Welcome to the Seminar!

www.biocomicals.com

Overview of Bioinformatics

Seminar “Informatics in Biochemistry”

By Lukas Jarosch and Leonhard Kohleick

Question Round

In what semester are you?

What is your name?

Let’s proceed with www.menti.com

What even is “Bioinformatics”?

Definition of Bioinformatics

> “*Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data*
>
> *- Wikipedia*”

Research fields in Bioinformatics

- **Genomics (DNA)**
  - genotype-phenotype patterns
  - protein homology
  - ancestral relationships

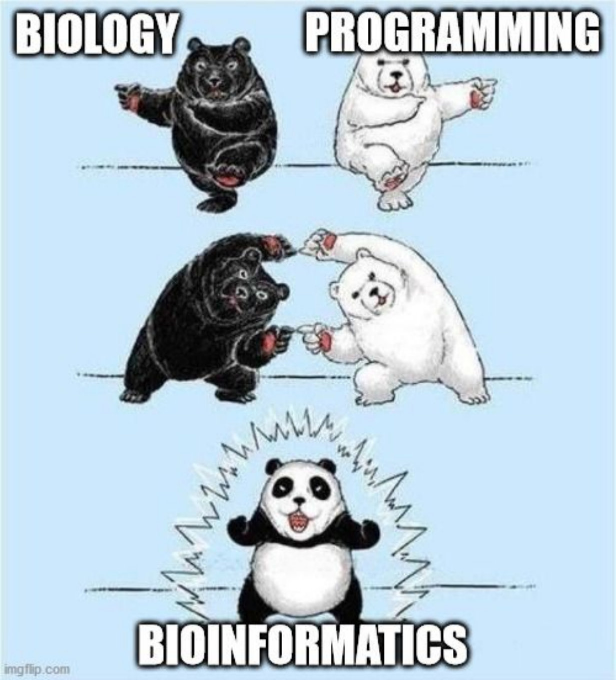
- **Transcriptomics (RNA)**
  - identifying expressed genes
  - understanding disease mechanisms
  - gene regulatory relationships

- **Proteomics (proteins)**
  - protein-protein interactions
  - post-translational modifications
  - biomarkers

- **Structural modeling**
  - protein structure prediction
  - simulating protein function
  - drug identification

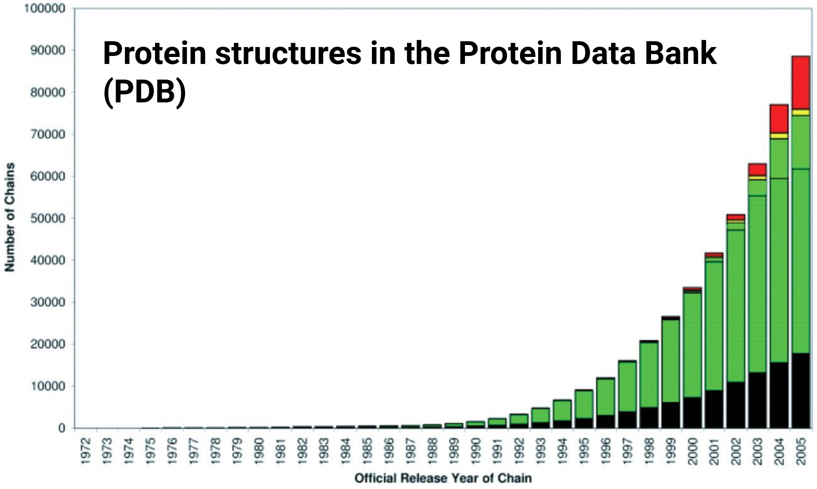
- **Image Analysis, Systems Biology, ...**

What skills does a Bioinformatician need?



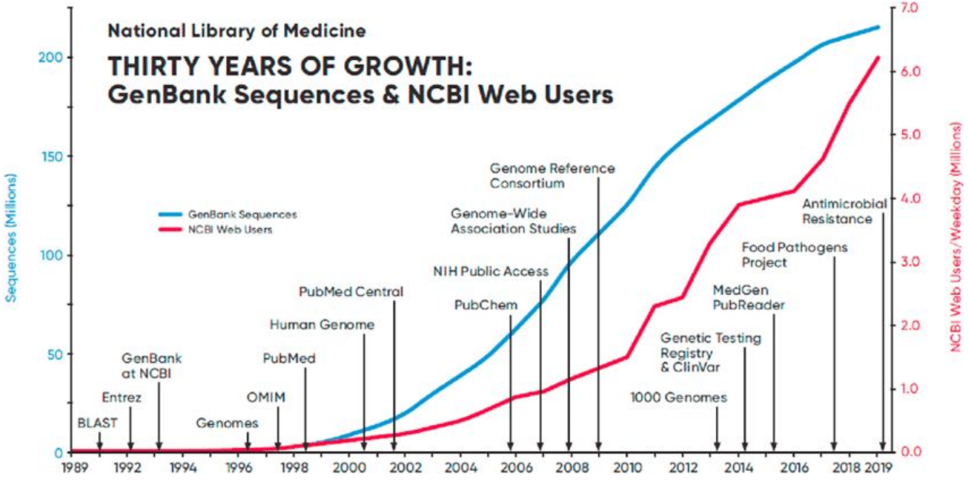
- Programming (mostly Python & R)
- Statistics
- Biology
- Physics

Why should you consider learning Bioinformatics?



Protein structures in the Protein Data Bank (PDB)

Greene et al., 2017



National Library of Medicine

**THIRTY YEARS OF GROWTH:
GenBank Sequences & NCBI Web Users**

https://www.nlm.nih.gov/about/2021CJ_NLM.pdf

Why should you consider learning Bioinformatics?

- Current trend towards **high-throughput technologies** that generate massive amounts of data
- Lots of exciting **new research fields**
- **PhD applications** increasingly require coding skills
- Coding is (mostly) **reproducible**
- Programming can **automate** a lot of tedious tasks

Sequence Alignments & Genomics

Sequence alignment

Protein 1

```
KINLKVIKNTLLFRAL
```

Protein 2

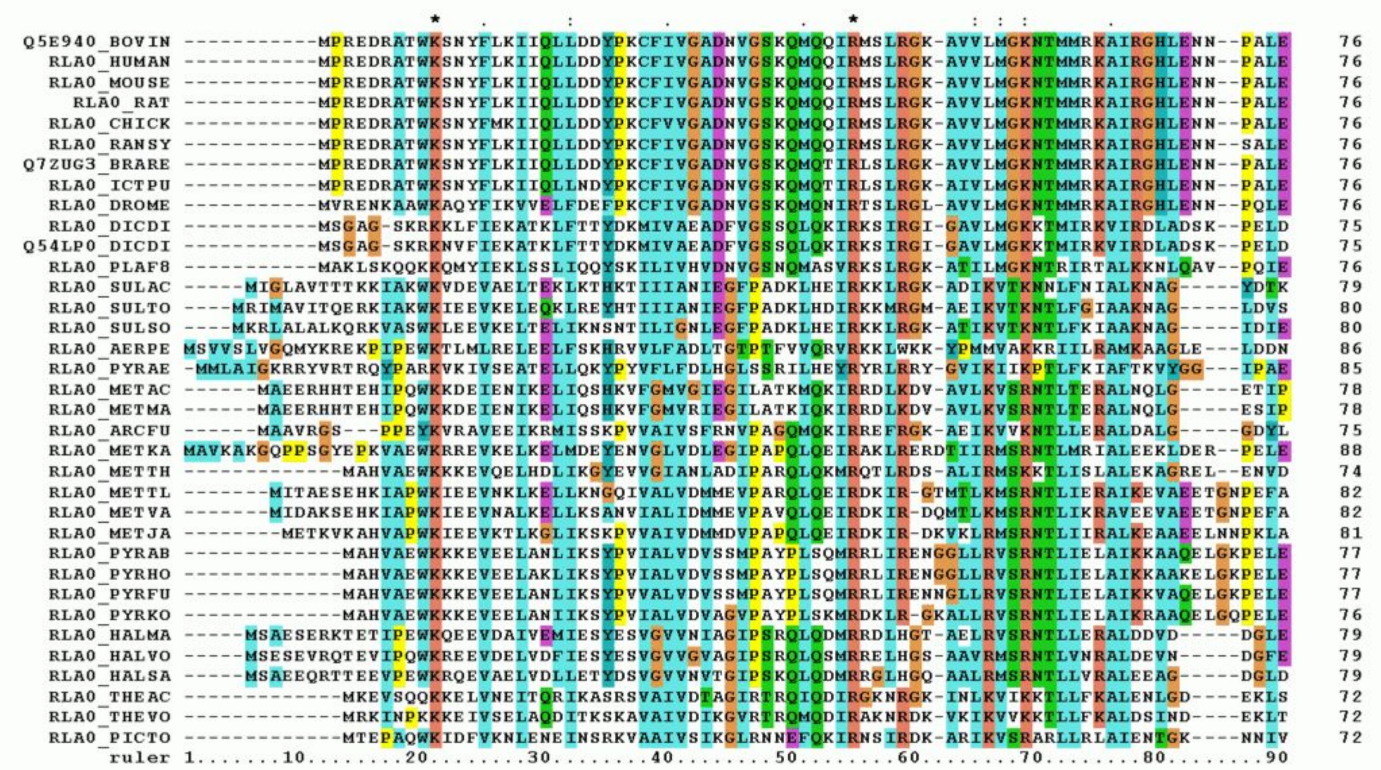
```
GKALLRVRNTLIELAI
```

Sequence alignment

Aligned sequences

```
K-INLKVIKNTLLFRAL
GKALLRV-RNTLIELAI
```

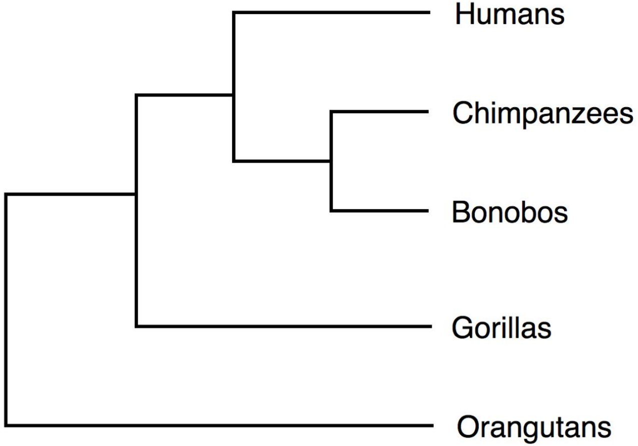
Multiple sequence alignment

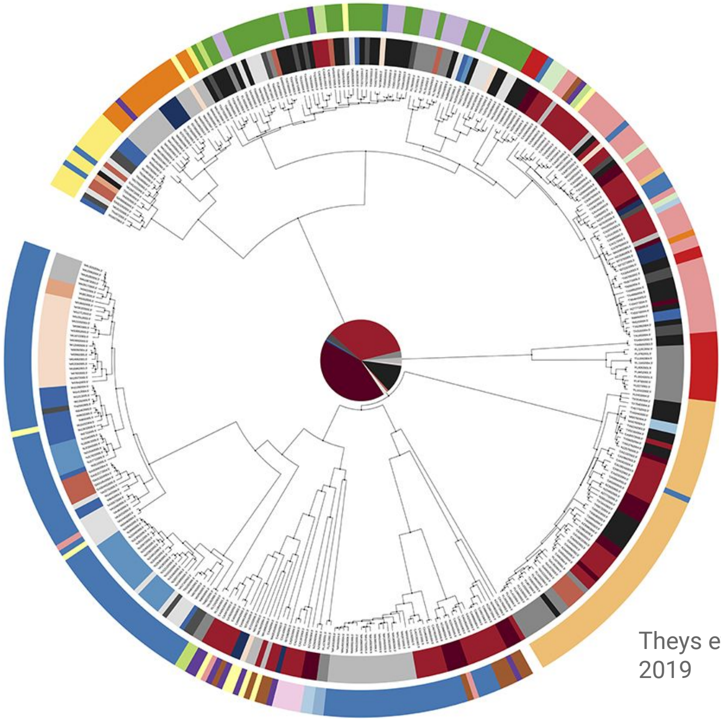


```
                                 *   .      :       .          .   *         : : :      .
Q5E940_BOVIN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
 RLA0_HUMAN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
 RLA0_MOUSE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
   RLA0_RAT -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
 RLA0_CHICK -----------MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
 RLA0_RANSY -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE   76
Q7ZUG3_BRARE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
 RLA0_ICTPU -----------MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE   76
 RLA0_DROME -----------MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE   76
 RLA0_DICDI -----------MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD   75
Q54LP0_DICDI -----------MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD   75
 RLA0_PLAF8 -----------MAKLSKQQKKQMYIEKLSSLIQQYSKILIVHVDNVGSNQMASVRKSLRGK-ATILMGKNTRIRTALKKNLQAV--PQIE   76
 RLA0_SULAC -----MIGLAVTTTKKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAG-----YDTK   79
 RLA0_SULTO ----MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG-----LDVS   80
 RLA0_SULSO ----MKRLALALKQRKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAG-----IDIE   80
 RLA0_AERPE MSVVSLVGQMYKREKPIPEWKTLMLRELEELFSKHRVVLFADLTGTPTFVVQRVRKKLWKK-YPMMVAKKRIILRAMKAAGLE---LDDN   86
 RLA0_PYRAE -MMLAIGKRRYVRTRQYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRRY-GVIKIIKPTLFKIAFTKVYGG---IPAE   85
 RLA0_METAC ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVGIEGILATKMQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ETIP   78
 RLA0_METMA ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKIQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ESIP   78
 RLA0_ARCFU ------MAAVRGS---PPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGQMQKIRREFRGK-AEIKVVKNTLLERALDALG-----GDYL   75
 RLA0_METKA MAVKAKGQPPSGYEPKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPQLQEIRAKLRERDTIIRMSRNTLMRIALEEKLDER--PELE   88
 RLA0_METTH -------------MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLISLALEKAGREL--ENVD   74
 RLA0_METTL -------MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPARQLQEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA   82
 RLA0_METVA -------MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVQLQEIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA   82
 RLA0_METJA --------METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIRDKIR-DKVKLRMSRNTLIIRALKEAAEELNNPKLA   81
 RLA0_PYRAB -------------MAHVAEWKKKEVEELANLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAQELGKPELE   77
 RLA0_PYRHO -------------MAHVAEWKKKEVEELAKLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAKELGKPELE   77
 RLA0_PYRFU -------------MAHVAEWKKKEVEELANLIKSYPVVALVDVSSMPAYPLSQMRRLIRENNGLLRVSRNTLIELAIKKVAQELGKPELE   77
 RLA0_PYRKO -------------MAHVAEWKKKEVEELANIIKSYPVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSRNTLIELAIKRAAQELGQPELE   76
 RLA0_HALMA -----MSAESERKTETIPEWKQEEVDAIVEMIESYESVGVVNIAGIPSRQLQDMRRDLHGT-AELRVSRNTLLERALDDVD-----DGLE   79
 RLA0_HALVO -----MSESEVRQTEVIPQWKREEVDELVDFIESYESVGVVGVAGIPSRQLQSMRRELHGS-AAVRMSRNTLVNRALDEVN-----DGFE   79
 RLA0_HALSA -----MSAEEQRTTEEVPEWKRQEVAELVDLLETYDSVGVVNVTGIPSKQLQDMRRGLHGQ-AALRMSRNTLLVRALEEAG-----DGLD   79
 RLA0_THEAC -------------MKEVSQQKKELVNEITQRIKASRSVAIVDTAGIRTRQIQDIRGKNRGK-INLKVIKKTLLFKALENLGD----EKLS   72
 RLA0_THEVO -------------MRKINPKKKEIVSELAQDITKSKAVAIVDIKGVRTRQMQDIRAKNRDK-VKIKVVKKTLLFKALDSIND----EKLT   72
 RLA0_PICTO -------------MTEPAQWKIDFVKNLENEINSRKVAAIVSIKGLRNNEFQKIRNSIRDK-ARIKVSRARLLRLAIENTGK----NNIV   72
      ruler 1.......10........20........30........40........50........60........70........80........90
```

https://upload.wikimedia.org/wikipedia/commons/thumb/7/79/RPLP0_90_ClustalW_aln.gif/1024px-RPLP0_90_ClustalW_aln.gif

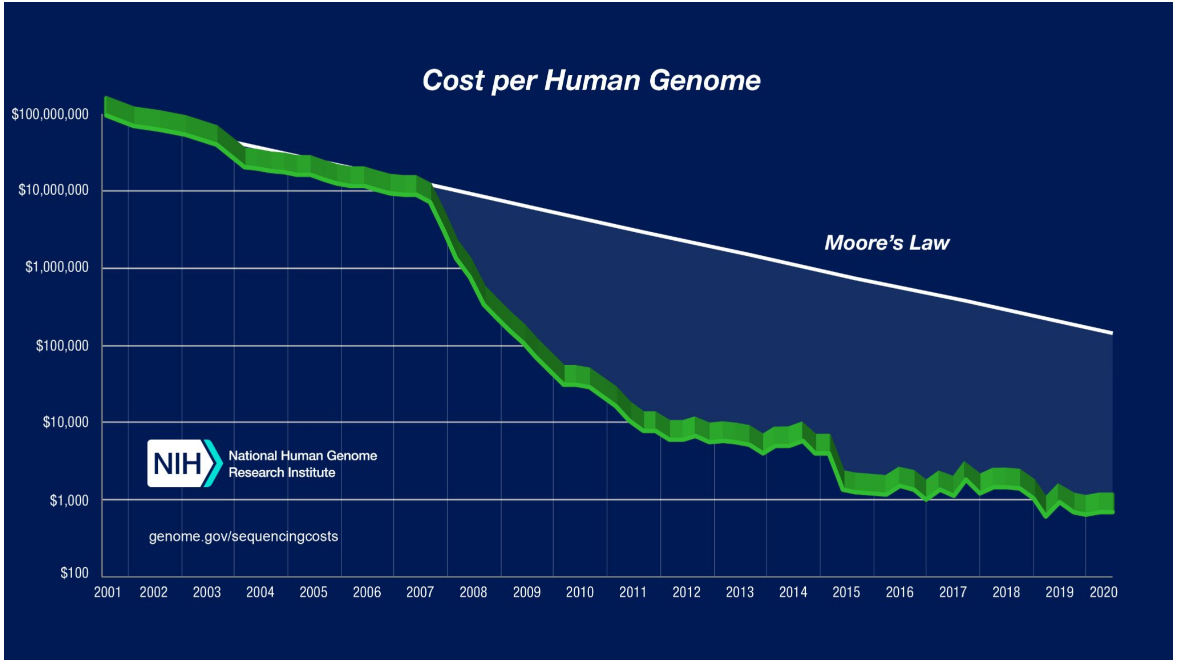
Phylogenetics

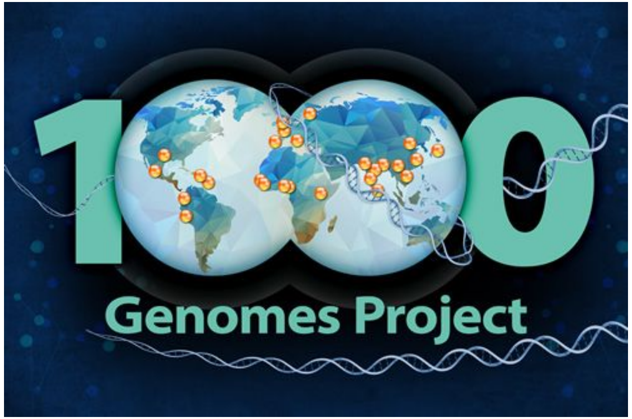




Theys et al., 2019

Large-scale genome sequencing

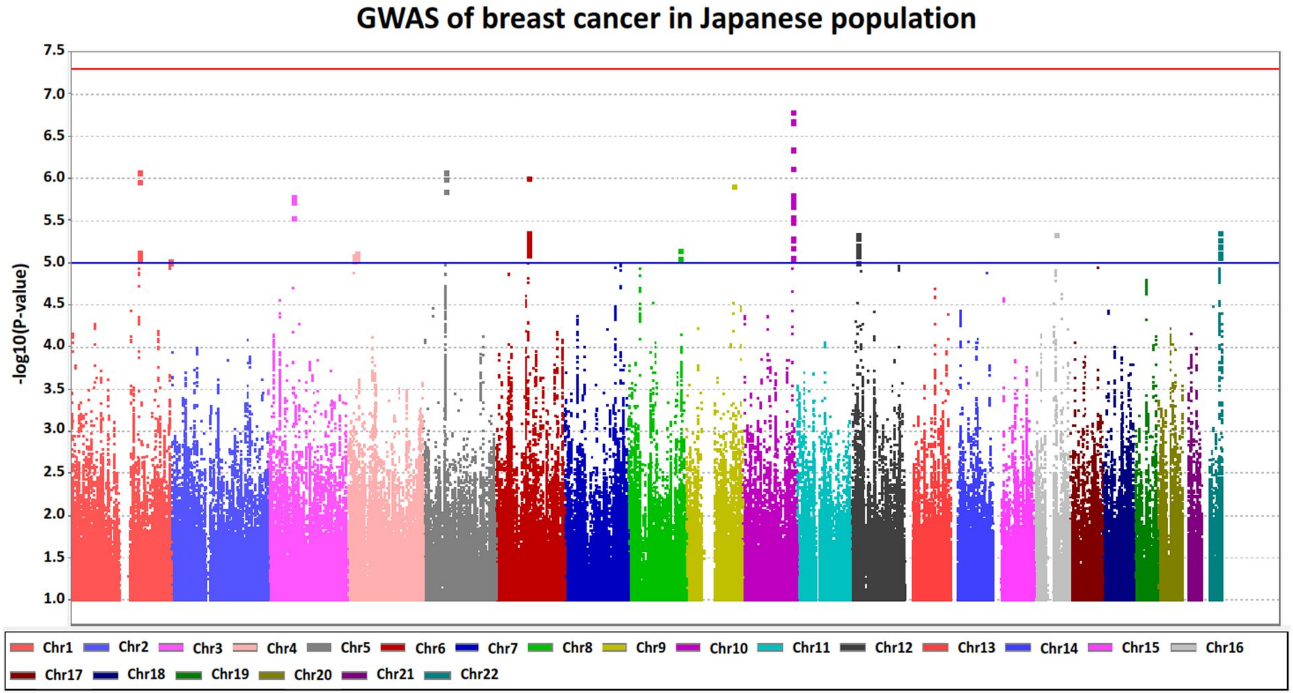




http://www.genome.gov/sites/default/files/genome-old/images/content/1000genomes.jpg

https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Genome-wide association studies (GWAS)

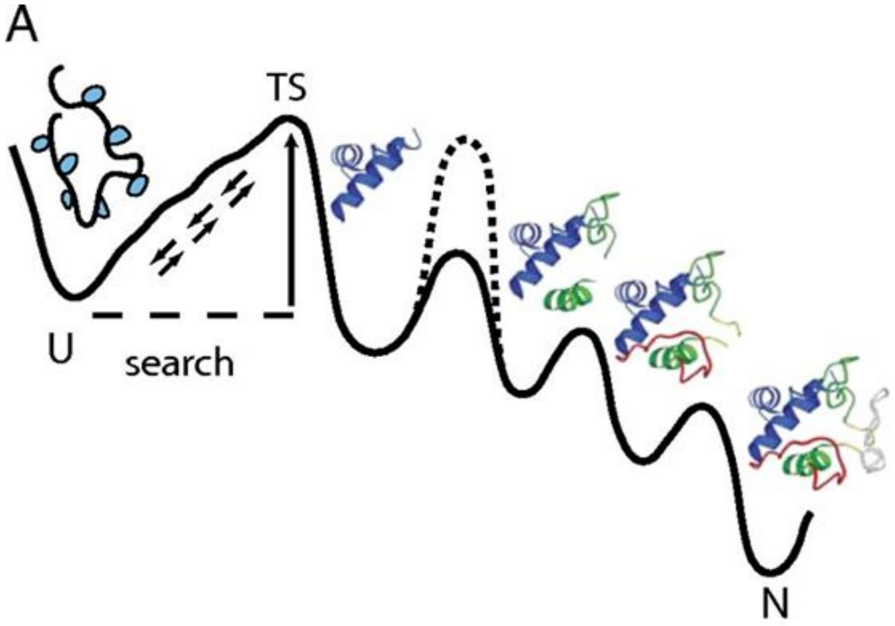


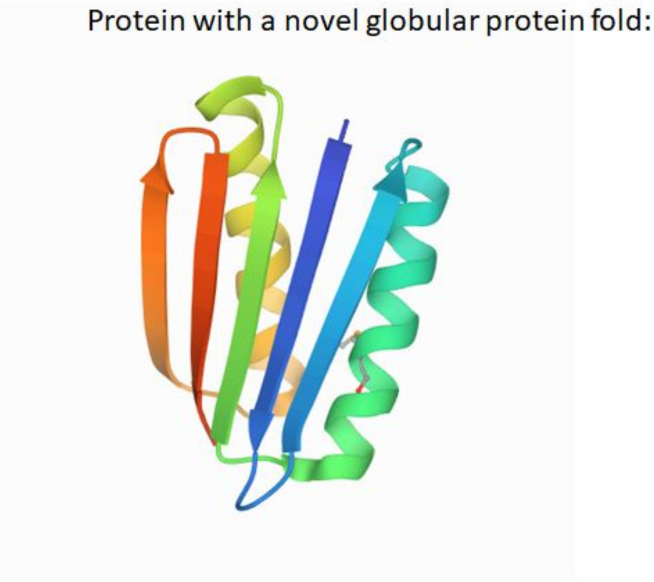
Low et al., 2013

Modeling

Protein engineering with Rosetta

Rosetta can predict protein folding and also has various design features



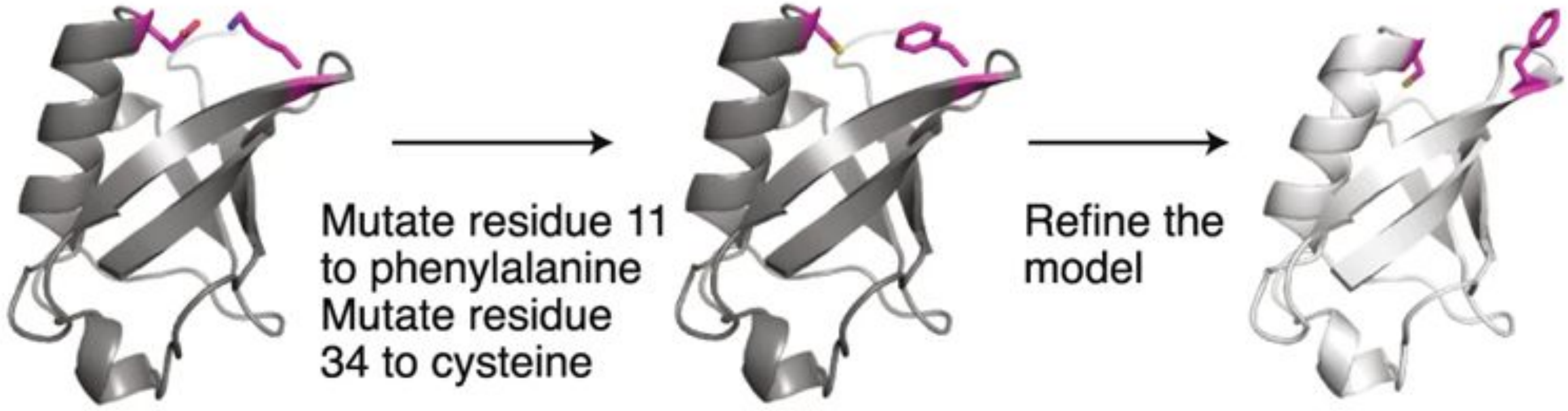

Protein with a novel globular protein fold:

TOP7

PDB-ID: 1QYS

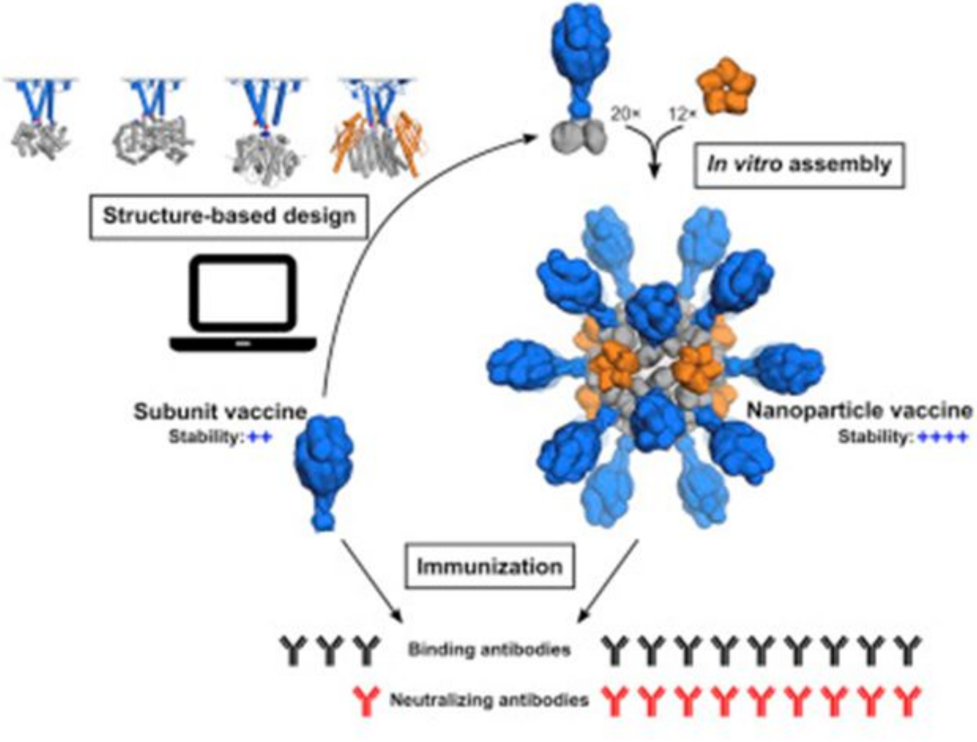
(Englander and Mayne, 2014) & (Kuhlman et al., 2003)

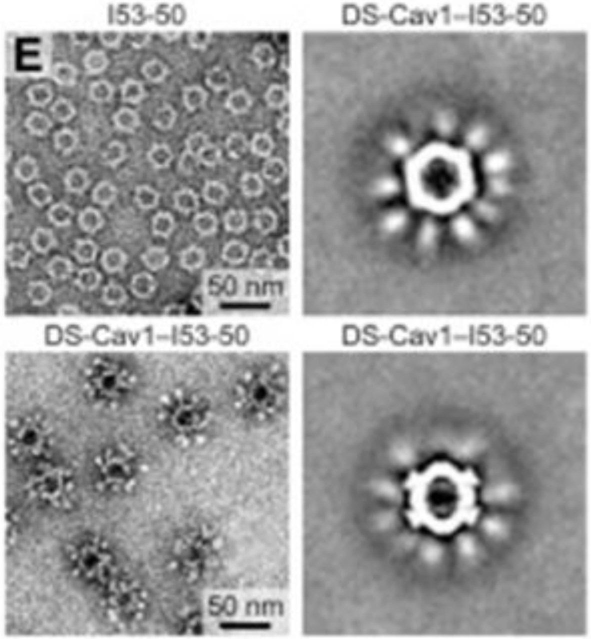
Designing a protein with Rosetta



(Leman et al., 2020)

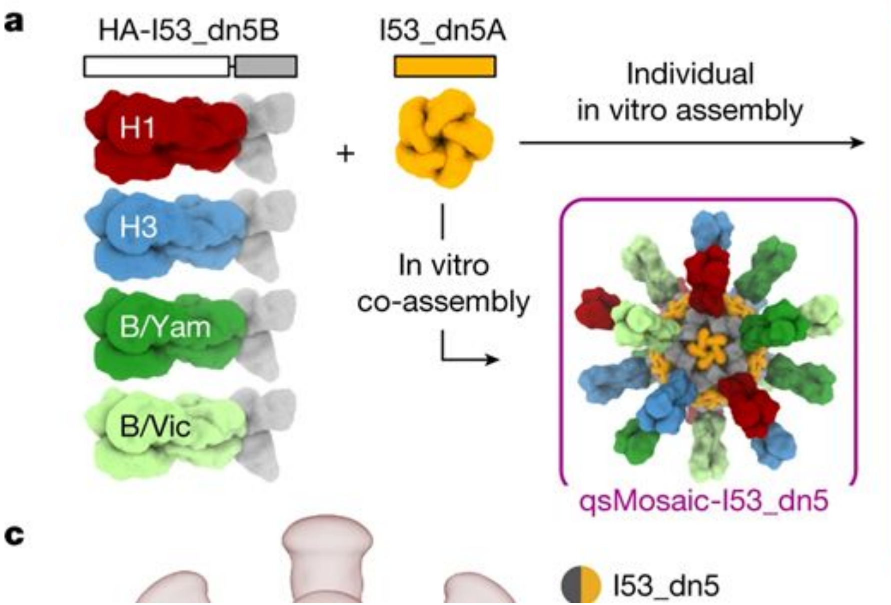
Vaccines designed with Rosetta (for RSV)

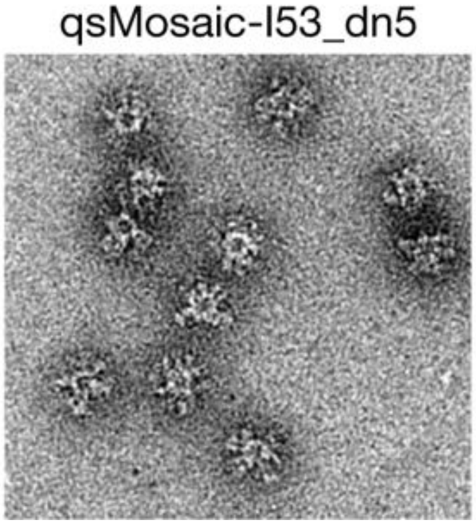




(Marcandalli et al., 2019)

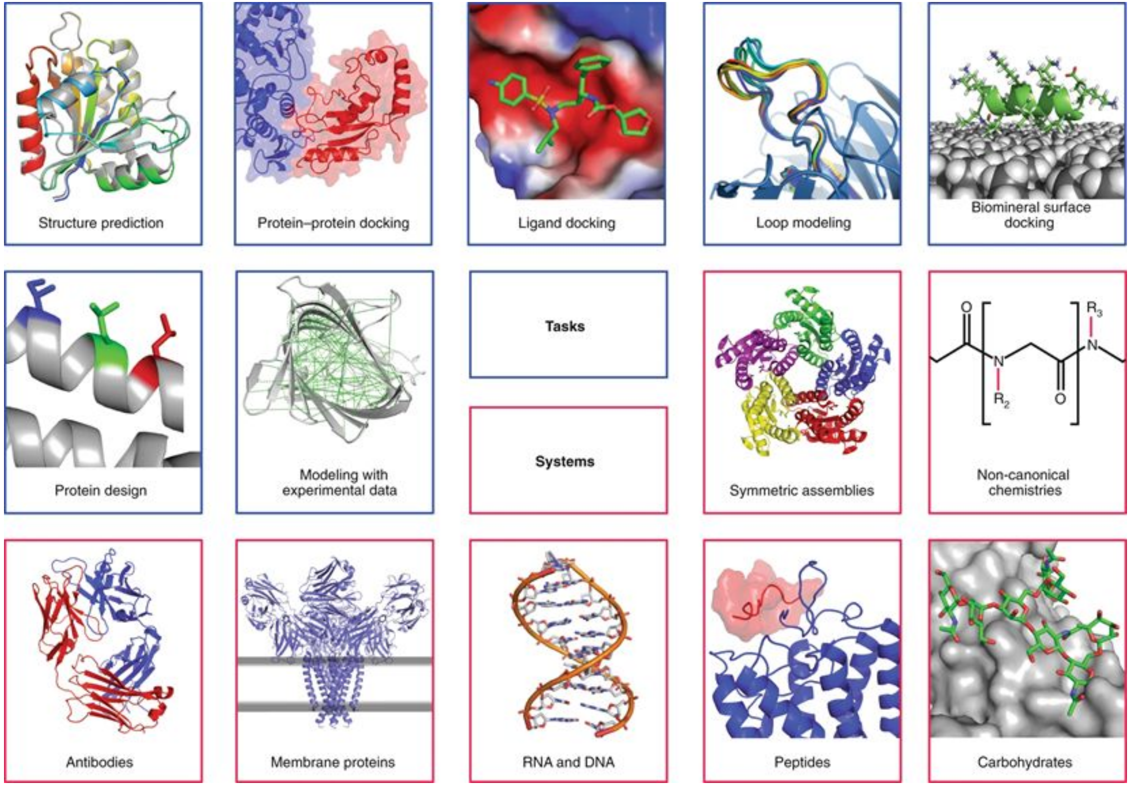
A new quadrivalent vaccine against influenza





(Boyoglu-Barnum et al., 2021)

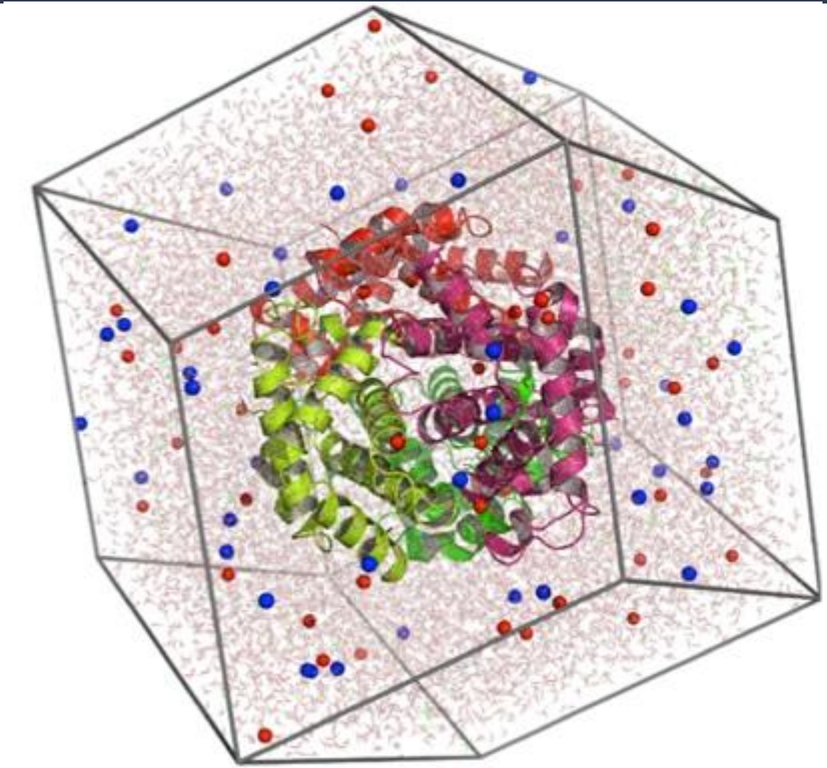
Rosetta has a lot of capabilities



(Leman et al., 2020)

Molecular Dynamics

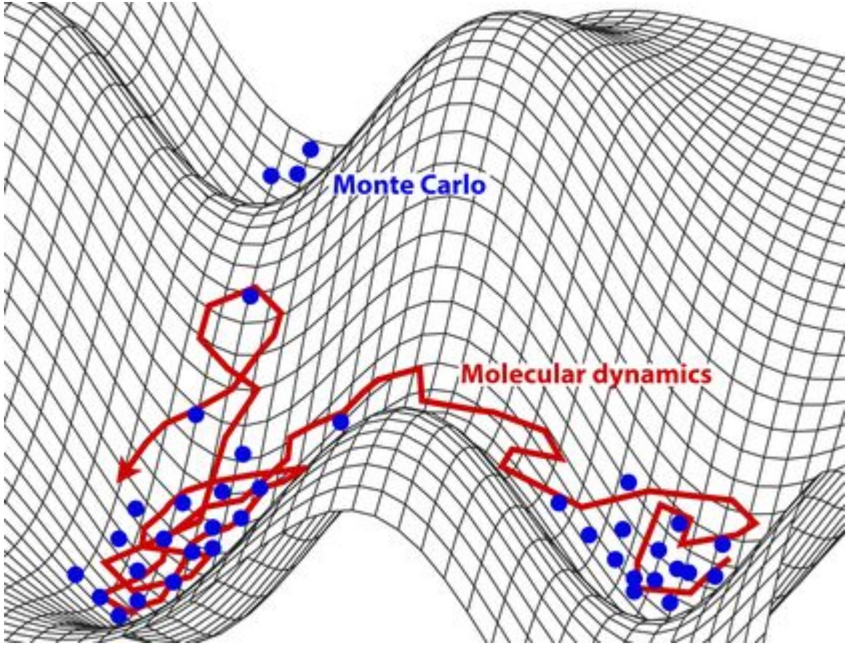
Studying physical movements of atoms and molecules



Hemoglobin

PLoS Comput Biol 2010,6(5): e1000774

Molecular dynamics searches for the lowest energy conformation



https://commons.wikimedia.org/wiki/File:Sampling_in_Monte_Carlo_and_molecular_dynamics.png

Use of Molecular Dynamics in Biochemistry

- Using Newton's laws of motion, we predict the position of each atom as a function over time

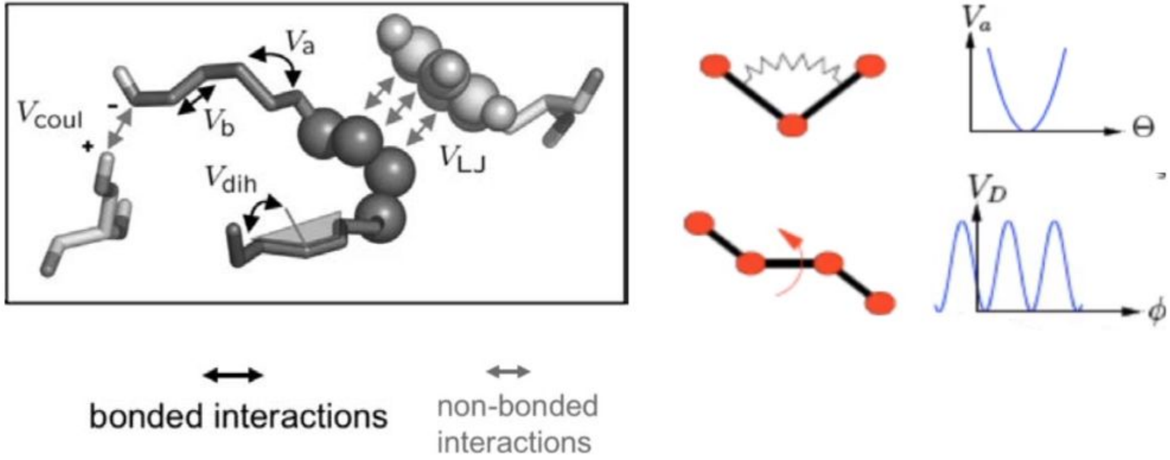
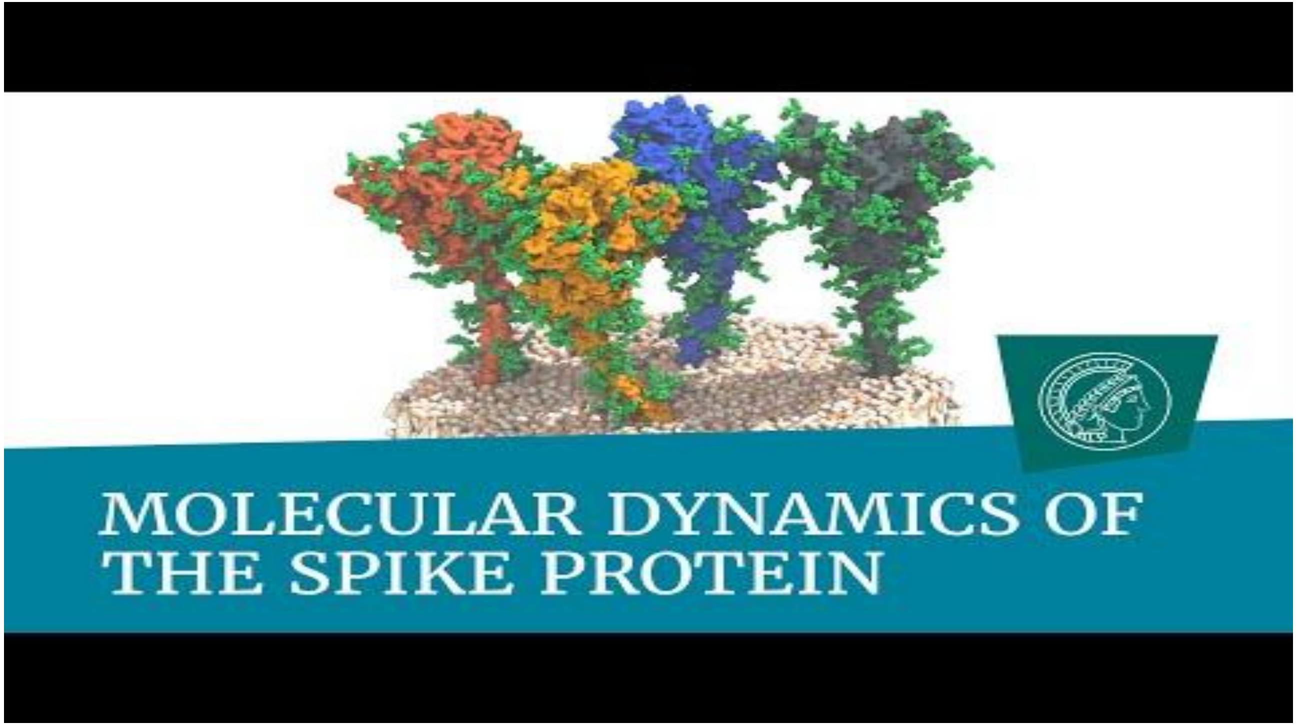


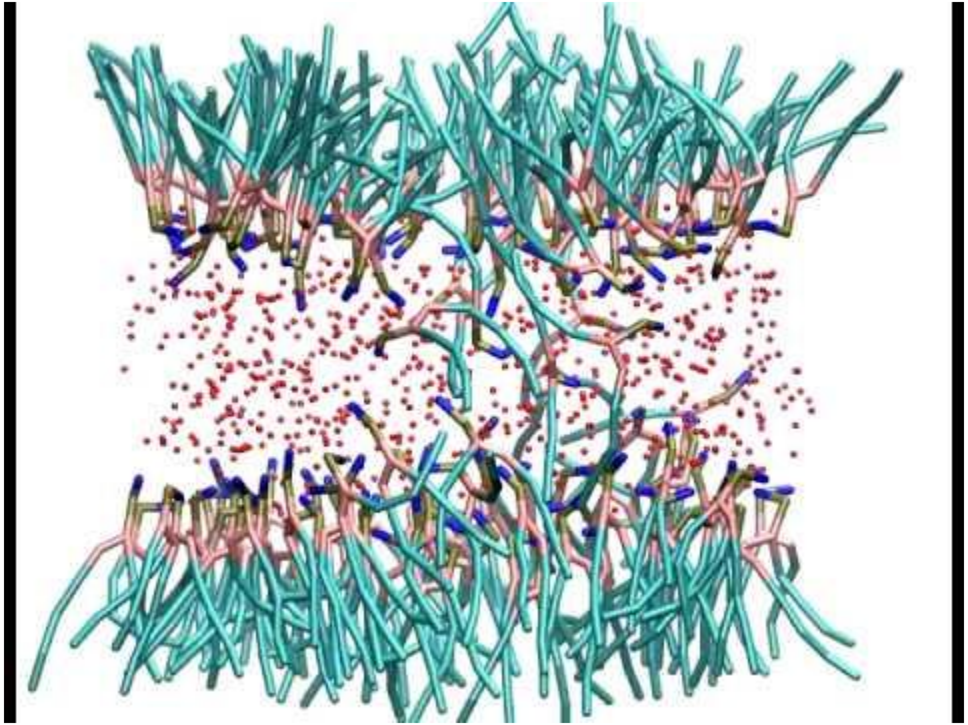
Figure taken from Prof. Frauke Gräther

Studying conformational flexibility and stability



https://www.youtube.com/watch?v=7AhQ19m2ok4

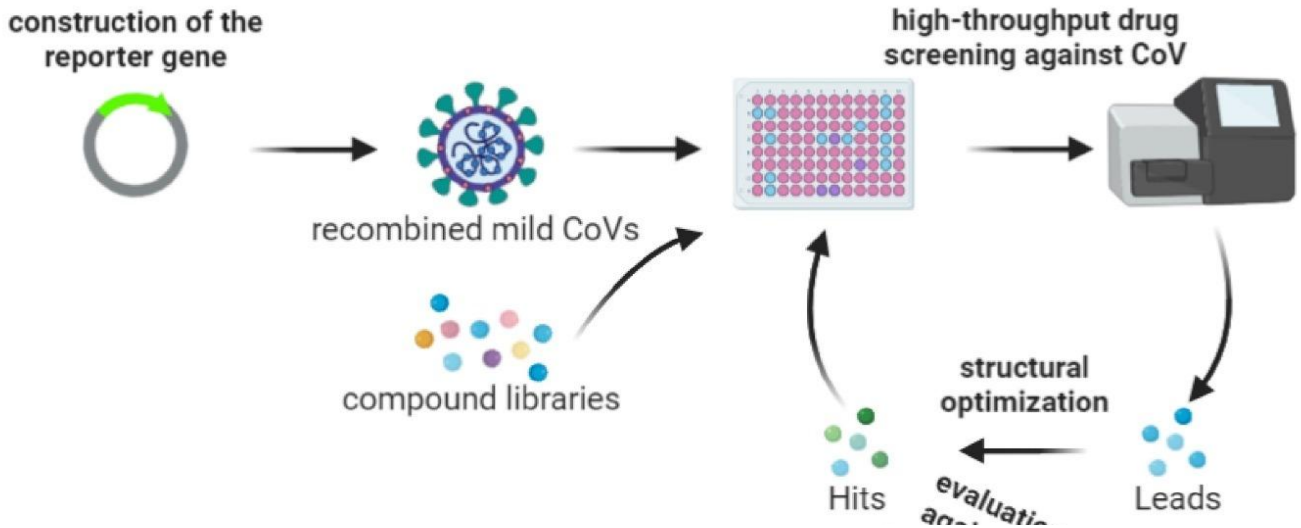
Modeling the lipid bilayer to better understand cell behaviour



https://www.youtube.com/watch?v=SbWh_XgCHyw

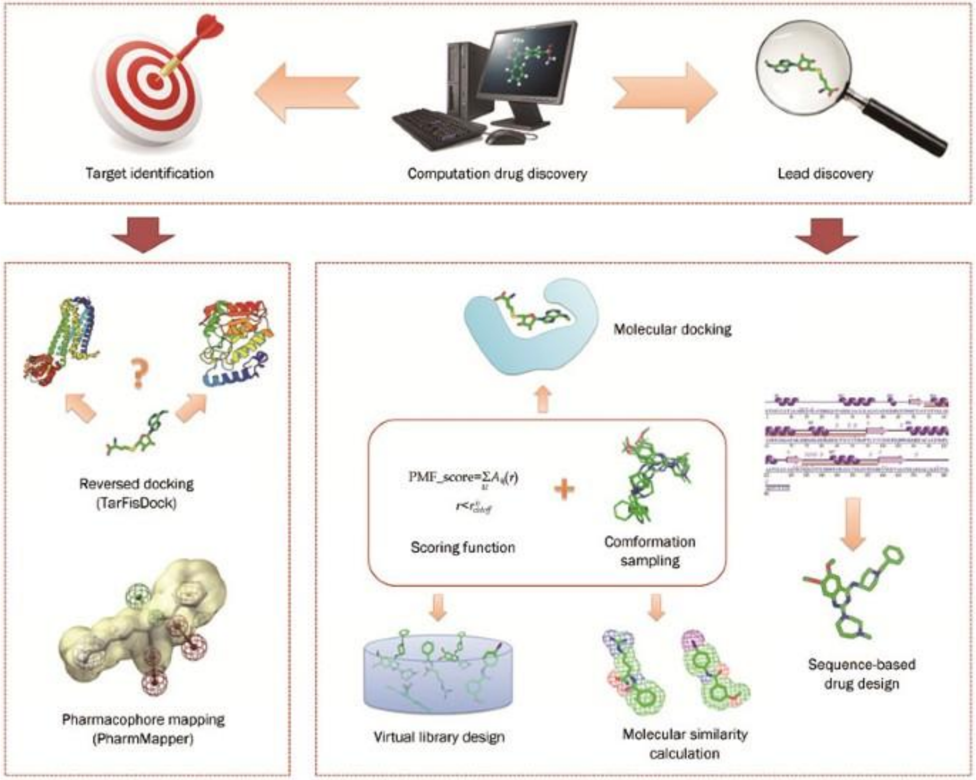
Drug screening

Normal drug screening workflow



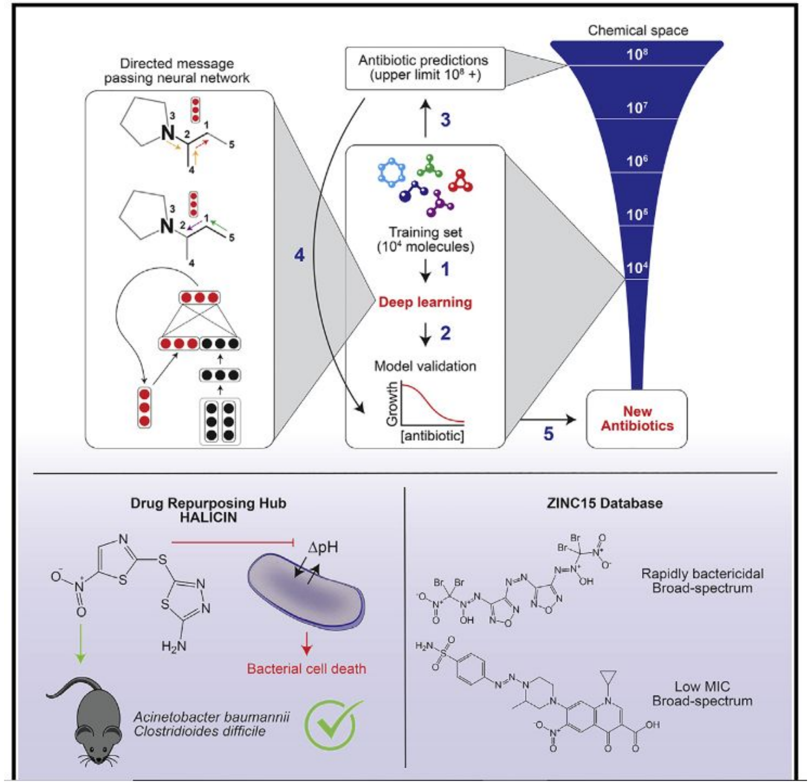
(Liu et al., 2020)

In silico screening reduces the number of molecules of interest



(Ou-Yang et al. 2012)

Deep learning yields new antibiotics candidates

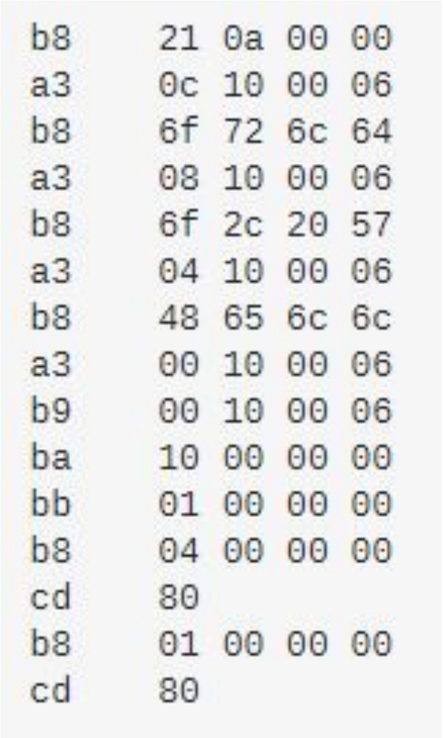


10^8 molecules screened

(Stokes et al., 2020)

Introduction to programming languages

From machine code to Python

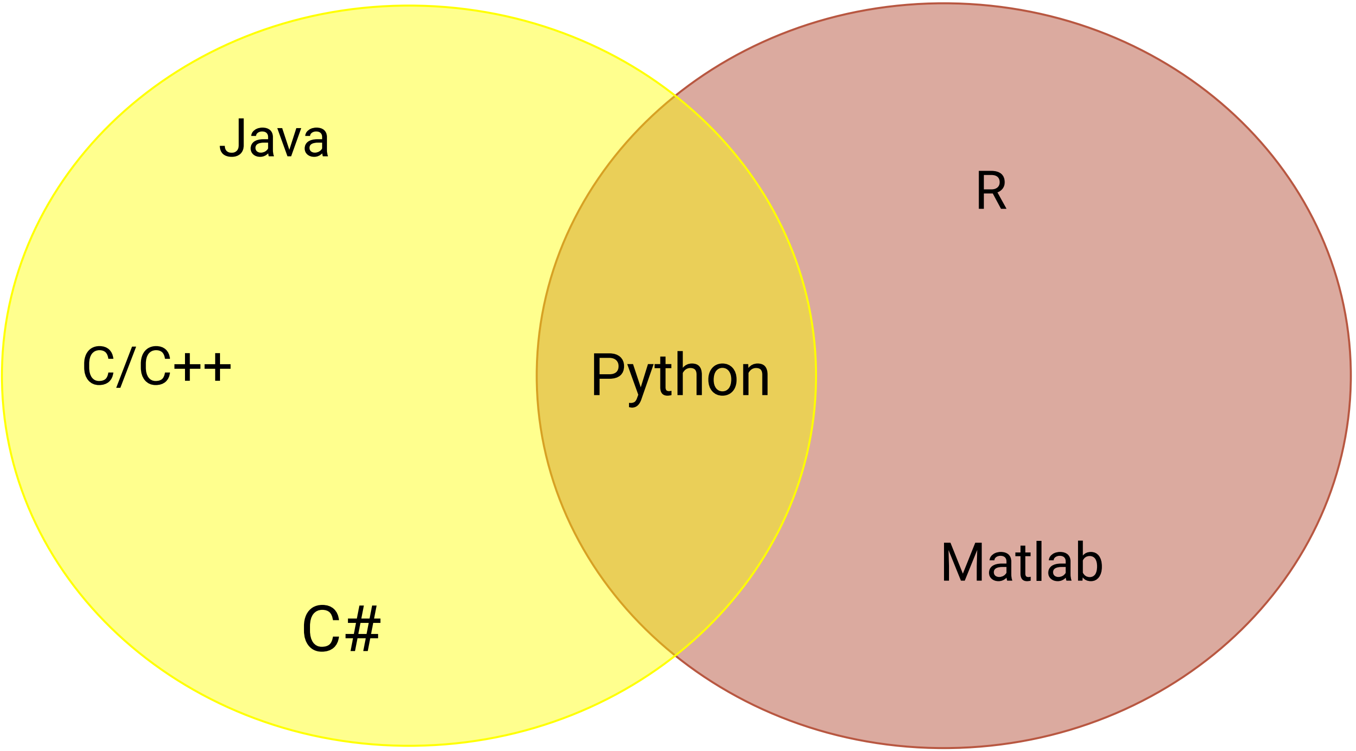


```
b8    21 0a 00 00
a3    0c 10 00 06
b8    6f 72 6c 64
a3    08 10 00 06
b8    6f 2c 20 57
a3    04 10 00 06
b8    48 65 6c 6c
a3    00 10 00 06
b9    00 10 00 06
ba    10 00 00 00
bb    01 00 00 00
b8    04 00 00 00
cd    80
b8    01 00 00 00
cd    80
```

```
print(“Hello World”)
```

Languages suited for general applications

Languages suited for research



Java

C/C++

C#

Python

R

Matlab

Why Python?

- Easy to Learn and Use
- Big Community
- Established in the Corporate World (primarily Google and YouTube)
- Versatility, Efficiency, Reliability, and Speed
- Hundreds of Libraries and Frameworks

Python vs. R

	Python	R
Purpose	multi purpose language	mainly statistical and data science applications
Learning curve:	easy to learn, linear learning curve	easy to learn, advanced functionalities can be difficult to use
Used by:	industry, academia, engineering	academics, scientists without programming skills, (few industries)
Visualization of Data	can be difficult	easy, straightforward

How to run Python?

The most relevant ways are:

- Simply executing .py files in your command line
- Using an IDE (Integrated Development Environment)
- The standard Python shell
- What we use in the course: **Jupyter notebook**

What is the “Jupyter Notebook”?

- Integrated framework to write:
  - Python code
  - Text/Markdown
  - Figures, Tables, animations

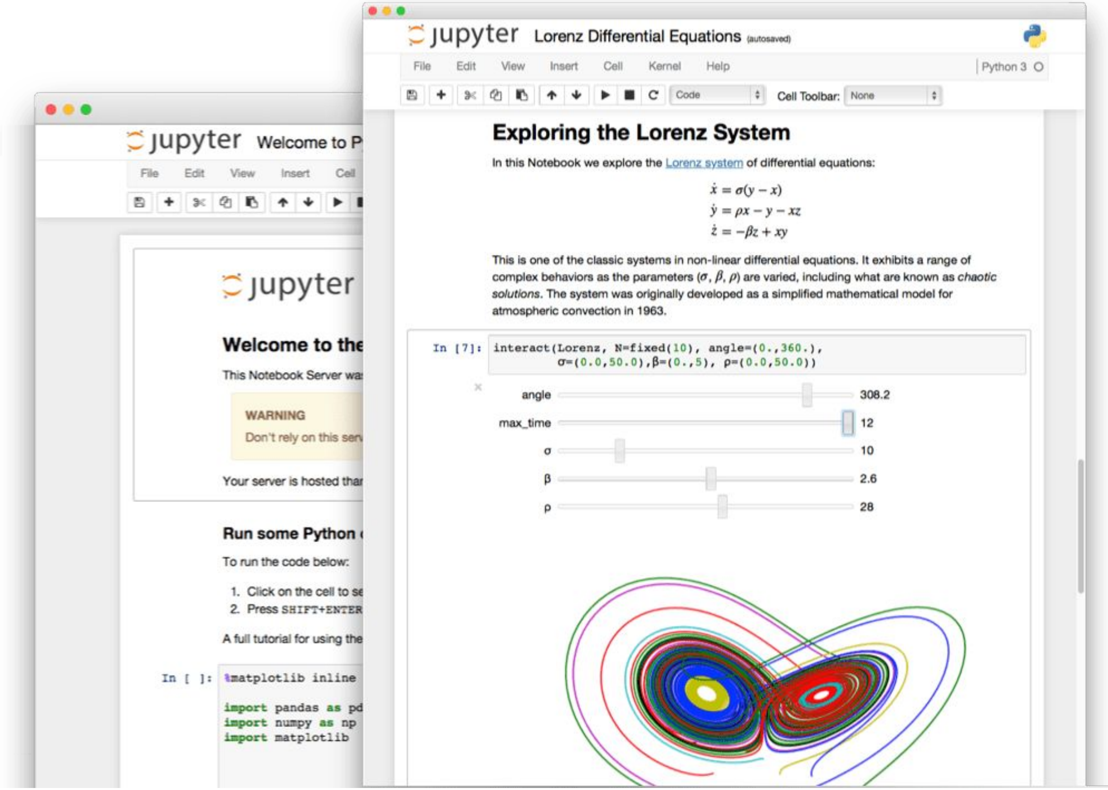
- Runs inside your browser
- Makes research reproducible and allows others to understand your code better

Jupyter Notebook

Editor that allows you to combine Python code with formatted text, plots, tables, and even animations.

Runs in your web-browser and is compatible with all operating systems

- ➡ **reproducible research**
- ➡ **great for sharing**



Welcome to the

This Notebook Server wa

WARNING

Don't rely on this serv

Your server is hosted thar

Run some Python

To run the code below:

1. Click on the cell to se
2. Press SHIFT+ENTER

A full tutorial for using the

```
In [ ]: %matplotlib inline

        import pandas as pd
        import numpy as np
        import matplotlib
```

Exploring the Lorenz System

In this Notebook we explore the Lorenz system of differential equations:

$$\dot{x} = \sigma(y - x)$$
$$\dot{y} = \rho x - y - xz$$
$$\dot{z} = -\beta z + xy$$

This is one of the classic systems in non-linear differential equations. It exhibits a range of complex behaviors as the parameters (σ, β, ρ) are varied, including what are known as *chaotic solutions*. The system was originally developed as a simplified mathematical model for atmospheric convection in 1963.

```
In [7]: interact(Lorenz, N=fixed(10), angle=(0.,360.),
                 σ=(0.0,50.0),β=(0.,5), ρ=(0.0,50.0))
```

angle 308.2

max_time 12

σ 10

β 2.6

ρ 28

https://jupyter.org/assets/jupyterpreview.png

Our roadmap for today:

Day 1 9:00-18:00 (Lunch Break 13:00-13:45)		
Topic	**Time**	**Break**
Talk “Overview of Bioinformatics”	1h	
Basic Variables and DataTypes	2h	15min
Conditional Clauses		
Lists and Tuple	2h	15min
Sets and Dictionaries		
Loops	2h	15min
Functions		

Schedule for tomorrow

Day 2 10:00 -15:00 (Lunch Break 12:45 - 13:30)		
Topic	**Time**	**Break**
Working with .txt and .csv files	2h	15 min
Data plotting		
Big Exercise 1	1h	10 min
Big Exercise 2	1h	
Guest talk by Jannik Buhr: “Data Science with R”	30min	

References and further reading material

References: Modeling

Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L., and Baker, D. (2003). Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. Science *302*, 1364–1368.

Englander, S.W., and Mayne, L. (2014). The nature of protein folding pathways. PNAS *111*, 15873–15880.

Boyoglu-Barnum, S., Ellis, D., Gillespie, R.A., Hutchinson, G.B., Park, Y.-J., Moin, S.M., Acton, O.J., Ravichandran, R., Murphy, M., Pettie, D., et al. (2021). Quadrivalent influenza nanoparticle vaccines induce broad protection. Nature *592*, 623–628.

Marcandalli, J., Fiala, B., Ols, S., Perotti, M., de van der Schueren, W., Snijder, J., Hodge, E., Benhaim, M., Ravichandran, R., Carter, L., et al. (2019). Induction of Potent Neutralizing Antibody Responses by a Designed Protein Nanoparticle Vaccine for Respiratory Syncytial Virus. Cell *176*, 1420-1431.e17

Leman JK et al. . Macromolecular modeling and design in Rosetta: recent methods and frameworks. Nat Methods. 2020 Jul;17(7):665-680. doi: 10.1038/s41592-020-0848-2. Epub 2020 Jun 1. PMID: 32483333; PMCID: PMC7603796.

Hub JS, Kubitzki MB, de Groot BL. Spontaneous quaternary and tertiary T-R transitions of human hemoglobin in molecular dynamics simulation. PLoS Comput Biol. 2010 May 6;6(5):e1000774. doi: 10.1371/journal.pcbi.1000774. PMID: 20463873; PMCID: PMC2865513.

Ou-Yang SS, Lu JY, Kong XQ, Liang ZJ, Luo C, Jiang H. Computational drug discovery. Acta Pharmacol Sin. 2012 Sep;33(9):1131-40. doi: 10.1038/aps.2012.109. Epub 2012 Aug 27. PMID: 22922346; PMCID: PMC4003107.

Liu J, Li K, Cheng L, Shao J, Yang S, Zhang W, Zhou G, de Vries AAF, Yu Z. A high-throughput drug screening strategy against coronaviruses. Int J Infect Dis. 2021 Feb;103:300-304. doi: 10.1016/j.ijid.2020.12.033. Epub 2020 Dec 14. PMID: 33333250; PMCID: PMC7832824.

Stokes JM et al. A Deep Learning Approach to Antibiotic Discovery. Cell. 2020 Feb 20;180(4):688-702.e13. doi: 10.1016/j.cell.2020.01.021. Erratum in: Cell. 2020 Apr 16;181(2):475-483. PMID: 32084340; PMCID: PMC8349178.