Protein Language Models in Directed Evolution

Russell Maguire, Kotryna Bloznelyte, Fikayo Adepoju, Matthew Armean-Jones, Shafiat Dewan, Akash Gupta, Frances Patricia Jones, Preet Lalli, Anna Schooneveld, Ece Ibrahimi, Stella Fozzard, David Berman, Luca Rossoni, Will Addison, Ian Taylor

Abstract

Background: Protein Development

- Key area in biotechnology
- Traditional methods are laborious and inefficient
  - Experimental directed evolution may require >10 000 screening samples
- ML-integrated protein engineering holds promise in *de* protein design and guided directed evolution
- Most ML models have lacked robustness

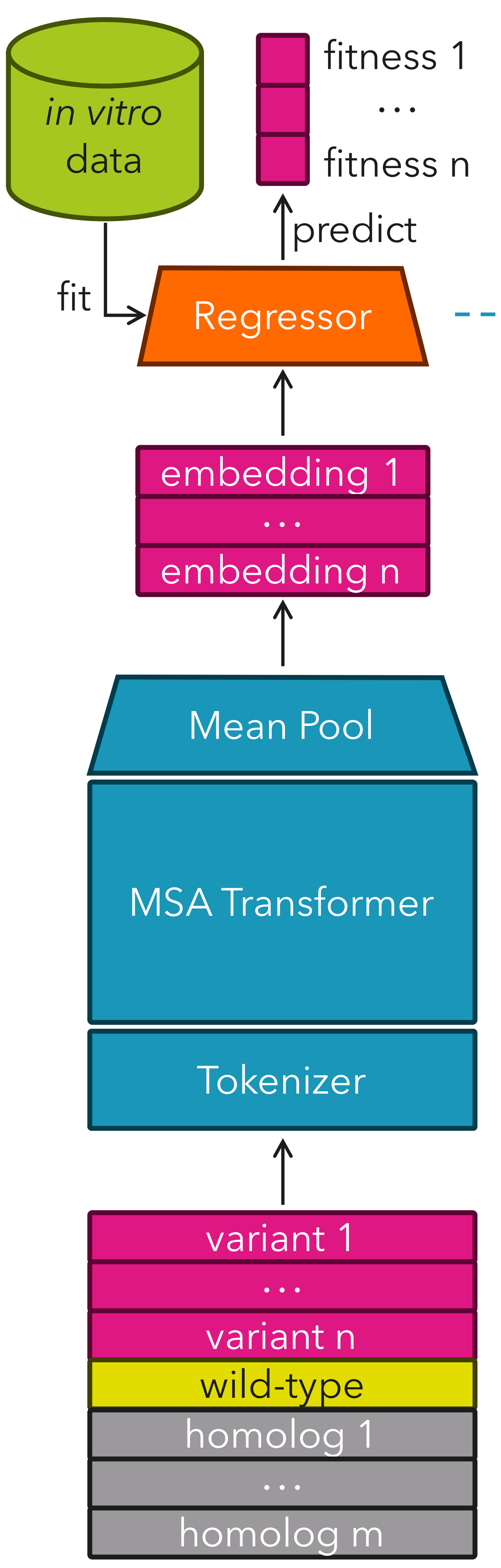
Our approach: novel framework + *in vitro* data + validation

- Few-shot protein language modelling
  - Adapt MSA transformer to include *in vitro* fitness measurements
- Extensive model evaluation & variant characterisation *in vitro*
- Improved *in silico* sequence predictions for high-performing variants
  - **62%** improved PET degradation over starting variant
  - **16%** improved PET degradation over current state-of-the-art

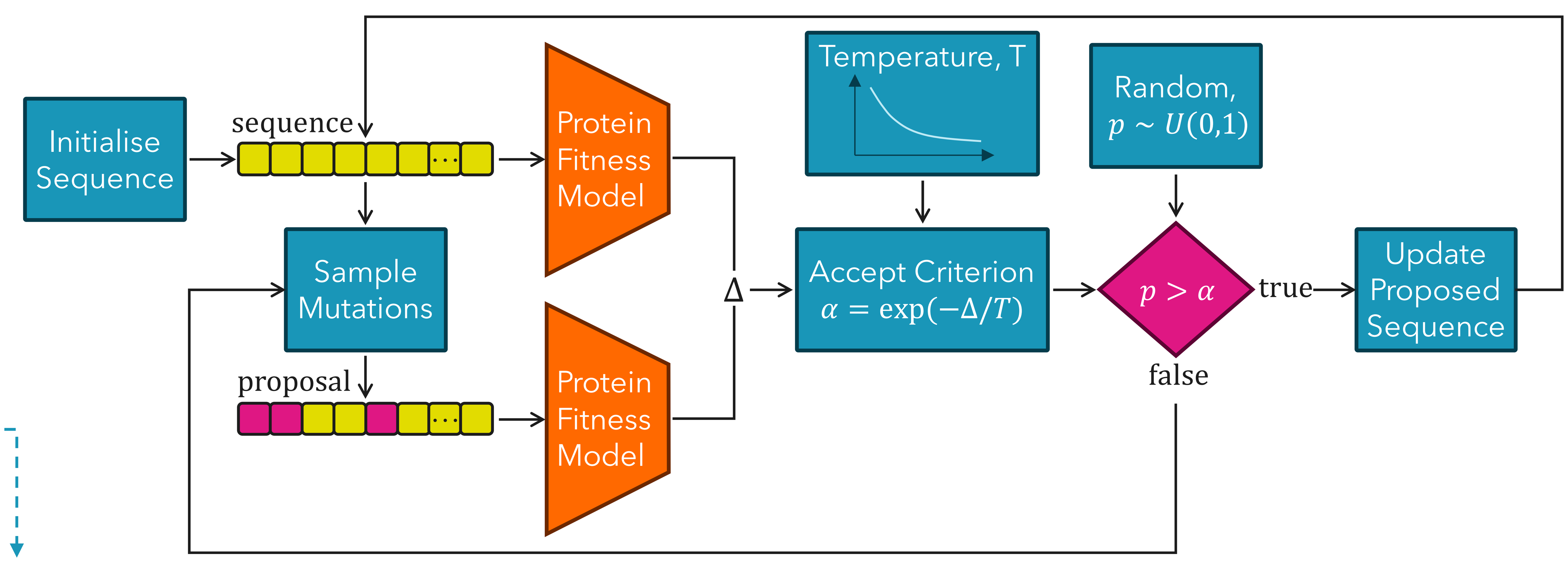
In Silico Mutagenesis via Simulated Annealing

Our Protein Fitness Model

A small dataset of protein fitness measurements are regressed from the final embeddings of MSA transformer



Traditionally, one scores mutations by computing the log odds ratio between the mutant and wild-type. This scales the fitness prediction relative to wild-type.



Experimental Setup

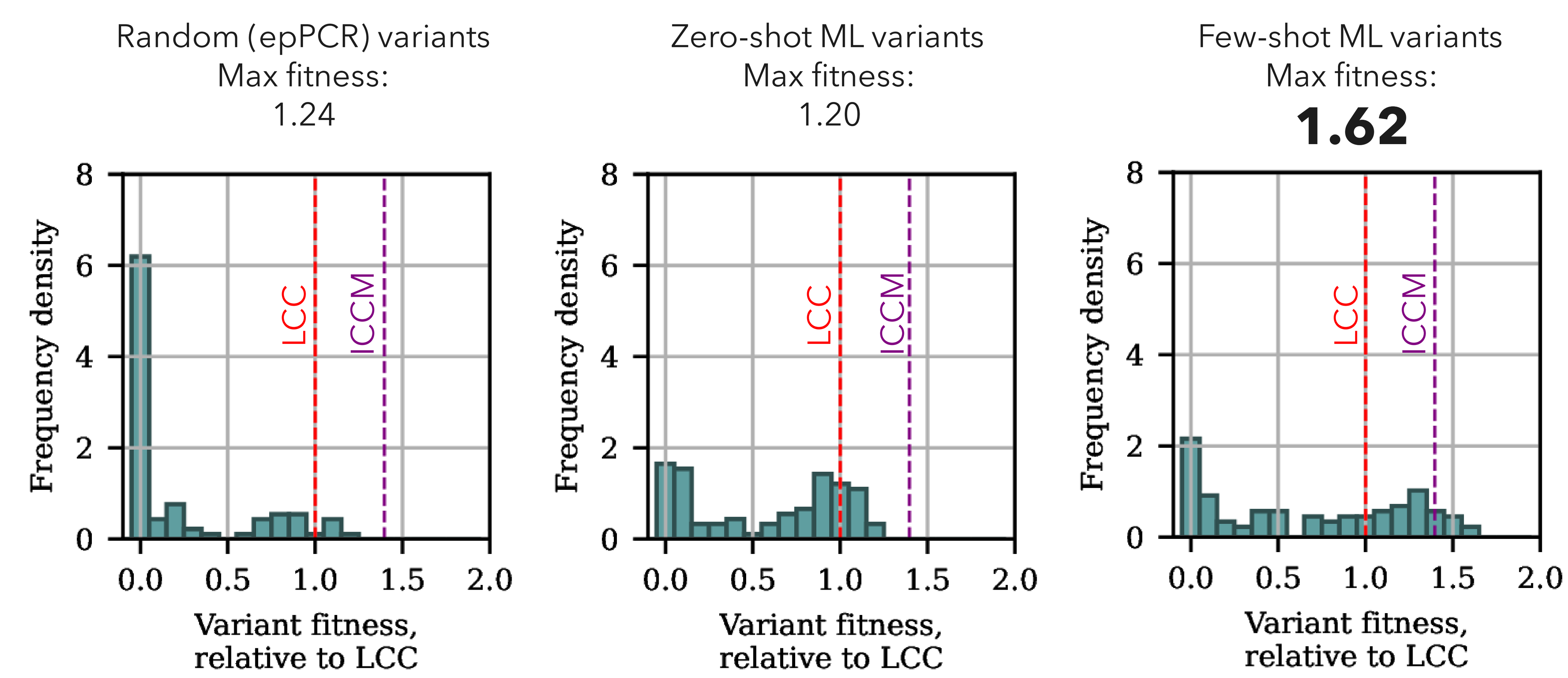
- Starting protein: leaf-branch compost cutinase (LCC)
- PET degradation assays: cell lysate was incubated with semi-crystalline PET for 72 h. Reaction products were quantified by HPLC and used as a measure of enzymatic activity.
- Thermostability assays: colorimetric reporter assay was used to measure enzymatic rates before and after heat treatment (2 h at 70°C). Thermostability was calculated as the ratio of heat-treated to non-heat-treated activity.

References

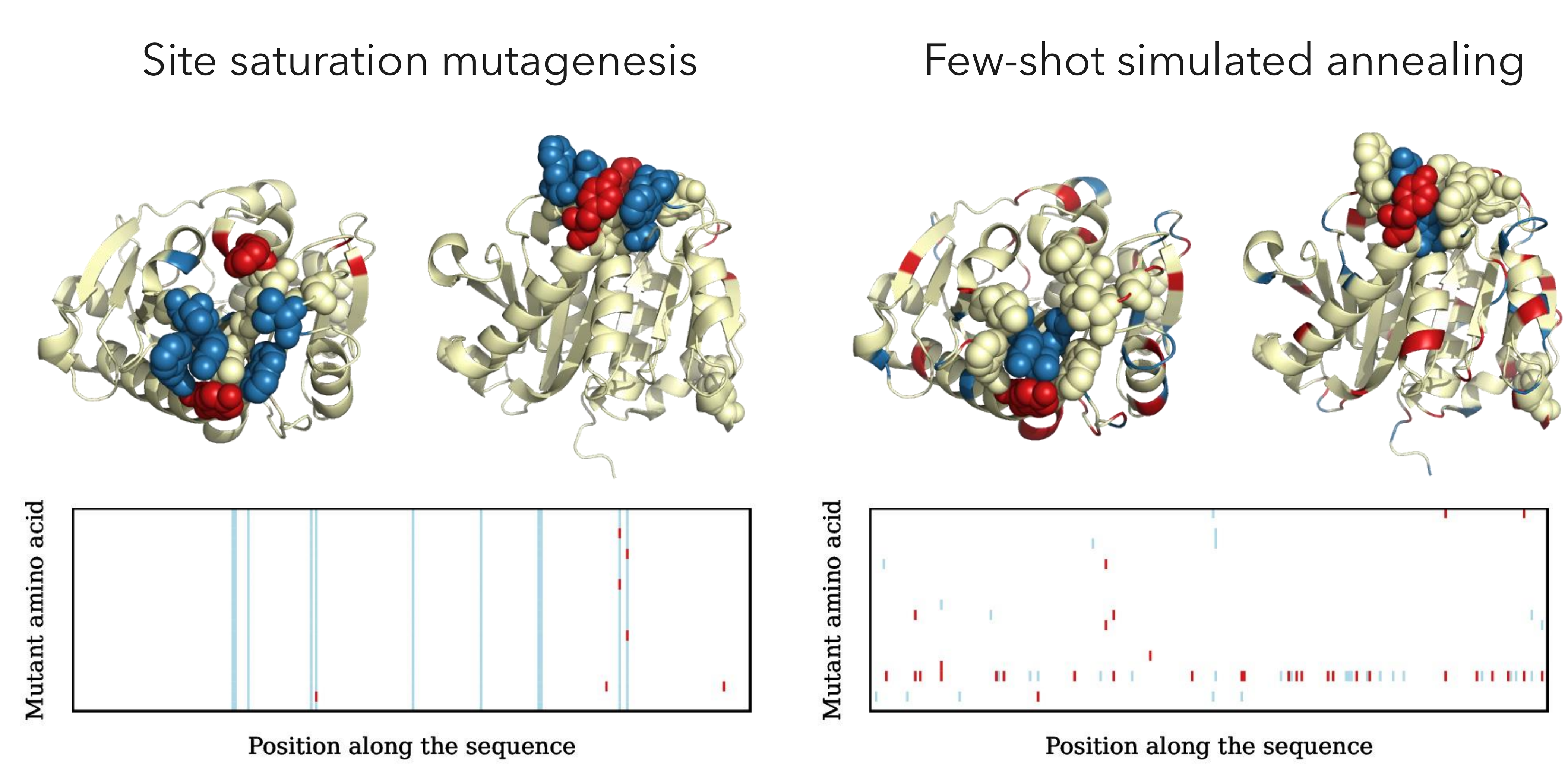
- Tournier V et al., Nature 2020;580(7802):216-219. doi:10.1038/s41586-020-2149-4
- Sulaiman S et al., *PDB*, 4EB0, 2012b. doi:10.2210/pdb4Eb0/pdb
- Rao R et al., bioRxiv 2021.02.12.430858; doi:10.1101/2021.02.12.430858

cambridge consultants
Part of Capgemini Invent

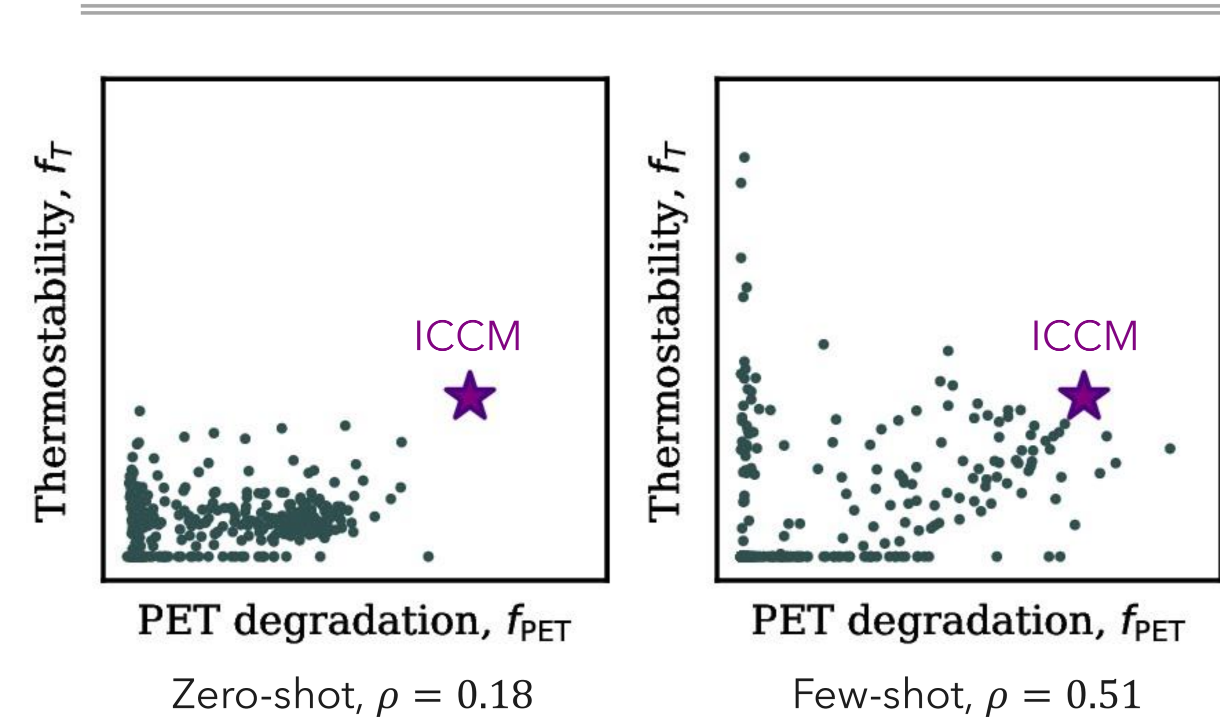
Results



ICCM is a high-performing variant engineered through site-directed mutagenesis



Mutation locations found in variants fitter than wild-type are shown in red; mutation locations found in variants less fit than wild-type are shown in blue. Highlighted on the crystal structure of LCC obtained from the protein databank, PDB ID: 4EB0



Conclusions

- **High success rate**: 39% of few-shot variants show improved fitness
- **Wide exploration** of protein space & scope for further improvement by recombining mutations
- Few-shot learning enables model tailoring to **diverse phenotypes**
- **High sample efficiency**: 240 *in vitro* examples used in 2nd round training