Deep2Full: Computational strategies to make complementary predictions of the effects of the massively parallel disease or antibiotic resistance causing mutations

Sruthi C. K.
Adviser : Meher K. Prakash
Theoretical Sciences Unit, JNCASR, Bangalore

DMPH 2019, ICTS

Background

Protein

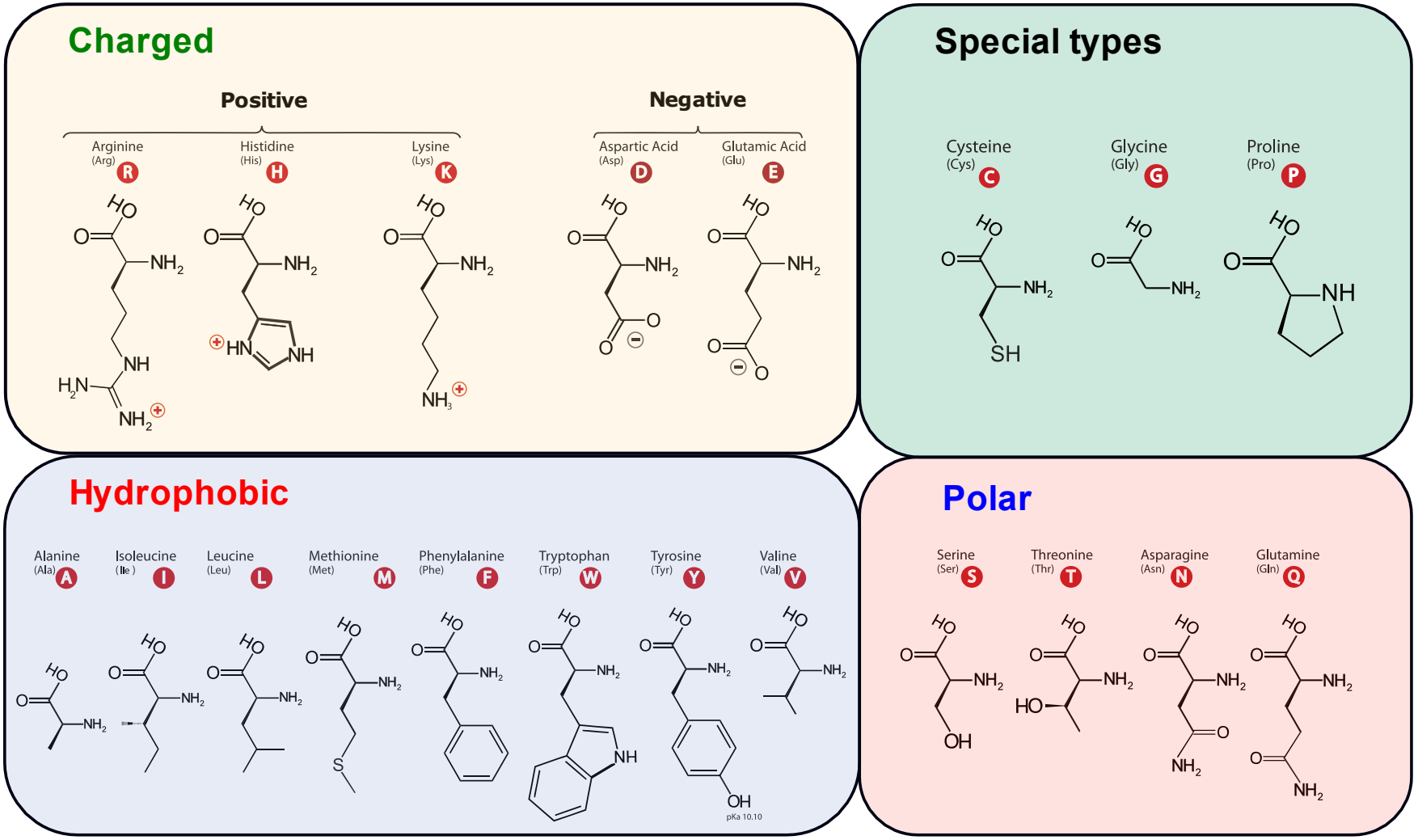
➢ Biological molecule which plays critical roles in cells

➢ Proteins can be enzymes, antibodies, messengers , transporters and also can function as support to structure of cells (structural proteins)

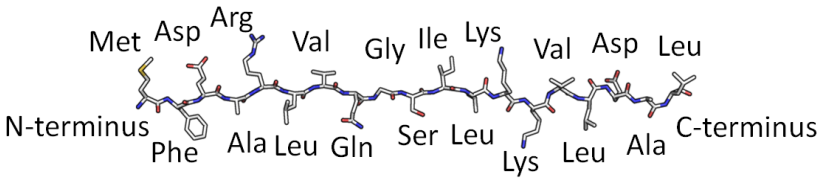
➢ Amino acids connected by peptide bond

➢ 20 amino acids (A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y)

Grouping of amino acids based on charge type

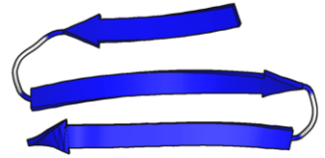


Structure of protein



MFDARLVQGSILKKVLVLDAL

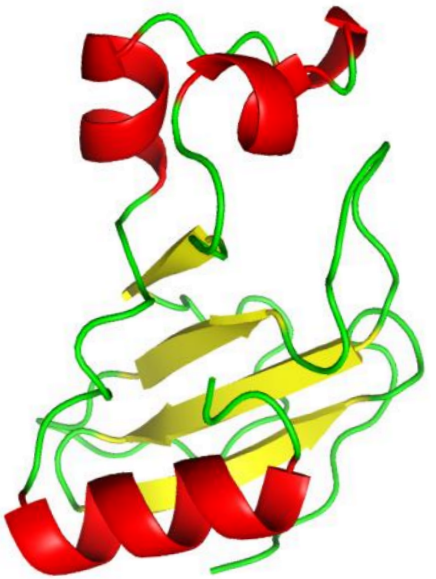
Primary



β-Sheet (3 strands)

α-helix

Secondary



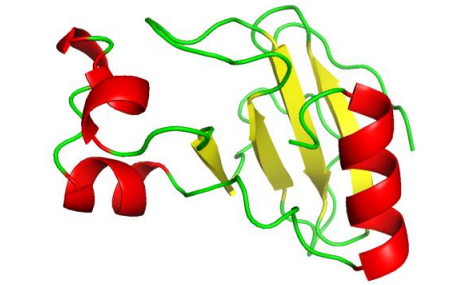
Tertiary

Amino acid mutations

MNNQRKKTARPSFNMLKRARNRVSTVSQLAKRFSKGLL

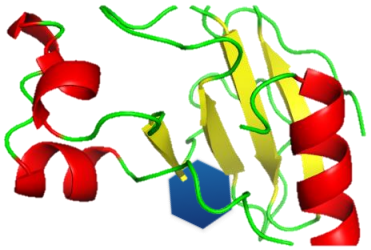
MNNQRKKTARPSFNMLKRARNRVSTVSQLAKRFSK**M**LL

Effect of mutation



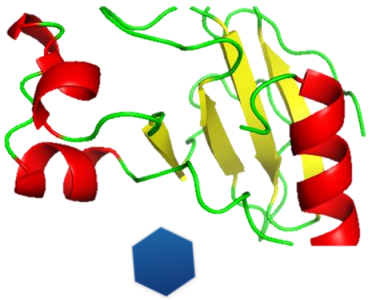
Mutation

Loss of structure



Mutation

Loss of function



Importance of understanding the effect of mutations

1) Disease causing mutations – Eg : Sickle cell anemia, cancers
2) Understanding of drug resistance

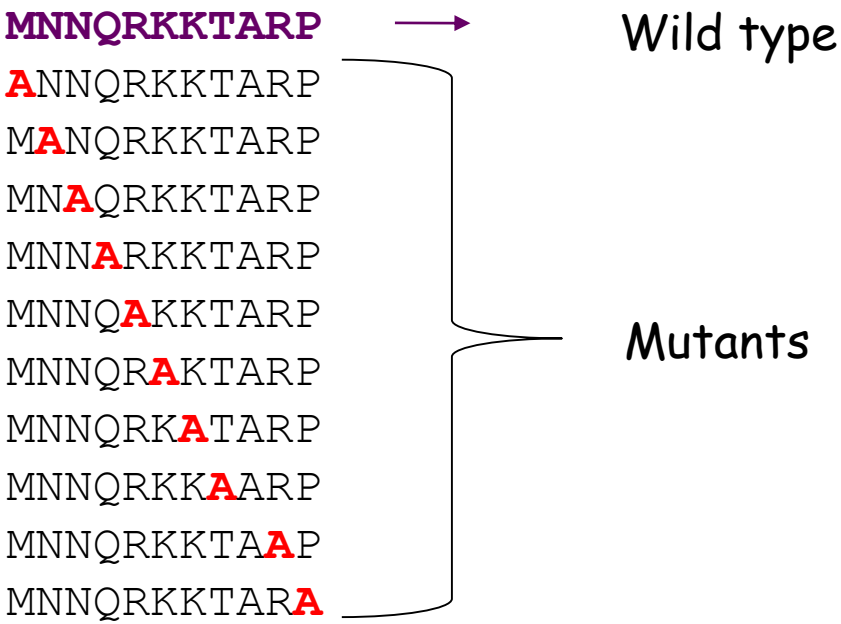
Mutational scan studies

Mutational studies

Old methods : few tens of mutations

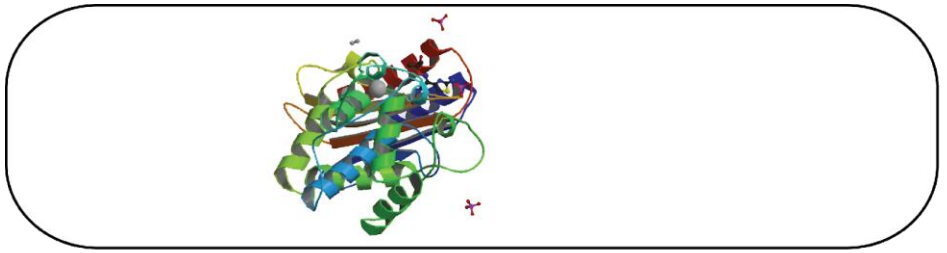
Alanine scan mutagenesis

Substitution of each wild-type amino acid with alanine



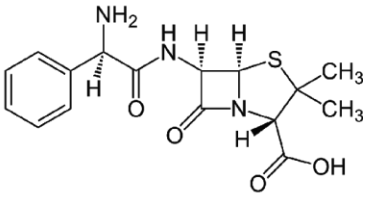
Deep mutational scan : ~10000 mutations

Deep mutational scan

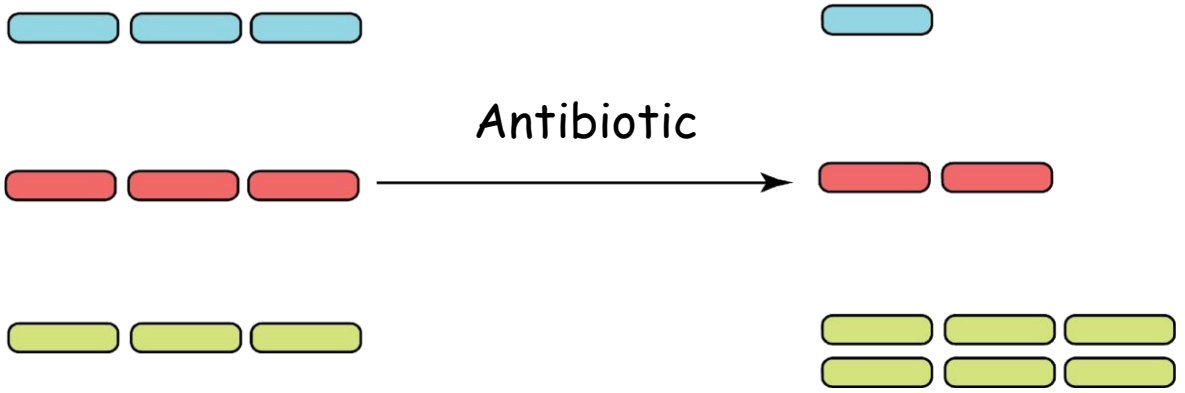


Beta-lactamase in E. coli

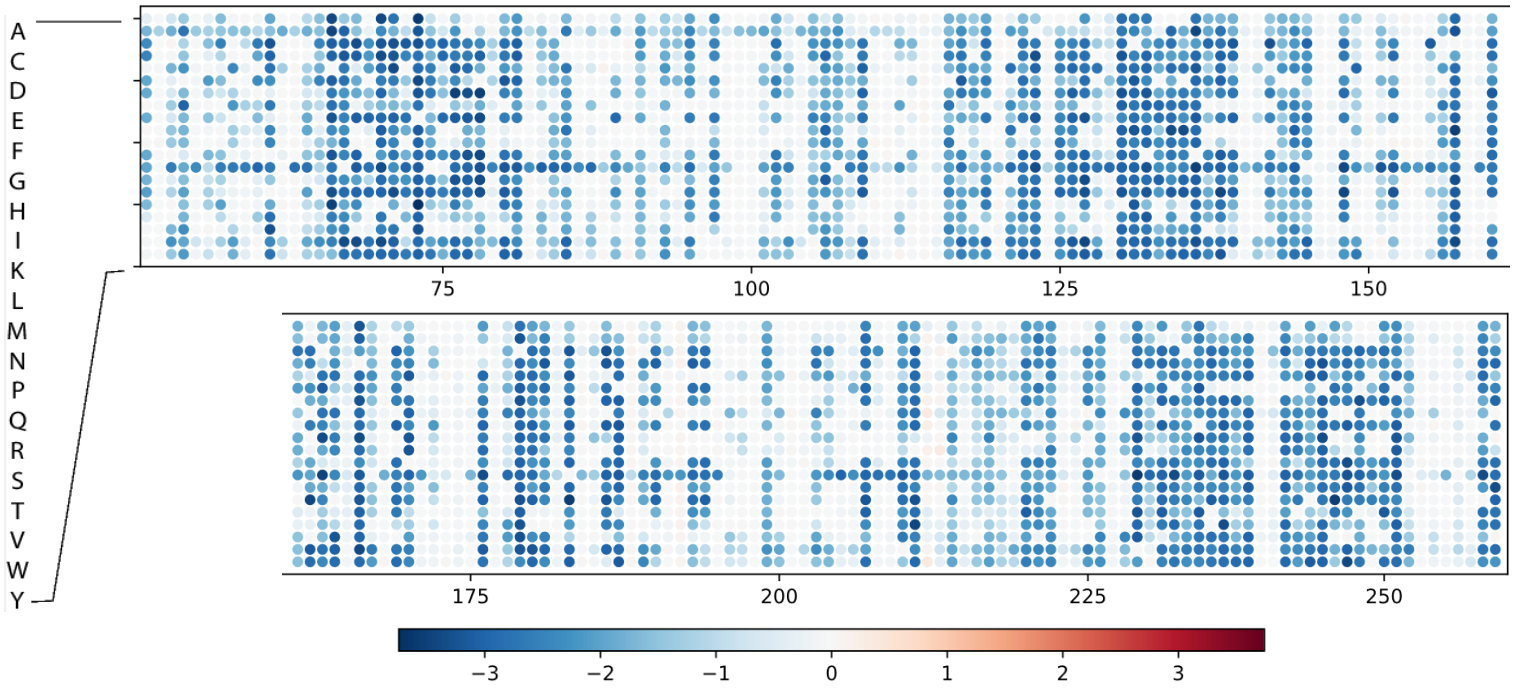
+



Ampicillin



Deep mutational scan



Length of protein = L (=263)
Number of single point substitutions = L*19

Deep mutational scan data : Relative fitness of 4997 single points mutations of TEM-1 beta-lactamase
Selection of mutants : Presence of the antibiotic ampicillin

M. A. Stiffler, D. R. Hekstra, and R. Ranganathan, "Evolvability as a function of purifying selection in tem-1 β-lactamase," Cell, vol. 160, no. 5, pp. 882–892, 2015

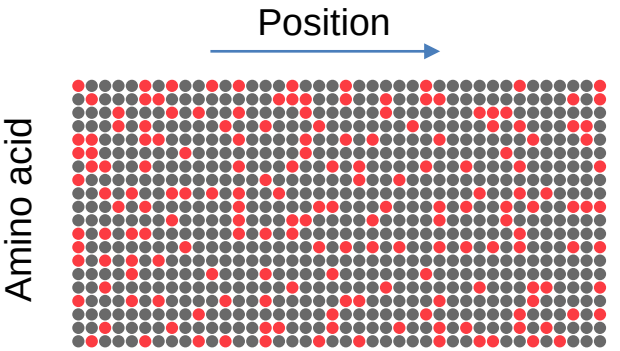
Do we need all of it?

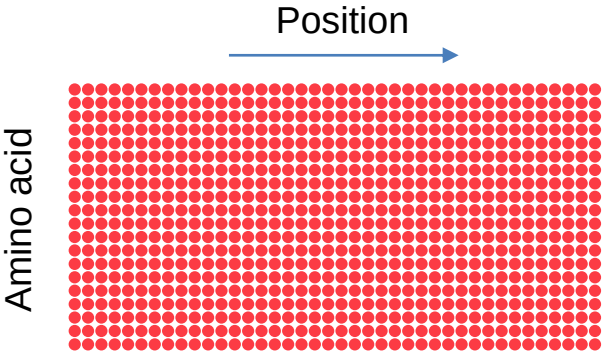
Predicting mutational effect

Mutational effect scores
for a subset of mutants

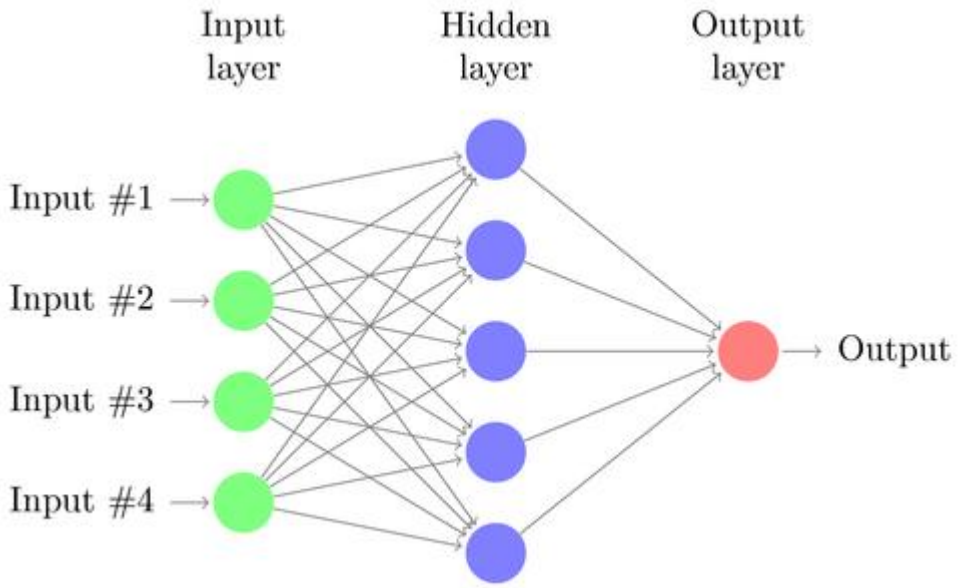
Computational models

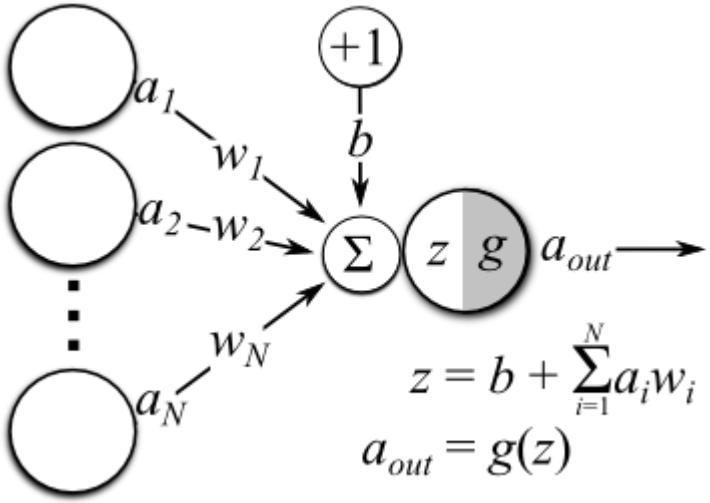
Mutational effect scores
for all other protein variants





Neural Network model





Division of data set

Training : Trained on this set to generate model
Validation : Prevents overfitting
Prediction : To check the predictability on a completely new set of data

Structure-sequence variables

(1) Solvent accessible surface area (SASA)
(2) Secondary structure
(3) Number of structural contacts
(4) Average commute time

Structure

(5) BLOSUM substitution matrix
(6) Hydrophobicity of the mutant
(7) Hydrophobicity of the amino acid in the wild type
(8) Position specific substitution matrix (PSSM) score for the amino acid after mutation,
(9) PSSM score for the wild-type amino acid
(10) Conservation of the amino acid.

Sequence

(11) Average correlation
(12) Degree centrality
(13) Betweenness centrality
(14) Closeness centrality
(15) Eigenvector centrality
(16) Impact factor
(17) Dependency factor

Co-evolution

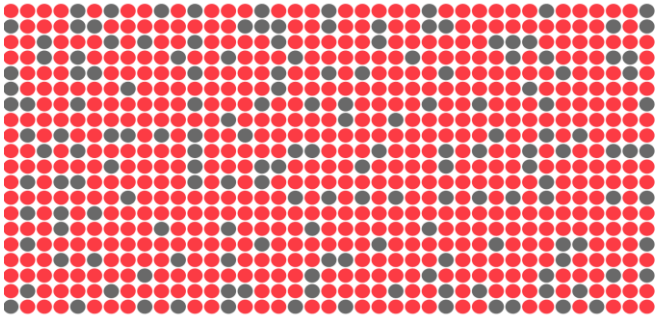
Which is the best and minimal experimental data to train the model ?

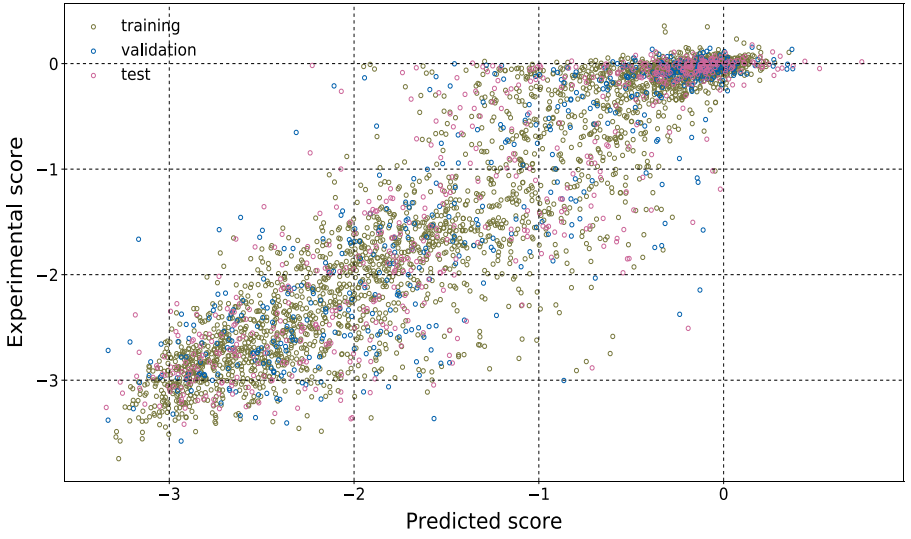
Random scan

Training set - Randomly selected mutations

Random mutagenesis

Random 85%

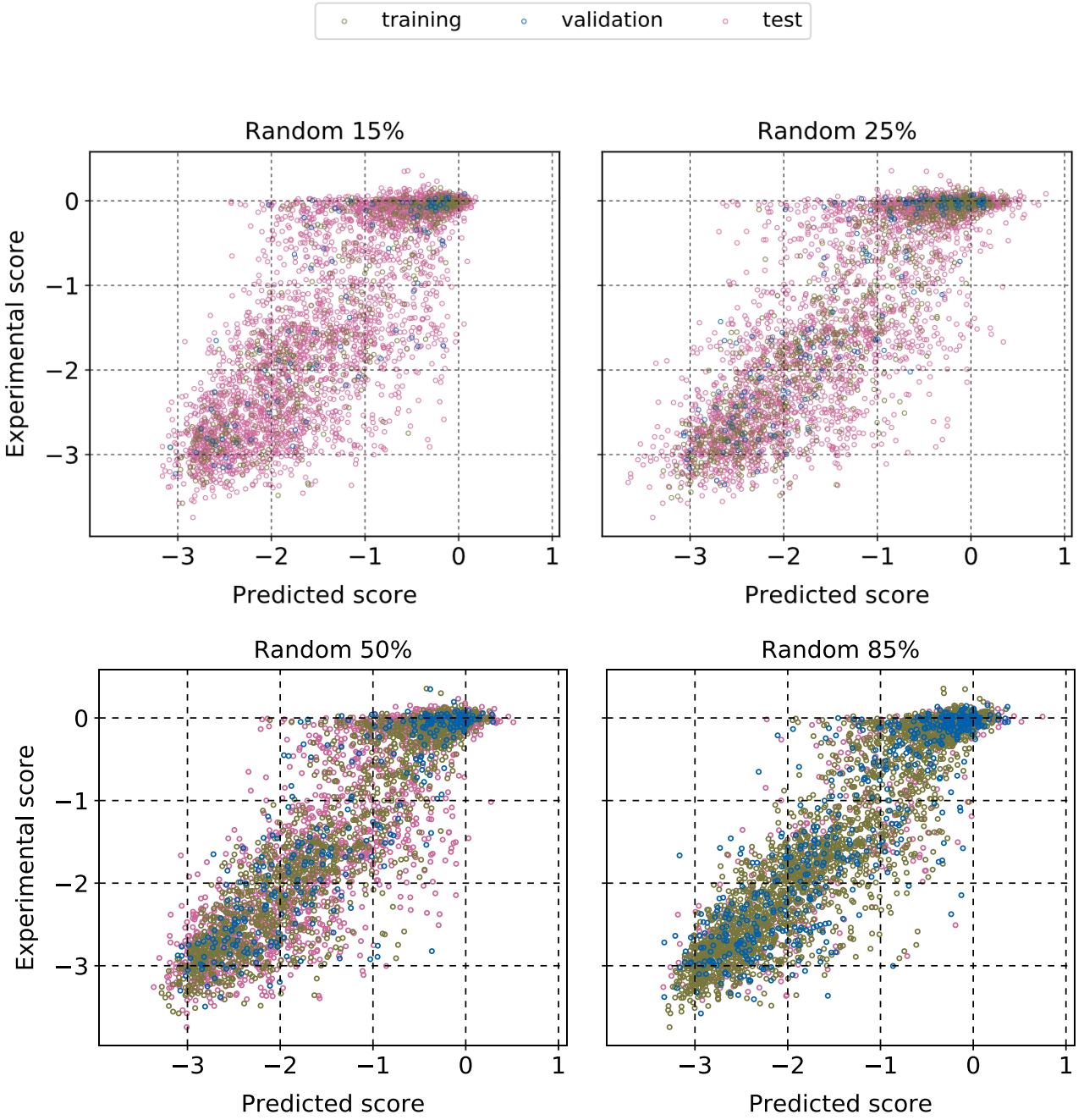




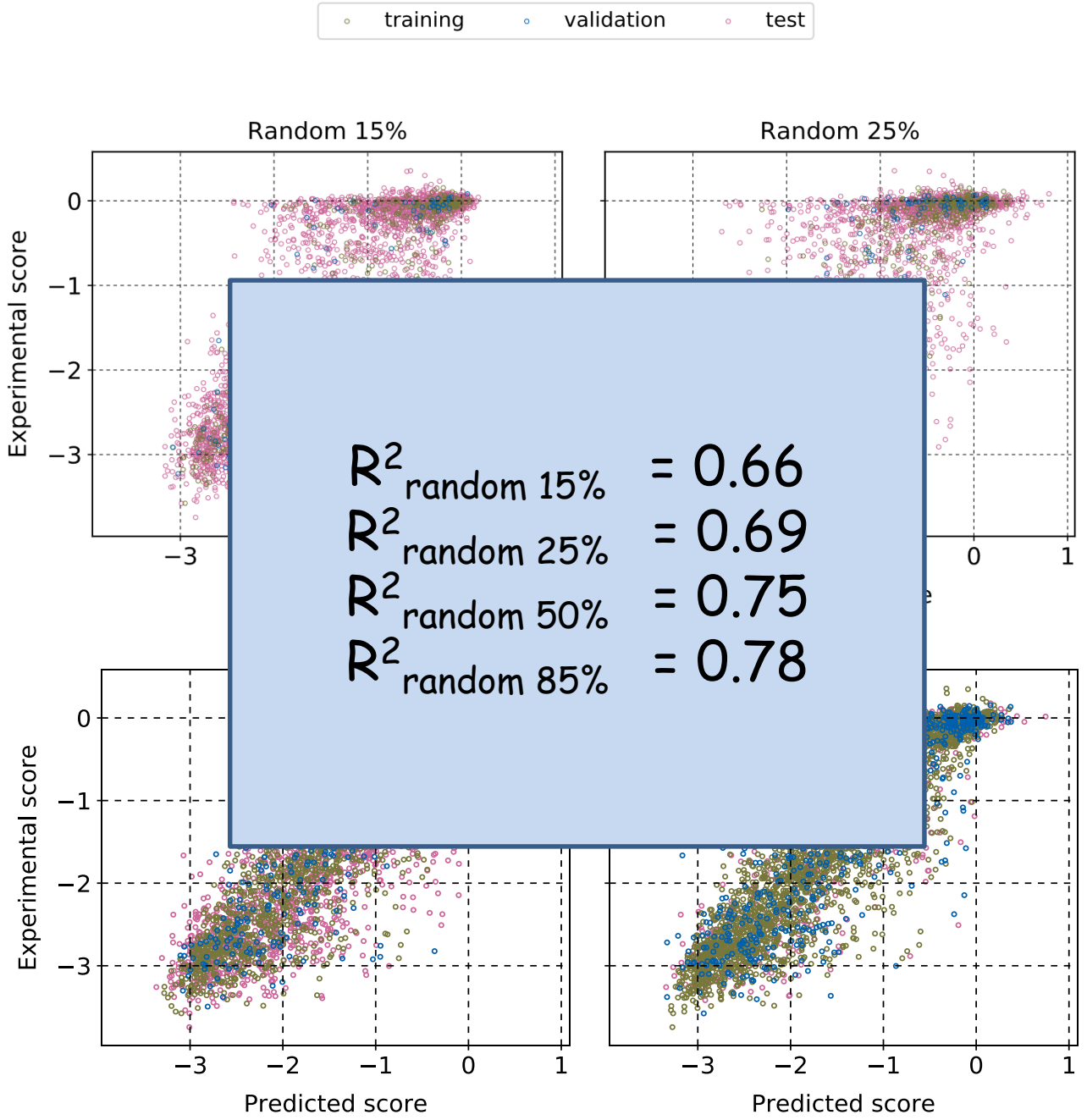
Prediction using the model trained on 85% data

$R^2_{training}$ = 0.87
$R^2_{validation}$ = 0.78
R^2_{test} = 0.78

Varying the training data set size



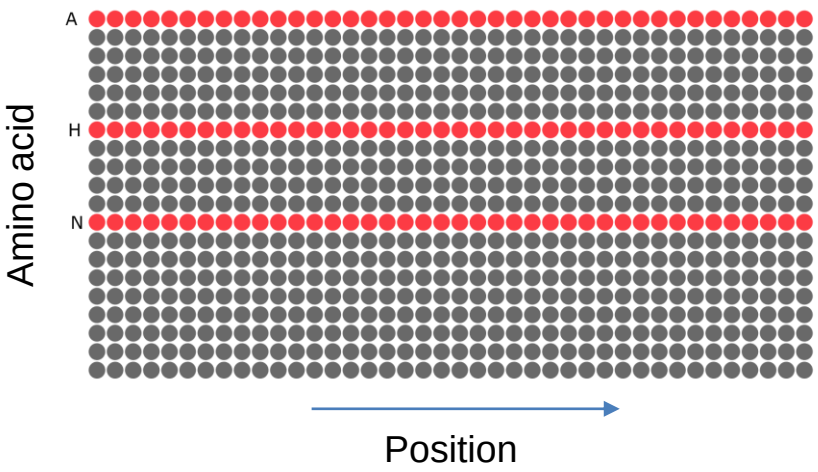
Varying the training data set size

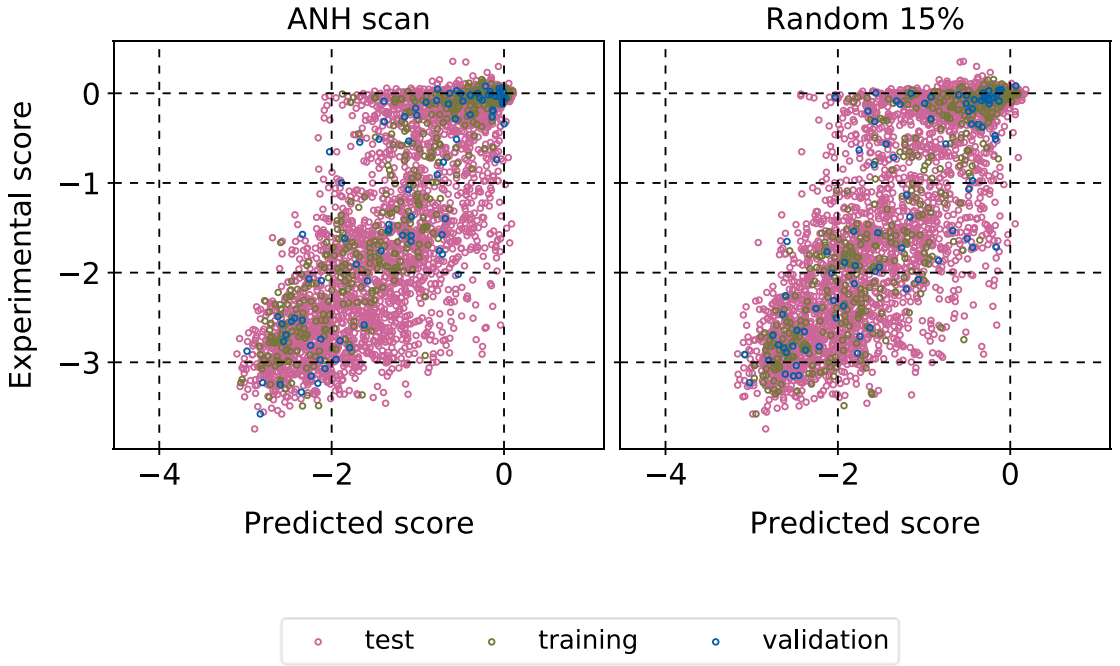


ANH scan (15%)

Training set – substitutions to **Alanine(A), Asparagine(N) and Histidine(H)**

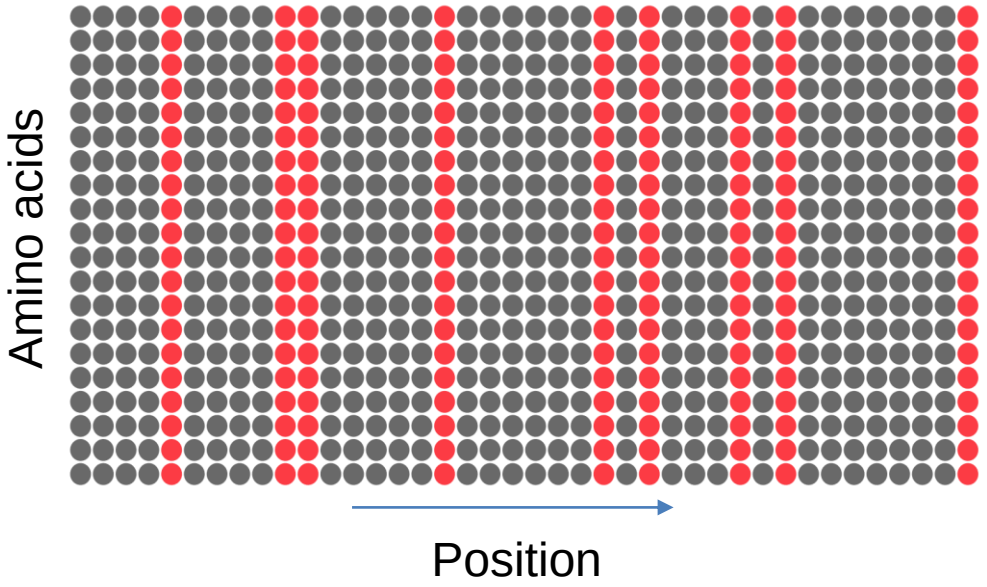
Site-directed mutagenesis





Position scans

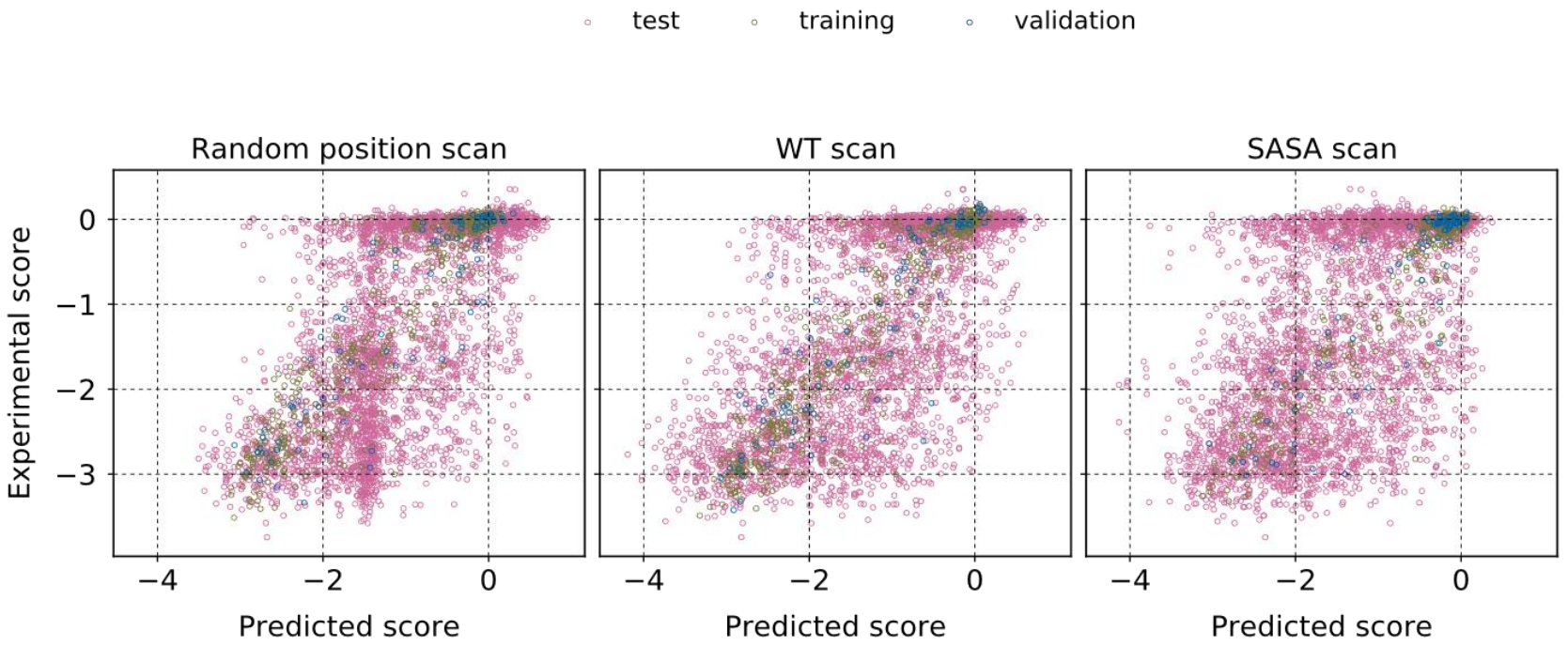
Training set – all 19 substitutions at a few chosen sites



How to choose positions

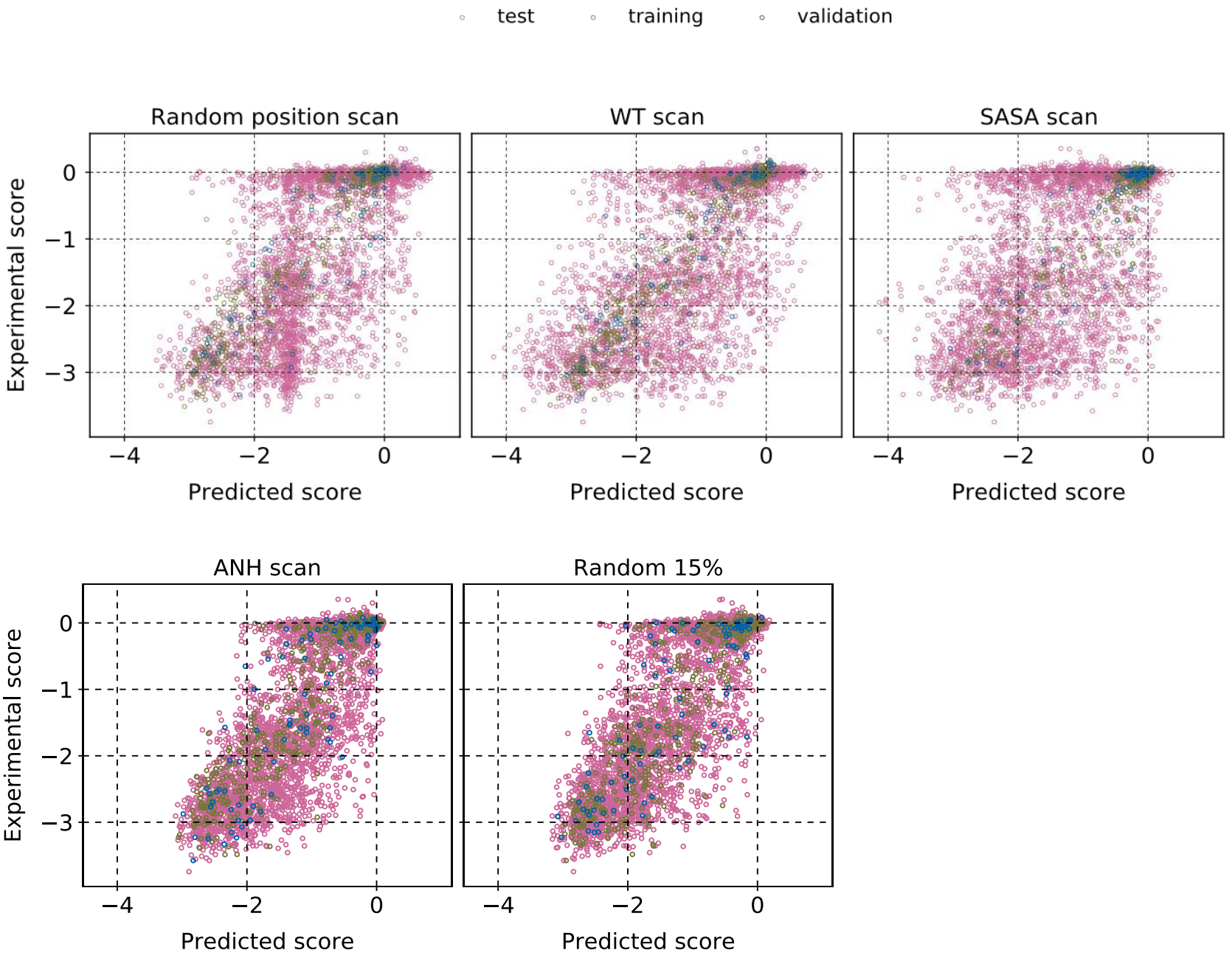
- Random (Random position scan)
- Based on the solvent exposure (SASA scan)
- Based on the wild type amino acid (WT scan)

Position scans

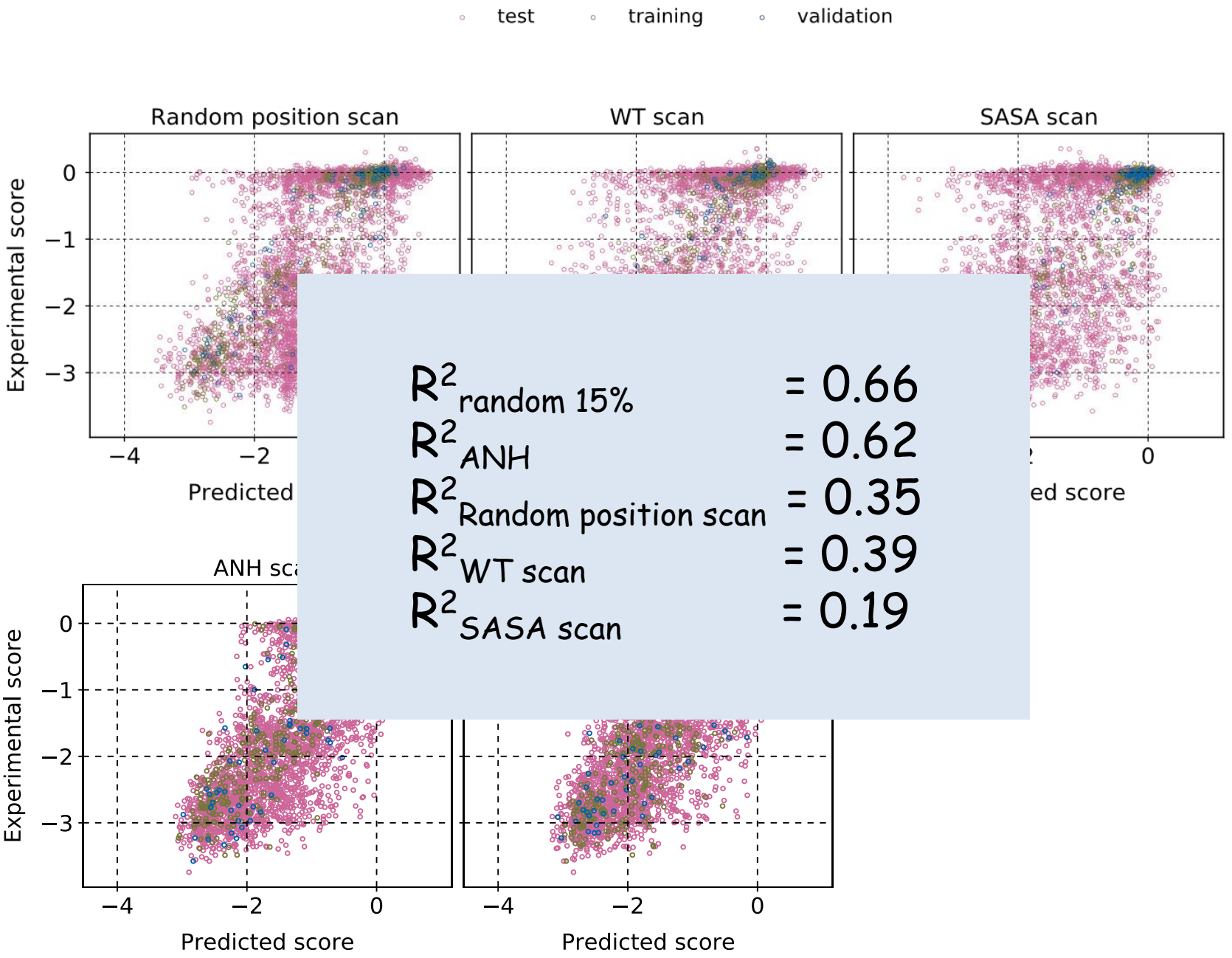


How do the predictive abilities of the models compare?

Position scans



Position scans



$R^2_{\text{random 15\%}} = 0.66$

$R^2_{\text{ANH}} = 0.62$

$R^2_{\text{Random position scan}} = 0.35$

$R^2_{\text{WT scan}} = 0.39$

$R^2_{\text{SASA scan}} = 0.19$

Prediction qualities quantified using Pearson correlation

Protein/ Scan	Random 15%	Random 25%	Random 50%	Random 85%
beta-lactamase	0.81	0.83	0.87	0.89
APH(3')-II	0.69	0.68	0.72	0.78
Hsp90	0.72	0.77	0.82	0.85
MAPK1	0.62	0.63	0.74	0.77
UBE2I	0.52	0.59	0.66	0.67
TPK1	0.24	0.23	0.26	0.42

Prediction qualities quantified using Pearson correlation

Protein/ Scan	ANH scan	Random 15%	Random position scan	WT scan	SASA scan
beta-lactamase	0.8	0.81	0.65	0.67	0.53
APH(3')-II	0.67	0.69	0.52	0.54	0.49
Hsp90	0.75	0.72	0.3	0.37	0.50
MAPK1	0.62	0.62	0.31	0.39	0.33
UBE2I	0.56	0.52	0.2	0.32	0.31
TPK1	0.25	0.24	0.13	0.19	0.10

Conclusions

- 15% of data could be useful for predicting the remaining experiments.

- This 15% can be random choice or what is called alanine-asparagine-histidine (ANH) scan we conceptualized.

References

1) M. A. Stiffler, D. R. Hekstra, and R. Ranganathan, "Evolvability as a function of purifying selection in tem-1 β-lactamase," Cell, vol. 160, no. 5, pp. 882–892, 2015

2) D. M. Fowler and S. Fields, "Deep mutational scanning: a new style of protein science," Nature methods,
vol. 11, no. 8, p. 801, 2014.

How does this biochemistry feed into public health models?

Acknowledgement

Thank You

Protein	Num. of amino acids.	Data availability
Beta-lactamase	263/208	4997/3952
AGK	264	4234
MAPK1	360	4470
HSP90	629/219	4021
TPK1	243	3181
UBE2I	159	2563