# Class 11: AlphaFold

Aadhya Tripathi (PID: A17878439)

## Table of contents



### Background

In this hands-on session we will utilize AlphaFold to predict protein structure from sequence (Jumper et al. 2021).

Without the aid of such approaches, it can take years of expensive laboratory work to determine the structure of just one protein. With AlphaFold we can now accurately compute a typical protein structure in as little as ten minutes.

The PDB database (the main repository of experimental structures) only has ~**250 thousand** structures, as we saw in the last lab. The main protein sequence database has **>200 million** sequences! Only 0.125% of known sequences have a known structure. This is called the "structure knowledge gap".

```
250000 / 200000000
```

```
[1] 0.00125
```

- Structures are much harder to determine than sequences.
- They are expensive (costing $1M on average).
- They take an average of 3-5 years to solve.

### EBI AlphaFold Database

The EBI has a database of pre-computed AlphaFold (AF) models called AFDB. This is growing all the time and can be useful to check before running AF ourselves.

## Running AlphaFold

We can download and run locally but we need a GPU. Or we can use "cloud" computing to run this on someone else's computer!

We will use ColabFold https://github.com/sokrypton/ColabFold

We previously found there was no AFDB entry for our HIV sequence.

```
>HIV-Pr-Dimer
PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGIGGFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQ
```

Here we will use AlphaFold2_mmseqs2.