An Exhaustive Analysis of Double InDel Mutants using Inverse Kinematics Robotics and Machine Learning

Katie Christensen, Sarah Coffland, Changrui Li

Research Mentors: Prof. Brian Hutchinson, Prof. Filip Jagodzinski

Computer Science Department, Western Washington University

Overview

Motivation: The number of possible mutants with double insertions/deletions (InDels) for even a small protein is combinatorially immense. Studies that seek to explore trends across a set of mutants rely on computational methods to generate mutants *in silico*. These methods are computationally time and data intensive.

Goal: Find and reveal hot spots that highlight pairs of residue locations where InDels have a pronounced and impactful effect while minimizing the computational resources required for this task.

Approach: Generate an exhaustive data set of all possible InDel mutants using an inverse kinematics robotics approach available in Rosetta. Generate heat maps of several energetics-based metrics output by Rosetta to reveal hot spots using a transformer neural network model of these phenomena.

Background

We study four proteins and the InDel locations and identities of amino acids introduced or removed.

Proteins:

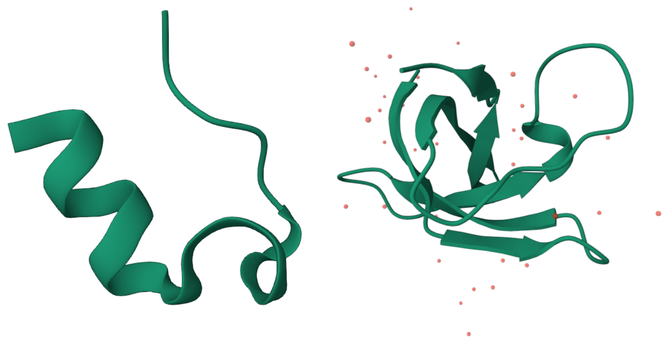


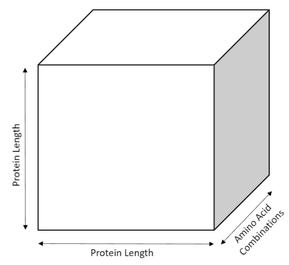
Figure 1: 1L2Y (left) and 1CSP (right)

Figure 2: 1CRN (left) and 1HHP (right)

Rosetta Scores: Widely used metrics for assessing the energy of protein folding and the quality of protein structure prediction.

Experimental Setup

The axes of the heat maps represent the insertion positions for each residue in the protein.



Each cell in the heat map contains either an accumulated count of the mutants with Rosetta Scores more than two standard deviations away from the mean of each respective score (outliers), or the average over all of the mutants for each respective Rosetta score.

Results

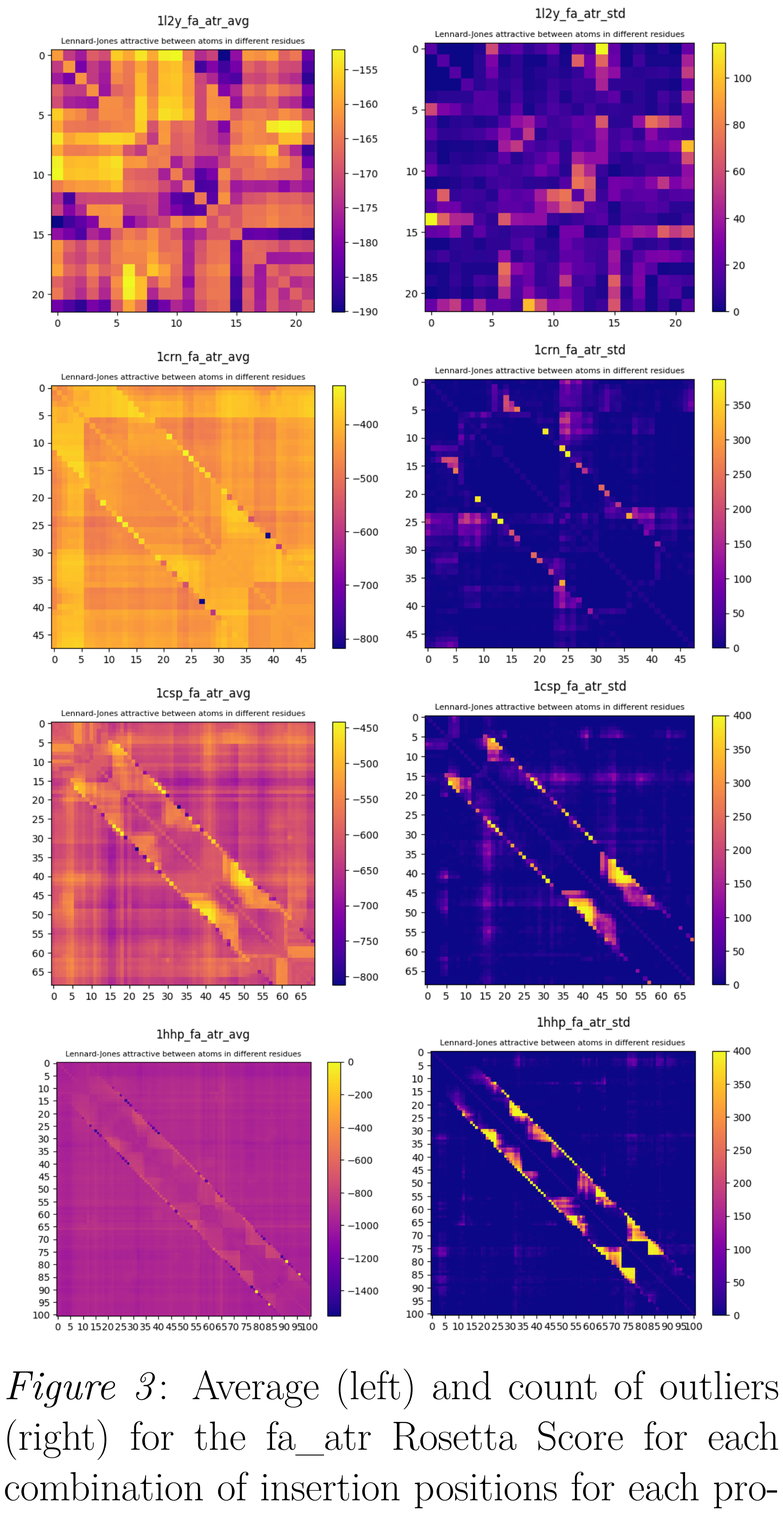


Figure 3: Average (left) and count of outliers (right) for the fa_atr Rosetta Score for each combination of insertion positions for each protein. fa_atr refers to the Lennard-Jones attractive between atoms for different residues.

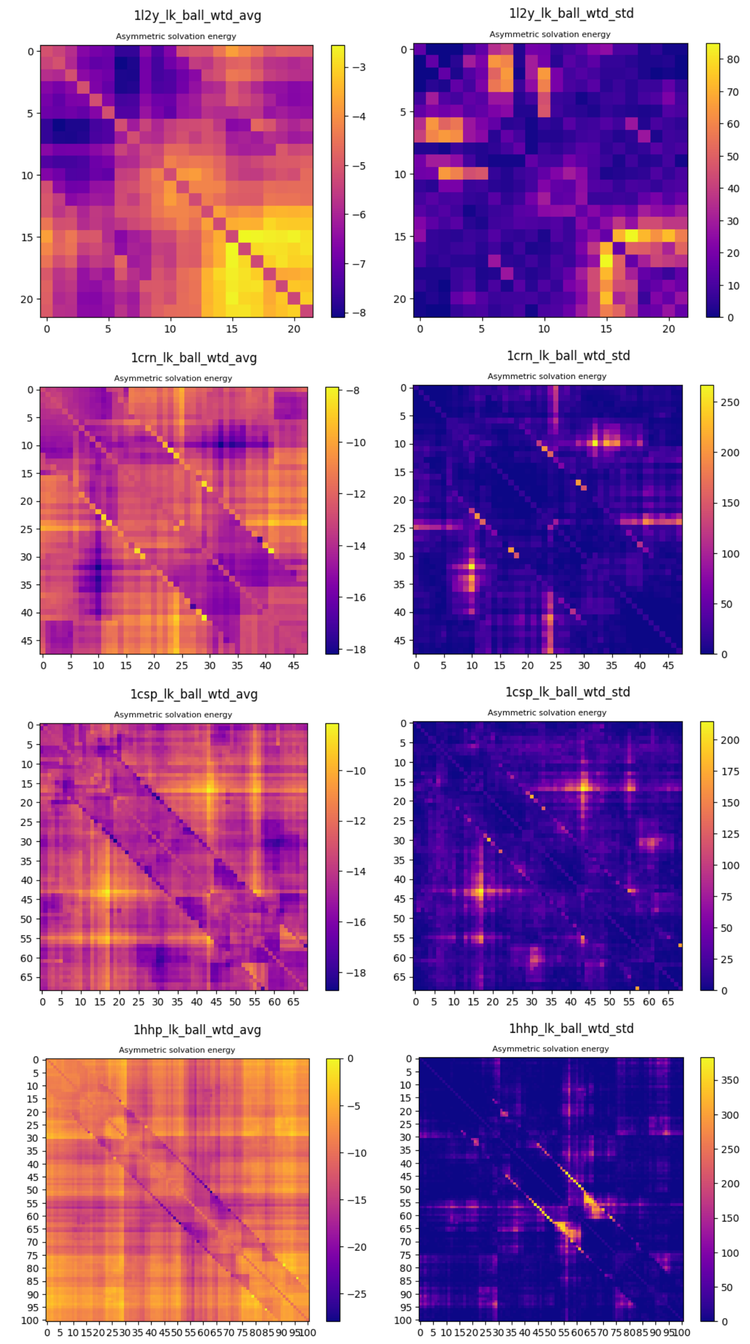


Figure 4: Average (left) and count of outliers (right) for the lk_ball_wtd Rosetta Score for each combination of insertion positions for each protein. The Rosetta Score lk_ball_wtd refers to the asymmetric solvation energy for that residue.

Future Work

We will develop a Deep Neural Network and a Transformer Neural Network that takes as input a mutant protein as four tokenized values:

$$[n_1, r_1, n_2, r_2]$$

Variable	Description
n_1	First InDel position
r_1	Identity of the amino acid at n_1
n_2	Second InDel position
r_2	Identity of the amino acid at n_2

and predicts the 20 Rosetta Scores for this mutant.