**DIA-BERT manual**

**Installation**

On Windows systems, download and unzip the zip file. Click on DIA-BERT.exe to run without installation.

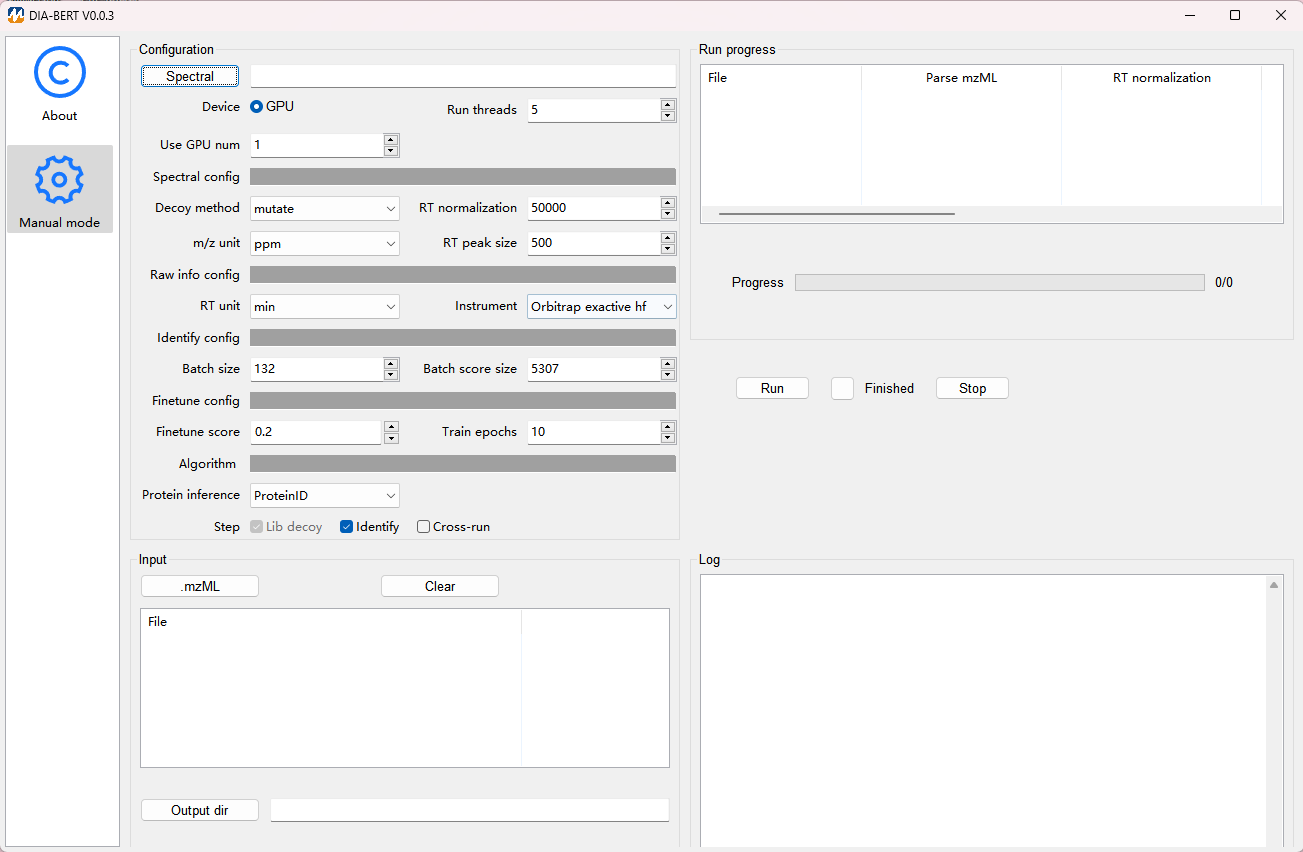
On **Linux**, download the file from the release. DIA-BERT runs install-free and requires no additional configuration of the environment.

**Hardware Requirements:**

* **Operating System:** Supports both Windows and Linux operating systems.
* **Processor:** A dual-core processor is recommended, but it can run on a single-core processor.
* **Memory:** 40GB or more is recommended. If the mass spectrometry files or library files to be identified are large, it is advised to use more memory.
* **Storage:** At least 100GB of available hard disk space is recommended.
* **Graphics Card:** A 40GB NVIDIA GPU with CUDA support or a V100 32GB GPU is recommended.

**Getting Started**

**GUI settings reference**



1. Click **Spectral** (in the **Configuration** pane), select your library. Currently, for library-base analysis, select comma-separated (.csv, .txt), tab-separated (.tsv, .xls, .xlsx) as spectral libraries. If the decoy field already exists in the spectrum library file, please manually set the decoy method to no; Otherwise, by default, Decoy method is set to mutate, and the corresponding decoy will be automatically generated.
2. Click **.mzML** (in the **Configuration** pane)**,** select mass spectrometry data files. The currently supported formats are .mzML.
3. Click **Instrument** (in the **Configuration** pane)**,** select the instrument of mass spectrometry data files. Additionally, Click **Protein inference** (in the **Configuration** pane)**,** adjust the fields (ProteinID or ProteinName) used by the protein inference.
4. Click **Output dir** (in the **Configuration** pane)**,** Specify **Main output** file path and click **Run**. The software will display the qualification progress in real time.

**Command-line reference**

./DIA-BERT

--rawdata\_file\_dir\_path=./data.txt

--lib=speclib.tsv

--out\_path=/user/identify/

--instrument="Orbitrap exactive hf"

--decoy\_method=mutate

--step\_size=20000;

|  |  |  |  |
| --- | --- | --- | --- |
| Parameter type | Parameter name | Parameter annotations | Default value |
| Required | --rawdata\_file\_dir\_path | The absolute path of the mzML file to be analyzed |  |
| Required | --lib | The absolute path to the library file |  |
| Required | --out\_path | The output path of the identification results; If the path does not exist, the program will automatically create it |  |
| Required | --decoy\_method | The decoy generation strategy of Identification：   * mutate | mutate |
| Optional | --instrument | The instrument type of the mass spectrometry file：   * Orbitrap exactive hf * Orbitrap exactive hf-x * Orbitrap exploris 480 * Orbitrap fusion lumos * Other | Other |
| Optional | --gpu\_devices | List of GPU indices to be used, separated by commas | auto |
| Optional | --n\_cycles | The search scope for the precursor identification | 100 |
| Optional | --step\_size | The number of batches to be scored at each time. The higher the value, the higher the overall efficiency. | Dynamically calculate based on GPU memory |
| Optional | --batch\_size | The number of precursors processed per batch | Dynamically calculate based on GPU memory |
| Optional | --raw\_rt\_unit | Units of retention time in mass spectrometry file:   * min * sec | min |
| Optional | --device | The type of device used:   * cuda | cuda |
| Optional | --n\_thread | Num of thread | 5 |
| Optional | --open\_cross\_quantification | Whether or not to correct quantitative values for multiple mass spectral files   * 1-open * 0-off | 0 |
| Optional | --protein\_infer\_key | Protein inference key   * ProteinID * ProteinName | ProteinID |
| Optional | --open\_identify | Do you want to initiate the appraisal process   * 1-open * 0-off | 1 |
| Optional | --train\_pkl\_size | The number of batches to be train at each time. The higher the value, the higher the overall efficiency. |  |
| Optional | --quant\_pkl\_size | The number of batches to be quant at each time. The higher the value, the higher the overall efficiency. |  |

**Input and output formats**

**Raw data formats**

Formats supported by DIA-BERT: .mzML. Other mass spec formats, such as Sciex .wiff, Bruker .d and Thermo .raw, can be converted to .mzML using the MSConvertGUI application from [ProteoWizard](http://proteowizard.sourceforge.net/download.html). We will support these formats in the future.

**Spectral library formats**

DIA-BERT supports comma-separated (.csv, .txt), tab-separated (.tsv, .xls, .xlsx) as spectral libraries. Important: the library must not contain non-fragmented precursor ions as 'fragments': each fragment ion must actually be produced by the peptide backbone fragmentation.

Required columns:

|  |  |  |  |
| --- | --- | --- | --- |
| Parameter type | Parameter name | Parameter annotations | Alias |
| Required | PeptideSequence | Sequence of precursor | * Sequence * StrippedPeptide |
| Required | FullUniModPeptideName | Modified sequence of precursor | * ModifiedPeptide * LabeledSequence * modification\_sequence |
| Required | PrecursorCharge | Charge of precursor | * Charge * prec\_z |
| Required | PrecursorMz | Mz of precursor | Q1 |
| Required | Tr\_recalibrated | arbitrary RT scale can be used | * iRT * RetentionTime * NormalizedRetentionTime * RT\_detected |
| Required | FragmentMz | Mz of fragment | * ProductMz * Q3 |
| Required | FragmentType | either y or b; for x and z fragments also specify fragment type as y, and for a and c - as b | * FragmentIonType * ProductType * ProductIonType * frg\_type |
| Required | LibraryIntensity | Intensity of fragment | * RelativeIntensity * RelativeFragmentIntensity * RelativeFragmentIonIntensity * relative\_intensity |
| Required | FragmentCharge | Charge of fragment | * FragmentIonCharge * ProductCharge * ProductIonCharge * frg\_z |
| Required | ProteinID | Identifiers for the protein isoforms | * UniProtIds * ProteinID * UniprotID * uniprot\_id |
| Required | FragmentNumber | Number of fragment | * frg\_nr * FragmentSeriesNumber |
| Optional | FragmentLossType | LossType of fragment:  NH3   * H2O * CO * noloss | * FragmentIonLossType * ProductLossType * ProductIonLossType |
| Optional | ProteinName | Name of protein | * Protein Name * Protein\_name * protein\_name |
| Optional | Gene | Name of Gene | * Genes * GeneName |
| Optional | decoy |  | Decoy |
| Optional | ExcludeFromAssay | Determine whether deletion is required when Quantification | ExcludeFromQuantification |

**Output formats**

Tabular data containing precursor and protein identifications are currently output, respectively. The output column names will be normalized to the following format:

**Main output file format**

{filename}\_precursor.csv

|  |  |
| --- | --- |
| Parameter name | Parameter annotations |
| FileName | Name of file |
| PrecursorID | ID of precursor |
| PeptideSequence | Sequence of precursor |
| PrecursorCharge | Charge of precursor |
| iRT | arbitrary RT scale can be used |
| RT | Retention time of precursor |
| PrecursorQuant | Quant of precursor |
| ProteinID | identifiers for the protein isoforms |
| ProteinName | Name of protein |

{filename}\_protein.csv

|  |  |
| --- | --- |
| Parameter name | Parameter annotations |
| FileName | Name of file |
| ProteinID | identifiers for the protein isoforms |
| ProteinName | Name of protein |
| ProteinQuant | Quant of protein |

**Cross-run output file format**

crossrun\_precursor.csv

|  |  |
| --- | --- |
| Parameter name | Parameter annotations |
| PrecursorID | ID of precursor |
| PeptideSequence | Sequence of precursor |
| PrecursorCharge | Charge of precursor |
| ProteinID | identifiers for the protein isoforms |
| ProteinName | Name of protein |
| FileName1 | Quant of precursor for file1 |
| FileName2 | Quant of precursor for file2 |
| FileName3 | Quant of precursor for file3 |

crossrun\_protein.csv

|  |  |
| --- | --- |
| Parameter name | Parameter annotations |
| ProteinID | identifiers for the protein isoforms |
| ProteinName | Name of protein |
| FileName1 | Quant of protein for file1 |
| FileName2 | Quant of protein for file2 |
| FileName3 | Quant of protein for file3 |

**Demo data**

Library with 50,000 precursors and one human tumor DIA mass spectrometry file

DIA-BERT identifications of demo data：

* precursor：23154
* protein：3853
* time required：16 min