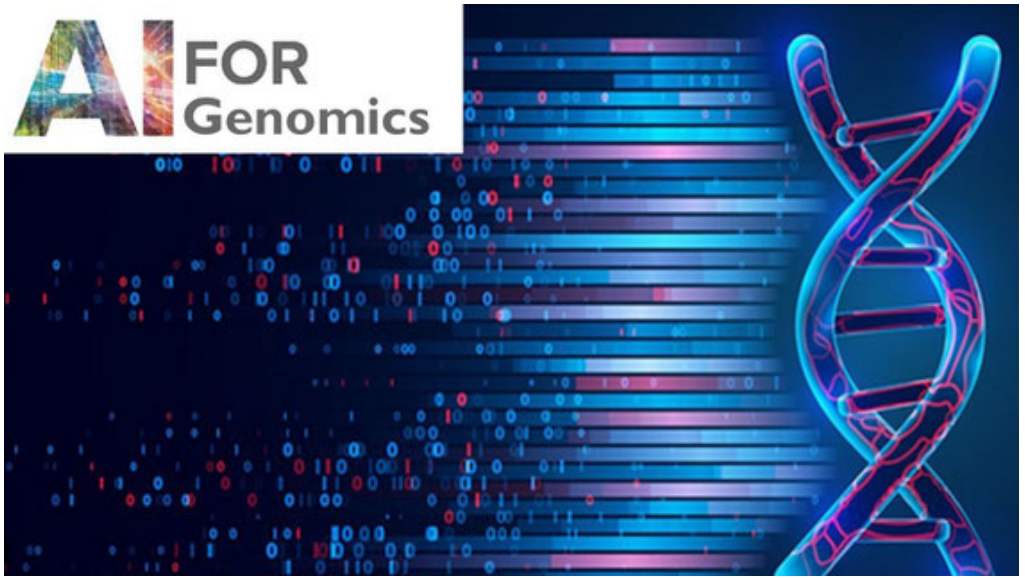
AI for Genomics: from CNNs and LSTMs to Transformers

Physalia course, online, 9-11 September 2025

Course outline and practical information

Nikolay Oskolkov, Group Leader (PI) at LIOS, Riga, Latvia



Physalia
Courses

About us

Organizer: Carlo Pecoraro, Physalia courses

info@physalia-courses.org

Instructor:

Dr. Nikolay Oskolkov, Lund University, NBIS SciLifeLab

Nikolay.Oskolkov@biol.lu.se

@NikolayOskolkov

@oskolkov.bsky.social

Personal homepage:
https://nikolay-oskolkov.com

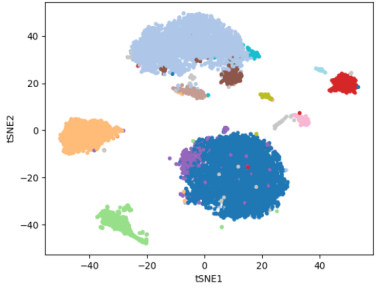
Brief introduction: who am I

2007 PhD in theoretical physics in Moscow

2011 medical genetics at Lund University

2016 working at NBIS SciLifeLab, Sweden

Single cell



Biomedical data integration

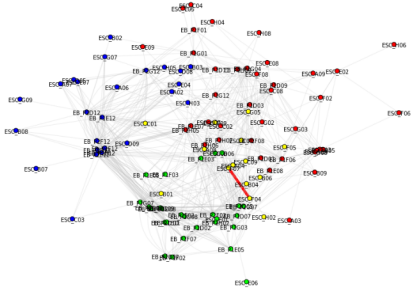
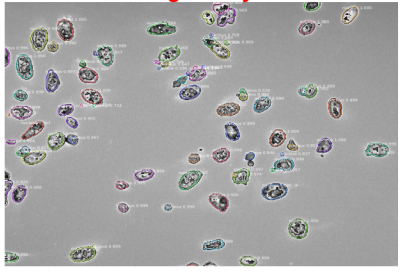
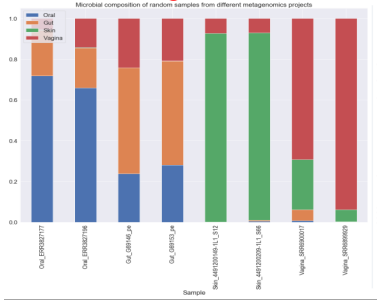


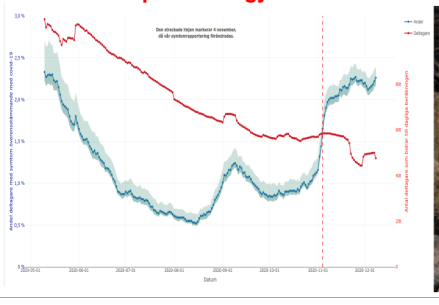
Image analysis



Metagenomics

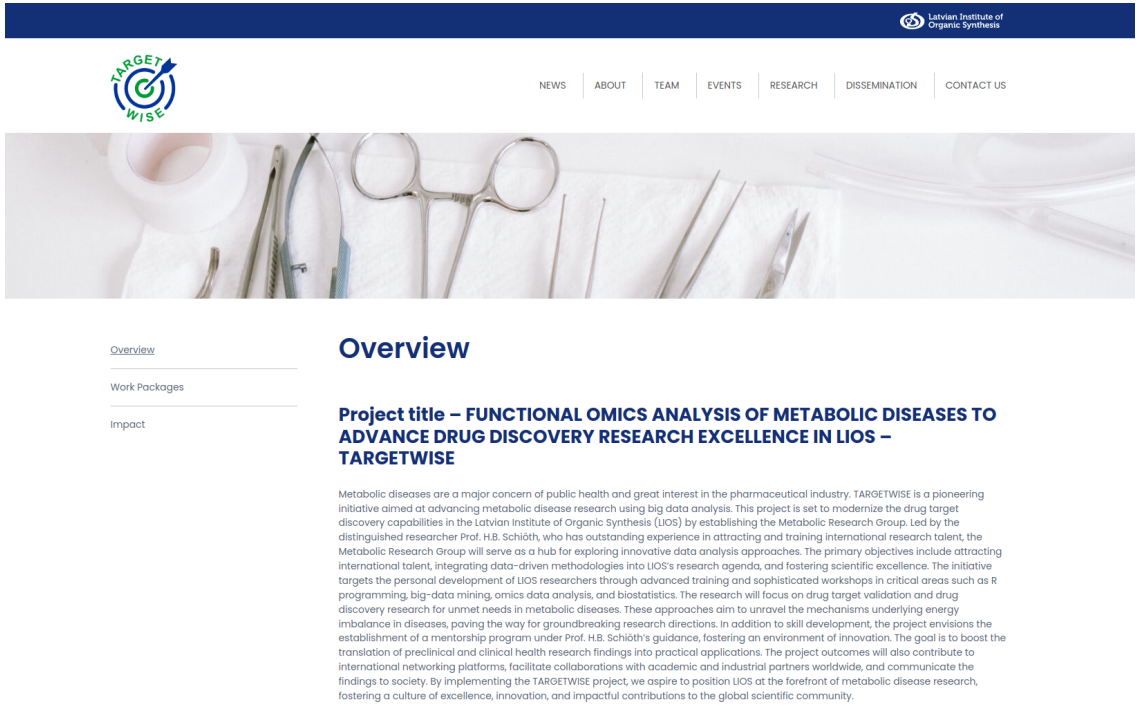


Epidemiology



Ancient microbiome

Starting own research group in Riga

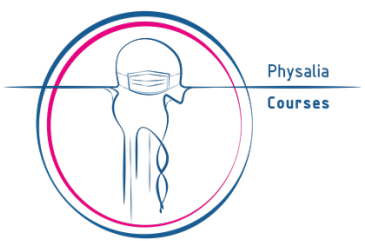


2 open postdoc positions, 2 PhD positions to be open this fall.

If you know anyone who might be interested, please contact me!

About you

- Name
- University / Institute / Company
- Research interest(s)
- Previous experience with computational analysis and bioinformatics
- Motivation to join the course
- Expectations from the course

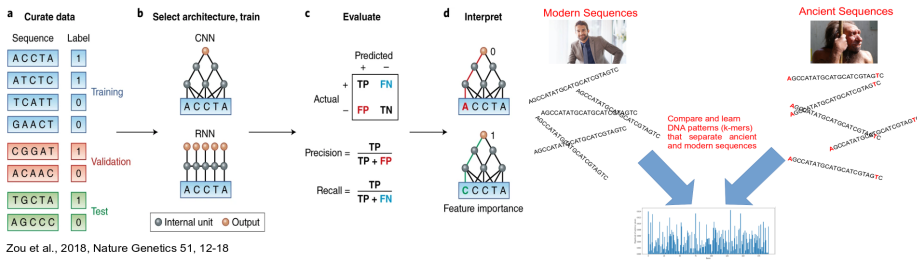


Course outline

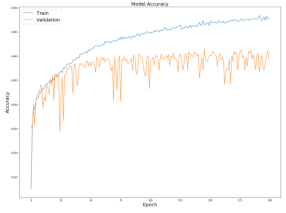
Day 1: introduction to machine and deep learning, applications of Artificial Neural Networks (ANNs) in Life Sciences, coding gradient descent and a vanilla ANN from scratch in R and Python

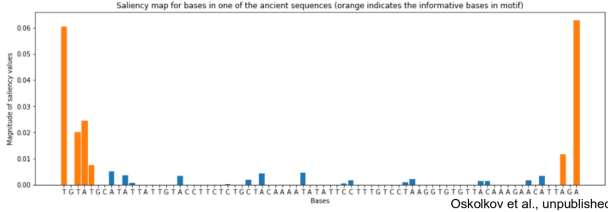
Day2: autoencoder ANN and its applications to single cell genomics and integration of heterogeneous data

Day3: convolutional (CNN) and recurrent (LSTM) neural networks, their applications in genomics and metagenomics, attention mechanism and Transformer

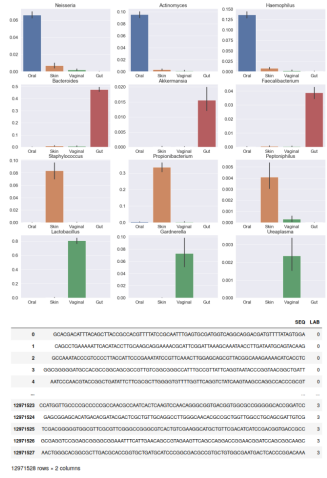


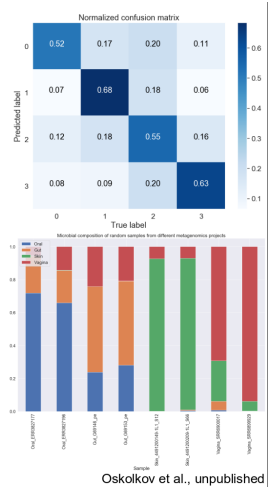
Zou et al., 2018, Nature Genetics 51, 12-18





Oskolkov et al., unpublished





Oskolkov et al., unpublished

Housekeeping rules

- To ask question please raise your hand, unmute yourself and ask. You can also ask questions in the zoom chat or slack channel for this seminar.
- Please keep your camera on as much as possible for better contact and communication.
- The course includes 6 lectures (~1h each) and 7 practicals (~1.5h each), there is a 15 min break after each lecture and each practical
- During practicals, we will use Rmarkdown and Jupyter notebooks. At the end of each practical, I will use html-versions of the notebooks to go through command lines with my explanations.
- The material is based on data and problems from computational biology, however the concepts discussed are general and can be applied for other types of data.

Practical information: GitHub and Zoom

The course will take place in Zoom from 2 pm to 8 pm (CET, Berlin time)

Links to the Zoom room will be posted in the Slack channel

The course GitHub repository containing lectures and exercises is:

https://github.com/NikolayOskolkov/Physalia_AI_Genomics

Please bookmark this address!

Physalia
Courses

Physalia_AI_Genomics

Public

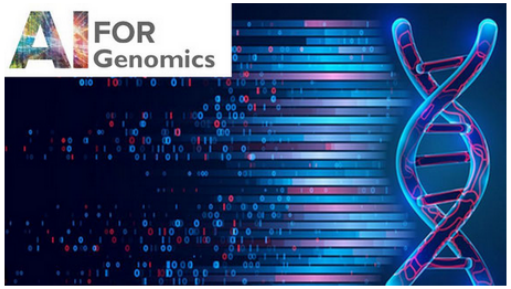
Pin Watch 0 Fork 0 Star 0

main 1 Branch 0 Tags Go to file t Add file Code

LeandroRitter modified links a4c6cef · 9 minutes ago 6 Commits

AI_Genomics_Oskolkov_ syllabus.pdf	added logo and modified the program	34 minutes ago
README.md	modified links	9 minutes ago
command-line-basics.md	added logo and modified the program	34 minutes ago
course_logo.jpg	added logo and modified the program	34 minutes ago

README



AI for Genomics

Instructor

- Dr. Nikolay Oskolkov, Group Leader (PI) at LIOS, Riga, Lativia

Course overview

This course explores the application of modern AI architectures—Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and Transformers—to genomic and metagenomic data. Students will gain practical experience through hands-on coding labs and interactive notebooks, learning how to model sequence data, extract biologically meaningful features, and interpret results. Emphasis is placed on real-world applications, including prediction of genomic functional elements, sequence classification and source tracking, as well as biological sequence generation.

Target audience and assumed background

About

This is a repository with the course material for Physalia AI for Genomics course

Readme
Activity
0 stars
0 watching
0 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 2

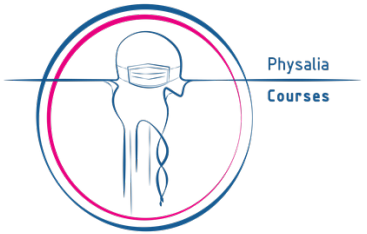
LeandroRitter Nikolay Oskolkov
NikolayOskolkov Nikolay Oskolkov

Practical information: Amazon Cloud (AWS EC2)

We will use the Cloud Computing service from Amazon, which we will access via `ssh` (secure shell protocol)

```
ssh -i multiomics.pem -X ubuntu@54.244.40.220
```

- The IP address changes every day
- Everyone is given a username, with a `home` and `shared` folders
  - List of usernames can be found in Slack
  - The `shared` folder is copy-only: do not delete, move, rename, or write



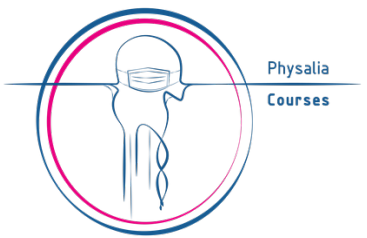
However, most of the time you will be using Rstudio and Google Colab (for Python exercises)!

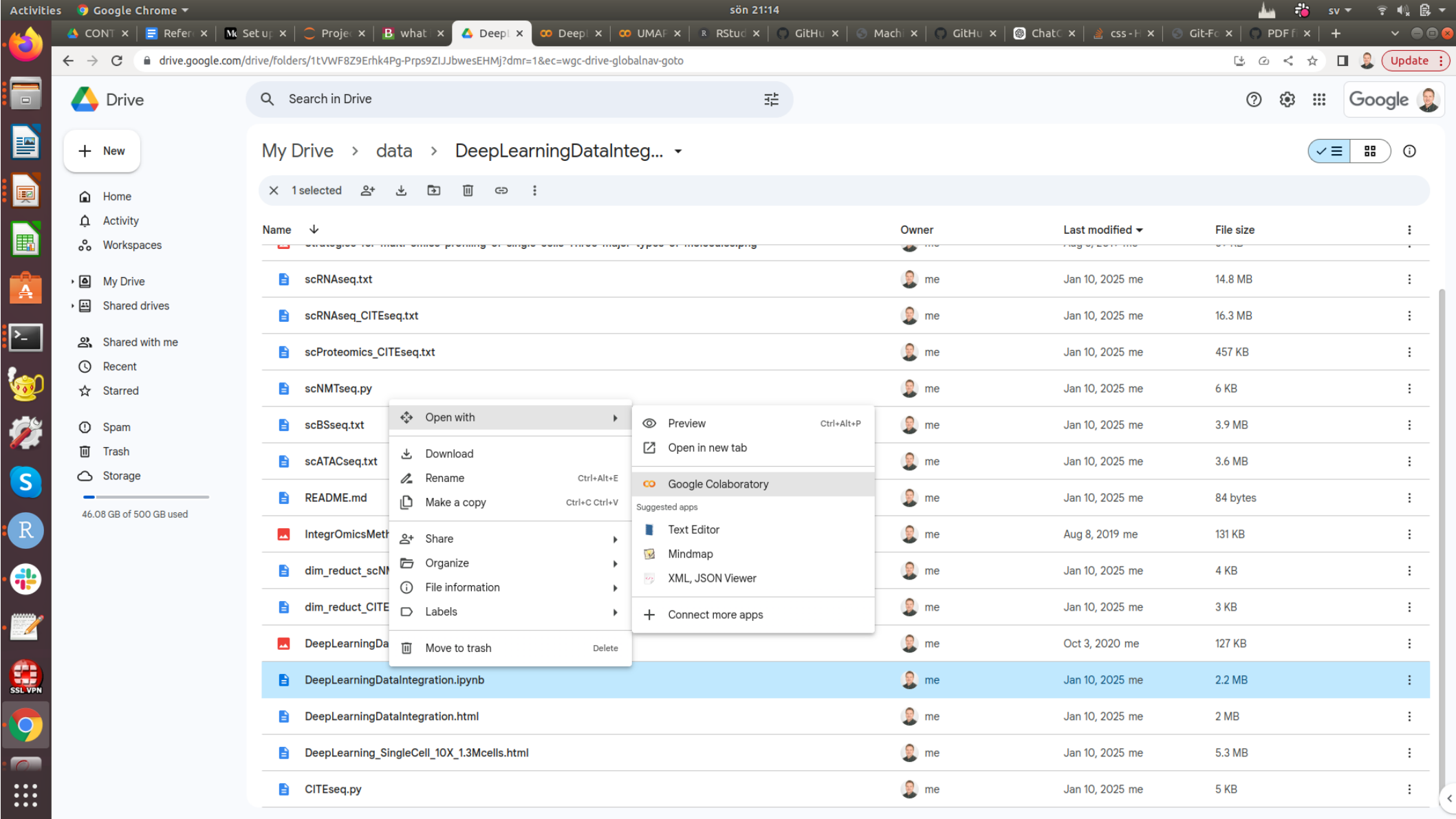
To run Jupyter notebooks on Google Colab please do the following:

1. Clone this repo: https://github.com/NikolayOskolkov/Physalia_AI_Genomics
2. Unzip the data.zip folder locally and upload it to your Google drive
3. Start the Jupyter notebook from the Google drive and mount google drive:

```
#Mount Google Drive to Colab

from google.colab import drive
drive.mount('/content/drive')

!ls -l /content/drive/MyDrive/

import os
os.chdir("/content/drive/MyDrive/data")
```

N.B. you need a Google account to be able to run Google Colab!




Activities
Google Chrome
sön 21:14
drive.google.com/drive/folders/1tVWF8Z9Erhk4Pg-Prps9ZIJJbwesEHMj?dmr=1&ec=wgc-drive-globalnav-goto
Update
Drive
Search in Drive
Google
New
Home
Activity
Workspaces
My Drive
Shared drives
Shared with me
Recent
Starred
Spam
Trash
Storage
46.08 GB of 500 GB used
My Drive › data › DeepLearningDataInteg...
1 selected
Name
Owner
Last modified
File size
scRNAseq.txt me Jan 10, 2025 me 14.8 MB
scRNAseq_CITEseq.txt me Jan 10, 2025 me 16.3 MB
scProteomics_CITEseq.txt me Jan 10, 2025 me 457 KB
scNMTseq.py me Jan 10, 2025 me 6 KB
scBSseq.txt me Jan 10, 2025 me 3.9 MB
scATACseq.txt me Jan 10, 2025 me 3.6 MB
README.md me Jan 10, 2025 me 84 bytes
IntegrOmicsMeth me Aug 8, 2019 me 131 KB
dim_reduct_scN me Jan 10, 2025 me 4 KB
dim_reduct_CITE me Jan 10, 2025 me 3 KB
DeepLearningDa me Oct 3, 2020 me 127 KB
DeepLearningDataIntegration.ipynb me Jan 10, 2025 me 2.2 MB
DeepLearningDataIntegration.html me Jan 10, 2025 me 2 MB
DeepLearning_SingleCell_10X_1.3Mcells.html me Jan 10, 2025 me 5.3 MB
CITEseq.py me Jan 10, 2025 me 5 KB
Open with
Download
Rename Ctrl+Alt+E
Make a copy Ctrl+C Ctrl+V
Share
Organize
File information
Labels
Move to trash Delete
Preview Ctrl+Alt+P
Open in new tab
Google Colaboratory
Suggested apps
Text Editor
Mindmap
XML, JSON Viewer
Connect more apps

Deep Learning for Data Integration

Biological and biomedical research has been tremendously benefiting last decade from the technological progress delivering DNA sequence (genomics), gene expression (transcriptomics), protein abundance (proteomics) and many other levels of biological information commonly referred to as OMICs. Despite individual OMICs layers are capable of answering many important biological questions, their combination and consequent synergistic effects from their complementarity promise new insights into behavior of biological systems such as cells, tissues and organisms. Therefore OMICs integration represents the contemporary challenge in Biology and Biomedicine.

```
[ ] from google.colab import drive
    drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

```
[ ] !ls -l /content/drive/My\ Drive/data/DeepLearningDataIntegration
```

```
total 49588
-rw------- 1 root root     5105 Jan 10 12:26 CITEseq.py
-rw------- 1 root root  2066933 Jan 10 12:26 DeepLearningDataIntegration.html
-rw------- 1 root root  1747228 Jan 10 12:41 DeepLearningDataIntegration.ipynb
-rw------- 1 root root   130233 Oct  3  2020 DeepLearningDataIntegration.jpg
-rw------- 1 root root  5598515 Jan 10 12:26 DeepLearning_SingleCell_10X_1.3Mcells.html
-rw------- 1 root root     3499 Jan 10 12:26 dim_reduct_CITEseq.py
-rw------- 1 root root     3616 Jan 10 12:26 dim_reduct_scNMTseq.py
-rw------- 1 root root   134229 Aug  8  2019 IntegrOmicsMethods.png
-rw------- 1 root root       84 Jan 10 12:26 README.md
-rw------- 1 root root  3772209 Jan 10 12:26 scATACseq.txt
-rw------- 1 root root  4082712 Jan 10 12:26 scBSseq.txt
-rw------- 1 root root     6141 Jan 10 12:26 scNMTseq.py
-rw------- 1 root root   467570 Jan 10 12:26 scProteomics_CITEseq.txt
-rw------- 1 root root 17128751 Jan 10 12:26 scRNAseq_CITEseq.txt
-rw------- 1 root root 15547715 Jan 10 12:26 scRNAseq.txt
-rw------- 1 root root    71008 Aug  8  2019 Strategies-for-multi-omics-profiling-of-single-cells-Three-major-types-of-molecules.png
-rw------- 1 root root     3647 Jan 10 12:26 tSNE_on_Autoencoder_CITEseq.py
-rw------- 1 root root     3635 Jan 10 12:26 tsne_on_autoencoder_scNMTseq.py
```

```
import os
os.chdir("/content/drive/My Drive/data/DeepLearningDataIntegration")
```

```
[ ] from IPython.display import Image
    Image('DeepLearningDataIntegration.jpg', width=2000)
```