SATURN report

May 8th 2023 – July 28th 2023

Papatheodorou Group
Supervisors: Yuyao Song, Dr. Irene Papatheodorou
Internship funded by the Embassy of France

Anaëlle Cossard
Paris-Saclay University

Cross-species integration of single cell RNA-seq data from the primary motor cortex between human and mouse using SATURN

To perform SATURN, I used the datasets we selected at the beginning of the internship. No big modifications of the anndata objects are required, only a filtering on the cells and the genes, adding a species column and a cell type column in the *.obs* dataframe plus putting the names of the genes as the *.var_names* of the anndata object. This part on the names is not specified but since the protein embeddings files are from *Ensembl*, the gene names required seems to be HGNC names at least for the human.

I runed several integrations with SATURN, mostly always with the default parameters (HVG = 8000, macrogenes = 2000). I tried first only for the human and mouse again, with the whole dataset (fig. 1), only *GABAergic* (fig. 2), only *Glutamatergic* (fig. 3) and only *Non-Neuronal* (fig. 4). These integrations might need more training, it should be possible to pass in the arguments a file with the *centroids*[1]. Unfortunately, this file is not generated automatically and when I tried, I had a memory issue[2], even with 150G on the GPU node. By doing that, the *centroids of the macrogenes* should not be calculated from scratch again but improved after each training.

Finally, I tried to perform the integration with the fly dataset as well, unfortunately there is a problem with the gene names[3]. This part of the scripts should be the part that take the gene names from the anndata object and match them with the ones in the proteome embeddings files. The index was first of type *CategoricalIndex* that I change to str and was instead of type *object*, but it should not be the problem since the mouse dataset also have an index of type *CategoricalIndex* that does not cause any problem. The gene names annotation is from *FlyBase* which seems to be the one use by *Ensembl*. Sadly, I did not have the time to investigate this problem.

[1] With the command --*centroids_init_path* that either indicates the path of the file or where it should be save when done for the first time

[2]
```
File "/nfs/research/irene/anaelle/Scripts/SATURN/model/saturn_model.py", line 105, in
forward
    expr = torch.zeros(batch_size, self.num_genes).to(inp.device)
RuntimeError: CUDA out of memory. Tried to allocate 166.00 MiB (GPU 0; 31.75 GiB total
capacity; 305.90 MiB already allocated; 26.69 MiB free; 330.00 MiB reserved in total by
PyTorch) If reserved memory is >> allocated memory try setting max_split_size_mb to avoid
fragmentation.  See documentation for Memory Management and PYTORCH_CUDA_ALLOC_CONF
```

[3]
```
File "/nfs/research/irene/anaelle/Scripts/SATURN/data/gene_embeddings.py", line 110, in
<setcomp>
    genes_to_use = {gene for gene in adata.var_names if gene.lower() in
genes_with_embeddings}
AttributeError: 'float' object has no attribute 'lower'
```

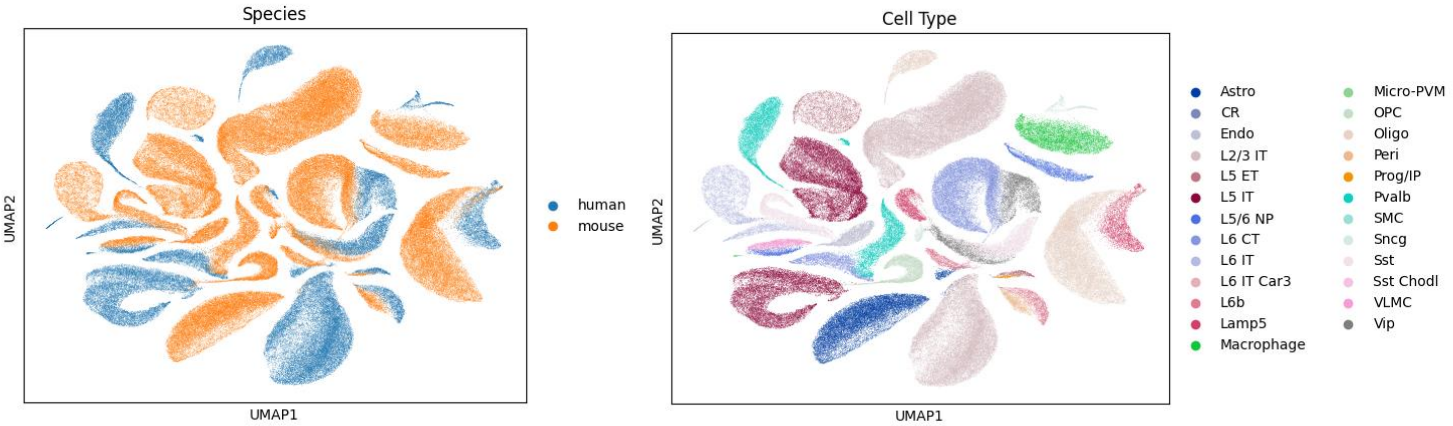

Figure 1: UMAP plots after the first training on human and mouse data, using the whole datasets. First plot is colored by species and second plot is colored by cell types.

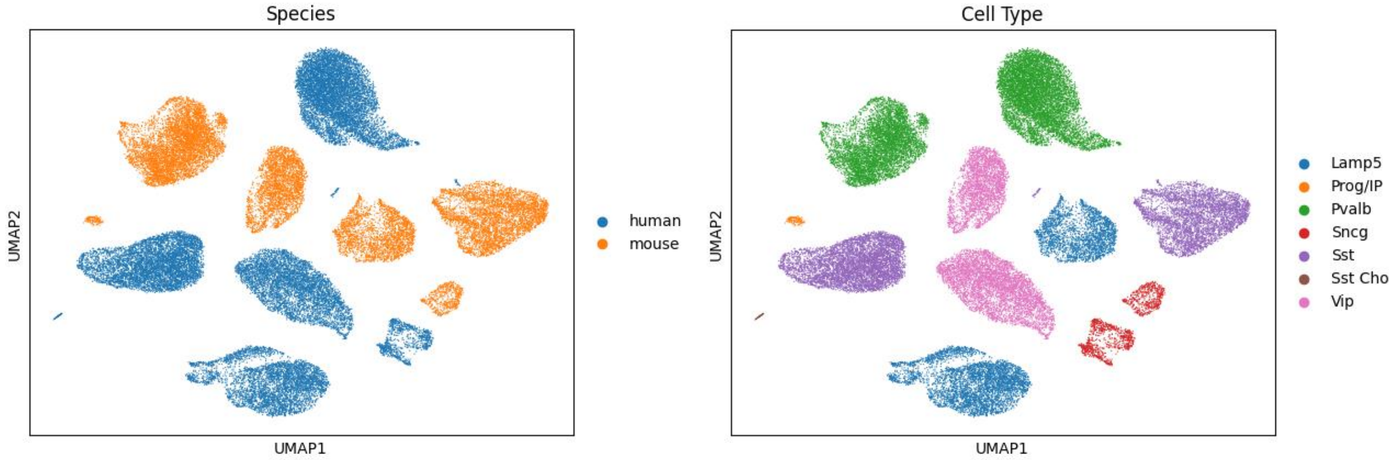

Figure 2: UMAP plots after the second training on human and mouse, only using GABAergic cells. First plot is colored by species and second plot by cell types.

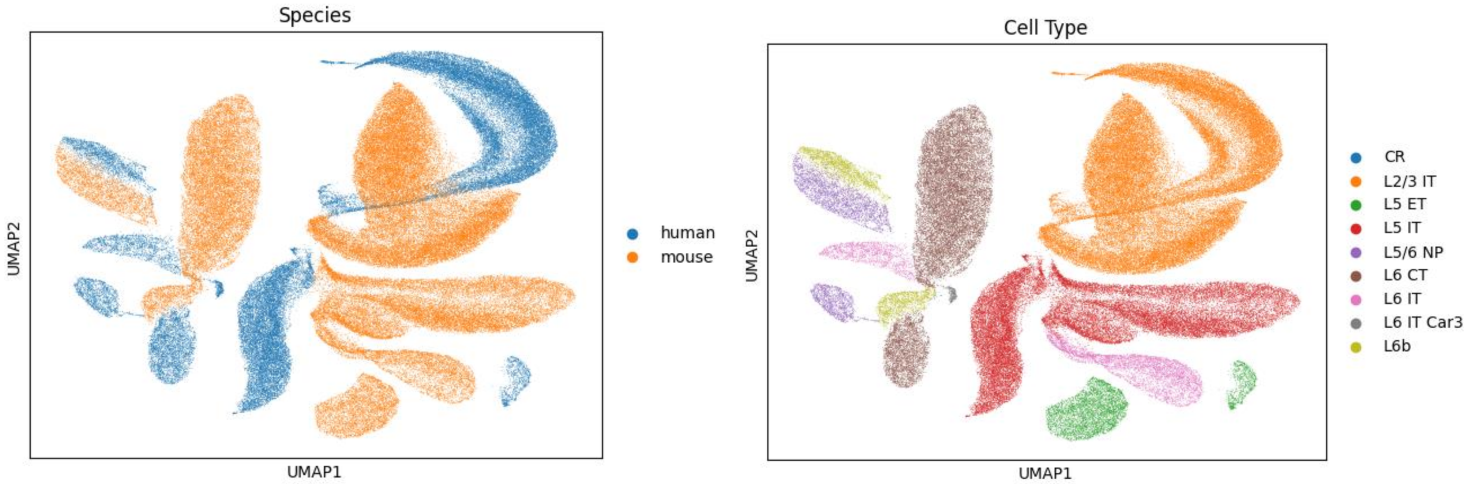

Figure 3: UMAP plot of the third training on human and mouse, only using Glutamatergic cells. First plot is colored by species and second plot by cell types.

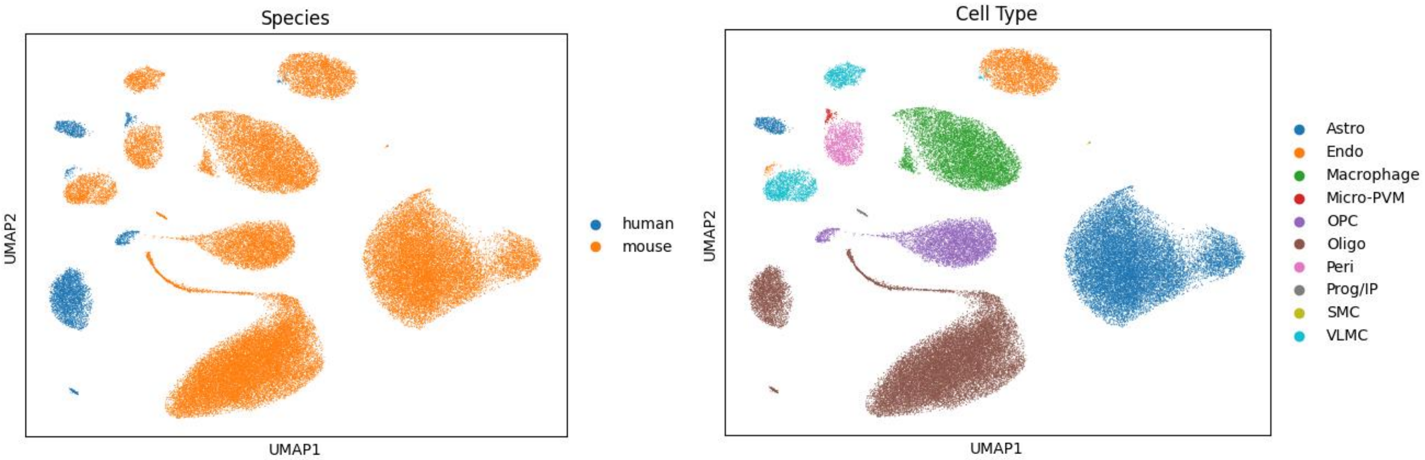

Figure 4: UMAP plot of the fourth training on human and mouse, only using Non-Neuronal cells. First plot is colored by species and second plot by cell types.