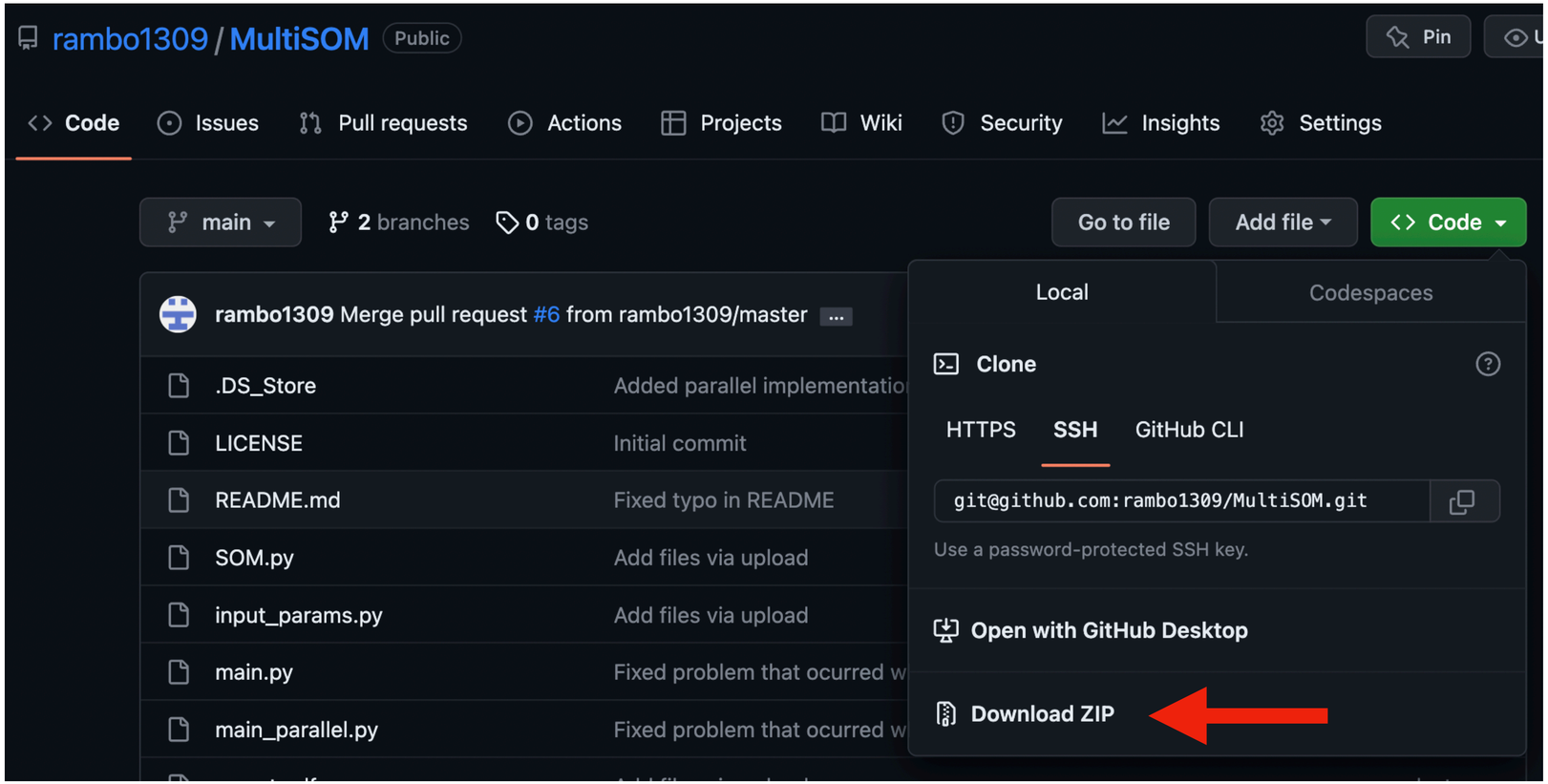
MultiSOM: Tutorial, tips and tricks

Introduction:

- Our goal is to extract patterns and useful information from data (data mining).
  → clustering
- Main steps:
  1. Feature and sample selection
  2. Model tuning
  3. Result interpretation
- This presentation focuses only on how to use the software.

How to download:

- Option 1: Clone directly from the repository (https://github.com/rambo1309/MultiSOM)
- Option 2: Download as zip



We need to have the MultiSOM script in the same folder as the files we want to analyse.

Feature and sample selection:

- Taking our time to carefully assess and select the input features is crucial.
- Building a feature pipeline is highly recommended when analysing multiple files.
- Noisy features can be previously averaged over neighbours for better results.

Selecting files:

- Select training sample and write its file path in the `'training_file'`.
- Select samples that are going to be analysed after training, and write their file paths in the `'file_list'`.
- Python's list comprehension can be used in `'file_list'`.
- Remember that the training file is not classified unless included in `'file_list'`.

input_params.py

Users > francoaquistapace > Desktop > Temp_Samples > Cu50Zr50_nanoindentation > input_params.py

```
PARAMS = {'file_list' : ['dump_GPa_20.indentacion.'+str(int(i*10000))+'_shear_strain.config'
                         for i in [0,8,16,24,32,40,48,56,100]],
          'search_pattern' : '',
          'training_file' : 'sliced_dump_GPa_0.indentacion.1000000_shear_strain.config',
```

Model tuning:

- Only use layers as needed, since they increase training time.
- The number of nodes, ’N’ is usually a very impactful hyper-parameter.
- As of today, the best way to find the optimal settings is by trial-and-error.
- Try adjusting one hyper-parameter at a time. To find what works and what doesn’t.

Defining model layers and hyper-parameters:

- For every layer we have to specify the following dictionary in the LAYERS section:

input_params.py ×

Users > francoaquistapace > Desktop > Temp_Samples > Cu50Zr50

```
LAYERS = [{'features' : ['gr_coord_10','c_vor1'],
           'scaling' : ['standard','normal'],
           'f' : 1,
           'sigma' : 0.7,
           'eta' : 1.0,
           'N' : 2,
           'Batched' : True,
           'Batch_size' : 100},
```

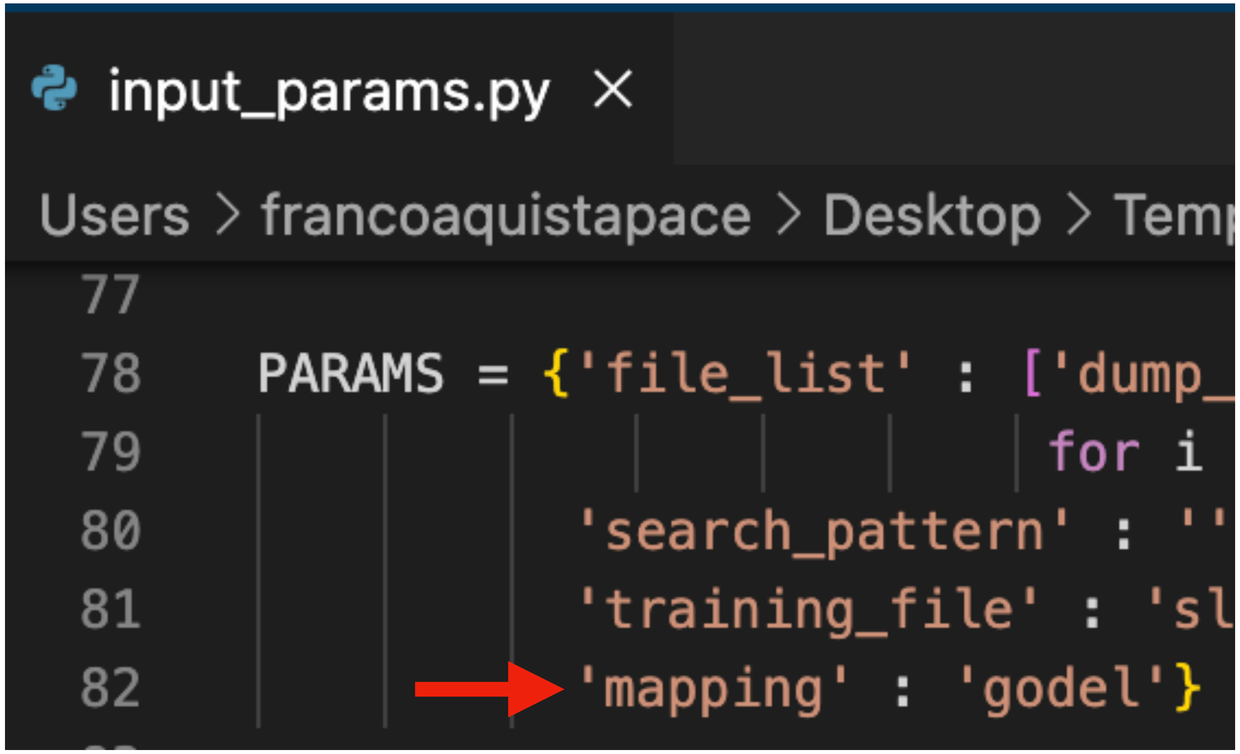
- We may use the same features in different layers.
- The ‘Batch_size’ parameter is only used when ‘Batched’ is set to True.

Result interpretation:

- Select the mapping style that best fits your post-processing pipeline.
- When results are analysed with OVITO, and the model has multiple layers, the ‘default’ mapping is recommended.
- For up to two layers, the ‘linear’ mapping can be easier to inspect visually.
- ‘godel’ mapping is only recommended when integer mapping is necessary.

Selecting output mapping:

- We need to specify one of three alternative mappings in the PARAMS section: ‘godel’, ‘linear’ or ‘default’.



- Then, the clustered results have to be inspected in some visualisation software, like OVITO.