Machine Learning for Dynamics-Aware Protein Sequence Design

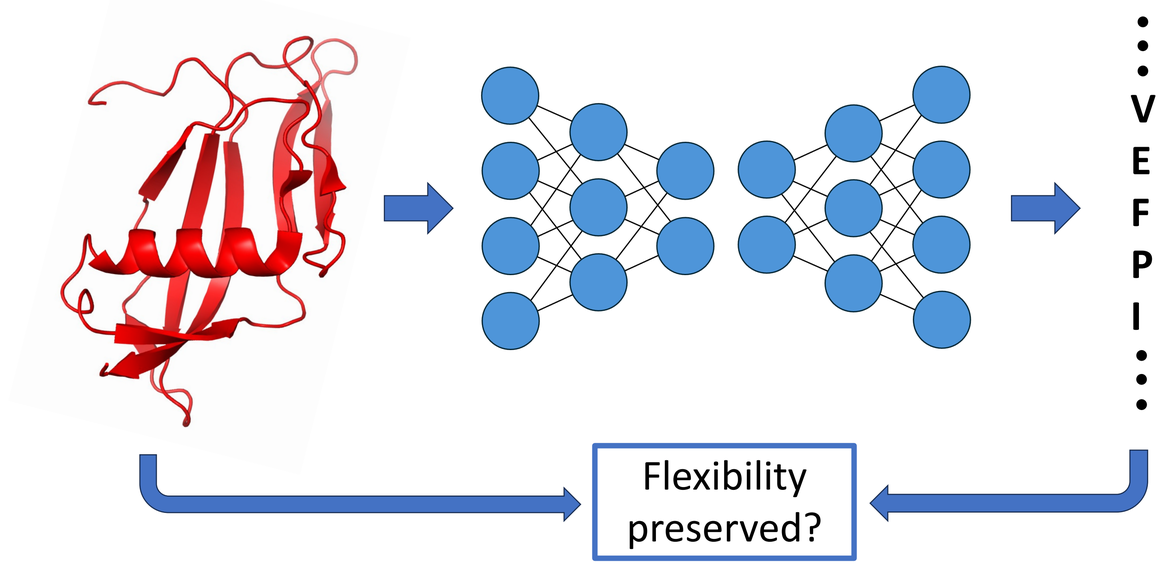
Petr Kouba, Joan Planas-Iglesias, Jiří Damborský, Jiří Sedlář, Josef Šivic , Stanislav Mazurenko

Motivation

The problem of finding a protein sequence for a required protein backbone structure, also known as **'inverse folding'**, is of high importance for protein design and protein engineering. It can be used to find **alternative sequences for existing protein structures**, helping the researchers navigate the space of protein mutations by identifying those that do not alter the structure. Recent methods, such as ProteinMPNN [1], tackled the core problem with success. However, the experimental validation of these methods brought a new question. **Do the inverse folding models overoptimize for the recovery of the original structure at the cost of losing the protein's native flexibility?** Apart from structure, **dynamics** is another important characteristic of a protein which should be accounted for [2]. This ability of proteins to change their structure is essential especially for most **enzymes** as it allows them to interact with the substrate and **perform biochemical reactions**.

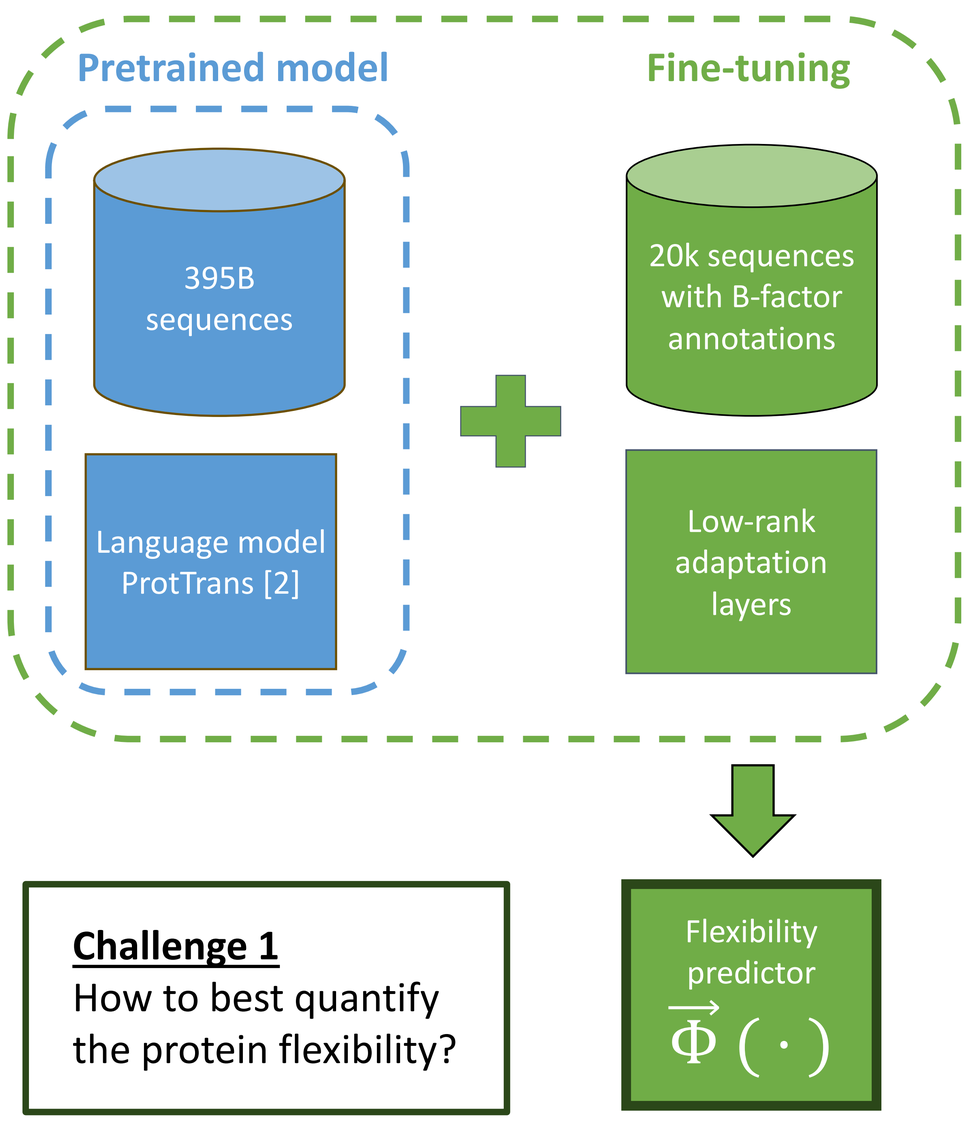
Goals

- Study the state-of-the-art inverse folding models and their capabilities to account for protein flexibility.
- Develop an inverse folding model capable of generating protein sequences according to prescribed flexibility.



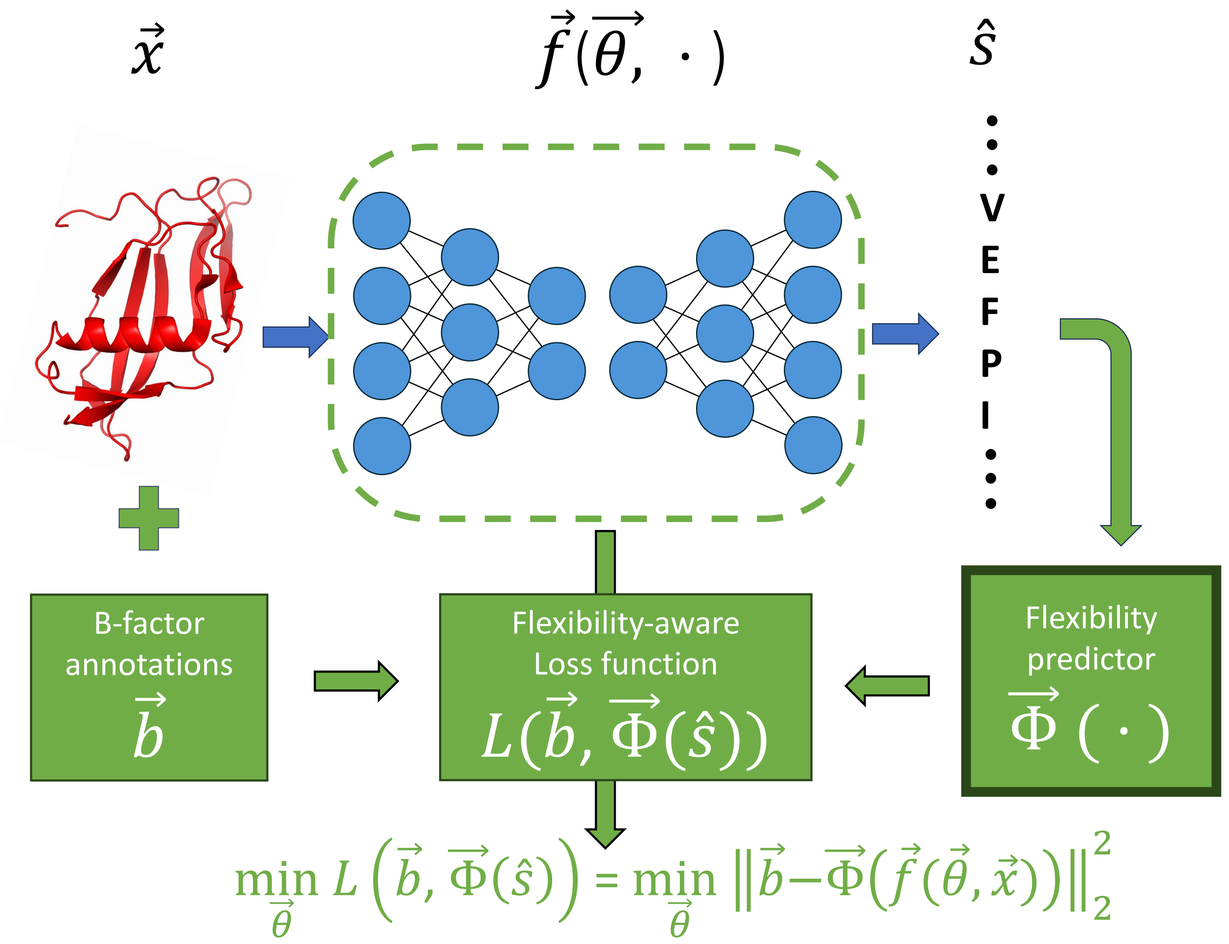
Approach

1) Train a sequence-to-flexibility predictor



Challenge 1
How to best quantify the protein flexibility?

2) Use the flexibility predictor to guide protein sequence design



Challenge 2
How to sample the sequence differentiably for end-to-end training?

Preliminary results

Challenge 1: Correlation of true and predicted flexibility using different representations of the residue flexibility

Flexibility representation	Pearson R	Spearman R
Raw C_α B-factors	0.47	-
Normalized C_α B-factors	0.56	0.54
Norm. SSE-avg. C_α B-factors	0.56	0.55

Challenge 2: Using Gumbel-Softmax to differentiably sample from the logits of ProteinMPNN enabled end-to-end training. The table shows first results on sequence recovery and flexibility preservation.

Fine-tuning	Seq. Recovery	Flexibility corr. (Pears./Spear.)
Batch size 1	0.49	0.52 / 0.54
Batch size 8	0.49	0.49 / 0.52

Next steps

- Demonstrate computationally that the flexibility is retained
- Apply the model to the design of the Staphylokinase protein for the wet lab validation of the method

References:
1. Dauparas et al. Robust Deep Learning–Based Protein Sequence Design Using ProteinMPNN. Science (2022)
2. Kouba et al. Machine Learning-Guided Protein Engineering. ACS Catalysis (2023)
3. Elnaggar et al. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. TPAMI (2022)

MUNI | RECETOX · LOSCHMIDT LABORATORIES · FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE · CZECH INSTITUTE OF INFORMATICS ROBOTICS AND CYBERNETICS CTU IN PRAGUE · LUMI · VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER · Brno Ph.D. Talent