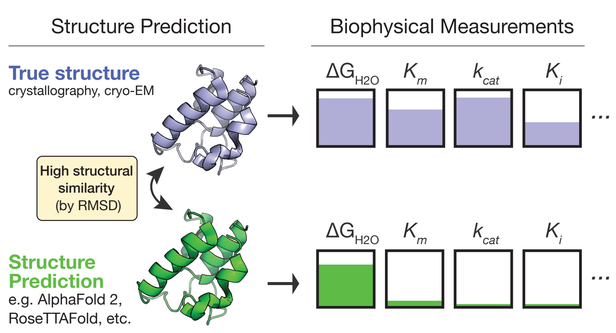
Leveraging novel protein language models to understand constraints on enzyme function and design

Micah Olivas[1,3], Clara Wong-Fannjiang[2], Craig Markin[3], Nikhil Naik[2], Polly Fordyce[1,3]

[1]Department of Genetics, Stanford University [2]Salesforce AI Research [3]Sarafan ChEM-H, Stanford University

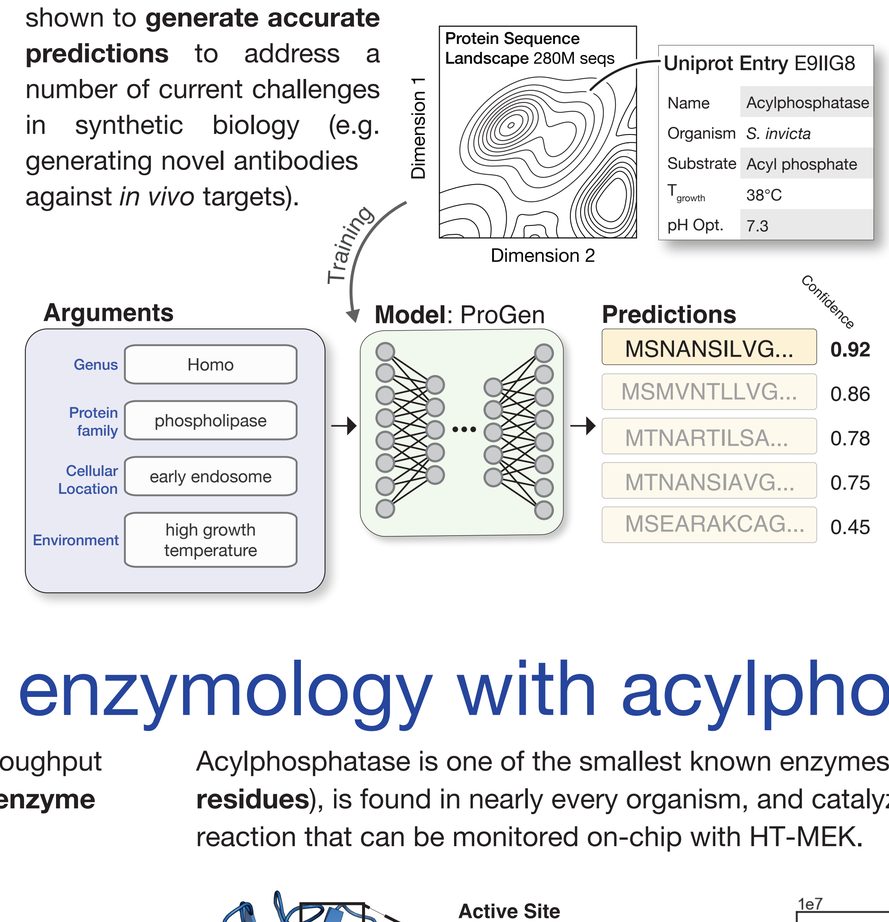
Introduction

Recent developments in protein structure prediction have drastically improved our ability to generate stable protein folds. Generating functional predictions is considerably more challenging, due in large part to a lack of large, biophysical datasets of protein function nearing the scale and dimensionality of datasets in the Protein Data Bank.

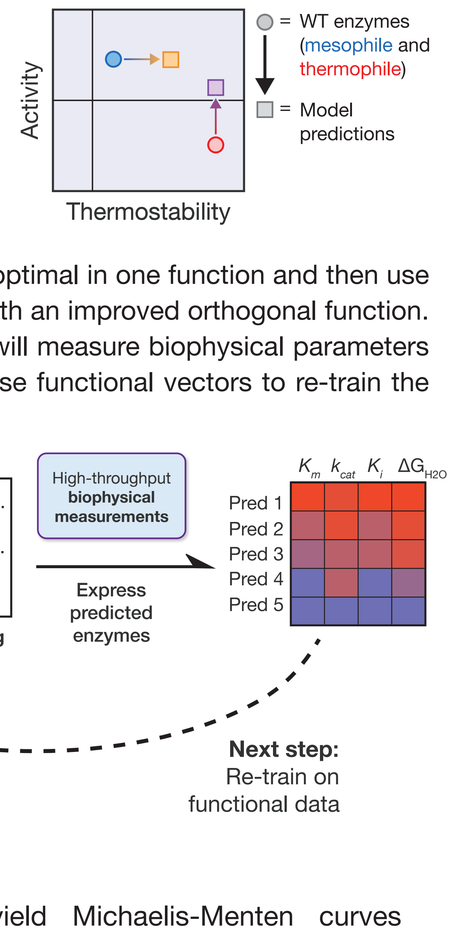


This suggests that improving protein functional prediction will require **improved predictive models** and a vastly expanded **landscape of biophysical training data.**

Protein language models learn an underlying amino acid sequence distribution from hundreds of millions of natural proteins. Best in class models such as ProGen sample a highly-parameterized joint probability distribution to generate sequences with a "high likelihood" of existing in nature. Even with a sparse set of inputs, these models have been shown to **generate accurate predictions** to address a number of current challenges in synthetic biology (e.g. generating novel antibodies against *in vivo* targets).



Ultimately, successful predictors will optimize multiple, orthogonal functions for high-confidence sequence predictions, for example thermostability and substrate turnover in an enzyme, as shown on the right. To acheive dual optimization across two functions, we propose selecting a starting sequence that is known to be optimal in one function and then use the predictor to generate sequences with an improved orthogonal function. To appraise ProGen performance, we will measure biophysical parameters across model generations and use these functional vectors to re-train the model.

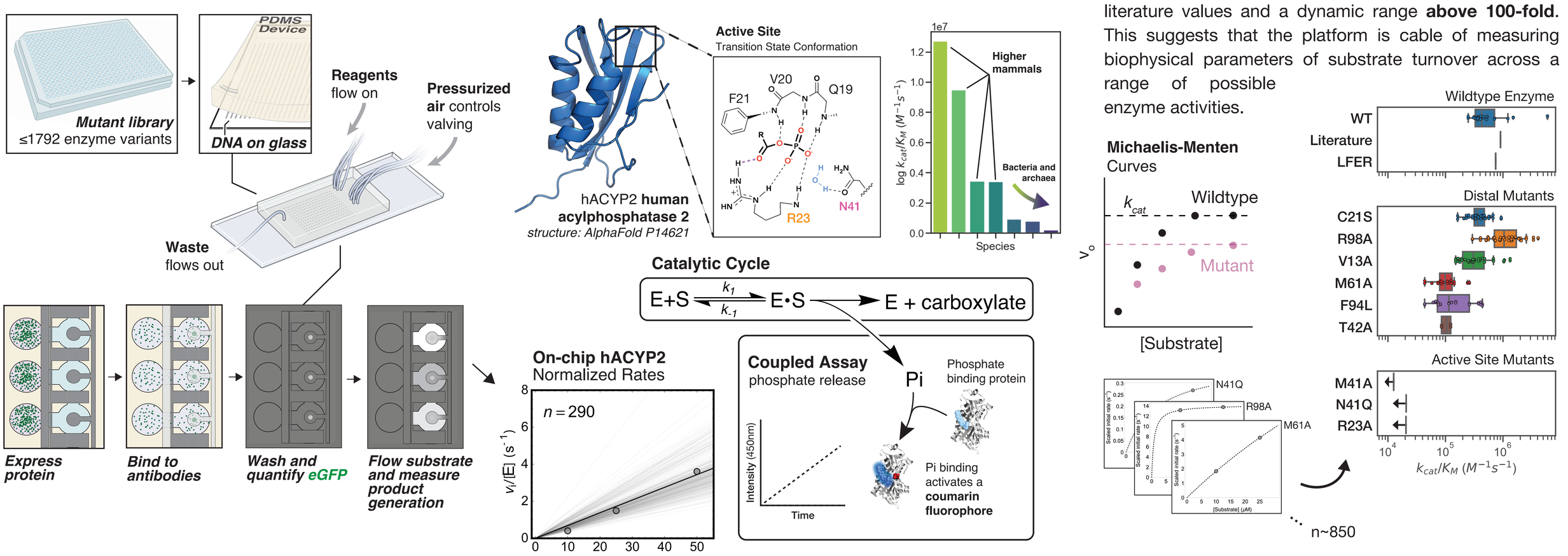


In vitro microfluidic enzymology with acylphosphatase

The HT-MEK microfluidic platform allows for high-throughput expression, purification, and kinetic assays of **1,800 enzyme variants** via time-resolved fluorescence microscopy.

Acylphosphatase is one of the smallest known enzymes (**100 residues**), is found in nearly every organism, and catalyzes a reaction that can be monitored on-chip with HT-MEK.

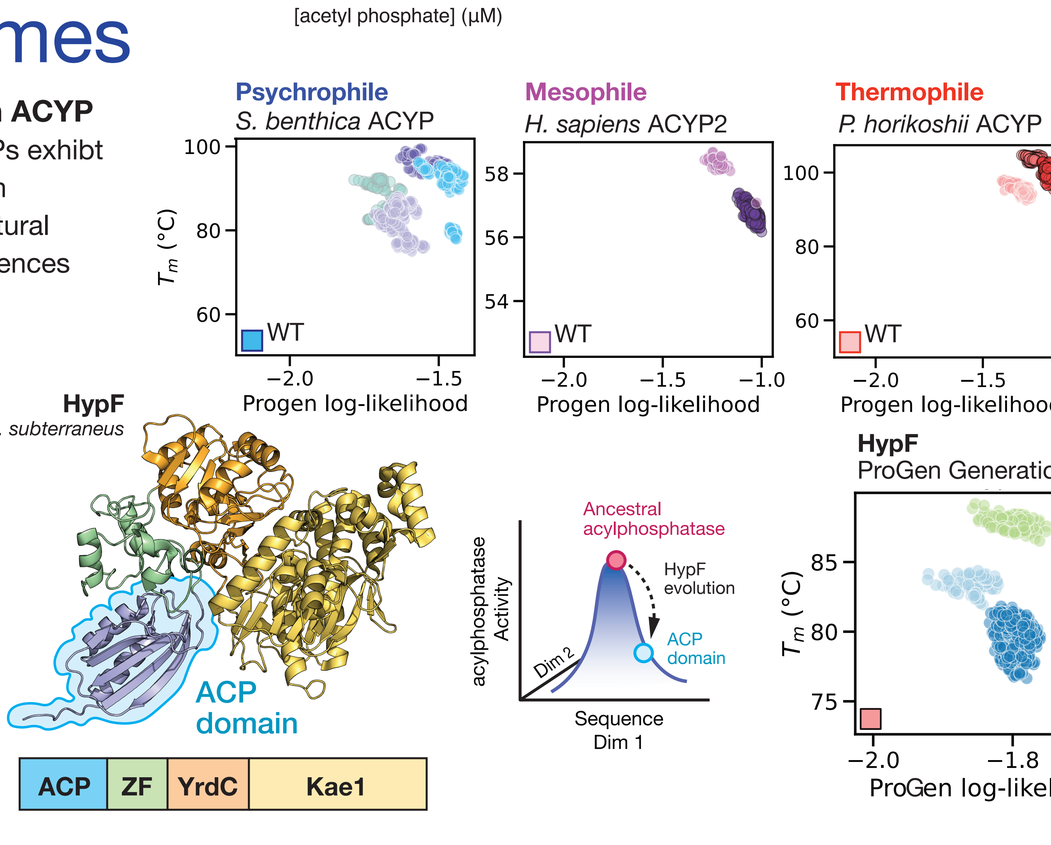
On-chip experiments yield Michaelis-Menten curves similar to those shown below. Measurements for k_{cat}/K_M across a library of 9 mutants show an ageement with literature values and a dynamic range **above 100-fold.** This suggests that the platform is cable of measuring biophysical parameters of substrate turnover across a range of possible enzyme activities.



Expected Outcomes

Optimizing activity and thermostability in ACYP orthologs Like other globular proteins, ACYPs exhibit variable thermostability across different growth temperatures. We used ProGen to improve natural extremophilic ACYP variants to generate sequences with improved thermostability and native acylphosphatase activity.

Reversing inactivation of an ACP domain in HypF A gene fusion event is thought to have fused an ancestral acylphosphatase to the N terminus of the carbamoyl transfersase HypF. Activity measurements of truncated ACP suggest that HypF evolution has selected against acylphosphatase function in the domain. We generated ProGen predictions with improved ensemble log-likelihood to identify activating residue-residue epistasis within the acylphosphatase topology.



Library Generation

To test these predictions, we designed a library of 250 WT and predicted sequences. We will use the HT-MEK platform to meaure many biophysical parameters across the library and investigate new hypotheses about basic enzyme function.

salesforce
CHAN ZUCKERBERG INITIATIVE
NATIONAL INSTITUTES OF HEALTH
ONO PHARMA FOUNDATION
Sarafan ChEM-H Stanford University