Protein design
Part 2: predicting mutation free energy

Anastassia Vorobieva

Introduction to protein design: predicting the effect of mutations on stability

- Introduction to protein design
- The principle behind DDG predictions
- Expected output
- Considerations for selecting a template PDB
- The Protein Repair One Stop Shop (PROSS)
- Fitness effect prediction with evolution-based protein language models

Introduction to protein design

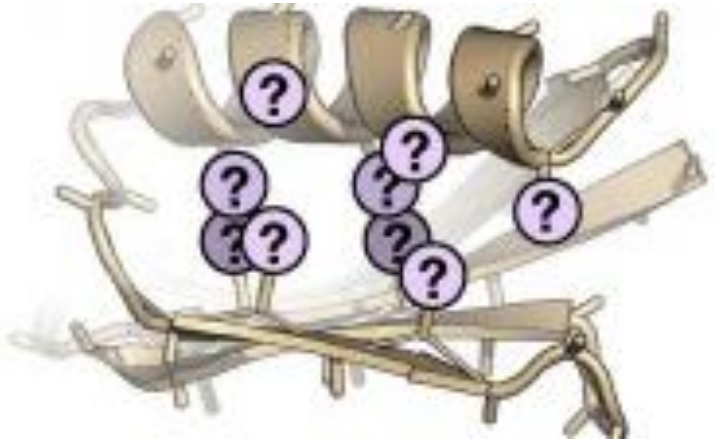
Understanding amino acid patterns, predicting the effect of mutations

Protein design

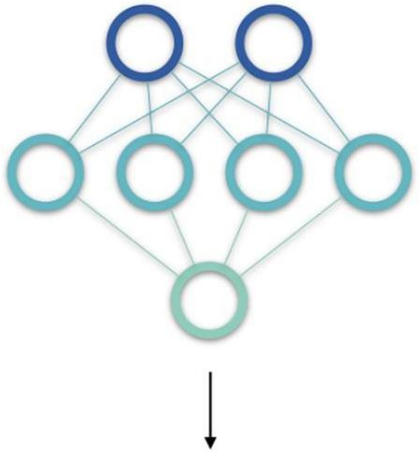
- **Protein design** is the process of manipulating a protein sequence to achieve a specific functions, structures, or stability.

Structure *vs* sequence-based protein design

Structure-based design



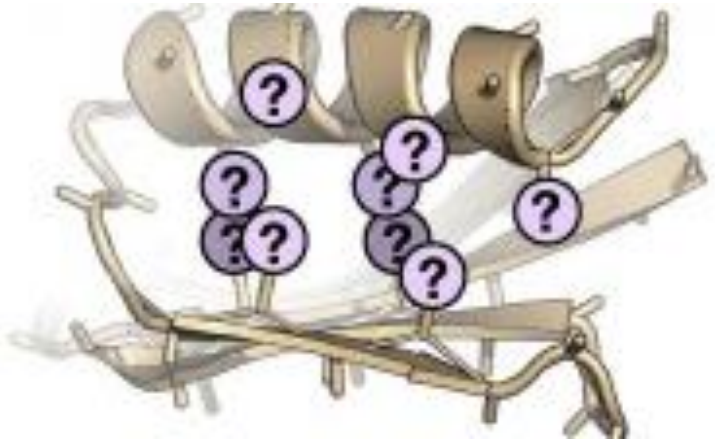
Direct sequence generation (e.g. LLMs)



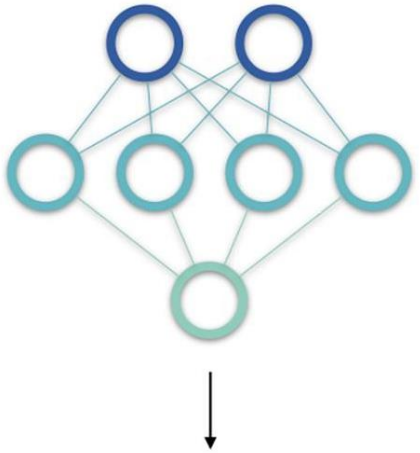
…MALKIPTHNHM…
…VFRDCEWS…
…WYIOPMNVGTDEW…

Structure *vs* sequence-based protein design

Structure-based design



Direct sequence generation (e.g. LLMs)



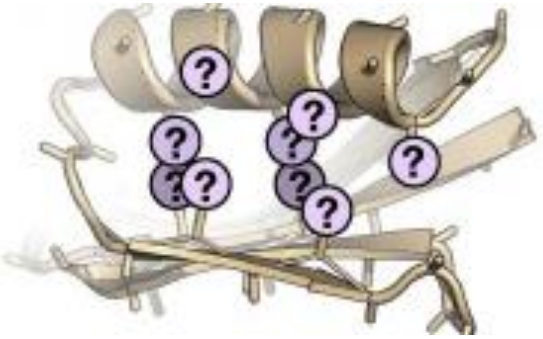
…MALKIPTHNHM…
…VFRDCEWS…
…WYIOPMNVGTDEW…

Protein design

- **Protein design** is the process of manipulating a protein sequence to achieve a specific functions, structures, or stability.

- **Applications** of protein design:
  - Biomedicine: design binders, improve enzyme stability for gene therapy
  - Biotechnology: improve the stability and function of enzymes for cleaning, chemical synthesis, bio-remediation
  - Synthetic biology

Protein design



- **Protein design** is the process of manipulating a protein sequence to achieve a specific functions, structures, or stability.

- **Applications** of protein design:
  - Biomedicine: design binders, improve enzyme stability for gene therapy
  - Biotechnology: improve the stability and function of enzymes for cleaning, chemical synthesis, bio-remediation
  - Synthetic biology

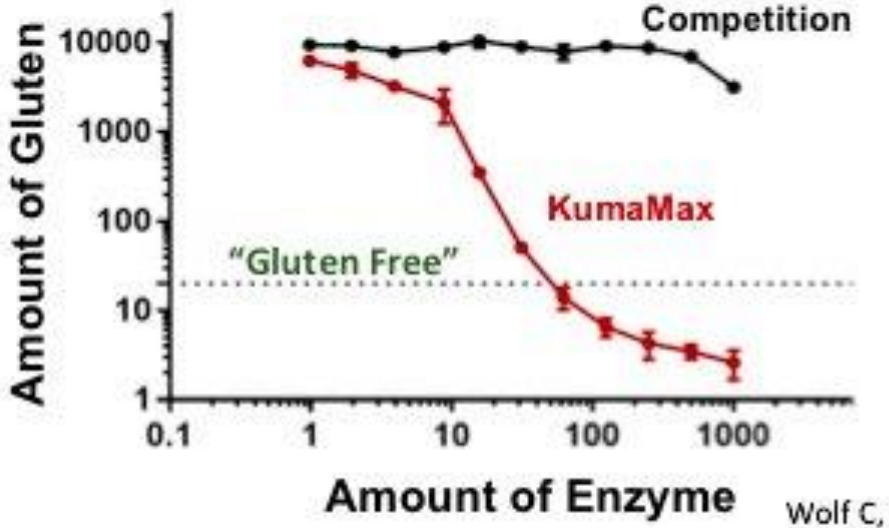
- Predicting the Effect of Mutations on Protein Stability and Function
  Understanding mutation effects can help in the design of proteins that are more stable, more active, or more specific to their targets

KumaMax: Designed Gluten Busting Enzyme Therapy for Celiac Disease

- Computationally re-designed enzyme
- Specific for gluten, destroys the immunogenic peptide
- Active only in low pH of the stomach
- Rapid and near complete degradation of gluten in beer, bread, etc.

Oral Pill

KumaMax
Designed Enzyme

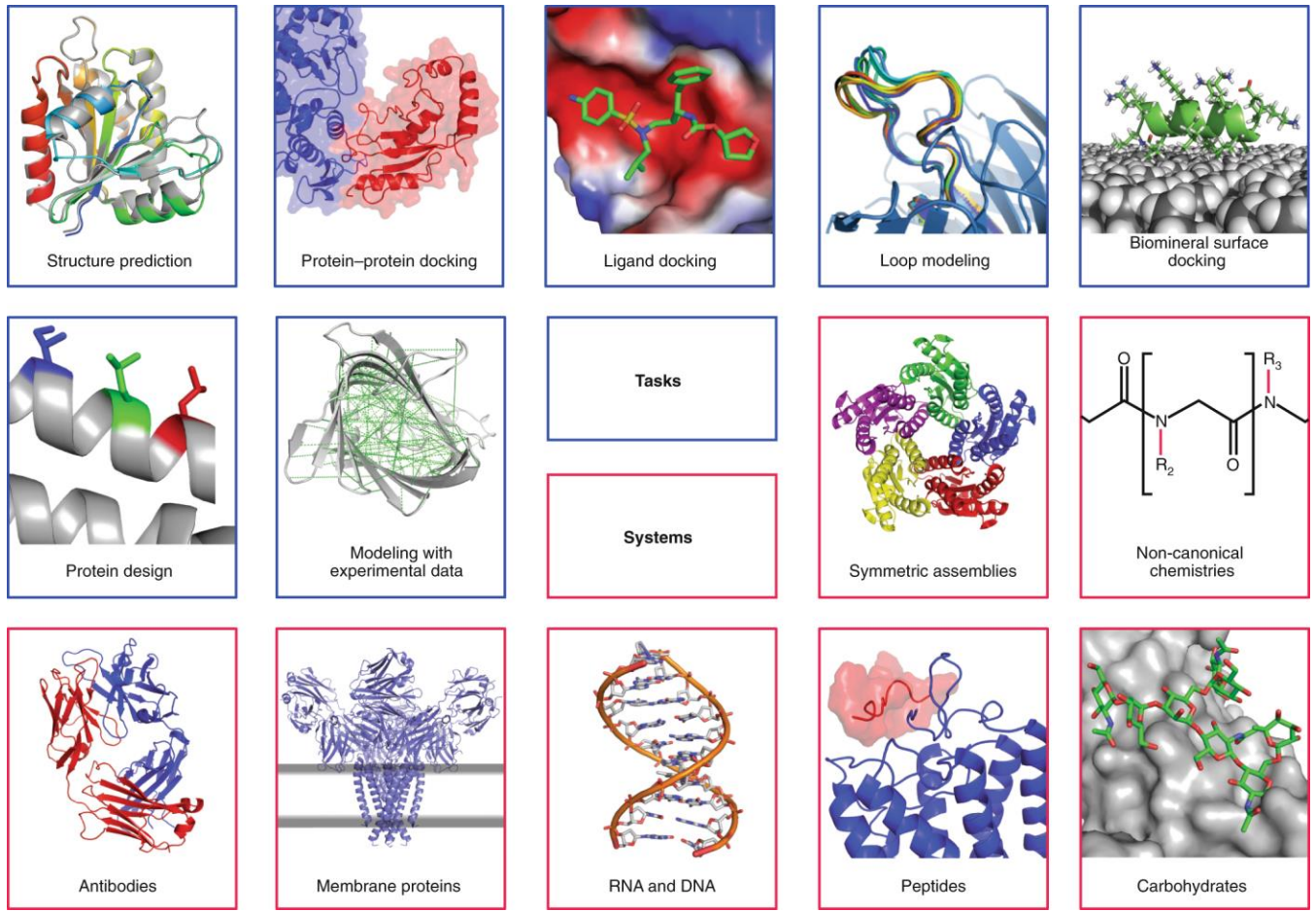


INSTITUTE FOR PROTEIN DESIGN
UNIVERSITY of WASHINGTON

Wolf C, et al. Engineering of Kuma030: A Gliadin Peptidase That Rapidly Degrades Immunogenic Gliadin Peptides in Gastric Conditions. J Am Chem Soc. 2015, 137:13106-13.

B. Preparing the PDB for Rosetta calculations

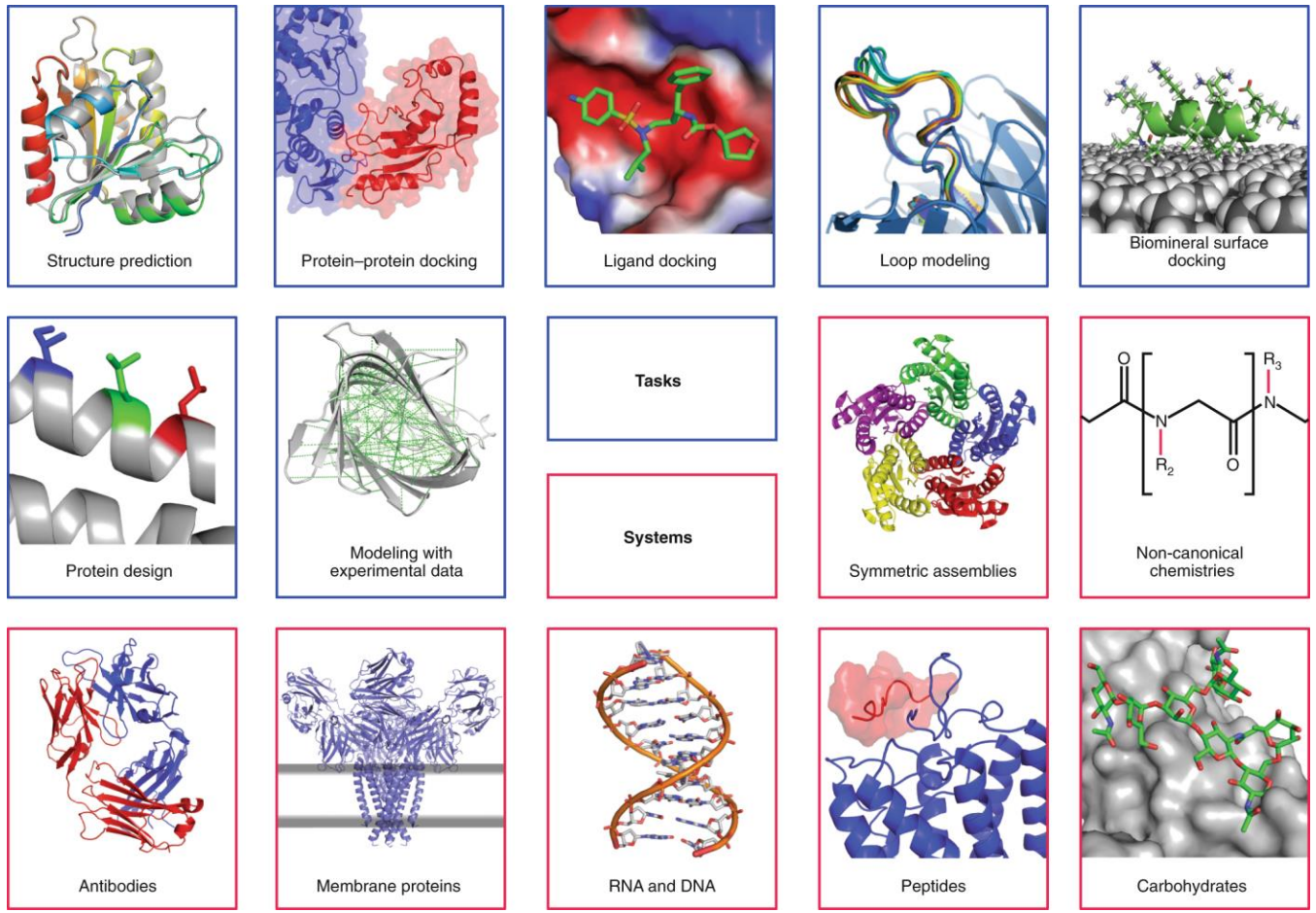
What is Rosetta ?



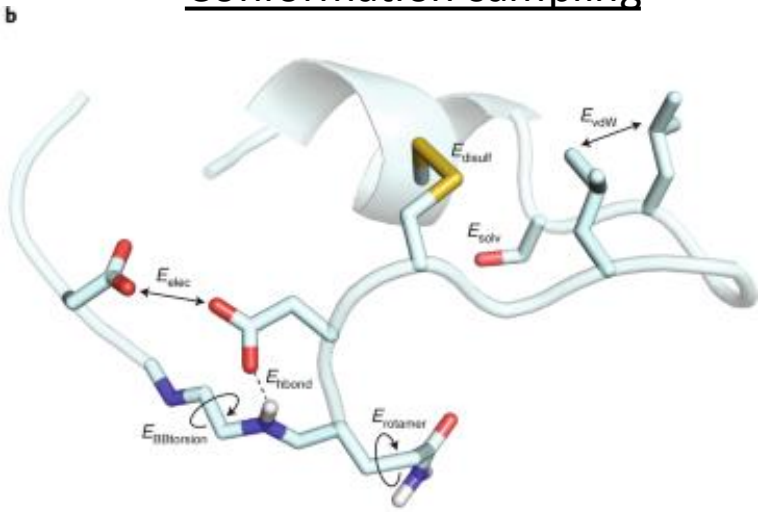
https://doi.org/10.1038/s41592-020-0848-2

B. Preparing the PDB for Rosetta calculations

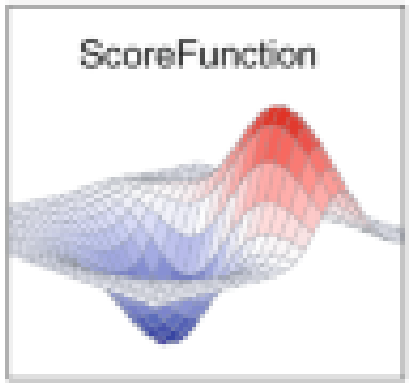
What is Rosetta ?



Conformation sampling



Scoring



E_{vdW}	Lennard–Jones for attractive or repulsive interaction
E_{hbond}	Hydrogen bonding allows buried polar atoms
E_{elec}	Electrostatic interaction between charges
E_{disulf}	Disulfide bonds between cysteines

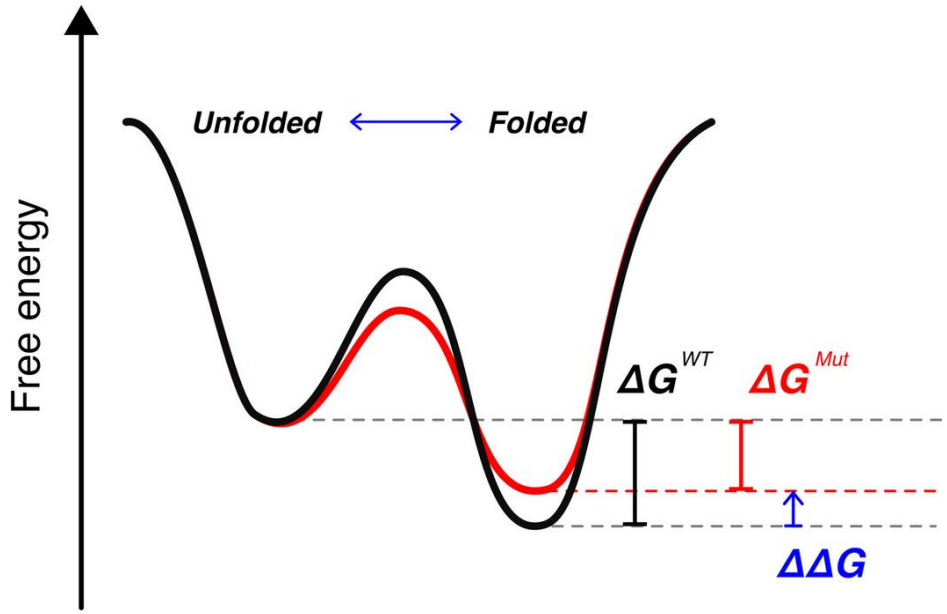
E_{solv}	Implicit solvation model penalizes buried polar atoms
$E_{BBtorsion}$	Backbone torsion preferences from main-chain potential
$E_{rotamer}$	Side-chain torsion angles from rotamer library
E_{ref}	Unfolded state reference energy for design

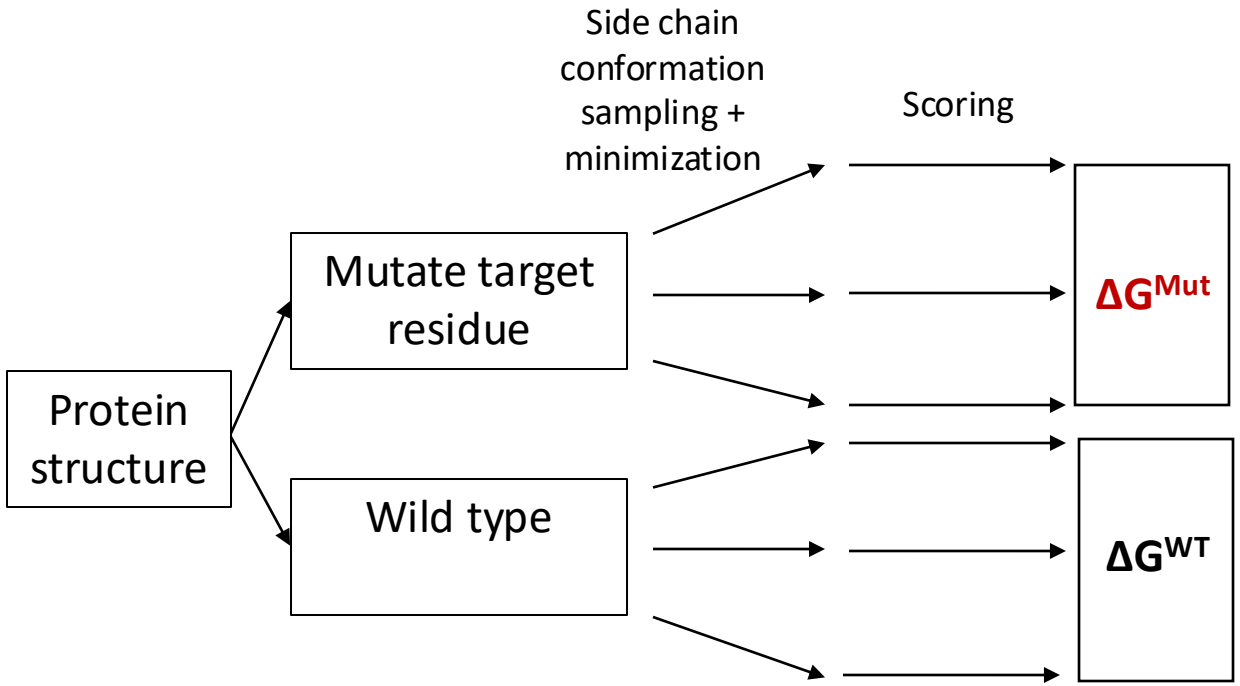
https://doi.org/10.1038/s41592-020-0848-2

The PDB file is a list of atom coordinates

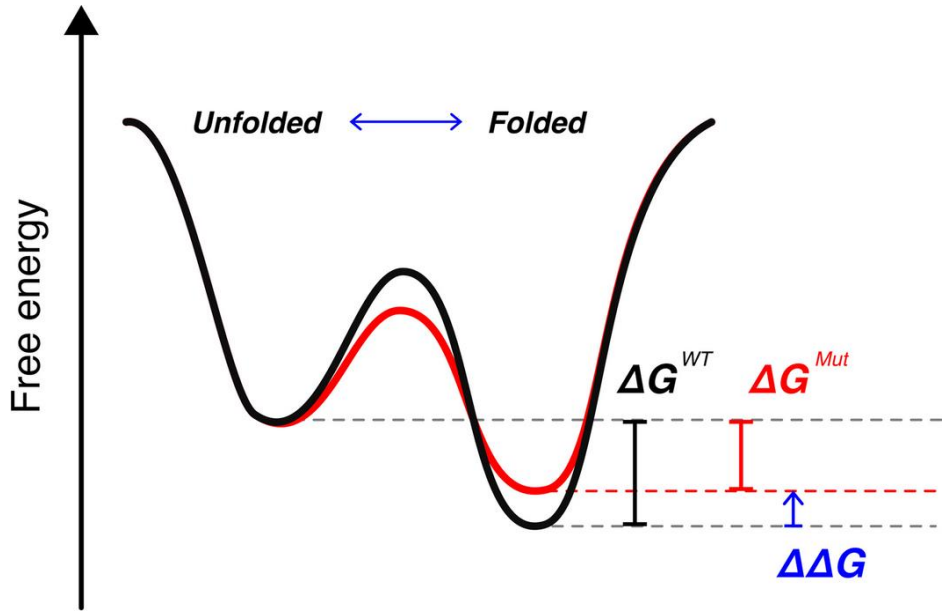
```
ATOM   8178  CD2 LEU A 207     135.703 156.448 175.146  1.00 59.65           C
ATOM   8179  N   LEU A 208     138.542 160.482 176.163  1.00 64.39           N
ATOM   8180  CA  LEU A 208     139.558 160.550 177.204  1.00 64.39           C
ATOM   8181  C   LEU A 208     139.523 161.898 177.917  1.00 64.39           C
ATOM   8182  O   LEU A 208     140.107 162.065 178.988  1.00 64.39           O
ATOM   8183  CB  LEU A 208     140.944 160.306 176.607  1.00 64.39           C
ATOM   8184  CG  LEU A 208     141.162 158.912 176.018  1.00 64.39           C
ATOM   8185  CD1 LEU A 208     142.520 158.825 175.340  1.00 64.39           C
ATOM   8186  CD2 LEU A 208     141.011 157.842 177.088  1.00 64.39           C
ATOM   8187  OXT LEU A 208     138.909 162.851 177.438  1.00 64.39           O
TER    8188      LEU A 208
HETATM 8189  O1' LMT E 301     138.158 131.617 150.769  1.00 20.00           O
HETATM 8190  C1  LMT E 301     137.918 131.876 152.147  1.00 20.00           C
HETATM 8191  C2  LMT E 301     137.062 133.128 152.260  1.00 20.00           C
HETATM 8192  C3  LMT E 301     137.071 133.678 153.676  1.00 20.00           C
HETATM 8193  C4  LMT E 301     136.376 135.025 153.714  1.00 20.00           C
HETATM 8194  C5  LMT E 301     135.935 135.347 155.125  1.00 20.00           C
HETATM 8195  C6  LMT E 301     135.530 136.801 155.206  1.00 20.00           C
HETATM 8196  C7  LMT E 301     136.343 137.543 156.242  1.00 20.00           C
HETATM 8197  C8  LMT E 301     135.478 138.063 157.375  1.00 20.00           C
HETATM 8198  C9  LMT E 301     136.194 137.846 158.690  1.00 20.00           C
HETATM 8199  C10 LMT E 301     135.627 138.700 159.797  1.00 20.00           C
HETATM 8200  C11 LMT E 301     135.963 138.052 161.121  1.00 20.00           C
HETATM 8201  C12 LMT E 301     135.929 139.084 162.215  1.00 20.00           C
```

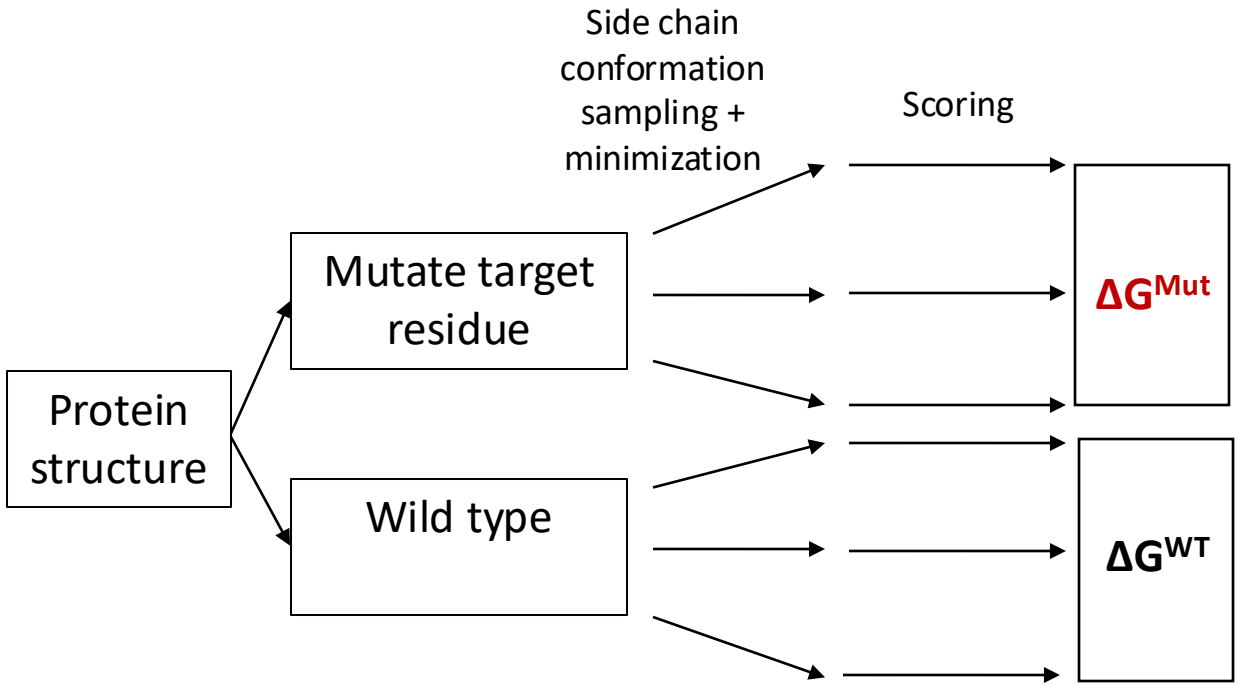
Running DDG calculations – the principle





Running DDG calculations – the principle

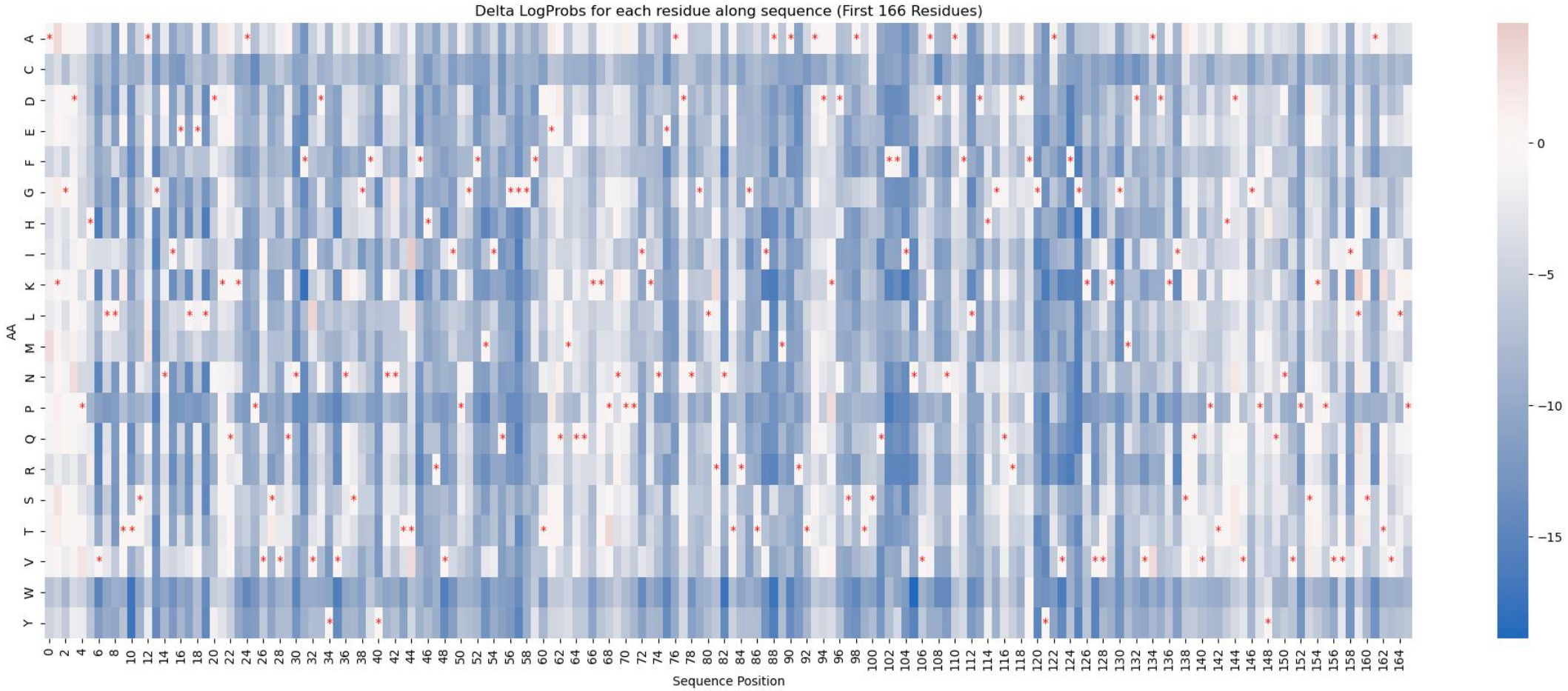




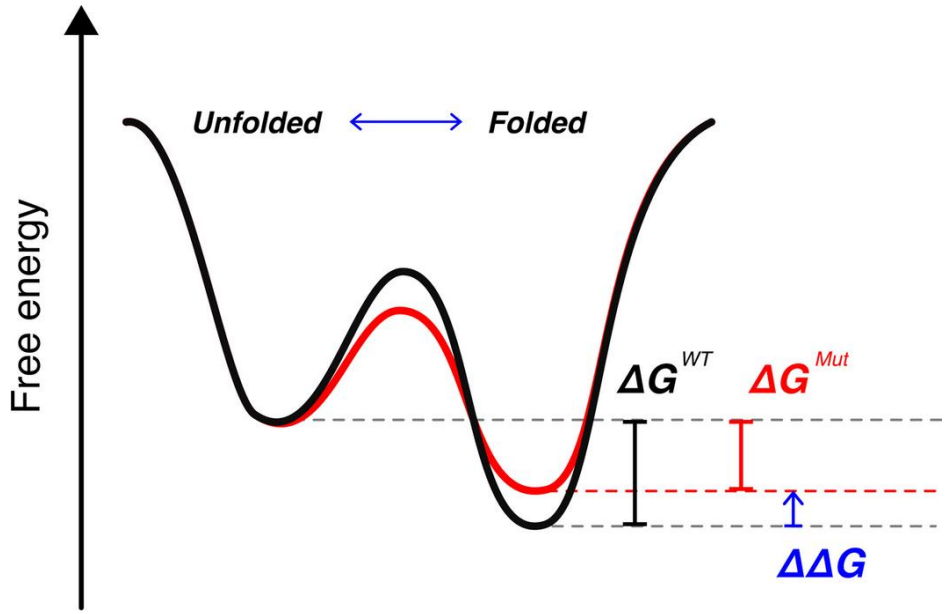
$$\Delta\Delta G = \frac{\sum_i MUT\_total\_score_i}{ddg\_iterations} - \frac{\sum_i WT\_total\_score_i}{ddg\_iterations}$$

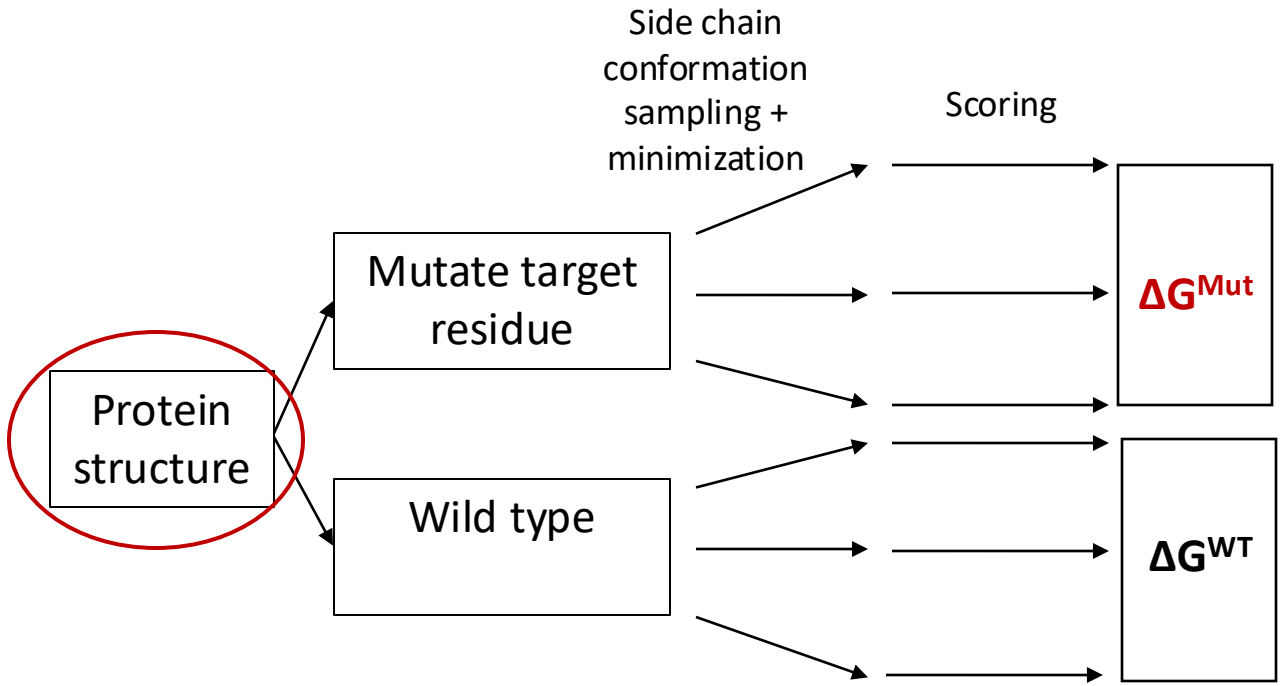
Tools: FoldX, **Rosetta**

Running DDG calculations – the **heatmap** output



Running DDG calculations – the principle

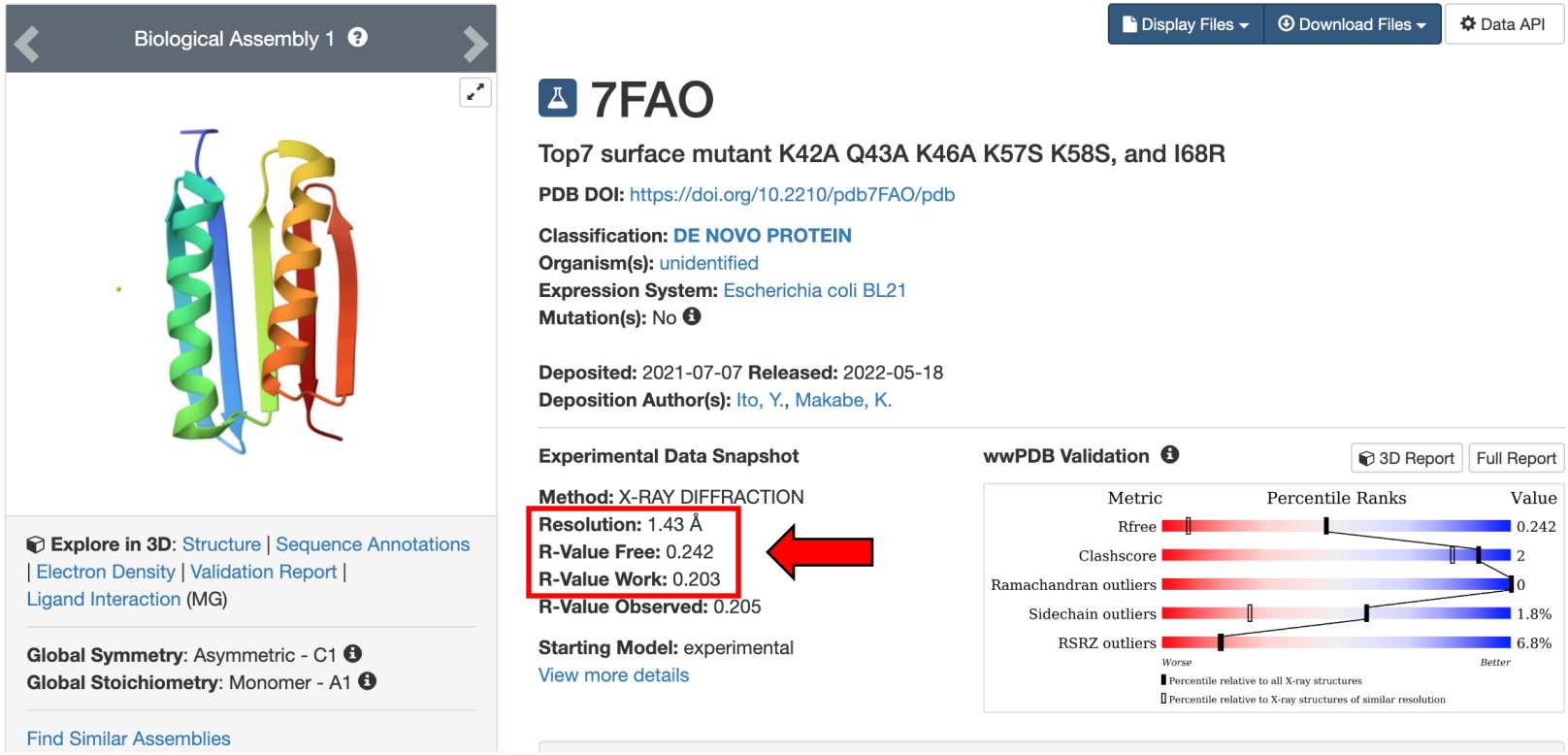




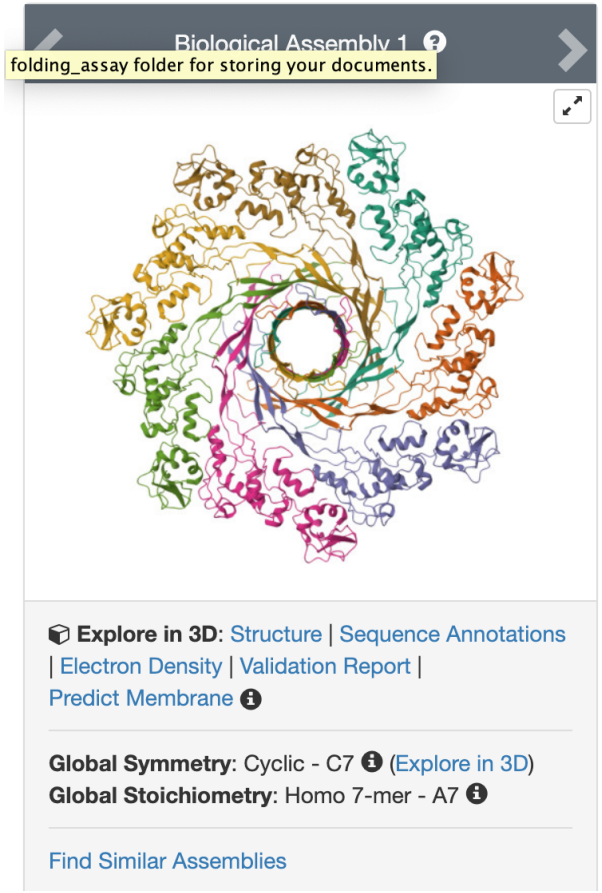
$$\Delta\Delta G = \frac{\sum_i MUT\_total\_score_i}{ddg\_iterations} - \frac{\sum_i WT\_total\_score_i}{ddg\_iterations}$$

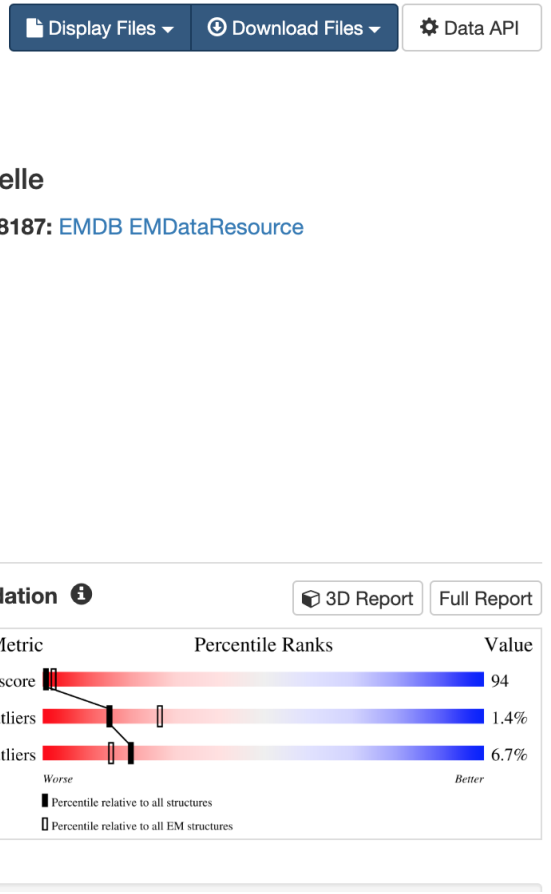
Tools: FoldX, **Rosetta**

1. Selecting your template PDB structure



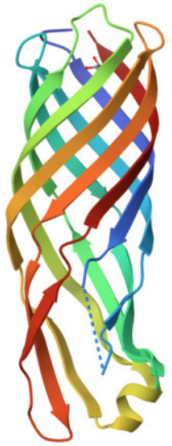
Problem: Low resolution structure





Problem: Poor refinement

Biological Assembly 1



Explore in 3D: Structure | Sequence Annotations | Electron Density | Validation Report | Ligand Interaction (GOL) | Predict Membrane

Global Symmetry: Asymmetric - C1
Global Stoichiometry: Monomer - A1

Find Similar Assemblies

Display Files | Download Files | Data API

2F1V

Outer membrane protein OmpW

PDB DOI: https://doi.org/10.2210/pdb2F1V/pdb

TMB12_3_nb3

Classification: MEMBRANE PROTEIN
Organism(s): Escherichia coli K-12
Expression System: Escherichia coli
Mutation(s): No
Membrane Protein: Yes OPM PDBTM MemProtMD

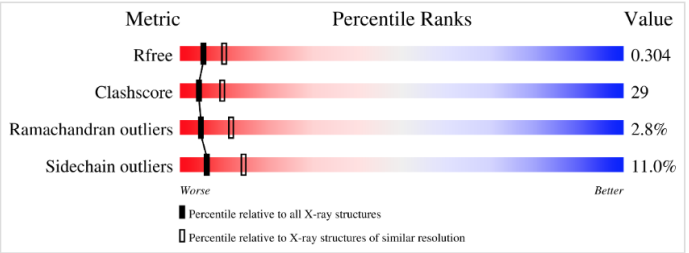
Deposited: 2005-11-15 **Released:** 2006-01-24
Deposition Author(s): van den Berg, B.

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.70 Å
R-Value Free: 0.314
R-Value Work: 0.292
R-Value Observed: 0.293

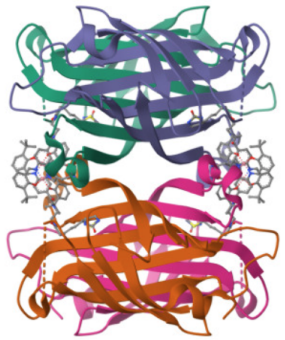
Starting Model: experimental
View more details

wwPDB Validation 3D Report | Full Report



Problem: Missing loops, alternative conformations (b-factor)

Biological Assembly 1



Explore in 3D: Structure | Sequence Annotations | Validation Report | Ligand Interaction (UFU)

Global Symmetry: Dihedral - D2 (Explore in 3D)
Global Stoichiometry: Homo 4-mer - A4

Find Similar Assemblies

Biological assembly 1 assigned by authors and generated by PISA (software)

Biological Assembly Evidence: gel filtration

Macromolecule Content

Display Files | Download Files | Data API

8QEX

Streptavidin variant with a cobalt catalyst for CH metal-catalyzed hydrogen-atom-transfer (M-HAT)

PDB DOI: https://doi.org/10.2210/pdb8QEX/pdb

Classification: METAL BINDING PROTEIN
Organism(s): Streptomyces avidinii
Expression System: Escherichia coli BL21(DE3)
Mutation(s): No

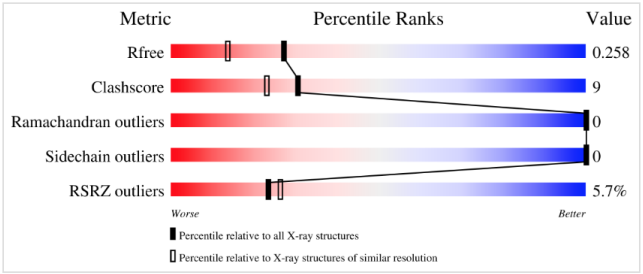
Deposited: 2023-09-01 **Released:** 2024-07-31
Deposition Author(s): Jakob, R.P., Chen, D., Ward, T.R.
Funding Organization(s): Not funded

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 1.90 Å
R-Value Free: 0.254
R-Value Work: 0.229
R-Value Observed: 0.230

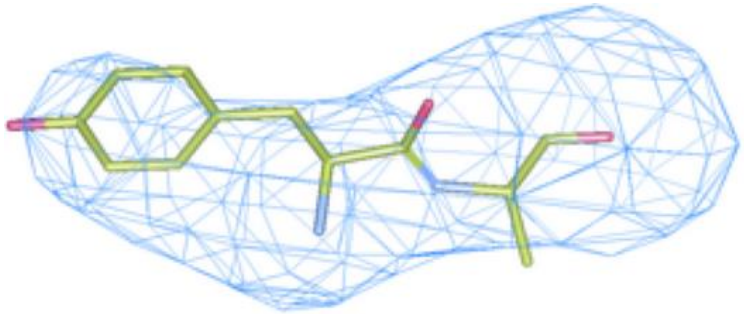
Starting Model: experimental
View more details

wwPDB Validation 3D Report | Full Report

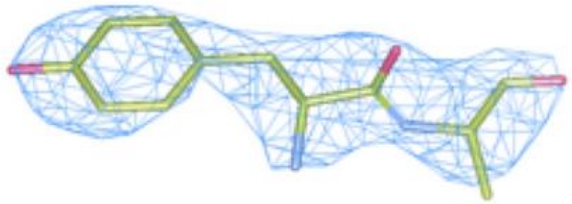


Ligand Structure Quality Assessment

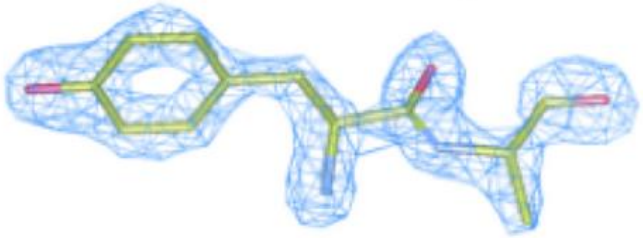
Problem: Ligand modelling!!



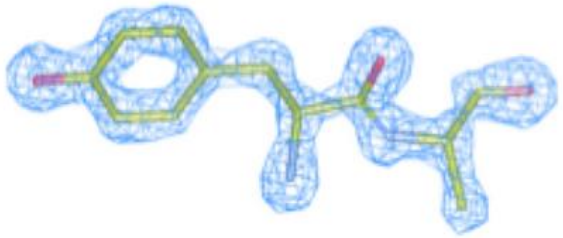
5.0 Å resolution, 2.5σ



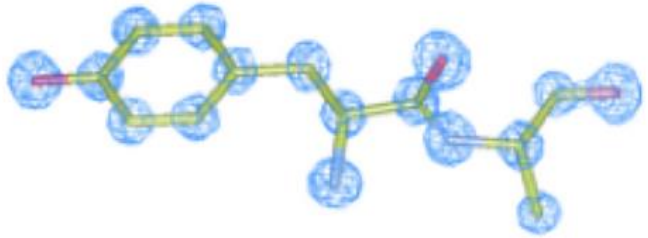
3.0 Å resolution, 6.5σ



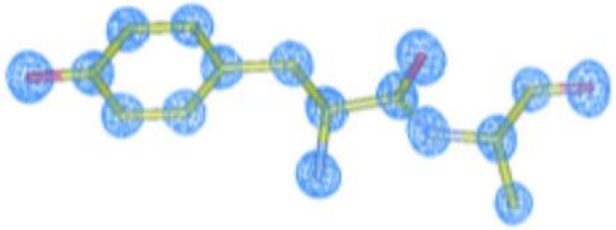
2.0 Å resolution, 8.5σ



1.5 Å resolution, 10σ



1.0 Å resolution, 14σ



0.6 Å resolution, 12σ

https://www.cresset-group.com/discovery/specific/protein-preparation/

Q: What are the different steps to prepare the protein structure for energy calculations?

A. The first step is to clean the PDB (remove non-protein atoms)

A. Clean PDB

```
ATOM   8178  CD2 LEU A 207     135.703 156.448 175.146  1.00 59.65           C
ATOM   8179  N   LEU A 208     138.542 160.482 176.163  1.00 64.39           N
ATOM   8180  CA  LEU A 208     139.558 160.550 177.204  1.00 64.39           C
ATOM   8181  C   LEU A 208     139.523 161.898 177.917  1.00 64.39           C
ATOM   8182  O   LEU A 208     140.107 162.065 178.988  1.00 64.39           O
ATOM   8183  CB  LEU A 208     140.944 160.306 176.607  1.00 64.39           C
ATOM   8184  CG  LEU A 208     141.162 158.912 176.018  1.00 64.39           C
ATOM   8185  CD1 LEU A 208     142.520 158.825 175.340  1.00 64.39           C
ATOM   8186  CD2 LEU A 208     141.011 157.842 177.088  1.00 64.39           C
ATOM   8187  OXT LEU A 208     138.909 162.851 177.438  1.00 64.39           O
TER    8188      LEU A 208
HETATM 8189  O1' LMT E 301     138.158 131.617 150.769  1.00 20.00           O
HETATM 8190  C1  LMT E 301     137.918 131.876 152.147  1.00 20.00           C
HETATM 8191  C2  LMT E 301     137.062 133.128 152.260  1.00 20.00           C
HETATM 8192  C3  LMT E 301     137.071 133.678 153.676  1.00 20.00           C
HETATM 8193  C4  LMT E 301     136.376 135.025 153.714  1.00 20.00           C
HETATM 8194  C5  LMT E 301     135.935 135.347 155.125  1.00 20.00           C
HETATM 8195  C6  LMT E 301     135.530 136.801 155.206  1.00 20.00           C
HETATM 8196  C7  LMT E 301     136.343 137.543 156.242  1.00 20.00           C
HETATM 8197  C8  LMT E 301     135.478 138.063 157.375  1.00 20.00           C
HETATM 8198  C9  LMT E 301     136.194 137.846 158.690  1.00 20.00           C
HETATM 8199  C10 LMT E 301     135.627 138.700 159.797  1.00 20.00           C
HETATM 8200  C11 LMT E 301     135.963 138.052 161.121  1.00 20.00           C
HETATM 8201  C12 LMT E 301     135.929 139.084 162.215  1.00 20.00           C
```

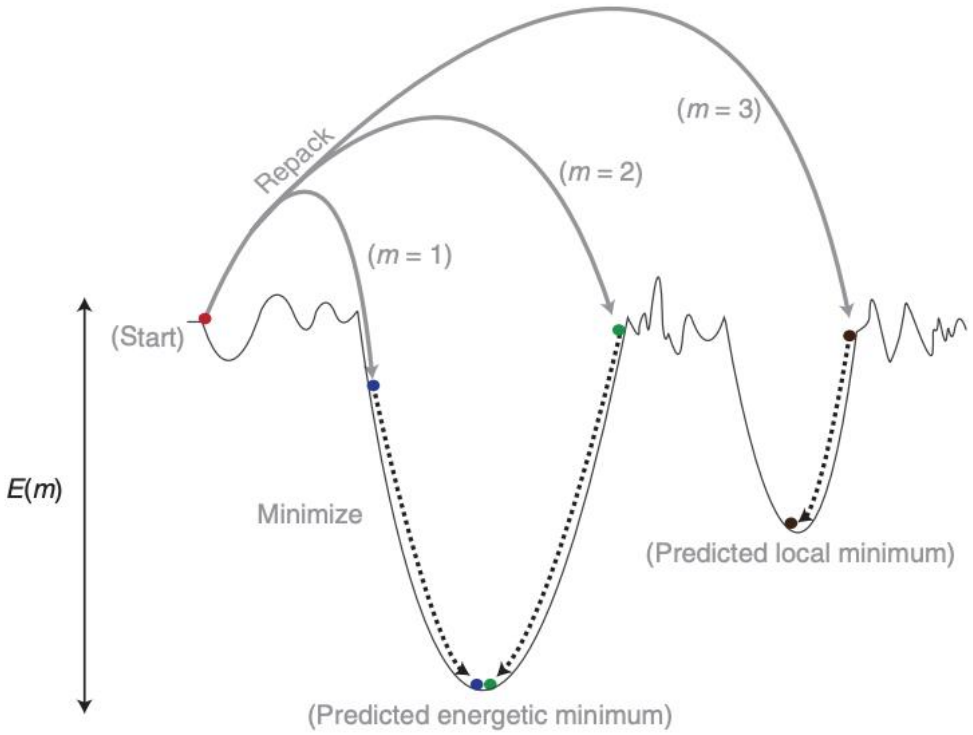
A. The first step is to clean the PDB (remove non-protein atoms)

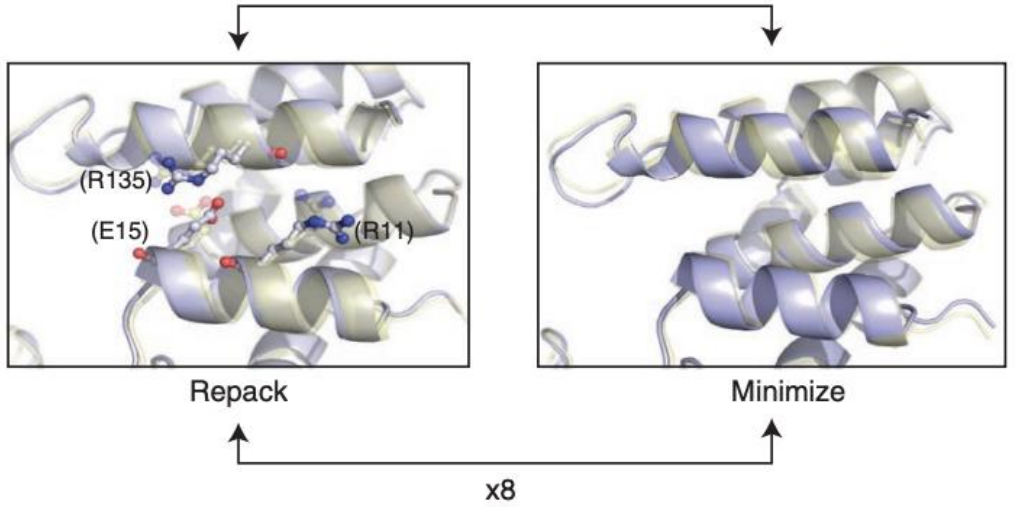
A. Clean PDB

```
ATOM   8178  CD2 LEU A 207     135.703 156.448 175.146  1.00 59.65           C
ATOM   8179  N   LEU A 208     138.542 160.482 176.163  1.00 64.39           N
ATOM   8180  CA  LEU A 208     139.558 160.550 177.204  1.00 64.39           C
ATOM   8181  C   LEU A 208     139.523 161.898 177.917  1.00 64.39           C
ATOM   8182  O   LEU A 208     140.107 162.065 178.988  1.00 64.39           O
ATOM   8183  CB  LEU A 208     140.944 160.306 176.607  1.00 64.39           C
ATOM   8184  CG  LEU A 208     141.162 158.912 176.018  1.00 64.39           C
ATOM   8185  CD1 LEU A 208     142.520 158.825 175.340  1.00 64.39           C
ATOM   8186  CD2 LEU A 208     141.011 157.842 177.088  1.00 64.39           C
ATOM   8187  OXT LEU A 208     138.909 162.851 177.438  1.00 64.39           O
TER    8188      LEU A 208
HETATM 8189  O1' LMT E 301     138.158 131.617 150.769  1.00 20.00           O
HETATM 8190  C1  LMT E 301     137.918 131.876 152.147  1.00 20.00           C
HETATM 8191  C2  LMT E 301     137.062 133.128 152.260  1.00 20.00           C
HETATM 8192  C3  LMT E 301     137.071 133.678 153.676  1.00 20.00           C
HETATM 8193  C4  LMT E 301     136.376 135.025 153.714  1.00 20.00           C
HETATM 8194  C5  LMT E 301     135.935 135.347 155.125  1.00 20.00           C
HETATM 8195  C6  LMT E 301     135.530 136.801 155.206  1.00 20.00           C
HETATM 8196  C7  LMT E 301     136.343 137.543 156.242  1.00 20.00           C
HETATM 8197  C8  LMT E 301     135.478 138.063 157.375  1.00 20.00           C
HETATM 8198  C9  LMT E 301     136.194 137.846 158.690  1.00 20.00           C
HETATM 8199  C10 LMT E 301     135.627 138.700 159.797  1.00 20.00           C
HETATM 8200  C11 LMT E 301     135.963 138.052 161.121  1.00 20.00           C
HETATM 8201  C12 LMT E 301     135.929 139.084 162.215  1.00 20.00           C
```

B. Relaxing the PDB for Rosetta calculations

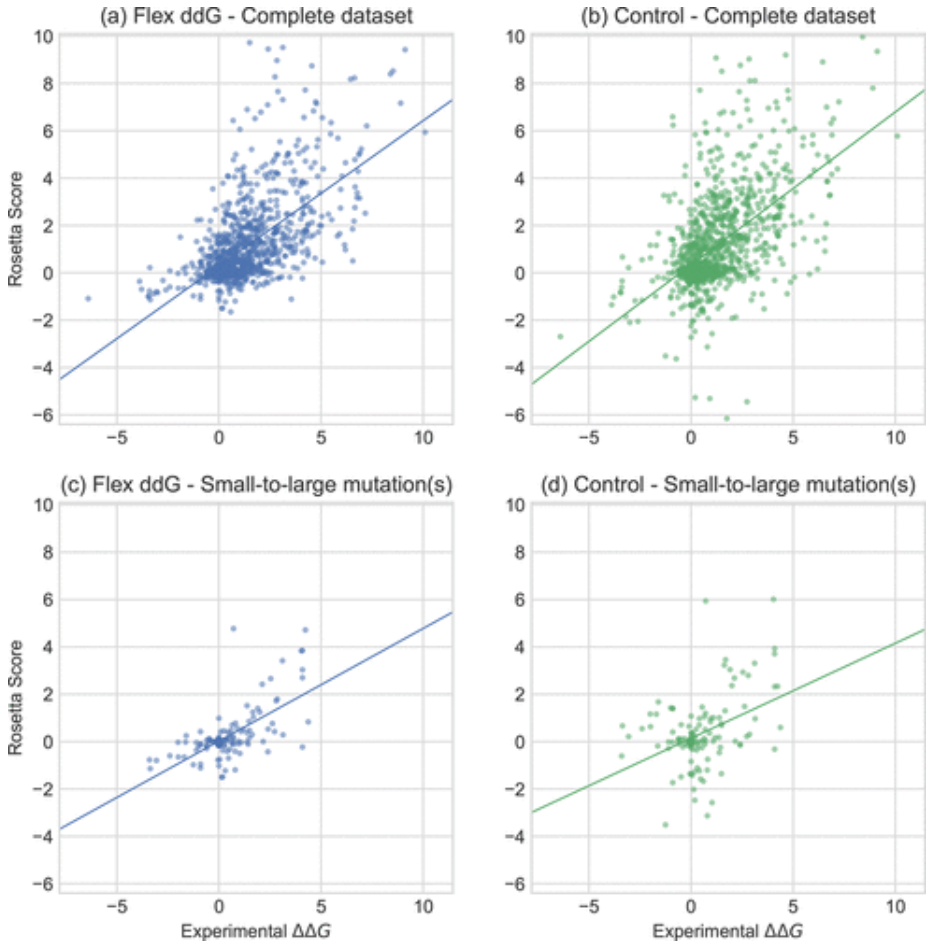
Rosetta FastRelax algorithm
~ simulated annealing



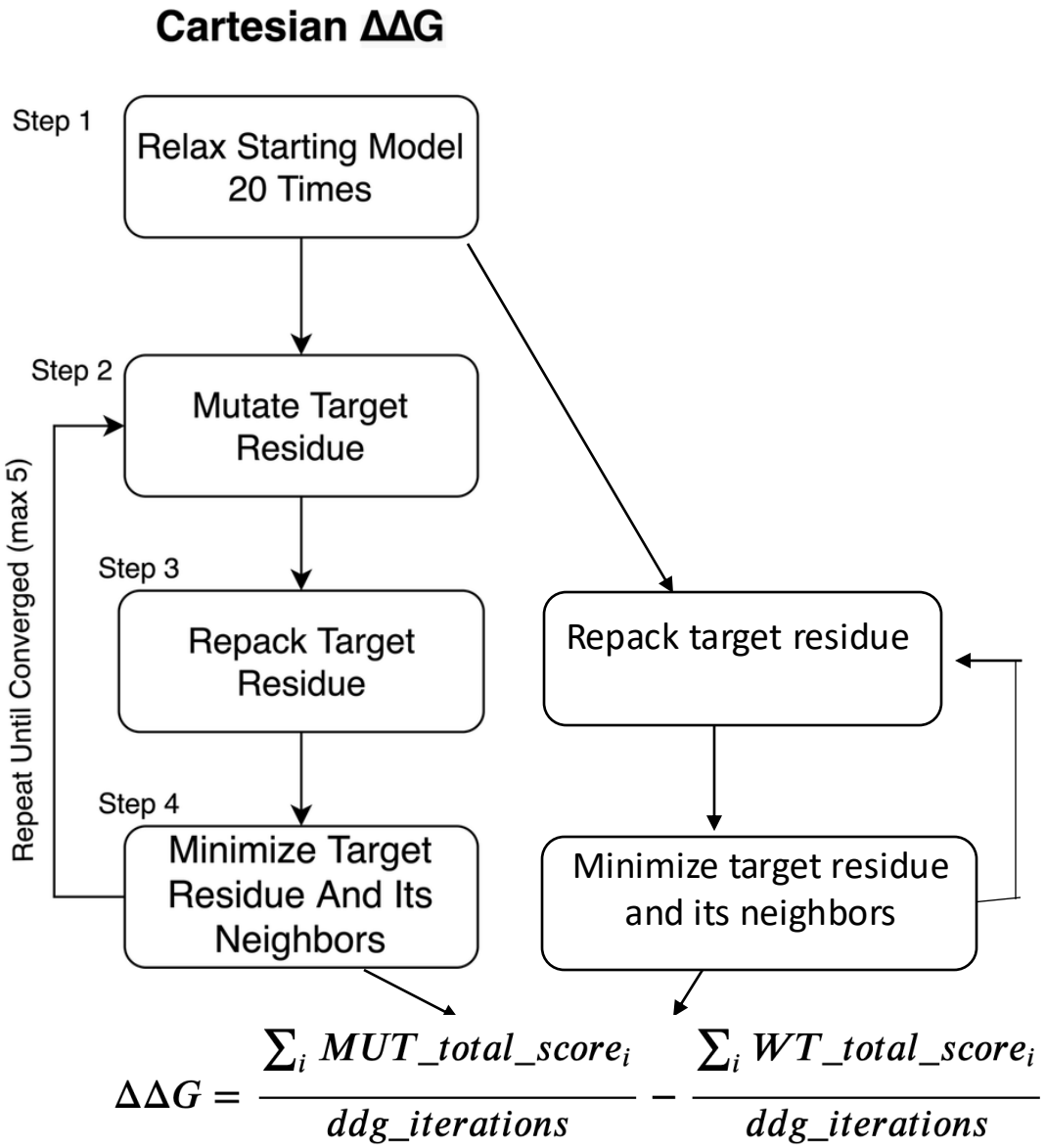


3. Running DDG calculations – Rosetta (energy model)

Rosetta's scorefunction was fitted to experimentally-determined DDGs



Benchmarks from Kortemme lab

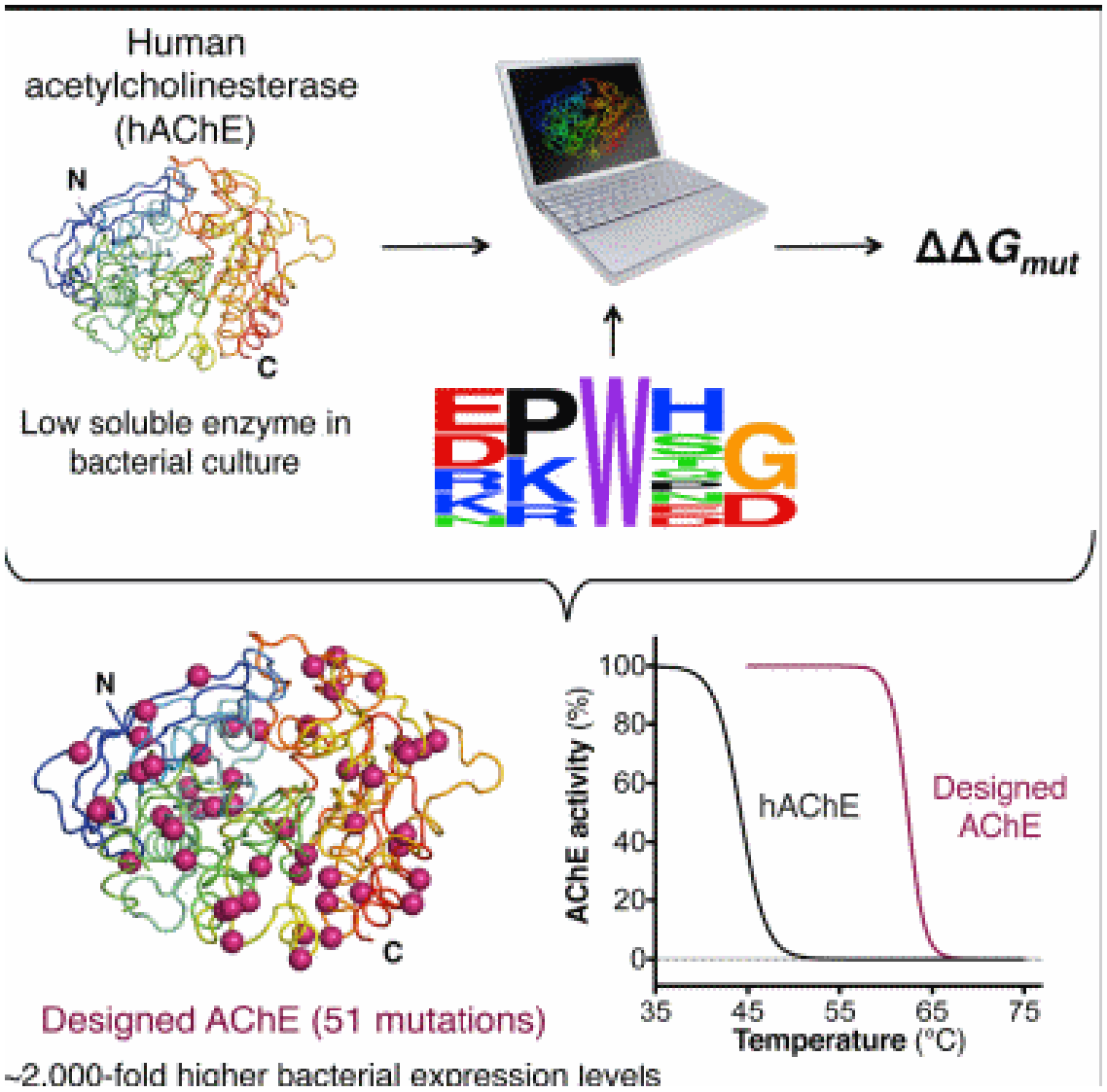


Q: Calculate the effect of mutations on protein-protein interactions?

Q: Calculate the effect of mutations on protein-protein interactions?

$$\Delta\Delta G_{interface} = \frac{\sum_{i=0}^{n} COMPLEX\_score_n}{n} - \frac{\sum_{i=0}^{n} SEPARARATE\_score_n}{n}$$

Running Combinatorial design - PROSS



Improves bacterial expression, stability and function of many eukaryotic (human) proteins

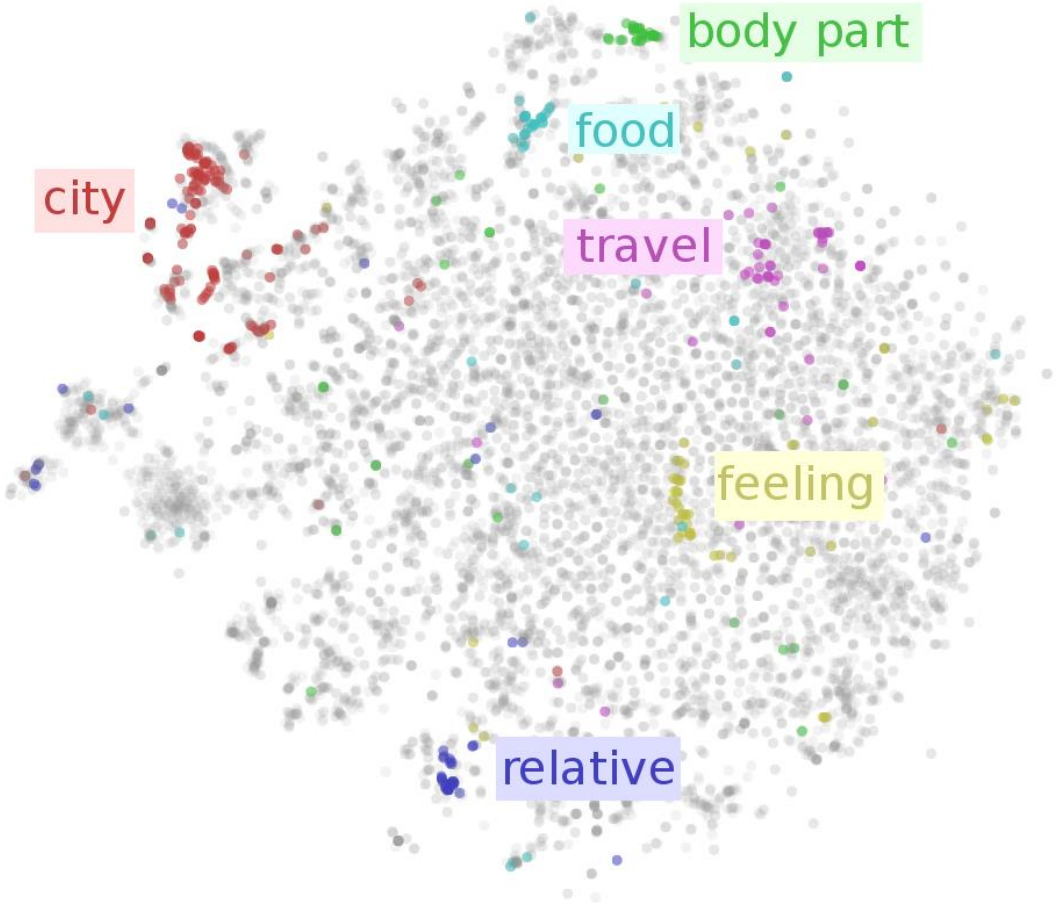
Paper: https://doi.org/10.1016/j.molcel.2018.08.033

Server: https://pross.weizmann.ac.il/

3. Large language models

E.g. ChatGPT

Words are embedded into a latent space based on **<u>semantics</u>**

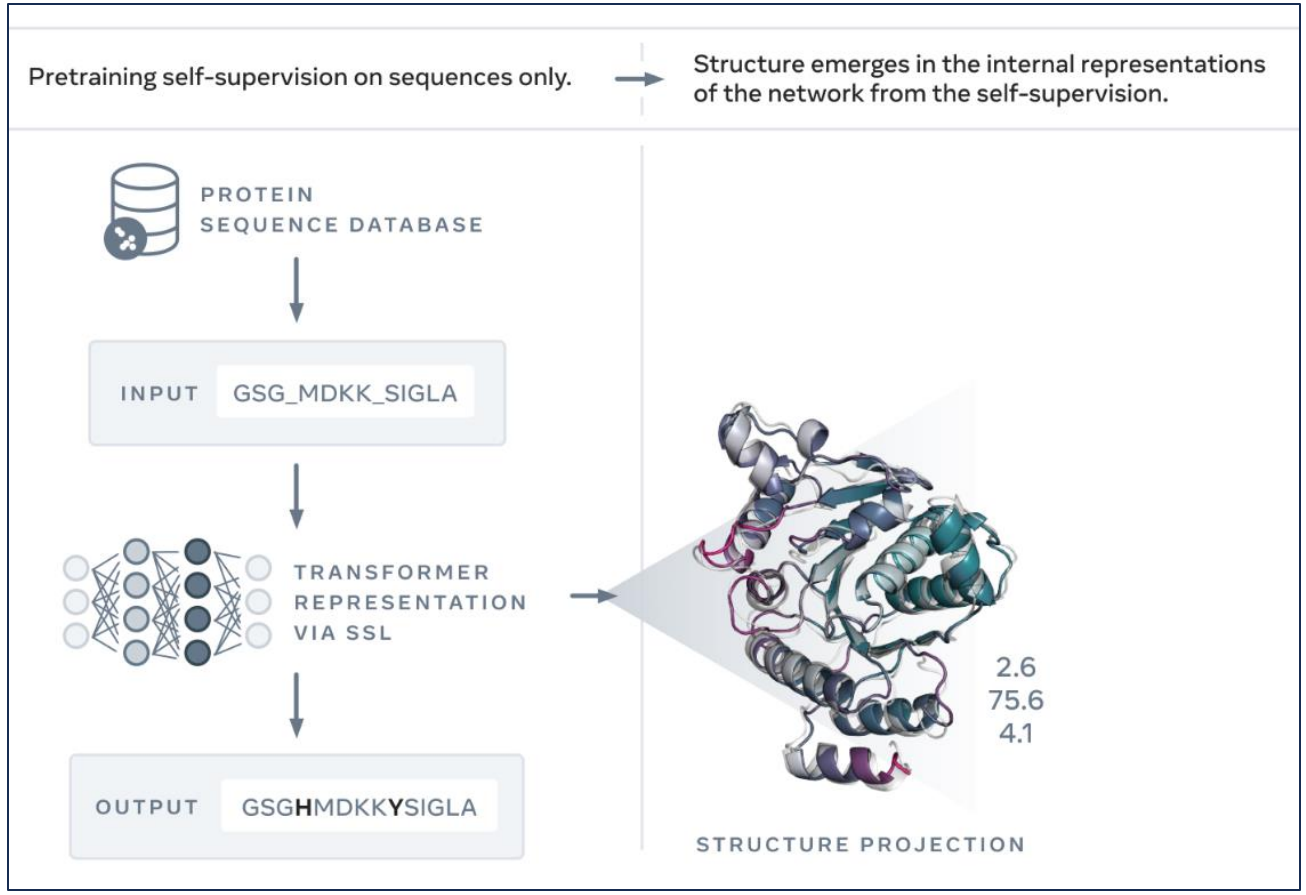


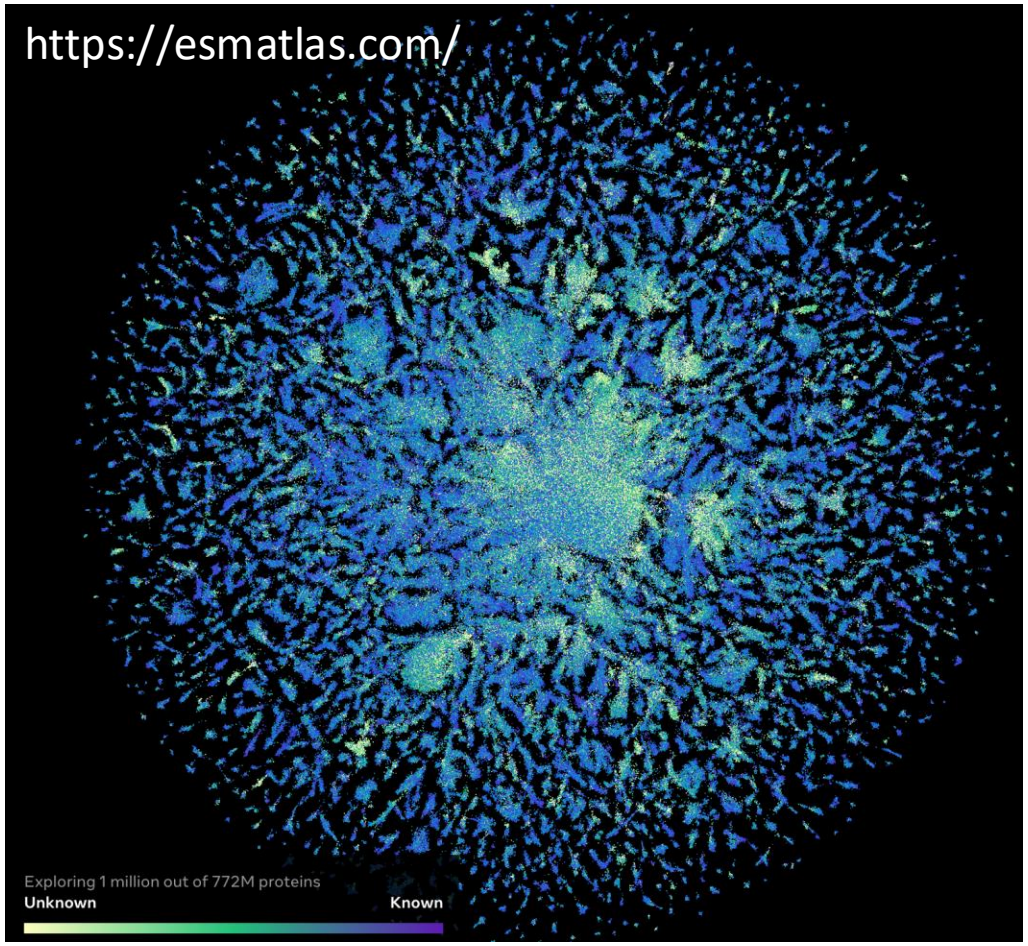
3. Protein language models

E.g. ESM models

Trained to capture evolution and structure relationships from large sequence databases

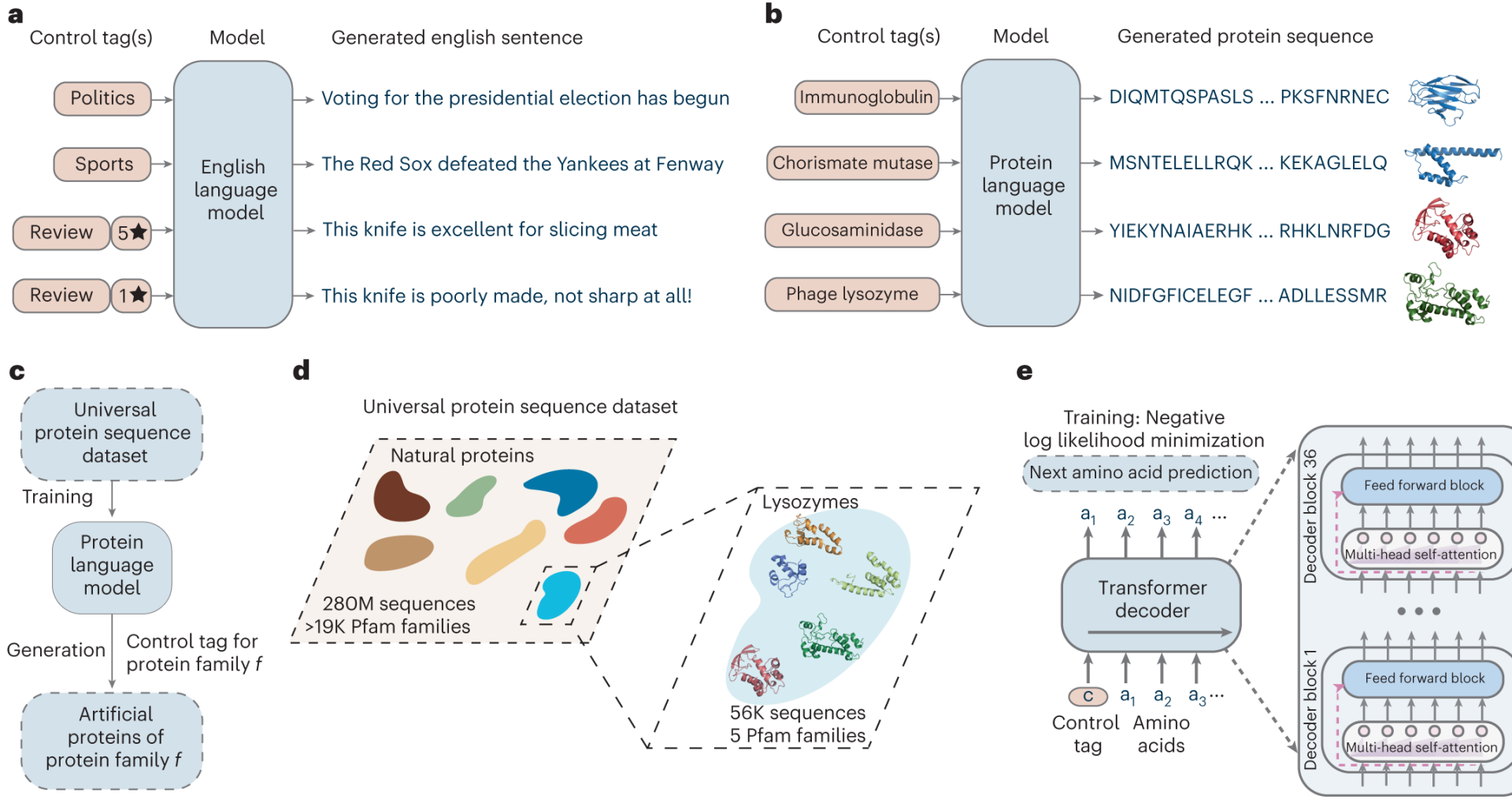
Enabled fast protein structure prediction (ESMFold) and building the ESM metagenomic atlas





DOI: 10.1126/science.ade2574

3. Running DDG calculations – LLMs (evolutive models)



A notebook to run a basic LLM fitness prediction will be sent after the class (high memory requirements)

https://doi.org/10.1038/s41587-022-01618-2

Next lecture: *de novo* protein design