DDGScan is an enhanced *in silico* protein mutation prediction and analysis system

Sun Jinyuan1, 2, Wu Bian\*1

1 CAS Key Laboratory of Microbial Physiological and Metabolic Engineering, State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, PR China

2 University of Chinese Academy of Sciences, Beijing 100101, PR China

# Abstract

Proteins are fundamental biomolecules carrying out various biological functions directly. With the boost from highly accurate structure prediction methods, the space of proteins with available structures expanded by three orders of magnitude. However, structural bioinformatic tools are not previously designed for analysis in high throughput manner. The rich atomic-level information of structure is essential to mutation effect prediction. DDGScan is designed to improve the ability to analyze and design mutations of interest. Integrating input structure cleaning, multiple backend software, and parallel scheduling of all prediction jobs, DDGScan makes in silico deep mutational scanning more efficient. We also proved that by the ensemble of various methods improved the prediction accuracy. Also, DDGScan is equipped with multiple post-analysis functions to help biologists to interpret the results more easily. DDGScan is freely available at https://github.com/JinyuanSun/DDGScan to the community.

# Introduction

Protein and mutation prediction are important

Proteins are involved various biological processes in life. Enzymes have been catalyzing biochemical reactions including DNA replication and protein synthesis for billions of years. In past decades, enzymes have been engineered for biotechnology purpose because of its remarkable efficiency and selectivity. Introducing mutations based on structural features was one of the most successful strategies to improve protein stability. Matsumura et al. introduced disulfide bonds to the phage T4 lysozyme and improved the stability by 11 degrees. Wijma et al. designed variant of limonene epoxide hydrolase with Tm 35 ˚C higher than the wildtype by combining FoldX and Rosetta to calculate Gibbs free energy change (∆∆G) of every possible mutation and design disulfide bonds *in silico*. Cui et al. combined mutation designed by Rosetta, FoldX, ABACUS and consensus analysis successfully improved the Tm of IsPETase by 31 ˚C. These successes indicating that the combination of mutation designed through different methodologies is benefit to enzyme stabilization.

Introduction of basis of Rosetta protocol, FoldX and ABACUS1, 2

Various computational tools have been developed and try to quantify the effect of single mutation on the protein folding stability. In order to achieve a balance between computation time and accuracy, FoldX and ABACUS adopted the fixed-backbone assumption. The FoldX software used linear combination of empirical terms to calculate free energy. As a complementary, the ABACUS adopted a statistical method to calculate the negative log likelihood as energy. Rosetta cartesian ddg protocol also uses empirical energy function to score structures but sampled more extensively with backbone conformations, which makes it more accurate but also time consuming.

The importance of repair, and md

To eliminate possible pitfalls in calculation raised from insufficient backbone conformation sampling, molecular dynamics and visual inspection are usually involved in stabilizing mutation identification prior to experimental characterizations. The root mean squared error (RMSD) per residue is a strong indicator of point mutation stability thus helpful to visual inspection.

The overall protocol

Based on previous enzyme stabilization practices, we developed the DDGScan package to further facilitates computation added enzyme stabilization and would also be helpful to various purpose related to protein structure mutations. This package has three main sub-modules: mutation scanning module calculates all possible mutations of given protein with selected backends in parallel; list distribution module calculates all selected mutations parallel; analysis and plot module will summarizing and visualizing calculation results. Structure preparation and results analysis were automated and computation was distributed to make the process easier and more efficient.

# Results

## DDGScan protocol and post calculation analysis

In general, DDGScan main protocol contains 3 steps as demonstrated in figure 1. First, the input structure is examined and missing loops are added according to users assigned sequence information for sequence record in downloaded RCSB PDB file. The loop modeling is achieved through the modeler program, 5 models were built and the model with the best energy score is passed to further calculations. Second, all required pre-processes are automatically preformed according to backend software selected by the user. For FoldX, the structure file is required to be repaired before run ∆∆G calculations. In the Rosetta cartesian\_ddg protocol, the protein structure is relaxed and the relaxed structure with the lowest rosetta energy score is selected to run ∆∆G calculations. As for ABACUS, ABACUS\_prepare and ABACUS\_S1S2 are ran before single mutation calculations. After pre-processes a slow and a fast protocol are provided. In slow protocol, all mutations were calculated parallelly except ABACUS1, because of no single mutation prediction function provided. At this stage, all result files were dumped to a separated directory. Because the Rosetta cartesian\_ddg is the most time consuming in the slow protocol, an alternative fast protocol is provided. In the fast protocol, the cartesian\_ddg is replaced by pmut\_scan\_parallel, this application applied a relatively low-resolution protocol with fixed backbone and only reports mutations within cutoff set by user. After the ∆∆G calculation, molecular dynamics simulations (MDs) are performed using openmm for all mutations predicted to be stabilizing. GPU acceleration is supported for MDs.

Upon the heavy calculations ended, all scores will be analyzed. For FoldX and Rosetta with multiple runs, both mean score and the lowest score are going to be reported. Besides a complete prediction result, mutations are selected and grouped by the followings: mutations with score below the cutoff, the best amino acid at each position, the best mutation at each position with predicted score below the cutoff.

## In silico deep mutational scan

The computation performance of DDGScan protocol was tested. As all protocols adopted only consider the local environment of the target mutation within defined cutoff, the time consumption grows with the protein length in a linear manner (Figure 2). For a typical globular protein shorter than 500 amino acids, DDGScan protocol takes less than 30 hours to perform the *in silico* DMS. Also, we examined the scaling performance when more threads were used. Comparing with pre-defined naïve job distribution, the DDGScan parallel all calculations more efficiently and very close to perfect scaling, which means the running time drops linearly with the increasing of threads.

## DDGScan ensemble is a better ∆∆G predictor

## Evaluation of DDGScan on ∆Tm classification

## Case study of the β1 domain of streptococcal protein G (Gβ1)

# Methods and materials

## Computation protocols for different backends

For FoldX, the input structure is repaired with the RepairPDB module, then the BuildModel is used and the number of independent runs is set to 5 by default for the balance of sufficient sampling and time cost. For Rosetta, the input structure is relaxed by 50 runs by default, the relaxed model with lowest Rosetta energy is picked. The cartesian\_ddg application is used with iterations set as 3 and repack residues within 9 Å of the mutation, flexible backbone neighbor is set as 1. For Rosetta pmut\_scan\_parallel, ABACUS1 and ABACUS2, the default setting is applied. For more detailed information, refer to our previously published protocol of GRAPE. The missing loop is closed with the Modeller, which is a template based homologous modelling software. The model with lowest energy in five generated models is picked.

## Repeated high throughput molecular dynamics simulations

For predicted stabilizing mutations, short MD simulations were performed to further sample possible conformations. Simulations were carried out with openmm, using the ff99SB-ILDN force fields. The TIP3P water model was used for simulations. The net charge of the system was neutralized with sodium or chloride ions. The system was initially subject to energy minimization, followed by equilibrating for 100 steps to increase the temperature to 300K. The step length was set to 2 fs and a snapshot was saved every 1000 step. The production run of 100 ps generate 50 frames and in total 250 frames with 5 independent repeats.

## Collection of benchmark data

# Discussions

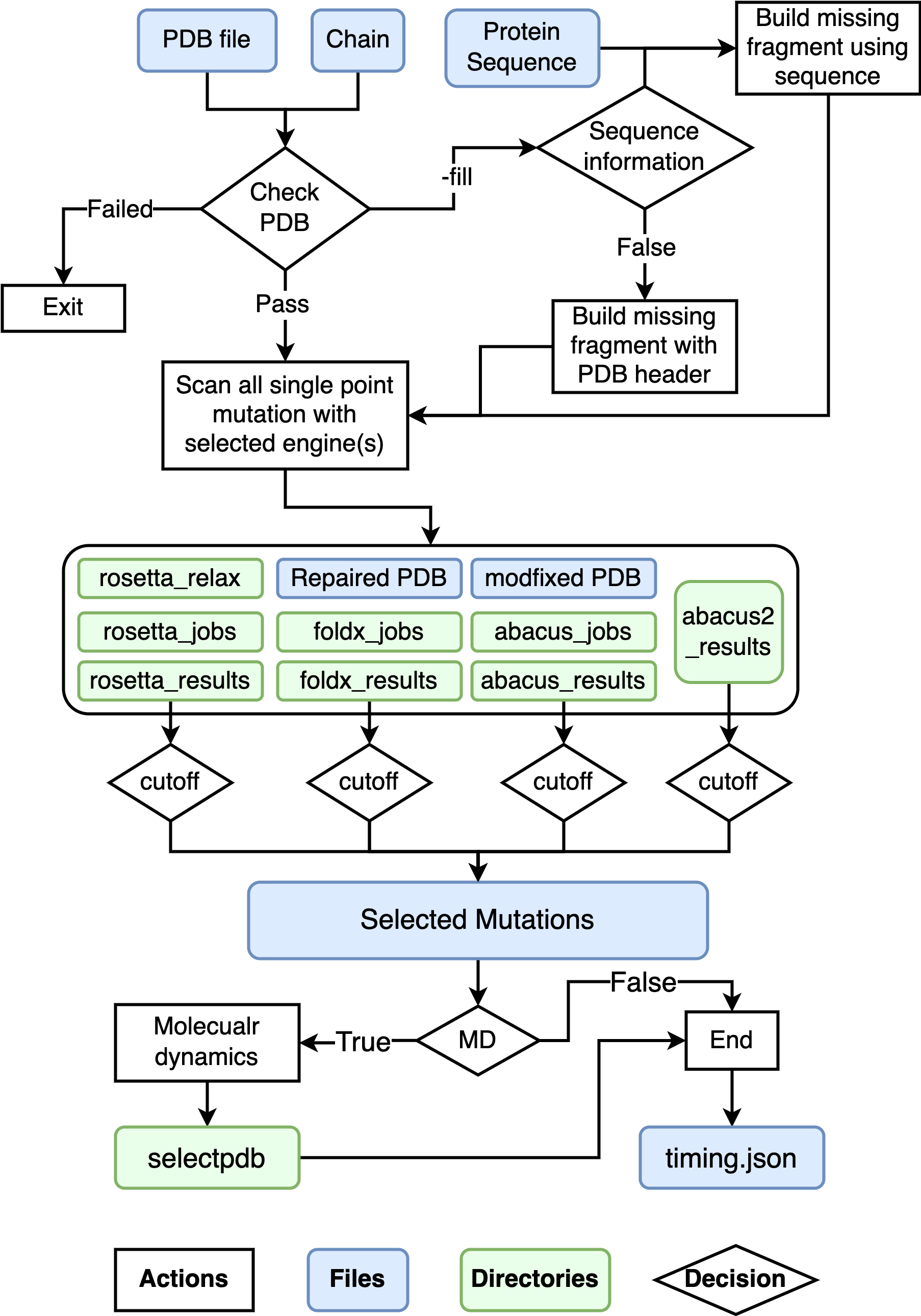


Figure 1.

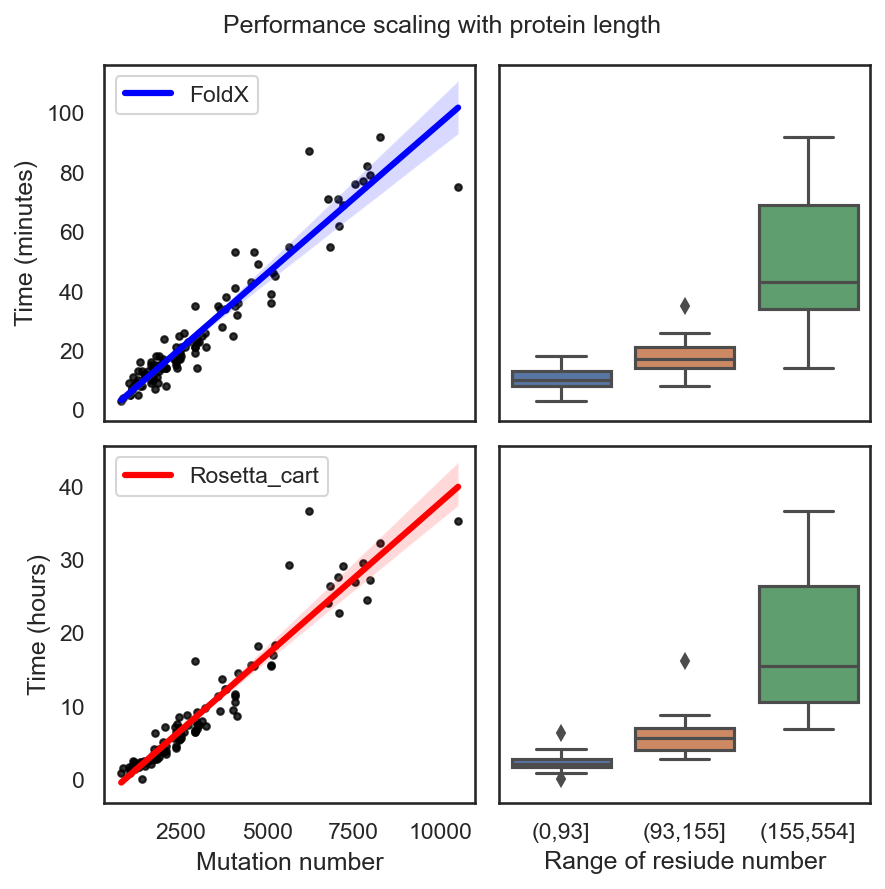


Figure 2

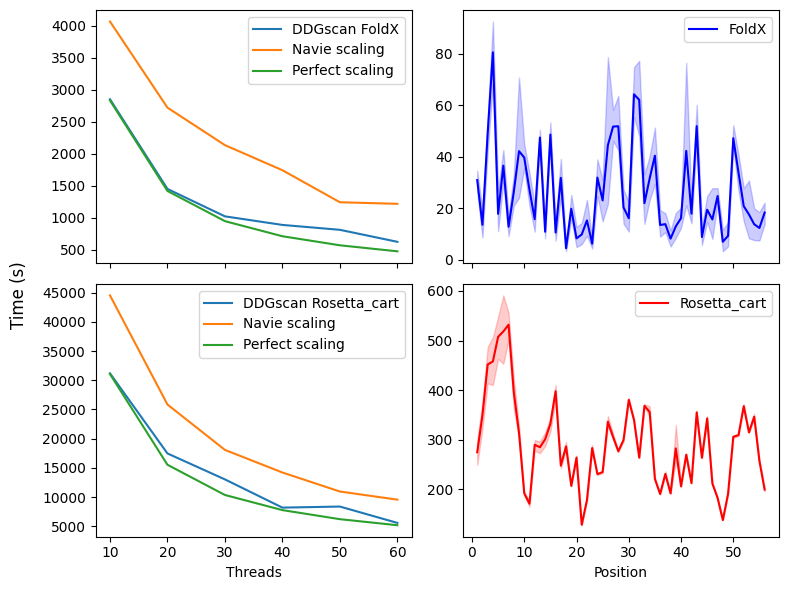


Figure 3

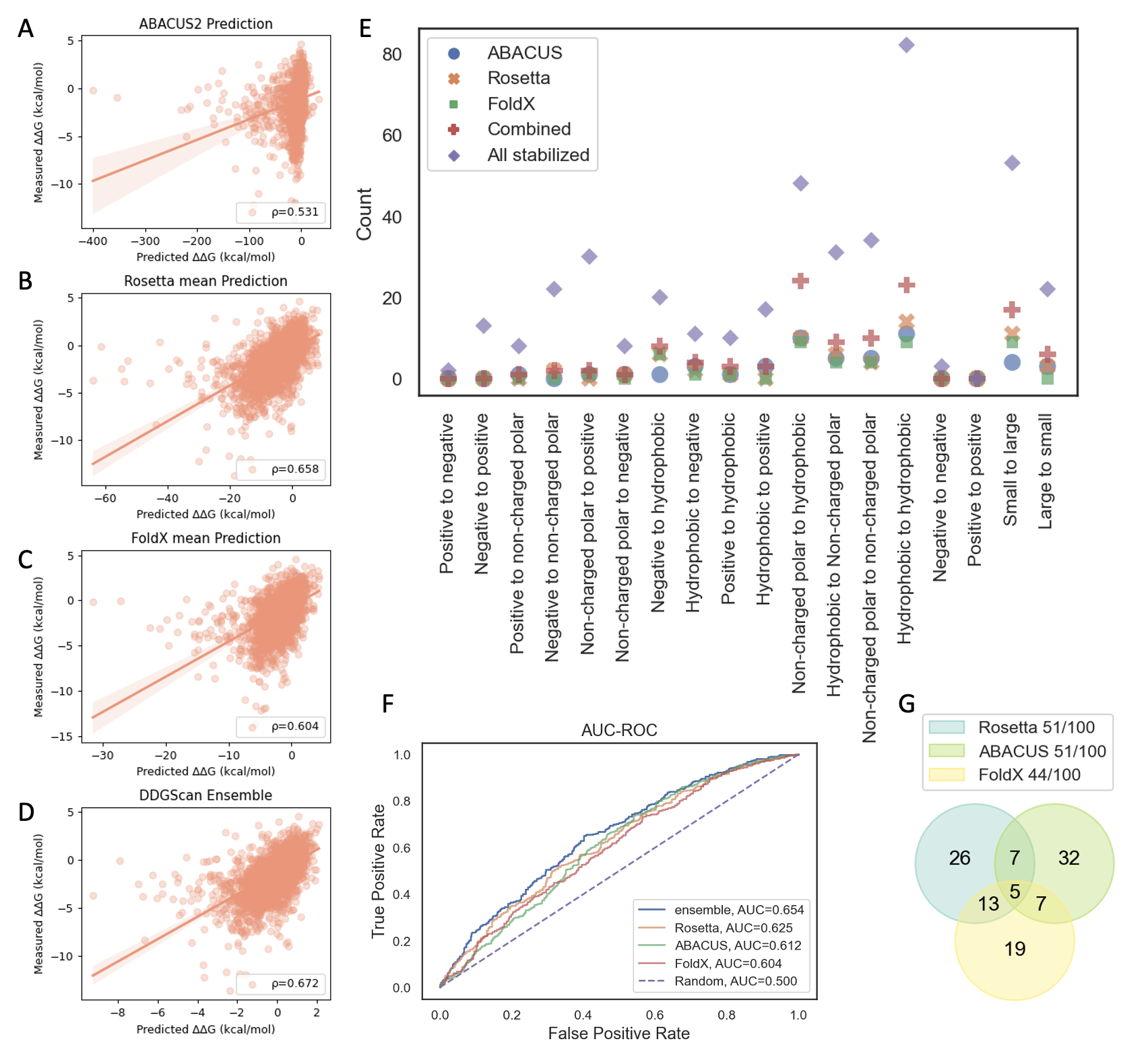


Figure 4

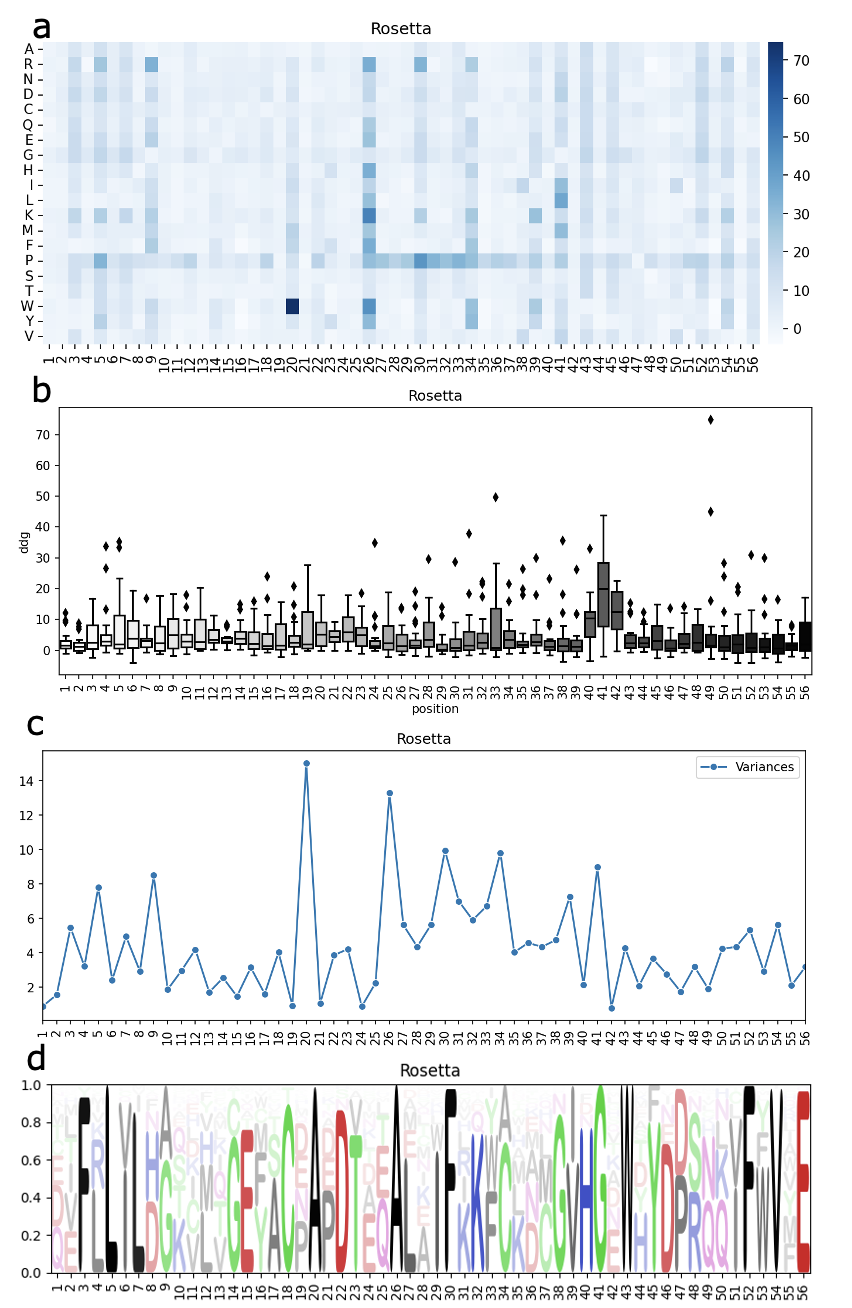


Figure 5

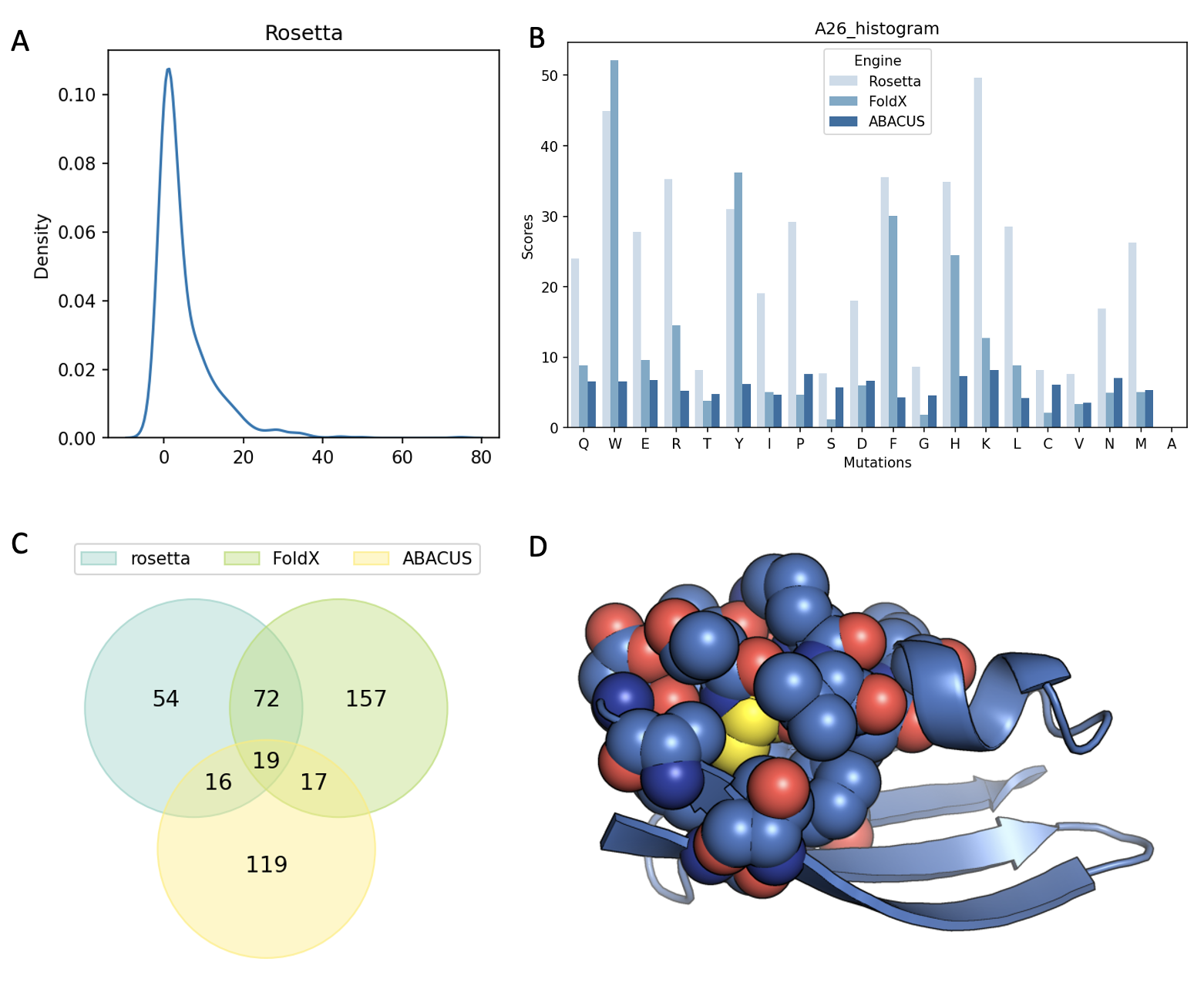


Figure 6