Tesorai Search: Large pretrained model boosts identifications in mass spectrometry proteomics without the need for Percolator.

Authors: Maximilien Burq[1], Dejan Stepec[1], Juan Restrepo[2], Jure Zbontar[1], Shamil Urazbakhtin[2], Bryan Crampton[1], Shivani Tiwary[1], Melissa Miao[1], Juergen Cox[2], Peter Cimermancic[1*],

[1] Tesorai
[2] Computational Systems Biochemistry Research Group, Max-Planck Institute of Biochemistry, Germany

[*]Corresponding author

Abstract:

Current database search algorithms for mass spectrometry proteomics typically fail to identify up to 75% of spectra. To mitigate the issue, tools such as the widely used Percolator and PeptideProphet leverage machine learning for rescoring. Though this can boost peptide and protein identification rates, the fact that they are trained to separate targets from decoys can lead to inaccurate false-discovery rate (FDR) estimates. In this paper, we propose a novel approach to peptide-spectrum matching, based on a pre-trained large deep learning model, which does not utilize decoys during training and does not require training a new model for every new sample. We trained the Tesorai model on over 100M real peptide-spectrum pairs and demonstrated that the approach performs robustly across a wide range of use cases including standard trypsin-digested human samples, immunopeptidomics, metaproteomics, single-cell and isobaric-labeled samples. In addition to providing robust FDR control, our method increases identifications by up to 170%. Furthermore, we incorporated it into a cloud-native, end-to-end system that can process hundreds of samples in just a few hours. To facilitate new discoveries, we made the model publicly available at www.tesorai.com.

Introduction

Mass spectrometry combined with computational analysis has revolutionized proteomics, enabling proteome-wide study of biological systems. In the widely used bottom-up approach, proteins are digested into peptides, which are then analyzed through liquid chromatography-tandem mass spectrometry (LC-MS/MS). The core of peptide (and by extension protein) identification in data-dependent acquisition involves matching fragment ion spectra to peptide sequences, a process underpinned by database searching. First-generation search engines such as Comet[1], MaxQuant [2], MSFragger[3], MS-GF+[4], Sage[5], and Sequest[6] rely on a hand-crafted heuristic for scoring. Because

they don't take the intensities of fragment ions into account, they typically fail to identify 75% of spectra[7].

Second-generation tools, such as Prosit [8], Fragpipe (https://fragpipe.nesvilab.org/) and others use deep learning to generate features such as fragment ion intensity predictions. These features are then combined using a semi-supervised learning approach to improve the matching. This approach, popularized by Percolator[9] and PeptideProphet[10], has significantly enhanced the analysis of mass spectrometry data. However, a pivotal concern with these tools lies in its handling of the false discovery rate (FDR)[11]. In particular, its dual use of decoys—both for training its machine learning model and estimating the FDR—can inadvertently lead to overfitting and insufficient FDR control. This issue of FDR control is not isolated to Percolator: it is also present in all other second-generation tools which rely on a similar approach. Though some fixes have recently been proposed [12] to mitigate these issues, they have not yet been evaluated for FDR correctness in use cases with large search spaces, such as immunopeptidomics or metaproteomics. These issues highlight the need for ongoing development to fully harness the potential of modern deep learning for proteomics in research and clinical applications.

In this paper, we propose a new approach to utilize deep learning for mass-spectrometry proteomics. Inspired by the success of large pre-trained models in various deep learning domains, such as natural language processing[13], [14], computer vision [15], [16], and speech recognition[17], we demonstrate that this strategy can also be leveraged to improve peptide-spectrum matching.

Our model, dubbed Tesorai Search, takes only the raw spectrum and peptide sequence as input, directly outputting a score that effectively distinguishes correct from incorrect PSMs – without requiring retraining on each new sample, and without ever seeing decoys during training. This approach also greatly simplifies processing by removing the need for custom steps such as spectrum de-isotoping, charge correction, or score recalibration. We first demonstrate in entrapment experiments that the model is well calibrated, and that its FDR control is correct. Next, we demonstrate the versatility of this scoring model across diverse experimental settings, increasing the number of identifications when compared to Prosit by 30% in a deep fractionated tryptic sample, by 5% in an immunopeptidomics use-case and by 50% in a metaproteomics dataset.

This approach shares ideas with [18] and [19] where the known peptide-spectrum identifications from first-generation search engines are used as a source of supervision to learn better peptide-spectrum scoring functions. Similarly, concurrent work by [20] proposed repurposing the Casanovo model for database search, using its posterior probability output as a score function. To the best of our knowledge, the resulting scoring models' performance have not been shown to outperform modern (second-generation) tools for database search. Relatedly, [21] used a similar training paradigm to learn to cluster unidentified spectra.

Results

Pretrained model paradigm.

Recent work by Freestone et al. [ref] has employed entrapment experiments to assess the efficacy of FDR control mechanisms of a number of peptide-spectrum rescoring tools, such as Percolator and PeptideProphet. They observe significant underestimation in the FDR control obtained from target-decoy competition, suggesting potential overestimation of peptide identifications.

"The problem, in a nutshell, arises because Percolator uses decoys in two fundamentally different ways: first to train a machine learning model to distinguish between correct and incorrect identifications, and second to estimate the FDR using the TDC procedure." [12]

This widespread approach in proteomics search engines, which requires re-training a machine-learning model for every new sample, stands in sharp contrast with the recent trend in other machine learning fields where large models are trained only once and then used on a wide range of modalities [14], [15], [16], [22]. As a result, we hypothesized that training a single large peptide-spectrum matching algorithm can improve FDR control while also increasing the number of identified peptides across a wide variety of use cases.

To test that hypothesis, we propose a new algorithmic paradigm, where the deep-learning module outputs a score – similar to Andromeda in MaxQuant, Hyperscore in Comet, etc – directly from a single tandem mass spectrum and an associated candidate peptide sequence (Fig. 1A, 1C). This is in contrast to second-generation tools which employ deep learning modules to predict a number of properties, such as ion intensity or retention time prediction, which then need to be combined into a score using a second-stage machine learning model – typically Percolator or PeptideProphet (Fig. 1B).

For training, we collected a representative dataset of around 200M tandem mass spectra, spanning 16000 independent LC/MS/MS runs, across 22 different studies. These raw MS/MS spectra were then searched with a combination of three first-generation database search engines, providing the top peptide sequences for each spectrum. The model was then trained to rank the target peptide sequences. Crucially, the training happens only **once,** and does not need retraining on the dataset of interest. Furthermore, the training process does **not** use any kind of decoy sequences. This is important as it ensures that real sequences are alternatively in the positive and in the negative classes, which prevents the model from simply memorizing the entire peptidomes of the species included in the training set, and enables learning the association between the peptide sequence and the tandem mass spectrum.

The model itself consists of three components. First, a spectrum encoder takes as input a list of centroided tandem mass spectra, in the form of two lists containing the m/z values and the measured ion intensities. It outputs a vector representation of the tandem mass spectrum. Second, a peptide

encoder, which takes the sequence of (possibly modified) residues in the candidate peptide, and outputs a vector representation of the peptide. Finally, a joint encoder combines the vector representations for both the peptide and the spectrum into a single number.

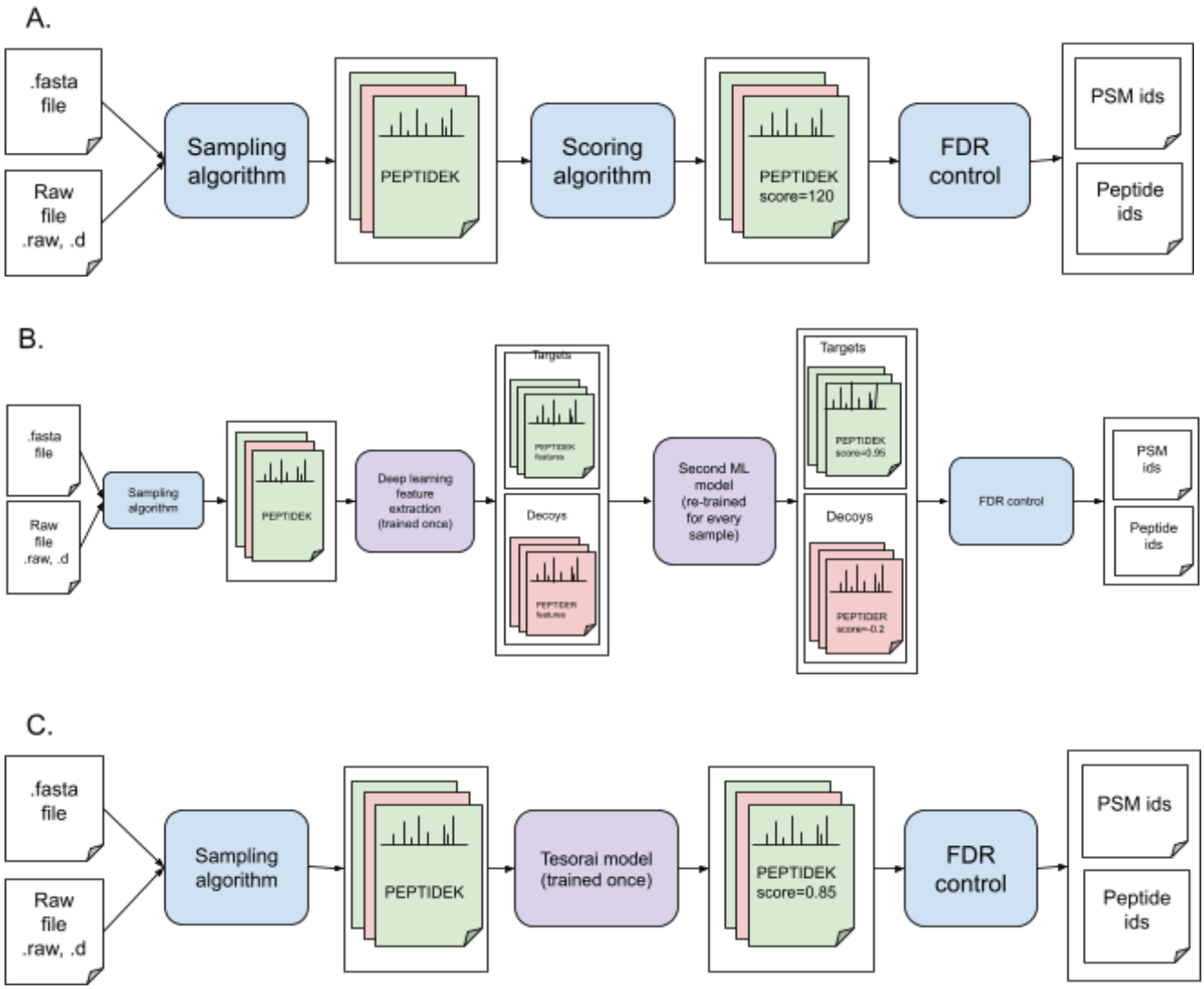

Figure 1. A: Schematic for a traditional (first generation) search engine. Note that the decoys are never used for anything other than to determine the FDR score threshold. **B.** Schematic for a second-generation semi-supervised ML tool. Steps in purple boxes typically leverage machine learning / deep learning algorithms. **C.** Schematic for our search engine, which closely resembles that of traditional search engines.

Accurate False-Discovery rate estimates

Accurate estimation of false-discovery rate (FDR) is typically done using an entrapment analysis. We followed[11], [12] and utilized the ISB18 dataset [23] with entrapment sequences from the Castor plant proteome. We selected three subsets, representing 25%, 50%, and 100% respectively, of the 48 proteins known to be in the sample (including 30 previously found contaminants). This resulting target search database was then combined with the whole proteome of the Castor plant as an entrapment set. This

enabled us to estimate the ability of our model to accurately estimate FDR even in the case where the correct target sequence is not in the search set.

The peptide-level FDR was estimated using the standard target-decoy competition method. The false discovery proportion (FDP) was computed using the "combined method" as detailed by Bo Wen et al.,[24] which provides an upper bound on FDP. The ratio of entrapment to target peptides are 668:1, 1202:1 and 2761:1 for the 25%, 50% and 100% experiments respectively. As a result, the upper-bound is very tight and can be readily used in place of the “paired” method (see Methods). In figure 2, we demonstrated that FDP closely tracks with the FDR threshold, across FDR cutoffs between 0.5% and 10%.

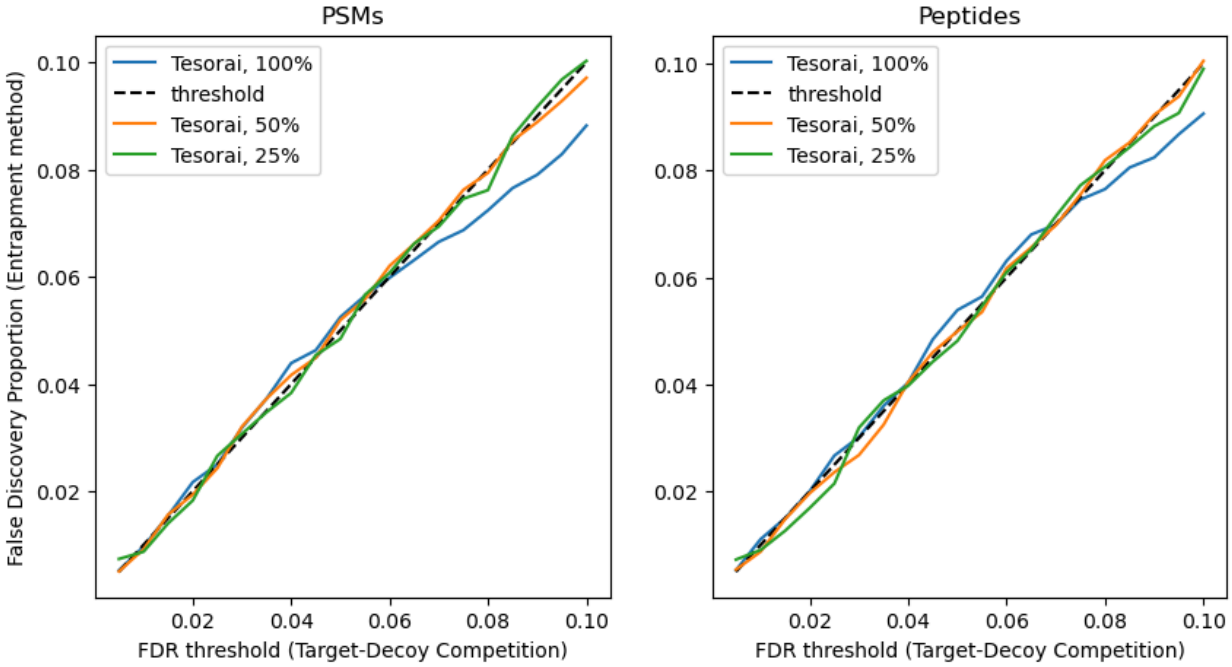

Figure 2. FDR calibration results. False-discovery proportion (entrapment method) as a function of False-discovery threshold (estimated using target-decoy competition).

Increased number of identifications across many common use cases.

Despite recurring concerns related to FDR control, second-generation rescoring tools are widely used because they are able to confidently identify an increased number of peptides in a given sample. This is especially true in instances where the set of peptides in the search database is particularly large or when spectrum quality is poor. We also observed in early experiments that comparing multiple search engines can be challenging, and that the relative performance – as measured by the number of identifications at 1% FDR – can vary from one dataset to another. As a result, we decided to provide comparisons on data from 7 distinct samples, representative of a variety of common use cases in mass-spectrometry proteomics. When possible, to ensure a fair comparison, we selected datasets that were used in the original publications of the search engines against which we benchmark.

We compare our identifications to those found by three leading and widely used search engines: MaxQuant[2], Prosit[8],[21],[22] and MSFragger [3]. To provide a uniform comparison, we re-ran MSFragger (v21.1) and MaxQuant (version 2.5.0.0) across all datasets. We were not able to re-run Prosit, so we report numbers computed from the data deposited on Pride whenever available. In Figure 3, we compare the set of identified peptides at 1% FDR for all three search engines which utilize machine learning (MSFragger, Prosit and Tesorai), against peptides identified by MaxQuant. We use vennbars as a way to measure both the ability to identify new peptides, and to avoid missing ones that are found by a widely used first-generation tool. Despite this variability, we find that our model consistently identified more peptides than MaxQuant, MSFragger and Prosit.

First we re-analyzed two samples which were used in [8] to measure the usefulness of second-generation tools: a deep-fractionated tryptic HeLa sample from [27] (Fig. 3A) and a gut sample designed to test metaproteomics use-cases (Fig. 3B). The HeLa sample was searched with a human fasta file, while the gut sample was searched with a concatenated fasta file Human and Bacteria proteins. Our approach identified 15% more peptides when compared to MaxQuant on the Hela sample, as well as 5% and 16% more identifications when compared to MSFragger and Prosit respectively. It is well known that second-generation tools can significantly increase identifications in cases where the search database is large, as is the case for metaproteomics use-cases. Accordingly, we observe a 86% increase in the number of peptides identified when compared to MaxQuant. Moreover, we identified 4% and 8% more peptides than Prosit and MSFragger.

Next, we looked at two immunopeptidomics samples. The first one, from [ref bassani-sternberg], was previously used in benchmarking Prosit and MSBooster. The second dataset was taken from MassIVE-KB[28] with identifier MSV000089312. Again, we observed significant increases in identifications (Fig 3C and 3D).

We then sought to evaluate our model's ability to generalize to unknown data distributions. Our training dataset consisted of only data from ThermoFisherScientific mass spectrometers so we chose a sample that was processed on a TimsTOF instrument. We observe strong performance (Fig. 3E), even when compared to MSFragger[3] which is known [ref] to work well on TOF data.

Next, we followed[25] and tested our model on a TMT-labeled sample[25]. Note that because of our end-to-end architecture, we were able to handle this use-case with the exact same model architecture and weights as the others. Although the Prosit model did not improve much on MaxQuant, we observe in figure 3F that our model is able to identify 39% more peptides (and 5% more than MSFragger).

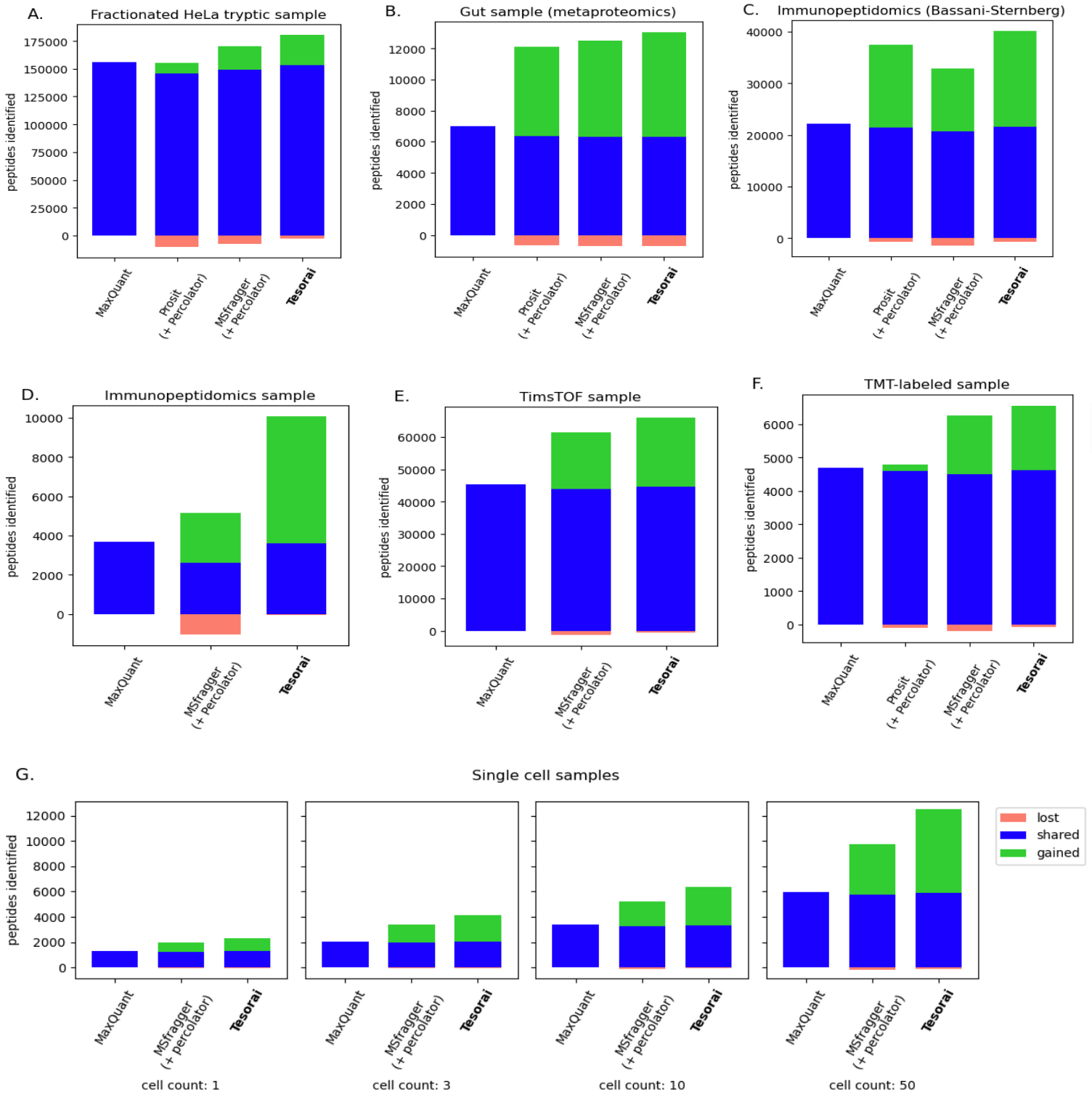

Figure 3: Peptide identifications, 1% FDR. Peptides identified by Prosit, MSFragger and Tesorai, compared to the peptides identified by MaxQuant. Blue indicates peptides identified by both MaxQuant and the model of interest. Green indicates new identifications. Red indicates peptides identified by MaxQuant only. **A.** Deep fractionated tryptic HeLa sample **[27]**. **B.** Gut sample (Metaproteomics search) **[8]**. **C.** Immunopeptidomics use-case **[29]**. **D.** Immunopeptidomics use-case; MSV000089312 by Nowatzky et al. **E.** TimsTOF sample**[30]**. **F.** TMT sample**[25] G.** Single-cell samples (NanoPots) **[31]**.

Finally, in order to test our model on noisy spectra, we selected a single-cell dataset from the nanoPOTS platform[31] [ref]. By analyzing the data from 1, 3, 10 and 50 cells (Fig, 3G), we observed that the number of peptides identified increases as more cells are pooled together, likely because of improved spectra quality. We see a consistent improvement of 17-23% in terms of unique peptide identifications in comparison with MSFragger and up to 100% in comparison with MaxQuant. This enables us to see the same number of peptides with 3-5x fewer cells when compared to MaxQuant.

Newly identified peptides

In figure 4A, we look into the score distributions for the Mel15 sample [29]. We observe that similar to Prosit, the Tesorai scores follow a bimodal distribution where the first mode for targets closely matches resembles the distribution of decoy PSMs. This bimodal distribution enables a cleaner separation of the scores, and makes the resulting identifications less sensitive to the choice of cutoff threshold based on target-decoy competition. We observe that most of the newly identified peptides would also be identified by MaxQuant at a 5% FDR cutoff (Figure 4B).

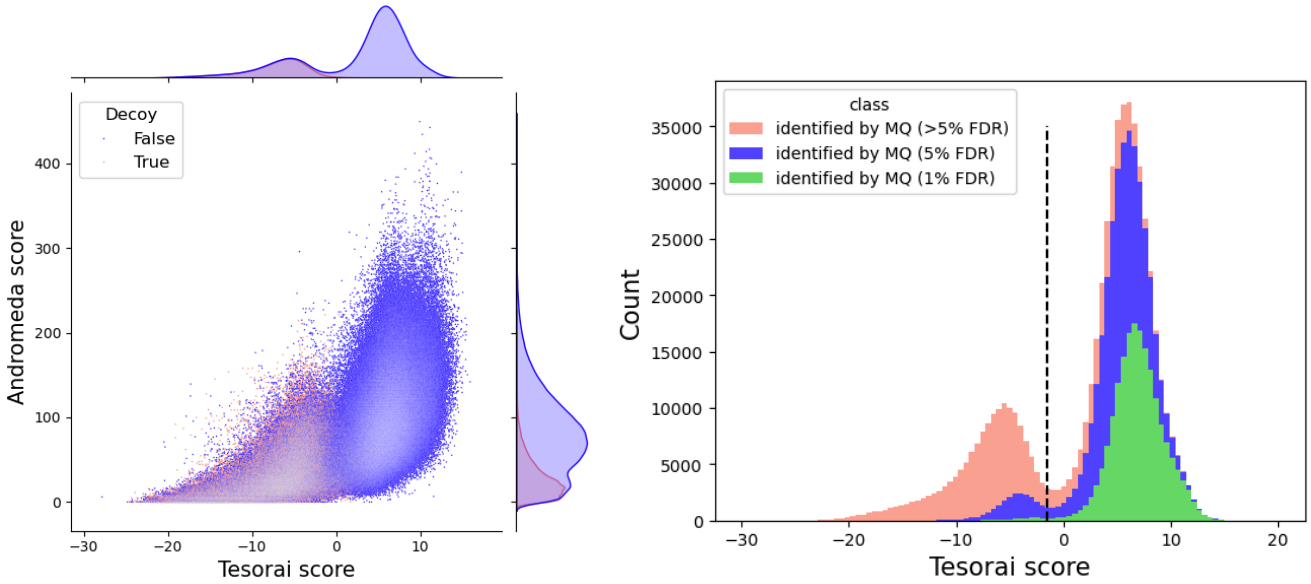

Figure 4. A. Joint distribution of the Andromeda score from MaxQuant, against the Tesorai score. **B.** Breakdown of Tesorai identified PSMs by whether they also were identified by MaxQuant at various FDR thresholds. The dashed line represents the 1% FDR cutoff for the Tesorai Score.

Cloud-based implementation allows for fast processing

The cloud-based implementation of Tesorai Search dramatically enhances the processing speed and accessibility of proteomic analyses. Built on a robust cloud platform, it allows users to perform database searches directly through a web browser without requiring local computational resources or software installations. By simply uploading raw spectral files and corresponding database files, users can leverage the Tesorai platform and model to process hundreds of samples in only a few hours. This

cloud-native approach not only simplifies the workflow, but also ensures that cutting-edge computational resources are available to researchers globally, facilitating rapid data processing and timely generation of results.

To assess the scalability of the cloud-based platform, we searched 4 subsets of [ref] against a standard human proteome (100k proteins, reviewed + unreviewed, downloaded from Uniprot on xx), with varying number of .raw files. Table 1 shows the end-to-end runtime, demonstrating the benefits of parallel scaling of the computation.

Number of .raw files	End-to-end runtime (minutes)	Runtime per file (seconds)
1	38	2280
10	41	246
100	46	27
500	48	5.7

Table 1. End-to-end runtimes. Runtimes for [ref] across various subsets of raw files.

Discussion

The findings presented in this paper demonstrate the significant potential of Tesorai Search, a large pretrained model, to enhance peptide-spectrum matching in mass spectrometry proteomics. This advancement is noteworthy as it achieves high accuracy without relying on the traditional target-decoy competition method, which has been a staple in proteomic analysis but is known for its limitations, such as potential overfitting and inaccurate false discovery rate (FDR) estimation.

Our approach represents a shift toward a more robust method of peptide identification. By utilizing a large-scale dataset of over 100 million real peptide-spectrum pairs for training, Tesorai Search not only maintains rigorous FDR control but also increases identification rates across various complex samples. This improvement is evident when compared to both first-generation search engine and machine learning-enhanced - second generation - proteomics tools. The increased identification rate—up to 170% in certain cases—suggests that Tesorai Search can accurately detect an increased diversity of peptides present in a sample, which holds the potential to lead to new biological insights. For example, in immunopeptidomics, the enhanced detection of peptides that are presented on the surface of cells by MHC molecules can provide critical information for cancer immunotherapy and vaccine development. The robustness of our model is highlighted by its performance across different types of samples, including standard trypsin-digested human samples and more challenging immuno-peptidomics samples. The ability to maintain high identification accuracy in these varied

contexts underscores the model's versatility and its potential to handle a broad spectrum of proteomic challenges.

Furthermore, the implementation of Tesorai Search within a cloud-native, end-to-end system facilitates its use in a wide range of laboratory settings, enabling rapid processing of hundreds of samples within a few hours. This accessibility, combined with the model's availability on a public platform, is poised to democratize high-quality proteomic analysis, allowing researchers with varying levels of computational expertise to conduct advanced proteomic studies.

However, despite these advancements, there are areas that warrant further investigation. For instance, while the model demonstrates improved identification rates, the biological relevance of these newly identified peptides remains to be thoroughly examined. Additionally, the impact of this approach on the long-term reproducibility of mass spectrometry data across different labs and instrumental setups is another important area for future research.

In conclusion, the Tesorai Search approach represents a step forward in the field of proteomics, offering a powerful tool for the accurate and efficient identification of peptides. Its innovative approach to peptide-spectrum matching, combined with its scalability and ease of use, holds promise for advancing our understanding of complex biological systems and fostering new discoveries in both research and clinical settings.

Methods

Model architecture and training

In Figure 5. We describe our procedure to train the end-to-end model. First, we identified 21 publicly available datasets, which were representative of a wide range of mass-spectrometry proteomics use cases. After searching them with three different first-generation search engines, we obtained a training dataset of 100 million tandem mass spectra. We centroided each raw spectrum and generated a list of (possibly modified) peptide sequences that were identified by one of the 3 search engines. For each (spectrum, sequence) pair, we took the ranked list of identified peptides, and trained our model to predict a single score. This score was then compared to the one from other peptides and a ranking loss was used for backpropagation. In order to improve generalization, we introduced various augmentations at the spectrum level.
Though we experimented with a variety of model architectures, in the end we settled on a simple one where we first encode the spectrum and sequence separately, then we combine them and use a joint encoder to output the final score.

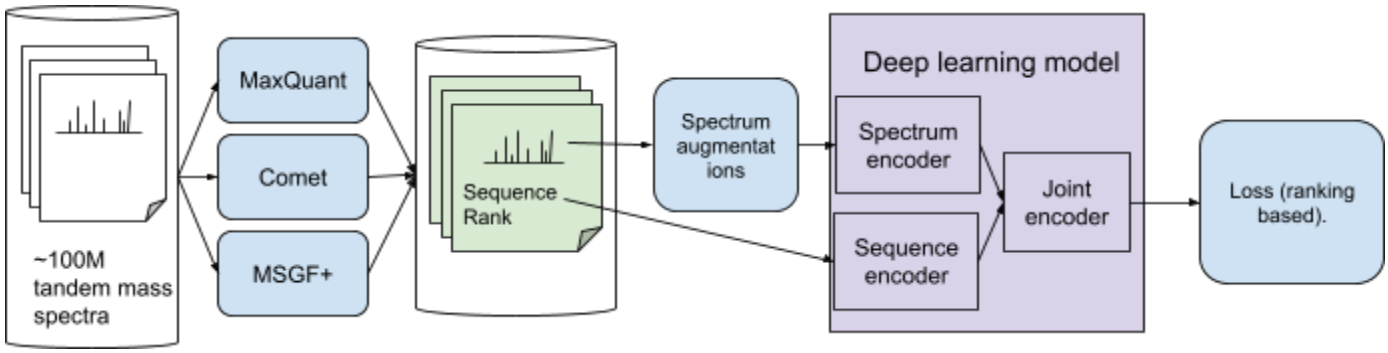

Figure 5. Model training framework. Publicly-available RAW files are searched using three search engines - MaxQuant, Comet and MSGF+. The identified peptides are ranked and we augment the spectra with artificial noise. The deep learning model is then trained to rank the top 3 sequences identified by each search engine.

Entrapment experiment to verify the accuracy of our FDR estimate.

The raw data for the entrapment analysis was accessed on March 15, 2024 at https://regis-web.systemsbiology.net/PublicDatasets/18_Mix/Mix_7/LTQ/RAW_Data/. Following [ref: Freestone et al.] we analyzed files 2-10 and excluded number 11. The Castor plant proteome was accessed from Uniprot on xx. The list of ISB18 proteins was downloaded from https://regis-web.systemsbiology.net/PublicDatasets/database/18mix.fasta.

Following [ref: Freestone], we used strict tryptic enzymatic in-silico digestion (this is our default), and set missed cleavages to 0 (default 2). To reduce the probability of a contaminant randomly matching with a peptide from the castor plant, we set the minimum peptide length to 8. All other settings were kept default. All peptide mappings accounted for Isoleucine to Leucine substitution.

Search engine configurations

MSFragger configuration

MSFragger analysis was performed within the Fragpipe v21.1 computational platform. Fragpipe includes Percolator and Philosopher tools by default for the downstream processing of the MSFragger search results. Decoys and contaminants were generated using Fragpipe. MSBooster and Protein inference (ProteinProphet) were disabled for all experiments. The provided `TMT10-MS3` Fragpipe workflow setting was used in the Prosit-TMT experiment. All other parameters were kept default across all the experiments.

Tesorai Search configuration

Setting	Default	Other options
enzyme	Trypsin/P	Na

allowed_missed_cleavage	2	0 to 2
enzyme_mode	SPECIFIC	UNSPECIFIC
include_contaminants	TRUE	FALSE
fdr_threshold	0.01	0.001 - 1
max_peptide_length	63 (25 if unspecific digestion)	6-15 (if unspecific digestion)
min_peptide_length	7	5-10

Table 2. Tesorai search default settings.

Processing of external data.

A. Bekker-Jensen (HeLa) tryptic dataset

Following [8], the Bekker-Jensen et al. [27] multi-protease dataset was downloaded from the PRIDE repository with the identifier PXD004452. Files mapping to the identifier `QE3_UPLC9_DBJ_SA_46fractions` were selected for analysis. Prosit results were obtained at PXD010871, in the zipped folder `Figure_2_3_5_Multiprotease_Dataset`, under the directory: `/trypsin/percolator_unzipped/prosit_target.psms`. MaxQuant results were obtained at PXD004452, under the directory: `SearchResults/msms.txt`.

Tesorai search results were without contaminants. All other settings were left to their default values. A human Swiss-Prot protein sequence database including annotated isoforms (downloaded May 3rd, 2024; 42421 protein sequences).

MSFragger results were run by us using the settings described above. Enzymatic digestion was set to `stricttrypsin`.

B. Bassani-Sternberg et al. (HLA I immunopeptides) dataset

Following [Wilhelm et al.], the HLA Class I sample from patient MEL15 [Bassani-Sternberg et al.] was downloaded from the PRIDE repository with the identifier PXD004894. Only files mapping to HLA Class I from patient MEL15 were used in the analysis. The sequence database (`HUMAN_2014`) was provided within the zipped `Search` folder in PXD004894, which was used in all experiments.

MaxQuant and Prosit results were obtained from PXD021398 [Wilhelm et al.]. MaxQuant results (`msms.txt`) were obtained from the zipped folder `Figure_5_Mel15_MaxQuant1`. Prosit results (`prosit_target.peptides`) were obtained from the zipped folder `Figure_5_Mel15_MaxQuant100_and_Rescoring`, under the directory `forPride_100%/rescoring_for_paper_2/percolator`.

Tesorai search results were with contaminants turned off. The enzyme mode was set to `UNSPECIFIC` with the minimum peptide length set to 7 and maximum to 15, following [Wilhelm et al.].

MSFragger results were run by us using the settings described above. Enzymatic digestion was set to `NONSPECIFIC`. The minimum peptide length was set to 8 and the maximum to 15, following [Wilhelm et al.].

C. Nowatzky et al. (HLA I immunopeptides) dataset

The Nowatzky et al. dataset was downloaded from MSV000089312. Only files mapping to identifiers `190514_H_FreyaLC_AD_Nowatzky_HLA_test` (KO and WT) were used in the analysis.

Tesorai search results were with contaminants turned off. The enzyme mode was set to `UNSPECIFIC` with the minimum peptide length set to 7 and the maximum to 15.

MSFragger results were run by us using the settings described above. Enzymatic digestion was set to `NONSPECIFIC`. The minimum peptide length was set to 7 and the maximum to 12.

D. Meier et al (HeLa sample, timsTOF)

Human cervical cancer cell (HeLa) dataset used in [30] was downloaded from the PRIDE repository with the identifier PXD010012. The 200ng, 100ms experiment was reproduced with the raw files provided in the zipped folder `HeLa_200ng_100ms_raw`. MaxQuant results (`msms.txt`) were obtained in the same PRIDE repository in the zipped folder `HeLa_200ng_100ms_txt`.

Tesorai search results were run with contaminants turned off. At the time of writing, raw data from Bruker instruments (.b) was not yet supported in the main platform. As a result, this dataset was processed with an experimental parsing and sampling method. The model rescoring was kept the same. All other settings were left to their default values. A reference human proteome (UP000005640) was downloaded from UniProt using one protein sequence per gene (March 26th, 2024; 20,590 protein sequences).

MSFragger results were run by us using the settings described above. Enzymatic digestion was set to `stricttrypsin`.

E. Gabriel et al (HeLa-Yeast, TMT)

Human Yeast Dilution TMT labeled dataset [25] was downloaded from the PRIDE repository with the identifier PXD030340. The HCD fragmentation on the Orbitrap mass analyzer (HCD/OT) experiment was reproduced. The `190416_FPTMT_MS3_HCDOT_R1` raw file was used in the analysis. MaxQuant (`msms.txt`) and MaxQuant with Andromeda (`andromeda_target.pepides`) results

were obtained in the same PRIDE repository in the zipped folder `HCD_OT_Without_rescoring`. Prosit rescored results (`prosit_target.peptides`) were obtained in the same PRIDE repository in the zipped folder `HCD_OT_With_rescoring`.

Tesorai search results were with contaminants turned off. All other settings were left to their default values. A reference human proteome (UP000005640) was downloaded from UniProt using one protein sequence per gene (March 26th, 2024; 20,590 protein sequences). A reference Baker's yeast proteome (UP000002311) was downloaded from UniProt using one protein sequence per gene (March 28th, 2024; 6,060 protein sequences).

MSFragger results were run by us using the settings described above. Enzymatic digestion was set to `stricttrypsin`. The provided `TMT10-MS3` Fragpipe workflow setting was used.

F. Williams et al. (Single-cell NanoPOTS)

Following [Yang et al.], the single-cell data from the nanoPOTS platform [31] was downloaded from the MassIVE repository under the identifier MSV000085230. We reproduced the experiment from [Yang et al.] using 1, 3, 10, and 50 cells from the MCF10A 30-minute experiment. We used the human Swiss-Prot protein sequence database provided at MSV000085230 for all the experiments (`UniprotKB_homosapiens_Swiss_Prot_122916`).

Tesorai search results were with contaminants turned on. All other settings were left to their default values. MSFragger results were run by us using the settings described above. Enzymatic digestion was set to `stricttrypsin`. For the MaxQuant results we used MaxQuant version 2.5.2.0 with default settings except for the following: proteon FDR equal to 1, label-free quantification turned on.

Cloud-based implementation

The ability to find vastly more peptides from the same samples, while rigorously controlling FDR, holds the promise to discover new biology. As a result, we've packaged the model into a full-fledged, publicly available search engine, Tesorai Search. The web-based architecture is designed to make running a database search intuitive and easy for anyone; it works on any web browser with no installation needed. All you need is .raw files in the manufacturer format and a .fasta file containing your search database (Fig. 6).

We enabled fast and scalable processing by running with high parallelism and leveraging on-demand, serverless Cloud resources. Each sample file is first run through multiple first-generation search algorithms at 100% FDR. Every file and algorithm is run on a separate Cloud machine, with optimized resources for that stage. The Tesorai Search model is then run on the first-generation search results, using T4 Nvidia GPUs, and the results are collected into a usable format with a graphical report.

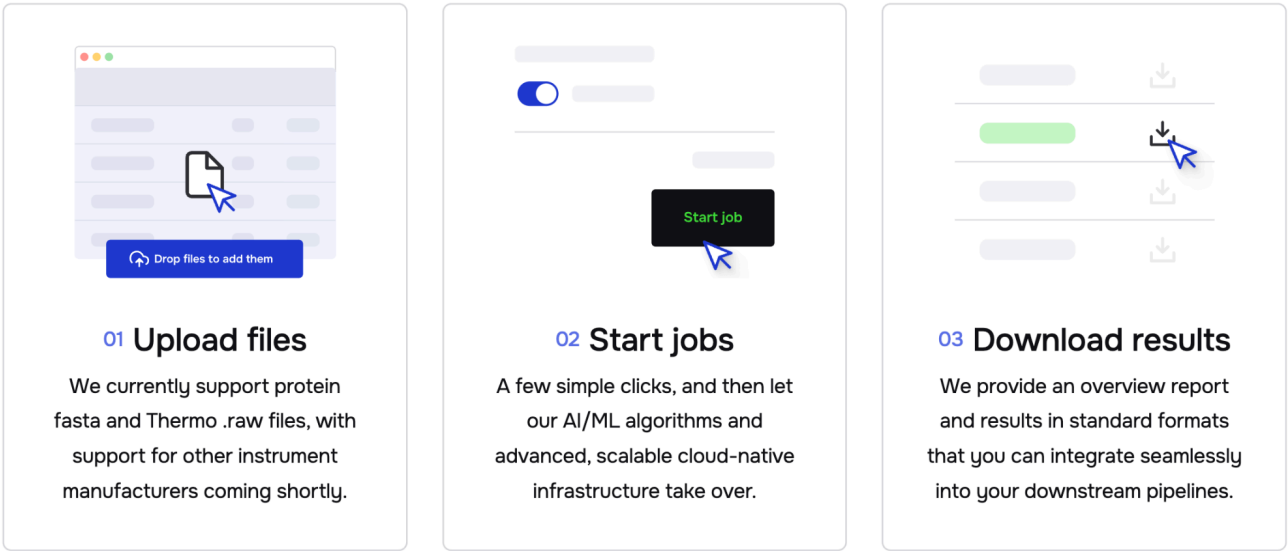

Figure 6. Visual summary of the Tesorai Search platform. Users can upload RAW and fasta files, configure jobs and download results in an intuitive and streamlined web interface.

Data availability

All raw data analyzed in this study were already publicly available. All results run by ourselves, from Tesorai Search, MaxQuant, and FragPipe will be made available on the Pride platform.

Code availability

The Tesorai platform, used to generate the main results, is available online at console.tesorai.com. The code used to process the results and generate figures and tables will be made available upon reasonable request.

References

[1] J. K. Eng, T. A. Jahan, and M. R. Hoopmann, “Comet: An open-source MS/MS sequence database search tool.,” *PROTEOMICS*, vol. 13, no. 1, pp. 22–24, doi: https://doi.org/10.1002/pmic.201200439.

[2] J. Cox and M. Mann, “MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification,” *Nat. Biotechnol.*, vol. 26, no. 12, pp. 1367–1372, 2008, doi: 10.1038/nbt.1511.

[3] A. T. Kong, F. V. Leprevost, D. M. Avtonomov, D. Mellacheruvu, and A. I. Nesvizhskii, “MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics,” *Nat. Methods*, vol. 14, no. 5, pp. 513–520, 2017, doi: 10.1038/nmeth.4256.

[4] S. Kim and P. A. Pevzner, “MS-GF+ makes progress towards a universal database search tool for proteomics,” *Nat. Commun.*, vol. 5, no. 1, pp. 5277–5277, 2014, doi: 10.1038/ncomms6277.

[5] M. R. Lazear, “Sage: An Open-Source Tool for Fast Proteomics Searching and Quantification at Scale,” *J. Proteome Res.*, vol. 22, no. 11, pp. 3652–3659, Nov. 2023, doi: 10.1021/acs.jproteome.3c00486.

[6] J. K. Eng, A. L. McCormack, and J. R. Yates, “An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database,” *J. Am. Soc. Mass Spectrom.*,

vol. 5, no. 11, pp. 976–989, Nov. 1994, doi: 10.1016/1044-0305(94)80016-2.
[7] J. Griss *et al.*, "Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets," *Nat. Methods*, vol. 13, no. 8, pp. 651–656, 2016, doi: 10.1038/nmeth.3902.
[8] S. Gessulat *et al.*, "Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning," *Nat. Methods*, vol. 16, no. 6, pp. 509–518, 2019, doi: 10.1038/s41592-019-0426-7.
[9] L. Käll, J. D. Canterbury, J. Weston, W. S. Noble, and M. J. MacCoss, "Semi-supervised learning for peptide identification from shotgun proteomics datasets," *Nat. Methods*, vol. 4, no. 11, pp. 923–925, 2007, doi: 10.1038/nmeth1113.
[10] A. Keller, A. I. Nesvizhskii, E. Kolker, and R. Aebersold, "Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search," *Anal. Chem.*, vol. 74, no. 20, pp. 5383–5392, Oct. 2002, doi: 10.1021/ac025747h.
[11] J. Freestone, W. S. Noble, and U. Keich, "Reinvestigating the Correctness of Decoy-Based False Discovery Rate Control in Proteomics Tandem Mass Spectrometry," *J. Proteome Res.*, Apr. 2024, doi: 10.1021/acs.jproteome.3c00902.
[12] J. A. Freestone, L. Käll, W. S. Noble, and U. Keich, "Semi-supervised learning while controlling the FDR with an application to tandem mass spectrometry analysis," *bioRxiv*, p. 2023.10.26.564068-2023.10.26.564068, Jan. 2024, doi: 10.1101/2023.10.26.564068.
[13] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," *CoRR*, vol. abs/2005.14165, 2020, [Online]. Available: https://arxiv.org/abs/2005.14165
[14] J. Devlin Chang, M. ,. Lee, K. ,. &. Toutanova, K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *ArXiv*, 2018.
[15] A. Radford Kim, J. W. ,. Hallacy, C. ,. Ramesh, A. ,. Goh, G. ,. Agarwal, S. ,. Sastry, G. ,. Askell, A. ,. Mishkin, P. ,. Clark, J. ,. Krueger, G. ,. &. Sutskever, I., "Learning Transferable Visual Models From Natural Language Supervision," *ArXiv*, 2021.
[16] athilde Caron and Hugo Touvron and Ishan Misra and Hervé Jégou and Julien Mairal and Piotr Bojanowski and Armand Joulin, "Emerging Properties in Self-Supervised Vision Transformers," *arXiv*, 2021.
[17] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision." 2022.
[18] W. Bittremieux, D. H. May, J. Bilmes, and W. S. Noble, "A learned embedding for efficient joint analysis of millions of mass spectra," *Nat. Methods*, vol. 19, no. 6, pp. 675–678, Jun. 2022, doi: 10.1038/s41592-022-01496-1.
[19] S. S. Schoenholz *et al.*, "Peptide-Spectra Matching from Weak Supervision." 2018. [Online]. Available: https://arxiv.org/abs/1808.06576
[20] V. Ananth, J. Sanders, M. Yilmaz, S. Oh, and W. S. Noble, "A learned score function improves the power of mass spectrometry database search," *bioRxiv*, 2024, doi: 10.1101/2024.01.26.577425.
[21] T. Altenburg, T. Muth, and B. Y. Renard, "yHydra: Deep Learning enables an Ultra Fast Open Search by Jointly Embedding MS/MS Spectra and Peptides of Mass Spectrometry-based Proteomics," *bioRxiv*, 2021, doi: 10.1101/2021.12.01.470818.
[22] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *ArXiv*, 2018.
[23] J. Klimek *et al.*, "The Standard Protein Mix Database: A Diverse Data Set To Assist in the Production of Improved Peptide and Protein Identification Software Tools," *J. Proteome Res.*, vol. 7, no. 1, pp. 96–103, Jan. 2008, doi: 10.1021/pr070244j.
[24] B. Wen, J. Freestone, M. Riffle, M. J. MacCoss, W. S. Noble, and U. Keich, "Assessment of false discovery rate control in tandem mass spectrometry analysis using entrapment," *bioRxiv*, p. 2024.06.01.596967-2024.06.01.596967, Jan. 2024, doi: 10.1101/2024.06.01.596967.
[25] W. Gabriel *et al.*, "Prosit-TMT: Deep Learning Boosts Identification of TMT-Labeled Peptides," *Anal. Chem.*, vol. 94, no. 20, pp. 7181–7190, May 2022, doi: 10.1021/acs.analchem.1c05435.
[26] M. Wilhelm *et al.*, "Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics," *Nat. Commun.*, vol. 12, no. 1, p. 3346, Jun. 2021, doi: 10.1038/s41467-021-23713-9.
[27] D. B. Bekker-Jensen *et al.*, "An Optimized Shotgun Strategy for the Rapid Generation of

Comprehensive Human Proteomes," *Cell Syst.*, vol. 4, no. 6, pp. 587-599.e4, Jun. 2017, doi: 10.1016/j.cels.2017.05.009.
[28] M. Wang, J. Wang, J. Carver, B. S. Pullman, S. W. Cha, and N. Bandeira, "Assembling the Community-Scale Discoverable Human Proteome.," *Cell Syst.*, vol. 7, no. 4, pp. 412-421.e5, Oct. 2018, doi: 10.1016/j.cels.2018.08.004.
[29] M. Bassani-Sternberg *et al.*, "Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry," *Nat. Commun.*, vol. 7, no. 1, pp. 13404–13404, 2016, doi: 10.1038/ncomms13404.
[30] F. Meier *et al.*, "Online Parallel Accumulation–Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer," *Mol. Cell. Proteomics MCP*, vol. 17, pp. 2534–2545, 2018.
[31] S. M. Williams *et al.*, "Automated Coupling of Nanodroplet Sample Preparation with Liquid Chromatography–Mass Spectrometry for High-Throughput Single-Cell Proteomics," *Anal. Chem.*, vol. 92, no. 15, pp. 10588–10596, Aug. 2020, doi: 10.1021/acs.analchem.0c01551.