Tesorai Platform:

AI-enhanced analysis of mass spectrometry data yields higher identifications and faster insights with an intuitive chat interface for DDA and DIA datasets.

Abstract

Mass spectrometry (MS)-based proteomics is a powerful tool for high-throughput protein identification and quantification. However, as data volumes increase, there is a growing need for: 1. Accurate and robust search algorithms that avoid overfitting and provide reliable false discovery rate (FDR) estimates. 2. Scalable and user-friendly platforms capable of processing thousands of files in hours, not weeks. 3. Versatile and accessible solutions for downstream analyses that are easy to use for all, regardless of computational experience.

Tesorai offers a comprehensive platform comprising two key solutions: Tesorai Search and Tesorai Chat. Tesorai Search directly addresses the bottlenecks in MS-based proteomics by combining high identification accuracy and sensitivity with a scalable, low-barrier interface. Early user feedback highlights significant time savings, often reducing processing time by several days to weeks per dataset. Tesorai Chat leverages recent advancements in AI to empower users, even those with limited computational experience, to process raw MS data and conduct differential analyses within an hour using best-in-class tools.

www.tesorai.com info@tesorai.com @tesorai_bio @tesorai

Introduction

Extracting insights from mass spectrometry (MS) proteomics data typically requires three stages of analysis (Figure 1):

- **Primary analysis –** Identification and quantification of peptidoforms/proteoforms using MS search algorithms.
- **Secondary analysis –** Selection of proteins or peptides relevant to the hypothesis (e.g., through differential expression analysis).
- **Tertiary analysis –** Interpretation of results within a biological context, for example via pathway enrichment or network analyses.

Currently, performing all three levels of proteomics analysis remains challenging, especially for scientists without dedicated coding expertise or support. Most laboratories rely on computational biologists to assemble and maintain complex pipelines stitched together from multiple tools. These workflows are often difficult to reproduce, cumbersome to maintain, and inaccessible to scientists less experienced in programming. Additionally, most MS analysis software, especially search algorithms, was originally designed for individual workstations and scales poorly with the large data volumes produced by modern MS instruments, which now generate an order of magnitude more data per sample than previous generations.

As studies become larger and analytical throughput becomes increasingly important, there is an urgent need for an integrated platform that can deliver end-to-end analysis directly from raw data. Tesorai solves this need by offering a scalable, cloud-based software platform that combines best-in-class search algorithms with an intuitive interface designed for scientists, including those across levels of computational expertise. Tesorai's platform integrates two innovative components (Figure 1):

1. **Tesorai Search:** A next-generation search engine for data-dependent acquisition (DDA) and data-independent acquisition (DIA) acquisition modes that redefines the fundamentals of peptide and protein identification, quantification, and scalability.
2. **Tesorai Chat:** A natural language-based interface for QC, secondary and tertiary analyses, data visualization, and collaboration.

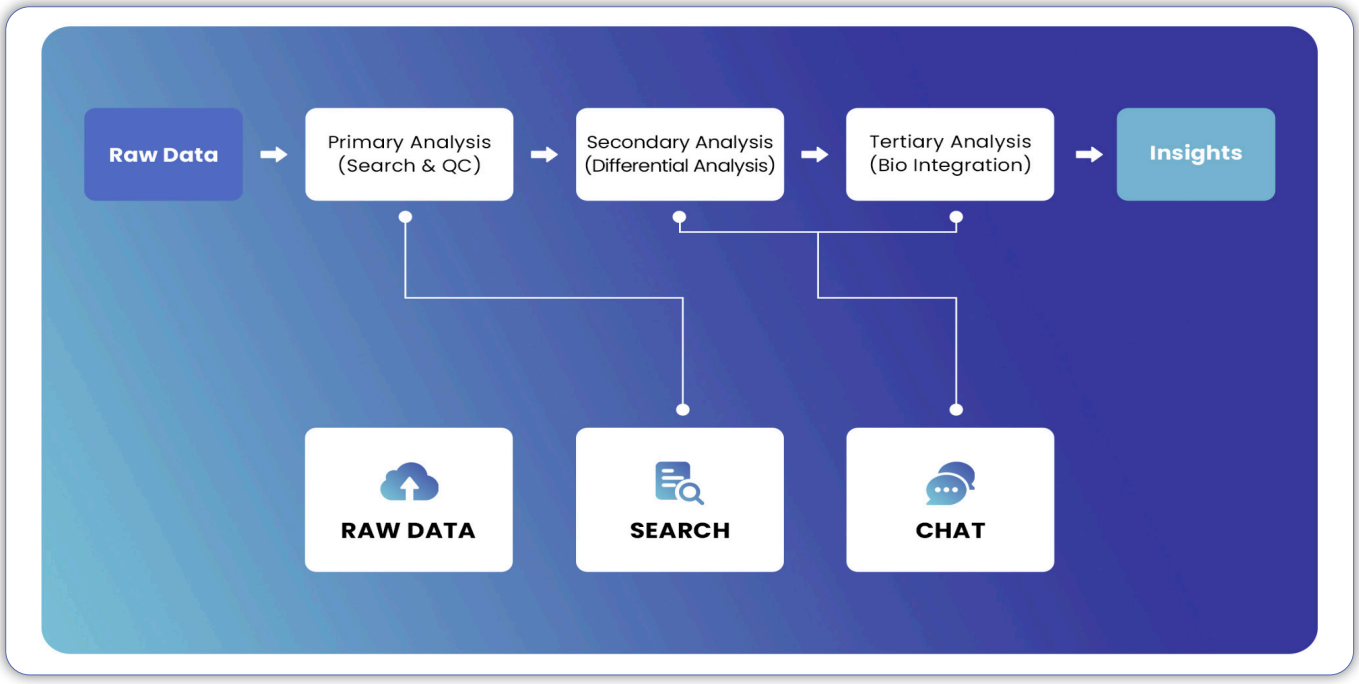


Figure 1: An overview schematic highlighting the three stages of common data workflows and how Tesorai platform facilitates them.

In order to assess the capabilities and effectiveness of the search algorithm and the integrated secondary and tertiary analysis, we investigated two distinct datasets:

1. A dia-PASEF tri-species dataset containing two groups of samples, each a triple-proteome mixture. We highlight high identification rates, accurate quantification, and easy visualization.
2. A dda-PASEF immunopeptidomics dataset. We highlight the ability to detect and deeply analyze potentially therapeutically relevant immunopeptides through end-to-end analysis, efficiently and with minimal hands-on time.

How does Tesorai work?

Step 1: Primary analyses - peptide and protein identification & quantification, and QC analyses

The first and most critical stage in the proteomics data workflow is primary analysis, where raw spectra are transformed into biologically meaningful and interpretable units, namely peptides and proteins. This stage underpins every subsequent step of discovery, as the accuracy, sensitivity, and reproducibility of identifications and quantifications directly determine the reliability of downstream analyses.

Tesorai Search is a next-generation search engine built on a novel, pretrained AI model that directly learns the complex relationship between the full spectra and associated peptide sequence, without relying on decoys. This simplified workflow requires only a few parameters, making it easy to use while performing robustly across DDA and DIA acquisition modes and data from different instruments (Burq et al. 2024). Designed to meet the needs of modern proteomics, Tesorai Search combines ease of use with state-of-the-art performance.

tesorai

Best-in-Class Specificity and Sensitivity

In peptide-spectrum matching (PSM), false positives are a persistent challenge. Conventional search engines often depend on simplified scoring functions that use only a subset of fragment ions, which can allow unrelated peptides to appear as plausible matches and inflate identifications. More recent machine learning–based rescoring tools add hundreds of engineered features and rely on decoy-based classifiers trained on the fly using the user's own data. While this approach can improve sensitivity, it also increases the risk of overfitting: the model may adapt too closely to dataset-specific noise or decoy-generation quirks, leading to underestimated false discovery rates (FDR) and reduced reproducibility across experiments. This problem has been observed in several prominent search engines for analysis of data from DIA (Wenn et al. 2025).

Tesorai Search improves specificity (i.e., accuracy of identifications) by design. The model is pretrained once on hundreds of millions of real PSMs and is never retrained on user data. This avoids dataset-specific overfitting while ensuring broad generalizability. Standard target-decoy competition is used purely for FDR estimation and not for model fine-tuning or training, yielding more reliable error control. The result is fewer false identifications, clearer separation between true and false PSMs, and more confident peptide calls, with Tesorai Search consistently recovering 12–68% more peptides than other commonly used search tools (Burq et al. 2024; Li et al. 2025). For example, applying the entrapment analysis to the dda-PASEF immunopeptidomics dataset, Tesorai Search demonstrates strong FDR control (Figure 2).

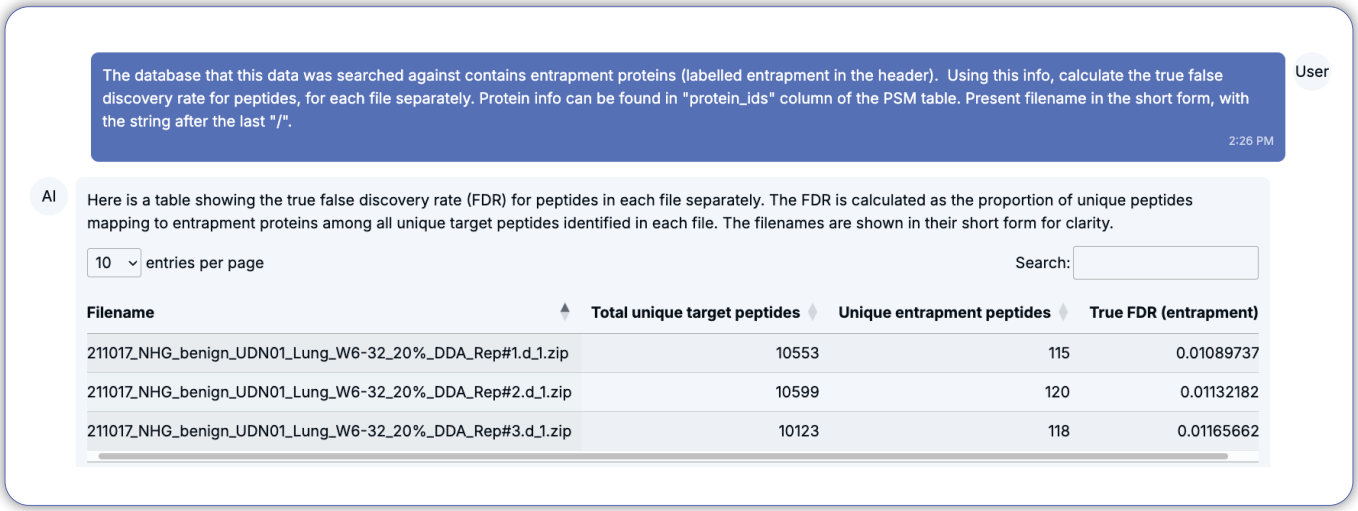


Figure 2: Tesorai Search effectively controls the FDR, even with difficult unspecific digests. Lung immunopeptidomic samples (PXD03878, Hoenisch Gravel 2023) were searched against an entrapment database containing the Human Canonical proteins as well as 5 entrapment proteins for every human protein as well as generated decoy proteins. Using the chat function, Tesorai when prompted can calculate the true FDR based on this entrapment experiment. As indicated, only 1% of entrapped peptides pass the standard 1% FDR filter, demonstrating Tesorai Search's robust FDR control.

Simplicity: One Model for All Workflows

Tesorai Search employs a **“one model to rule them all”** paradigm:

- Compatible with diverse instruments, acquisition approaches, fragmentation methods, and digestion strategies, including Bruker’s timsTOF instruments.
- Requires minimal parameter setup, lowering the barrier to entry for non-specialists.
- Facilitates uniform workflows that reduce user error and improve reproducibility across laboratories.

This simplicity enables high-quality analyses with minimal hands-on time and no coding required, benefiting both experimental scientists and bioinformaticians (Figure 3).

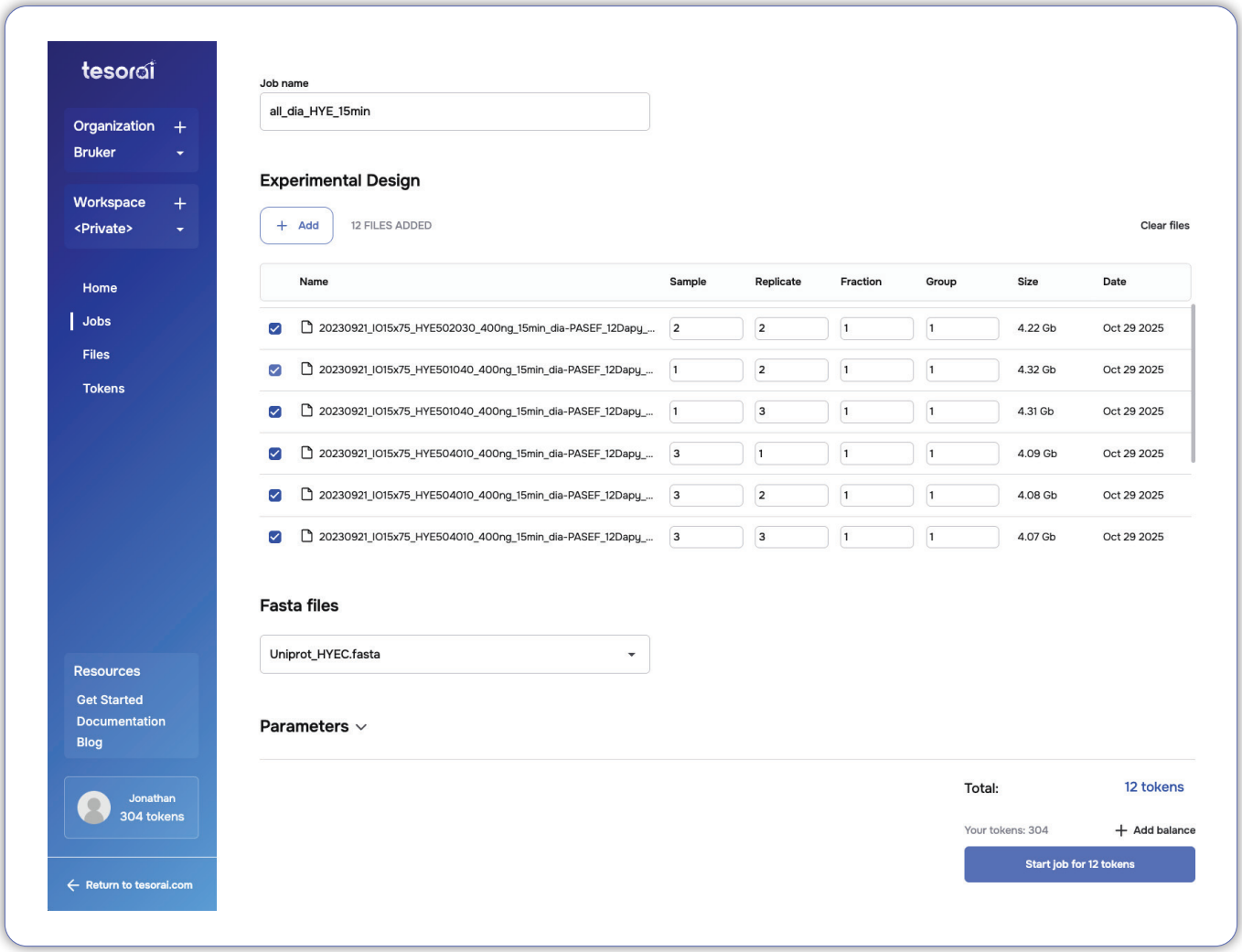


Figure 3: This screenshot shows the Tesorai Search job setup page. Its intuitive design allows for efficient search job configuration, boosting platform efficiency and productivity.

Scalability, especially with large timsTOF datasets

Tesorai Search is built to handle modern data volumes, processing thousands of large files within hours through a distributed, cloud-native infrastructure. With scalability as a core principle, it transforms primary analysis from a bottleneck into a streamlined, high-throughput process that's ready for the era of petabyte-sized datasets (Figure 4).

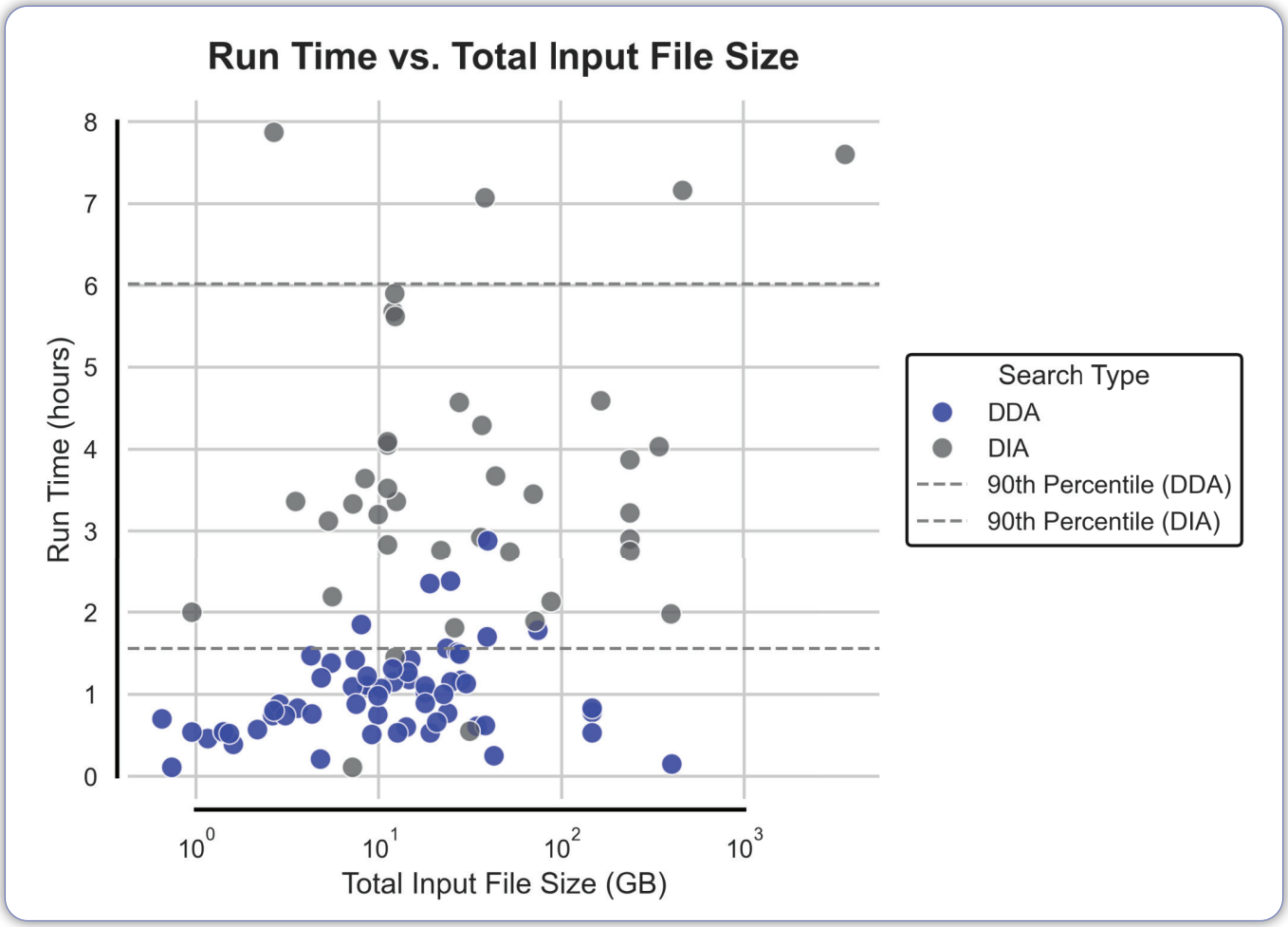


Figure 4: Tesorai Search is extremely scalable both on DDA and DIA datasets. For example, the dda-PASEF immunopeptidomics dataset of 100 Gb was processed in 1.1 hour. Another DIA dataset of 4 Tb was processed in only 7.5 hours.

Quality Control and Transparency

Tesorai Search provides comprehensive QC analyses through interactive dashboard reports that includes standard QC metrics, target-decoy score distribution, and annotated spectra (Figure 5). Additionally, custom analyses and deeper dives can now be performed seamlessly through Tesorai Chat (Figures 6 and 7).

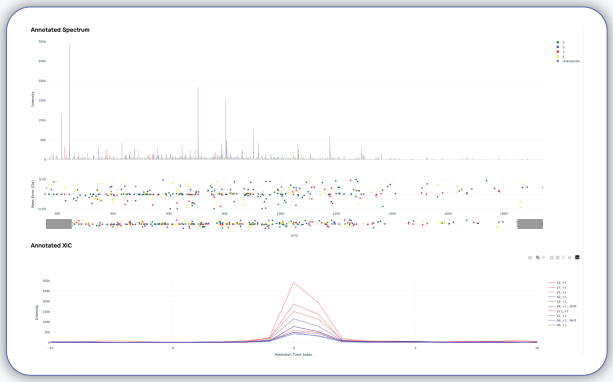


Figure 5: The Tesorai Search interactive dashboard offers a comprehensive view of annotated spectra, facilitating quick assessment of PSM confidence. It enables easy review of annotated peaks in MS2 scans, XICs, and mass errors.

Figure 6: Tesorai Chat can compute missed-cleavage rates for various conditions, such as different proteome ratios in the dia-PASEF tri-species dataset.

Assessing quantification performance is effortless with Tesorai Chat. For example, simply describing the desired plot to Chat allows visualization of quantification accuracy and precision using the mixed proteome dataset, as illustrated in Figure 7.

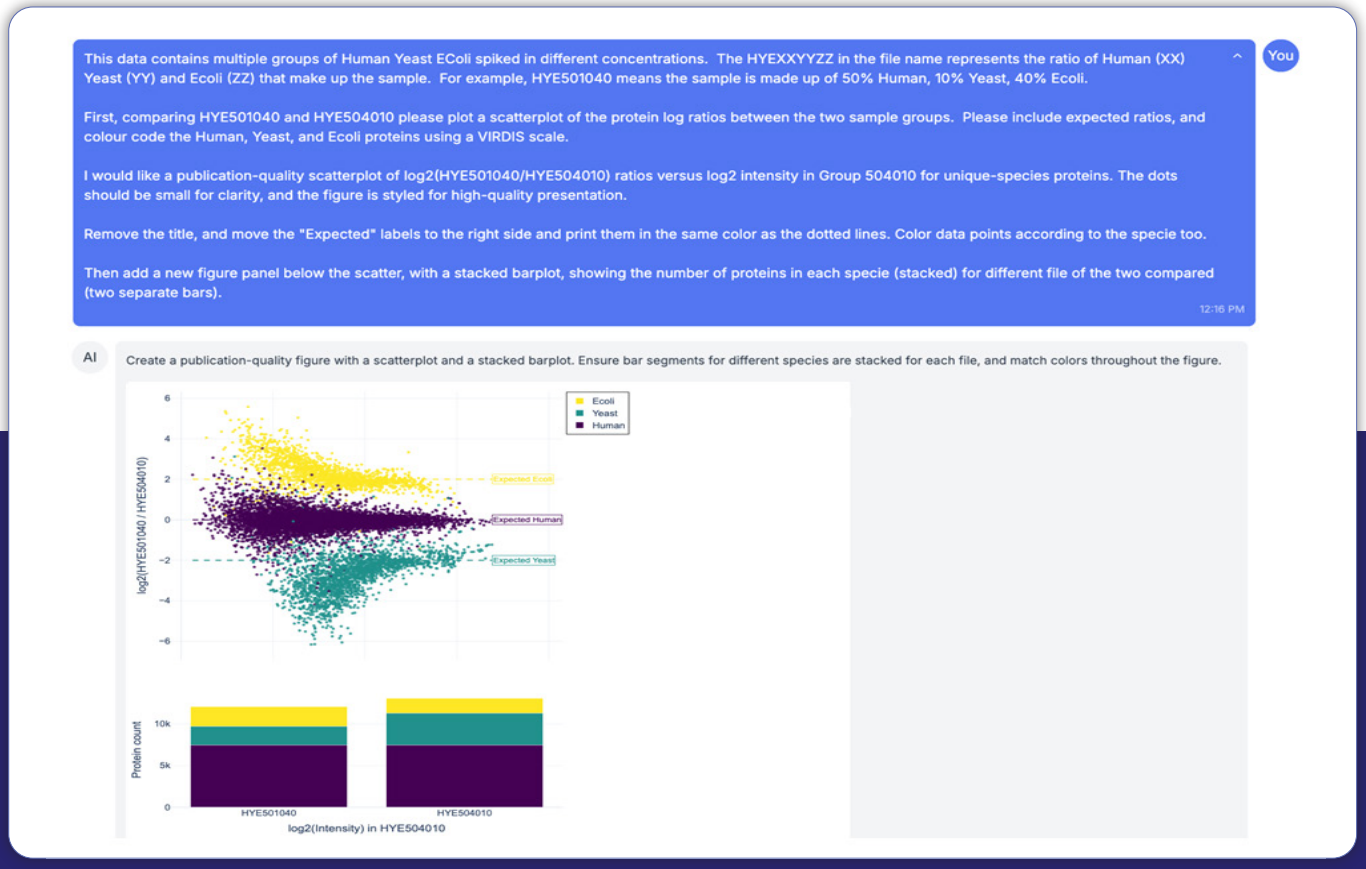


Figure 7: Tesorai Chat can derive complex plots to assess quantification accuracy and precision. These QC tools ensure that researchers can have confidence in their results and communicate their rigor clearly and convincingly to peers and collaborators.

Step 2: Secondary analysis - selecting the subset of proteins or peptides relevant to the hypothesis

After primary identification and quantification of peptides and proteins from raw MS data, the secondary analysis phase focuses on determining which proteins or peptides are biologically and experimentally relevant. This phase involves multiple steps to identify subsets of peptides or proteins that display systematic behavior across biological conditions. Depending on the experimental design, this process can include:

- **Peptide/protein deep dives:** Visualization of various peptide or protein features to uncover biologically relevant patterns.
- **Abundance-based filtering:** Selecting analytes above a certain intensity threshold.
- **Post-translational modification (PTM)–specific filtering:** Isolating peptides carrying defined PTMs (e.g., phosphorylation, acetylation) to focus on regulated proteoforms.
- **Differential abundance testing:** Performing statistical comparisons between groups (e.g., treated vs. control or disease vs. healthy), using tests such as Welch's t-test, ANOVA, or linear models. Multiple testing correction (e.g., Benjamini–Hochberg FDR) is then applied to control false discovery rates.
- **Time-course and regression analyses:** Modeling longitudinal or dose–response experiments using mixed models, spline regression, or other parametric frameworks to capture dynamic proteomic changes.
- **Multivariate and machine learning–based feature selection:** Training classification or regression models (e.g., random forests, elastic nets, or gradient-boosted trees) to predict experimental or clinical phenotypes, followed by assessing feature importance to identify key discriminative analytes.

Through these steps, the proteome can be distilled into a subset of biologically relevant protein candidates that are both statistically robust and interpretable, serving as the basis for downstream functional interpretation and validation.

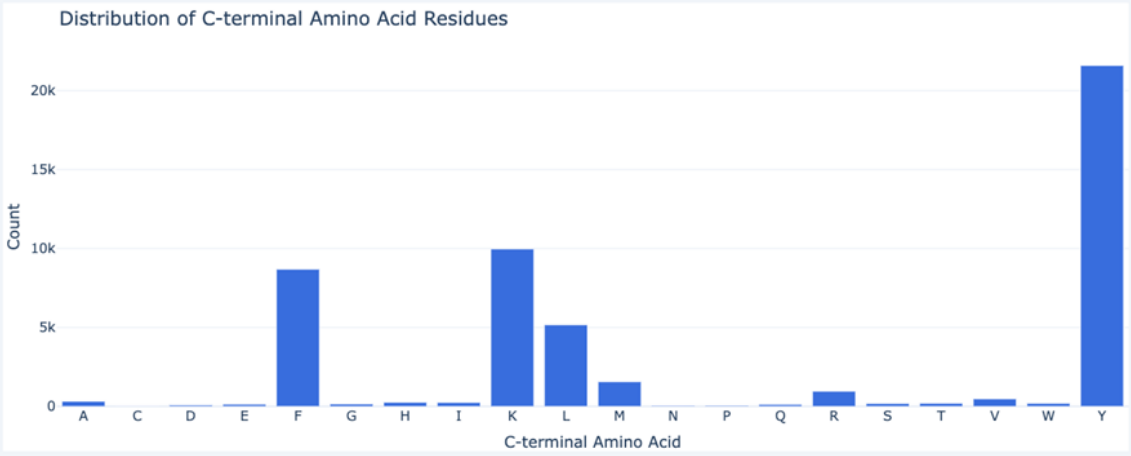
Secondary Analysis in Tesorai Chat

Tesorai Chat provides an integrated, interactive environment for performing analyses directly from processed MS data tables generated by Tesorai Search. Using natural-language queries, researchers can execute complex analytical workflows that would traditionally require multiple software packages and/or scripting environments. Some easy-to-execute examples include:

- Visualization of peptide properties (Figure 8)

User: Can you generate barplot showing distribution of C-terminal amino acid residues? 10/1, 11:38 PM

AI: Here is a bar plot showing the distribution of C-terminal amino acid residues from all identified peptides in the experiment. This helps visualize which amino acids are most frequently found at the C-terminus of peptides.

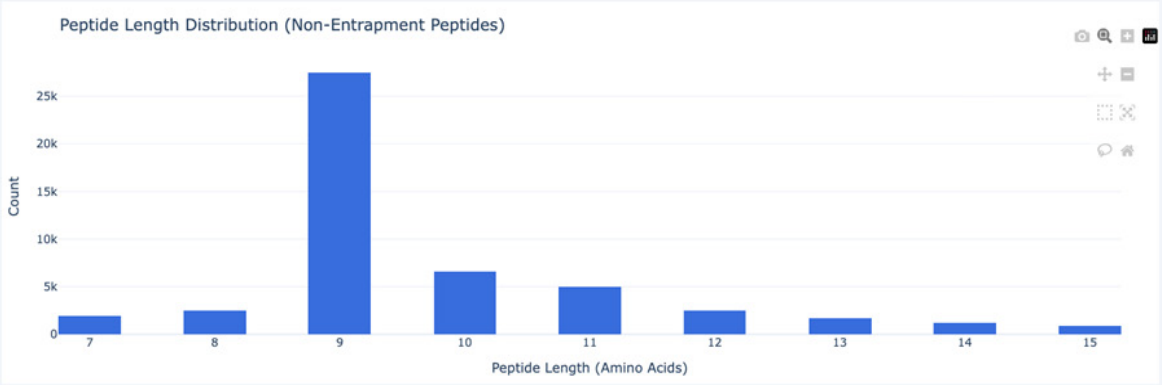


View Code

10/1, 11:38 PM

User: Generate peptide length distribution plot for peptides that aren't from entrapment. 2:29 PM

AI: Here is a histogram showing the distribution of peptide lengths for peptides that are not from entrapment proteins. Only target peptides are included, and entrapment peptides have been excluded.



User: Generate a logo like plot for distribution of amino-acid residues in all 9-mers. 2:30 PM

AI: Here is a logo-like plot showing the distribution of amino acid residues at each position for all 9-mer peptides (peptides of length 9) that are not from entrapment proteins. This visualization helps to identify residue preferences at each position in the 9-mer peptides.

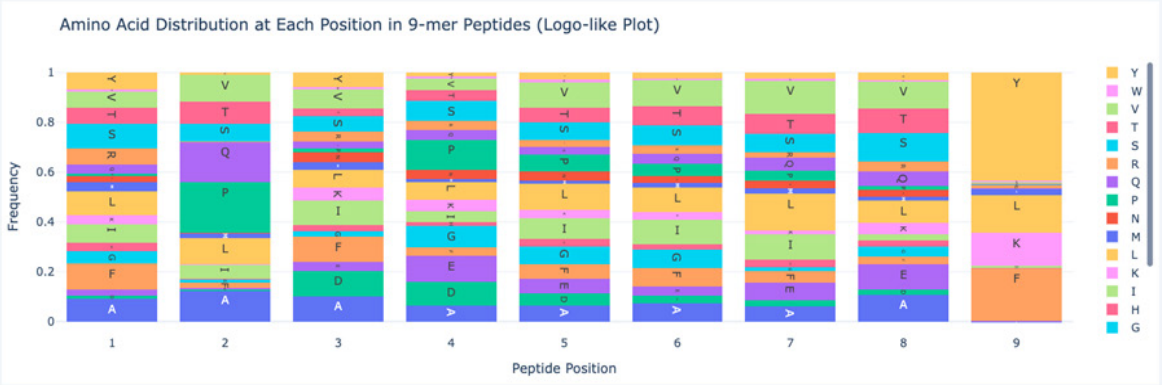


Figure 8: Tesorai Chat is capable of determining the distribution of C-terminal amino acid residues within the identified peptides from immunopeptidomics data sets (top); the distribution of immunopeptide lengths (center); and distribution of amino-acid residues at each position (bottom).

- Selecting the most abundant analytes (or proteins with the highest sequence coverage)
- Identifying statistically significant proteins between treatment groups (Figure 9)

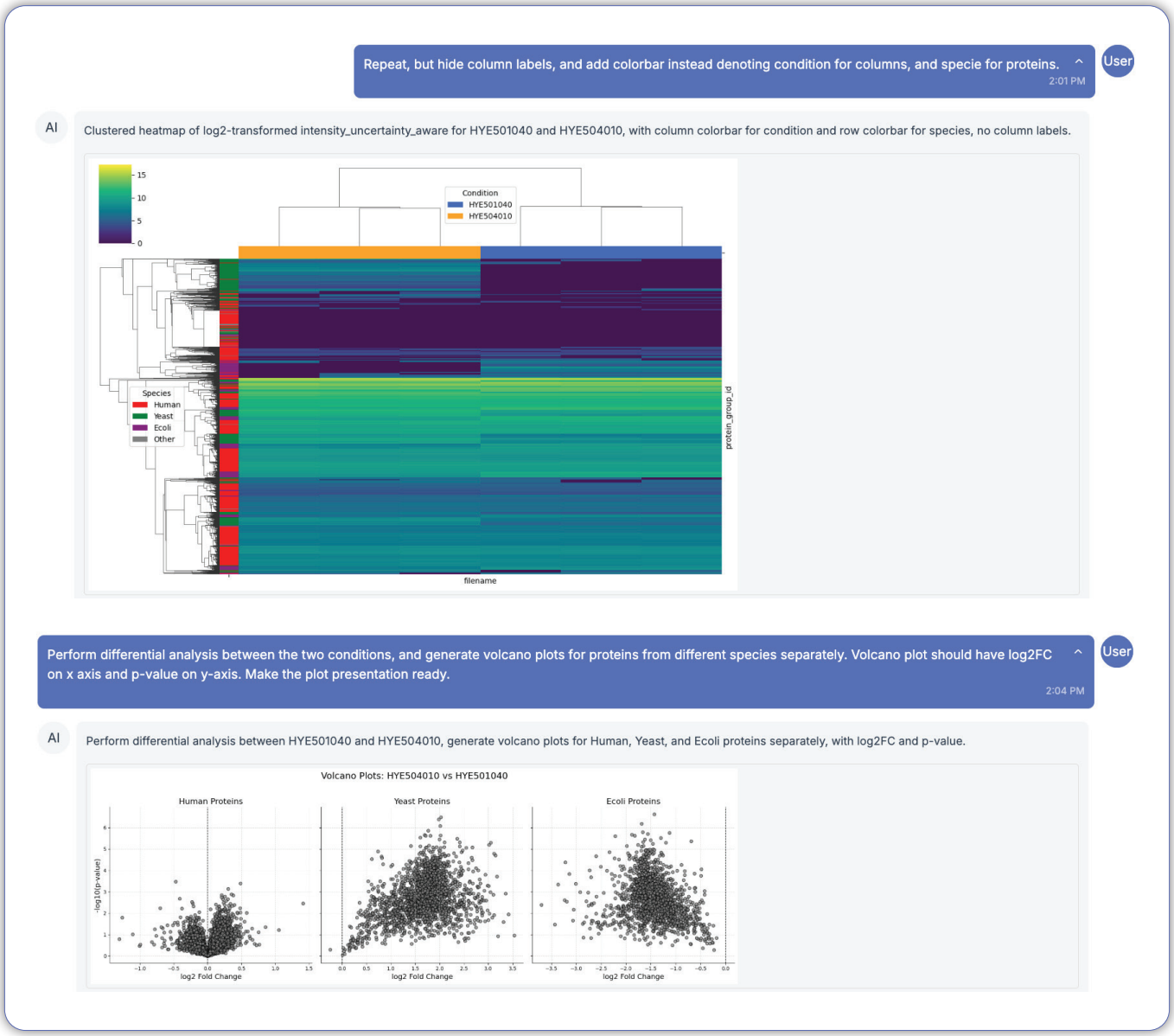


Figure 9: Tesorai Chat empowers researchers and analysts with tools for data exploration and interpretation. It generates annotated clustered heatmaps for visual representation of complex data relationships, helping identify patterns and outliers and batch effects (top). Furthermore, Tesorai Chat can conduct statistical differential analysis to pinpoint significant differences between data groups, offering a comprehensive platform for data understanding (bottom).

- Training machine learning models for classification or regression and visualizing feature importance (Figure 10)

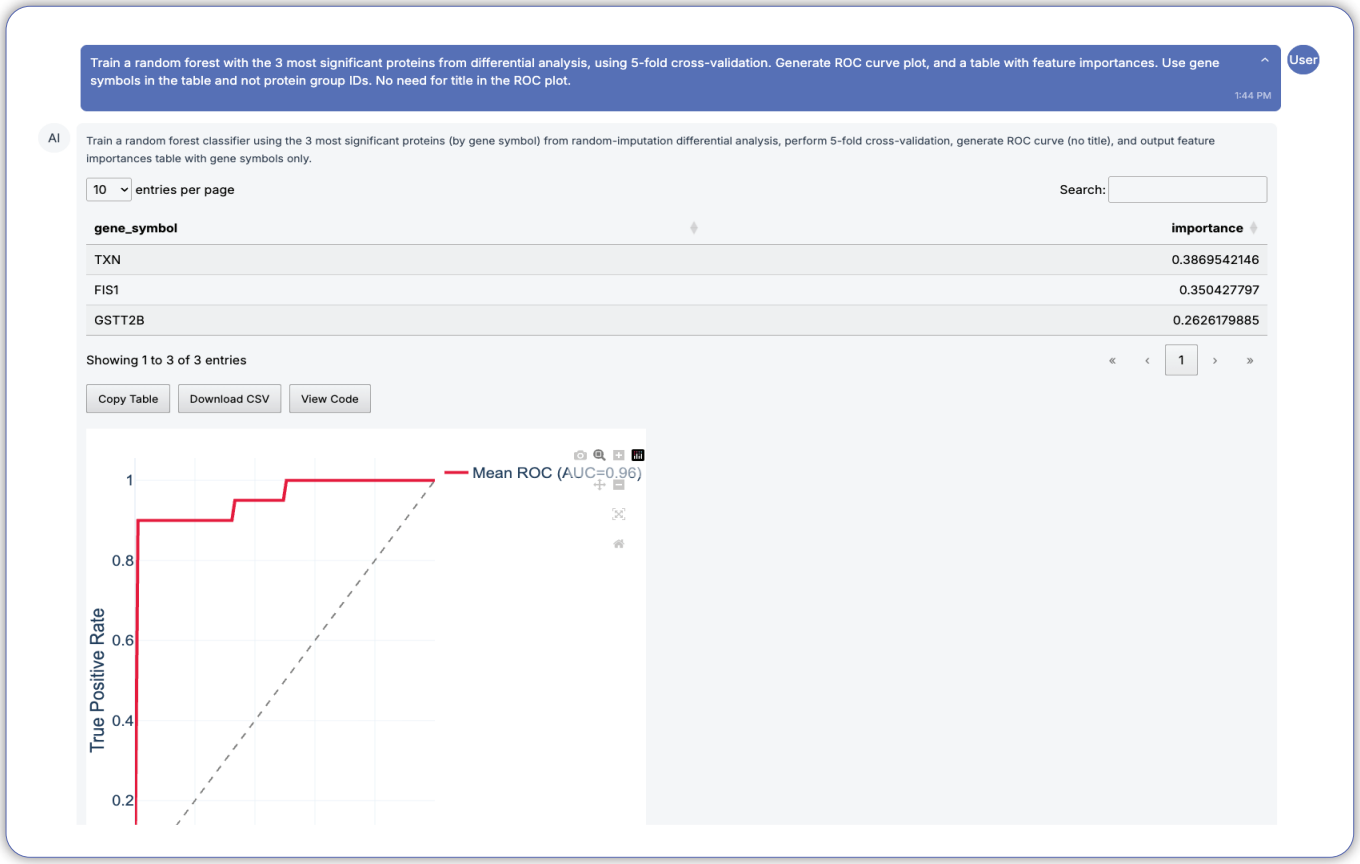


Figure 10: Tesorai Chat is capable of training and cross-validating standard machine learning models, including random forest and support-vector machines. It can also output feature importances for proteins or other covariates utilized in an analysis.

By unifying these capabilities in a single conversational and computational interface, Tesorai Chat lowers the barrier to advanced proteomic data analysis. It ensures methodological transparency by explicitly reporting statistical tests, FDR thresholds, and model parameters, with the complete analysis code available for download to support reproducibility and further exploration. This design also supports rapid iteration and hypothesis refinement.

Unlike conventional software, Tesorai Chat provides full flexibility and interactivity throughout the analytical workflow. Users can request cosmetic changes to visualizations, such as adjusting font sizes, color palettes, or axis labels, directly through natural-language commands, without re-running code or exporting plots. Scientists can also ask Chat to explain the exact methods and algorithms used to perform each analysis and then modify those parameters in subsequent queries. Moreover, Chat can automatically draft text for the Methods section of a manuscript, describing the precise analytical procedures applied.

Traditionally, secondary analysis required close collaboration between experimental scientists and computational specialists, often involving manual data exports, scripting, and multiple software tools, which made iteration slow and limited accessibility. Tesorai Chat removes these barriers: scientists can now perform advanced statistical and machine learning analyses without writing code, while still leveraging state-of-the-art algorithms and open, transparent methods. Acting as a virtual mentor, Chat can guide analytical choices, recommend appropriate statistical tests or modeling strategies, and suggest next steps when uncertainty arises. The result is a faster, more reproducible, and more inclusive approach to proteomic data analysis, placing best-in-class computational tools directly in the hands of every researcher.

Step 3: Tertiary analysis – placing these results into a biological context

After identifying statistically significant or otherwise relevant proteins and peptides in the secondary analysis stage, tertiary analysis contextualizes these findings by integrating proteomic results with external biological knowledge bases and orthogonal data types. This includes mapping identified analytes to known pathways, protein and small-molecule interaction networks, and tissue- or cell-type profiles to uncover the higher-level biological meaning behind observed quantitative patterns.

The outcome of tertiary analysis is a systems-level view of the proteomic data, revealing biological processes, pathways, or disease mechanisms associated with observed changes. This stage bridges quantitative measurement with biological interpretation, enabling hypothesis generation, mechanistic modeling, and prioritization of follow-up experiments.

Tertiary Analysis in Tesorai Chat

Tesorai Chat seamlessly integrates tertiary analysis into the same conversational framework used for earlier stages. Once a subset of significant proteins has been identified, users can ask Chat to perform enrichment, network, or contextual analyses without needing to export data or write customized scripts.

Examples include:

- Performing pathway enrichment analysis on a list of significant proteins and visualizing results as bar charts or network diagrams (Figure 11)

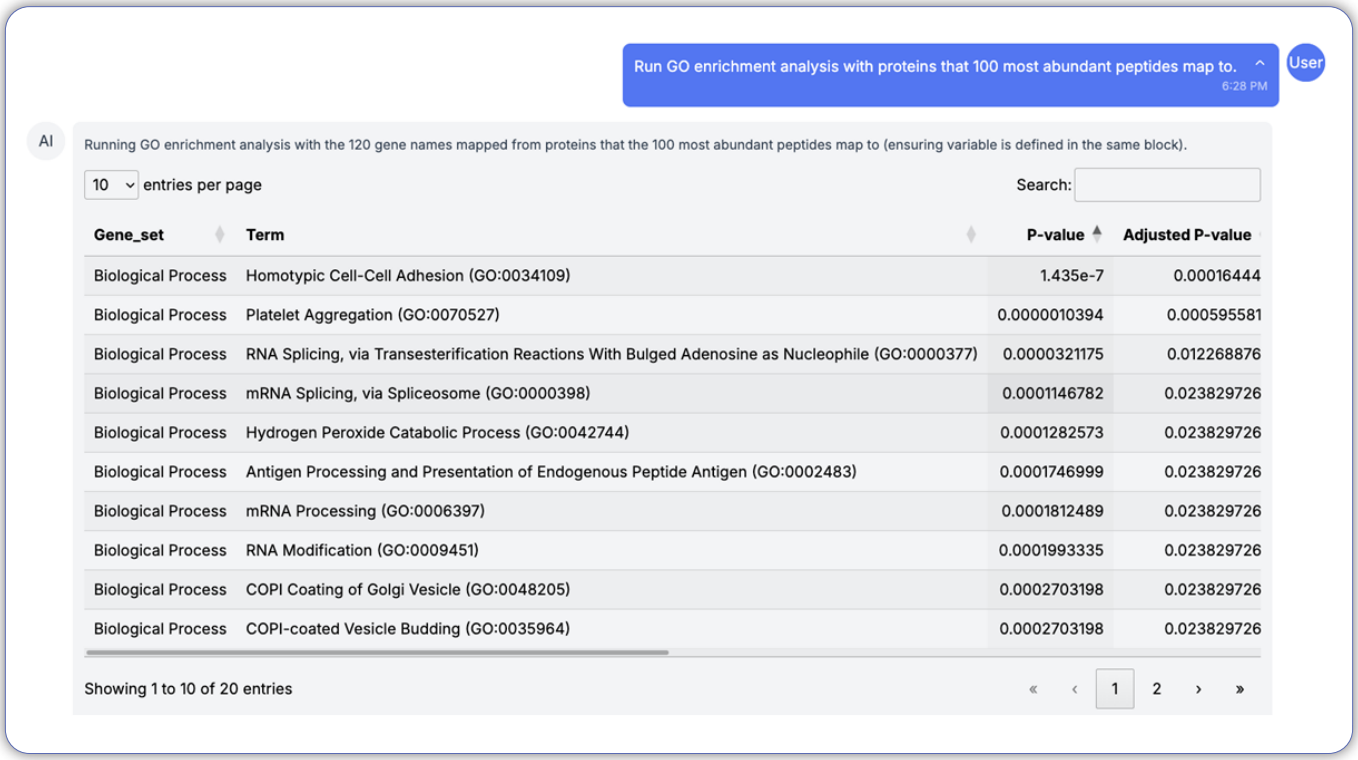


Figure 11: Tesorai Chat enhances differential analysis outputs by integrating them with various knowledge graphs. It offers gene set enrichment analysis, including GO enrichment.

- Exploring known protein–protein interactions among significant hits using STRING-based networks (Figure 12)

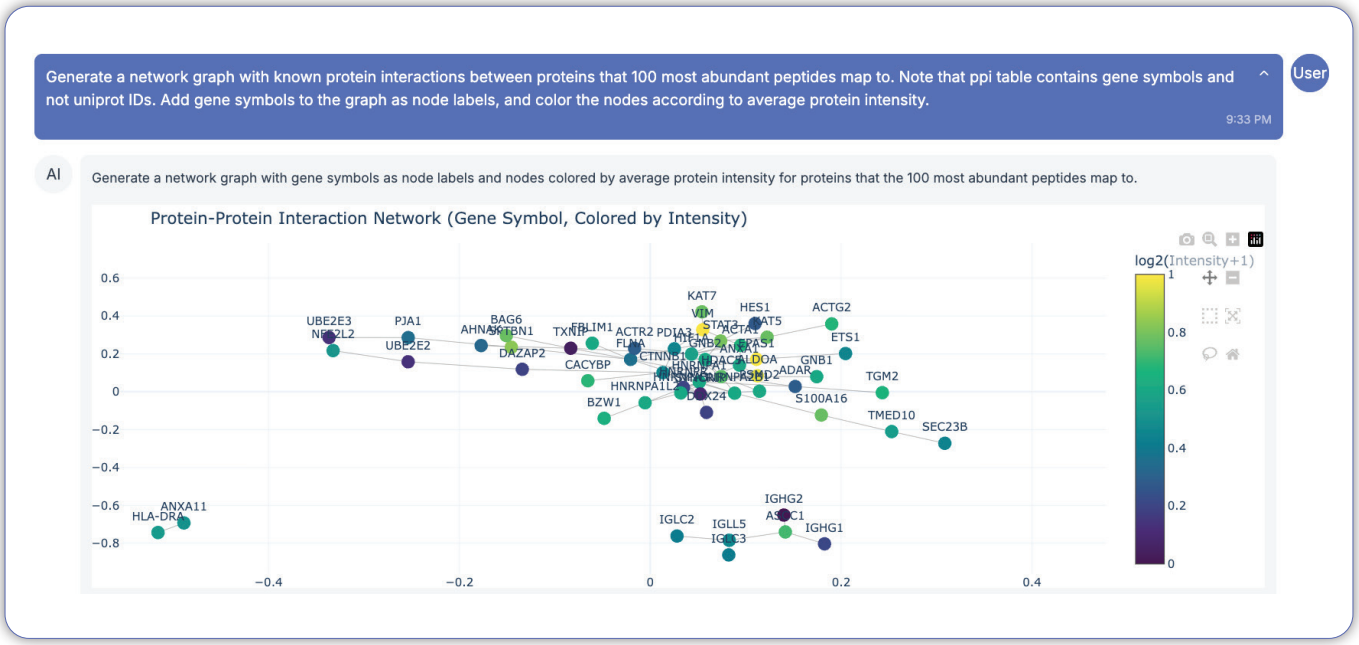


Figure 12: Tesorai Chat offers integration with and visualization of known protein-protein interaction networks, layered with additional information for each node (e.g. protein intensity displayed here) or edges.

- Checking tissue or cell-type expression of top candidates using Human Protein Atlas data

Chat automatically handles background selection, statistical testing (e.g., Fisher's exact test, FDR correction), and integration with major databases through reproducible APIs and cached reference data. It reports all parameters used and allows scientists to request further detail or modify criteria in subsequent queries.

Tertiary analysis represents the culmination of the MS data interpretation pipeline, transforming quantitative outputs into biological understanding. What previously required coordination between multiple tools, databases, and domain experts can now be achieved through a single, conversational interface. With Tesorai Chat, scientists not only gain access to advanced enrichment and network algorithms, but also an intelligent analytical partner that recommends appropriate approaches, explains results, and guides next experimental steps. In doing so, it closes the loop between computation and biology, accelerating the path from data to discovery.

A Note to Our Users

Tesorai Chat delivers advanced AI-assisted analysis for proteomics by combining validated algorithms with curated scientific databases to provide reliable, high-quality results. Our team continually tests and refines the platform to uphold the highest standards of accuracy and performance. We welcome feedback from our users to help Tesorai Chat evolve further. If you have suggestions or notice anything unexpected, please contact us at info@tesorai.com. Your input directly helps shape the future of AI-assisted proteomics analysis.

References

Maximilian Burq, Dejan Stepec, Juan Restrepo, Jure Zbontar, Shamir Urazbakhtin, Bryan Crampton, Shivani Tiwary, Rehan Chinoy, Melissa Miao, Jürgen Cox, Peter Cimermancic. (2024). Back to Basics: Spectrum and Peptide Sequence are Sufficient for Top-tier Mass Spectrometry Proteomics Identification. doi: *https://doi.org/10.1101/2024.08.19.606805*

Guangyuan Li, Omar U Guzmán-Bringas, Aman Sharma, Maxence Dellacherie, Palak Sekhri, Rachel Yamin, Dejan Stepec, Maximilien Burq, Ioana Clotea, Ethan Tardio, Aswin Natarajan, Zachary Harpaz, Xinya Liu, David Requena, Darren Taylor, Beatrix M. Ueberheide, Michelle Krogsgaard, C. Russell Y. Cruz, Peter Cimermancic, Mark Yarmarkovich. (2025) A pan-cancer atlas of therapeutic T cell targets. doi: *https://doi.org/10.1101/2025.01.22.634237*

Bo Wen, Jack Freestone, Michael Riffle, Michael J. MacCoss, William S. Noble, Uri Keich. Assessment of false discovery rate control in tandem mass spectrometry analysis using entrapment. Nat Methods 22, 1454–1463 (2025). doi: *https://doi.org/10.1038/s41592-025-02719-x*

Hoenisch Gravel, N., Nelde, A., Bauer, J. et al. TOFIMS mass spectrometry-based immunopeptidomics refines tumor antigen identification. Nat Commun 14, 7472 (2023). doi: *https://doi.org/10.1038/s41467-023-42692-7*

tesorai

www.tesorai.com info@tesorai.com @tesorai_bio @tesorai