DEEP LEARNING INDABAX
Tunisia 2023

Sequence-based predictor for the impact of mutation on protein stability

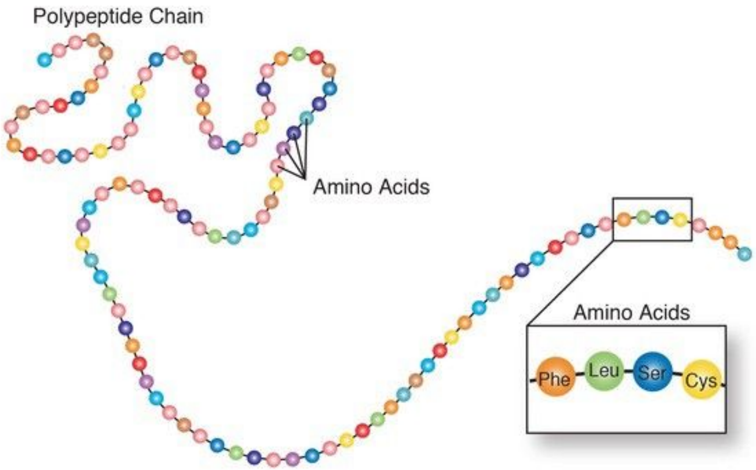
IEEE
SUPCOM STUDENT BRANCH

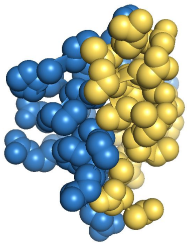
ZiNDI

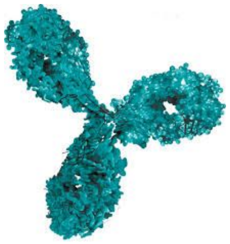
InstaDeep™

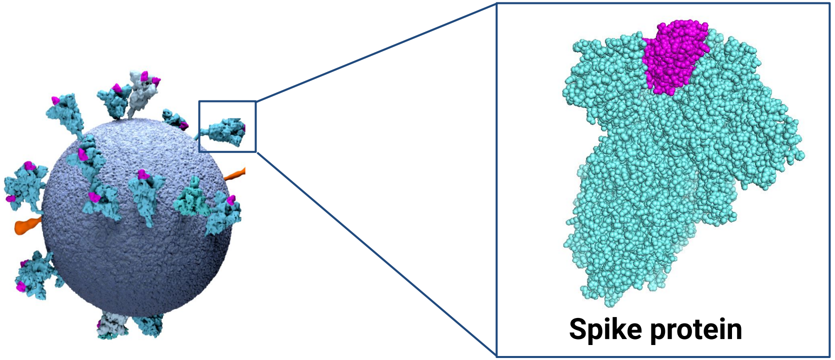
Introduction

- **Proteins** are the building blocks of life itself.
- They are constituted of multiple chains of **amino acids**.
- The latter are molecules that, when aligned, represent a **peptide chain**.



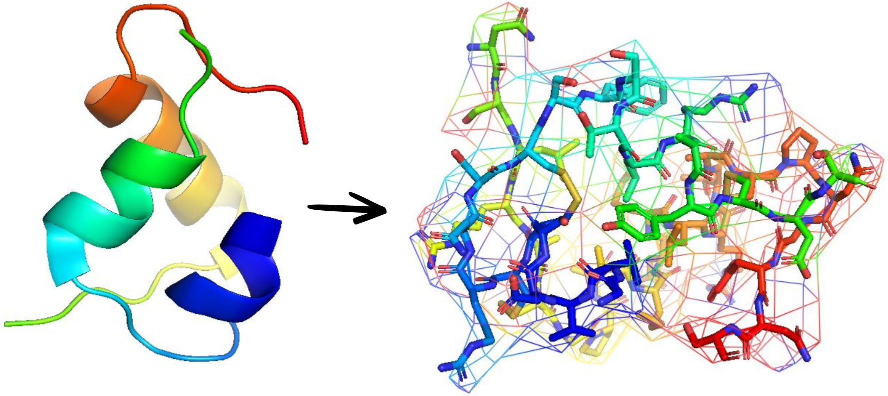

Insulin

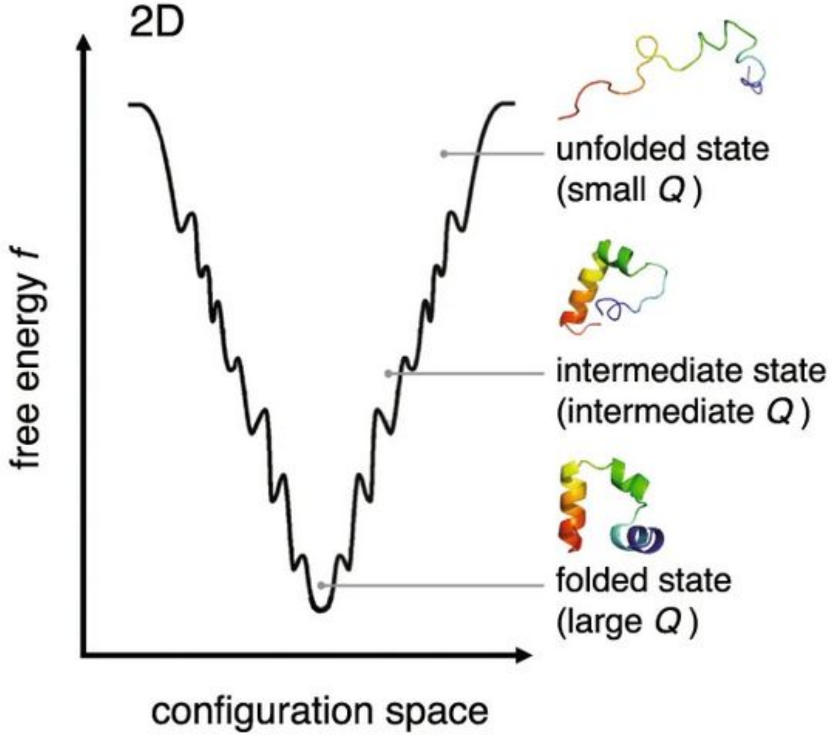

Antibody


Spike protein

Folding free energy Landscape

- Proteins are made up of different types of atoms, including carbon (**C**), hydrogen (**H**), oxygen (**O**), nitrogen (**N**), and sometimes sulfur (**S**).
- These atoms contribute significantly to protein folding and stability



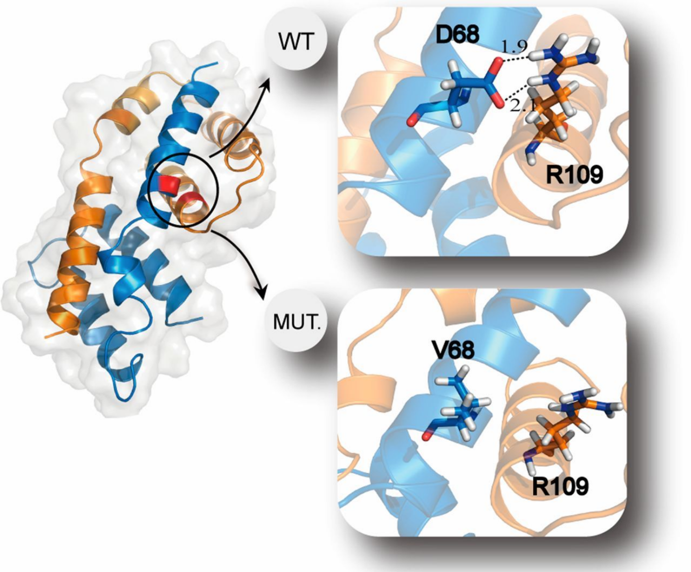


Protein Mutation

- **Mutations** are changes in genetic material
- Point mutation are chemical changes in just one base pair of a gene
- The change of a single nucleotide in a DNA strand can lead to the production of mutant protein
- The normal gene/proteins are called **Wild Type**

DNA seq: A A T G C A T A T G C A
mRNA seq: U U A C G U A U A C G U
Wt seq: leu Arg Ile Arg

DNA seq: A A T **T** C A T A T G C A
mRNA seq: U U A A G U A U A C G U
Mut seq: leu Ser Ile Arg

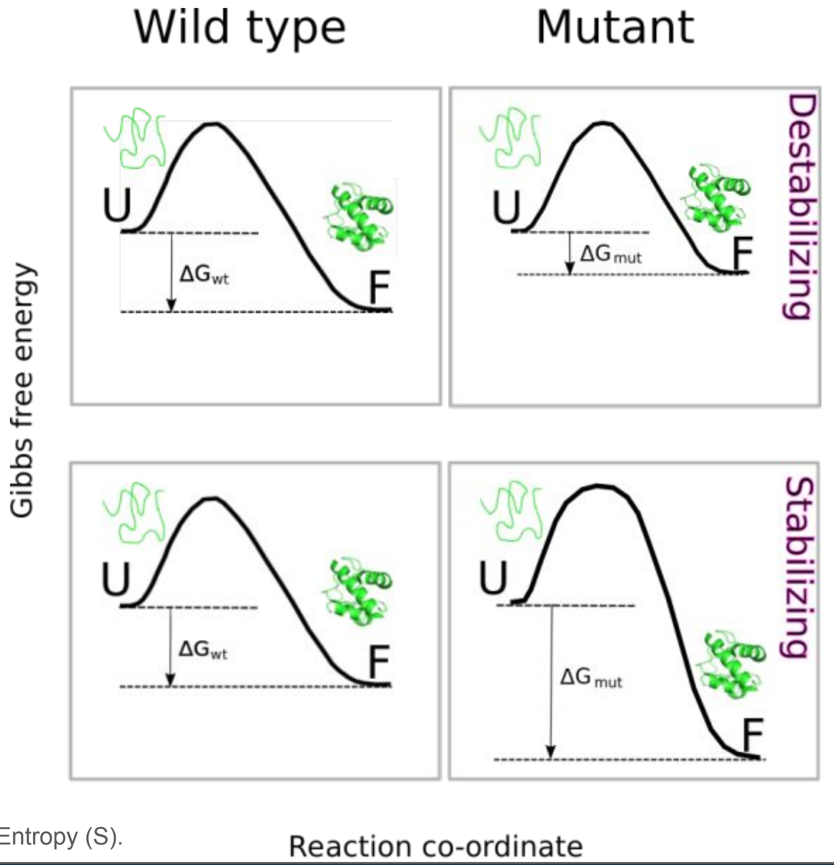


A deeper look at DDG for proteins

- Delta Delta G (**DDG**) is a metric for predicting how a single point mutation will affect protein stability.
- DDG (**change in Gibbs free energy**) is a measure of the change in energy between the folded and unfolded states (DGfolding) and the change in DGfolding when a point mutation is present.
- An excellent predictor of whether a point mutation will be favorable in terms of **protein stability**.

$$\Delta\Delta G = \Delta G^{\text{mutant}} - \Delta G^{\text{wt}}$$

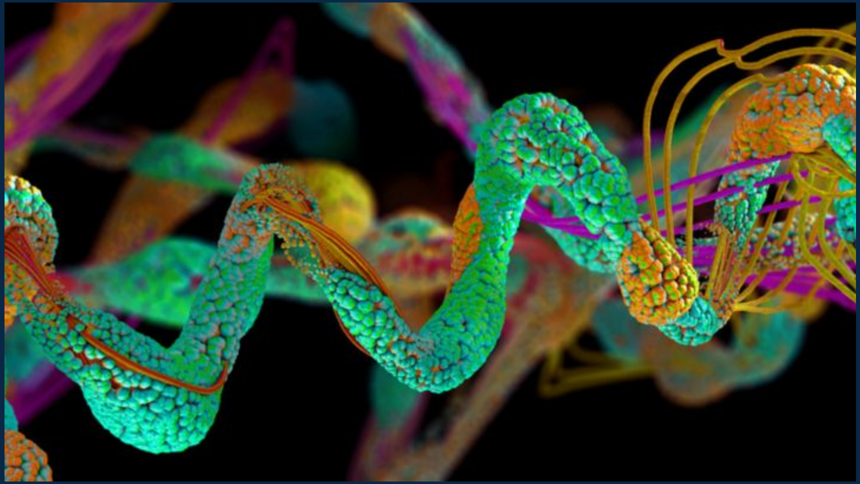
$$\Delta G = \Delta G_{\text{folded}} - \Delta G_{\text{unfolded}}$$



N.B: As a reminder, Gibbs free energy (G) = Enthalpy (H) – Temperature (T) x Entropy (S).

Sequence-based predictor for the impact of mutation on protein stability

Predict the thermodynamic folding stability of a protein (DDG) in response to a single amino acid mutation



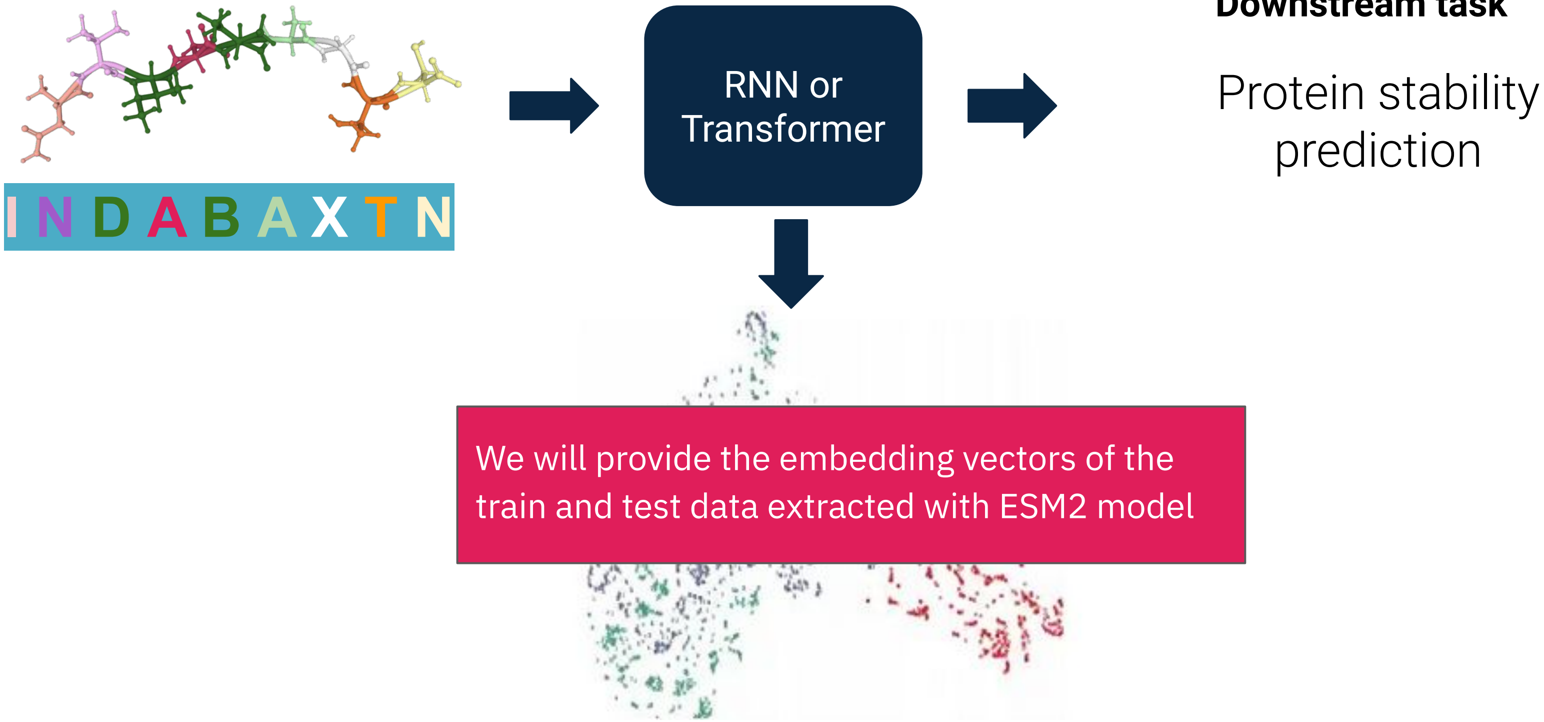
The data

- **≈340K** high-quality sequences with annotated labels

Variable definitions

- **ID:** Indicated the column index.
- **pdb_id:** It contains the 4 characters that represents the PDB structure or otherwise, something like “*HHH-rd1-0142*” if the structure was generated by Rosetta.
- **mutation:** Mutation applied to the wt_sequence in this pattern; XnY given X is the wild type amino acid(wt_aa), n is the position number of the amino acid that will be replaced(mutation_pos) and Y is the new amino acid(mut_aa).
- **wt_seq:** Wild Type sequence. The natural form, appearance or strain existing in the wild protein sequence.
- **mut_seq:** Mutant sequence. A protein sequence that has undergone a change or mutation from the natural form, appearance, or strain existing in the wild protein sequence.
- **ddg:** Delta Delta G is a metric for predicting how a single point mutation will affect protein stability.

The proposed solution : a Language model



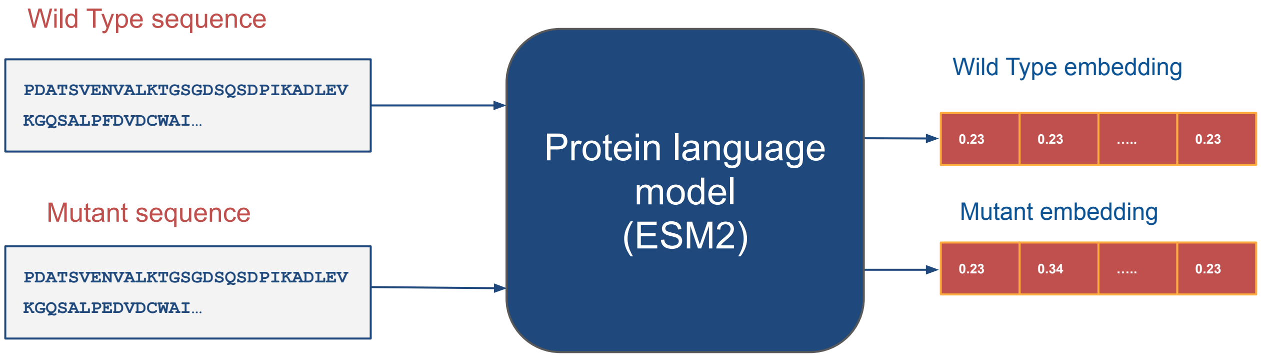
RNN or Transformer

Downstream task

Protein stability prediction

We will provide the embedding vectors of the train and test data extracted with ESM2 model

Embeddings for the protein sequences



Wild Type sequence

PDATSVENVALKTGSGDSQSDPIKADLEV KGQSALPFDVDCWAI...

Mutant sequence

PDATSVENVALKTGSGDSQSDPIKADLEV KGQSALPEDVDCWAI...

Protein language model (ESM2)

Wild Type embedding

0.23	0.23	.....	0.23

Mutant embedding

0.23	0.34	.....	0.23

We will provide you with this embedding data!

Starter notebook

colab

Link: https://bit.ly/indabax

Thanks and Good Luck!

DEEP
LEARNING
INDABAX