Some proteins can be **translated** into an alphabet of **10** amino acids and still keep the **same structure**.

Predicting Protein Folding in Reduced Alphabet Protein Sequences

Hugo Hrbáň, David Hoksza
hugohrban2@gmail.com, david.hoksza@matfyz.cuni.cz
Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Background

- It is hypothesized, that over **4.5 billion years ago**, proteins relied on an alphabet of only **ten**, so-called ***early*** or ***prebiotic*** amino acids (*ADEGILPSTV*), and the rest of the standard alphabet developed later through biosynthetic pathways.
- Our project focuses on two objectives.

1. Analysis of a combinatorial **library** of early-alphabet proteins. Each sequence had one of three labels: *folded*, *misfolded* or *not expressed*.
   - We searched for structural homologs, but found **no correlation** between the hits and the experimental labels.
   - We attempted to train a **classification language model** that could predict foldability and expression.
   - The data proved **too noisy** for the model to learn this task.
2. **Main objective** was therefore **to find out** if the ***early*** alphabet can create **proteins with similar structure to modern proteins**, while using *ESMFold* to predict folding of ***early*** proteins.
   - Details of the procedure follow.

Methods

Iterative replacement of late amino acids with early ones. The following replacement strategies were explored:

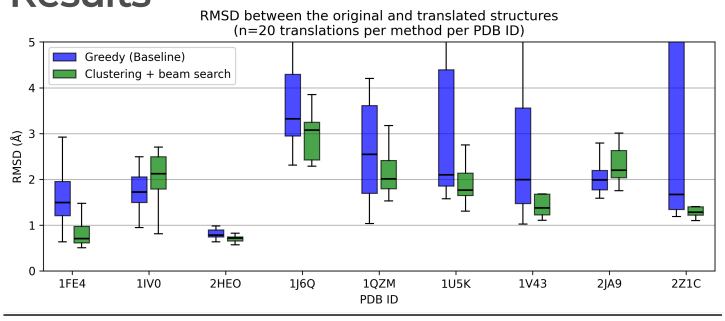
Greedy Translation (Baseline)

- **Randomly** select a **position** with a late residue to mutate.
- **Predict folds** of all 10 possible variants using *ESMFold*.
- **Select** the one with **lowest *RMSD*** to the original structure.
- Repeat until **all** residues are ***early***.

Clustering + Beam Search

- Create **contact map** from late C-β coordinates in PDB structure.
- **Hierarchical clustering** of *late* residues.
- For each cluster, do **Beam Search**:
  - **Candidates** ← All possible mutations on all possible positions.
  - **Evaluate** all candidates (weighted combination of *RMSD* and model confidence – *pLDDT*).
  - **Keep top w** candidates.
  - Repeat until **all** residues **in cluster** are ***early***.

Results



PDB ID	Baseline (Greedy)			Clustering + Beam Search		
	RMSD (Å)	pLDDT	TM-score	RMSD (Å)	pLDDT	TM-score
1FE4	1.50	66.2	0.86	**0.71**	**92.2**	**0.95**
1IV0	**1.72**	75.2	0.68	2.12	**78.7**	0.68
2HEO	0.78	89.2	0.94	**0.72**	**92.9**	**0.95**
1J6Q	3.32	62.5	0.78	**3.08**	**82.9**	**0.82**
1QZM	2.59	67.2	0.85	**2.08**	**70.0**	**0.89**
1U5K	2.10	74.4	0.85	**1.76**	**78.8**	**0.88**
1V43	1.99	84.2	0.87	**1.40**	**86.8**	**0.89**
2JA9	**1.99**	67.5	**0.90**	2.20	**70.2**	0.88
2Z1C	1.67	85.8	0.86	**1.28**	**89.6**	**0.87**

Values in the table are median out of n=20 translations.

Observations

- For most structures, beam search is more accurate and robust; Larger beam size (w) yields better structures, but longer time.
- For structures with **unobserved residues** (*1U5K*, *1V43*), we optimize RMSD with respect to the initial prediction of ESMFold.
- For larger structures, translation can take several minutes (***1V43*** length = 372: **Greedy**: *20* min., **Beam search** (w=5): *85* min. on H100 GPU).

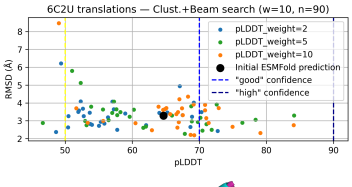
Design Candidate – Case Study

Original:

- PDB ID: **6C2U**
- **P-loop** (Walker-A motif) protein
- **ATP**-binding
- Length: **115**
- Num. late residues: **30**

Translation:

- RMSD: **2.2 Å**
- TM-score: **0.82**
- pLDDT: **79.1**



```
MRVIVVIVGPSGAGKTTLDELARKAKEEVPDAEIRTVTTKEDAKRVAEEAERRNADIVVIVGPSGSGKSTLAKIVKKIIARAGAKTIEVTTEEELRKAVAKARGSWSLEHHHHHH
 ||||||||||||| |||||||  | |||||||| |||| ||| ||||||  |||||||||||||| |||| ||  ||| ||| ||||||||||  ||| | || |||
ITVIVVIVGPSGAGPTTLDELAAPASEEVPDAEILTVTTDEDAIEVAEEAESAAADIVVIVGPSGSGPSTLADIVSIIIATAGAVTIEVTTEEELPGAVADAEGSVSLEAAEITEV
```

Summary

- We developed an **algorithm** to iteratively **translate** a protein from the standard alphabet into the ***early*** alphabet of **10** residues while keeping the original structure.
- The ***early*** alphabet can support **modern folds**.
- **Beam search** and **structure prediction** are used as heuristics to explore the sequence space.
- **Dataset** based on ongoing research by *Hlouchová et al.*

This work was supported by the Czech Science Foundation (GAČR) grant number 23-07349S, and the ELIXIR CZ Research Infrastructure (ID LM2018131, MEYS CR).

Scan to see visualizations of the translation process! →

hugo.matfyz.cz/eccb2025