Predicting disruptions in protein conformational landscapes

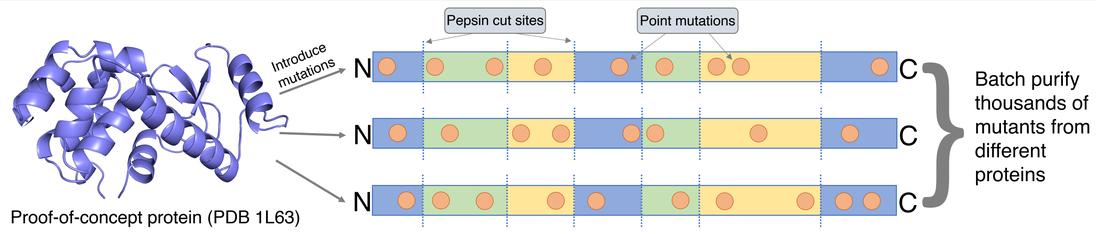
Andrew Reckers and Anum Glasgow

Department of Biochemistry, Columbia University Irving Medical Center

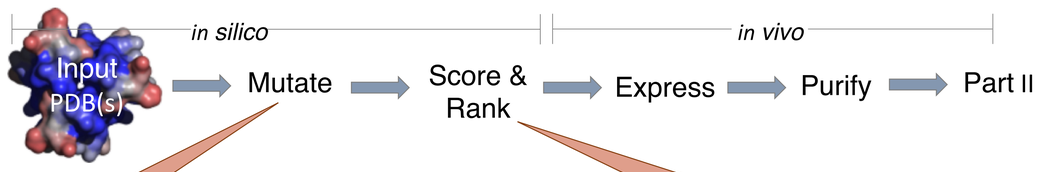
Can we predict how mutations and perturbations affect protein conformational landscapes?

- *Background:* Allosteric perturbations can change the protein conformational landscape. These changes can affect protein function in the form of interrupted interactions, stability, PTM formation, and catalytic activity, potentially leading to disease.
- *Challenge:* Allosteric changes to a protein's conformation are difficult to observe, and there are no methods to predict how allosteric perturbations affect protein structural dynamics. Static structures do not provide dynamic data, while data from hydrogen-deuterium exchange with mass spectrometry (HDX/MS) are too low-throughput to allow for prediction of allosteric effects.
- *Proposed solution:* Develop a high-throughput HDX/MS pipeline to observe protein conformational response to mutations. With this data, we can train a machine learning model to predict conformational changes beyond the initial training set.

Part I: Designing a mutagenized, barcoded protein library



1. Computationally design a mutagenized protein library such that, following pepsin digest, each peptide has a unique mutation, enabling deconvolution by MS
2. Develop a pipeline for batched scoring, expression and purification of these mutations from multiple easily expressed proteins

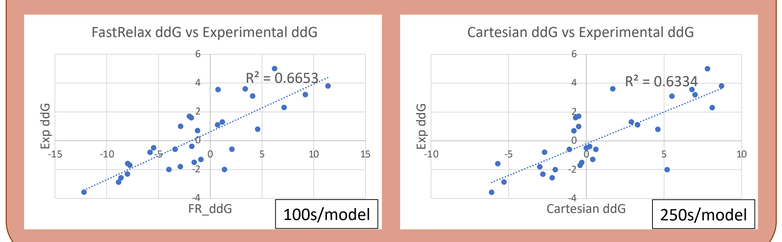


Mutation methods:
1. Random point mutation every n-mer

2. Machine learning model for sequence prediction from backbone structure (ProteinMPNN)

3. Coevolution-biased mutations

Tested scoring methods against 30 experimentally determined ΔΔGs:

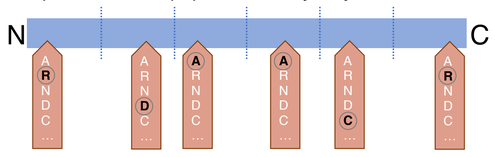


Part I preliminary data:

Method 1: Random mutations in every n-mer

- *Workflow:*

1 random substitution in every 12-residue segment of a proof-of-concept protein: T4 Lysozyme



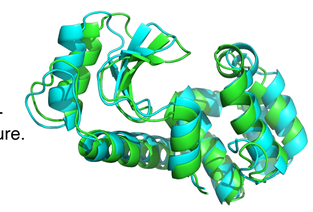
Protein design 70k mutant structures using FastRelax

Rank structures by lowest ΔΔG

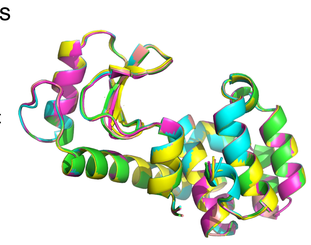
Rank	ddG
1	-4.64
2	-2.11
3	-0.92
4	-0.92
5	0.41

Orthogonal computational validation of the top 30 mutants: RoseTTAFold as a method for structure prediction from mutant sequences

Designed (green) aligned to RoseTTA-Folded (blue) structure. RMSD = 1.9Å



5 RoseTTAFold output models aligned to each other. Mean RMSD = 0.15Å



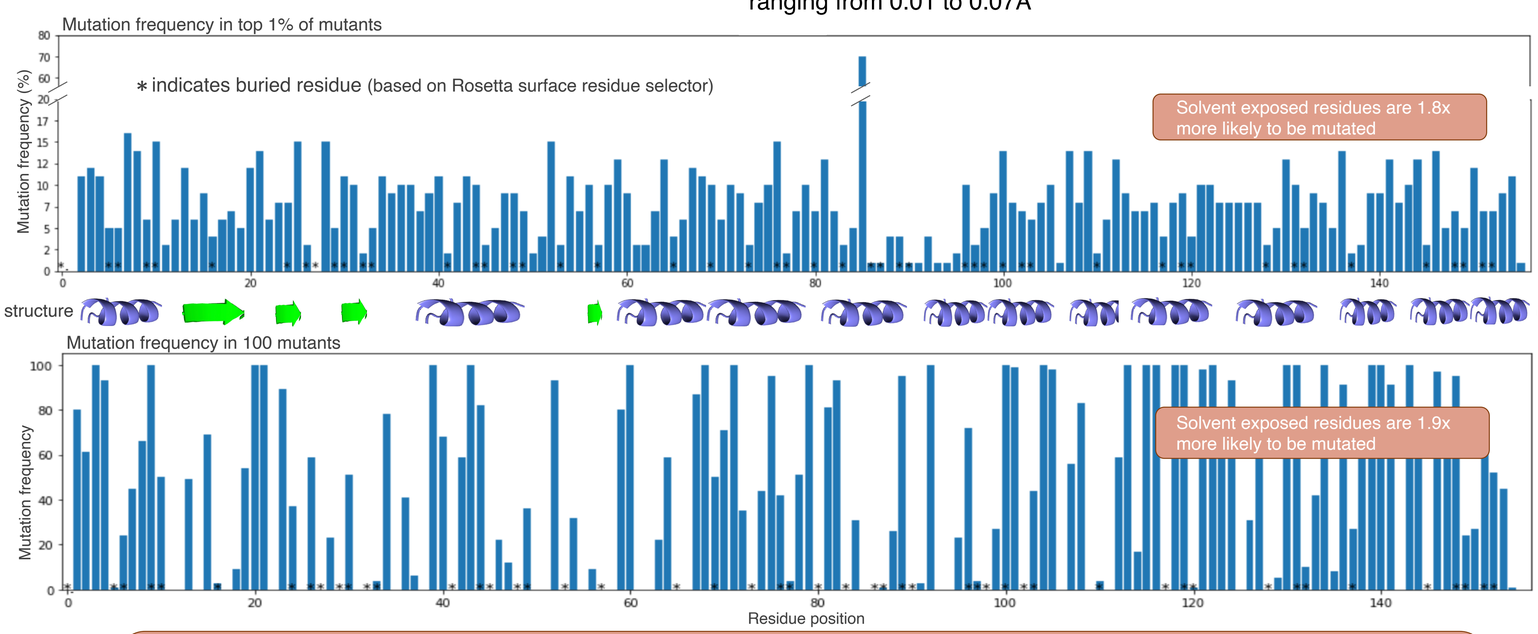
- RMSDs of design vs RoseTTAFold structures range from 0.9Å to 2.1Å, with a mean of 1.6Å
- All 5 models produced by RoseTTAFold for each mutant were similar to each other, with an SD ranging from 0.01 to 0.07Å

Order 30 genes. Clone & express

Lyse cells and check soluble fractions for presence of T4 Lysozyme to assess if it is folding properly

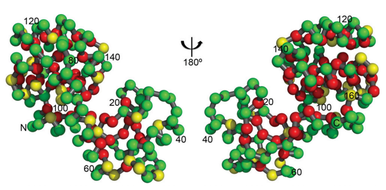
	Insoluble	Soluble
Wildtype		
Mutant 1		
Mutant 19		

Results: a few of the 30 mutants have soluble protein, but the majority do not.

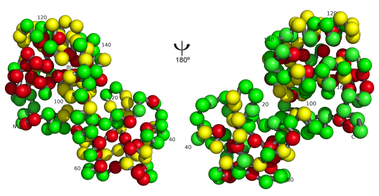


Results:
- ✓ After relaxing, 99 of the 100 sequences had lower REUs than all 70k random mutants (maybe indicating a higher likelihood of folding *in vitro*)
- ✓ Mutations occur throughout the protein sequence, but little diversity between generated sequences
  - Need to introduce further diversity to generate unique protein library

Comparing published activity data with mutation frequency in top 1% of mutants:



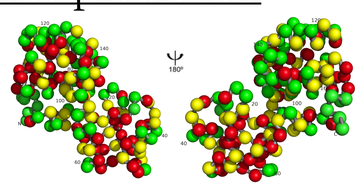
Activity assays (Baase, et al)[3]
Red = intolerant to mutations
Yellow = somewhat tolerant to mutations
Green = tolerant to mutations



Random mutation pipeline
Red = least likely to mutate
Yellow = somewhat likely to mutate
Green = most likely to mutate

Method 2: ML sequence prediction

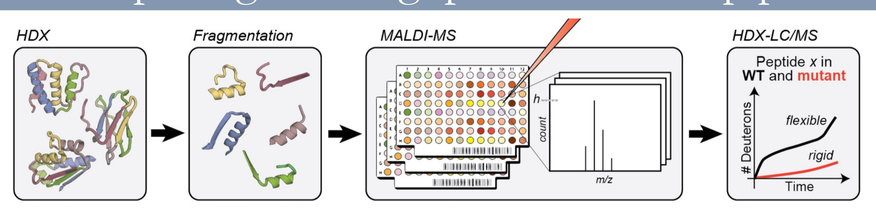
- The newly published Protein MPNN method[2] uses machine learning to generate sequences for a given protein backbone
- We generated 100 MPNN sequences from a relaxed T4 lysozyme PDB



MPNN mutation pipeline
Red = least likely to mutate
Yellow = somewhat likely to mutate
Green = most likely to mutate

- This tool may allow us to develop a mutagenized protein library that is more representative of naturally-occurring proteins, and may be more likely to fold correctly than the random mutation approach

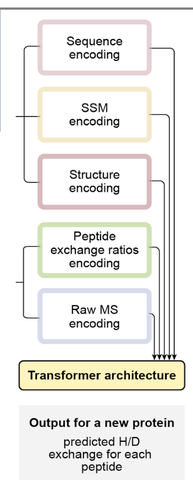
Part II: Develop a high-throughput HDX/MS pipeline



- The batched, purified library of mutagenized proteins is deuterium exchanged, pepsin digested, and then screened by iterative matrix-assisted laser desorption/ionization (MALDI)-MS followed by HDX/MS
- Screening by MALDI-MS allows for rapid identification of:
  - The peptides in each well (helpful for subsequent HDX/MS analysis)
  - Peptides with mutations that alter the relative rigidity of the secondary structures in the protein
- These selected peptides are sent to HDX/MS for a longer, more thorough analysis of protein dynamics

Part III: Build a machine learning model to predict effects of mutations

- The HDX/MS data, alongside sequence and structural information used in protein structure prediction methods, will be encoded to build a neural net for predicting protein conformational landscapes
- We will initially test our model to validate conformational changes in mutants from our library that were not in our training set, and then move on to larger proteins that include domains with similar topologies
- Accurate predictions may enable the discovery of currently invisible disease mechanisms and the design of highly specific allosteric therapeutics



References:
1. Changeux, Jean-Pierre. "50 years of allosteric interactions: the twists and turns of the models." *Nature reviews. Molecular cell biology* vol. 14,12 (2013): 819-29. doi:10.1038/nrm36952.
2. Dauparas, J, et al., bioRxiv 2022.06.03.494563; doi.org/10.1101/2022.06.03.494563
3. Baase, Walter A et al. "Lessons from the lysozyme of phage T4." Protein science : a publication of the Protein Society vol. 19,4 (2010): 631-41.
Pdb icon: www.biorxiv.org/content/biorxiv/suppl/2016/11/01/084269.DC1/084269-1.pdf
Flask: www.bioicons.com

COLUMBIA
COLUMBIA UNIVERSITY IRVING MEDICAL CENTER
GLASGOW LAB
NIH NIGMS