

# DeepDIA Demo: Spectral Library Generation with Detectability Prediction

Using deep learning to generate in silico spectral libraries from proteome databases with detectability filtering for data-independent acquisition (DIA) analysis.

## 1. System Requirements

In this demo, spectral library generation has been tested on a workstation with Intel Xeon E5-2690 v3 CPU, 16 GB RAM, and Microsoft Windows Server 2016 Version 1607 (OS Build 14393.2430) operating system with the following softwares:

- Anaconda 4.2.0 (Python 3.5.2).
- Keras 2.2.4 and TensorFlow 1.11.
- Microsoft R Open 3.5.1.
- RStudio 1.1.447.
- R packages `rjson` and `readr`.
- Protein Digestion Simulator (<https://omics.pnl.gov/software/protein-digestion-simulator>).

DIA data analysis is performed on a workstation with Intel Core i9-7960X CPU, 128 GB RAM, and Microsoft Windows 10 Version 1809 (OS Build 17763.503) 64-bit operating system with the following softwares:

- Spectronaut 13.3.190726.

## 2. Demo Data

LC-MS/MS data of 3 DIA technical replicates of 2 h gradient of HeLa cells on Q Exactive HF are available at ProteomeXchange (<http://proteomecentral.proteomexchange.org/>) with the data set `PXD005573`. (Bruderer, R. et al. Mol. Cell. Proteomics 2017, 16, 2296-2309.)

- Fig1\_MP-DIA-120min120kMS1-22W30k-8dppp\_MHRM\_Ro1.raw
- Fig1\_MP-DIA-120min120kMS1-22W30k-8dppp\_MHRM\_Ro2.raw
- Fig1\_MP-DIA-120min120kMS1-22W30k-8dppp\_MHRM\_Ro3.raw

SwissProt *Homo sapiens* database (2018-04 version, 20,301 entries), pre-trained models, and in silico spectral libraries are available at ProteomeXchange/iProX with identifier `PXD014108/IPX0001628000`.

- `swissprot_human_201804_validated.fasta`

Pre-trained models for Q Exactive HF trained with HeLa and HEK-293 data from [PXD005573](#) are available at ProteomeXchange/iProX with identifier [PXD014108/IPX0001628000](#) .

- `HeLa.model.zip`
- `HeLa_HEK293.model.zip`

They are also provided in `data/models` folder.

Generated spectral libraries and the saved projects from Spectronaut are also available at ProteomeXchange/iProX with identifier [PXD014108/IPX0001628000](#) .

- `SwissProtHuman_detectability50_prediction.library.zip`
- `SwissProtHuman_detectability50_prediction.kit`
- `HeLa_SwissProtHuman_detectability50_prediction.sne.zip`
- `HeLa_SwissProtHuman_detectability50_prediction.csv.zip`

## 3. Spectral Library Generation

### 3.1. Protein Digestion

A peptide list can be generated by in silico digestion from protein sequences (FASTA). Protein digestion can be performed using Protein Digestion Simulator.

Tryptic (no Proline Rule) is selected as digestion enzyme. Digestion is performed with the following parameters:

- Max Miss Cleavages: 2
- Minimum Residue Count: 7
- Maximun Fragment Mass: 6000
- Minimun Fragment Mass: 0

DeepDIA only supports peptide sequences with standard amino acids (ACDEFGHIKLMNPQRSTVWY) and length  $\leq 50$ .

Start RStudio, ensure package `readr` has been installed.

```
install.packages("readr")
```

Open `deepdetect/R/init.R` and run the script by clicking [Source](#).

```
source("{PATH_TO_CODE}/deepdetect/R/init.R")
```

Set the protein sequence directory as working directory and run

`deepdetect/R/get_peptides_from_ProteinDigestionSimulator_result.R`.

```
setwd("{PATH_TO_DATA}")  
source("{PATH_TO_CODE}/deepdetect/R/get_peptides_from_ProteinDigestionSimulator_result.R")
```

The peptide list file is generated in the working directory.

- `swissprot_human_201804_validated.peptide.csv`

## 3.2. Predict MS Detectability

In this demo, a pre-trained model `data/model/detectability` is used for MS detectability prediction.

Copy the peptide list file `swissprot_human_201804_validated.peptide.csv` to the model directory.

Run `deepdetect/py/predict.py` in the directory.

```
cd {PATH_TO_MODEL}  
python {PATH_TO_CODE}/deepdetect/py/predict.py
```

The detectability file is generated in the working directory.

- `swissprot_human_201804_validated.prediction.detectability.csv`

In this demo, peptides with detectability score  $< 0.5$  are ruled out.

Open `deepdetect/R/filter_peptides_by_detectability.R` and run the script by clicking [Source](#).

```
source("{PATH_TO_CODE}/deepdetect/R/filter_peptides_by_detectability.R")
```

The filtered peptide list file is generated in the working directory.

- `swissprot_human_201804_validated_detectability50.peptide.csv`

## 3.3 Generate Spectral Library

Follow the instruction described in *DeepDIA Demo: Spectral Library Generation From Peptide Lists* to predict MS/MS spectra and iRT, and generate a spectral library.

Expected run time depends on the number of peptides and the performance of the computer. In this demo, this command may take up to 1 day.

An entrapment strategy has been used to evaluate identification error rates by adding proteins from other organisms to the libraries. The entrapment protein sequence databases and in silico spectral libraries are available at ProteomeXchange/iProX with identifier [PXD014108/IPX0001628000](#) .

- [swissprot\\_Caenorhabditis\\_elegans\\_201903.fasta](#)
- [swissprot\\_Dictyostelium\\_discoideum\\_201903.fasta](#)
- [swissprot\\_Escherichia\\_coli\\_strain\\_K12\\_201903.fasta](#)
- [swissprot\\_Saccharomyces\\_cerevisiae\\_strain\\_ATCC\\_204508\\_S288c\\_201903.fasta](#)
- [SwissProt\\_entrapment\\_detectability50\\_prediction.library.zip](#)
- [SwissProt\\_entrapment\\_detectability50\\_prediction.kit.zip](#)

## 4. DIA Analysis

DIA analysis can be performed following the instruction described in *DeepDIA Demo: Spectral Library Generation From Proteome Databases*.

For benchmarking purpose, the saved projects using a sample-specific spectral library generated by data-dependent acquisition (DDA) experiments, and directDIA are also available at ProteomeXchange/iProX with identifier [PXD014108/IPX0001628000](#) .

- [HeLa\\_DDA.kit](#)
- [HeLa\\_DDALib.sne.zip](#)
- [HeLa\\_DDALib.csv.zip](#)
- [HeLa\\_SwissProtHuman\\_directDIA.sne.zip](#)
- [HeLa\\_SwissProtHuman\\_directDIA.csv.zip](#)