Mini AlphaFold – Small-Scale Protein Structure & Drug Discovery AI

Track: VC big bets (Healthcare)

1. Motivation / Goal to Achieve

AlphaFold, which recently earned its creators a Nobel Prize, has transformed biology by predicting protein structures from amino acid sequences with remarkable accuracy. It's trained on massive datasets with huge compute budgets—an approach out of reach for most teams.

Goal:
This challenge flips the problem: instead of building a massive, general-purpose model, **create a lightweight AI system** that can predict or assist with protein structure–related tasks **using smaller datasets, fewer parameters, or domain-specific focus**. Your model should aim for a **realistic, narrowly scoped subproblem** in protein structure or drug discovery where small models can still have practical value.

2. Possible Subproblem Directions (Pick One)

1. **Secondary Structure Prediction**
    - Predict coarse protein folding states (alpha-helix, beta-sheet, coil) from sequence without doing full 3D folding.
2. **Ligand–Protein Binding Affinity Estimation**
    - Given a protein structure and a small molecule ligand, predict whether they bind strongly.
3. **Protein Family Classification**

Organizing team: Linn Bieske, Kai Wiederhold, Lisa Sklyarova, Nico Fröhlich

    - Classify a protein into known structural or functional families from its amino acid sequence.

4. **Small-Molecule Docking Scoring** *(stretch)*

    - Score candidate drug molecules for binding to a target protein, using simplified docking models.

3. Core Features (MVP)

1. **Data Ingestion & Preprocessing** – Load protein sequence/structure data from public datasets; clean, tokenize, and normalize it.

2. **Lightweight Model Training** – Use smaller architectures (CNNs, RNNs, transformers with ≤50M parameters) or fine-tune pre-trained embeddings (e.g., ESM-2, ProtBERT).

3. **Evaluation & Visualization** – Output predictions with confidence scores; visualize protein secondary structures or binding affinity graphs.

4. Stretch Goals (Optional)

1. **Generative Approach** – Explore diffusion models (e.g., stable diffusion adapted to 3D molecular structures) for generating plausible protein conformations.

2. **Web or Notebook Demo** – Interactive interface to input sequences and get predictions instantly.

5. Hints & Resources

Modeling Approaches

- Pre-trained protein language models:

    - **ESM-2** (Meta AI) – amino acid sequence embeddings.

    - **ProtBERT** – transformer trained on UniProt sequences.

- Small CNN/RNN classifiers for secondary structure prediction.

- Diffusion models for molecular generation:

    - DiffDock (molecule docking), ProteinMPNN (protein design).

Datasets

- **Protein Data Bank (PDB)** – experimentally determined 3D protein structures.
- **CB513** – curated secondary structure benchmark.
- **BindingDB** – binding affinities for protein–ligand pairs.
- **ChEMBL** – bioactive molecules with drug-like properties.
- **SCOPe** – protein structural classification.

Visualization Tools

- PyMOL, UCSF Chimera, Mol* viewer for 3D structures.

Compute-Friendly Options

- Limit to short sequences or small protein families.
- Use transfer learning from embeddings to avoid training from scratch.

6. Evaluation Criteria

- Feasibility: Model runs within hackathon compute limits.
- Accuracy: Competitive results on a held-out test set for the chosen subproblem
- Innovation: Creative use of small models or pre-trained embeddings to achieve meaningful results.
- Usability: Clear visualization and interpretability of outputs.

7. Why It Matters

Drug discovery timelines are long and costly. While large-scale AlphaFold-like models push the science forward, **smaller, specialized models** can democratize protein research—letting startups, academic labs, and even hackathon teams tackle meaningful biology questions on modest compute. Mini AlphaFold challenges participants to show how **lean AI approaches** can still deliver breakthroughs in healthcare innovation.

Organizing team: Linn Bieske, Kai Wiederhold, Lisa Sklyarova, Nico Fröhlich