

# PRIDE Data Downloader

This documentation provides basic information on the usage and design of the PRIDE Data Downloader and Converter. The PRIDE Data Downloader is a tool implemented in Python 3.7 that offers access to PRIDE projects and converts the contained PSMs into a portable .csv format.

## I. INTRODUCTION

The tool is used as a command line utility with the goal of downloading, extracting, parsing and converting PSMs from PRIDE projects. The user is able to perform feature selection on the PRIDE database. Configuration of the tools parameters is accomplished by utilizing the command line arguments or the accompanied .ini configuration file.

## II. CONCEPT

The tool first requests the specified number of projects with the desired settings from the PRIDE archive utilizing the PRIDE RESTful API. The json response is decoded and the enclosed download links for the individual assays are extracted. The files (mzid and mgf tuples) are downloaded and saved with the json description of the project. Afterwards the parsing and processing of the downloaded file tuples starts. The MZID and MGF Parser extract the relevant features which are matched to form a CSV file containing the calculated data (y and b series matches, length calculations, modifications per length ...) for each PSM.

## III. IMPLEMENTATION

The tool is implemented in Python 3.7 and utilizes a variety of additional modules to perform the tasks described in the concept. The MZID parser is implemented utilizing the xml.sax module of Python. MGF are parsed by evaluating the content of each line. The tool is able to utilize multiple processes for generating the CSV files. Multiprocessing is implemented with a data parallel approach and the multiprocessing module of Python.

## IV. REQUIREMENTS

Besides the official modules provided by Python the downloader also utilizes external modules available through PyPi.

External Package	Usage	URL
Pyteomics	Provides the masses of molecules and amino acids for the CSV calculations.	<a href="https://pyteomics.readthedocs.io/en/latest/">https://pyteomics.readthedocs.io/en/latest/</a>
numba	Offers JIT-compiled functions in llvm for a significant performance increase.	<a href="https://numba.pydata.org/">https://numba.pydata.org/</a>

## V. CONFIGURATION

The tool can be configured by utilizing the command line parameters or the config.ini file. The tool also utilizes a modifications.csv in the top level folder of the downloader. This file contains all modifications that are either valid or invalid for processing with this tool. During the execution of the tool a debug.log is created by the Python logging module.

```
python pride_data.py <Arguments>
```

Option	Description
accessions	Specify certain projects by accessions to download.
csv	Generates a csv file for each available file tuple!
memory	Limits the RAM for the program to the given ratio of the available RAM!
number	Maximal number of projects per page with fitting metadata to include.
pages	Maximal number of project pages to search.
single_file	Only download a single file tuple for each available project!
instruments	MS/MS instruments used in projects. String used by PRIDE
json	Generates a json file from each available project!
species	Species evaluated in projects. NCBI Taxonomy ID
folder	Folder containing downloaded data relative to the python script!
submission	SubmissionType for projects.
ini	Executes the setting in the config.ini and disregards command line arguments!