PROTEIN LANGUAGE MODELS: WHAT IS EVEN A "LANGUAGE"?

PLMs for Predicting Mutation Stability Changes

Author: Jakub Vlk, Supervision: Miloš Musil

Motivation & Context

Proteins are masterpieces of biochemistry and quantum mechanics, yet our understanding is stalled by a critical shortage of any data. Experimental measurements of protein stability are not only resource-heavy but also often suffer from high variance and poor replicability, making it hard to combine them. This scarcity has ignited a high-stakes race to build robust prediction tools capable of navigating these inconsistencies to unlock the next frontier of biological design. We present a fine-tuned PLM that leverages massive, curated datasets to overcome these limitations and achieve precise stability prediction

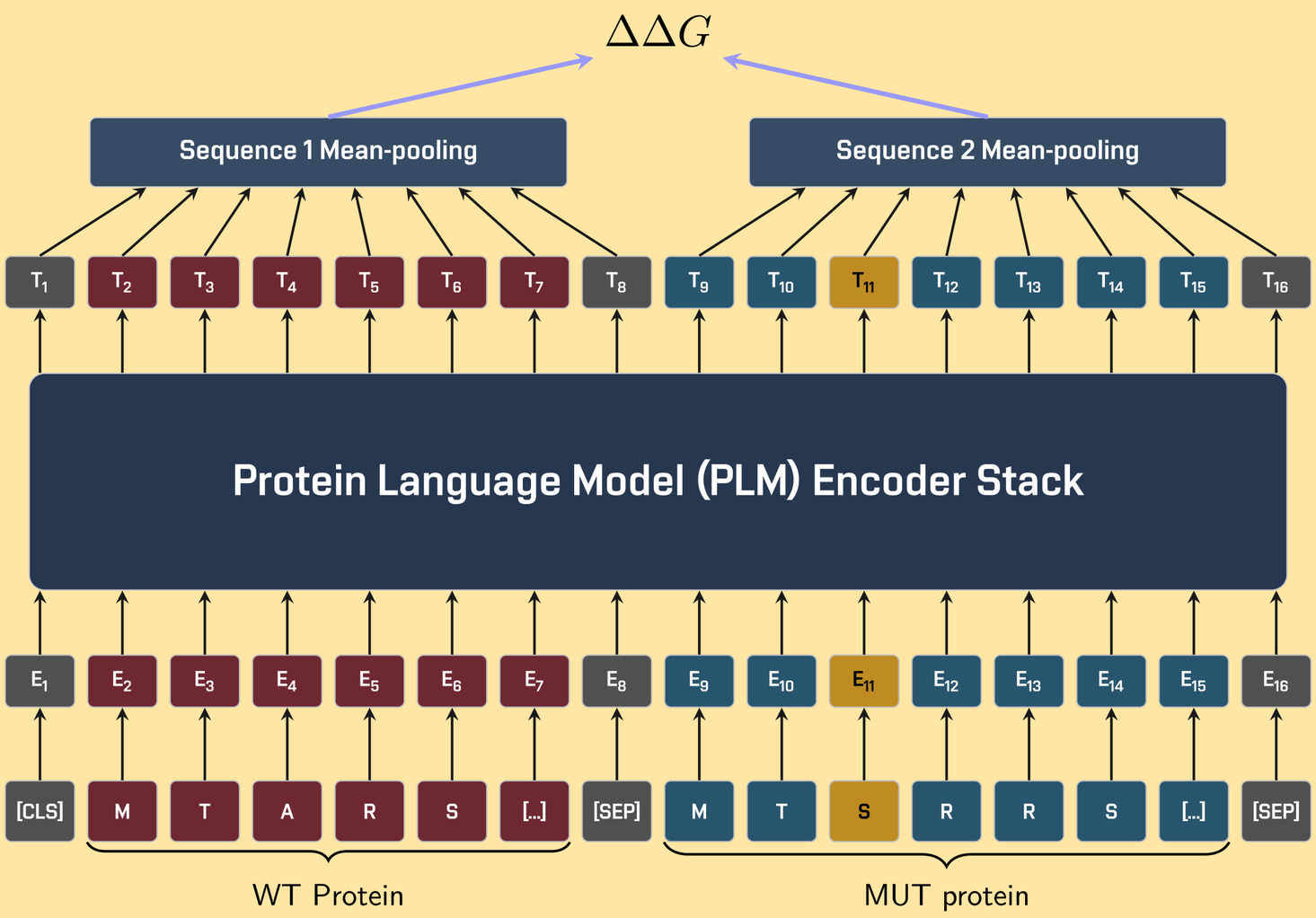


Fig. 1: Architecture's scheme overview

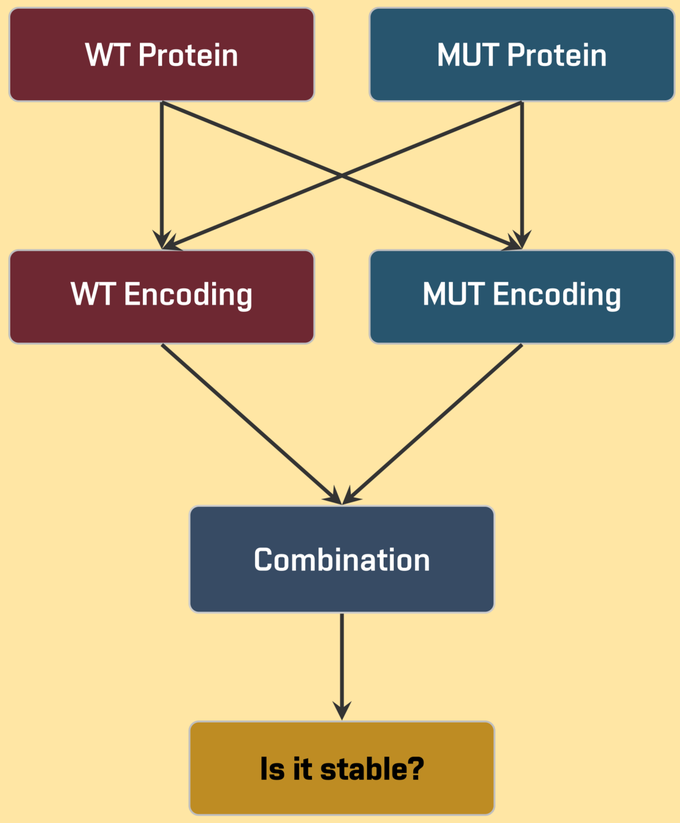
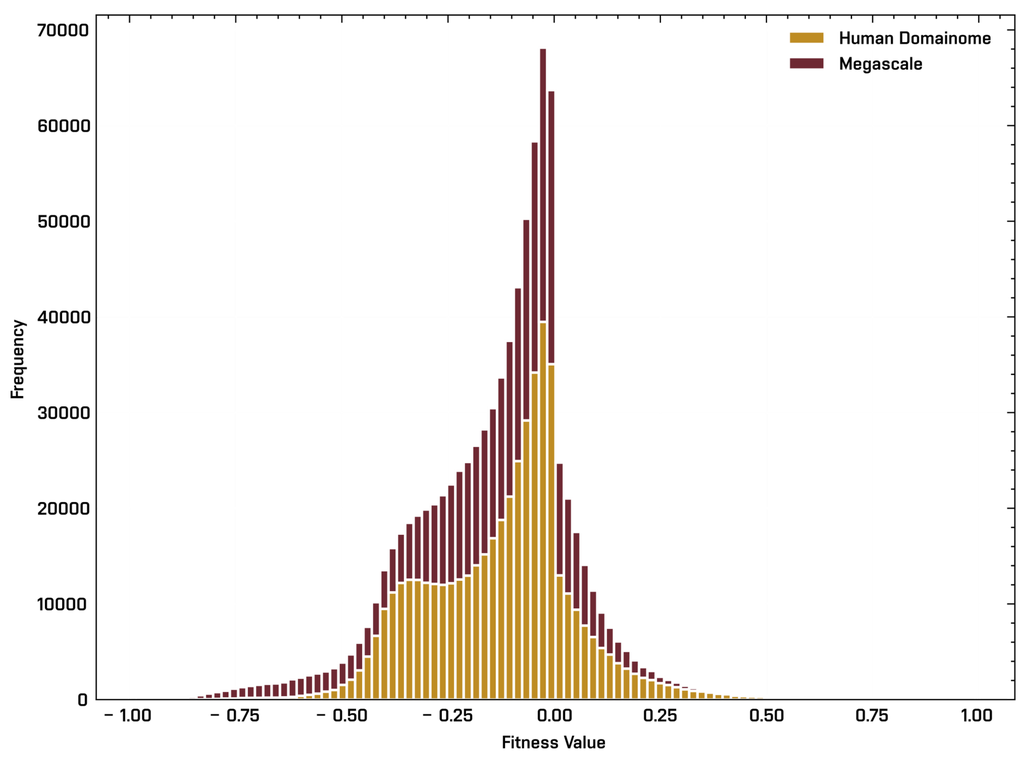


Fig. 3: Processing pipeline scheme.

Dataset

Fig. 2: Distribution of Fitness Values.



Total size: 864 033

Protein length (min/avg/max): 44 / 547 / 8,525

Stabilizing share: 20%

Results

Tool	Pearson	MCC	Ter. MCC
Prime	0.51	0.32	0.24
ThermoMPNN	**0.64**	**0.45**	**0.47**
ProtBERT*	0.31	0.05	0.19

Tab. 1: S350 results.

Tool	Pearson	MCC	Ter. MCC
Prime	0.10	0.13	0.16
ThermoMPNN	0.38	**0.35**	**0.30**
ProtBERT*	**0.57**	-0.14	0.19

Tab. 2: Benchstab results.

*Developed by authors of this poster.

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Excel@FIT 2026