PROTEIN EVALUATION AND TARGETED ASSAY DEVELOPMENT USING CATALOG AND PEPTIDE BROWSER

The Ohio State University

Alex Joyce

ABSTRACT

Proteomics datasets can be complex to interpret, even in their processed forms when reduced to data matrices. To facilitate ease of access, two programs have been developed for two different target audiences. The first, CATalog, aims to make biofluid data available to clinicans for optimal biofluid selection and to serve as a baseline dataset for proteins of interest. The second, Peptide Browser, aims to make targeted assay development much more accessible by providing biologists with the necessary figures of merit that they need to determine the best peptides for their assay.

CONTENTS

1. Data Accessibility in Proteomics 3

    1-1: Survey of Current Proteomics Platforms 3

2. Data Collection and Initial Processing 5

3. Protein Exploration through CATalog 6

    3-1: CATalog Program Structure 7

    3-2: Investigating a Specific Protein of Interest 8

    3-3: Identifying Proteins Associated with Specific GO Terms 9

4. Targeted Assay Development with Peptide Browser 11

    4-1: Database and Program Structure 12

    4-2: Calibration Curves 13

    4-3: Interpretation of LOD and LOQ 14

    4-4: Experimental Data 15

Appendix A: Schematic of Instrumentation 18

Appendix B: Data Acquisition 19

Appendix C: Initial Processing 21

Appendix D: More Details on EncylopeDIA 22

Appendix E: Peptide Browser Program Structure 23

References 24

ONE: Data Accessibility in Proteomics

Interpreting processed proteomics data remains a significant challenge, even for researchers familiar with biological datasets. While mass spectrometry-based proteomics provides rich quantitative and qualitative insights, the processed results are often presented as dense tables of peptide and protein identifications, statistical metrics, and confidence scores—requiring domain expertise to extract meaningful conclusions. Traditionally, these datasets are distributed in the form of large Excel spreadsheets, which, while comprehensive, can be difficult to navigate and interpret for those without a background in proteomics.

To facilitate the dissemination of two proteomic datasets (fig 1) generated through tandem mass spectrometry, we have developed two user-friendly web-based platforms: CATalog, designed for clinicians to assist in protein selection, and Peptide Browser, which bridges the gap between biologists and instrument operators by simplifying targeted assay design. In this report, we discuss the challenges of interpreting processed proteomics data, the motivation behind developing these platforms, the methods used to generate the datasets, and how these tools improve data accessibility and usability.

A

Entry	Protein	Gene	AU	BU	CU	DU	EU	HU	GU	FU	BP	AP	CP	DP	EP	FP	GP
A0A5F5XC	Alpha-1-B	A1BG	31.3022	31.01467	32.28593	31.29986	31.22713	31.23362	32.30432	31.57507	31.47495	31.2439	31.37721	31.04954	31.37672	32.19719	31.54548
M3W0W4	Alpha-1-B	A1BG	19.8943	19.38823	20.47157	15.89057	15.84983	22.37332	23.09663	23.2305	15.4932	15.5589	17.95917	22.03437	21.22049	23.07406	22.7822
M3W3E7	Alpha-2-m	A2M PZP	25.25537	23.38098	25.26956	24.01522	24.58062	24.253	23.73027	23.66938	26.55741	25.27921	25.97455	25.22411	25.61347	26.50115	24.93321
A0A5F5Y1	Alpha-2-m	A2M PZP	25.44943	27.13949	24.82691	25.06006	28.99808	25.67655	28.10796	27.87625	30.68382	30.12956	31.00195	30.71086	30.59361	30.95099	30.37994
A0A5F5Y3	Alpha-2-m	A2M PZP	20.91451	19.99337	17.72418	17.59459	16.39557	19.23405	13.67135	14.19765	21.26607	22.26781	21.4243	17.16387	18.48532	17.71043	18.33721
A0A337S3	Alpha-2-m	A2ML1	22.35916	21.88088	21.09341	21.09989	24.83838	21.69763	22.90946	22.49598	26.58834	25.94046	26.57527	27.45799	26.63648	27.05787	26.80605
M3WN23	ATP bindin	ABCA6	20.19058	20.29625	20.83587	12.18237	16.16466	20.20884	21.37277	21.29095	11.23883	12.84915	12.472	12.58186	10	12.41046	10
A0A337S6	ATP bindin	ABCB9	20.98336	22.12493	20.80871	20.81384	21.46013	21.87961	23.46564	23.24842	23.91505	22.03228	25.14687	22.74032	23.44285	22.039	21.11833
A0A337RX	Putative pr	ABHD14B	25.50441	25.14283	25.0411	27.04927	25.90998	25.2098	25.96472	26.48494	22.01596	19.37539	20.82068	20.55368	19.72687	21.28781	21.87275
A0A337SD	ABI family	ABI3BP	21.42868	19.63318	20.13427	22.00841	20.50838	19.46856	18.77468	19.15224	19.60372	18.82175	18.41579	18.27213	18.44061	19.09218	20.38934
A0A337S1	Costars fa	ABRACL	17.85805	18.05548	17.423	18.40372	16.55828	17.51478	18.49716	17.76629	18.56669	20.30425	16.96605	19.51702	21.03646	18.11467	16.58243

B

Peptide	X0	X0pt01	X0pt0215	X0pt0467	X0pt1	X0pt215	X0pt467	X1	X2pt15	X4pt67	X10	X21pt5	X46pt8	X68pt1	X100		lod	loq
AAAAAAPV	-5	-2.61384	-5	-1.63438	-1.68278	-1.27011	-0.87173	-1.29385	-0.91503	-1.47896	-1.94132	-1.49223	-0.69371	-2.51564		0	-2.88798	0
AAAAATVV	-1.07354	-1.71091	-0.73502	-1.91476	-1.8227	-1.90542	-2.49218	-2.03092	-2.09202	-1.86159	-1.80332	-0.47872	-0.68568	-0.43074		0	-0.48816	0
AAAADGEI	-3.30724	-5	-5	-2.59001	-2.5609	-2.71741	-2.87497	-3.20122	-3.46461	-1.64214	-1.15204	-0.70138	-0.35051	-0.1097		0	-2.48412	-1.40801
AAAADTGI	-2.73617	-2.72546	-5	-2.27169	-2.71405	-5	-2.69673	-3.52433	-1.41629	-2.12998	-1.76698	-1.04258	-1.03644	-0.85908		0	-1.83025	-0.0709
AAAADTGI	-1.22144	-5	-5	-2.33925	-0.2038	-1.60729	-1.66638	-0.50464	-0.28853	-0.24883	-0.12527	-1.09386	-1.03146	-1.81667		0	-1.64172	0
AAAAGALV	-4.15853	-2.0577	-5	-3.23631	-2.02262	-4.18164	-2.60639	-2.62915	-3.08737	-3.15099	-3.29723	-4.021	-0.38526	-0.25804		0	-3.6007	0
AAAAGEHK	-2.57118	-1.65725	-5	-1.57863	-2.22521	-1.7953	-1.25211	-5	-3.46863	-5	-4.06392	-1.82389	-0.11479	-0.09279		0	-1.03527	0
AAAAGLGI	-3.59437	-5	-5	-3.14145	-1.77622	-1.84439	-2.91534	-2.29138	-2.87635	-1.71169	-1.27212	-2.26594	-0.28802	-0.26305		0	-2.61439	-0.29181
AAAANIDE	-0.40001	-0.85708	-1.46804	0	-0.72552	-5	-1.15619	-0.9993	-0.36925	-1.35154	-1.50339	-0.13874	-1.29094	-0.32196		0	0	0
AAAAPAGK	-3.86359	-4.24333	-5	-2.49079	-3.12701	-2.67878	-3.188	-2.34413	-5	-3.36364	-2.13482	-1.81287	-0.35807	-0.19811		0	-1.87619	0
AAAASAAE	-2.05948	-2.6314	-2.78417	-2.21194	-1.85279	-2.28808	-5	-2.53929	-3.20211	-2.50246	-1.80297	-0.65977	-0.23412	0		0	-1.56133	0
AAAATETS	-3.20999	-5	-5	-2.71604	-2.97753	-4.22575	-3.01923	-2.79893	-2.53174	-4.52248	-1.37982	-0.79176	-0.26972	-0.0703		0	-2.46224	-0.84628
AAADADVI	-1.1683	-1.16074	-1.03836	-1.83164	-2.23682	-3.5715	-1.64125	-3.17391	-1.73111	-3.13902	-2.0237	-0.91213	-0.22269	-0.25403		0	-0.98217	0
AAADDGEI	-5	-2.59965	-2.49258	-5	-5	-2.25466	-5	-5	-5	-5	-2.5884	-1.24285	-0.52588	-0.13524		0	-1.78901	0
AAADDGEI	-5	-5	-5	-5	-5	-5	-5	-3.23616	-5	-1.33656	-2.57081	-0.7264	-0.32856	-0.12505		0	-2.7048	-0.00207
AAADFATP	-2.21723	-1.54682	-2.06959	-2.80191	-2.78201	-1.42863	-1.84943	-3.26787	-3.55394	-2.3652	-5	-0.06056	-1.65247	0		0	-1.01643	0

Figure 1: Spreadsheets that compose the core of the mass spectrometry data we wish to distribute. A) Cat biofluid data showing log2 fold change values for relative abundance across samples. B) Mouse proteome data showing log10 normalized intensity values for individual peptides.

1-1: Survey of Current Proteomics Platforms

A wide variety of online or downloadable platforms exist to hold, process, or visualize proteomics data. UniProt [1], ProteomicsDB [2], The Human Protein Atlas [3], and The Global Proteome Machine [4] hold curated datasets with the goal of being easily accessible and searchable. Other data repositories such as MassIVE [5], Panorama [6], and PRIDE [7] focus on the dissemination of published data from many different types of experiments, with tools such as ProteomeXChange [8] facilitating user submissions. Scaffold [9], Simplifi [10], Amica [11], Limelight [12], and PeptideShaker [13] are multipurpose tools designed to help visualize, validate,

and interpret proteomics data, with a focus on users that are generating the data itself. Concurrently, PeptideAtlas [14], Picky [15], Passport [16], the CPTAC Assay Portal [17], and Skyline [18] assist users in designing targeted assays to identify proteins of interest.

These platforms were designed with one or more specific purposes in mind, however, they can be roughly divided into four categories. Curated databases such as Proteomics DB are easily accessible and searchable, however, the data they contain may not be directly translatable to detailed proteomic assays. Published databases such as Panorama (targeted quantitative data) or MASSIVE (untargeted, discovery data) are great for submitting and distributing data the data that we have generated, however, utilizing the data in the form that would be submitted to these repositories is not immediately useful to the end user.

Postprocessing tools such as Amica and Limelight are useful for data interpretation and visualization and have been developed with ease of use in mind. However, they are focused on the user analyzing existing data rather than developing new datasets. Furthermore, platforms that aim to help targeted assay development either lack support for the quantitative information that we wish to distribute (PeptideAtlas and Picky) or require significant domain knowledge to utilize effectively (CPTAC Assay Portal and Skyline).

TWO: Data Collection and Initial Processing

Both datasets were collected using tandem mass spectrometry (MS/MS) coupled with a high-performance liquid chromatography (HPLC) system. The instrument used was a hybrid quadrupole-orbitrap set to acquire data in a data-independent fashion. Briefly, thermo raw files acquired from the instrument were converted to mzml files [19] using msConvert [20]. The search engine EncylopeDIA [21] was then used to identify both peptides and proteins in the mzML file. It reported these identifications in the form of a quantitative report which was further processed for dissemination through the data's corresponding platform. A schematic of this workflow is shown in figure 2. Additional details on this process along with visualizations of an example dataset are in appendix A.

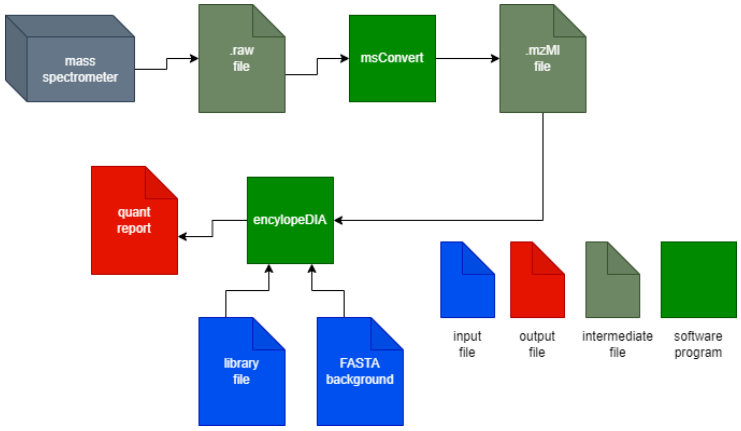


Figure 2: Schematic of the data processing workflow. Output from the instrument (.raw) was converted to a mzML file before being searched and quantified using EncylopeDIA.

THREE: Protein Exploration through CATalog

The first dataset that we aim to publish is a set of protein information obtained from eight cats. Each of these cats has a trio of paired samples from one of three biofluids: urine, serum, and plasma, constituting twenty-four samples per protein. These values can also be averaged together

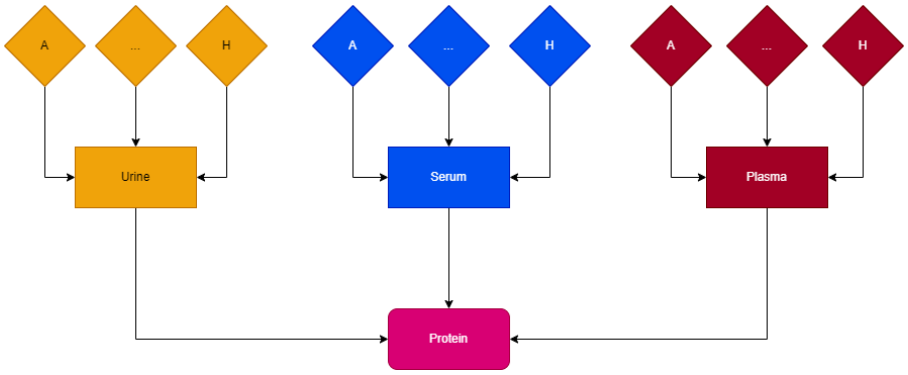


Figure 3: Diagram of the CATalog database structure. Each protein contains 24 total samples across three biofluids. Each biofluid contains a sample from eight cats, A through H.

At a more granular level, the data is presented as a series of proteins stored in a table format. Each row of the table represents a single protein; each protein has three values associated with it that are the average relative abundance. These values represent quantitative information that can assist clinicians in determining which biofluid a given protein of interest is most abundant in. To disseminate this data, we have developed the program CATalog, which is an interactive dashboard written in the R programming language that uses the Shiny interface for both the GUI and webhosting components.

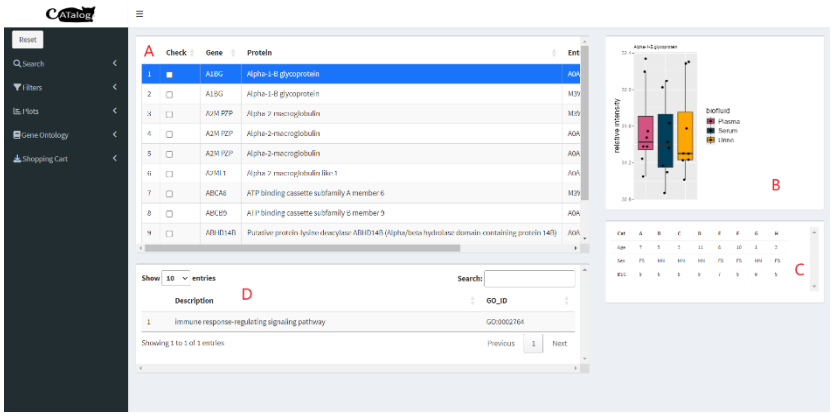


Figure 4: Screenshot of the main CATalog dashboard. Each row in the primary display table (A) represents a protein. Selecting a row brings up that protein's barplot (B) which contains relative abundance information from eight samples across three biofluids. Demographic information (C) for each sample is displayed on a table below the barplot. Gene ontology (GO) information (D) is featured in the table below the primary display table.

	Check	Gene	Protein	Entry	Urine	Plasma	Serum
1	☐	A1BG	Alpha-1-B glycoprotein	A0A5F5XCT0	31.5	31.6	31.5
2	☐	A1BG	Alpha-1-B glycoprotein	M3W0W4	20	19.9	20.2
3	☐	A2M PZP	Alpha-2-macroglobulin	M3W3E7	24.3	25.7	25.6
4	☐	A2M PZP	Alpha-2-macroglobulin	A0A5F5Y1U9	26.6	30.6	30.6
5	☐	A2M PZP	Alpha-2-macroglobulin	A0A5F5Y328	17.5	19.4	19.7
6	☐	A2ML1	Alpha-2-macroglobulin like 1	A0A337S3K5	22.3	26.7	26.8
7	☐	ABCA6	ATP binding cassette subfamily A member 6	M3WN23	19.1	11.6	11.3
8	☐	ABCB9	ATP binding cassette subfamily B member 9	A0A337S6A5	21.8	22.8	18.3
9	☐	ABHD14B	Putative protein-lysine deacylase ABHD14B (Alpha/beta hydrolase domain-containing protein 14B)	A0A337RXE3	25.8	20.8	21

Figure 5: Display of the database on the CATalog main dashboard. Each protein constitutes a row, with associated UniProt entries as well as their respective genes. Values for each biofluid are the averaged log2 fold change for intensities across the eight cats.

3-1: CATalog Program Structure

CATalog was written in the R programming language, implementing the Shiny package for GUI development. This framework separates the core program (app.R) into two main sections: the front-end GUI and the back-end server. The primary database is read from a .csv file in the project directory (also functioning as the git repository) by the 'setup' script that executes when the program is run. This script also reads in any necessary functions and loads any other required packages, abstracting much of the preparation code away from the core app.

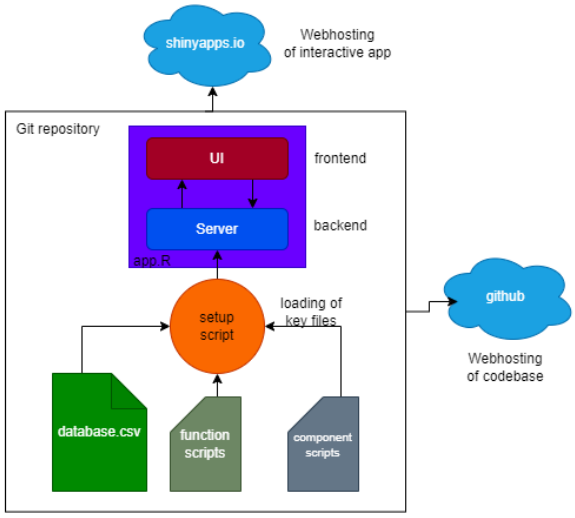


Figure 6: CATalog program structure. External dependencies are initialized from a setup script upon execution of the core app.

Webhosting of the app is done through shinyapps.io, which provides an easy way to distribute the app and subsequently the dataset through a single link. Similarly, the codebase is held on GitHub for version control. Both of these links are included in appendix # of this document.

3-2: Investigating a Specific Protein of Interest

Leptin is a peptide hormone primarily used by the body to maintain a normal weight and regulate metabolism [22], as such, it is known to be related to obesity [23]. Bob is a researcher studying leptin and wishes to determine what the baseline should be for the protein. Using the protein search function, he notices that the boxplot shows two cats that have elevated levels of leptin in plasma.

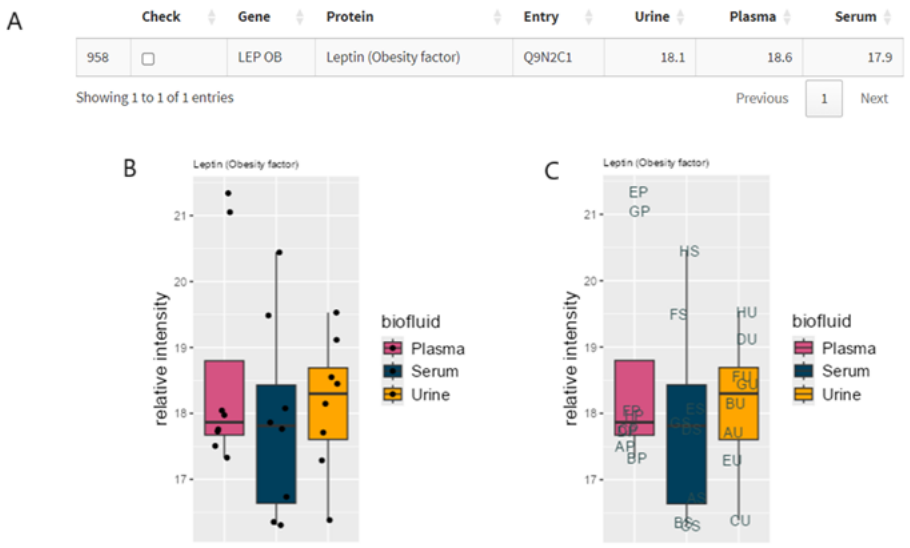


Figure 7: Leptin data from CATalog. A) Row of the main database corresponding to the protein, B) Unannotated boxplot showing the spread of values across each biofluid; each dot represents a different cat, C) annotated boxplot used to connect samples to demographic information. Cats E and G show an abnormally high relative abundance in plasma.

An annotation option connects each point on the boxplot to their respective sample, with the first letter corresponding to the cat and the second letter (P, S, and U) being the biofluid. The demographic legend below the boxplot indicates that cats E and G have higher than average BCS (body composition score, 7 and 6 respectively, the average is 5). This shows that leptin is upregulated in the plasma of obese cats.

Bob opts to filter these cats out of the dataset, removing them from the sample pool. This changes both the boxplot and the values seen in the foreground database. These changes highlight the importance of demographic effects on the data, as they can produce alterations in control datasets, potentially skewing the results. These results show Bob that obesity can be a potentially interfering factor when analyzing leptin or other metabolism-related proteins.

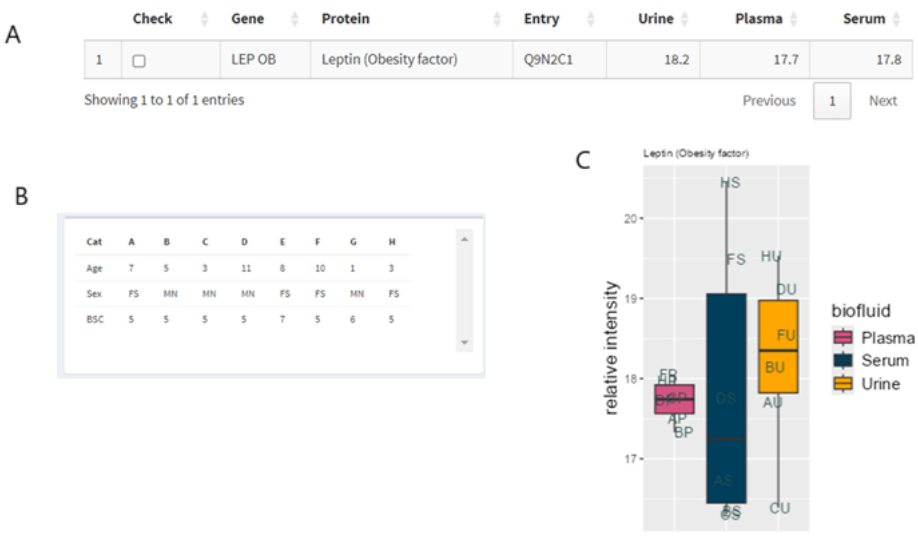


Figure 8: Leptin data after applying demographic filters. A). Average relative abundance values change due to the removal of cats E and G. B). Demographic table; cats E and G have a higher than average BSC value, indicating obesity. C). Updated boxplot to reflect the removal of those two cats.

Identifying Proteins Associated with Specific GO Terms

Gene ontology (GO) is a project that aims to annotate the functional relevance of genes across several species [24]. For the purposes of CATalog, a joint database of GO terms for *felis catus* (taxon ID: 9685) was constructed from UniProt. Terms include: biological process, cellular compartment, and molecular function. Alice is a researcher studying insulin resistance in cats, a known risk factor for diabetes [25]. Searching for 'insulin' in CATalog queries all three GO fields, returning a set of relevant proteins.

When assessing the fold change of these proteins across the three biofluids, Alice notices that the majority of the proteins are most abundant in urine. This information suggests that Alice should sample urine for her study, as such samples may provide the most accurate quantitative information for a targeted assay due to this abundance. Due to this, she applied a filter that removes all proteins that are not most abundant in urine (i.e., those whose log fold changes are more than one unit away from the maximum).

Entry	Urine	Plasma	Serum
A0A5F5XF81	19.1	13.9	14.6
M3XCH2	17	14.3	15.9
A0A2I2UB75	20.5	15.9	16.2
M3WF53	21.9	19.3	19.7
M3WLD9	20.7	19.2	19
M3WGH2	23.4	23.8	24.4
M3W8G3	26	20.9	21.2
M3W5W5	24.7	20.6	21.3
M3WSI8	29.8	29.1	29.2

Figure 9: Select proteins that have GO annotations relating to insulin. Notice that many proteins have a high relative abundance in urine compared to other biofluids.

Knowing that these proteins are concentrated in the urine and that they are related to insulin levels, Alice wants to know if they are associated with kidney function as well. The basis for this is that diabetes is a risk factor for chronic kidney disease (CKD) in humans [26] but is unproven in cats [27], therefore, she applies another GO filter on the set of insulin-related proteins in urine for ‘kidney’, resulting in a final set of proteins to target for an assay.

FOUR: Targeted Assay Development with Peptide Browser

The second dataset that we wish to publish is a comprehensive set of peptides linked to proteins from the mouse proteome. This data was obtained using parallel reaction monitoring to serve as a reference knowledge base for peptide quantitation. To facilitate, our dataset contains this information in the form of figures of merit, values that help determine how confidently a peptide can be detected and if it is truly quantifiable or not. This dataset was made accessible through Peptide Browser (figure 10), a platform similar in structure and architecture to CATalog, with the goal to assist in targeted assay development and to bridge the gap between biologists outside the proteomics field and core facility members who run the instrument.

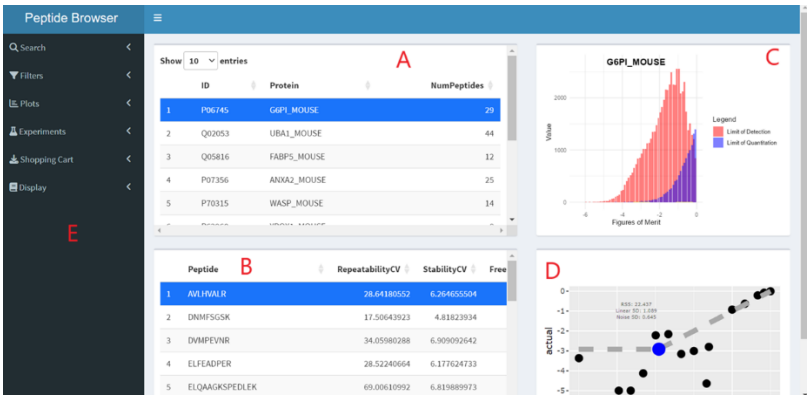


Figure 10: Screenshot of the main PeptideBrowser display. Protein selection is done through the top table (A), which brings up a table of associated peptides on the bottom table (B). A global histogram shows (C) shows the total distribution of figures of merit for all peptides in the database, with annotations for the selected protein. Calibration curves (D) displays quantitative information for the selected peptide. A sidebar (E) allows the user to select different options, such as viewing gene ontology information or filtering the datasets.

4-1: Database and Program Structure

The core of the Peptide Browser database is composed of a peptide portion and a protein portion along with several accessory datasets. Each protein in the database has a set of associated peptides, which in turn have quantitative data (figures of merit) alongside a set of replicate data (coefficient of variation) and FAIMS data. A more detailed breakdown of the program's structure is featured in appendix E.

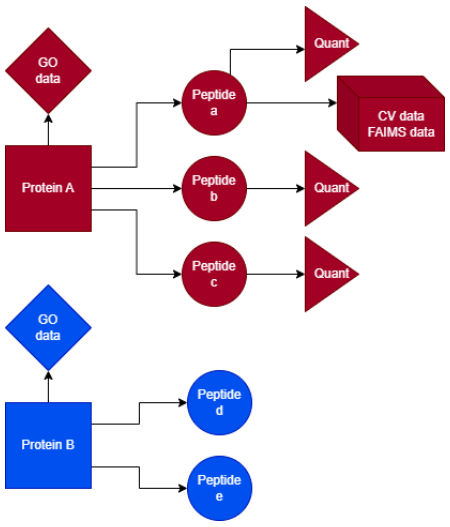


Figure 11: Data structure of the Peptide Browser database. Proteins are associated with GO terms as well as a set of peptides. Each peptide contains both quantitative figures of merit as well as CV based figures of merit.

To display all required information to the user, the lower table is bimodal. By default, selecting a protein from the upper table (fig 12a) will show the list of peptides associated with that protein (fig 12b, extracted from the peptide foreground). Subsequently selecting a peptide will show the calibration curve (fig 12c) or experiment plot, if that option is selected.

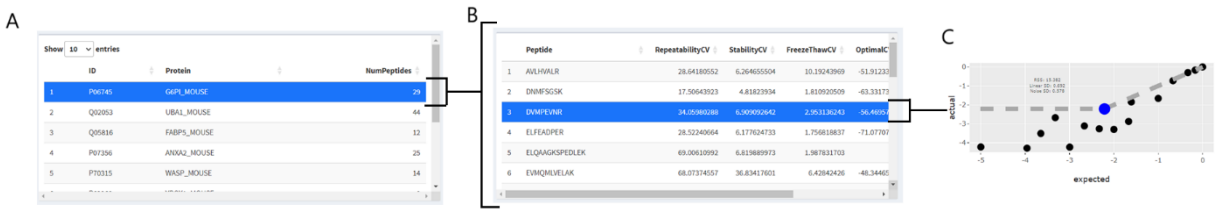


Figure 12: Peptide selection in PeptideBrowser. A) protein display table. B) peptide display table; this shows the peptides associated with the currently selected protein (G6PI). C) calibration curve for the selected peptide in B).

By choosing ‘gene ontology’ from the ‘Show lower table as’ option in the sidebar, the user can view similar GO information to what is available in CATalog. As with CATalog, this information appears in the lower table when the user selects a protein.

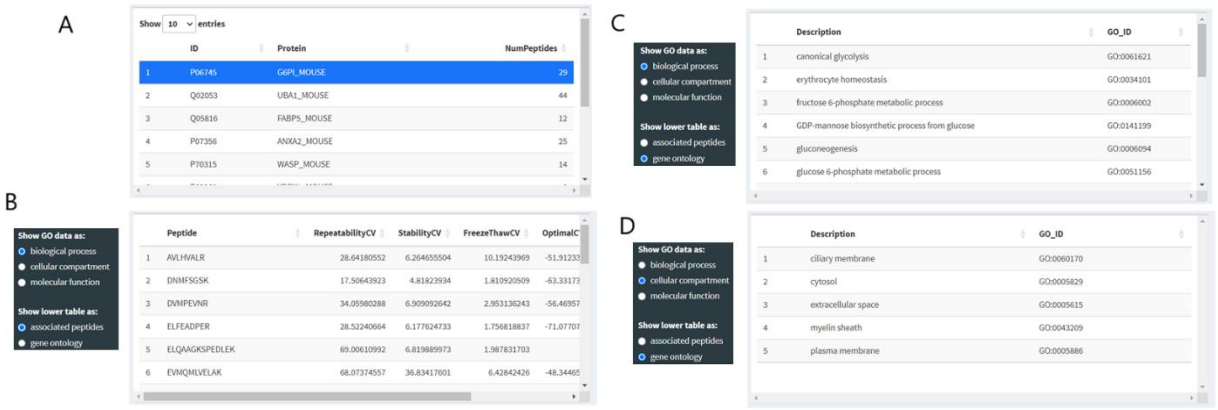


Figure 13: Gene ontology information in PeptideBrowser. Both the gene ontology information and the associated peptides for a given protein share the same lower table, with the type of information being toggled by the user through a sidebar button. A) main protein display showing the currently selected protein. B) lower display showing the associated peptides. C) lower display showing the GO information for G6PI in the 'biological process' category. D) lower display showing the GO information for G6PI in the 'cellular compartment' category.

4-2: Calibration Curves

Quantitative information contained in PeptideBrowser is derived from a dilution series used to build a calibration curve. The purpose of these curves is to measure the relationship between concentration (independent, known variable) and response (dependent, unknown variable). A stable relationship between concentration and the measured signal can be used to develop a model and predict the concentration for any given signal.

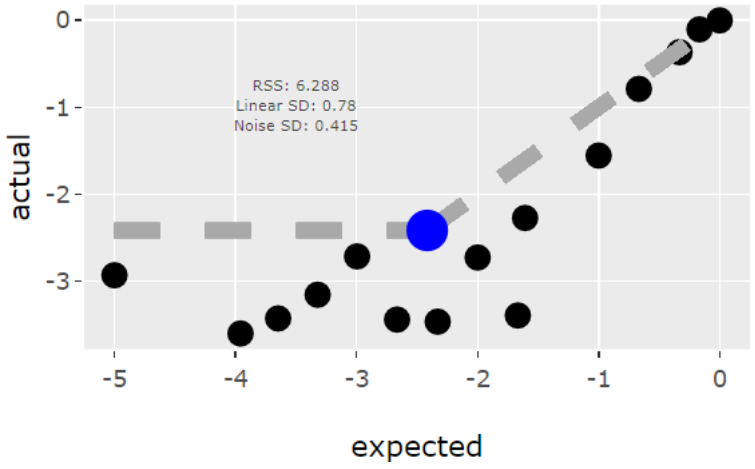


Figure 14: calibration curve for the peptide EVMQMLVELAK from the protein G6PI. Each black dot on the curve represents a

In the context of targeted proteomics, the calibration curve can be divided into two segments [28]. The linear (signal) portion is comprised of a linear model that attempts to fit the data to a slope of 1. The noise portion has a slope of 0 is composed of all points that have not been fit to the linear model. A blue dot on the graph indicates the transition point between these two regions. This is the limit of detection (LOD) which is the lowest amount of analyte that can be reliably detected [29]. Similarly, the limit of quantitation (LOQ) describes the lowest amount of analyte that can be reliably quantified.

These values were derived using a curve-fitting approach through a script in the EncylopeDIA program. This script allocated N – 1 points to be 'signal' and a linear model was fit, this repeated for each N until a curve had been fit for every possible allocation. The best fit was the curve that minimized the residual sum of squares (RSS) values. LOD is then determined by finding the largest y-value in the noise portion to establish the transition point. Another value, limit of quantitation, is calculated by adding a constant, 3 * σ_S, to the LOD, where σ_S is the signal's standard deviation.

4-3: Interpretation of LOD and LOQ

Consider the example peptides shown below.

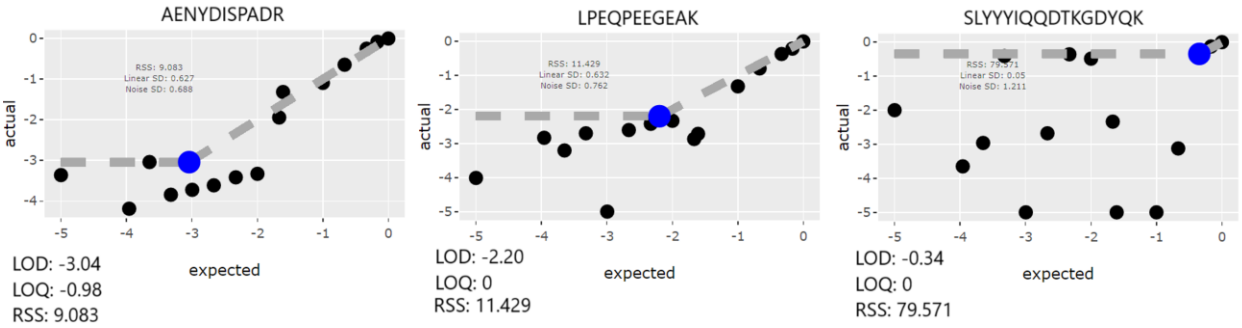


Figure 15: calibration curves for several select peptides. Quality of a peptide for targeted assays largely depends on minimizing three key metrics: LOQ, LOD, and RSS.

Of these peptides, only one is quantifiable, the first; an LOQ value of 0 indicates that the peptide cannot be quantifiable using existing data. The first peptide subsequently has the lowest LOD/LOQ and RSS values. Other peptides may share similar metrics, however, there is little consistency in their standard deviation values. Due to this variability, standard deviations are not considered an adequate assessment of peptide quality in PeptideBrowser. The platform incorporates these values into its analysis pipeline through a series of filters accessible from the sidebar. Values can be entered into a set of boxes to set thresholds while the filters themselves can be applied individually. An additional filter can also be applied to remove all peptides that have LOQ values of 0.

To help the user determine what a good cutoff point for the maximum LOD/LOQ values should be, he can look at a distribution of all the peptides associated with G6PI (glucose 6 phosphate isomerase [30]), which has 29 total peptides. A 'local' distribution shows the counts of peptides across the distribution of LOD/LOQ values for that protein. Selecting a peptide will annotate the local distribution based on the LOD (yellow) and LOQ (green) value of the selected peptide.

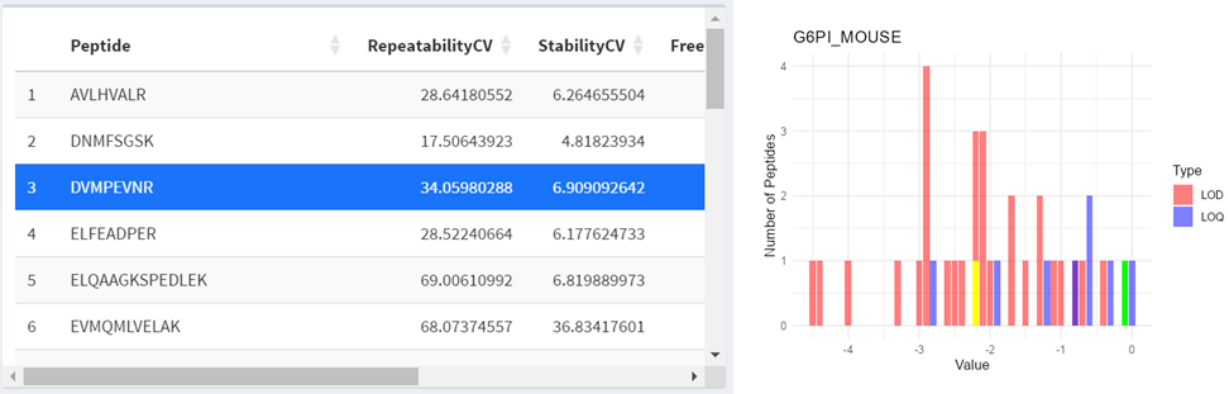


Figure 16: local distribution of LOD (red) and LOQ (blue) values in G6PI. Selecting a peptide annotates the distribution by highlighting the bin where that peptide's LOD (yellow) and LOQ(green) values are located.

To find additional proteins that one would want to consider for their assay, they can search for 'glucose' or 'gluconeogenesis' in the GO dataset to obtain a list of proteins related to the same pathway as G6PI. Additionally, a similar distribution of LOD/LOQ values can be viewed globally. If they selects a protein from these search results, this global annotation, which shows the figures of merit for all peptides in the dataset, will be annotated. This annotation highlights the bins containing the peptides for the selected protein, allowing one to quickly determine how easily a protein is to assay prior to viewing detailed peptide information.

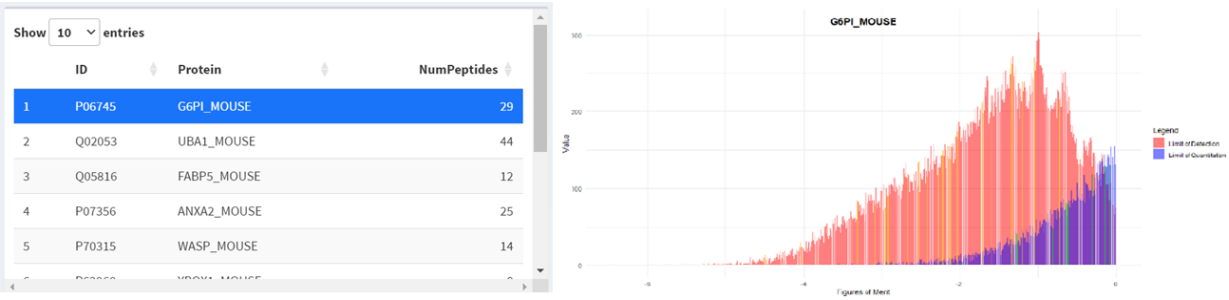


Figure 17: Global distribution of LOD/LOQ values for all peptides in the database. The bins that contain values corresponding to peptides associated with G6PI are highlighted. The main purpose of this is to quickly let the user look over each protein and find those that are easily detectable and/or quantifiable.

4-4: Experimental Data

PeptideBrowser also contains several additional figures of merit from a series of different experiments. The first three of these datasets, freeze-thaw, stability, and repeatability are focused

on finding the coefficient of variation, which is a ratio of standard deviation to mean. These values are displayed in the lower table with their corresponding peptides. Ideally, a peptide should have a low coefficient of variation across these three measurements, which would indicate less variability across repeated measurements.

	Peptide	RepeatabilityCV	StabilityCV	FreezeThawCV	OptimalCompVoltage
1	AVLHVALR	28.64180552	6.264655504	10.19243969	-51.91233749
2	DNMFSGSK	17.50643923	4.81823934	1.810920509	-63.33173482
3	DVMPEVNR	34.05980288	6.909092642	2.953136243	-56.46957179
4	ELFEADPER	28.52240664	6.177624733	1.756818837	-71.07707383
5	ELQAAGKSPEDLEK	69.00610992	6.819889973	1.987831703	
6	EVMQMLVELAK	68.07374557	36.83417601	6.42842426	-48.34465852

Figure 18: Experimental data as shown in the peptide display table. It should be noted that there are certain peptides that do not have specific types of experimental data, these are shown as empty cells.

The final experimental dataset relates to FAIMS (High-field asymmetric waveform ion mobility spectrometry), which is an apparatus attached to the front of a mass spectrometer; the purpose of which is to serve as an additional form of ion separation by applying electrical fields. It is commonly used in conjunction with liquid chromatography (LC) to help with peptide identification in complex mixtures [31]. Compensation voltage (called 'CompVoltage' in PeptideBrowser to avoid confusion with the coefficient of variation, CV) is a type of DC current that allows ions to pass through the device by compensating for ion drift [32]. These datasets can be found under the experiments tab, where a button can be used to replace the calibration curve with the desired experimental plot.

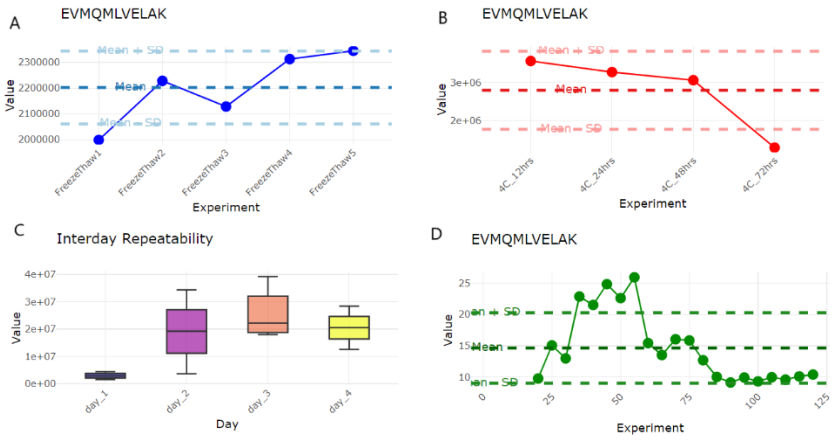


Figure 19: Plots of the experimental data corresponding to the peptide EVMQMLVELAK from GP6I. A) Freeze-thaw. B) Stability. C) Interday repeatability. D) Optimal FAIMS compensation voltage.

The repeatability dataset (fig 19c) consists of four measurements taken on four different days, giving sixteen total datapoints. To reduce the complexity of this data, this dataset was split into two different types of plots: an interday boxplot which shows the spread of data across each day and intraday line plots which show how each measurement within a given day compares to other measurements. Ideally, there should be low variation between measurements.

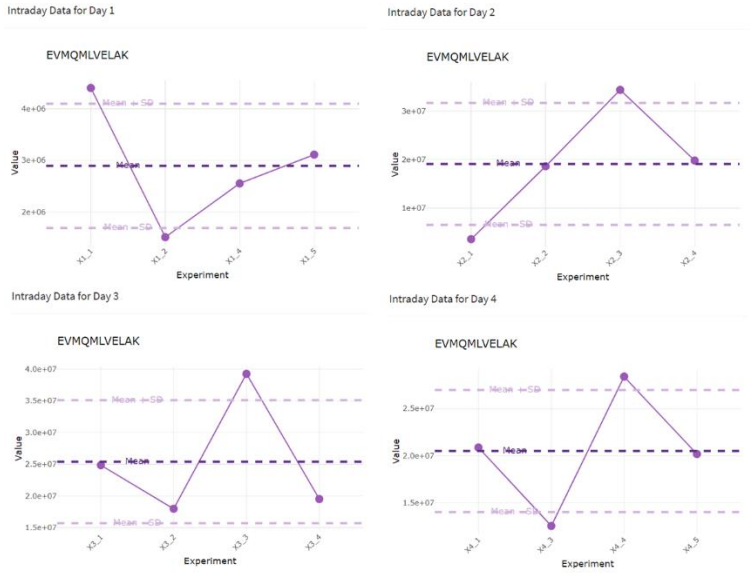


Figure 20: Intraday plots for the G6PI peptide EVMQMLVELAK. These plots are viewable by clicking on the boxes that make up figure 28c.

Appendix A: Schematic of Instrumentation

Fig 21 shows a diagram of the instrument used to collect these datasets. Tryptic peptides are separated using the HPLC system before being ionized through electrospray ionization (ESI) and injected into the instrument. The quadrupole serves as the first mass analyzer and selects for the precursor(s) of interest based on the experiment type. These ions pass through the C-trap before being stored in the ion-routing multipole where they are fragmented using collision-induced dissociation (CID). When enough product ions for a given precursor or set of precursors have accumulated, they are released from the collision cell and pass through the C-trap once again; this structure focuses them into the orbitrap, which is the second mass analyzer. This device generates MS2 level spectra which are then used for sequencing and subsequent protein identification.

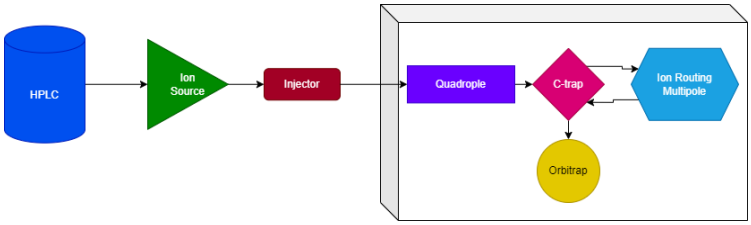


Figure 21: Example of a hybrid quadrupole-orbitrap mass spectrometer. Peptides elute from the column (HPLC) and are ionized through the ion source (i.e., ESI). The injector allows them to enter the system where precursors are analyzed (MS1) using a quadrupole. The ions move through the C-trap to the ion routing multiple, which is a collision cell. Following fragmentation, ions move back into the C-trap where they are focused and moved into the orbitrap to be analyzed (MS2).

This method selects specific peptides (precursors) using an initial mass analyzer and fragments them to produce product ions. These product ions are then analyzed using a second mass analyzer and the spectra (MS2) are then reported. One of the main purposes of this process is to sequence peptides in bottom-up experiments, which is vital for identification [33].

Appendix B: Data Acquisition

Proteomic data collected using mass spectrometry exists in a three-dimensional space. The following figure shows the structure of an example dataset collected using data-dependent acquisition (DDA) visualized using the program mzMine [34].

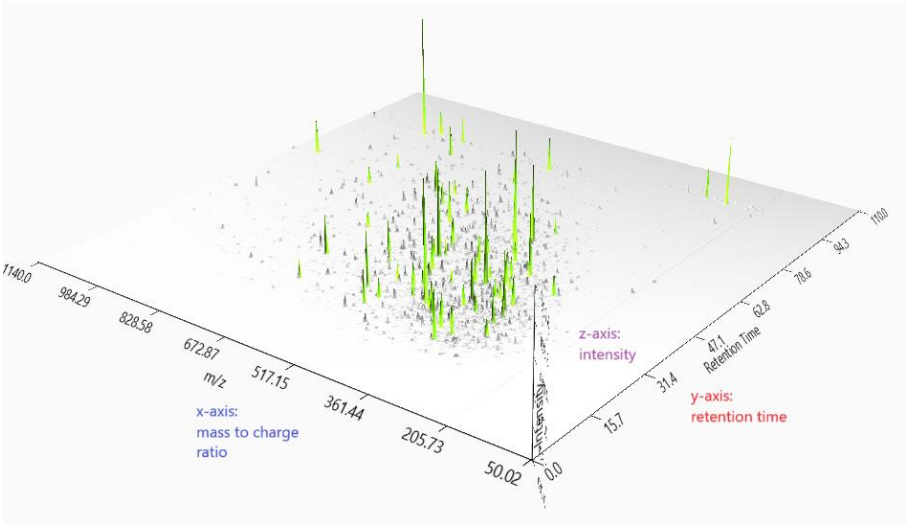
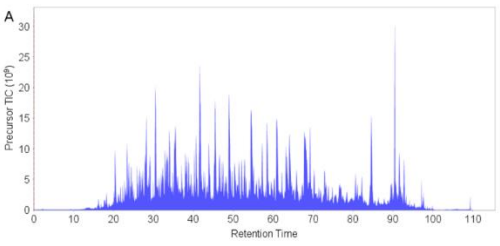


Figure 22: Visualization of a DDA dataset. This data contains both a retention time axis (Y) and a m/z axis (X). These twoaxes share a common intensity axis (Z).

Using this example, the y-axis contains the retention time which shows how long each peptide takes to leave the column [35]. In order to select peptides for fragmentation, the instrument must take periodic scans across retention time space which can be visualized using the raw file browser in EncylopeDIA. If the visualization is done across retention time space, then a total ion chromatogram (TIC) can be formed by summing the intensities of all precursors across the span of an experiment (fig 23a). Scans across this space are done on the millisecond scale (fig 23b), producing a precursor scan each time (fig 23a).



#	Spectrum Name	Scan Start Time	Precursor M/Z
1	controllerType=0 c...	0.003	-1
2	controllerType=0 c...	0.009	-1
3	controllerType=0 c...	0.017	-1
4	controllerType=0 c...	0.024	-1
5	controllerType=0 c...	0.028	252.643
6	controllerType=0 c...	0.029	202.635
7	controllerType=0 c...	0.03	310.193
8	controllerType=0 c...	0.03	223.64
9	controllerType=0 c...	0.031	216.632
10	controllerType=0 c...	0.031	210.135
11	controllerType=0 c...	0.032	303.185
12	controllerType=0 c...	0.033	244.664
13	controllerType=0 c...	0.033	266.659
14	controllerType=0 c...	0.034	296.177
15	controllerType=0 c...	0.035	207.627
16	controllerType=0 c...	0.035	206.619

Figure 23: Dataset from figure 22 visualized using the RAW file browser from EncylopeDIA. A) total ion chromatogram (TIC) of the data. B) Summary table of scans from the data.

A precursor scan is a type of mass spectrum that looks across the m/z axis to analyze what precursors are eluting off the column at that particular point in time (MS1 spectrum). In a DDA experiment, if there are precursors that exceed a given intensity threshold, they are selected for fragmentation, producing a product ion spectrum (MS2).

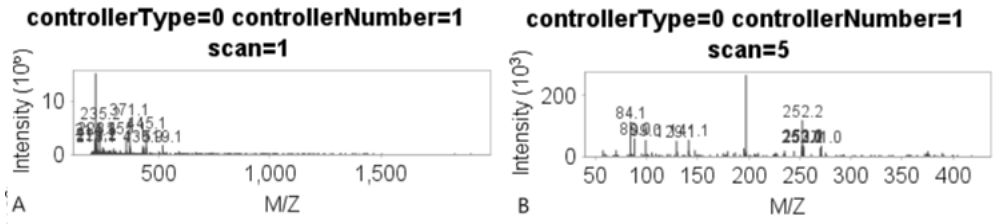


Figure 24: Mass spectra from the dataset in figures 22 and 23. Right- precursor spectrum (MS1). Left- product spectrum (MS2).

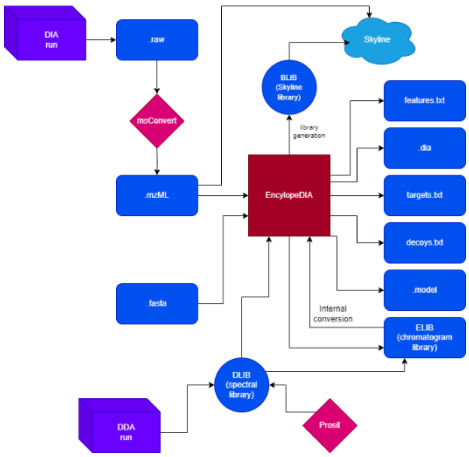
If we subdivide the m/z axis into regularly sized windows (isolation windows) and select every peptide contained within those windows for fragmentation, then the acquisition method is deemed to be data-independent (DIA). Using a similar framework, isolation windows can be drawn around predetermined precursor masses in a method of targeted proteomics known as parallel reaction monitoring (PRM).

Appendix C: Initial Processing

The initial format of the raw data depends on the instrument. For example, Thermo Fisher instruments generally produce .raw files through their Xcalibur software. In order to bridge these disparate file formats, the XML based format mzML was developed [19]. Conversion to mzML can be done through software such as msConvert [20]. There are several conversion algorithms available in msConvert, with one of the most commonly used algorithms being peak picking, a process that takes large amounts of spectral data and distills it down into a much more manageable state [35].

Appendix D: More Details on EncylopeDIA

Figure # shows the main workflow of EncylopeDIA. Briefly, the raw output is converted from its original file format (i.e., .raw) to an mzML file through MSConvert. This file is then loaded into EncylopeDIA alongside a library file and a .fasta file (background file for protein identification). The type of library file used depends on the type of experiment. DDA experiments use spectral libraries (DLIB) that are generated computationally through Prosit [36] or from a previous DDA run. DIA experiments use chromatogram libraries (ELIB) that are generated from DLIB files using an internal converter. Scoring is conducted using a variety of different features including several chromatogram-based scoring metrics such as co-elution, fragment ion intensity correlation, and peak shape similarity along with a primary score based on the X!Tandem hyperscore.

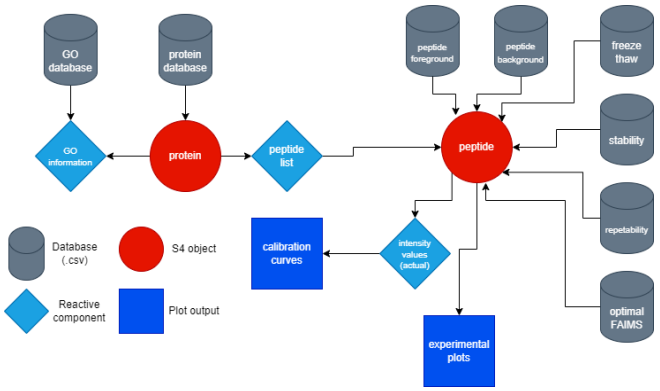


These scores are then aggregated and submitted to Percolator [37], a support vector machine (SVM) algorithm that determines the false discovery rate (FDR) through both machine learning and target-decoy analysis [38]. Percolator is implemented in both first-pass and final-pass steps with a retention time alignment step in the middle at 1% FDR to determine the highest scoring retention time point.

The primary outputs from EncylopeDIA are two text files (targets.txt and decoys.txt, derived from target-decoy analysis) that contain a list of the PSMs from the analysis along with their respective percolator scores, posterior error probabilities (PEP, how likely is a PSM to be incorrect), and q-values (false discovery rate (FDR) corrected p-values). A .dia file is a SQL-based summation of a .raw file used to quickly obtain information regarding isolation windows. Finally, a .model file contains the weights and other parameters used in the linear discriminant analysis to allow for repeat experiments.

Appendix E: Peptide Browser Program Structure

As with CATalog, the data itself is contained within the project directory as a set of .csv files in the 'data' subdirectory. The peptide portion is further subdivided into a 'background' dataset that contains the complete set of intensity values for each peptide and a 'foreground' portion, which contains coefficient of variation (CV) information and connecting information to the protein database. The 'foreground' database is what is shown to the user in the lower table.



Summary table of the CATalog database:

Filename	Role	Description
go_database.csv	Presented to the user	Contains GO information for proteins in the database.
peptide_background.csv	Hidden from the user	Used to construct the calibration curve; contains a complete dilution series as well as metadata.
peptide_foreground.csv	Presented to the user	Contains coefficient of variation values and optimal compensation voltage for each peptide (if applicable). Linked to both the protein database and the peptide background.
protein_database.csv	Presented to the user	Populates the top table and facilitates protein selection. This database also connects to both the GO database and the peptide foreground.
freeze_thaw.csv	Hidden from the user	Contains the complete data for the freeze-thaw experiment.
repeatability.csv	Hidden from the user	Contains the complete data

		for the repeatability experiment. This dataset is further divided into ‘interday’ and ‘intraday’ experiments.
stability.csv	Hidden from the user	Contains the complete data for the stability experiment.
optimal_cv.csv	Hidden from the user	Contains information regarding the optimal FAIMS compensation voltage for each peptide.

References

1. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D115-9. doi: 10.1093/nar/gkh131. PMID: 14681372; PMCID: PMC308865.
2. Schmidt T, Samaras P, Frejno M, Gessulat S, Barnert M, Kienegger H, Krcmar H, Schlegl J, Ehrlich HC, Aiche S, Kuster B, Wilhelm M. ProteomicsDB. Nucleic Acids Res. 2018 Jan 4;46(D1):D1271-D1281. doi: 10.1093/nar/gkx1029. PMID: 29106664; PMCID: PMC5753189.
3. Thul PJ, Lindskog C. The human protein atlas: A spatial map of the human proteome. Protein Sci. 2018 Jan;27(1):233-244. doi: 10.1002/pro.3307. Epub 2017 Oct 10. PMID: 28940711; PMCID: PMC5734309.
4. Fenyö D, Eriksson J, Beavis R. Mass spectrometric protein identification using the global proteome machine. Methods Mol Biol. 2010;673:189-202. doi: 10.1007/978-1-60761-842-3_11. PMID: 20835799; PMCID: PMC3757509.
5. Choi M, Carver J, Chiva C, Tzouros M, Huang T, Tsai TH, Pullman B, Bernhardt OM, Hüttenhain R, Teo GC, Perez-Riverol Y, Muntel J, Müller M, Goetze S, Pavlou M, Verschueren E, Wollscheid B, Nesvizhskii AI, Reiter L, Dunkley T, Sabidó E, Bandeira N, Vitek O. MassIVE.quant: a community resource of quantitative mass spectrometry-based proteomics datasets. Nat Methods. 2020 Oct;17(10):981-984. doi: 10.1038/s41592-020-0955-0. Epub 2020 Sep 14. PMID: 32929271; PMCID: PMC7541731.
6. Sharma V, Eckels J, Taylor GK, Shulman NJ, Stergachis AB, Joyner SA, Yan P, Whiteaker JR, Halusa GN, Schilling B, Gibson BW, Colangelo CM, Paulovich AG, Carr SA, Jaffe JD, MacCoss MJ, MacLean B. Panorama: a targeted proteomics knowledge base. J Proteome Res. 2014 Sep 5;13(9):4205-10. doi: 10.1021/pr5006636. Epub 2014 Aug 18. PMID: 25102069; PMCID: PMC4156235.
7. Perez-Riverol Y, Bai J, Bandla C, García-Seisdedos D, Hewapathirana S, Kamatchinathan S, Kundu DJ, Prakash A, Frericks-Zipper A, Eisenacher M, Walzer M, Wang S, Brazma A, Vizcaíno JA. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Res. 2022 Jan 7;50(D1):D543-D552. doi: 10.1093/nar/gkab1038. PMID: 34723319; PMCID: PMC8728295.
8. Vizcaíno JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Ríos D, Dianes JA, Sun Z, Farrah T, Bandeira N, Binz PA, Xenarios I, Eisenacher M, Mayer G, Gatto L, Campos A, Chalkley RJ, Kraus HJ, Albar JP, Martinez-Bartolomé S, Apweiler R, Omenn GS, Martens L, Jones AR, Hermjakob H. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotechnol. 2014 Mar;32(3):223-6. doi: 10.1038/nbt.2839. PMID: 24727771; PMCID: PMC3986813.
9. Searle BC. Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. Proteomics. 2010 Mar;10(6):1265-9. doi: 10.1002/pmic.200900437. PMID: 20077414.
10. Wilson J, Palmeri J, Pappin D. SimpliFi: a data-to-meaning analytics engine to bring omics understanding to all. J Biomol Tech. 2020 Aug;31(Suppl):S1–2. PMCID: PMC7424641.

11. Didusch S, Madern M, Hartl M, Baccarini M. amica: an interactive and user-friendly web-platform for the analysis of proteomics data. BMC Genomics. 2022 Dec 9;23(1):817. doi: 10.1186/s12864-022-09058-7. PMID: 36494623; PMCID: PMC9733095.
12. Riffle M, Zelter A, Jaschob D, Hoopmann MR, Faivre DA, Moritz RL, Davis TN, MacCoss MJ, Isoherranen N. Limelight: An Open, Web-Based Tool for Visualizing, Sharing, and Analyzing Mass Spectrometry Data from DDA Pipelines. J Proteome Res. 2025 Mar 4. doi: 10.1021/acs.jproteome.4c00968. Epub ahead of print. PMID: 40036265.
13. Farag YM, Horro C, Vaudel M, Barsnes H. PeptideShaker Online: A User-Friendly Web-Based Framework for the Identification of Mass Spectrometry-Based Proteomics Data. J Proteome Res. 2021 Dec 3;20(12):5419-5423. doi: 10.1021/acs.jproteome.1c00678. Epub 2021 Oct 28. PMID: 34709836; PMCID: PMC8650087.
14. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R. The PeptideAtlas project. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D655-8. doi: 10.1093/nar/gkj040. PMID: 16381952; PMCID: PMC1347403.
15. Zauber H, Kirchner M, Selbach M. Picky: a simple online PRM and SRM method designer for targeted proteomics. Nat Methods. 2018 Feb 28;15(3):156-157. doi: 10.1038/nmeth.4607. PMID: 29489744.
16. https://panoramaweb.org/Passport/project-begin.view?pageId=Passport
17. Whiteaker JR, Halusa GN, Hoofnagle AN, Sharma V, MacLean B, Yan P, Wrobel JA, Kennedy J, Mani DR, Zimmerman LJ, Meyer MR, Mesri M, Boja E, Carr SA, Chan DW, Chen X, Chen J, Davies SR, Ellis MJ, Fenyö D, Hiltke T, Ketchum KA, Kinsinger C, Kuhn E, Liebler DC, Liu T, Loss M, MacCoss MJ, Qian WJ, Rivers R, Rodland KD, Ruggles KV, Scott MG, Smith RD, Thomas S, Townsend RR, Whiteley G, Wu C, Zhang H, Zhang Z, Rodriguez H, Paulovich AG. Using the CPTAC Assay Portal to Identify and Implement Highly Characterized Targeted Proteomics Assays. Methods Mol Biol. 2016;1410:223-36. doi: 10.1007/978-1-4939-3524-6_13. PMID: 26867747; PMCID: PMC5017244.
18. MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics. 2010 Apr 1;26(7):966-8. doi: 10.1093/bioinformatics/btq054. Epub 2010 Feb 9. PMID: 20147306; PMCID: PMC2844992.
19. Deutsch EW. Mass spectrometer output file format mzML. Methods Mol Biol. 2010;604:319-31. doi: 10.1007/978-1-60761-444-9_22. PMID: 20013381; PMCID: PMC3073315.
20. R, Mallick P. Data Conversion with ProteoWizard msConvert. Methods Mol Biol. 2017;1550:339-368. doi: 10.1007/978-1-4939-6747-6_23. PMID: 28188540.
21. Searle BC, Pino LK, Egertson JD, Ting YS, Lawrence RT, MacLean BX, Villén J, MacCoss MJ. Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. Nat Commun. 2018 Dec 3;9(1):5128. doi: 10.1038/s41467-018-07454-w. PMID: 30510204; PMCID: PMC6277451.
22. Dornbush S, Aeddula NR. Physiology, Leptin. [Updated 2023 Apr 10]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK537038/
23. Benbaibeche H, Bounihi A, Koceir EA. Leptin level as a biomarker of uncontrolled eating in obesity and overweight. Ir J Med Sci. 2021 Feb;190(1):155-161. doi: 10.1007/s11845-020-02316-1. Epub 2020 Jul 17. PMID: 32681271.

24. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25-9. doi: 10.1038/75556. PMID: 10802651; PMCID: PMC3037419.
25. Clark M, Hoenig M. Feline comorbidities: Pathophysiology and management of the obese diabetic cat. J Feline Med Surg. 2021 Jul;23(7):639-648. doi: 10.1177/1098612X211021540. PMID: 34167340; PMCID: PMC10812123.
26. Kumar M, Dev S, Khalid MU, Siddenthi SM, Noman M, John C, Akubuiro C, Haider A, Rani R, Kashif M, Varrassi G, Khatri M, Kumar S, Mohamad T. The Bidirectional Link Between Diabetes and Kidney Disease: Mechanisms and Management. Cureus. 2023 Sep 20;15(9):e45615. doi: 10.7759/cureus.45615. PMID: 37868469; PMCID: PMC10588295.
27. Pérez-López L, Boronat M, Melián C, Saavedra P, Brito-Casillas Y, Wägner AM. Assessment of the association between diabetes mellitus and chronic kidney disease in adult cats. J Vet Intern Med. 2019 Sep;33(5):1921-1925. doi: 10.1111/jvim.15559. Epub 2019 Jul 15. PMID: 31305000; PMCID: PMC6766521.
28. LK, Searle BC, Yang HY, Hoofnagle AN, Noble WS, MacCoss MJ. Matrix-Matched Calibration Curves for Assessing Analytical Figures of Merit in Quantitative Proteomics. J Proteome Res. 2020 Mar 6;19(3):1147-1153. doi: 10.1021/acs.jproteome.9b00666. Epub 2020 Feb 24. PMID: 32037841; PMCID: PMC7175947.
29. https://www.chromatographyonline.com/view/limit-detection
30. Lu Y, Yu SS, Zong M, Fan SS, Lu TB, Gong RH, Sun LS, Fan LY. Glucose-6-Phosphate Isomerase (G6PI) Mediates Hypoxia-Induced Angiogenesis in Rheumatoid Arthritis. Sci Rep. 2017 Jan 9;7:40274. doi: 10.1038/srep40274. PMID: 28067317; PMCID: PMC5220294.
31. Swearingen KE, Moritz RL. High-field asymmetric waveform ion mobility spectrometry for mass spectrometry-based proteomics. Expert Rev Proteomics. 2012 Oct;9(5):505-17. doi: 10.1586/epr.12.50. PMID: 23194268; PMCID: PMC4777519.
32. Aksenov AA, Kapron J, Davis CE. Predicting compensation voltage for singly-charged ions in high-field asymmetric waveform ion mobility spectrometry (FAIMS). J Am Soc Mass Spectrom. 2012 Oct;23(10):1794-8. doi: 10.1007/s13361-012-0427-6. Epub 2012 Aug 8. PMID: 22872526.
33. Neagu AN, Jayathirtha M, Baxter E, Donnelly M, Petre BA, Darie CC. Applications of Tandem Mass Spectrometry (MS/MS) in Protein Analysis for Biomedical Research. Molecules. 2022 Apr 8;27(8):2411. doi: 10.3390/molecules27082411. PMID: 35458608; PMCID: PMC9031286.
34. https://www.chromatographytoday.com/news/autosamplers/36/breaking-news/understanding-the-difference-between-retention-time-and-relative-retention-time/31166
35. Bauer C, Cramer R, Schuchhardt J. Evaluation of peak-picking algorithms for protein mass spectrometry. Methods Mol Biol. 2011;696:341-52. doi: 10.1007/978-1-60761-987-1_22. PMID: 21063959.
36. Gessulat S, Schmidt T, Zolg DP, Samaras P, Schnatbaum K, Zerweck J, Knaute T, Rechenberger J, Delanghe B, Huhmer A, Reimer U, Ehrlich HC, Aiche S, Kuster B, Wilhelm M. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nat Methods. 2019 Jun;16(6):509-518. doi: 10.1038/s41592-019-0426-7. Epub 2019 May 27. PMID: 31133760.
37. The M, MacCoss MJ, Noble WS, Käll L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. J Am Soc Mass Spectrom. 2016

Nov;27(11):1719-1727. doi: 10.1007/s13361-016-1460-7. Epub 2016 Aug 29. PMID: 27572102; PMCID: PMC5059416.

38. Elias JE, Gygi SP. Target-decoy search strategy for mass spectrometry-based proteomics. Methods Mol Biol. 2010;604:55-71. doi: 10.1007/978-1-60761-444-9_5. PMID: 20013364; PMCID: PMC2922680.