Normalizing SPTAN1 Across Different Experiments

Harsh Patel, Aryan H Nair, Cynthia Zhu, Jerry Gerber, Jennifer Li

Agenda

1

Background

2

Showcase

3

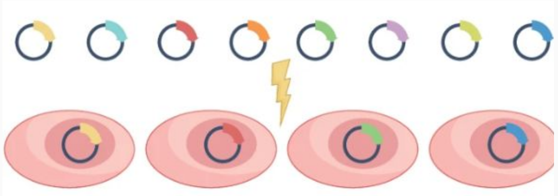
Implications

MaveDB - Multiplexed Assays of Variant Effect (MAVE) database

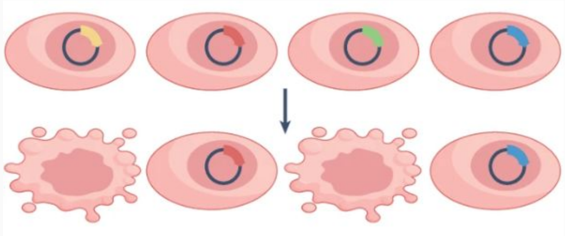
MAVE procedure

1) Select model system

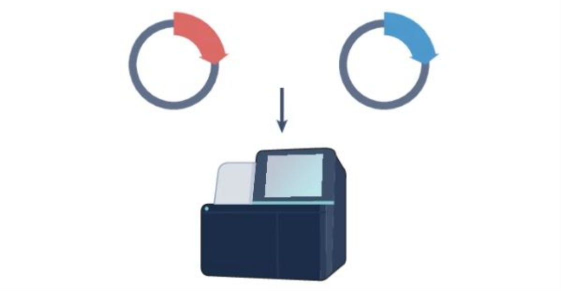
2) Introduce variations to target gene/protein sequence



3) Screen variants for phenotypic effect

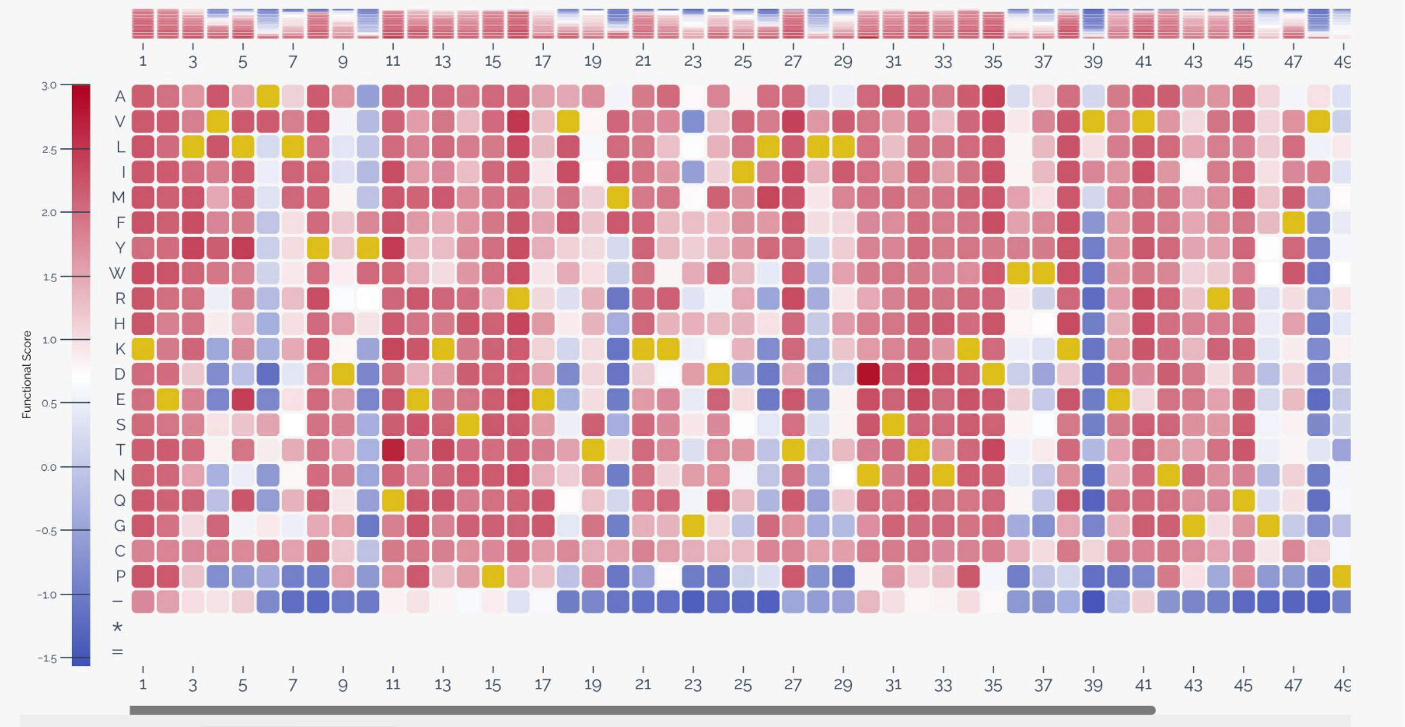


4) Sequence selected variants



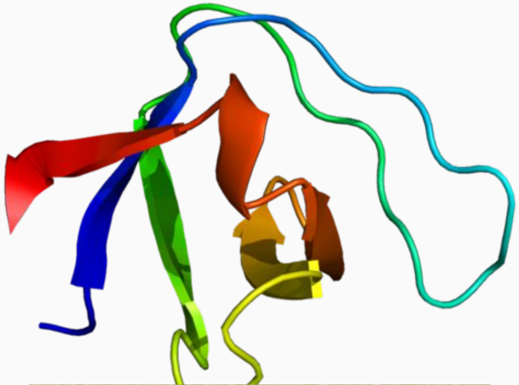
5) Compute functional scores and create visualization

hgvs_pro	score
p.Gly1Gln	0.75628101
p.Gly1Glu	0.62041306
p.Gly1Asn	0.72601136
p.Gly1Ile	0.70457362
p.Gly1Trp	0.67974701
p.Gly1Tyr	0.78640319
p.Gly1Phe	0.91414386
p.Gly1Pro	0.81780164
p.Gly1Cys	1.00312161
p.Gly1delinsGlyGly	0.74653632
…	….



Gene of interest: SPTAN1

- SPTAN1 gene
  - encodes alpha spectrin protein
  - mutations associated with early infantile epileptic encephalopathy-5
- MaveDB
  - 42 experiments
  - Selected for folding using proteases
  - Scored based on log10 K50 value and dG
- Experimental differences
  - Methods - Protease used (trypsin, chymotrypsin, or combination)
  - Species (gallus gallus and homo sapiens)



Normalizing the scores across different experiments:

Different experiments have different functional scores for the same mutation to the same protein.

How do we make a useful comparison between them?

.

Normalizing the scores across different experiments:

Answer:

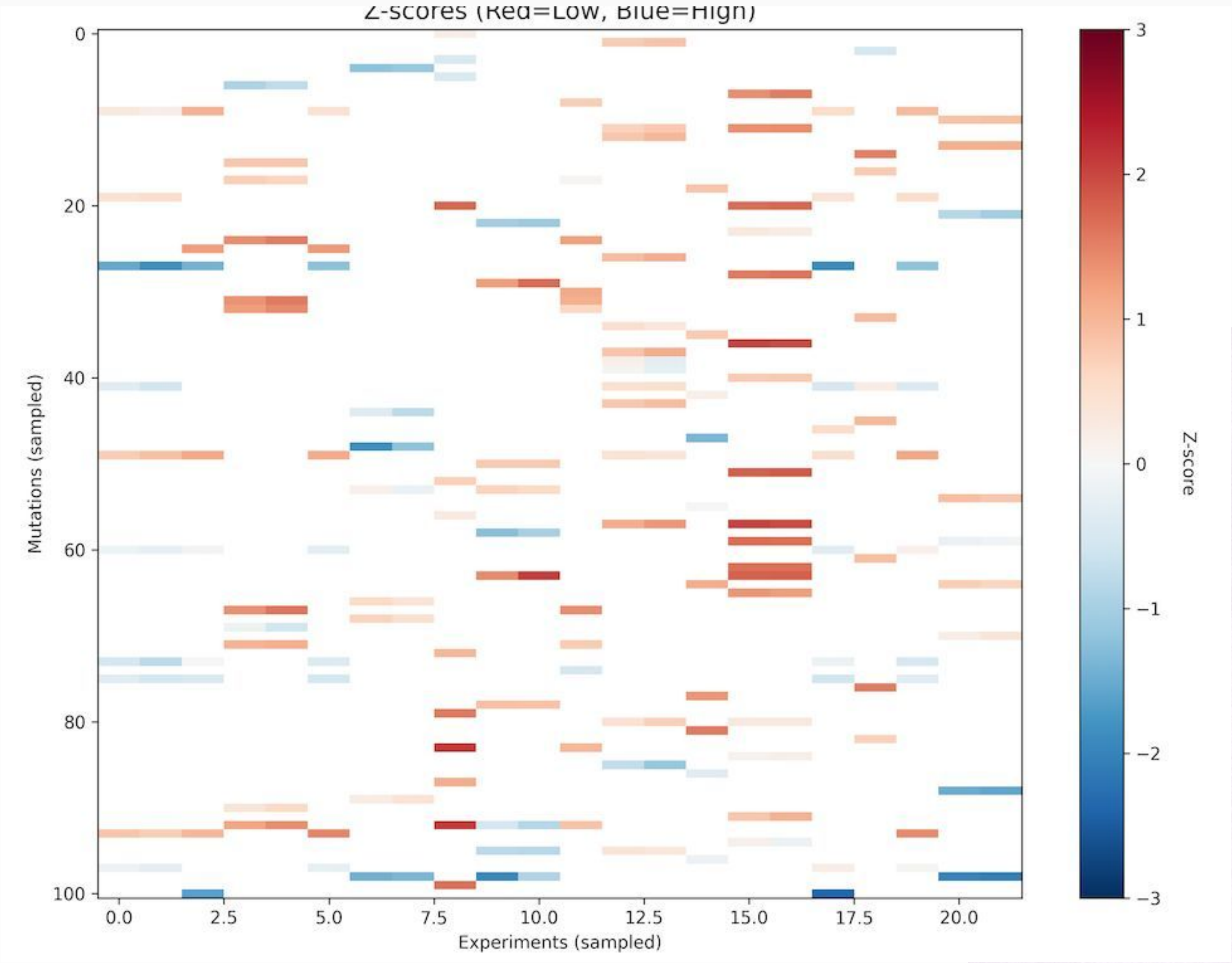
Z-Score

We find the Z-Score of the scores across experiments to normalize the data.

.

Lots of Missing Data

Comparing Averages?

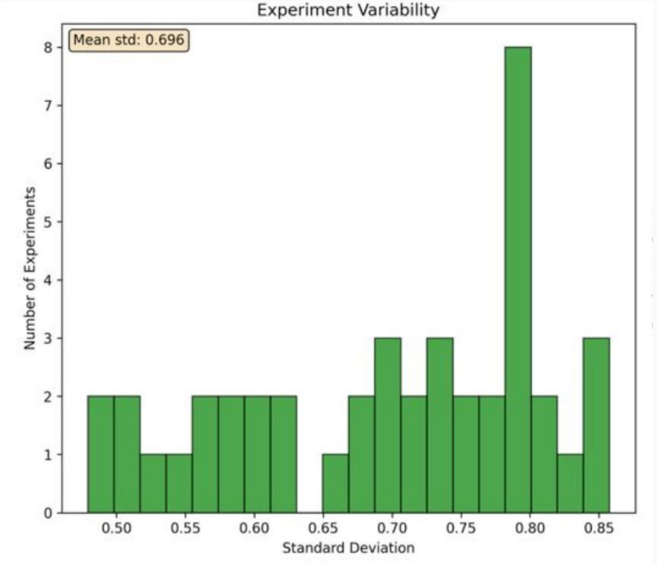


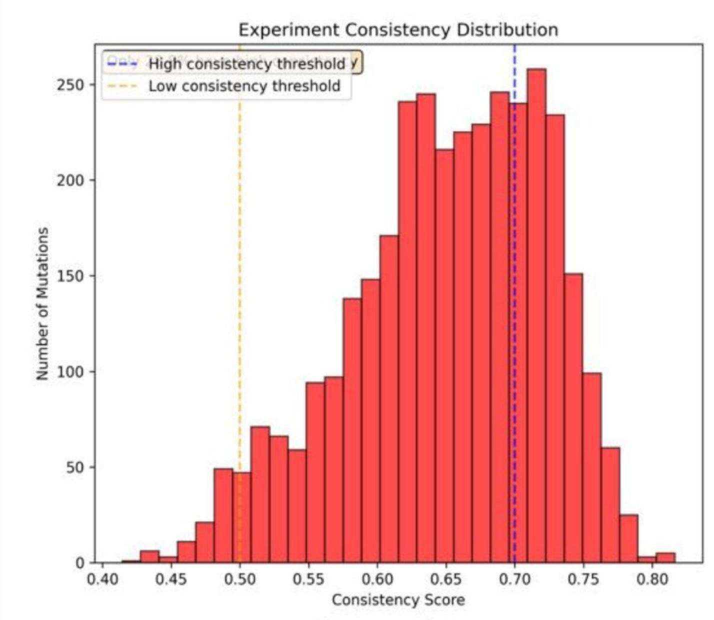
Consistency

Create a Metric:

1 / (1 + Sigma)

Threshold of 70%





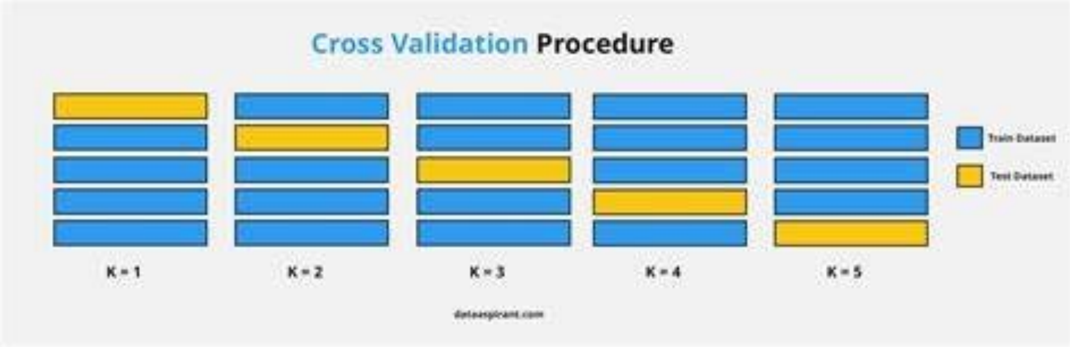
Takeaway One: only 60.2% of the data has high consistency indicating room for improvement

Filling in missing values

Only ~30% of Mutations are well covered (>5) experiments

Tackle filling in with K-Nearest Neighbours with Cross Validation

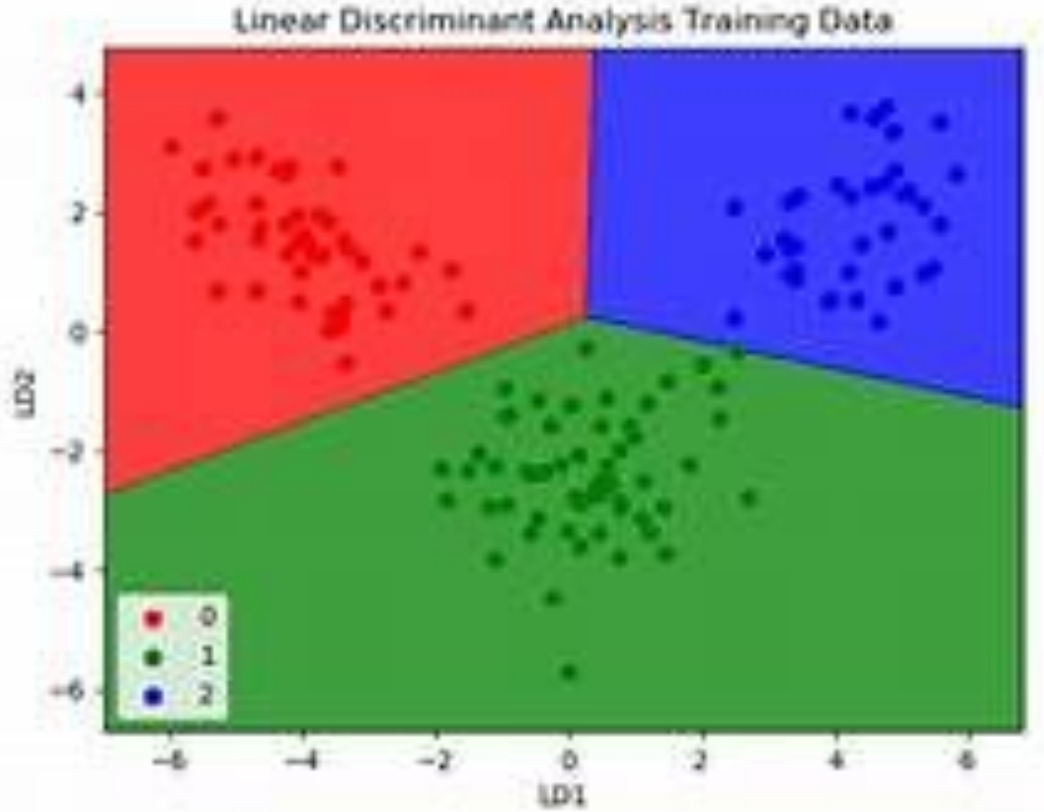
R^2 = 0.953 (Likely overfitted, since data is limited)



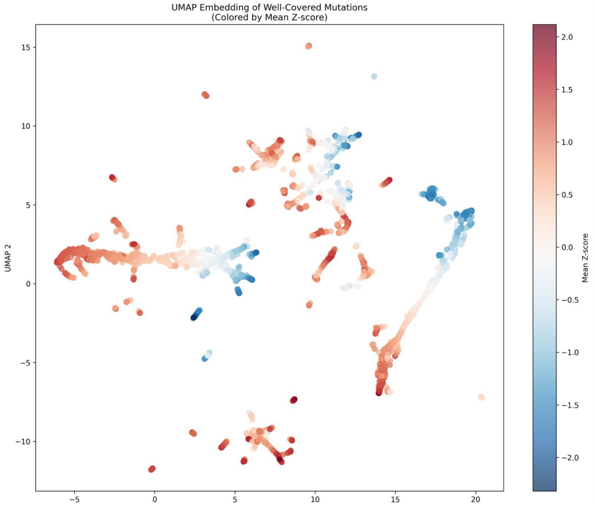
Takeaways/Improvements

Try using a generative model for imputation like a modified version of linear discriminant analysis

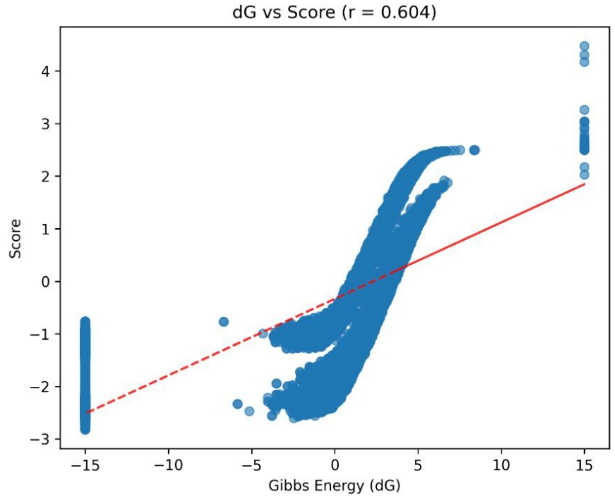
Found Literature that suggests it might follow Dirichlet or Boltzmann Distributions



Look into the geometry of the data


UMAP Embedding of Well-Covered Mutations
(Colored by Mean Z-score)

Look into change in gibbs energy if that is available


dG vs Score (r = 0.604)

Thank you