# Designing active and thermostable enzymes with sequence-only predictive models

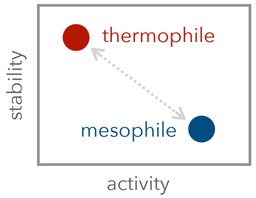

Clara Fannjiang[1,2], Micah Olivas[3], Eric R. Greene[4], Craig J. Markin[5], Bram Wallace[2], Ben Krause[2], Margaux M. Pinney[6], James S. Fraser[4], Polly M. Fordyce[3,7], Ali Madani[8], Nikhil Naik[2]

[1]Dept. of Electrical Engineering & Computer Sciences, UC Berkeley [2]Salesforce Research [3]Dept. of Genetics, Stanford University
[4]Dept. of Bioengineering & Therapeutic Sciences, UCSF [5]Dept. of Biochemistry, Stanford University [6]Dept. of Biochemistry & Biophysics, UCSF
[7]Dept. of Bioengineering, Stanford University [8]Profluent Bio


**How can we use predictive models of fitness to design proteins**
**(i) that satisfy *multiple* properties**
**(ii) when these models are not always trustworthy?**

## Case study: designing active, more thermostable enzymes

- broadly applicable goal, *e.g.* for industrial applications
- natural enzymes often exhibit trade-off
- existing methods: PROSS, consensus




## Our general approach

$f_i, i = 1, \dots m$ : predictive model of $i$-th fitness function
$\text{TrustRegion}_i, i = 1, \dots, m$ : region of sequence space on which we trust $f_i$

We use a Metropolis-Hastings algorithm to sample novel sequences from the distribution

$$p^\star(x) \propto \begin{cases} \exp(\sum_{i=1}^m \lambda_i \cdot f_i(x)) & \text{if } x \in \bigcap_{i=1}^m \text{TrustRegion}_i \\ 0 & \text{otherwise} \end{cases}$$

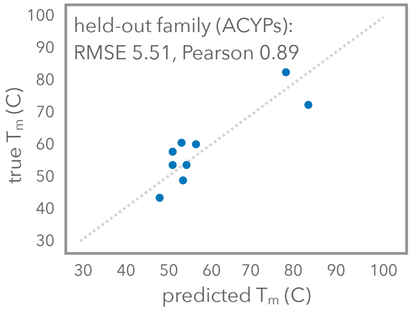
which is the solution to the following optimization problem:

$$\begin{aligned} \arg\max_{p \in \mathcal{P}} \quad & H(p) \\ \text{subject to} \quad & \mathbb{E}_p[f_1(x)] \geq \tau_1, \\ & \dots \\ & \mathbb{E}_p[f_m(x)] \geq \tau_m, \\ & \text{support}(p) \subseteq \bigcap_{j=1}^m \text{TrustRegion}_j \end{aligned}$$

## Predictive models and trust regions for…

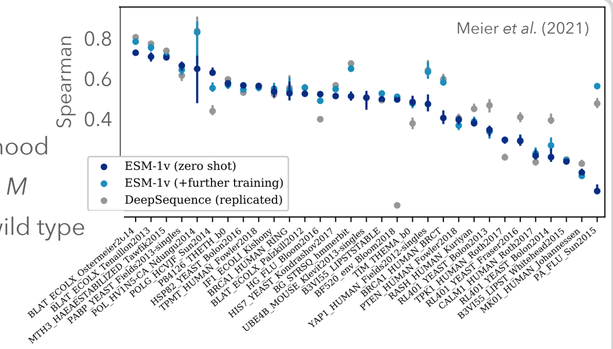
### Thermostability

- meltome: melting points for 34k+ protein sequences
- NN on top of ESM-1b embeddings
- trust region: sequences classified as in-distribution with meltome sequences



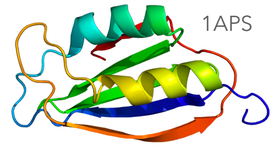

### Activity

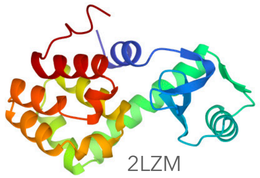
- ProGen2 log-likelihood
- trust region: within *M* mutations from a wild type




We also expect the ProGen2 log-likelihood's correlation with activity to depend on the wild type—specifically, **on the extent to which activity drove evolutionary pressure on the wild type**.

## Wild-type enzymes




### Acylphosphatase (ACYP)

- human ACYP2
- *P. horikoshii* (thermophile)
- *S. benthica* (psychrophile)
- ACYP-like domain in hypF

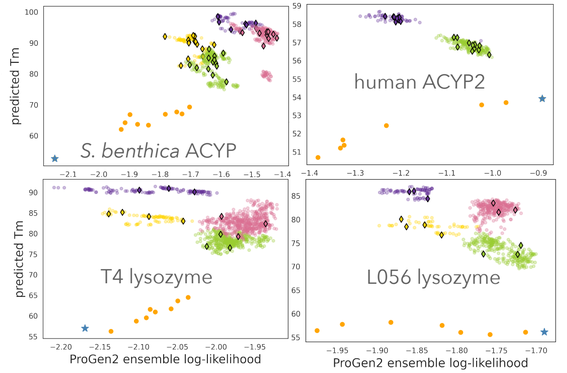
### Lysozyme

- phage T4
- L056 (previously designed)
- L070 (previously designed)
- *B. intermedia*

## Predicted behavior of designed sequences

Experimental characterization is in progress and forthcoming

- low λ, high ID to WT
- high λ, high ID
- low λ, low ID
- high λ, low ID
- PROSS
- wild type (WT)





**References**

Goldenzweig *et al.* (2016), *Mol. Cell*
Jarzab *et al.* (2019), *Nat. Methods*
Madani *et al.* (2022), *Nat. Biotech.* (to appear)
Markin & Mokhtari *et al.* (2021), *Science*
Meier *et al.* (2021), NeurIPS
Nijkamp & Ruffolo *et al.* (2022), arXiv:2206.13517
Pinney *et al.* (2021), *Science*
Rives *et al.* (2021), *PNAS*
Sun *et al.* (2022), ICML