Tracking Dataset Reuse in Proteomics: A Comprehensive Analysis of PRIDE Archive Downloads

Suresh Hewapathirana[1], Jingwen Bai[1], Chakradhar Bandla[1], Selvakumar Kamatchinathan[1], Deepti J Kundu[1], Nithu Sara John[1], Boma Brown-Harry[1], Nandana Madhusoodanan[1], Marc Riera Duocastella[1], Juan Antonio Vizcaíno[1], Yasset Perez-Riverol[1,*]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

*Corresponding author: yperez@ebi.ac.uk

Abstract

Understanding how scientific datasets are accessed and reused is essential for resource planning, impact assessment, and the evaluation of open science policies. Here we present the PRIDE Archive download tracking infrastructure and a comprehensive analysis of download records from the PRIDE proteomics database (2020-2025), spanning 34,085 datasets accessed from 213 countries. The infrastructure includes `nf-downloadstats`, a scalable Nextflow pipeline for processing large-scale download logs, and DeepLogBot, a deep-learning bot detection framework. After removing 88.0% of download traffic identified as automated, analysis of the remaining 19.1 million genuine downloads reveals that the United States leads PRIDE data reuse (26.8%), followed by the United Kingdom (23.6%) and Germany (22.5%). Downloads grew steadily from 1.2 million in 2021 to 5.8 million in 2025, with FTP and HTTP alternating as the dominant download protocol. Dataset reuse follows a highly concentrated distribution, with the top 1% of datasets accounting for 43.3% of all downloads. These findings provide actionable insights for repository infrastructure planning and demonstrate the importance of bot-aware analytics for scientific data resources.

Availability: The `nf-downloadstats` pipeline is available at `https://github.com/PRIDE-Archive/nf-downloadstats` and DeepLogBot at `https://github.com/ypriverol/deeplogbot`, both under the Apache 2.0 license.

Keywords: bot detection, data reuse, download analytics, PRIDE, proteomics

1 Introduction

The Proteomics Identification database (PRIDE) is a leading public mass spectrometry-based proteomics data repository [1]. As a founding member of the ProteomeXchange consortium, PRIDE enables researchers to share and access high-quality proteomics datasets globally, promoting transparency, reproducibility, and data reuse. Aligned with the FAIR principles - Findable, Accessible, Interoperable, and Reusable [2] - PRIDE supports open science by ensuring that public datasets are well-annotated and machine-readable. These principles are essential for maximizing the value of shared scientific data.

Understanding how public datasets are reused is essential for assessing their scientific impact and informing data-driven policies. While citations in scholarly publications offer one indicator of reuse, download statistics provide a more immediate and granular view of data demand and utility. In our previous work [3], we demonstrated that usage metrics can serve as complementary indicators of impact, supporting improved data stewardship, resource allocation, and funding decisions. Beyond measuring impact, download statistics are critical for designing more effective data infrastructures. Insights into download behavior can inform the optimization of

data access protocols, guide the prioritization of metadata curation and visualization features, and identify high-value datasets for targeted annotation or integration efforts. As public data volumes continue to grow, usage-driven strategies become increasingly important for improving dataset discoverability and reuse. For example, frequently downloaded datasets - particularly those used as community benchmarks - could be prioritized for enhanced metadata annotation (e.g., SDRF sample descriptions [4]) and enriched with curated tags and keywords, making them easier to find through search interfaces. Users could then combine these annotations with download counts to identify the most relevant and community-validated datasets for their analyses. Similarly, repositories can leverage download patterns to allocate faster transfer services and optimized storage for high-demand datasets, ensuring that the most reused data remain readily accessible.

Despite their importance, systematic tracking of dataset downloads remains a major challenge across bioinformatics resources, PRIDE and other ProteomeXchange partners. Barriers include the absence of a standardized infrastructure for logging access events, technical complexities in aggregating usage data across distributed and heterogeneous transfer systems (e.g., FTP, HTTP), and ongoing concerns related to user privacy and data protection. Compounding these challenges, automated bot traffic contaminates download statistics - studies estimate that bots account for 30-70% of all internet traffic (https://cpl.thalesgroup.com/ppc/application-security/bad-bot-report), and scientific repositories are particularly attractive targets due to their open-access policies and valuable content [5]. Without accounting for this contamination, any analysis of repository usage risks drawing conclusions from inflated and distorted metrics. As bioinformatics resources continue to scale in both size and complexity, robust download analytics will become increasingly vital - not only for measuring impact - but also for enabling smarter, user-informed development of open data platforms.

Here, we present the PRIDE Archive downloads infrastructure, which includes `nf-downloadstats`, a large-scale Nextflow [6] workflow for processing extensive traffic logs, and DeepLogBot, a bot detection framework that implements two complementary algorithms to identify and remove automated traffic. Additionally, we have developed an infrastructure that integrates download statistics directly into the PRIDE Archive interface, allowing users to sort datasets by total downloads - both raw and normalized by file count - and perform percentile-based analyses. Using these tools on 159.3 million download records from 2020 to 2025, we characterize global proteomics data reuse patterns, examining geographic distribution across countries and regions, temporal evolution of download activity, shifts in download protocol preferences, and the concentration of dataset reuse across the PRIDE collection. The workflow and resulting data are open-source and can be used by research labs and data depositors in grant reports and publications.

2 Methods

2.1 PRIDE Download Logs

Download logs are stored in a secure file system as compressed, comma-delimited text files. Each log entry includes a timestamp, dataset accession, filename, anonymized and non-reversible IP hash, download status, user country, download protocol (Globus, HTTP, Aspera, FTP), and dataset type. No personal or directly identifiable user data is stored. Log files are organized hierarchically by Protocol, Public/Private access, Year, Month, and Day. Individual log files can be large, ranging from 1 GB to 237 GB (Supplementary Notes Section S1).

2.2 nf-downloadstats: Log Processing Pipeline

Due to the large volume and size of download log files, we developed `nf-downloadstats` (`https://github.com/PRIDE-Archive/nf-downloadstats`), an open-source Nextflow workflow

for large-scale processing of the original anonymized log files (Figure 1A). Each processing step is implemented as an independent Python module, performing tasks such as removing incomplete transfers and filtering for PRIDE-specific records. A custom log file parser addresses the heterogeneity and scale of download log data; due to inconsistencies in log entries such as variable column structures and incomplete records, additional filters retain only complete transfers associated with PRIDE datasets.

To efficiently process the large volume of log files, parallelization is employed using a high-performance computing (HPC) environment managed via Slurm. Log files are processed in batches, with batch sizes and filtering criteria defined in a user-friendly YAML configuration file. The output is a consolidated 4.7 GB Parquet file containing 159,327,635 individual download records spanning January 2020 through January 2025. Each record includes the download date, geographic location (derived from IP geolocation), country, dataset accession, filename, download method (protocol), and an anonymized user identifier. The data covers 35,528 unique dataset accessions accessed from 235 countries.

For analysis, individual download events are aggregated at the *location* level, where each location represents a unique geographic coordinate. This aggregation produces 71,133 location profiles, each characterized by behavioral features including download volume, user counts, temporal patterns, and access characteristics.

2.3 Bot Detection Framework

We implemented two complementary bot detection algorithms, each combining Isolation Forest anomaly detection [7] with a distinct classification strategy. Both methods share a common feature extraction pipeline that computes 90+ behavioral features per location, organized into four categories: activity features (download counts, user statistics), temporal features (hourly/yearly entropy, working hours ratio), behavioral features (burst patterns, circadian deviation, coordination scores), and discriminative features (file exploration patterns, user authenticity scores). Full feature descriptions are provided in Supplementary Notes Section S3.

The *rule-based* method applies YAML-configurable threshold patterns to classify each location into one of three categories: **bot** (automated scraping or crawling), **hub** (legitimate automation such as institutional mirrors or CI/CD pipelines), or **organic** (human researchers). Patterns are evaluated sequentially, with the first match determining classification. This approach prioritizes interpretability and configurability. The *deep* method augments rule-based classification with additional behavioral feature engineering, including bot interaction features (download concentration, temporal irregularity, composite bot score) and bot signature features (request velocity, access regularity, session anomaly patterns). These 40+ additional features enable more nuanced separation of the three categories. The method incorporates a two-stage classification pipeline: Stage 1 separates organic from automated traffic, and Stage 2 distinguishes malicious bots from legitimate automation (hubs) using a discriminative scoring system with behavioral validation.

To evaluate both algorithms, we constructed a ground truth dataset of 1,411 labeled locations (88 bots, 44 hubs, 1,279 organic) using high-confidence heuristic criteria (Supplementary Notes Section S5). Both methods were evaluated on a 1-million record sample; we computed precision, recall, and F1 score per category, with 1,000-iteration bootstrap confidence intervals for the macro F1. The Deep method achieves the highest macro F1 score (0.775, 95% CI: 0.731-0.818), with perfect bot recall (1.000) and strong hub detection (F1 = 0.718), while the Rules method provides higher bot precision (0.506) but detects fewer hubs (F1 = 0.275, Supplementary Notes Section S6.)

We applied the Deep method - the best-performing algorithm - to the full dataset (Figure 1A), classifying each of the 71,133 locations as bot, hub (legitimate automation), or organic. Locations classified as bot (53.1%, representing 88.0% of download traffic) were removed prior to all downstream analyses (Figure 1B; Supplementary Notes Section S7). The remaining 33,354

non-bot locations (organic users and institutional hubs) account for 19.1 million downloads across 34,085 datasets and 213 countries. While the framework effectively removes the vast majority of automated noise, some borderline locations and download patterns inevitably fall near the decision boundary; a fraction of genuine downloads may be misclassified as bots (false positives) and, conversely, some low-volume automated activity may persist in the filtered data (false negatives). Users should therefore treat the cleaned download counts as robust estimates rather than exact values.

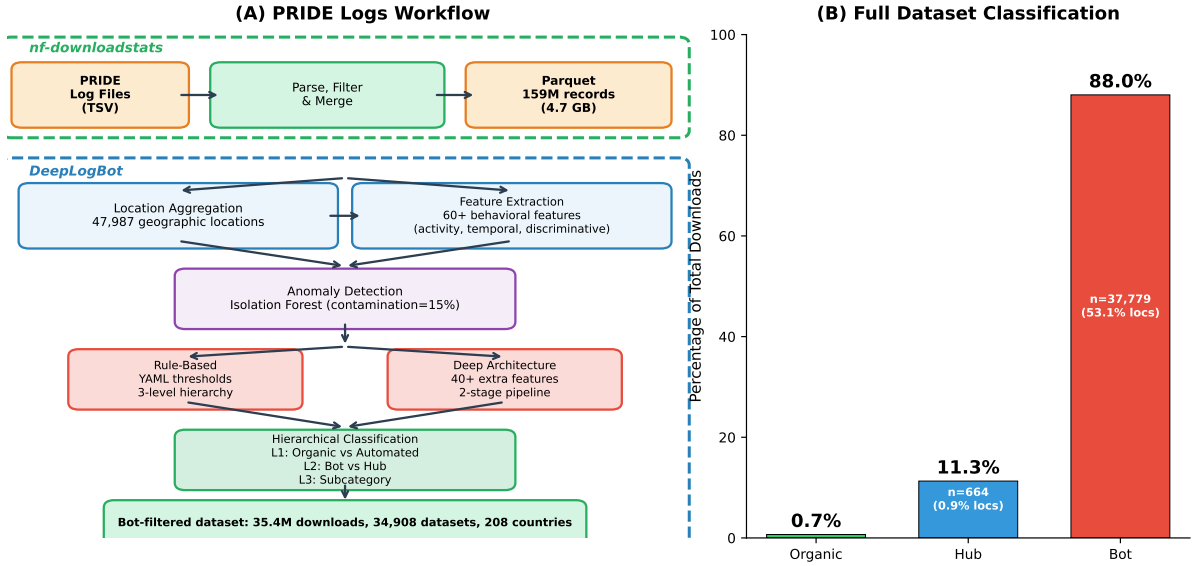

Figure 1: PRIDE Logs Workflow and classification results. (A) End-to-end pipeline comprising two components: `nf-downloadstats` (green dashed box) processes raw PRIDE download logs into a consolidated Parquet file, and DeepLogBot (blue dashed box) aggregates records to geographic locations and classifies them via two complementary algorithms (rule-based and deep architecture) using Isolation Forest anomaly detection and hierarchical classification. (B) Download share by classification category on the full dataset: bots account for 88.0% of all traffic despite representing 53.1% of locations.

2.4 PRIDE Download Statistics Visualization

The aggregated download statistics produced by `nf-downloadstats` are stored in MongoDB and Elasticsearch to enable fast searching and visualization within the PRIDE Archive web interface. For each dataset, the total number of downloads is displayed alongside a gradient bar indicating its download percentile, with higher intensity representing datasets in the top 1% of downloads (Figure 2). Additionally, a yearly trend chart is provided for each dataset, illustrating download activity over time. Datasets can be sorted by both total downloads per project and a normalized metric that accounts for the number of files within each project, enabling users to identify highly reused datasets for benchmarking or reanalysis.

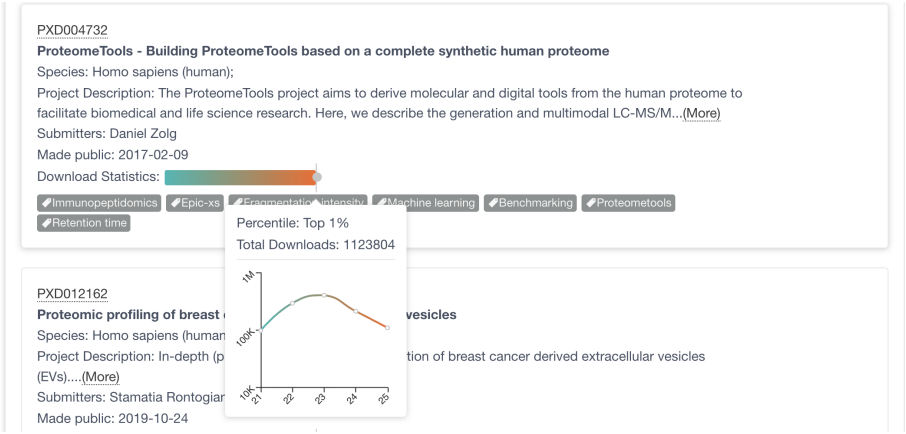

Figure 2: PRIDE Archive download statistics integration. Each dataset displays a gradient percentile bar and total download count. A popup trend chart shows yearly download activity, enabling users to assess dataset popularity and reuse trends directly within the PRIDE interface.

3 Results

3.1 Global PRIDE Usage Patterns

The PRIDE Archive serves users in 213 countries, demonstrating truly global reach (Figure 3). The top five countries by download volume are the United States (5.1M downloads, 26.8%), the United Kingdom (4.5M, 23.6%), Germany (4.3M, 22.5%), South Korea (869K, 4.6%), and Canada (691K, 3.6%). Europe accounts for the largest share of downloads (55.7%), followed by the Americas (30.7%) and Asia (11.4%).

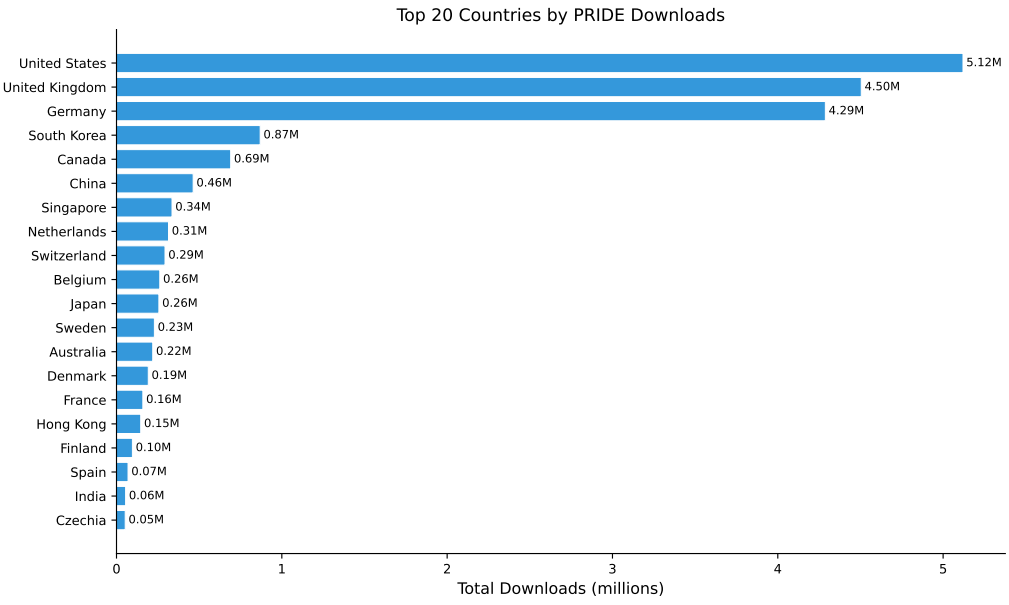

Figure 3: Geographic distribution of PRIDE downloads. Top 20 countries by download volume after bot removal (2020-2025).

To characterize the relationship between user base size and download intensity, we plotted total downloads against unique users for the top 50 countries (Figure 4A). Download patterns vary considerably across countries. Some, such as France (10,844 users, 15 downloads/user) and Canada (9,124 users, 76 downloads/user), show broad user bases with moderate per-user

activity, suggesting predominantly individual researchers. Others, such as Hong Kong (176 users, 832 downloads/user) and Singapore (703 users, 478 downloads/user), exhibit high per-user averages that likely reflect a small number of heavy institutional users or hub-like access points concentrating the median. Most countries fall on a spectrum between these extremes. This variation highlights that aggregate download volume alone does not fully capture the nature of data reuse, and that the balance between distributed individual access and concentrated institutional access differs markedly across regions.

Although European countries account for the majority of PRIDE downloads, yearly trends reveal shifting dynamics (Figure 4B). Germany showed a pronounced peak in 2023 (2.4M downloads) before declining, while the United Kingdom peaked in 2022 (1.8M) with subsequent moderation. Notably, PRIDE usage is growing in low- and middle-income countries (Figure 4C) including the grow in countries like India (55K downloads), Mexico (23K), and Brazil (12K). This growth suggests that PRIDE is increasingly serving as a resource for researchers in developing nations, supporting broader global participation in proteomics data reuse.

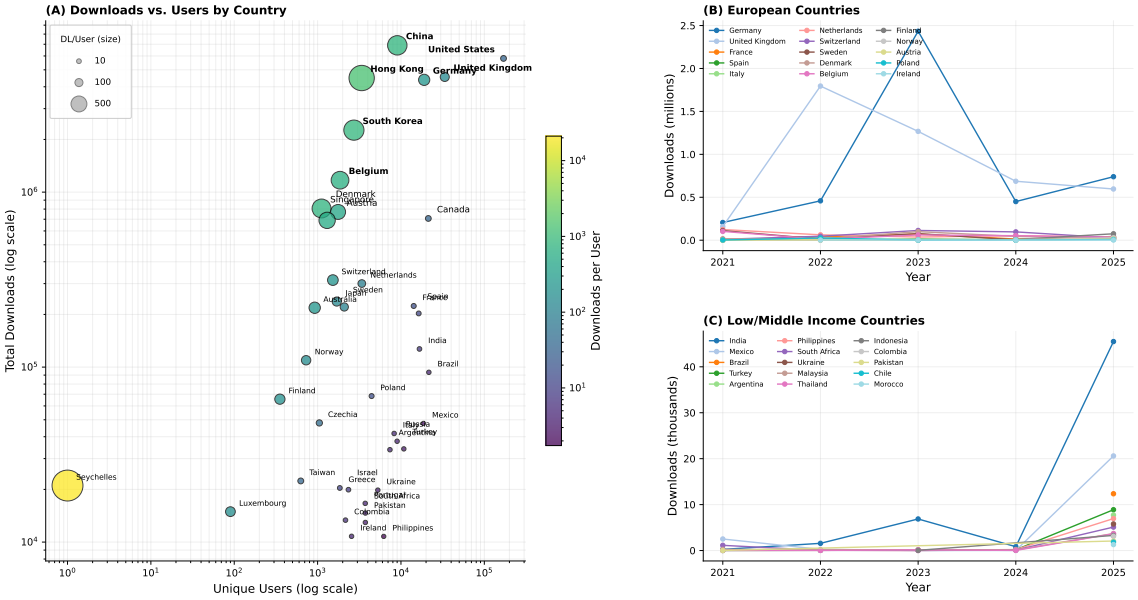

Figure 4: Global country-level download patterns. (A) Total downloads versus unique users for the top 50 countries by download volume; bubble size and color represent downloads per user on a log scale, revealing broad individual access (lower right) versus concentrated institutional access (upper left). (B) Yearly download trends for the top 15 European countries showing shifting dominance. (C) Yearly download trends for low- and middle-income countries, demonstrating accelerating adoption since 2024.

3.2 Temporal Trends

Download activity has grown substantially over the study period (Figure 5). Annual downloads increased from 1.2 million in 2021 to 3.5 million in 2022, peaking at 5.8 million in 2025 (after removing bots). The number of unique datasets accessed per year grew from 14,879 in 2021 to 18,621 in 2024, reflecting both the growing PRIDE collection and increasing data reuse.

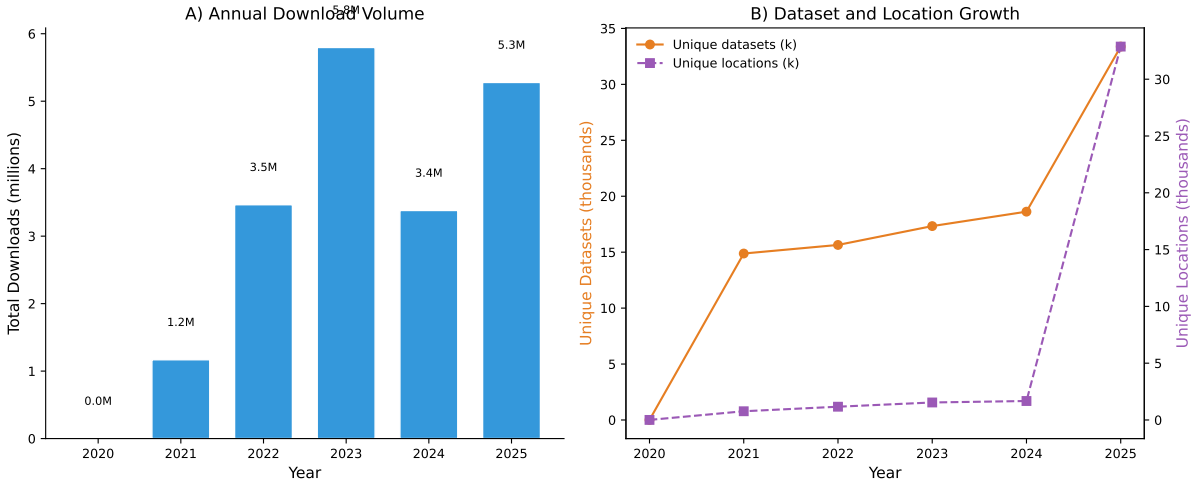

Figure 5: Temporal trends in PRIDE usage. (A) Annual download volume in millions. (B) Growth in unique datasets accessed and unique locations per year.

3.3 Protocol Usage

Download protocol preferences have shifted over the study period (Figure 6A). FTP was the dominant protocol in 2021, accounting for 65% of genuine downloads. HTTP overtook FTP in 2022 (54%), but FTP surged again in 2023 (79%), likely driven by institutional hub traffic that relies on FTP for bulk transfers. By 2025, HTTP re-emerged as the leading protocol (69%), reflecting broader adoption of web-based download tools. Despite superior transfer performance for large files, advanced protocols such as Aspera and Globus remain in early adoption stages, accounting for 3.3% and 1.0% of 2025 downloads respectively. To lower adoption barriers, we released `pridepy` [8] in March 2025, a Python-based command-line tool that abstracts protocol complexity and enables seamless switching between FTP, Aspera, and Globus transfers with a single command. A monthly breakdown of 2025 downloads (Figure 6B) shows emerging Aspera usage alongside sustained Globus adoption, indicating that providing user-friendly tooling can facilitate the transition to high-performance transfer protocols in scientific data repositories.

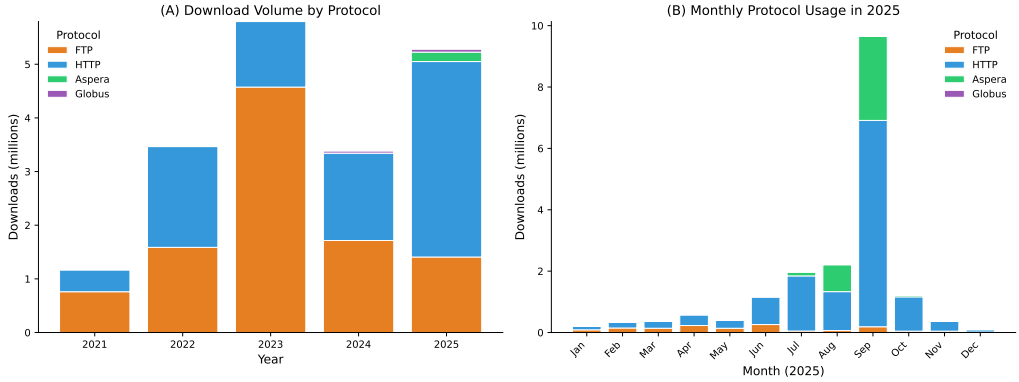

Figure 6: Download protocol usage. (A) Annual download volume by protocol (2021-2025), showing the alternation between FTP and HTTP dominance. (B) Monthly breakdown of 2025 showing protocol usage patterns.

3.4 Download Concentration

Dataset reuse follows a highly skewed distribution characteristic of heavy-tailed systems (Figure 7A). The Gini coefficient of 0.84 indicates substantial inequality: the top 1% of datasets (341

datasets) account for 43.3% of all downloads, the top 10% account for 77.2%, and the bottom 50% of datasets collectively represent only 3.1% of downloads. The median dataset has received 85 downloads, while the most popular exceeds 355,000. The rank-frequency distribution reveals a characteristic long tail, with download counts dropping steeply beyond the top 1% of datasets.

Importantly, the most downloaded datasets are not one-time events but show sustained reuse over multiple years (Figure 7B). Of the top 25 datasets, most have been actively downloaded in at least 4 of the 5 years covered (2021-2025), and PXD000001 - the first dataset deposited in PRIDE - has been downloaded every year. Several datasets exhibit pronounced temporal spikes (e.g., PXD021013 with 303K downloads, PXD029360 with 148K), likely reflecting their use as benchmarks in specific studies or community challenges. Others maintain steady download rates across years (e.g., PXD004732, PXD000561), suggesting their role as long-term reference datasets for the proteomics community. A ranking of the top 20 most downloaded datasets and extended analyses are provided in Supplementary Notes Section S8.

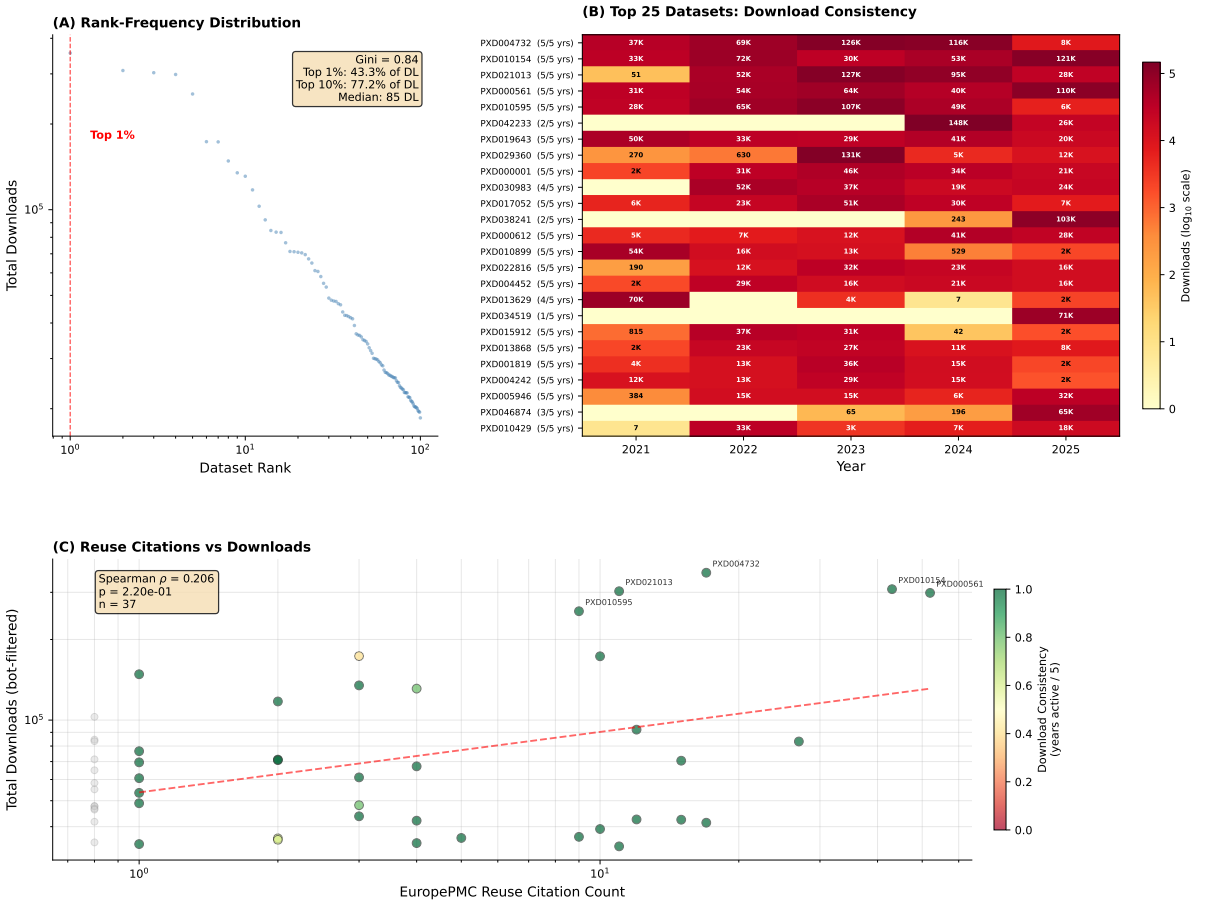

Figure 7: Dataset reuse concentration and consistency. (A) Rank-frequency distribution on log-log scale; the dashed red line marks the top 1% boundary, beyond which download counts drop sharply (Gini = 0.84). (B) Download consistency heatmap for the top 25 most downloaded datasets (2021-2025, after bot removal); color intensity represents download count on a $\log_{10}$ scale. Most top datasets show sustained reuse across 4-5 years, indicating their role as community reference and benchmark datasets. (C) Relationship between EuropePMC reuse citations and download volume for the top 50 most downloaded datasets (n−1 correction to exclude the original submission); point color indicates download consistency (fraction of years active out of five), grey points represent datasets with no independent reuse citations.

To assess whether download popularity reflects broader scientific impact, we queried EuropePMC for reuse citations of each dataset accession among the top 50 most downloaded

datasets, subtracting the original submission publication to count only independent reuse mentions (Figure 7C). Of these, 37 datasets (74%) have been cited in at least one independent publication beyond their original submission, with a median of 2 reuse citations and a maximum of 52 (PXD000561). While the correlation between download volume and reuse citation count shows a positive trend, it does not reach statistical significance (Spearman $\rho = 0.206$, $p = 0.22$), suggesting that high download counts do not simply mirror publication visibility. This indicates that sustained, multi-year download activity - rather than raw download volume alone - is the stronger signal of genuine community adoption and complements publication-based impact metrics.

3.5 Download Hubs

Our classification identified 664 download hubs distributed across 58 countries (Figure 8), accounting for 18.0 million downloads. These hubs represent institutions that systematically and continuously reanalyze public proteomics data [9], including institutional mirrors, research infrastructure nodes, and data aggregation services. The United States hosts the most hubs (155), followed by Germany (99), Japan (46), and the Netherlands (38), with total hub download volume led by the United States (4.7M), the United Kingdom (4.5M), and Germany (4.2M). The geographic spread of hubs - spanning all continents - demonstrates that systematic data reuse is not confined to a few centers but is a global phenomenon. Hub characteristics vary widely: some operate with very few users but extremely high per-user download rates (e.g., Dresden with 275K downloads/user from 8 users, consistent with a mirror), while others involve hundreds of users accessing data at moderate intensity (e.g., Melbourne with 186 users).

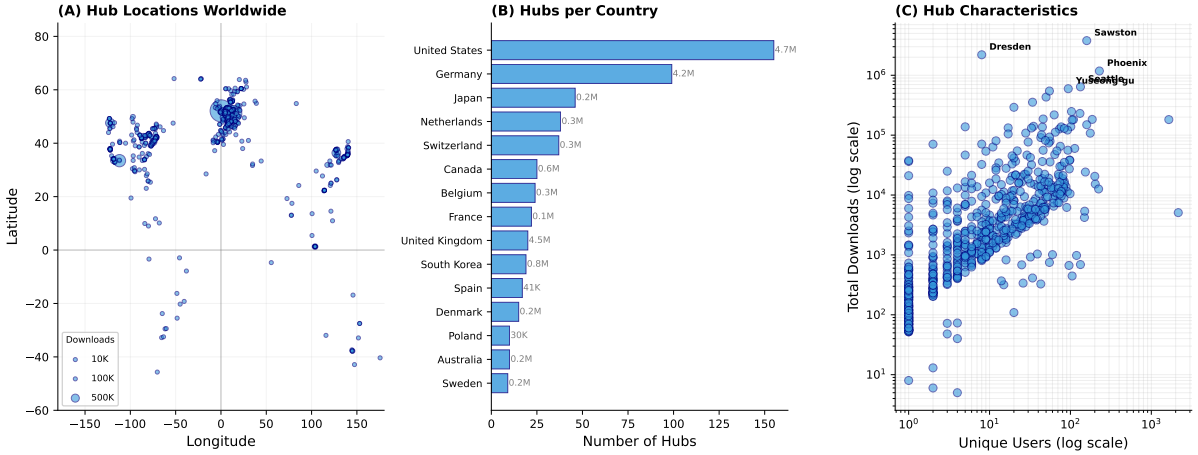

Figure 8: Distribution and characteristics of 664 identified download hubs. (A) Geographic locations, with bubble size proportional to download volume. (B) Top 15 countries by number of hubs, annotated with total hub download volume. (C) Hub diversity: unique users versus total downloads on log-log scale, showing the range from single-user mirrors to multi-user reanalysis infrastructure.

3.6 File Type Download Patterns

More than 81% of all downloads originate from five countries, and developing countries are largely absent from the top 10. Analysis of file type download patterns across regions reveals distinct usage profiles (Figure 9): raw instrument files dominate downloads in all regions, accounting for 72-73% of traffic in East Asia and North America. LMIC countries show a lower raw file proportion (54%) with a corresponding increase in result files and processed spectra [9]. This imbalance highlights that most users currently need to download and reprocess raw data from

scratch, even when search engine results already exist within the submission - underscoring the need for better infrastructure to make analysis results more discoverable and independently downloadable.

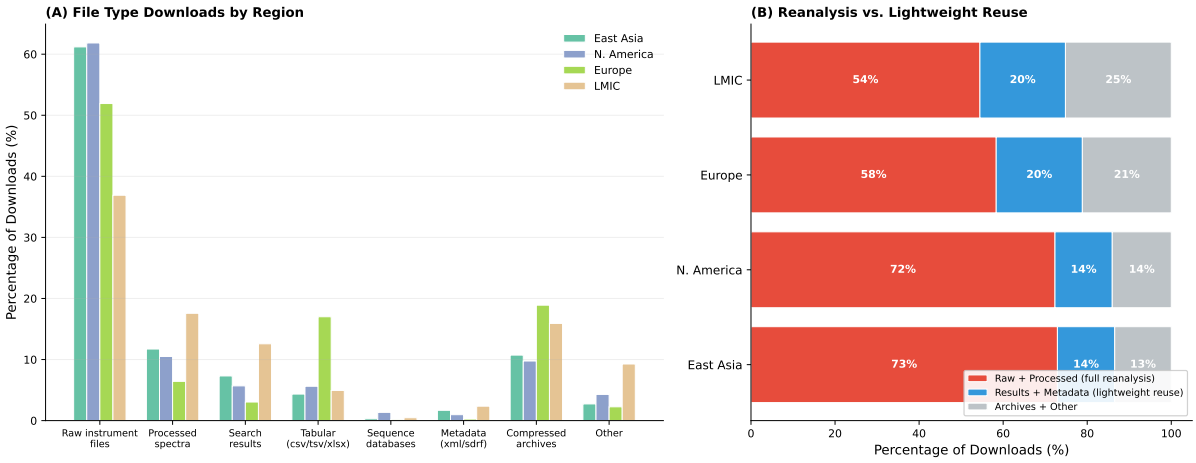

Figure 9: File type download patterns by region (after bot removal). (A) Percentage of downloads by file type category across four regions. LMIC countries show lower raw file usage and higher result/processed spectra downloads compared to East Asia and North America. (B) Aggregated view: raw + processed spectra (full reanalysis capability) versus results + metadata (lightweight reuse). LMIC countries have the lowest reanalysis-oriented profile (55%) and the highest lightweight reuse proportion (15%).

4 Discussion

That 88.0% of PRIDE download traffic is automated is consistent with the high bot prevalence observed across scientific data repositories, which offer unauthenticated, persistent, and predictable access to valuable content [10]. Without filtering, the most "popular" datasets may simply be the most bot-targeted, rendering raw download counts unreliable as impact indicators - a consequential problem given that download statistics are increasingly used by funding agencies, institutions, and researchers as evidence of data impact alongside traditional citation metrics [3]. DeepLogBot demonstrates that a scalable deep learning approach combining Isolation Forest anomaly detection with behavioral classification can effectively recover genuine usage signals from heavily contaminated logs. The framework processes over 159 million records, generalizes across diverse bot behaviors without manual rule tuning, and distinguishes legitimate automation (institutional hubs) from harmful scraping - a distinction that rule-based methods alone handle poorly (hub F1 = 0.275 vs. 0.718). As AI-driven platforms that perform large-scale automated reanalysis become more prevalent, the boundary between harmful scraping and beneficial programmatic access will become increasingly blurred, and repositories will need adaptive classification schemes that evolve alongside legitimate automation patterns.

PRIDE downloads have grown from 1.2 million in 2021 to 5.8 million in 2025, confirming accelerating data reuse across a geographically broad user base (213 countries). Download intensity varies markedly: some countries exhibit broad individual user bases (e.g., United States with 64K users), while others show concentrated institutional access (e.g., Hong Kong with 832 downloads/user, Singapore with 478 downloads/user), suggesting that the nature of reuse - individual exploration versus systematic reanalysis - differs between research communities. The 664 download hubs we identified reveal a global infrastructure of institutional data consumers, from single-user mirrors performing full-repository synchronization to multi-user reanalysis centers processing hundreds of datasets. This hub distribution provides an empirical map of where pro-

teomics bioinformatics infrastructure exists and can inform ProteomeXchange decisions about mirror placement, edge caching, and regional resource allocation - for instance, countries with growing user bases but no local hubs (e.g., India, Brazil, Mexico) may benefit from targeted infrastructure support. FTP and HTTP have alternated as the dominant protocol, with the shift to HTTP dominance in 2025 (69%) reflecting both broader individual adoption and the growing importance of API-based programmatic access; advanced protocols (Aspera, Globus) remain in early stages (1.4% of genuine traffic), but tools such as `pridepy` [8] should lower adoption barriers as datasets continue to grow in size.

Dataset reuse is highly concentrated (Gini = 0.84), with the top 1% of datasets accounting for 43.3% of all downloads. While community reference datasets such as ProteomeTools (PXD004732) show sustained multi-year reuse - likely because its comprehensive synthetic peptide spectral libraries serve as training data for machine learning models, retention time predictors, and spectral library search engines across the field - the "long tail" of rarely downloaded datasets should not be disregarded: these datasets may gain future value through meta-analyses, machine learning applications, or integration into multi-omics studies. Repositories can better serve both ends of this distribution by investing in improved discoverability - richer metadata, curated tags, and recommendation systems - alongside prioritized access for high-demand datasets. Regional differences in file type usage, with LMIC countries showing higher reliance on processed results rather than raw files, suggest that computational capacity and bandwidth constraints shape reuse patterns. The dominance of raw file downloads across all regions (Figure 9A) indicates that researchers currently lack easy access to analysis results within submissions, forcing them to re-download and reprocess raw data even when search engine outputs already exist. To address this, the PRIDE team is developing dedicated infrastructure for discovering, browsing, and downloading result and analysis files independently of the full raw dataset, enabling researchers with limited computational resources to directly access quantification tables, identification lists, and processed spectra without the overhead of re-running search engines. In parallel, the PRIDE team will prioritize SDRF sample metadata annotation [4] for the most downloaded and community-relevant datasets identified in this study, making these high-impact submissions immediately reusable through standardized experimental design descriptions. Several complementary efforts support this vision: `quantms` [11] generates standardized reanalysis outputs from public datasets, the PTMExchange initiative (`https://www.proteomexchange.org/ptmexchange`) provides curated post-translational modification results, and the PRIDE team is collaborating with developers of widely used search engines - including DIA-NN [12], MaxQuant [13], and MSFragger [14] - to define standardized submission guidelines that ensure result files, quantification tables, and metadata are structured for immediate reuse. More broadly, the `nf-downloadstats` pipeline and DeepLogBot framework are applicable to any open data repository facing similar challenges, including genomics (ENA/SRA), structural biology (PDB), and metabolomics (MetaboLights) resources.

5 Conclusion

We present the PRIDE Archive download tracking infrastructure and the first comprehensive analysis of download patterns from the PRIDE proteomics archive, covering 159 million records over five years. The infrastructure comprises `nf-downloadstats`, a scalable Nextflow pipeline for processing large-scale download logs, and DeepLogBot, a bot detection framework with two complementary algorithms achieving up to 0.775 macro F1. After removing 88.0% of traffic identified as automated, we obtain reliable usage metrics for 19.1 million genuine downloads spanning 34,085 datasets.

Our analysis reveals a globally distributed user base led by the United States, the United Kingdom, and Germany, a transition from FTP to HTTP-based access with emerging adoption of high-throughput protocols (Aspera, Globus), and a highly concentrated dataset reuse distri-

bution. On average, any PRIDE dataset file has been downloaded at least 30 times from 2021 to 2025, and more than 96% of the datasets in PRIDE have been downloaded at least once.

A particularly noteworthy finding is the identification of 664 download hubs distributed across 58 countries, accounting for 18.0 million downloads (11.3% of total traffic). These hubs represent research groups and institutions that systematically reanalyze public proteomics data - whether to complement their own in-house experiments or to build community-wide resources such as `quantms` [11], PeptideAtlas [15], GPMDB [16], Scop3P [17], and MatrisomeDB [18]. The global distribution of these hubs reinforces the role of PRIDE as a centralized, standardized, and reliable repository for proteomics data worldwide: rather than requiring data to be replicated and stored across multiple national or regional archives, the community benefits from a single curated resource from which data can be accessed and reanalyzed anywhere in the world.These findings provide evidence for the growing impact of open proteomics data and offer actionable insights for repository development.

The PRIDE team, through `pridepy` [8] and ongoing infrastructure development, will continue releasing tools and features that enable researchers to discover, query, and download result files - including protein and peptide identifications, quantification tables, and processed spectra - independently of the full raw dataset. This is particularly important for researchers in low- and middle-income countries, who, as our file type analysis shows, rely more heavily on processed results than on raw files. Beyond standard community file formats such as mzIdentML and mzTab, we will collaborate with developers of widely used search engines to improve the representation and standardization of result-level information deposited in PRIDE, ensuring that analysis outputs are structured for immediate reuse.

The highly skewed reuse distribution - where the top 1% of datasets account for 43.3% of all downloads while half of all datasets collectively represent only 3.1% - highlights the need for improved discoverability of valuable but underutilized datasets. To address this, PRIDE will invest in richer metadata annotation through SDRF sample descriptions [4] for the most downloaded and community-relevant datasets, deploy quality control reports generated by tools such as pmultiqc [19], and develop recommendation systems that surface relevant datasets based on experimental similarity rather than popularity alone. These efforts aim to lower the barrier to finding and reusing the "long tail" of datasets that may be highly relevant to specific research questions but currently lack the visibility to attract broad download activity.

More broadly, the `nf-downloadstats` pipeline and DeepLogBot framework are freely available and applicable to any open data repository facing similar challenges, including genomics (ENA/SRA), structural biology (PDB), and metabolomics (MetaboLights) resources.

Data and Code Availability

The `nf-downloadstats` pipeline is available at `https://github.com/PRIDE-Archive/nf-downloadstats` and the DeepLogBot software at `https://github.com/ypriverol/deeplogbot`, both under the Apache 2.0 license. Download log data is available upon request from the PRIDE team.

Funding

This work was supported by EMBL core funding; Wellcome Trust [208391/Z/17/Z, 223745/Z/21/Z]; Biotechnology and Biological Sciences Research Council [APP9749, BB/S01781X/1, BB/T019670/1, BB/V018779/1, BB/X001911/1, BB/V018779/1]. Funding for open access charge: Wellcome.

Acknowledgements

S.H. implemented the nextflow workflow; and the collected the data; J.B implemented web interface for the downloads components; C.B, S.K. implemented the integration of the statistics components in the backend of PRIDE and databases; D.J.K, N.S.J., B.B-H., N.M. contributed to review the manuscript; the data generated and curate some of the datasets; M.R.D. generated the infrastructure for logs anonimization and provided the logs files to PRIDE team, J.A.V. review the manuscript, Y.P-R design the study; developed the bot detection framework; performed the analysis and wrote the manuscript. We thank the PRIDE team for their support and feedback on the development of the download tracking infrastructure and analysis. We also wants to thanks professor Bernard Kuster for the original discussion about this topic in 2024 during the 2024 HUPO conference in Dresden

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Yasset Perez-Riverol, Chakradhar Bandla, Deepti J Kundu, Selvakumar Kamatchinathan, Jingwen Bai, Suresh Hewapathirana, Nithu Sara John, Marc Riera Duocastella, Maria D Vibranovski, Henning Hermjakob, and Juan Antonio Vizcaíno. The PRIDE database at 20 years: 2025 update. *Nucleic Acids Research*, 53(D1):D543–D553, 2025. doi: 10.1093/nar/gkae1011.

[2] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, 2016. doi: 10.1038/sdata.2016.18.

[3] Yasset Perez-Riverol, Andrey Zorin, Gaurhari Dass, Manh-Tu Vu, Rui Xu, Henning Hermjakob, and Juan Antonio Vizcaíno. Quantifying the impact of public omics data. *Nature Communications*, 10(1):3512, 2019. doi: 10.1038/s41467-019-11461-w.

[4] Chengxin Dai, Anja Füllgrabe, Julianus Pfeuffer, Elizaveta M Solovyeva, Jingwen Deng, Pablo Moreno, Selvakumar Kamatchinathan, Deepti Jaiswal Kundu, Nancy George, Silvie Fexova, et al. A proteomics sample metadata representation for multiomics integration and big data analysis. *Nature Communications*, 12(1):5854, 2021. doi: 10.1038/s41467-021-26111-3.

[5] Michael C Orr, John S Ascher, and John Pickering. AI bots threaten online scientific infrastructure. *Nature*, 641(8064):852, 2025. doi: 10.1038/d41586-025-01602-1.

[6] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319, 2017. doi: 10.1038/nbt.3820.

[7] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008. doi: 10.1109/ICDM.2008.17.

[8] Selvakumar Kamatchinathan, Suresh Hewapathirana, Chakradhar Bandla, and Yasset Perez-Riverol. pridepy: A Python package to download and search data from PRIDE database. *Journal of Open Source Software*, 10(107):7563, 2025. doi: 10.21105/joss.07563.

[9] Yasset Perez-Riverol. Proteomic repository data submission, dissemination, and reuse: key messages. *Expert Review of Proteomics*, 19(7-12):297–310, 2022. doi: 10.1080/14789450.2022.2160324.

[10] Imperva. Bad bot report 2023: The account takeover edition. Technical report, Imperva Inc., 2023. URL `https://www.imperva.com/resources/reports/2023-bad-bot-report/`. Annual analysis of automated bot traffic patterns across the internet.

[11] Chengxin Dai, Julianus Pfeuffer, Hong Wang, Ping Zheng, Lukas Käll, Timo Sachsenberg, Vadim Demichev, Mingze Bai, Oliver Kohlbacher, and Yasset Perez-Riverol. quantms: a cloud-based pipeline for quantitative proteomics enables the reanalysis of public proteomics data. *Nature Methods*, 21(9):1603–1607, 2024. doi: 10.1038/s41592-024-02343-1.

[12] Vadim Demichev, Christoph B. Messner, Spyros I. Vernardis, Kathryn S. Lilley, and Markus Ralser. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods*, 17(1):41–44, 2020. doi: 10.1038/s41592-019-0638-x.

[13] Jürgen Cox and Matthias Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12):1367–1372, 2008. doi: 10.1038/nbt.1511.

[14] Andy T. Kong, Felipe V. Leprevost, Dmitry M. Avtonomov, Dattatreya Mellacheruvu, and Alexey I. Nesvizhskii. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nature Methods*, 14:513–520, 2017. doi: 10.1038/nmeth.4256.

[15] Frank Desiere, Eric W Deutsch, Nichole L King, Alexey I Nesvizhskii, Parag Mallick, Jimmy Eng, Sharon Chen, James Eddes, Sandra N Loevenich, and Ruedi Aebersold. The PeptideAtlas project. *Nucleic Acids Research*, 34(suppl_1):D655–D658, 2006. doi: 10.1093/nar/gkj040.

[16] Robertson Craig, John P Cortens, and Ronald C Beavis. Open source system for analyzing, validating, and storing protein identification data. *Journal of Proteome Research*, 3(6):1234–1242, 2004. doi: 10.1021/pr049882h.

[17] Pathmanaban Decoster, Eliane Nkuipou-Kenfack, Tim Van Den Bossche, Gerben Menschaert, Lennart Martens, Kris Gevaert, Bert Coornaert, Mathieu Versele, Elvis Ndah, Michael C Costanzo, et al. Scop3P: a comprehensive resource of human phosphosites within their full context. *Journal of Proteome Research*, 22(1):106–118, 2022. doi: 10.1021/acs.jproteome.2c00167.

[18] Xinhao Shao, Clarissa D Gomez, Nandini Kapoor, James M Considine, Christopher Grams, Yu (Tom) Gao, and Alexandra Naba. MatrisomeDB 2.0: 2023 updates to the ECM-protein knowledge database. *Nucleic Acids Research*, 51(D1):D1519–D1530, 2022. doi: 10.1093/nar/gkac1009.

[19] Qi-Xuan Yue, Chengxin Dai, Selvakumar Kamatchinathan, Chakradhar Bandla, Henry Webel, Asier Larrea, Wout Bittremieux, Julian Uszkoreit, Tom David Müller, Jinqiu Xiao, Juergen Cox, Philip Ewels, Vadim Demichev, Oliver Kohlbacher, Timo Sachsenberg, Chris Bielow, Mingze Bai, and Yasset Perez-Riverol. pmultiqc: An open-source, lightweight, and metadata-oriented QC reporting library for MS proteomics. *bioRxiv*, 2025. doi: 10.1101/2025.11.02.685980.