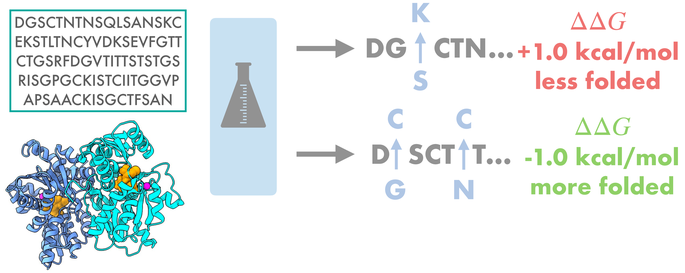
Predicting a Protein's Stability under a Million Mutations

TEXAS IFML
The University of Texas at Austin

Jeffrey Ouyang-Zhang Daniel J. Diaz Adam Klivans Philipp Krähenbühl

NEURAL INFORMATION PROCESSING SYSTEMS

Overview



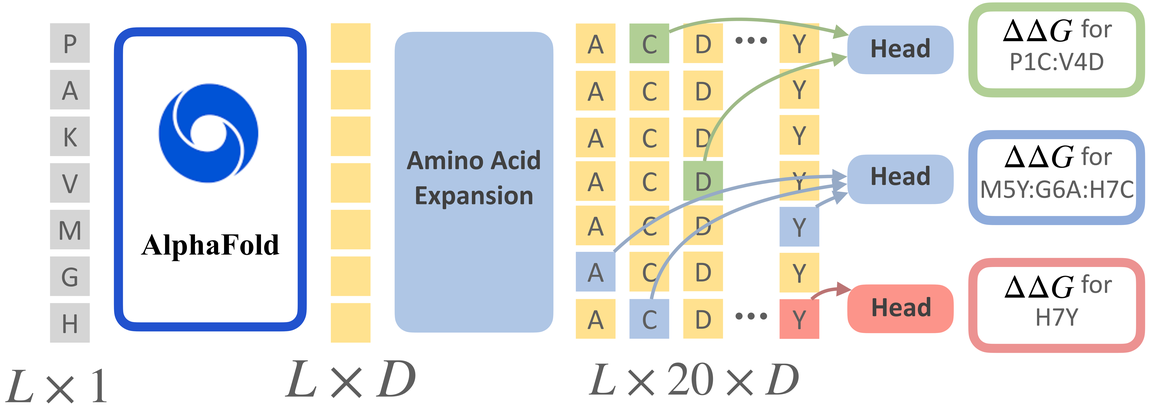
Given: protein sequence
Goal: find stabilizing mutations
Why? engineer protein-based biotechnologies

Problem

- combinatorial number of mutations
- stabilizing mutations are rare ($\sim 3\,\%$)
- epistatic (nonlinear) interaction between mutations

Method

Mutate Everything: one pass of protein backbone + parallel decoding



Loss

$$L(x, \{\mu_i, \Delta\Delta G_i\}) = \frac{1}{N} \sum_i ||f_\theta(x, \mu_i) - \Delta\Delta G_i||$$

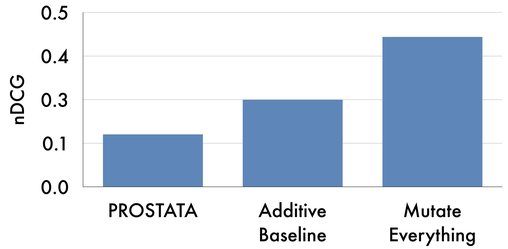
x: protein sequence
μ_i: mutation i for protein x
$\Delta\Delta G_i$: change in thermodynamic stability under μ_i

Benefits

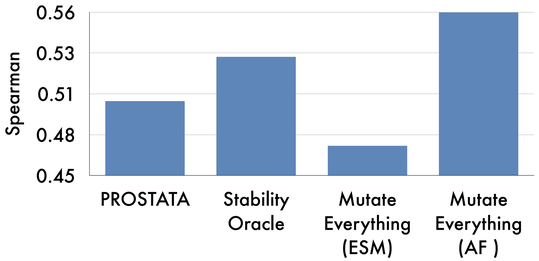
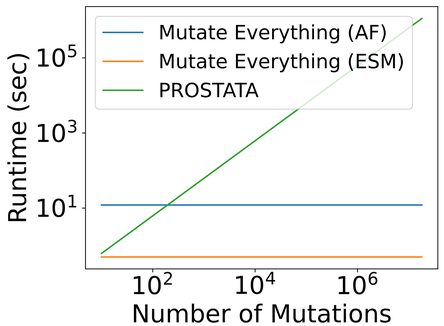
- highly accurate
- fast and parallel
- predicts higher-order (> 2) mutations

Results

cDNA Double Mutations



S669 Single Mutations



Runtime