DiNovo User Guide

(Last updated: Oct 29th, 2024)

CONTENT

1 Overview

DiNovo is a software tool for automated, high-coverage and confidence de novo peptide sequencing from tandem mass spectrometry data, supporting mirror-protease approach that include trypsin and lysargiNase, lysC and lysN, etc.

The kernel of DiNovo is written in Python3 and the graphical user interface is implemented in Java language in Windows Systems.

If you have any question, please send e-mail to: yfu@amss.ac.cn, piyuzhou@amss.ac.cn

2 Installation

Hardware

A computer with a 64-bit version Windows 10 (and above) is required to run DiNovo, and 16GB of RAM or more. For using MirrorNovo, a GPU with 4GB memory at least is required.

Environment

For using MirrorNovo, anaconda environment is required. You need to download and install anaconda, and then edit the environment variables as shown in Fig 1. Moreover, downloading and installing CUDA, cuDNN (suitable for your GPU), and editing environment variables are needed, too (Fig 2).

Tips: CUDA version is compatible with GPU, while cuDNN and pytorch (will be mentioned below) versions are compatible with CUDA.

Download

Download the latest version of DiNovo and unzip it.

Preparation

For using MirrorNovo in first time, open the command line window in the path of unzipped DiNovo folder, and run this three commands:

```
conda env create -f requirements.yml
conda activate mirrornovo_env
conda install pytorch==[w] torchvision==[x] torchaudio==[y] cudatoolkit=[z] –c pytorch –c nvidia
```

The last command means that install the packages of pytorch suited for your computer's GPU and CUDA. [w] [x] [y] [z] are the version number of the packages(see PyTorch to choose the command for yourself).

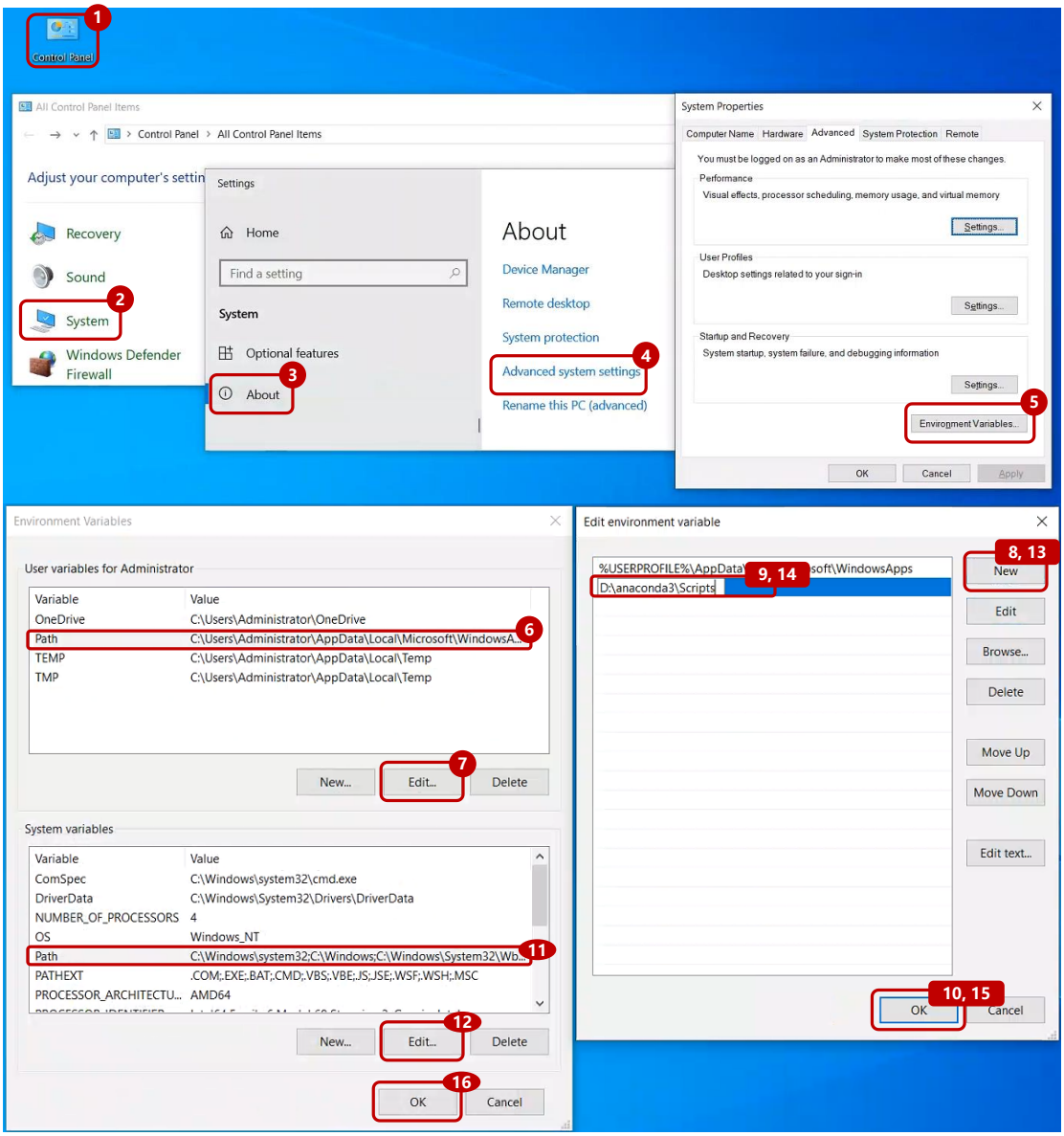
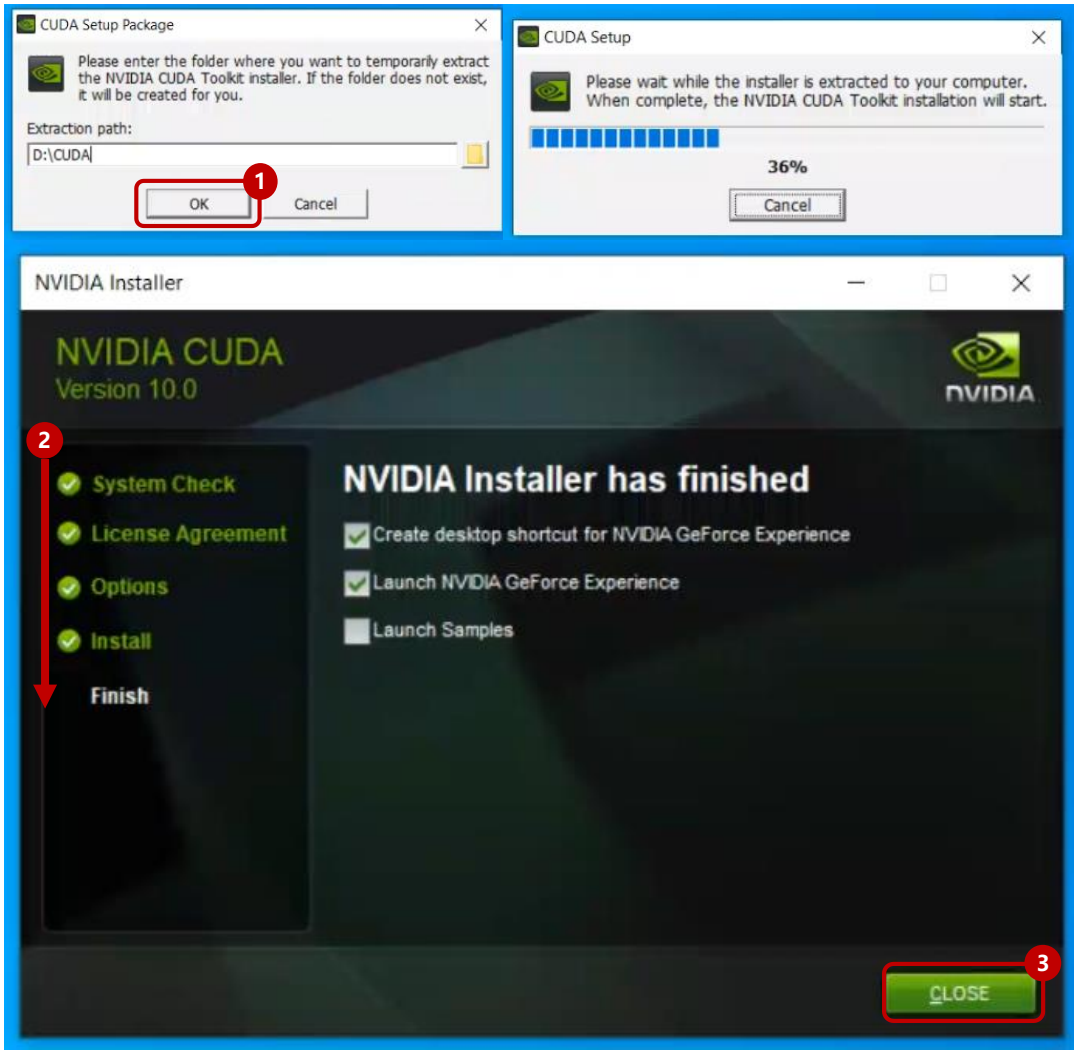
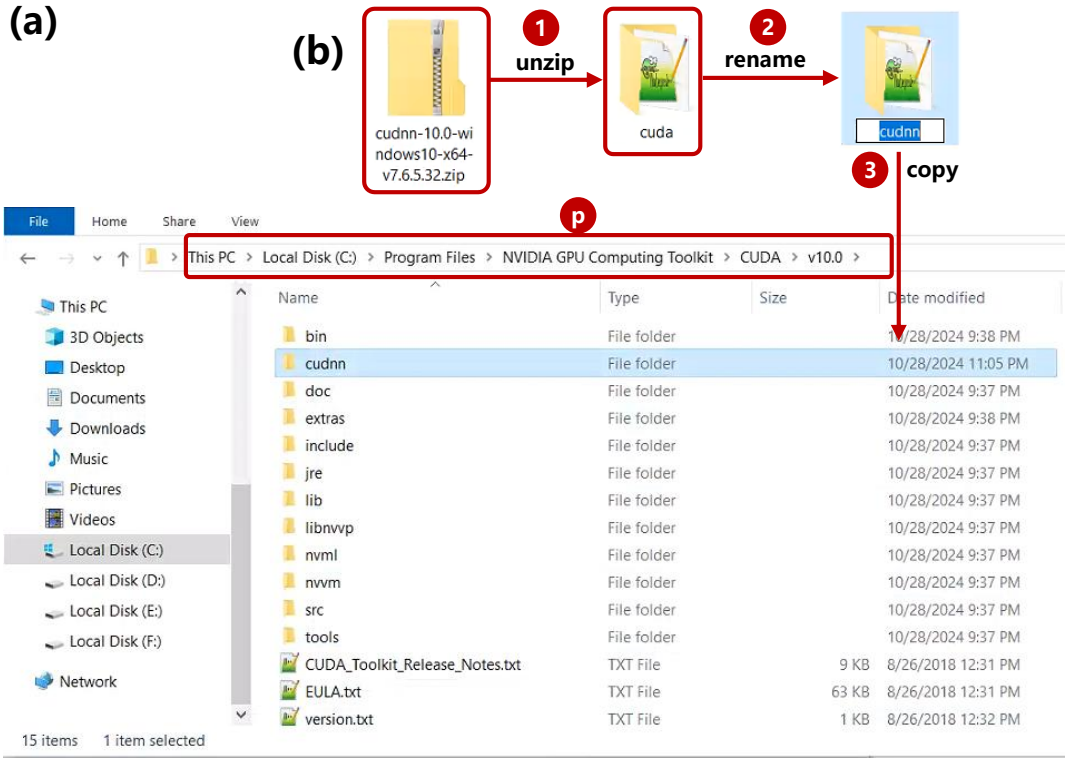


Figure 1. Editing environment variables for anaconda

In step 9 and 14, what we need to type in is the scripts folder address under the anaconda installation directory.


CUDA Setup Package
Please enter the folder where you want to temporarily extract the NVIDIA CUDA Toolkit installer. If the folder does not exist, it will be created for you.
Extraction path:
D:\CUDA
OK
Cancel
1
CUDA Setup
Please wait while the installer is extracted to your computer. When complete, the NVIDIA CUDA Toolkit installation will start.
36%
Cancel
NVIDIA Installer
NVIDIA CUDA
Version 10.0
NVIDIA
2
System Check
License Agreement
Options
Install
Finish
NVIDIA Installer has finished
Create desktop shortcut for NVIDIA GeForce Experience
Launch NVIDIA GeForce Experience
Launch Samples
3
CLOSE


(a)
(b)
cudnn-10.0-windows10-x64-v7.6.5.32.zip
1
unzip
cuda
2
rename
cudnn
3
copy
p
File
Home
Share
View
This PC > Local Disk (C:) > Program Files > NVIDIA GPU Computing Toolkit > CUDA > v10.0 >
This PC
3D Objects
Desktop
Documents
Downloads
Music
Pictures
Videos
Local Disk (C:)
Local Disk (D:)
Local Disk (E:)
Local Disk (F:)
Network
Name
Type
Size
Date modified
bin File folder 10/28/2024 9:38 PM
cudnn File folder 10/28/2024 11:05 PM
doc File folder 10/28/2024 9:37 PM
extras File folder 10/28/2024 9:38 PM
include File folder 10/28/2024 9:37 PM
jre File folder 10/28/2024 9:37 PM
lib File folder 10/28/2024 9:37 PM
libnvvp File folder 10/28/2024 9:37 PM
nvml File folder 10/28/2024 9:37 PM
nvvm File folder 10/28/2024 9:37 PM
src File folder 10/28/2024 9:37 PM
tools File folder 10/28/2024 9:37 PM
CUDA_Toolkit_Release_Notes.txt TXT File 9 KB 8/26/2018 12:31 PM
EULA.txt TXT File 63 KB 8/26/2018 12:31 PM
version.txt TXT File 1 KB 8/26/2018 12:32 PM
15 items
1 item selected

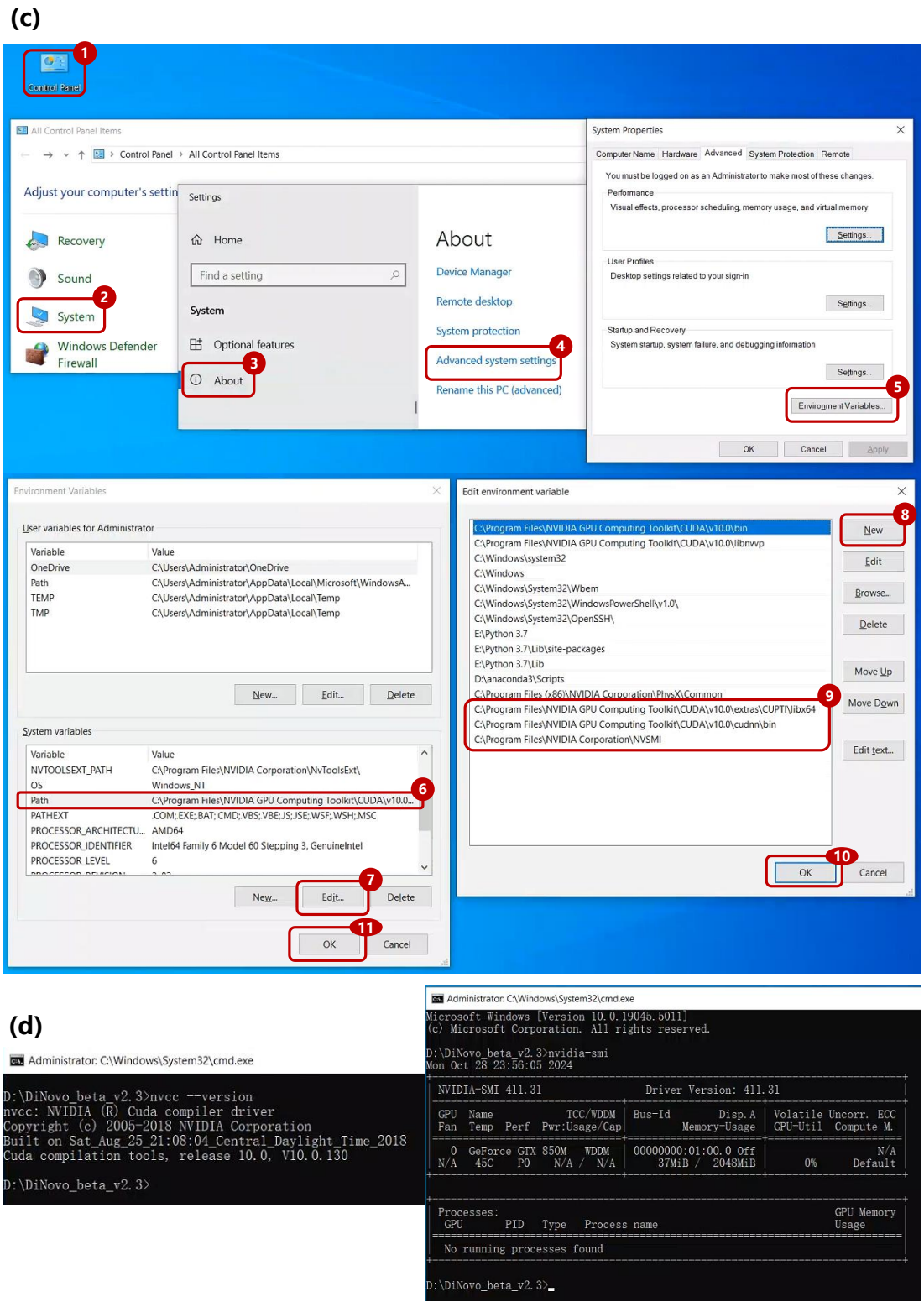


Figure 2. Example of installation process for CUDA and cuDNN. (a). Installation process of CUDA 10.0. (b). Installation process of cuDNN 7.6.5. In step 3, we need to copy the folder *cudnn* to the path *p*. (c). Editing the environment variables, and we need to type in 3 variables in step 8-9. (d). Checking if the installation was successful.

3 Configuration and Quick Start

Double click *DiNovo.exe* and set the configuration of your task, then run DiNovo.

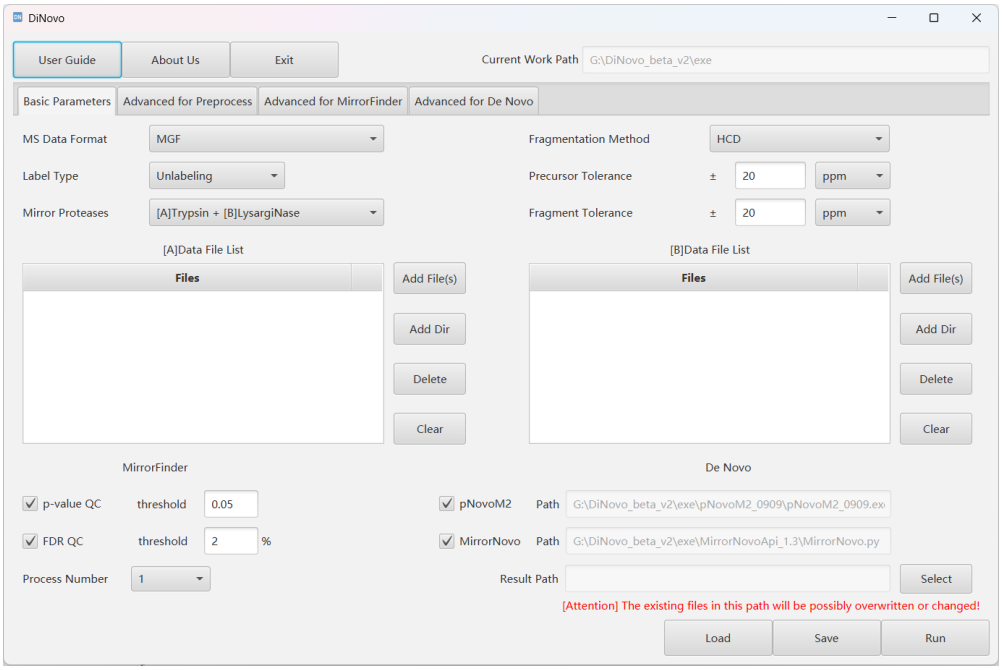


Figure 4. User interface of DiNovo.

Panel Introduction

There are four panels in GUI for configuration, including one for basic parameters and three for advanced parameters.

In basic parameters panel, you can directly start the analysis task after setting the basic parameters by click the button *Run*. If you want to run DiNovo by command line window, you could click the button *Save* to save the parameters file(s) after setting over. Clicking button *Load* could auto set these parameters by loading existed parameters file.

If there is any more refined parameter you want to adjust, you could set it in the related advanced parameters panel.

Command Line Mode

For using DiNovo by command line, open the command line window in the path of unzipped DiNovo folder, and run command:

DiNovoKernel.exe [param file path]

[param file path] is the path of configuration file of parameters(.cfg).

GUI Parameters Detail

The following is explanation of the meanings of parameters(Tab 1).

Table 1. Explanation of parameters in DiNovo

Panel	**Parameter**	**Explain**
Basic Parameters	MS Data Format	Format of Mass Spectrometry Data. MGF is only available now. Recommend using the data exported by pParse.
	Fragmentation Method	Method of fragmentation in MS/MS Data. HCD is only available now.
	Label Type	Label type of specified data. There are 2 available options: 1. Unlabeling. 2. NeuCode Labeling(K_{602} & K_{080} and R_{004} & R_{040}).
	Mirror Proteases	Mirror proteases approaches. There are 2 available options: 1. [A]Trypsin + [B]LysargiNase 2. [A]LysC + [B]LysN
	Precursor Tolerance	Maximum mass tolerance of precursor ions.
	Fragment Tolerance	Maximum mass tolerance of fragment ions.
	[A]Data File List	Data list of the sample digested by protease A.
	[B]Data File List	Data list of the sample digested by protease B.
	p-value QC & threshold	Quality control for mirror spectral pairs based on *p*-value. Threshold 0.05 is recommended.
	FDR QC & threshold	Quality control for mirror spectral pairs based on FDR. Threshold 2% is recommended.
	pNovoM2	De novo sequencing with pNovoM2.
	MirrorNovo	De novo sequencing with MirrorNovo.
	Process Number	Parallel accelerating for MirrorFinder and pNovoM2.
	Result Path	Specifying the path of result folder.
Advanced for Preprocess	Intensity enhancement by isotopic cluster	Add up the intensity of peaks in same isotopic cluster
	Remove precursor ions and related ions	Remove the precursor and its related ions.
	Remove common natural loss ions	Remove common natural loss ions (–H2O and –NH3).
	Consider precursor ion's charge	When precursor ion's charge is *c*, if considering this option, enumerate MS2 peak's charge from 1 to c, otherwise *c*-1.
	Reserve [x] peaks at most	1st round filtering, global noise peaks filter of preprocess.
	Reserve [x] peaks at most per [y] Da	2nd round filtering, local noise peaks filter of preprocess.
	Output preprocessed mgf	Output preprocessed MS data in .mgf format.
	Using double peak only	NeuCode detection approach.
	Using classification model	NeuCode detection approach. [ATTENTION]If it is chose to identify NeuCode peaks, DiNovo does preprocessing only.
	Maximum double peak intensity ratio(>1.0)	Parameter of "Using double peak only", that the maximum intensity ratio for identifying double peaks.
	Output NeuCode-labeling-annotated mgf	Output NeuCode-labeling-annotated mgf. There is a new line per output spectrum, 0 means peaks are identified noise, 1 means unlabeled, and 2 means the NeuCode-labeled.

Panel	Parameter	Explain
Advanced for MirrorFinder	Type A1 (type of mirror spectral pair)	[A]-protease peptide xxxxK Kxxxx [B]-protease peptide
	Type A2 (type of mirror spectral pair)	[A]-protease peptide xxxxR Rxxxx [B]-protease peptide
	Type A3 (type of mirror spectral pair)	[A]-protease peptide xxxx xxxx [B]-protease peptide
	Type B (type of mirror spectral pair)	[A]-protease peptide xxxxR Kxxxx [B]-protease peptide
	Type C (type of mirror spectral pair)	[A]-protease peptide xxxxK Rxxxx [B]-protease peptide
	Type D (type of mirror spectral pair)	[A]-protease peptide xxxxK xxxx [B]-protease peptide
	Type E (type of mirror spectral pair)	[A]-protease peptide xxxxR xxxx [B]-protease peptide
	Type F (type of mirror spectral pair)	[A]-protease peptide xxxx Kxxxx [B]-protease peptide
	Type G (type of mirror spectral pair)	[A]-protease peptide xxxx Rxxxx [B]-protease peptide
	Maximum Delta RT	The maximum delta of retention time of identified mirror spectral pair.
	Method (of Decoy Generation)	Only “Shifted Expected Score Bin” is supported now.
	Shifted [x] Da	Shifted the expected score mass bin with [x] Da.
Advanced for De Novo	Keep Top-N Result	The max reporting number of most possible peptides inferred per spectrum/spectral pair.
	Minimum Precursor Mass	The minimum precursor mass for de novo sequencing.
	Maximum Precursor Mass	The maximum precursor mass for de novo sequencing.
	Combine pNovoM2 and MirrorNovo result	Rescore the results of pNovoM2 by MirrorNovo, and combine them all(To be implemented).
	De novo sequencing from single-enzyme spectra	De novo sequencing not only from mirror-enzyme spectral pairs, but also all single-enzyme spectra
	pNovoM2 Mode	Special function of pNovoM2’s single-enzyme de novo sequencing mode(To be implemented).
	MirrorNovo Batch Size	For accelerating MirrorNovo. Setting it 2 to 10 needs about 4 to 10 GB GPU memory.
	Mapping the de novo peptide sequences to user-specific database	After finished sequencing task, the de novo sequencing results will be mapped with protein database.
	Database path	Specifying a database file (.fasta format) of protein sequence for mapping de novo sequencing results.

4. Output

After the program finished, the output window will prompt user with “finished!” and there will be several output files under the output path user specified.

you can use text editors (such as Notepad++, Sublime Text or Excel, etc.) to open the file(s) and view its contents. The following tables are details explanations about ALL result files and their

formats (Tab 2).

Table 2. Annotations of DiNovo output files

Result From	Output File Name	Explain
Preprocess	[x][y].mgf	Output preprocessed spectra in mgf format. [x] is original data name, and [y] means the approach of processing.
MirrorFinder	[MirrorFinder]MirrorSpecRes.txt	Output result of mirror spectra passing QC threshold(s).
	[MirrorFinder]MirrorSpecDis.txt	Output result of mirror spectra not passing QC threshold(s).
pNovoM2	[pNovoM2]MirrorSpecSeq.txt	De novo sequencing result from mirror-protease data.
	[pNovoM2]SingleSpecSeq_A.txt	De novo sequencing result from A-protease data.
	[pNovoM2]SingleSpecSeq_B.txt	De novo sequencing result from B-protease data.
	[pNovoM2]TOP1_Database_Coverage.txt	Mapping Top-1 de novo results to user-specified database.
	[pNovoM2]TOPN_Database_Coverage.txt	Mapping Top-N de novo results to user-specified database.
MirrorNovo	[MirrorNovo]MirrorSpecSeq.txt	De novo sequencing result from mirror-protease data.
	[MirrorNovo]SingleSpecSeq_A.txt	De novo sequencing result from A-protease data.
	[MirrorNovo]SingleSpecSeq_B.txt	De novo sequencing result from B-protease data.
	[MirrorNovo]TOP1_Database_Coverage.txt	Mapping Top-1 de novo results to user-specified database.
	[MirrorNovo]TOPN_Database_Coverage.txt	Mapping Top-N de novo results to user-specified database.

Here are some format explanations for the output results below (Tab 3, Tab 4).

Table 3. Description of output mirror spectra results

Column Name	Description
A_TITLE	Title of A-protease spectrum.
B_TITLE	Title of B-protease spectrum.
A_CHARGE	Charge state of peptide in A-protease spectrum.
B_CHARGE	Charge state of peptide in B-protease spectrum.
A_PM	Single-charge-state precursor mass of peptide in A-protease spectrum.
B_PM	Single-charge-state precursor mass of peptide in B-protease spectrum.
DELTA_PM	Delta of precursor masses of this spectral pair.
TARGET_P-VALUE	p-value of this spectral pair with target score bin.
DECOY_P-VALUE	p-value of this spectral pair with shifted score bin.
MIRROR_ANNO	Annotation of mirror spectra type.
TARGET_SCORE	Matching score of this spectral pair with target score bin.
DECOY_SCORE	Matching score of this spectral pair with shifted score bin.
FDR	Estimated FDR of mirror spectra. *It is only shown in [MirrorFinder]MirrorSpecRes.txt.

Table 4. Description of output de novo sequencing results

Column Name	Description
A_TITLE	Title of A-protease spectrum in mirror spectra. This column exists in MirrorSpecSeq only.
B_TITLE	Title of B-protease spectrum in mirror spectra. This column exists in MirrorSpecSeq only.
TITLE	Title of spectrum. This column exists in SingleSpecSeq only.
MIRROR_TYPE	The type of mirror spectra. This column exists in MirrorSpecSeq only.
CAND_RANK	The rank number of candidate sequence.
SEQUENCE	Sequence of candidate de novo sequencing peptide. In MirrorSpecSeq, it is combined sequence.
MODIFICATIONS	Site and name of modifications in one candidate de novo sequencing peptide.
PEPTIDE_SCORE	Score of peptide and spectra matching.
AA_SCORE	Score of AA (in candidate peptide) and spectra matching. It is shown in MirrorNovo's result only.
*Decoy	Label of that if this sequence is a decoy sequence in database
*Target (Protein/Top1Coverage/TopNCoverage)	Mapping info of this candidate sequence in database. This column exists in SingleSpecSeq only.
*TargetSuffix (Protein/Top1Coverage/TopNCoverage)	Mapping info of this candidate sequence (suffix means A-protease sequence) in database. This column exists in MirrorSpecSeq only.
*TargetPrefix (Protein/Top1Coverage/TopNCoverage)	Mapping info of this candidate sequence (prefix means B-protease sequence) in database. This column exists in MirrorSpecSeq only.
*SiteTag	Binary number tag of that if there is any evidence in every fragment site. In MirrorSpecSeq, it refers to A-protease peptide.
*MissNum	Number of fragment sites without any evidence in mirror spectra(number of 0 in SiteTag).
*Coverage	Coverage of theoretical fragment sites. Calculating it with (len(SiteTag)-MissNum) / len(SiteTag).

*It is only shown when setting mapping to database.