

Machine learning, Galaxy and more

Anup Kumar

Bioinformatics group, University of Freiburg, Freiburg, Germany

26th October 2021

Agenda

- Basics of machine learning
- Machine learning in Galaxy
- Ongoing machine learning projects - Jupyterlab editor for ML and predicting protein evolution in SARS-COV2 sequences using deep learning

Basics of Machine learning

Machine learning (ML)

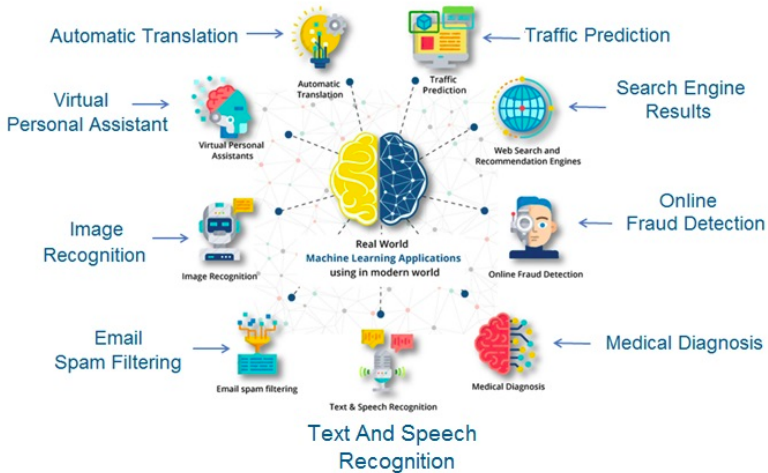
- ML - computer program that learns rules from data
- Use rules to distinguish patterns
- Rules are mathematical functions
- Learn on existing (training) data, predict unknown outputs (of test data)
- ML algorithms work on numbers and not text or characters
- Example task: handwritten digit recognition



1

¹https://en.wikipedia.org/wiki/MNIST_database

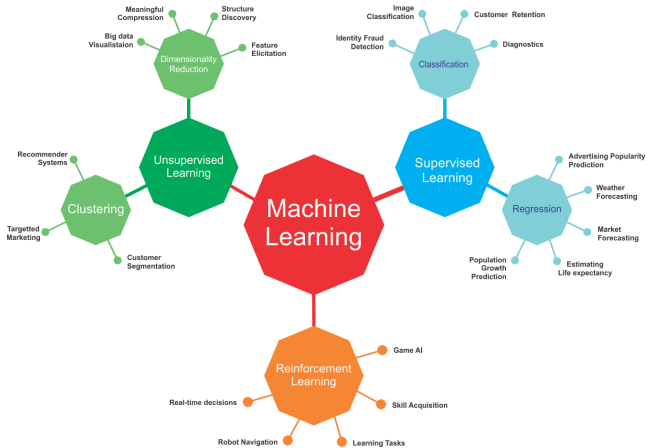
General applications of ML



2

²<https://www.learncomputerscienceonline.com/what-is-machine-learning/>

Types of ML



3

³<https://skillix.com/list-of-machine-learning-algorithms/>

Supervised learning (Classification)

- Labeled data
- Features (gender, height, weight, index)
- Labels/classes/targets/output (status)

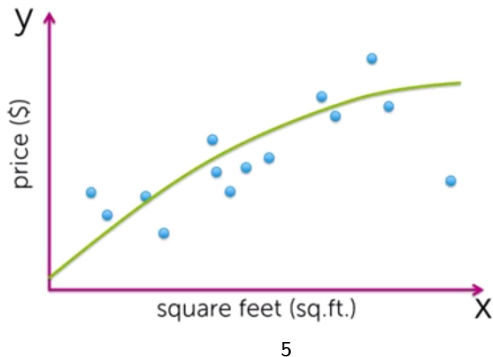
	Gender	Height	Weight	Index	Status
0	Male	174	96	4	Obesity
1	Male	189	87	2	Normal
2	Female	185	110	4	Obesity
3	Female	195	104	3	Overweight
4	Male	149	61	3	Overweight
5	Male	189	104	3	Overweight
6	Male	147	92	5	Extreme Obesity
7	Male	154	111	5	Extreme Obesity
8	Male	174	90	3	Overweight
9	Female	169	103	4	Obesity

4

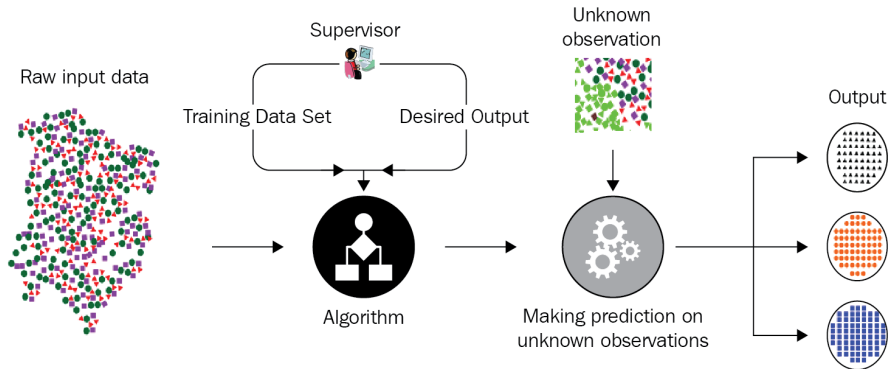
⁴<https://www.kaggle.com/yersever/500-person-gender-height-weight-bodymassindex>

Supervised learning (Regression)

- Labeled data
- Class is a real number instead of a category
- Example: house price prediction



Pipeline for supervised learning



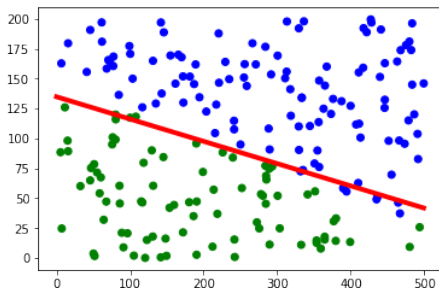
6

Algorithms for supervised learning

- Linear models
- Support vector machines
- Decision trees
- Ensemble models
- Neural networks
- ...

Linear models

- Learn straight line decision boundary
- Easy to use and fast
- Don't learn non-linear features

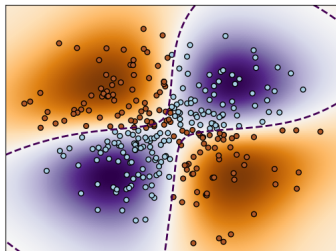


7

⁷<https://stats.stackexchange.com/questions/436827/why-does-linear-non-logistic-regression-work-as-a-linear-classifier-what-classi>

Non-linear models

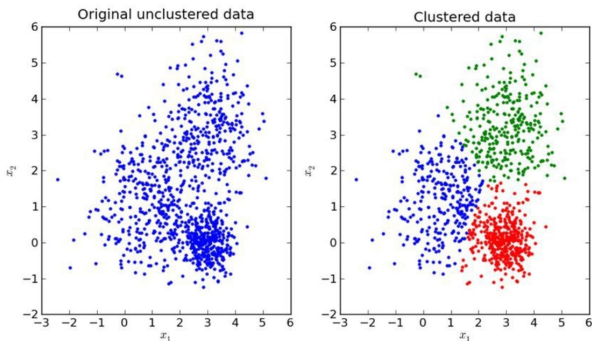
- Many times, patterns can only be separated by non-linear boundaries
- Linear models are not sufficient
- Need algorithms to learn non-linear features in data
- Examples: support vector machines, k-nearest neighbours, decision trees, ensemble methods ...



8

Unsupervised learning

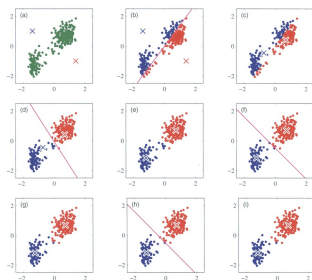
- Datasets have no labels - no supervision
- Extract structures in datasets
- Unsupervised approaches - clustering, dimensionality reduction, ...



9

Clustering

- Group data points based on similarity
- Similarity is determined by a notion of closeness
- Iterative process
- Types of clustering: k-means, hierarchical clustering, density-based spatial clustering of applications with noise (DBSCAN)

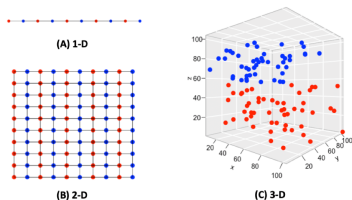


10

¹⁰<http://dendroid.sk/2011/05/09/k-means-clustering/>

Dimensionality reduction

- Number of dimensions \gg number of samples
- High dimensional dataset - data points become farther from one another
- Curse of dimensionality - hard to generalise on all combinations of a large number of dimensions
- May lead to high-variance or overfitting
- Remedy - remove noisy or insignificant dimensions
- Approaches: principal component analysis (PCA), autoencoders (Neural network)

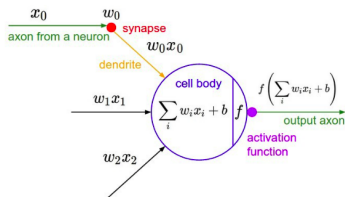
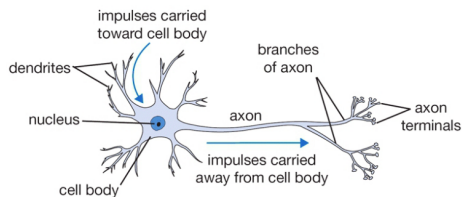


11

¹¹<https://cofactorgenomics.com/curse-of-dimensionality-wk-16/>

Artificial neural networks

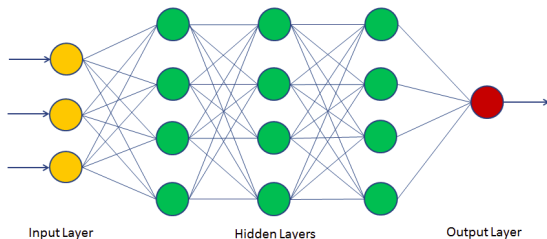
- Inspired by biological neurons
- Dendrites and axons carry signals
- In an artificial neural network, neural network edges carry data to and from neurons



12

Artificial neural networks

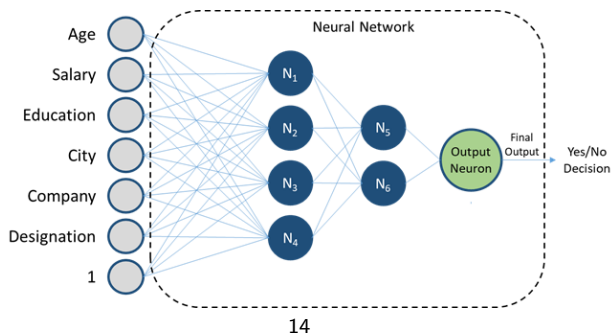
- Neural network has an architecture
- Layers - input, hidden, output,
- Loss function - mean squared error, cross-entropy loss.
- Optimiser - adam, adadelata, rmsprop ...
- Types of layers - recurrent, dropout, convolutional ...
- Types of activations - tanh, sigmoid, softmax ...



13

Artificial neural networks

- Input layer - receive data
- Number of neurons = number of features
- Hidden layer - number of neurons or layers not fixed, depends on the problem being solved. Responsible for learning complex patterns
- Output layer - compute output as a class or real number



¹⁴<https://www.datacamp.com/community/tutorials/neural-network-models-r>

General recommendations for using ML algorithms

- Preprocess datasets - outliers, incorrect labels, standardise features by scaling, encoding, imputing missing values
- Split datasets - train, test, validation and K-fold cross-validation
- Use right algorithm - start with simple and then move to complex
- Fix data imbalance
- Tune hyperparameters
- Look for overfitting
- Evaluate accuracy for each class (for classification)

Machine learning in Galaxy

Galaxy Europe

- Online platform for numerous (>2000) scientific tools running on large compute resources including GPUs as well as large storage
- Accelerates scientific, especially bioinformatics research
- Public infrastructure
- Open-source community, contributors across the globe
- Over 200 tutorials (hands-on materials) showing usage of tools in different scientific analyses ¹⁵

¹⁵<https://training.galaxyproject.org/>

COVID-19 Research

Want to learn the best practices for the analysis of SARS-CoV-2 data using Galaxy? Visit the [Galaxy SARS-CoV-2 portal](#). We mirror all public SARS-CoV-2 data from ENA in a [Galaxy data library](#) for your convenience. The Galaxy community has created [COVID-19 dedicated training materials](#). Please check our [recent activities](#) for more details. If you need help submitting your data to public archives, like ENA, please [get in touch](#). We will support you in sharing your data.

"Anyone, anywhere in the world should have free, unhindered access to not just my research, but to the research of every great and enquiring mind across the spectrum of human understanding." – Prof. Stephen Hawking

News

- Oct 23, 2021
UseGalaxy.eu Tool Updates for 2021-10-23
- Oct 18, 2021
Training Infrastructure Feedback from Dr. Theodora Tsirka
- Oct 16, 2021
UseGalaxy.eu Tool Updates for 2021-10-16
- Oct 13, 2021
BY-COVID: A new EU project for pandemic preparedness
- Oct 12, 2021
UseGalaxy.eu Use Case: cellular specification, differentiation and morphogenesis of the mucociliary epithelium
- Oct 11, 2021
UseGalaxy.eu Use Case: microRNAs in heart disease

Events

- Oct 28, 2021
Galaxy Developer Roundtable: Image analysis in Galaxy - pain points and lessons learnt
- Nov 1, 2021
Single-Cell RNAseq Training Course 2021
- Nov 4, 2021
Genome Annotation & Galaxy Large Data Handling workshop
- Nov 8, 2021 - Nov 12, 2021
ELIXIR BioHackathon Europe
- Nov 9, 2021 - Nov 10, 2021
Protein-ligand docking training for the Galaxy India community
- Nov 16, 2021 - Nov 17, 2021
5. NRZ-Authent Expertinnen- und Expertenworkshop

Currently Running and Queued Jobs

UseGalaxy.eu: The European Galaxy instance

OPEN CHAT

History

search datasets

Clade assignment

68 shown, 4 deleted
36.3 GB

- 72: Nextclade on data 45 (FASTA alignment)
- 71: Nextclade on data 45 (Auspice v2 tree)
- 70: Nextclade on data 45 (JSON report)
- 69: Nextclade on data 45 (TSV report)
- 68: Nextclade on data 34 (FASTA alignment)
- 67: Nextclade on data 34 (Auspice v2 tree)
- 66: Nextclade on data 34 (JSON report)
- 65: Nextclade on data 34 (TSV report)
- 64: radiopaedia.org_covid-19-pneumonia-7_857_03_0-dcm.nii
- 63: Nextclade on data 25 (FASTA alignment)
- 62: Nextclade on data 25 (Auspice v2 tree)
- 61: Nextclade on data

16

RESEARCH ARTICLE

Galaxy-ML: An accessible, reproducible, and scalable machine learning toolkit for biomedicine

Qiang Gu^{1,2}, Anup Kumar³, Simon Bray³, Allison Creason^{1,2},
Alireza Khanteymoori³, Vahid Jalili^{1,2}, Björn Grüning³, Jeremy Goecks^{1,2*}

1 Department of Biomedical Engineering, Oregon Health & Science University, Portland, Oregon, United States of America, **2** The Knight Cancer Institute, Oregon Health & Science University, Portland, Oregon, United States of America, **3** Bioinformatics Group, Department of Computer Science, University of Freiburg, Freiburg, Germany

* goecksj@ohsu.edu

**OPEN ACCESS**

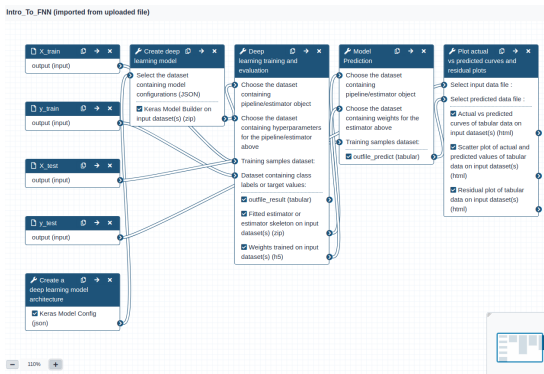
Citation: Gu Q, Kumar A, Bray S, Creason A, Khanteymoori A, Jalili V, et al. (2021) Galaxy-ML: An accessible, reproducible, and scalable machine learning toolkit for biomedicine. *PLoS Comput Biol* 17(6): e1009014. <https://doi.org/10.1371/journal.pcbi.1009014>

Abstract

Supervised machine learning is an essential but difficult to use approach in biomedical data analysis. The Galaxy-ML toolkit (<https://galaxyproject.org/community/machine-learning/>) makes supervised machine learning more accessible to biomedical scientists by enabling them to perform end-to-end reproducible machine learning analyses at large scale using only a web browser. Galaxy-ML extends Galaxy (<https://galaxyproject.org>), a biomedical computational workbench used by tens of thousands of scientists across the world, with a suite of tools for all aspects of supervised machine learning.

ML in Galaxy

- 20 - 30 ML tools powered by scikit-learn and tensorflow
- ML tools - classifiers, regressors, data preprocessors, visualizations, hyperparameter tuners, pipeline builders
- Workflow of tools
- Long running ML training on Galaxy infrastructure (using multiple CPUs, GPUs)



ML tutorials in Galaxy

Statistics and machine learning

Statistical Analyses for omics data and machine learning using Galaxy tools

Requirements

Before diving into this topic, we recommend you to have a look at:

- [Introduction to Galaxy Analyses](#)

Material

Q x

Lesson	Slides	Hands-on	Input dataset	Workflows	Galaxy tour	Galaxy instances
Age prediction using machine learning		-				
Basics of machine learning		-				
Classification in Machine Learning		-				
Clustering in Machine Learning		-				
Deep Learning (Part 1) - Feedforward neural networks (FNN)		-				
Deep Learning (Part 2) - Recurrent neural networks (RNN)		-				
Deep Learning (Part 3) - Convolutional neural networks (CNN)		-				
Interval-Wise Testing for omics data		-				
Introduction to deep learning		-				
Introduction to Machine Learning using R interactive tools		-				
Machine learning: classification and regression		-				
PAP1A/PDK1_OG: PanCancer Aberrant Pathway Activity Analysis Machine learning Pan-cancer Cancer biomarkers Chromines and tumor suppressor genes		-				
Regression in Machine Learning		-				
Text-mining with the SinText toolset interactive tools		-				

Our projects with machine learning

Galaxy Jupyterlab editor for ML

- Jupyter notebook - popular editor
- Scientific computing, data science, machine learning, learn to code ...
- Simple and fast way to create prototypes
- No need for any package installation
- Easy to share any analysis
- Runs on web

The screenshot displays the Galaxy Jupyterlab interface. At the top, there is a menu bar with options: File, Edit, View, Run, Kernel, Git, Tabs, Settings, Help. On the right, there are status indicators for Log Out, CPU usage (60%), and Memory usage (440 / 32768 MB). Below the menu is a file browser on the left with a search bar and a list of files: /, data (2 minutes ago), elyra, notebooks (2 minutes ago), outputs, and home_page.ipynb (2 minutes ago). The main area shows a Jupyter notebook titled 'home_page.ipynb'. The notebook content includes:

- Welcome to the Galaxy's GPU enabled Interactive Jupyterlab for Artificial intelligence (AI).**
- A paragraph: "Jupyter notebook is powered by the latest JupyterLab, Tensorflow, Scikit-learn, Pandas, Numpy, Scipy, Seaborn, Matplotlib and many more which can be used to prototype and develop machine learning and deep learning solutions executing on Galaxy's NVIDIA GPUs. The docker container used can be found at [docker image](#)."
- Core features**
 - Run AI programs on **GPUs**
 - Pre-installed** packages for AI
 - Integrated with **Git** version control
 - Elyra AI workflow** of notebooks
 - Shareable** AI models via Open Neural Network Exchange ([ONNX](#))
 - Run **Galaxy tools** using [BioBlend APIs](#)
- Introduction**

Jupyterlab notebooks are extremely popular with data scientists and researchers to explore datasets from multiple fields of studies and develop **prototypes**. The notebooks come integrated with a lots of packages such as NumPy, Statsmodel, Pandas, Scikit-learn, Tensorflow, Matplotlib which expedite prototyping and provide useful insights into the datasets. In a notebook, each rectangular box is known as a **cell** which executes **Python code** written in it.

```
[2]: a = 5
     b = 10
     c = a + b
```

Execute the above cell by clicking on "CTRL+Enter". Writing a variable such as "c" in a cell and executing it prints the output of that variable.

```
[3]: c
```

```
[3]: 16
```
- Working with GPUs**

GPU computation is configured in this notebook via the docker image which allows deep learning programs runs much faster.
- Data Science**

The field of data science incorporates multiple tasks to understand a dataset such as **plotting** to visualize different features, **data manipulation** to impute missing values, finding

At the bottom of the notebook, there are status indicators: Simple, 0, 1, Python 3 | idle, Fully initialized, Mem: 440.40 / 32768.00 MB, Mode: Command, Ln 1, Col 1, home_page.ipynb.

²⁰https://live.usegalaxy.eu/?tool_id=interactive_tool_ml_jupyter_notebook

Features of Galaxy Jupyterlab

- Base container - jupyter/tensorflow-notebook:latest ²¹
- CUDA and cuDNN packages for nvidia GPUs, tensorflow for GPU, pre-installed ML and DL packages
- Create, share and reuse ML/DL models - ONNX ²²
- Git integration
- Workflow of notebooks - Elyra AI ²³
- Connect to Galaxy histories, datasets using bioblend ²⁴
- Miscellaneous - dashboards for CPU, GPU, memory utilization, collapse/expand sections, notebook as voila ...
- Docker image ²⁵

²¹<https://hub.docker.com/r/jupyter/tensorflow-notebook/>

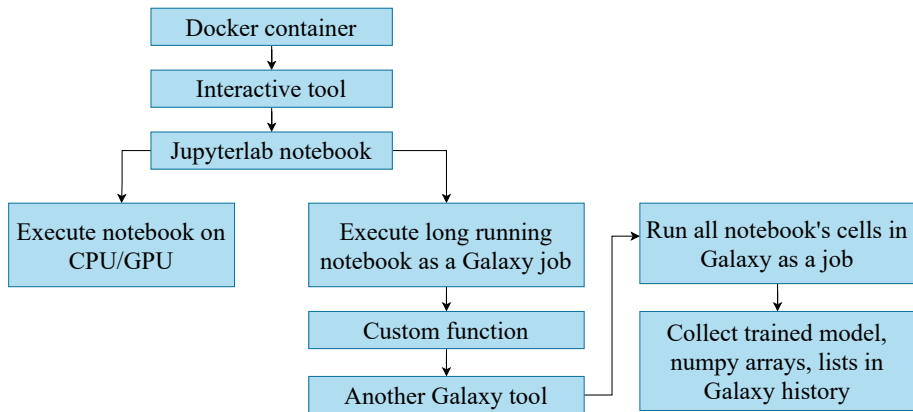
²²<https://onnx.ai/>

²³<https://github.com/elyra-ai/elyra>

²⁴<https://bioblend.readthedocs.io/en/latest/>

²⁵<https://github.com/anupruez/ml-jupyter-notebook>

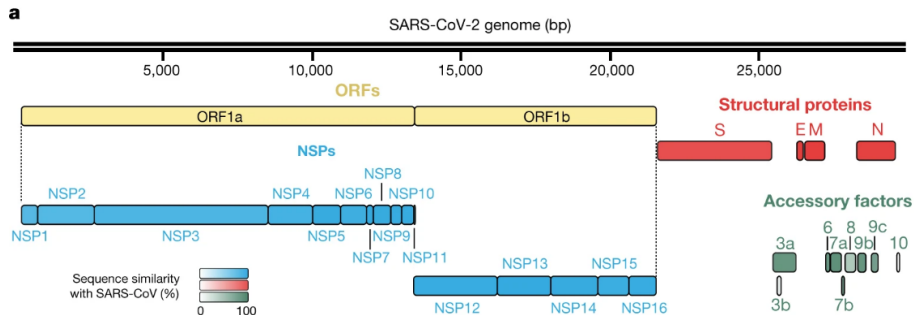
Architecture of Galaxy Jupyterlab



Prediction of protein evolution (amino acid substitutions) in SARS-COV2 sequences

- Spike protein and amino acid (AA) mutations
- Nextclade clades
- Sequence to sequence learning
- Generative adversarial networks (GANs)
- Comparison of true and generated SARS-CoV-2 AA sequences
- Substitutions from generated AA sequences (future substitutions)?

Spike protein (S)

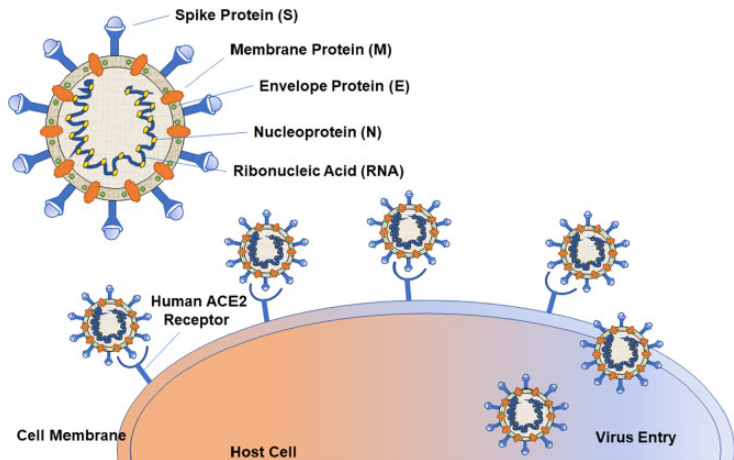


26

- Non-structural and structural proteins

²⁶<https://www.nature.com/articles/s41586-020-2286-9>

Spike protein (S)



27

Spike protein (S)

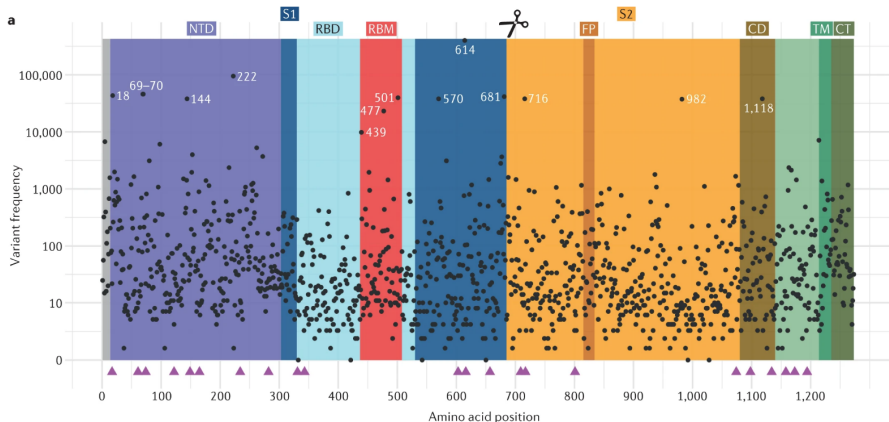
- Binds to the host cell
- Mutations may impact infectivity, transmissibility
- D614G: enhances viral replication ²⁸
- N439K: enhances the binding affinity for the ACE2 receptor and reduces the neutralizing activity of antibodies ²⁹
- Y453F: increased ACE2-binding affinity ³⁰
- ...

²⁸<https://www.nature.com/articles/s41586-020-2895-3>, <https://covariants.org/variants/20B.S.732A>

²⁹<https://www.nature.com/articles/s41579-021-00573-0>

³⁰<https://www.nature.com/articles/s41579-021-00573-0>

Frequency of spike mutations (substitutions and deletions)

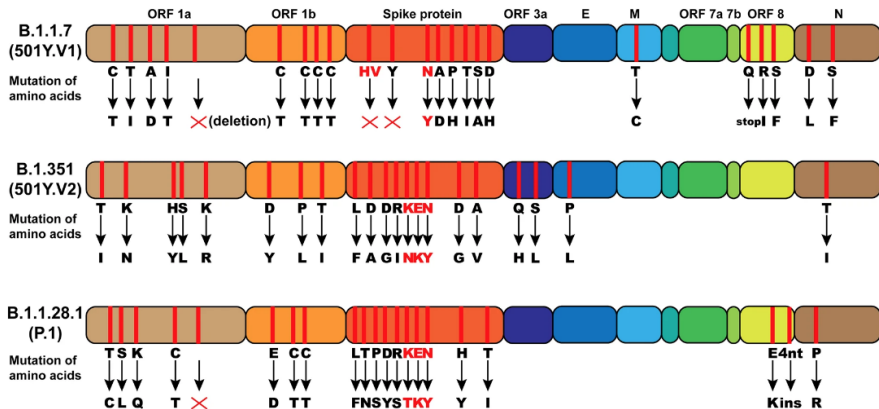


31

- 426,623 genomes, 5106 substitutions

³¹<https://www.nature.com/articles/s41579-021-00573-0>

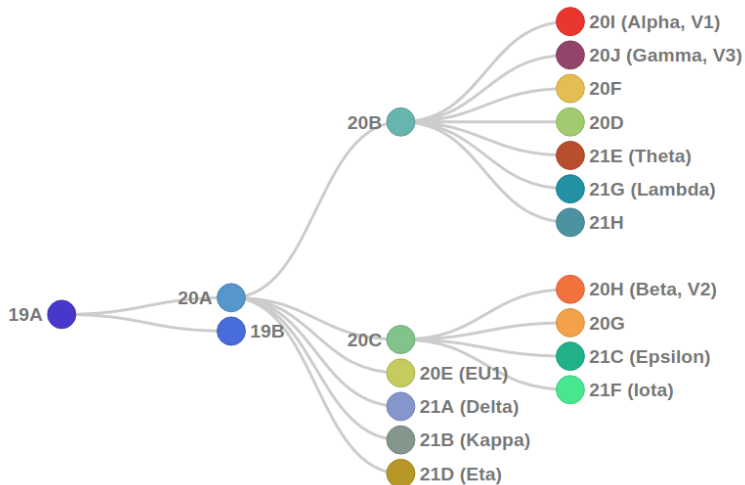
Spike mutations in lineages



32

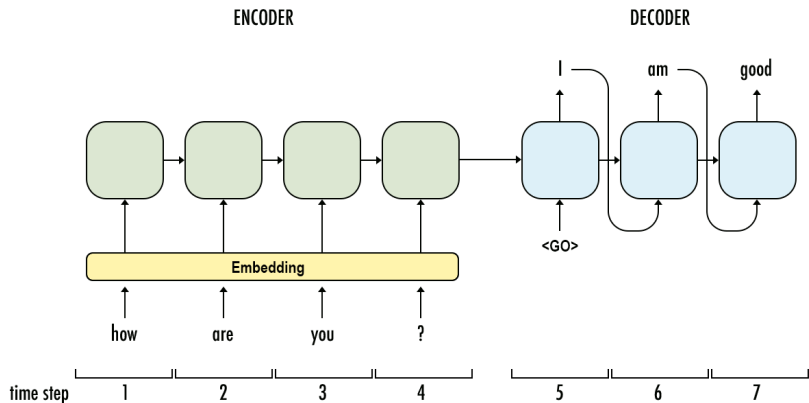
³²<https://www.nature.com/articles/s41392-021-00644-x>

Nextclade clades



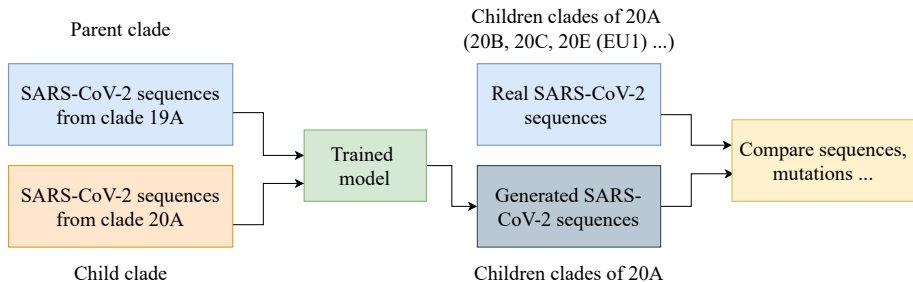
33

Sequence to sequence learning



34

Sequence to sequence learning with SARS-CoV-2 sequences



Generative Adversarial Networks (GANs)

- Generator - generates data (sequences)
- Generator network - sequence to sequence encoder-decoder network
- Discriminator - discriminates between real and generated data (sequences)
- Discriminator network - sequential network to predict either real (true) or generated (false)
- Generator and Discriminator - make each other better over training iteration
- Applications - improve astronomical images ³⁵, reconstruct 3D model of object from images ³⁶, age face photographs ³⁷, language translation³⁸, ...

³⁵<https://arxiv.org/pdf/1702.00403.pdf>

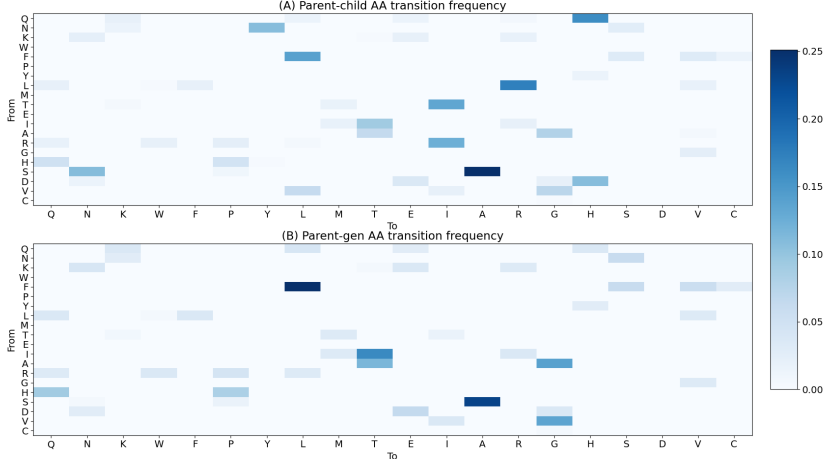
³⁶<http://3dgan.csail.mit.edu/>

³⁷<https://arxiv.org/pdf/1702.01983.pdf>

³⁸<https://arxiv.org/pdf/1704.06933.pdf>

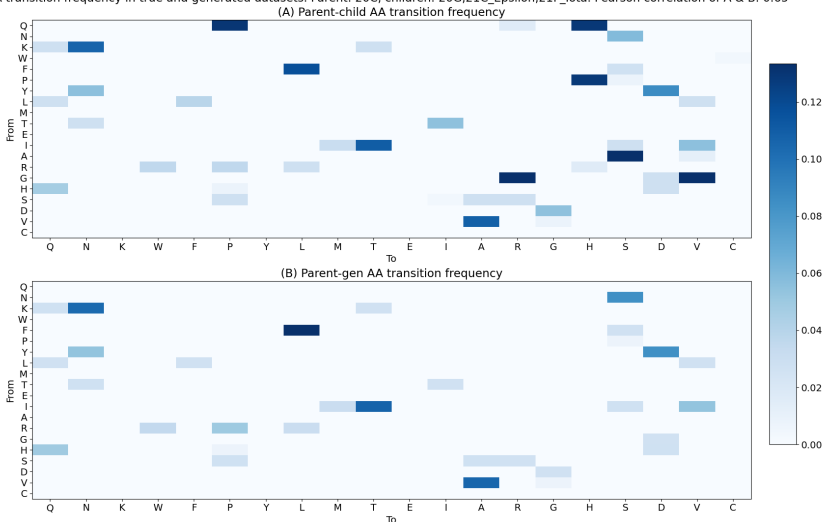
Prediction of protein evolution (for children clades of 20B)

AA transition frequency in true and generated datasets. Parent: 20B, children: 20I_Alpha,20F,20D,21G_Lambda,21H. Pearson correlation of A & B: 0.68



Prediction of protein evolution (for children clades of 20C)

AA transition frequency in true and generated datasets. Parent: 20C, children: 20G,21C_Epsilon,21F_Iota. Pearson correlation of A & B: 0.65



Thank you for your attention. Questions?