Designing active and thermostable enzymes with sequence-only predictive models

Clara Fannjiang¹,², Micah Olivas³, Eric R. Greene⁴, Craig J. Markin⁵, Bram Wallace², Ben Krause², Margaux M. Pinney⁶, James S. Fraser⁴, Polly M. Fordyce³,⁷, Ali Madani⁸, Nikhil Naik²

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

Our general approach
\[ f_i, i = 1, \ldots, m : \text{predictive model of } i\text{-th fitness function} \]
\[ \text{TRUST_REGION}, i = 1, \ldots, m : \text{region of sequence space on which we trust } f_i \]

We use a Metropolis-Hastings algorithm to sample novel sequences from the distribution
\[ p^*(x) \propto \exp\left(\sum_{i=1}^{m} \lambda_i \cdot f_i(x)\right) \] if \( x \in \bigcap_{i=1}^{m} \text{TRUST_REGION} \)
otherwise

which is the solution to the following optimization problem:
\[ \arg\max_{p \in P} H(p) \]
such that
\[ B_i[f_i(x)] \geq \tau_i, \]
\[ \ldots \]
\[ B_m[f_m(x)] \geq \tau_m, \]
support\( (p) \subseteq \bigcap_{i=1}^{m} \text{TRUST_REGION} \)

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

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

Predictive models and trust regions for…

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

- 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

References