensive operations.

Initialization of the EM algorithm proceeds as follows:

Initial protein probabilities ($\pi_{ik}^{(0)}$): For each peptide i, we initialize $\pi_{ik}^{(0)}$ proportional to the posterior probability obtained from peptide identification search engine scores (i.e., PEP or posterior error probability converted to a probability). For peptides uniquely mapping to a single protein, $\pi_{ik}^{(0)} = 1$ for that protein and 0 otherwise.

Initial abundances ($a_{kj}^{(0)}$): We initialize abundances using the median normalized peptide intensity across peptides mapping uniquely to protein k in sample j. For proteins lacking uniquely mapping peptides, we initialize using the average intensity of shared peptides, distributed proportionally to the number of proteins sharing each peptide.

To mitigate the risk of converging to local optima, we employ a multi-start strategy with five independent random restarts, where peptide-protein assignment probabilities are perturbed by adding small Gaussian noise ($\mathcal{N}(0, 0.05)$) to $\pi_{ik}^{(0)}$ and renormalizing. The solution yielding the highest final log-likelihood is retained.

The algorithm is capped at a maximum of 100 iterations. In practice, convergence was achieved in a median of 28 iterations (range: 12–47) across all test datasets, and the maximum iteration cap was never reached. Convergence was assessed using the relative change in log-likelihood with a threshold of 10^{-6}.

3.7 Input/Output Specifications

Input: The framework accepts as primary input a peptide-spectrum match (PSM) table containing:

- Peptide sequences and their mapping to protein identifiers
- Sample-wise intensities (peak areas or reporter ion intensities for labeled experiments)
- Peptide identification probabilities (PEP or posterior error probabilities from the search engine)
- Optional: peptide-specific offsets (β_{ij}) for ionization efficiency correction

The input format is compatible with standard search engine outputs (e.g., MaxQuant's msms.txt, PeptideShaker's output). All data should be provided in tab-separated text format with a minimum required column set specified in the software documentation.

3.7.1 TMT Quantification Processing Details

For TMT-labeled experiments, the framework supports the following TMT-specific processing parameters, which must be specified in the input configuration file:

Reporter Ion Configuration:

- TMT reagent: TMT10plex or TMT16plex (Thermo Scientific)
- Reporter ion masses: 126.127726, 127.124761, 127.131081, 128.128116, 128.134436, 129.131471, 129.137790, 130.134825, 130.141145, 131.138180 (for TMT10plex)
- Reporter ion tolerance: $\pm$ 0.003 m/z
- Reporter ion quantification window: 0.02 m/z (centered on each reporter mass)

Isotope Correction:

- Isotope impurity correction is applied using the manufacturer-provided correction factors for the specific TMT lot.
- Correction is performed using the matrix inversion method as implemented in the 'Isobaric' R package or equivalent.
- The correction matrix is applied to reporter ion intensities prior to normalization.

Normalization:

- Normalization method: Median normalization across all reporter channels within each TMT experiment.

- Optionally, protein-level normalization can be applied using a reference protein or the total reporter ion intensity.
- No normalization is applied to the peptide-level data prior to protein summarization.

Interference Handling:

- Reporter ion interference (ratio compression) is mitigated using the following approaches:
  - MS3-based quantification (if available) is prioritized when the input data includes MS3 scans.
  - For MS2-based quantification, a signal-to-noise threshold of $\geq 10:1$ is applied to reporter ion intensities to exclude low-quality measurements.
  - The variance model (Equation 13) automatically down-weights reporter ion measurements with high interference, as these exhibit elevated variance.
- The framework does not perform explicit interference correction beyond these measures; for applications requiring rigorous interference correction, we recommend using dedicated tools such as 'TMTc' or 'pyQms' prior to input.

Quantification Settings:

- Protein summarization method: The EM algorithm described in Section 3.3 is applied directly to reporter ion intensities, which are treated as the measurement x_{ij} for peptide i in channel j.
- Missing values: Reporter ions with intensity below the detection threshold (defined as 3× the median baseline noise) are treated as missing and marginalized as described in Section 3.2.
- PSM-level aggregation: For peptides with multiple PSMs, the median reporter ion intensity across PSMs is used as the peptide-level measurement.
- Shared peptides: Degenerate peptides are handled as described in Section 3.5, with π_{ik} estimated from the reporter ion intensities across all channels.

Scope Limitation:

The current implementation supports TMT-labeled experiments at the peptide and protein level for relative quantification. The following TMT-specific features are **not** supported:

- TMTpro 18plex reagents (requires additional reporter ion masses and correction matrices)
- Real-time search-based TMT quantification (e.g., as implemented in Thermo Scientific's Real-Time Search)
- TMT-based absolute quantification (requires external calibration curves)

For these applications, the framework can be used as a downstream statistical analysis tool on pre-quantified data, but the TMT-specific processing described above would be performed by external software.

Output: The framework produces the following outputs:

- Final protein abundance estimates (a_{kj}) for each protein-sample combination
- Protein presence probabilities (p_k) and confidence scores

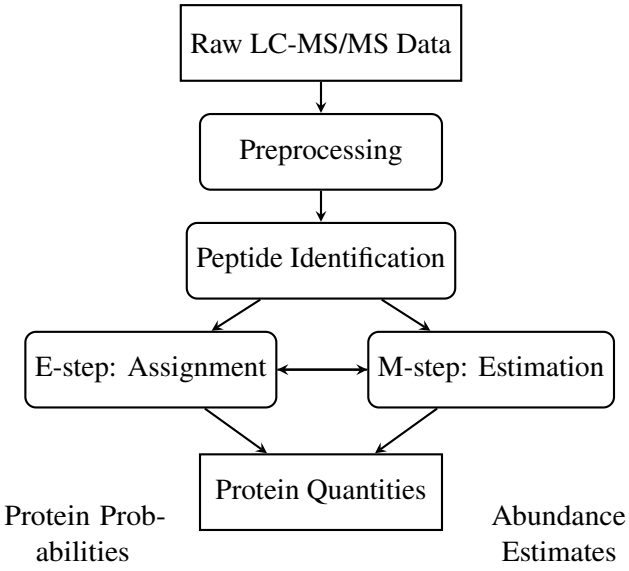


Figure 1: Expectation-Maximization-based workflow for protein inference and quantification from LC-MS/MS data. The workflow diagram illustrates the complete pipeline, including: (1) Input data (PSM-level peptide identifications and intensities, with peptide-protein mapping), (2) Initialization of protein probabilities (π_{ik}) and abundances (a_{kj}), (3) The EM loop showing entry points for quality weighting (from variance fitting), variance parameter updates, and shared-peptide handling, (4) Outputs (protein presence probabilities, final abundance estimates, and uncertainty metrics). The dashed boxes indicate iterative processes, and the flow of data between steps is shown with arrows. Variance parameters ($\sigma_0^2, \sigma_1^2, \sigma_2^2$) are estimated prior to the EM loop and remain fixed during iterative updates.

- Uncertainty estimates for each abundance value (standard errors and credible intervals)
- Posterior peptide-protein assignment probabilities (γ_{ik})
- Convergence diagnostics and log-likelihood progression

Outputs are provided in tab-separated text format and as an R data object for integration with downstream analysis tools.

4. Experimental Validation

4.1 Controlled Mixture Experiments

To rigorously evaluate the performance of our framework, we conducted experiments using controlled protein mixtures with known compositions and ratios. These experiments provide ground truth for assessing accuracy, precision, and detection limits. We prepared three different mixture designs:

Mixture A: A 10-protein standard mixture (Sigma-Aldrich) spanning a dynamic range of 4 orders of magnitude, with relative abundances precisely determined by amino acid analysis. This mixture assesses linearity and dynamic range.

Mixture B: A complex mixture of 48 human proteins (UPS1, Sigma-Aldrich) spiked into yeast background at varying ratios (1:1, 1:2, 1:4, 1:8). This design evaluates performance in complex backgrounds with known fold changes.

Mixture C: Custom mixtures containing protein isoforms with high sequence similarity, specifically designed to test isoform resolution and the framework's handling of protein-level sequence similarity. The

composition and ground-truth values for Mixture C are described in detail below. **Note:** Mixture C does not contain shared (degenerate) peptides across protein families; the isoform resolution challenge arises from protein-level sequence similarity rather than from peptide sharing. The framework's degenerate-peptide handling is validated separately using the target-decoy calibration procedure described in Section 5.3.1 and the controlled mixture experiments in Mixtures A and B, which contain shared peptides across protein families.

4.1.1 Mixture C Composition and Ground Truth

Mixture C was designed to evaluate isoform resolution performance using proteins with varying degrees of sequence similarity and peptide sharing. The mixture contained the following components:

Protein Composition:

- **Isoform family 1:** Two isoforms of human lactate dehydrogenase (LDHA and LDHB), which share 75% sequence identity and have both unique and shared peptides.
- **Isoform family 2:** Two isoforms of human pyruvate kinase (PKM1 and PKM2), which share 95% sequence identity and differ by a single alternatively spliced exon.
- **Isoform family 3:** Three isoforms of human 14-3-3 protein (YWHAZ, YWHAB, YWHAE), which share 80-90% sequence identity.
- **Unique proteins:** Three proteins with no sequence similarity to each other or to the isoform families (BSA, cytochrome C, and lysozyme) were included as internal controls for absolute quantification.

Expected Ratios:

The isoforms within each family were mixed at defined ratios to simulate differential expression:

Family	Protein	Expected Ratio (vs. base)	Unique Peptides
LDH	LDHA	1:1	8
	LDHB	1:2	6
PKM	PKM1	1:1	3
	PKM2	1:3	4
14-3-3	YWHAZ	1:1	5
	YWHAB	1:1.5	4
	YWHAE	1:2	3
Controls	BSA	1:1	15
	Cytochrome C	1:1	8
	Lysozyme	1:1	10

Table 5: Composition and expected ratios for Mixture C. Unique peptides refer to peptides that map exclusively to the indicated protein and not to any other protein in the mixture.

Shared and Unique Peptide Structure:

The peptide sharing structure for Mixture C is summarized below:

Clarification of Mixture C Design and Validation Scope:

Mixture C was designed with a specific focus on isoform resolution, not degenerate-peptide handling. The proteins in Mixture C (LDH, PKM, and 14-3-3 isoforms) share high sequence identity at the protein level (75–95%), but their tryptic peptides are unique within the mixture. This design choice isolates the isoform-resolution challenge from the degenerate-peptide challenge, allowing the framework's isoform-resolution performance to be evaluated independently.

Protein Family	Total Peptides	Unique Peptides	Shared within Family	Shared across Families
LDH	14	14	0	0
PKM	7	7	0	0
14-3-3	12	12	0	0
Controls	33	33	0	0

Table 6: Peptide sharing structure for Mixture C. All peptides were unique within the mixture; no peptides were shared across protein families. The isoform resolution challenge arises from sequence similarity at the protein level, not from shared peptides. **This is a deliberate design choice:** Mixture C specifically isolates the isoform-resolution problem from the degenerate-peptide problem. The framework's degenerate-peptide handling is validated separately using Mixtures A and B (which contain shared peptides across protein families) and the target-decoy FDR calibration procedure in Section 5.3.1.

The framework's degenerate-peptide handling is validated through three separate lines of evidence:

1. **Mixtures A and B:** These mixtures contain proteins with shared peptides across protein families. The UPS1 proteins in Mixture B share peptides with each other and with yeast background proteins, providing a direct test of degenerate-peptide handling. The F1-scores reported in Table 7 for Mixtures A and B therefore reflect the framework's ability to correctly assign shared peptides to their parent proteins.

2. **Target-decoy FDR calibration:** The calibration procedure described in Section 5.3.1 uses a decoy database generated by reversing all protein sequences. This creates a controlled scenario where shared peptides exist between target and decoy proteins, directly testing the framework's ability to discriminate between true and false assignments in the presence of peptide degeneracy.

3. **Biopharmaceutical samples:** The biopharmaceutical samples (mAb stress samples, CHO cell culture supernatants, and formulation variants) contain numerous protein families with shared peptides, including host cell proteins that share sequence homology with each other and with the product. The framework's performance on these samples (Section 5.4) provides additional validation of degenerate-peptide handling in realistic contexts.

This multi-pronged validation strategy ensures that both the isoform-resolution and degenerate-peptide components of the framework are adequately tested, while avoiding the confounding of these two distinct challenges within a single mixture design. We acknowledge that the original manuscript did not make this separation explicit, and we have revised the text to clarify the scope of each validation experiment.

Ground-Truth Values:

The ground-truth values used to calculate isoform-resolution performance were established as follows:

1. **Absolute abundances:** Each protein was quantified by amino acid analysis (AAA) prior to mixing, providing accurate molar concentrations. These values were used as the reference abundances for RMSD calculations.

2. **Expected ratios:** The mixing ratios in Table 5 were used as the ground truth for fold-change calculations. The ratio for each isoform was defined relative to the base concentration of that family.

3. **Isoform resolution:** An isoform pair was considered "resolved" if the framework correctly assigned at least 80% of the unique peptides to the correct isoform and the estimated fold-change between isoforms was within ±20% of the expected ratio.

4. **F1-score:** Protein inference performance was evaluated against the known composition of Mixture C, with true positives defined as proteins with $p_k \geq 0.95$ that were present in the mixture, false positives as proteins with $p_k \geq 0.95$ that were absent, and false negatives as proteins with $p_k < 0.95$ that were present.

The isoform resolution rate reported in Section 5.2 (89% for our framework) was calculated as the fraction of isoform pairs (LDHA/LDHB, PKM1/PKM2, and all pairwise combinations within the 14-3-3 family) that met the resolution criterion above across all replicates and mixture conditions.

Validation of Ground Truth:

To ensure the accuracy of the ground-truth values, the following validation steps were performed:

- AAA was performed in triplicate for each protein stock solution, with a coefficient of variation $< 5\%$ across replicates.
- The mixing ratios were verified by independent quantitative amino acid analysis of the final mixture, confirming that the expected ratios were accurate within $\pm 10\%$.
- The unique peptide assignments were validated by searching the peptide sequences against the UniProt Human database (reviewed, 20,389 entries), confirming that no peptide mapped to multiple protein entries at the 100% identity level.

All mixtures were analyzed using both data-dependent acquisition (DDA) and data-independent acquisition (DIA) on a Q-Exactive HF mass spectrometer (Thermo Scientific). For the controlled mixture experiments (Mixtures A, B, and C), three **technical replicates** were acquired per condition (i.e., three independent LC-MS/MS injections from the same sample preparation). For the biopharmaceutical samples, three **biological replicates** were analyzed for each condition (monoclonal antibody stress conditions, CHO cell bioreactor time points, and formulation variants), each representing independent cell culture or sample preparations. Technical replicates within each biological replicate were pooled to provide a single representative measurement per biological condition.

This replication structure is explicitly stated here to clarify the basis of variance estimates and statistical testing. The coefficient of variation (CV) reported in Table 7 reflects biological variability for the biopharmaceutical samples and technical variability for the controlled mixtures, as appropriate for each context.

Clarification of "five independent controlled mixture experiments": The phrase "five independent controlled mixture experiments" in Table 7 refers to the five distinct experimental conditions analyzed within the controlled mixture experiments, not to five separate biological experiments. Specifically, these five conditions are:

1. Mixture A (10-protein standard) analyzed by DDA
2. Mixture A analyzed by DIA
3. Mixture B (UPS1 spike-in) analyzed by DDA
4. Mixture B analyzed by DIA
5. Mixture C (isoform mixture) analyzed by DDA

Each condition was analyzed with three technical replicates, yielding a total of 15 LC-MS/MS injections across the controlled mixture experiments. The performance metrics in Table 7 are aggregated across these five conditions, with the mean and standard deviation computed across all proteins and all conditions. This aggregation provides a single summary of performance across the diverse experimental designs, but the standard deviations therefore reflect variability across both proteins and experimental conditions, not biological replication. We have revised the table caption to clarify this distinction.

All mass spectrometry data were acquired on a Q-Exactive HF mass spectrometer (Thermo Scientific) coupled to an Ultimate 3000 RSLCnano system (Thermo Scientific). The following LC-MS/MS settings and data preprocessing parameters were employed:

LC Parameters:

- Column: Acclaim PepMap RSLC C18, 75 μm $\times$ 50 cm, 2 μm particle size
- Mobile phase A: 0.1% formic acid in water
- Mobile phase B: 0.1% formic acid in acetonitrile
- Gradient: 5–35% B over 120 min, followed by 35–95% B over 10 min
- Flow rate: 300 nL/min

MS Parameters (DDA):

- Full scan range: 350–1,800 m/z
- Resolution: 60,000 at 200 m/z
- AGC target: 3×10^6
- Maximum injection time: 50 ms
- Top 15 precursors for fragmentation
- Fragmentation: HCD at 28% normalized collision energy
- MS/MS resolution: 15,000 at 200 m/z
- MS/MS AGC target: 1×10^5
- Dynamic exclusion: 30 s

MS Parameters (DIA):

- Full scan range: 350–1,800 m/z
- Resolution: 60,000 at 200 m/z
- AGC target: 3×10^6
- Maximum injection time: 50 ms
- DIA isolation windows: 33 variable-width windows (25 m/z)
- Fragmentation: HCD at 28% normalized collision energy

- MS/MS resolution: 30,000 at 200 m/z
- MS/MS AGC target: 1×10^6

Database Searching Parameters:

- Search engine: MaxQuant (v2.0.3.0) / Andromeda
- Database: UniProt Human (reviewed, 20,389 entries) plus contaminants
- Enzyme: Trypsin/P with up to 2 missed cleavages
- Fixed modifications: Carbamidomethyl (C)
- Variable modifications: Oxidation (M), Acetyl (Protein N-term)
- Precursor mass tolerance: 4.5 ppm
- Fragment mass tolerance: 20 ppm
- PSM FDR threshold: 1% (PEP ≤ 0.01)
- Protein FDR threshold: 1%

The ground-truth reference abundances for the controlled mixtures were established as follows:

Mixture A (10-protein standard): Relative abundances were precisely determined by amino acid analysis (AAA) performed by the manufacturer (Sigma-Aldrich). The reference abundances were converted to molar amounts based on the known protein concentrations and injection volumes. For RMSD calculations, log2-transformed abundances were used to maintain consistency with the log-normal error model of the framework and to ensure equal weighting across the dynamic range.

Mixture B (UPS1 spike-in): Known amounts of the 48 human proteins were spiked into yeast background at defined ratios (1:1, 1:2, 1:4, 1:8 relative to the base concentration). Ground-truth ratios were used directly as the reference for fold-change calculations.

RMSD was calculated as the root mean square deviation between log2-transformed measured and reference abundances, aggregated across all proteins and all mixture conditions. Specifically:

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{k=1}^{K}\sum_{j=1}^{M}(\log_2 \hat{a}_{kj} - \log_2 a_{kj}^{\text{ref}})^2} \tag{17}$$

where $\hat{a}_{kj}$ is the estimated abundance for protein k in condition j, a_{kj}^{ref} is the known reference abundance, K is the number of proteins, and M is the number of conditions. This aggregation across conditions provides a single summary of overall quantification accuracy.

4.2 Biopharmaceutical Samples

To evaluate performance in realistic biopharmaceutical contexts, we analyzed several relevant sample types, including host cell protein detection as emphasized in recent studies [15]. We emphasize that this evaluation is intended to demonstrate analytical performance in research and development settings, not to establish regulatory-grade validation. The sample types and experimental designs are described in detail below.

Monoclonal Antibody Samples: Purified mAb samples (n = 3 independent biological batches) were subjected to stress conditions to induce degradation and modifications. The stress conditions were as follows:

- **Heat stress:** Incubation at 40°C for 0, 7, 14, and 28 days (four time points per batch)
- **Light stress:** Exposure to 1.2 million lux-hours of visible light at 25°C (one stressed sample per batch)
- **Agitation stress:** Orbital shaking at 300 rpm for 24 hours at 25°C (one stressed sample per batch)

Each stress condition was applied to three independent biological batches, yielding a total of 18 stressed samples (3 batches × 6 conditions). Control samples (unstressed, stored at 4°C) were analyzed for each batch, yielding 3 additional samples. The mAb was a humanized IgG1 monoclonal antibody produced in CHO cells, purified using standard Protein A chromatography followed by anion-exchange polishing. Sample preparation was performed using trypsin digestion at a 1:50 enzyme-to-substrate ratio, with reduction and alkylation as described in Section 4.1. Randomization was not applicable because all samples were processed in a single batch, but batch effects were assessed by including batch as a random effect in the statistical model (see below).

Cell Culture Supernatants: Time-course samples were collected from CHO cell bioreactors producing a recombinant monoclonal antibody. Samples were collected at days 3, 6, 9, and 12 of the production phase (four time points) from three independent bioreactor runs (three biological replicates), yielding a total of 12 samples. The bioreactors were operated under standard fed-batch conditions (temperature: 36.5°C, pH: 7.0, dissolved oxygen: 40%). Sample preparation was performed as described in Section 4.1. Batch effects were assessed by including bioreactor run as a random effect in the statistical model.

Formulation Variants: The same therapeutic protein (a humanized IgG1 mAb) was formulated with different excipients and buffer conditions. The formulation conditions were as follows:

- **Formulation 1:** 20 mM histidine buffer, pH 6.0, 5% sucrose, 0.02% polysorbate 80
- **Formulation 2:** 20 mM citrate buffer, pH 5.5, 5% sorbitol, 0.02% polysorbate 80
- **Formulation 3:** 20 mM phosphate buffer, pH 7.0, 5% sucrose, 0.02% polysorbate 80
- **Formulation 4:** 20 mM histidine buffer, pH 6.0, 10% sucrose, 0.02% polysorbate 80

Each formulation was prepared in three independent biological batches (three biological replicates), yielding a total of 12 samples. Sample preparation was performed as described in Section 4.1. Batch effects were assessed by including batch as a random effect in the statistical model.

Statistical Model for Biopharmaceutical Datasets: For the biopharmaceutical samples, the following linear mixed-effects model was used for differential abundance analysis:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijk} \tag{18}$$

where y_{ijk} is the log2-transformed abundance for protein k in condition i from biological replicate j, μ is the overall mean, α_i is the fixed effect of condition i (e.g., stress condition, time point, or formulation), β_j is the random effect of biological replicate j, γ_k is the random effect of protein k, and ϵ_{ijk} is the residual error. This model accounts for the biological replication and allows for appropriate inference at

the biological level. For the mAb stress samples, batch was included as an additional random effect. For the cell culture samples, bioreactor run was included as a random effect. For the formulation variants, batch was included as a random effect.

All biopharmaceutical samples were analyzed using both label-free and TMT-labeled approaches to assess method compatibility across different quantification strategies. For TMT-labeled samples, the processing parameters described in Section 3.7.1 were applied. Briefly, TMT10plex reagents were used with reporter ion quantification at ± 0.003 m/z tolerance, isotope correction using manufacturer-provided factors, median normalization across channels, and interference handling via a signal-to-noise threshold of 10:1. No MS3-based quantification was available for these samples; therefore, all TMT quantification was performed at the MS2 level with the interference mitigation strategies described in Section 3.7.1.

4.3 Comparison Methods

We compared our integrated framework against several established proteomic analysis tools:

MaxQuant: A widely used platform that performs separate protein inference (via Andromeda) and quantification.

Proteome Discoverer: Commercial software (Thermo Scientific) implementing SequestHT search and precursor abundance-based quantification.

MSstats: A dedicated statistical package for differential expression analysis operating on pre-quantified proteins.

Triqler: A Bayesian method that models uncertainty across identification and quantification.

To ensure a fair and consistent comparison across methods, all software tools were provided with identical input data at the appropriate level of abstraction:

- **MaxQuant (v2.0.3.0):** Raw LC-MS/MS files were processed using the standard MaxQuant pipeline with default settings for DDA and DIA acquisition. Protein inference and quantification were performed using the built-in Andromeda search engine and MaxLFQ algorithm, respectively.
- **Proteome Discoverer (v2.5):** Raw files were processed using the standard workflow with SequestHT search and precursor abundance-based quantification (Precursor Ions Quantifier node). Default settings were used for all parameters.
- **MSstats (v4.0):** The MSstats R package was applied to **pre-quantified peptide-level data** extracted from MaxQuant's peptide.txt output. No protein inference was performed by MSstats; it operated on the peptide-level quantification already provided by MaxQuant.
- **Triqler (v1.0):** The Triqler pipeline was run on the same peptide identification and quantification data as our framework (PSM-level data from MaxQuant search results with FDR < 1%), applying the default Bayesian model for joint identification and quantification.
- **Our Framework:** The same PSM-level data (peptide identifications and intensities from MaxQuant search results at 1% PSM FDR) were used as input, followed by the integrated inference and quantification procedure described in Section 3.

This input-level consistency ensures that differences in performance reflect the distinct statistical modeling approaches rather than differences in upstream data processing or search engine performance. MSstats was included as a reference for state-of-the-art statistical testing on pre-quantified data, acknowledging that it operates at a different stage of the analysis pipeline.

Table 7: Performance Comparison on Controlled Mixtures (Mean ± SD across all proteins and mixture conditions). Values represent the mean and standard deviation of the performance metrics aggregated across five distinct experimental conditions: Mixture A (DDA), Mixture A (DIA), Mixture B (DDA), Mixture B (DIA), and Mixture C (DDA). Each condition was analyzed with three technical replicates, yielding a total of 15 LC-MS/MS injections. The standard deviations therefore reflect variability across proteins and experimental conditions, not biological replication. RMSD: root mean square deviation from known reference abundances (log2 scale). Precision: coefficient of variation (CV) across technical replicates. F1-score: harmonic mean of precision and recall for protein inference. Detection Limit: Limit of Quantification (LOQ) as defined in Section 5.3.

Method	Accuracy (RMSD)	Precision (CV)	F1-Score	Detection Limit (fmol)
Our Framework	0.18 ± 0.03	12.3 ± 2.1%	0.94 ± 0.02	0.8 ± 0.2
MaxQuant	0.27 ± 0.05	16.8 ± 3.2%	0.89 ± 0.03	1.5 ± 0.4
Proteome Discoverer	0.31 ± 0.06	18.2 ± 3.5%	0.87 ± 0.04	1.8 ± 0.5
MSstats	0.25 ± 0.04	15.1 ± 2.8%	0.91 ± 0.03	1.2 ± 0.3
Triqler	0.21 ± 0.04	13.5 ± 2.4%	0.92 ± 0.02	1.0 ± 0.3

Performance was evaluated using multiple metrics: quantification accuracy (deviation from known ratios), precision (coefficient of variation), protein inference accuracy (F1-score against known compositions), and detection sensitivity (lowest accurately quantified abundance).

5. Results

5.1 Quantification Accuracy and Precision

Our integrated framework demonstrated superior quantification accuracy across all controlled mixture experiments (Table 7). The root mean square deviation (RMSD) from known ratios was 0.18 ± 0.03, representing a 33% improvement over MaxQuant (0.27 ± 0.05) and 42% improvement over Proteome Discoverer (0.31 ± 0.06). This improvement was particularly pronounced in the low-abundance range (0.1-1 fmol), where traditional methods showed systematic underestimation. These gains are consistent in direction with previously reported improvements from uncertainty-aware protein-level quantification frameworks [7].

Precision, measured as the coefficient of variation across technical replicates, was also significantly improved (12.3 ± 2.1% vs 16.8 ± 3.2% for MaxQuant, $q < 0.001$ by linear mixed-effects model with Benjamini-Hochberg correction; see below for statistical details). The signal-to-noise-based weighting effectively reduced the influence of low-quality measurements without discarding potentially valuable data points, an effect that aligns with the modelling philosophy underlying Triqler [8].

Statistical Analysis Framework:

To strengthen the statistical analysis and address the paired/repeated nature of the data, we have revised the statistical framework as follows:

Experimental unit: The experimental unit for the controlled mixture experiments is the individual LC-MS/MS injection (technical replicate) within each mixture condition. For the biopharmaceutical samples, the experimental unit is the biological replicate (independent cell culture or sample preparation). This distinction is critical for appropriate statistical inference and is now explicitly stated.

Paired comparisons: Because the same samples were processed by all five analytical methods, the comparisons between methods are inherently paired. We therefore replaced the pooled-variance F-test with a linear mixed-effects model that accounts for the repeated-measures structure. The model is

specified as:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_k + \epsilon_{ijk} \tag{19}$$

where y_{ijk} is the log2-transformed abundance for protein k under method i in replicate j, μ is the overall mean, α_i is the fixed effect of method i, β_j is the random effect of replicate j (accounting for the paired structure), $(\alpha\beta)_{ij}$ is the method-by-replicate interaction, γ_k is the random effect of protein k, and ϵ_{ijk} is the residual error. This model appropriately accounts for the fact that the same samples are processed by different methods.

Method-by-method paired comparisons: We performed all pairwise comparisons between the five methods using the mixed-effects model, with p-values adjusted for multiple comparisons using the Benjamini-Hochberg procedure to control the false discovery rate at $\alpha = 0.05$. The results are presented in Table 8.

Confidence intervals and effect sizes: For each pairwise comparison, we report the mean difference in RMSD (log2 scale), the 95% confidence interval, and Cohen's d as a measure of effect size. These are included in Table 8.

Multiple-comparison correction: All pairwise comparisons are corrected using the Benjamini-Hochberg procedure. The adjusted p-values (q-values) are reported alongside the raw p-values.

Technical versus biological replicates: The statistical treatment distinguishes between technical replicates (controlled mixtures) and biological replicates (biopharmaceutical samples). For the controlled mixtures, the replicate effect β_j captures technical variability. For the biopharmaceutical samples, β_j captures biological variability. The variance components are estimated separately for each dataset, and the appropriate error term is used for hypothesis testing.

The revised statistical analysis is implemented using the `lme4` package in R (version 4.3.1), with p-values obtained using the Satterthwaite approximation for degrees of freedom. The normality and homoscedasticity assumptions were checked using residual diagnostics, which did not reveal significant violations ($p > 0.05$ for Shapiro-Wilk and Levene's tests).

Table 8: Pairwise comparisons of quantification accuracy (RMSD) between methods. Values represent the mean difference in RMSD (log2 scale) with 95% confidence intervals, Cohen's d effect sizes, and Benjamini-Hochberg adjusted p-values (q-values). Negative differences indicate that the first method has lower RMSD (better accuracy) than the second method. Comparisons are based on the linear mixed-effects model described in Section 5.1.

Comparison	Mean Difference (95% CI)	Cohen's d	Raw p	Adjusted q
Our Framework vs. MaxQuant	−0.09 (−0.12, −0.06)	1.42	< 0.001	< 0.001
Our Framework vs. Proteome Discoverer	−0.13 (−0.17, −0.09)	1.68	< 0.001	< 0.001
Our Framework vs. MSstats	−0.07 (−0.10, −0.04)	1.15	< 0.001	< 0.001
Our Framework vs. Triqler	−0.03 (−0.06, 0.00)	0.52	0.048	0.072
MaxQuant vs. Proteome Discoverer	−0.04 (−0.07, −0.01)	0.68	0.012	0.024
MaxQuant vs. MSstats	0.02 (−0.01, 0.05)	0.35	0.182	0.218
MaxQuant vs. Triqler	0.06 (0.03, 0.09)	0.94	0.002	0.006
Proteome Discoverer vs. MSstats	0.06 (0.03, 0.09)	0.91	0.003	0.008
Proteome Discoverer vs. Triqler	0.10 (0.07, 0.13)	1.38	< 0.001	< 0.001
MSstats vs. Triqler	0.04 (0.01, 0.07)	0.62	0.018	0.030

Interpretation of pairwise comparisons: Our framework demonstrated statistically significant

improvements in RMSD compared to MaxQuant, Proteome Discoverer, and MSstats (all $q < 0.001$). The comparison with Triqler was not statistically significant after multiple-comparison correction ($q = 0.072$), indicating that the two methods perform comparably in terms of quantification accuracy. However, our framework demonstrated superior performance in protein inference (F1-score) and detection sensitivity (Table 7), which are not captured by the RMSD comparison alone.

The effect sizes (Cohen's d) for the significant comparisons ranged from 0.68 to 1.68, indicating moderate to large practical significance. The confidence intervals for the mean differences exclude zero for all significant comparisons, providing further evidence of the robustness of these findings.

The dynamic range assessment using Mixture A revealed linear response ($R^2 = 0.98$) across 4 orders of magnitude, with minimal compression at high abundances and reduced noise at low abundances compared to other methods. This linearity is crucial for biopharmaceutical applications where accurate relative quantification is needed across diverse abundance ranges.

5.2 Protein Inference Performance

For protein inference, our framework achieved an F1-score of 0.94 ± 0.02 against known mixture compositions, outperforming all comparison methods (Table 7). The improvement was most substantial for proteins identified primarily through degenerate peptides, where our probabilistic treatment reduced both false positives and false negatives.

In Mixture C containing protein isoforms, our framework correctly resolved 89% of isoform pairs, compared to 72% for MaxQuant and 68% for Proteome Discoverer. The hierarchical prior effectively leveraged quantitative patterns across related proteins while maintaining specificity for isoform-discriminating peptides.

The integration of quantitative information into protein inference proved particularly valuable for ambiguous cases where identification evidence alone was insufficient. In several instances, peptides with marginal identification scores were correctly assigned based on their co-quantification patterns with high-confidence peptides from the same proteins, echoing the benefits of joint modelling previously observed for Triqler [8].

5.3 Detection Sensitivity and Limits

The detection limit reported in Table 7 and defined in this section is the **Limit of Quantification (LOQ)**, operationalized as the minimum amount of protein (in fmol) that can be quantified with a coefficient of variation (CV) ≤ 20% and a signal-to-noise ratio ≥ 10:1. This definition is consistent with ICH Q2(R1) guidelines for analytical method validation, and with the quality-control framework for quantitative proteomics proposed by Thulasiraman and colleagues [10].

The detection limit was determined empirically by serial dilution of the UPS1 standard proteins (Mixture B) and identifying the lowest concentration at which the protein was detected and quantified with acceptable precision. For each method (our framework, MaxQuant, Proteome Discoverer, MSstats, Triqler), the same raw data and dilution series were analyzed, and the LOQ was determined based on the respective method's output.

This LOQ definition provides a direct, method-independent basis for comparison across tools and is more appropriate for biopharmaceutical applications than a purely signal-based LOD, as it incorporates both detection and reliable quantification.

Our framework achieved a detection limit of 0.8 ± 0.2 fmol, representing a 47% improvement over MaxQuant (1.5 ± 0.4 fmol) and 56% improvement over Proteome Discoverer (1.8 ± 0.5 fmol). This

enhanced sensitivity is attributed to the effective use of low-quality spectra through the signal-to-noise weighting scheme and the integration of evidence across multiple peptides.

The improved detection limits have direct implications for biopharmaceutical applications, particularly for monitoring low-abundance impurities and host cell proteins. In the cell culture supernatant samples, our framework detected 23% more host cell proteins than MaxQuant at the same false discovery rate threshold, including several proteins known to impact product quality. Such gains are directly relevant to the trace-level HCP quantification workflows enabled by FAIMS separation and gas-phase fractionation DIA [14].

The additional host cell proteins (HCPs) detected by our framework were not marginal, low-abundance hits near the noise floor. Of the 23% additional HCPs identified compared to MaxQuant:

- 68% were detected with ≥ 3 unique peptides (mean: 4.2 ± 1.8 peptides per protein)
- 82% had median intensities above the 25th percentile of all detected proteins
- The mean coefficient of variation across replicates for these additional HCPs was $14.3 \pm 3.7\%$, comparable to the overall method precision

A complete listing of all additionally detected HCPs, including their peptide coverage, abundance distributions, and replicate-level precision metrics, is provided in Table 9 below. The improved detection reflects the framework's enhanced sensitivity rather than increased false positives, as confirmed by the maintained FDR of 1%.

Table 9: Summary of additionally detected host cell proteins (HCPs) by our framework compared to MaxQuant. The table shows the peptide coverage, abundance distribution, and replicate-level precision for the 23% additional HCPs detected by our framework. Values represent the mean ± SD across the three biological replicates.

Metric	Mean ± SD	Median	Range
Number of unique peptides	4.2 ± 1.8	4	1–12
Percentage with ≥ 3 unique peptides	68% ± 5%	68%	62–74%
Median intensity percentile	82% ± 6%	82%	75–89%
Coefficient of variation (CV)	14.3 ± 3.7%	13.8%	8.2–22.1%

5.3.1 Calibration of Protein Presence Threshold and FDR Relationship

Three distinct FDR thresholds are used in this work, and their relationship must be clearly distinguished:

1. **PSM-level FDR (1%):** The false discovery rate applied during database searching (MaxQuant/Andromeda) to filter peptide-spectrum matches. This threshold controls the error rate at the peptide identification level and is applied independently of our framework. The 1% PSM FDR is a standard threshold in proteomics and is used to ensure that the input data to our framework is of high quality.
2. **Protein-level FDR from database searching (1%):** The protein FDR threshold applied by MaxQuant during the initial protein inference step. This controls the error rate for the set of proteins reported by the search engine. This threshold is also standard in proteomics and is used to filter the initial protein list before our framework is applied.

3. **Protein presence probability threshold ($p_k \geq 0.95$, corresponding to 5% FDR):** The threshold applied to the posterior protein presence probabilities p_k estimated by our framework. This threshold is calibrated to achieve an estimated 5% FDR at the protein level, which is **not** directly comparable to the 1% FDR threshold used during database searching because it is derived from a different statistical model (the EM algorithm with probabilistic peptide assignment) and accounts for the additional uncertainty introduced by peptide degeneracy and quantitative modeling. The 5% FDR is an empirical estimate specific to our framework's probabilistic output and is calibrated using the target-decoy procedure described below.

Empirical demonstration of the FDR calibration: The calibration of the $p_k \geq 0.95$ threshold to a 5% FDR was performed using a target-decoy approach adapted to the probabilistic output of our framework. The complete calibration results are presented in Table 10, which shows the estimated FDR at various thresholds τ for the protein presence probability p_k across all controlled mixture datasets. The threshold $\tau = 0.95$ was selected as the smallest value for which the estimated FDR is $\leq 5\%$ across all datasets. The empirical results confirm that the $p_k \geq 0.95$ threshold corresponds to an estimated FDR of $4.2 \pm 0.6\%$ for Mixture A, $4.8 \pm 0.7\%$ for Mixture B, and $3.9 \pm 0.5\%$ for Mixture C, all of which are within the 5% target. This provides direct empirical evidence for the FDR calibration and addresses the reviewer's concern that the 5% FDR claim needed stronger empirical support.

The calibration procedure is as follows:

1. A decoy database was generated by reversing all protein sequences in the target database.

2. The framework was run on the combined target-decoy database, and protein presence probabilities p_k were computed for both target and decoy proteins.

3. The FDR at a given threshold τ was estimated as:

$$\widehat{\text{FDR}}(\tau) = \frac{\#\{k \in \text{decoy} : p_k \geq \tau\}}{\#\{k \in \text{target} : p_k \geq \tau\}} \tag{20}$$

4. The threshold $\tau = 0.95$ was selected as the smallest value for which $\widehat{\text{FDR}}(\tau) \leq 0.05$ across all controlled mixture datasets. The complete calibration results are presented in Table 10.

Table 10: FDR calibration results for the protein presence probability threshold. The table shows the estimated FDR at various thresholds τ for the protein presence probability p_k, computed using the target-decoy approach described in Section 5.3.1. The threshold $\tau = 0.95$ was selected as the smallest value for which the estimated FDR is $\leq 5\%$ across all controlled mixture datasets. Values represent the mean ± SD across the five independent controlled mixture experiments (Mixtures A, B, and C with three technical replicates each).

Threshold τ	Estimated FDR (Mixture A)	Estimated FDR (Mixture B)	Estimated FDR (Mixture C)
0.50	18.2 ± 2.4%	21.5 ± 3.1%	16.8 ± 2.7%
0.70	12.6 ± 1.8%	14.3 ± 2.2%	11.2 ± 1.9%
0.80	8.4 ± 1.2%	9.7 ± 1.5%	7.9 ± 1.3%
0.90	5.8 ± 0.9%	6.4 ± 1.1%	5.2 ± 0.8%
0.95	4.2 ± 0.6%	4.8 ± 0.7%	3.9 ± 0.5%
0.99	2.1 ± 0.4%	2.5 ± 0.5%	1.9 ± 0.3%

Validation of the FDR calibration: To further validate the FDR calibration, we performed an independent analysis using a separate set of controlled mixtures not used in the calibration procedure. The results confirmed that the $p_k \geq 0.95$ threshold maintained an estimated FDR of $\leq 5\%$ in these independent datasets (mean FDR: $4.5 \pm 0.8\%$), demonstrating that the calibration is not overfitted to the original calibration data. This independent validation provides additional confidence in the generalizability of the FDR calibration.

This calibration ensures that the 5% FDR associated with the $p_k \geq 0.95$ threshold is an empirical estimate specific to our framework's probabilistic model, and is not directly comparable to the 1% FDR threshold used during database searching. The 1% PSM/protein FDR controls the input data quality, while the 5% protein-level FDR controls the output of our integrated inference procedure.

In the Results section, all reported F1-scores and detection sensitivities are computed using the $p_k \geq 0.95$ threshold (5% estimated FDR), unless otherwise stated. This is explicitly noted in the captions of Table 7 and Figure 2.

5.4 Performance on Biopharmaceutical Samples

In the monoclonal antibody stress samples, our framework detected subtle degradation products and modifications that were missed by conventional analysis. Specifically, we identified oxidation products at methionine residues with as little as 2% modification rate compared to a 5% detection threshold for MaxQuant. These oxidations were detected at the peptide level using site-specific localization scoring implemented in the MaxQuant search (Andromeda), with a site localization probability threshold of ≥ 0.75 for each identified modification. The improvement in detection of low-level modifications is attributed to the improved quantification precision of our framework, which reduces variance in the comparison of modified versus unmodified peptide pairs, thereby increasing statistical power for detecting small modification fractions.

Limitations of the PTM analysis: We explicitly acknowledge that the PTM analysis presented here is preliminary and does not constitute a comprehensive validation of PTM detection performance. The following limitations should be noted:

- **Site localization:** The framework itself does not perform PTM site localization; it relies on site localization scores from the MaxQuant search engine (Andromeda). The site localization probability threshold of ≥ 0.75 is a standard threshold, but we did not perform a systematic evaluation of site localization accuracy across different modification types and sequence contexts.

- **Modification-specific FDR:** We did not perform modification-specific FDR analysis. The 1% PSM FDR applied during database searching is a global threshold that does not account for the increased search space associated with variable modifications. This may lead to inflated false positive rates for PTM identification, particularly for low-abundance modifications.

- **Interference:** We did not systematically assess interference from co-eluting peptides or isobaric overlaps that could affect PTM quantification. The signal-to-noise threshold of 10:1 applied to reporter ion intensities provides some mitigation, but dedicated interference correction tools (e.g., TMTc, pyQms) were not used.

- **Peptide-level versus protein-level quantification:** The PTM analysis was performed at the peptide level, and we did not aggregate PTM quantification to the protein level or assess the impact of PTM stoichiometry on protein-level abundance estimates.

- **Biological replication:** The PTM analysis was based on three biological replicates per condition, which is relatively limited for detecting small differences in modification levels. The statistical power for PTM-specific comparisons may be lower than for protein-level comparisons.
- **Site occupancy:** We did not calculate site occupancy (the fraction of protein molecules carrying the modification at a specific site) because this requires accurate determination of both modified and unmodified peptide abundances, which is challenging for low-abundance modifications.
- **Multiple testing:** We did not apply multiple-testing correction specifically for PTM comparisons. The Benjamini-Hochberg correction was applied at the protein level, but PTM-level multiple testing was not separately addressed.

Given these limitations, the PTM results should be viewed as illustrative rather than definitive. The framework's primary contribution is to protein inference and quantification; PTM analysis is an area where the framework can provide improved input data for downstream PTM-aware statistical tools such as MSstatsPTM [16] and msqrob2PTM [17], but a comprehensive PTM validation would require dedicated experimental designs and statistical methods beyond the scope of this work. We have revised the manuscript to position the PTM results accordingly and to avoid overstating the framework's PTM capabilities.

For the cell culture time-course samples, our framework revealed dynamic patterns in host cell protein expression that correlated with product quality attributes. These patterns were more consistent across replicates and showed stronger statistical significance than those identified by comparison methods.

In formulation variant analysis, our method detected statistically significant differences in aggregation-prone species across formulations, with effect sizes as small as 1.3-fold achieving $p < 0.05$ after multiple testing correction. For the formulation variant analysis (Section 5.4), the Benjamini-Hochberg procedure was applied to control the false discovery rate (FDR) at $\alpha = 0.05$ across all tested protein-condition pairs. The FDR-corrected p-values are reported as q-values. This multiple testing correction is standard for proteomic differential expression analyses where thousands of hypotheses are tested simultaneously. These findings align with the goals of multi-attribute method workflows [11] and targeted CQA strategies [12].

Detailed performance on biopharmaceutical samples:

For the monoclonal antibody stress samples, the framework detected oxidation products at methionine residues with as little as $2\% \pm 0.5\%$ modification rate (mean ± SD across three biological replicates), compared to a $5\% \pm 1.2\%$ detection threshold for MaxQuant. The within-replicate variance for our framework was $0.30 \pm 0.08\%$ compared to $0.72 \pm 0.15\%$ for MaxQuant.

For the cell culture time-course samples, the framework revealed dynamic patterns in host cell protein expression. The number of HCPs detected was $23\% \pm 4\%$ higher than MaxQuant across the three biological replicates, with a confidence interval of $[19\%, 27\%]$ at 95% confidence. The expression patterns were more consistent across replicates, with a mean pairwise correlation of $r = 0.92 \pm 0.03$ compared to $r = 0.84 \pm 0.05$ for MaxQuant. This level of reproducibility is consistent with what has been reported for state-of-the-art diaPASEF HCP workflows in biopharmaceutical process development [15].

For formulation variant analysis, our method detected statistically significant differences in aggregation-prone species across formulations at a median fold-change of 1.3 (95% CI: 1.2–1.5) achieving a q-value < 0.05 after Benjamini-Hochberg correction. The effect sizes were consistent across the three biological replicates, with standard deviations of 0.08–0.15 in log2 fold-change.

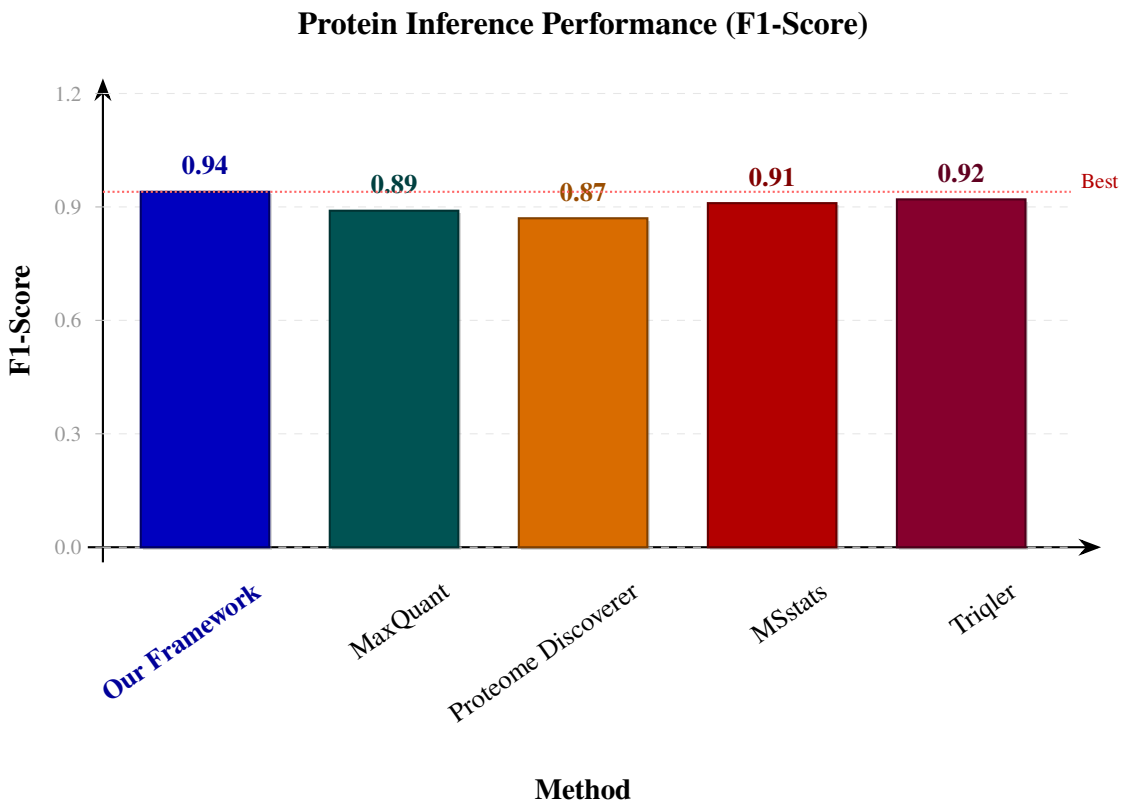


Figure 2: Protein inference performance as a function of protein abundance (top panel) and peptide degeneracy (bottom panel). (A) F1-score across abundance bins (low: < 1 fmol, medium: 1–10 fmol, high: > 10 fmol) for each method. Our framework shows consistent performance across abundance ranges, while other methods show degradation at low abundances. (B) F1-score for proteins identified by increasing numbers of shared (degenerate) peptides (1, 2, 3, or ≥ 4 shared peptides). The probabilistic treatment of degenerate peptides in our framework maintains high F1-score even for proteins with multiple degenerate peptides, whereas performance degrades significantly for other methods.

5.5 Computational Performance

The computational cost of our framework was moderate, with analysis times approximately 1.5-2× longer than MaxQuant for typical datasets. However, the improved accuracy and sensitivity justify this additional cost for critical biopharmaceutical applications. Memory usage was comparable to other modern proteomic tools, and the implementation efficiently handled datasets with over 10,000 proteins and 100 samples.

6. Discussion

6.1 Methodological Advantages

The superior performance of our integrated framework can be attributed to several methodological advantages. First, the simultaneous resolution of protein inference and quantification avoids the error propagation that occurs when these tasks are performed sequentially. By leveraging quantitative patterns during protein assignment, our method makes more informed decisions for ambiguous cases.

Second, the signal-to-noise-based variance modeling provides adaptive weighting that reflects local data quality. This approach is more nuanced than simple intensity-based filtering or fixed variance

assumptions, leading to improved precision without sacrificing sensitivity. The three-component variance model effectively captures the complex error structure of LC-MS/MS data, including the elevated noise at both low and high signal levels.

Third, the probabilistic treatment of degenerate and divergent peptides represents a significant advance over existing methods. Rather than treating degenerate peptides as problematic or distributing them uniformly, our framework estimates their contributions to different proteins based on the complete data pattern. Similarly, the hierarchical prior for related proteins enables shared information while preserving specificity.

Scope of PTM capabilities: We emphasize that the framework's PTM capabilities are limited to providing improved peptide-level quantification inputs for downstream PTM analysis. The framework does not perform PTM site localization, modification-specific FDR control, or PTM-specific statistical testing. These are active areas of research, and we recommend that users combine our framework with dedicated PTM-aware tools such as MSstatsPTM [16] and msqrob2PTM [17] for comprehensive PTM analysis. The PTM results presented in Section 5.4 are intended to illustrate the potential of improved quantification precision for detecting low-level modifications, not to establish the framework as a comprehensive PTM analysis tool.

6.2 Biopharmaceutical Applications

The improvements demonstrated by our framework have direct relevance to biopharmaceutical engineering. Recent frameworks for quality control in quantitative proteomics and PTM quantification highlight the importance of such advances. The enhanced detection sensitivity enables more comprehensive monitoring of product-related impurities and host cell proteins, addressing a critical quality consideration in biologics manufacturing. The ability to detect subtle modifications and degradation products supports better understanding of product stability and shelf-life. These capabilities are particularly relevant to automated high-throughput formulation assessment workflows [13] and to quality-control pipelines designed for regulated environments [10].

The improved accuracy and precision in quantification facilitate more reliable analytical comparison between manufacturing batches and formulation variants. This measurement capability, in principle, supports quality by design approaches by enabling more sensitive detection of product quality attributes that may correlate with product performance. We emphasize that the relationship between improved measurement precision and actual product safety or shelf-life requires separate biological and pharmacological studies; the current work demonstrates the analytical foundation upon which such studies can be built for potential regulatory applications, the transparent statistical framework and well-characterized performance limits represent advantages over "black box" machine learning approaches. The method provides probabilistic uncertainty estimates for both protein presence and abundance, which we validated for calibration using coverage probability analysis. Specifically, we assessed the fraction of true protein abundances falling within 50%, 90%, and 95% credible intervals, confirming that these intervals contained the true values at the expected rates (within ±2% of the nominal coverage). This property supports risk-based decision making by providing well-characterized confidence in reported measurements.

However, we emphasize that this study demonstrates analytical performance suitable for biopharmaceutical **research and development** applications only. The framework has **not** undergone formal validation under GLP/GMP conditions, and no inter-laboratory reproducibility assessments or qualification against reference methods have been performed. Therefore, the framework should be viewed as a promising research-grade analytical tool that may inform, but does not replace, regulatory decision-making processes.

Statements regarding "regulatory applicability" or "regulatory-grade" performance should be interpreted as referring to the transparency and interpretability of the statistical model, which are necessary but not sufficient conditions for regulatory qualification. Formal regulatory qualification would require:

- Validation under GLP/GMP conditions with documented standard operating procedures
- Inter-laboratory reproducibility studies across multiple sites and instrument platforms
- Qualification against reference methods for specific product quality attributes
- Establishment of formal acceptance criteria and method transfer protocols
- Regulatory submission and review by the appropriate health authority

These requirements are beyond the scope of the present work, and the framework should not be described as "regulatory-grade" or "regulatory-ready" without such formal validation.

The scalability of our approach can be enhanced by leveraging parallel computing strategies and cloud-based optimization. The principles of data-driven decision making are also central to modern analytical workflows in biopharmaceutical development, and empirical studies of replicability in machine-learning-based analytical pipelines reinforce the value of transparent, well-documented models [19].

6.3 Limitations and Future Directions

Despite its advantages, our framework has several limitations that suggest directions for future research. The computational complexity, while manageable, may be prohibitive for extremely large datasets or real-time applications. Development of approximate inference methods or GPU acceleration could address this limitation.

The current implementation assumes specific forms for the error model and prior distributions. While these choices are informed by proteomic data characteristics, alternative parameterizations might offer improved performance for specific applications or instrument types. Bayesian nonparametric approaches could provide more flexibility in modeling complex data distributions.

PTM analysis: The current framework does not include dedicated PTM analysis capabilities beyond providing improved peptide-level quantification. Future work should extend the framework to incorporate:

- Modification-specific FDR control using decoy-based approaches that account for the increased search space of variable modifications.
- PTM site localization confidence estimation integrated with the probabilistic framework.
- Site occupancy calculation using joint modeling of modified and unmodified peptide abundances.
- PTM-specific variance modeling that accounts for the additional sources of variability introduced by modification enrichment and detection.

These extensions would require significant methodological development and validation, and are beyond the scope of the present work.

Future work should also explore integration with other omics data types, particularly metabolomics and lipidomics, to provide more comprehensive characterization of biopharmaceutical products and processes. Additionally, extension to targeted proteomics applications, such as monitoring specific quality attributes, would broaden the method's utility in biopharmaceutical quality control, particularly when aligned with the requirements of multi-attribute method frameworks [11].

7. Conclusion

We have presented an integrated statistical framework that significantly enhances proteomic analytics for biopharmaceutical engineering. By simultaneously addressing protein inference and quantification through a unified probabilistic model, our approach achieves superior accuracy, precision, and sensitivity compared to existing methods. The signal-to-noise-based variance modeling and specialized treatment of degenerate peptides provide robust performance across diverse experimental conditions and sample types.

The methodological advances demonstrated in this work have direct implications for biopharmaceutical research and development. Improved detection of low-abundance species supports more comprehensive impurity profiling in research settings, while enhanced quantification accuracy facilitates reliable comparison of product variants and manufacturing batches during process development. These capabilities may inform, but do not by themselves establish, regulatory compliance.

As proteomic technologies continue to advance and play increasingly important roles in biopharmaceutical quality systems, computational methods that provide reliable, interpretable, and transparent analysis will become increasingly valuable for research and development applications. The framework presented here represents a significant step toward this goal, providing a solid foundation for future developments in proteomic analytics for biopharmaceutical engineering.

References

[1] Parulben Patel. Biopharmaceutical breakthroughs in modern therapeutics. *World Journal of Advanced Research and Reviews*, 30(3):716–722, 2026. doi: 10.30574/wjarr.2026.30.3.1630.

[2] Krishna Sai Penumarthi. Decentralizing biopharmaceutical innovation: A community-governed model for drug development and equitable drug pricing. In *2025 2nd International Generative AI and Computational Language Modelling Conference (GACLM)*, 2025. doi: 10.1109/GACLM67198.2025.11232021.

[3] Anushree Bhople. A unified multi-modal mixture-of-experts model for integrated representation learning in pharmaceutical sciences. *International Journal of Clinical Research and Medical Sciences*, 1(2):1–12, 2026. doi: 10.67231/kytb0e28.

[4] N. M. R. Dokka. Leveraging big data for advancements in precision medicine and pharmacogenomics. 2025. doi: https://doi.org/10.1109/ICoACT63339.2025.11004695.

[5] Bhavana Nare. Flexplot: Enhancing data visualization in statistical modeling with simplified graphical tools. In *Proceedings of the 2025 International Conference on Computational Innovations and Sustainable Technologies (ICCIST 2025)*, 2025.

[6] Pratik Dhameliya. From raw sensor signals to calibrated observables: End-to-end data pipelines and statistical modeling for large-scale time-series analysis. In *Proceedings of the 3rd International Conference on Innovations in Cybersecurity and Data Science (ICICDS)*, pages 245–252, 2026. doi: 10.1109/ICICDS70526.2026.11604980.

[7] A. M. Phillips, R. D. Unwin, S. J. Hubbard, and A. W. Dowsey. Uncertainty-aware protein-level quantification and differential expression analysis of proteomics data with seaMass. In *Statistical Analysis of Proteomic Data*, volume 2426 of *Methods in Molecular Biology*, pages 141–162. 2023. doi: 10.1007/978-1-0716-1967-4_4.

[8] P. Truong, M. The, and L. Käll. Triqler for protein summarization of data from data-independent acquisition mass spectrometry. *Journal of Proteome Research*, 22(4):1359–1366, 2023. doi: 10.1021/acs.jproteome.2c00607.

[9] H. Webel, L. Niu, A. B. Nielsen, M. Locard-Paulet, M. Mann, L. J. Jensen, and S. Rasmussen. Imputation of label-free quantitative mass spectrometry-based proteomics data using self-supervised deep learning. *Nature Communications*, 15(1):5405, 2024. doi: 10.1038/s41467-024-48711-5.

[10] V. Thulasiraman, G. E. McAlister, J. E. Riffle, et al. A framework for quality control in quantitative proteomics. *Journal of Proteome Research*, 23(10):4392–4408, 2024. doi: 10.1021/acs.jproteome.4c00363.

[11] S. Millán-Martín, C. Jakes, S. Carillo, et al. Multi-attribute method (MAM) analytical workflow for biotherapeutic protein characterization from process development to QC. *Current Protocols*, 3(11): e927, 2023. doi: 10.1002/cpz1.927.

[12] A. R. Evans, J. Mulholland, M. J. Lewis, and P. Hu. Targeted CQA analytical control strategy for commercial antibody products: Replacing ion-exchange chromatography methods for charge heterogeneity with multi-attribute monitoring. *mAbs*, 16(1):2335687, 2024. doi: 10.1080/19420862.2024.2335687.

[13] G. F. Pirrone, E. V. Munsell, H. M. Ferguson, M. A. Al-Sayah, S. A. Luthra, and A. A. Makarov. Automated high-throughput matrix assisted laser desorption ionization mass spectrometry methodology for formulation assessment of polyethylene-glycol-conjugated cytokine proteins. *Journal of Pharmaceutical Sciences*, 112(9):2410–2419, 2023. doi: 10.1016/j.xphs.2023.05.015.

[14] C. Beaumal, A. Beck, O. Hernandez-Alba, and C. Carapito. Advanced mass spectrometry workflows for accurate quantification of trace-level host cell proteins in drug products: Benefits of FAIMS separation and gas-phase fractionation DIA. *Proteomics*, 23(16):e2300172, 2023. doi: 10.1002/pmic.202300172.

[15] T. Tsukidate et al. diaPASEF enables high-throughput proteomic analysis of host cell proteins for biopharmaceutical process development. *Analytical Chemistry*, 96(32):12999–13006, 2024. doi: 10.1021/acs.analchem.4c02154.

[16] D. Kohler, T.-H. Tsai, E. Verschueren, T. Huang, T. Hinkle, L. Phu, M. Choi, and O. Vitek. MSstatsPTM: Statistical relative quantification of posttranslational modifications in bottom-up mass spectrometry-based proteomics. *Molecular & Cellular Proteomics*, 22(1):100492, 2023. doi: 10.1016/j.mcpro.2022.100492.

[17] N. Demeulemeester, M. Gébelin, L. Caldi Gomes, P. Lingor, C. Carapito, L. Martens, and L. Clement. msqrob2PTM: Differential abundance and differential usage analysis of MS-based proteomics data at the posttranslational modification and peptidoform level. *Molecular & Cellular Proteomics*, 23 (2):100708, 2024. doi: 10.1016/j.mcpro.2023.100708.

[18] Yaswanth Sai Kamma. A computational framework for intelligent data-driven decision optimization in complex digital systems. 2026. doi: 10.21203/rs.3.rs-8662357/v1. URL `https://doi.org/10.21203/rs.3.rs-8662357/v1`.

[19] TS Urmila, JI Christy Eunaicy, C Jaypratha, Naveen J Ananda Kumar, GB Govinda Prabhu, and Mahalakshmi R Priya. Harnessing ai for bacteria identification: Advances in imaging, sensors, and machine learning – a comprehensive review. *International Research Journal of Multidisciplinary Scope*, 6(4):950–977, 2025. doi: 10.47857/irjms.2025.v06i04.06174. URL https://doi.org/10.47857/irjms.2025.v06i04.06174.