# PEAKS DIA Spectral Library Search and Direct Database Search Workflow for Sensitive and Reproducible Identification of Proteins from DIA Mass Spectra

Dan Maloney, Application Scientist
Bioinformatics Solutions Inc., Waterloo, Canada

## Abstract:

Data independent acquisition (DIA) mass spectrometry (MS) has been developed to improve the reproducibility of protein identification within complex datasets. PEAKS offers a workflow that includes both a spectral library search and a database search algorithm designed for the complexities of DIA mass spectra. Through the DIA analysis of Human embryonic kidney (HEK) cells, spectral library search was shown to provide the most reproducible identification results compared to the typical data dependent acquisition (DDA) method. Direct database search using DIA data was shown to be the most sensitive approach compared to the other methods tested. While the combination of spectral library search and database search in a workflow provides a method that is both accurate and reproducible.

## Introduction:

DIA MS collects fragment ions from predefined mass ranges covering a large fraction of the total mass range in the associated unfractionated MS data. DDA MS targets individual ionized peptides for fragmentation and collects a narrow mass window to produce a spectrum intended to only contain fragment ions from the target ionized peptide. Thus, DIA MS produces unbiassed, reproducible yet complex spectra often containing fragment ions from several co-eluting peptides (Fig 1).

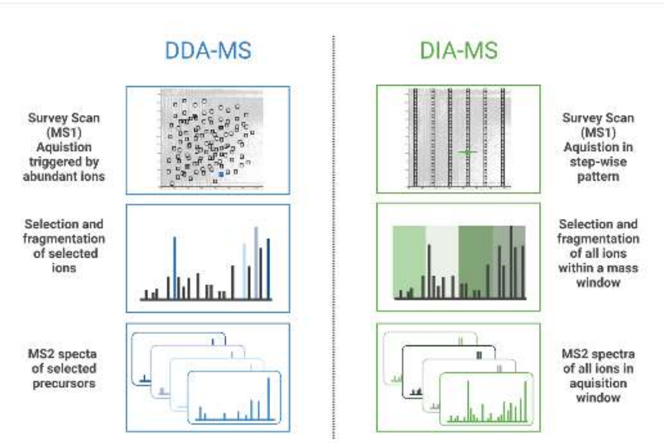
DDA-MS
DIA-MS
Survey Scan (MS1) Aquistion triggered by abundant ions
Selection and fragmentation of selected ions
MS2 specta of selected precursors
Survey Scan (MS1) Aquistion in step-wise pattern
Selection and fragmentation of all ions within a mass window
MS2 spectra of all ions in aquisition window

Fig 1. DDA-MS vs. DIA-MS comparison

To analyze the complex spectra produced in DIA datasets, PEAKS provides two analysis methods: spectral library search, and database search. PEAKS' spectral library search identifies spectra in the DIA experiment that match the characteristics of previously identified peptides from DDA spectra. That includes the fragment ion pattern, indexed retention time (iRT), and ion mobility (IM) details. PEAKS database search for DIA searches the dataset with an in silico generated spectral library directly from the protein sequence database. In silico peptide details including the fragment ion pattern, indexed retention time, and ion mobility are predicted using deep learning. These methods can be run separately or in a workflow. When run as a workflow, previously identified peptides are first matched using the spectral library search. Then, spectra that aren't confidently matched within a chosen false discovery rate (FDR) are re-analyzed using a database search.

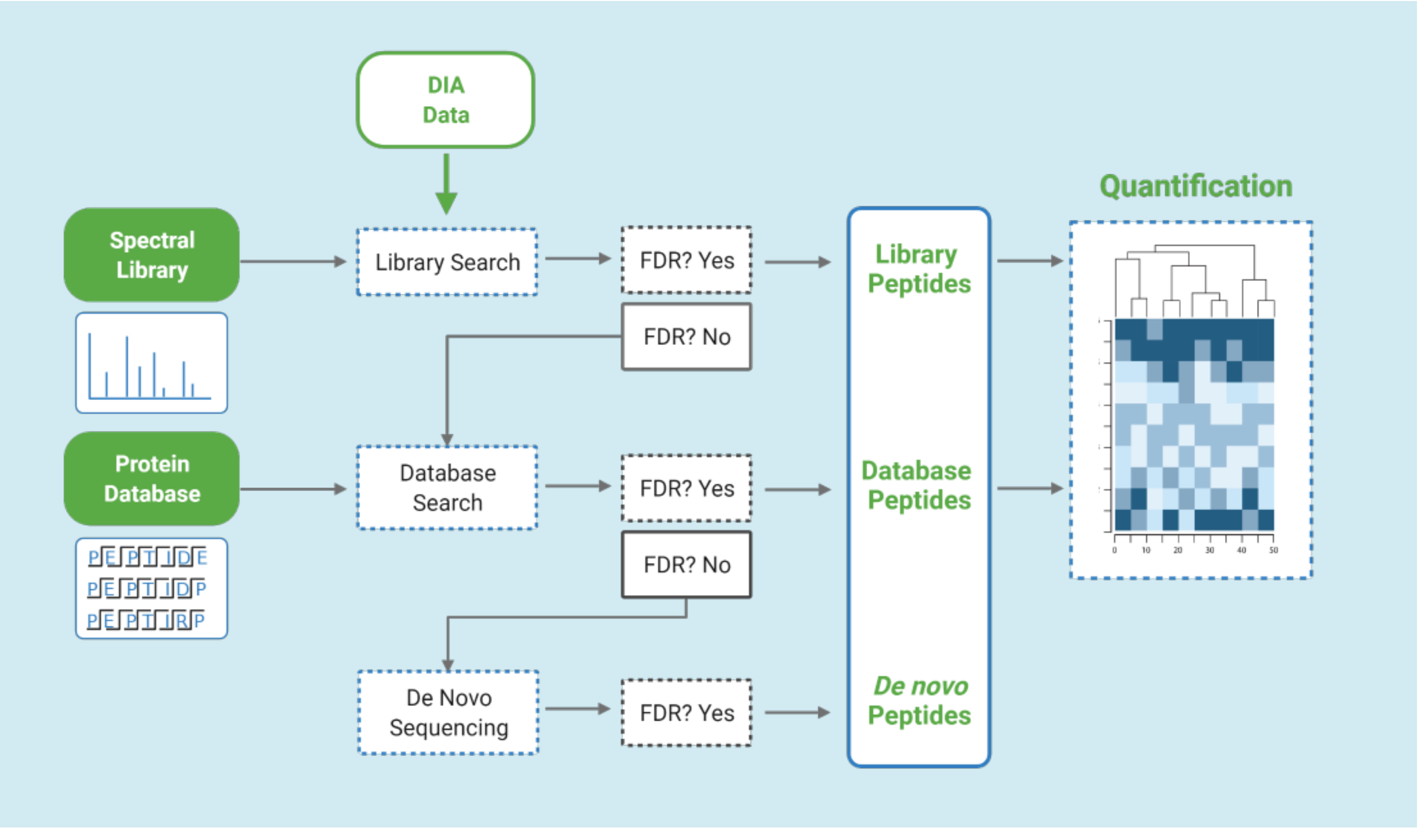

Fig 2. PEAKS Online Xpro DIA Workflow

## Methods:

The PXD008235 dataset was re-analyzed using new PEAKS Online Xpro (January 2022 build). Three microgram samples of HEK293 protein digest were analyzed on a Q-Exactive mass spectrometer (Thermo Fisher) using multiple methods[1]. Eleven replicates were analyzed using a DDA method; 8 were used for spectral library creation (table 1), and 3 were used to test the reproducibility and sensitivity of the DDA method and searched using PEAKS DB (table 2). Three replicates were analyzed using a DIA method and searched using PEAKS spectral library search and database search for DIA (table 3, table 4).

| Parameter | Parameter |
|---|---|
| Precursor mass tolerance | 10 ppm |
| Fragment mass tolerance | 0.02 Da |
| Fixed PTMs | Carbamidomethylation |
| Digestion enzyme | Trypsin |
| Enzyme specificity | Specific |
| Maximum missed cleavages | 3 |
| Peptide spectrum match FDR | 1% |
| Database | Reviewed Uniprot Homo sapiens |

Table 1: Spectral library generation parameters and DDA analysis parameters

| Parameter | Parameter |
|---|---|
| Peptide FDR | 0.1% |
| Protein group FDR | 0.1% |
| Minimum unique peptides | 1 |
| All others equal to table 1 | See table 1 |

Table 2: DDA analysis paramers

| Parameter | Parameter |
|---|---|
| Precursor mass tolerance | 10 ppm |
| Fragment mass tolerance | 0.02 Da |
| Peptide length range | 7-30 |
| Library | 8 DDA replicates from Table 1 |
| Peptide FDR | 0.1% |
| Protein Group FDR | 0.1% |
| Minnimum unique peptides | 1 |
| Database | Reviewed Uniprot Homo sapiens |

Table 3: DIA Spectral Library search parameters

| Parameter | Parameter |
|---|---|
| Precursor mass tolerance | 10 ppm |
| Fragment mass tolerance | 0.02 Da |
| Fixed PTMs | Carbamidomethylation |
| Digestion enzyme | Trypsin |
| Enzyme specificity | Specific |
| Maximum missed cleavages | 3 |
| Peptide Length range | 7-30 |
| Peptide FDR | 0.1% |
| Protein Group FDR | 0.1% |
| Minimum unique peptides | 1 |

Table 4: DIA database search parameters

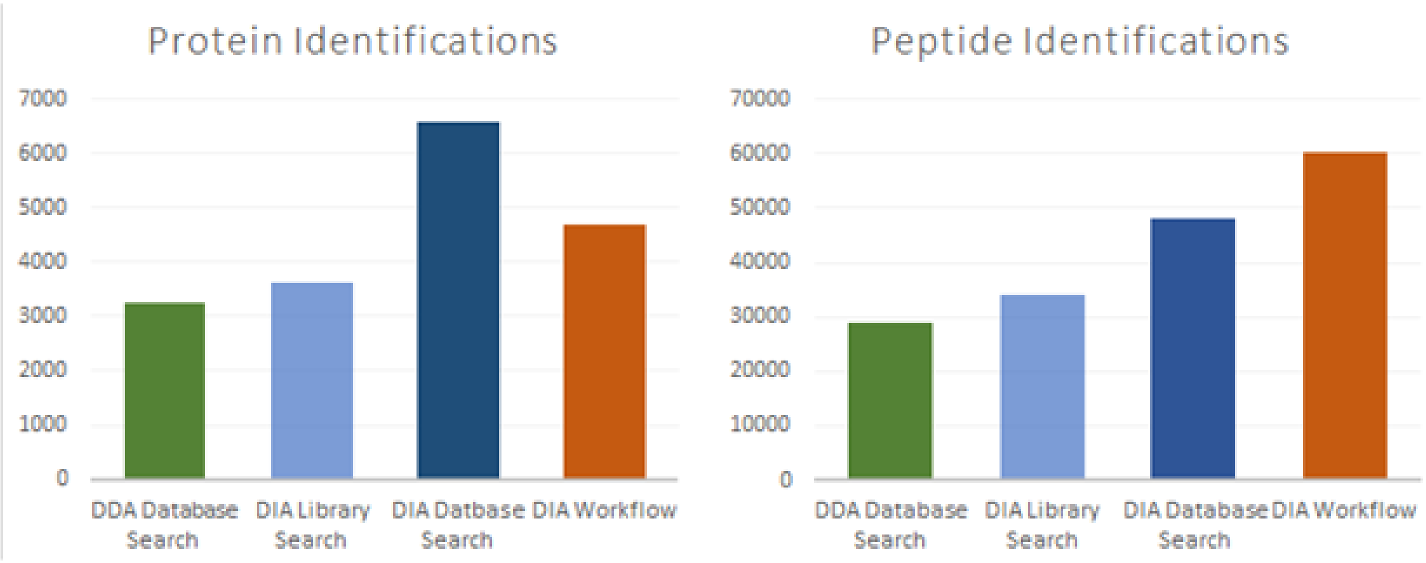

Fig 3. Protein and peptide total number of identifications over three replicates of HEK lysate samples. Protein identifications were within a 0.1% protein group FDR. Peptide identifications were within a 0.1% peptide FDR

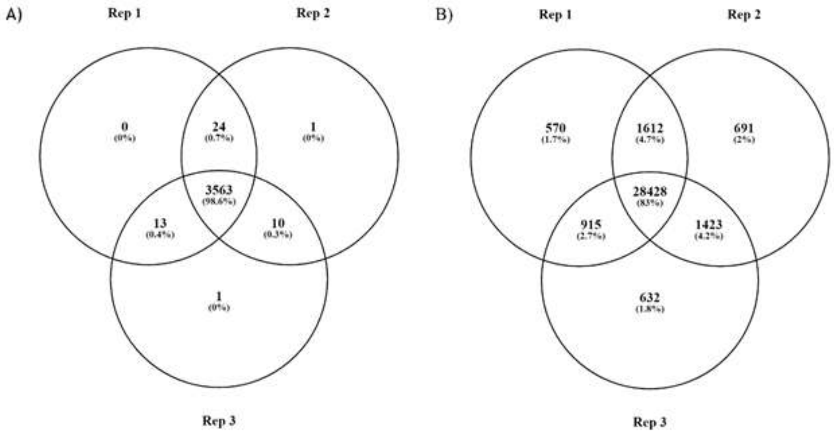

Fig 4. DIA Library Search reproducibility. A peptide/protein was identified in the replicate if it was identified in one spectrum within a 0.1% peptide FDR threshold.

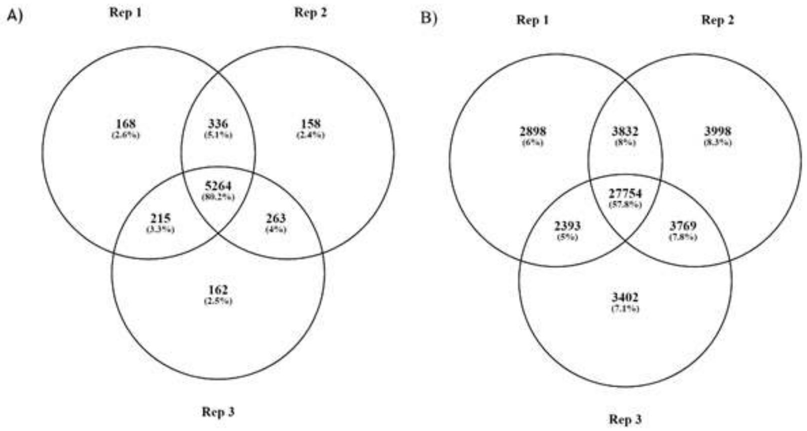

Fig 5. : A) DIA database search protein reproducibility. B) Peptide reproducibility. A peptide/protein was identified in the replicate if it was identified in one spectrum within a 0.1% peptide FDR threshold.

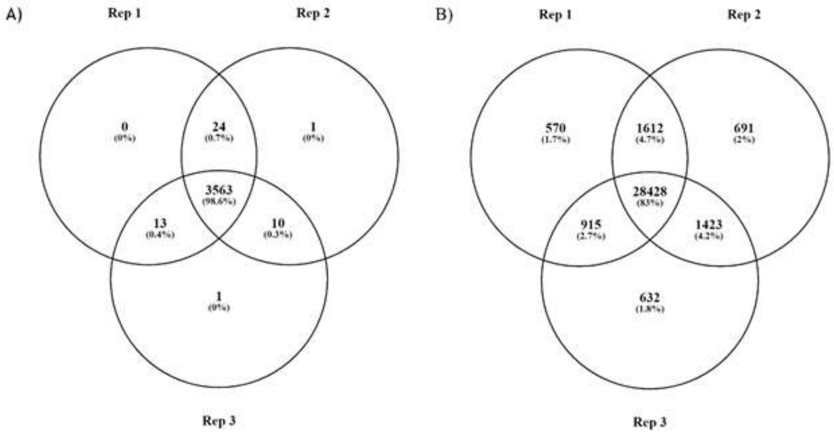

Fig 6. A) DIA Workflow (Spectral library search then DB search) protein reproducibility. B) Peptide reproducibility. A peptide/protein was identified in the replicate if it was identified in one spectrum within a 0.1% peptide FDR threshold.

## Results:

Database search using a DIA instrument method provided the most identified proteins at a 0.1% protein group FDR. The DIA workflow provided the most identified peptides at a 0.1% peptide FDR (Fig 3).

Across the three replicates, identification and reproducibility was assessed for each of the approaches. An identification within a replicate was accepted if the protein or peptide was identified by at least one spectrum in the replicate within a 0.1% peptide FDR. Library search proved to be the most reproducible with 98.6% reproducibility across three replicates at the protein level and 83% reproducibility at the peptide level (Fig 4).

The trade-off between sensitivity and reproducibility is clear in the DIA database search results as it was the most sensitive, but lest reproducible (Fig 4). The DIA workflow strikes a compromise between reproducibility and sensitivity with 96.5% reproducibility of 4536 protein identifications (Fig 5). In comparison to the DDA database search method, the DIA search methods appear to be more reproducible and sensitive than DDA protein identification (Fig 6).