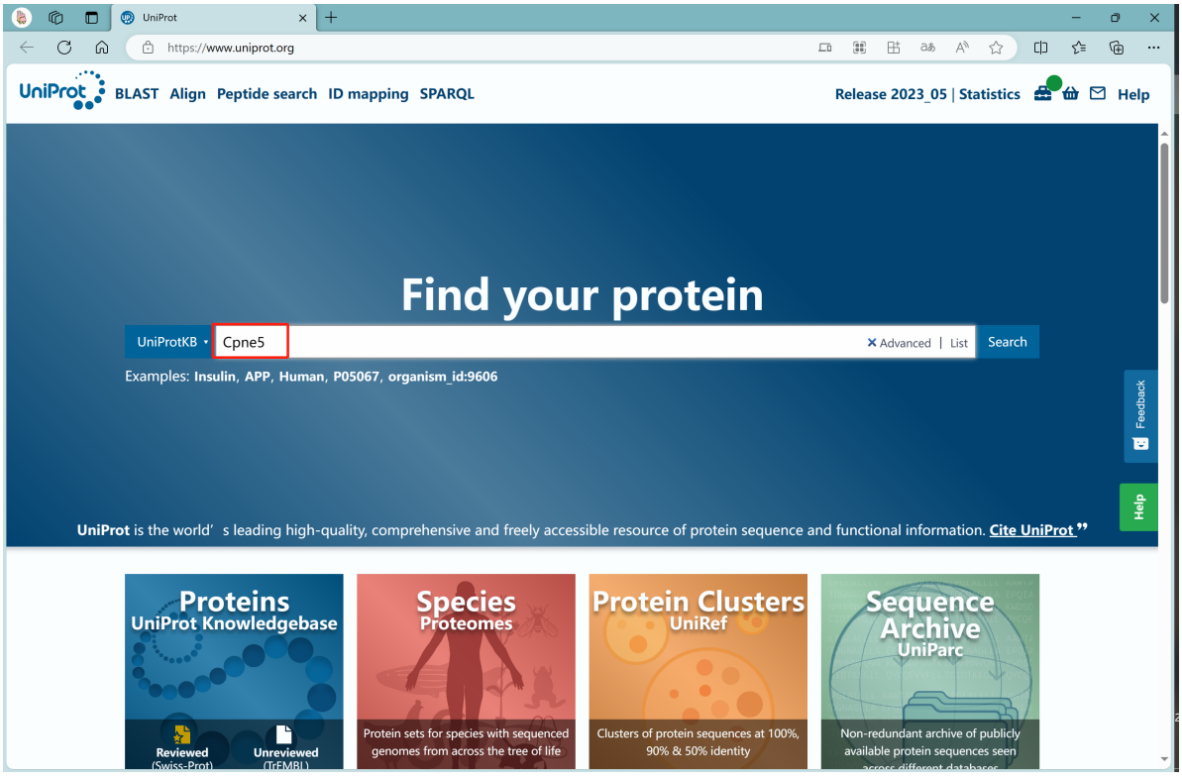
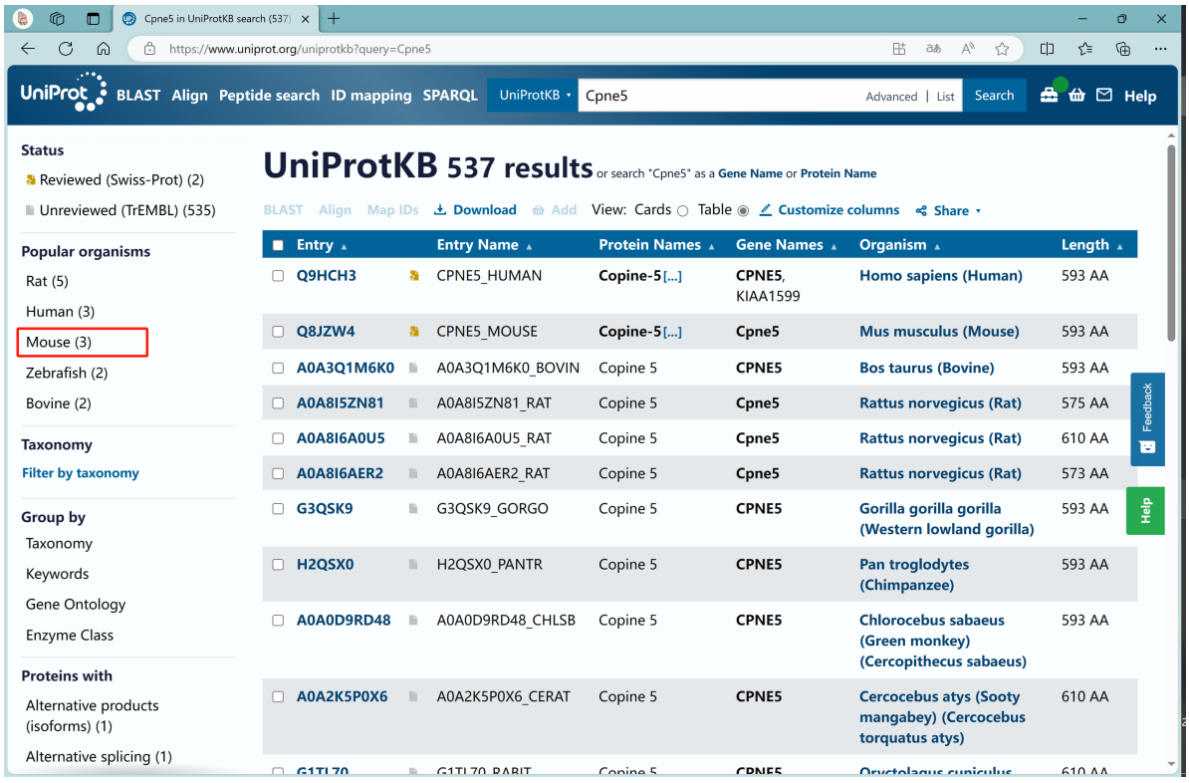
This part of the tutorial will guide you how to obtain the ESM-2 embedding of the gene.

Query

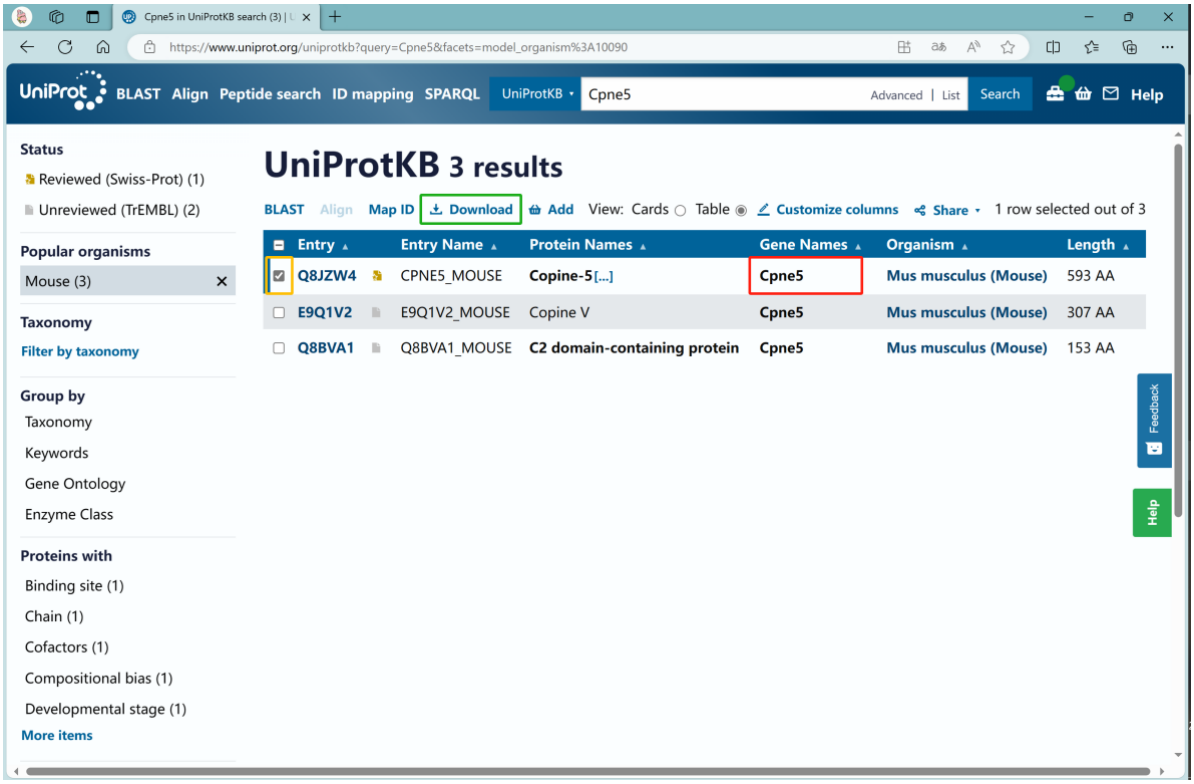
Go to UniProt, input the gene name in the search box (**Cpne5(Mouse)** is an example here), and **press Search**:



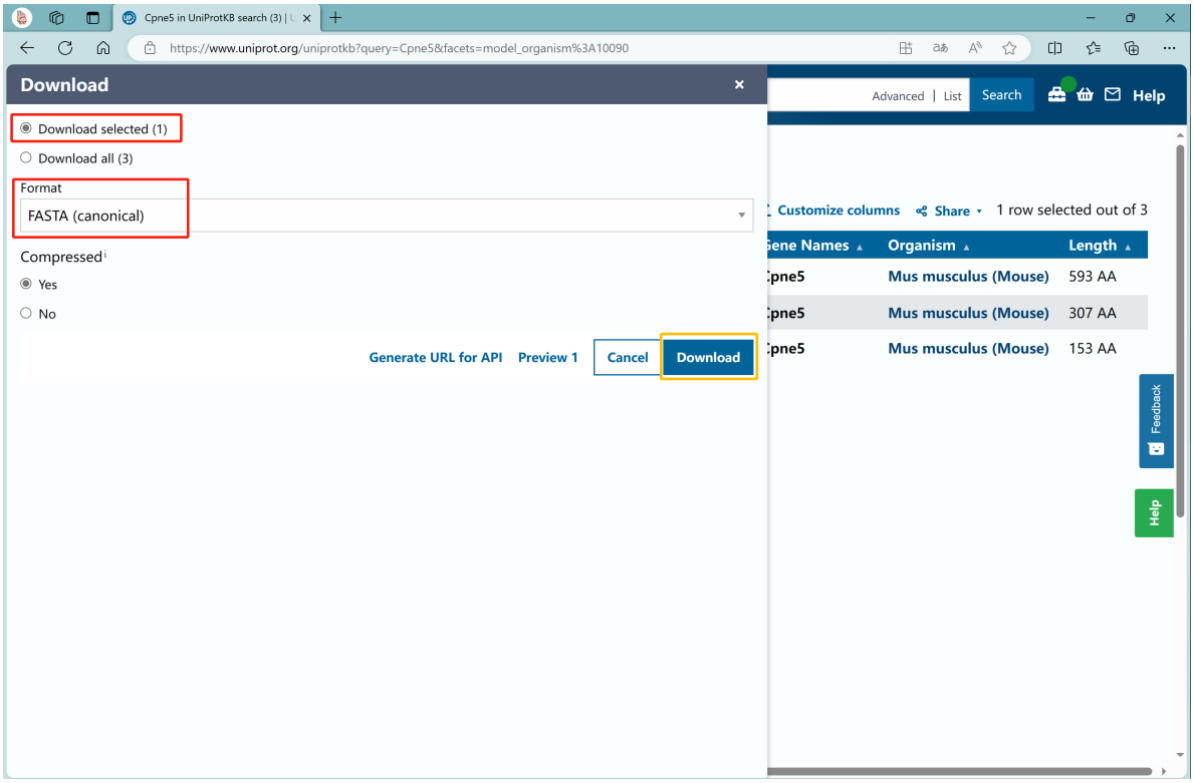
Then **select Mouse(or else you need)** in the left filter box:



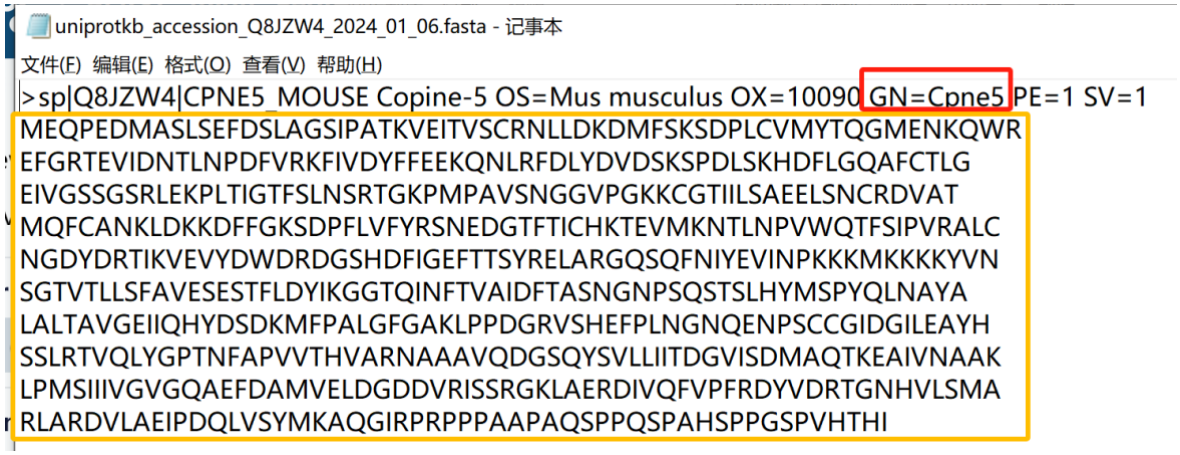
Make sure the red box is Cpne5, **click the yellow box** (select the first one if the Gene Names are the same), then **click Download** in the green box:



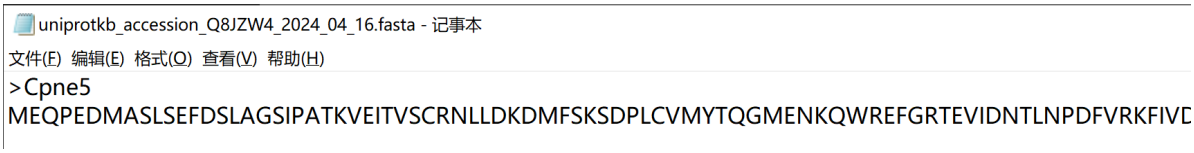
After determining that the red box is the FASTA format for downloading the selected 1 gene (this is the default), **click the yellow box** to download:



When you download and unzip it, you get a file with the gene name in red box and the protein expressions in yellow box:



Extract the gene name and protein expression, the following files are obtained:



ESM-2 Embedding

Installed the ESM-2 model (you can refer to ESM-2), then **enter the following command:**

```
cd esm-main/

python scripts/extract.py esm2_t36_3B_UR50D
uniprotkb_accession_Q8JZW4_2024_04_16.fasta examples/data/some_proteins_emb_esm2
--repr_layers 36 --include mean per_tok
```

Where `uniprotkb_accession_Q8JZW4_2024_04_16.fasta` is the gene name and protein expression file processed above.

After running successfully, the gene embedding file (in this case, `Cpne5.pt`) is generated in the `esm-main/examples/data/some_proteins_emb_esm2/` directory.

After converting all the genes of your dataset to some *.pt files, **put them in a folder** (such as `./pt/`) and **run the following python code**:

```
import os
import torch
import pickle
import pandas as pd

df = pd.DataFrame()

for path, dir_lst, file_lst in os.walk(r'./pt'):
    for file_name in file_lst:
        data = torch.load(open(os.path.join(path, file_name), 'rb'))
        print(data['label'])
        print(data['representations'][36][-1])
        df.insert(df.shape[1], data['label'], data['representations'][36]
[-1].numpy())
```

```
pickle.dump(df, open('emb.pkl', 'wb'))
```

Then the ESM-2 embedding file `emb.pkl` of the genes can be generated.