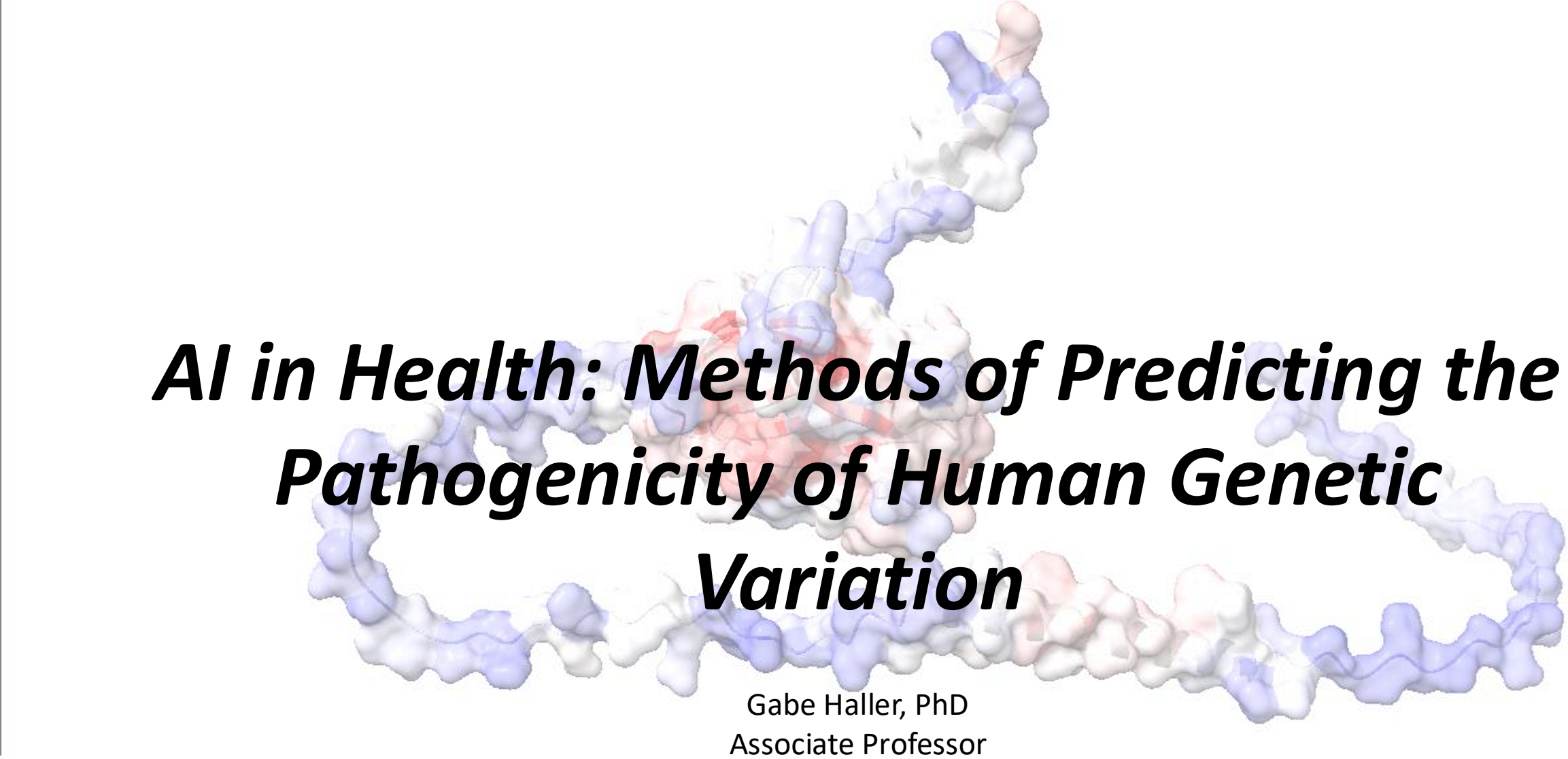


AI in Health: Methods of Predicting the Pathogenicity of Human Genetic Variation

Gabe Haller, PhD
Associate Professor
Departments of Neurosurgery, Neurology and Genetics
September 15, 2026

Current Methods of Pathogenicity Prediction

- **AlphaMissense**
  - AlphaMissense is a deep learning model by DeepMind built on AlphaFold's protein structure predictions. It estimates the pathogenicity of missense variants by considering structural and evolutionary information.

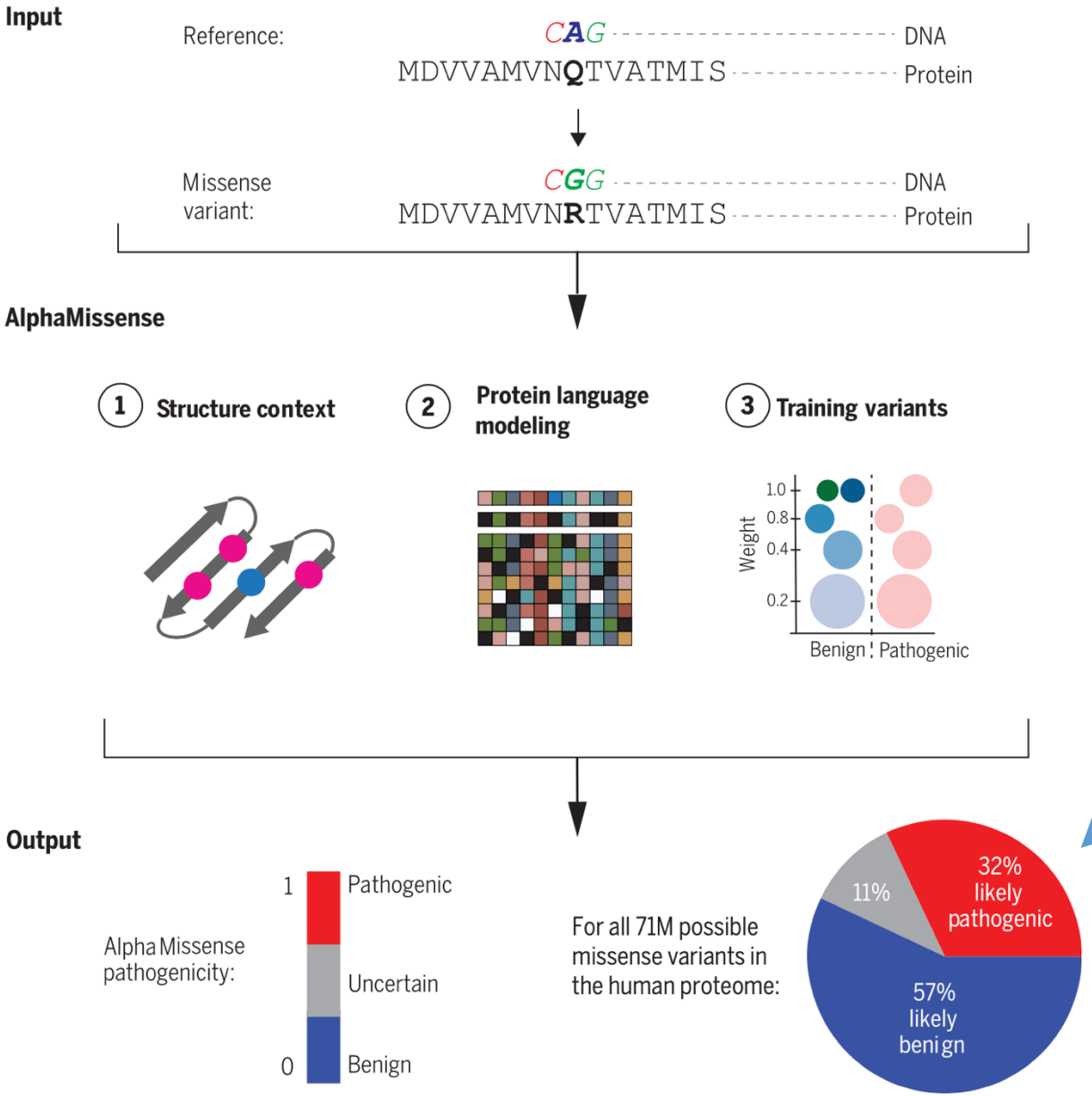
- **REVEL** (Rare Exome Variant Ensemble Learner)
  - Ensemble method combining outputs from multiple tools (e.g., PolyPhen-2, SIFT, MutationTaster). It predicts missense variant pathogenicity with high accuracy. Scores range 0-1, with higher values more likely pathogenic.

- **CADD** (Combined Annotation Dependent Depletion)
  - CADD integrates multiple annotations into a single metric to estimate the relative pathogenicity of variants. It uses machine learning trained on simulated and real variants. Scores >20 suggest potential pathogenicity.

- **PolyPhen-2** (Polymorphism Phenotyping v2)
  - PolyPhen-2 predicts the effect of amino acid substitutions on protein function using sequence conservation, structural features, and functional annotations. Output: benign, possibly damaging, probably damaging.

- **SIFT** (Sorting Intolerant From Tolerant)
  - SIFT predicts whether an amino acid substitution affects protein function based on sequence homology and conservation. Scores ≤0.05 indicate deleterious variants, >0.05 tolerated.

AlphaMissense

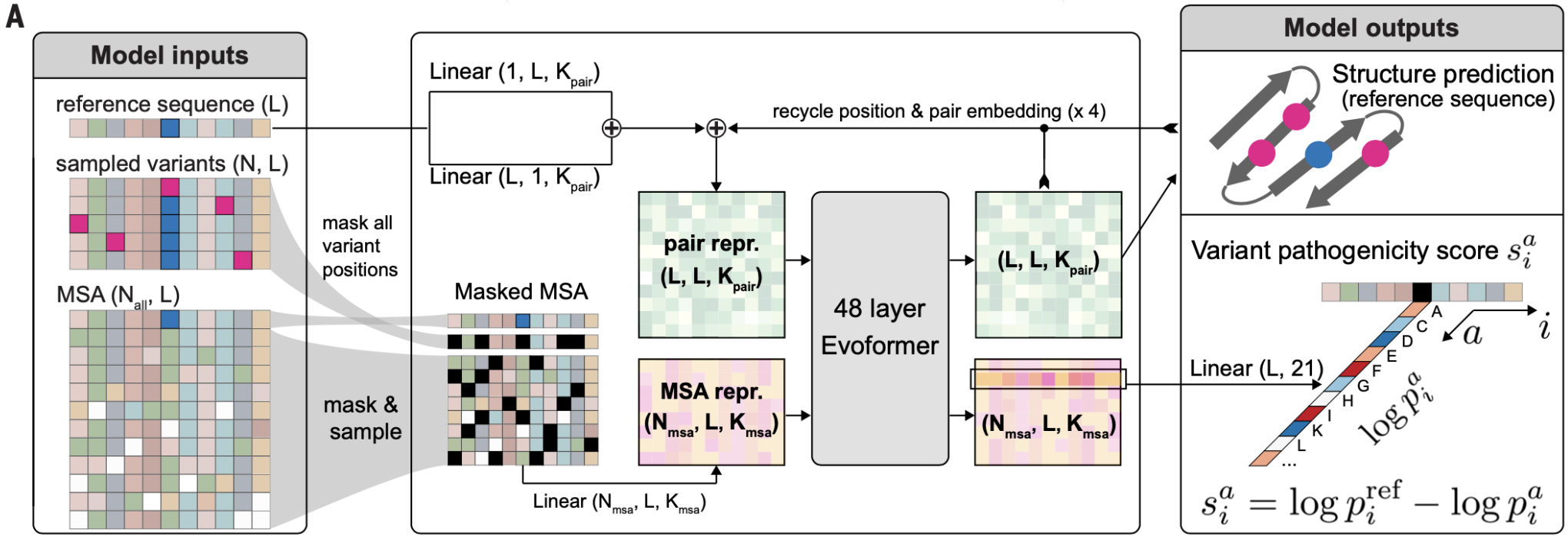


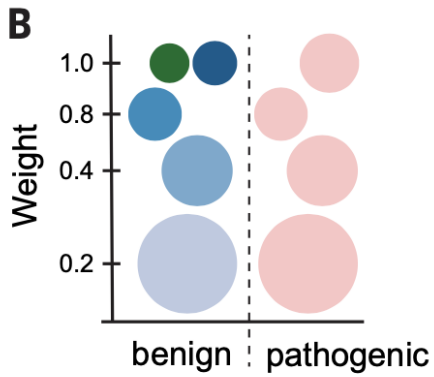
A bit high. 12-20% of DMS results show pathogenic function scores

AlphaMissense Motivation

- Prediction of the pathogenicity of novel variants is essential for precision medicine
- Prediction methods remain highly inaccurate
- MAVE data is growing as a source of positive control data
- Integrating known pathogenic/benign variants in prediction leads to leaks between training and test sets (previous methods)
- Training only on frequency (assuming observed variants are benign and unobserved are pathogenic) is riddled with false positives and false negatives (previous methods)

Overview of AlphaMissense





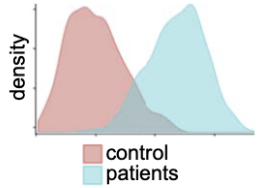
Primates:
- MAF > 1e-3, total 9.7e4

Humans:
- MAF > 2e-4, total 1.2e5
- 1e-5 < MAF < 2e-4, total 1.1e6
- sample from unobserved, match counts of benign

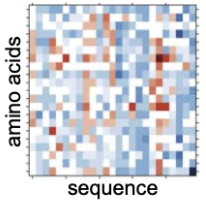
C



Known benign and pathogenic variants (ClinVar)

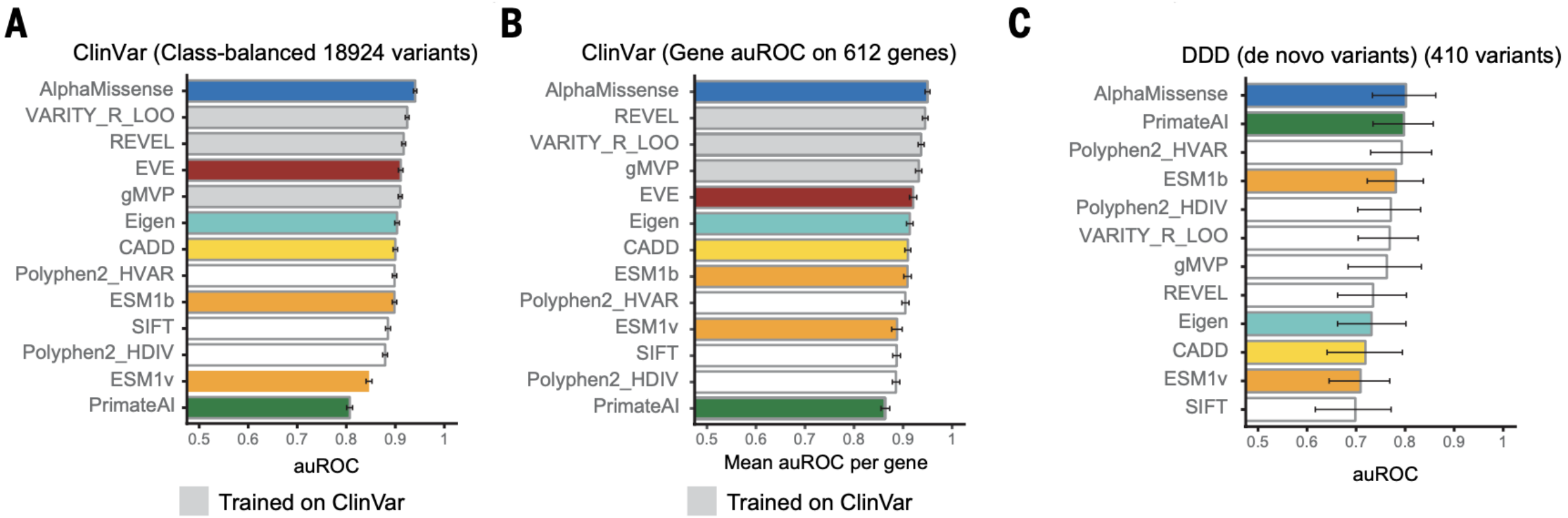


De novo variants of rare disease



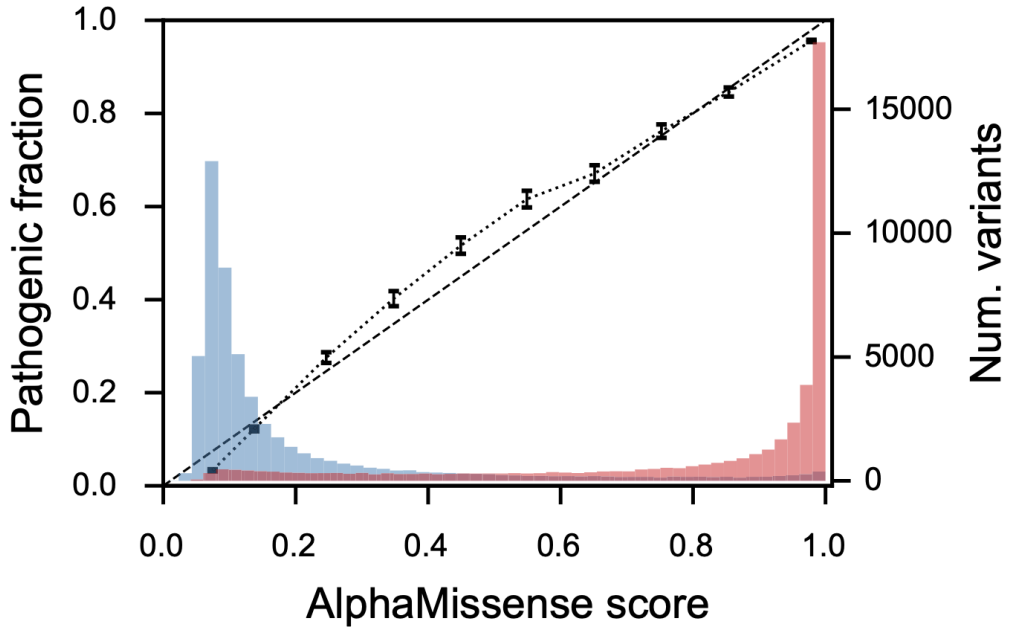
Multiplexed assay of variant effect (ProteinGym)

Figure 2: Performance of AlphaMissense on clinically curated classification benchmarks.

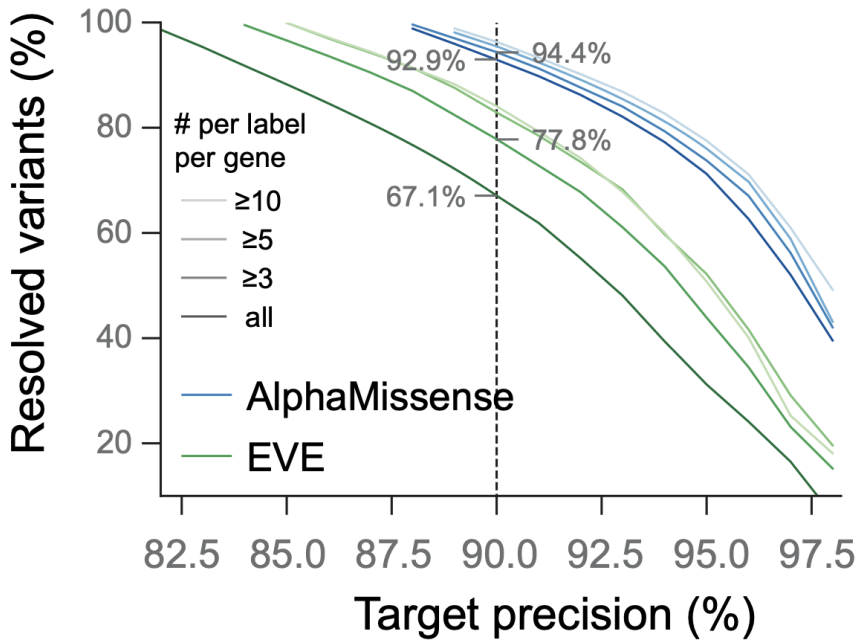


Performance of AlphaMissense on clinically curated classification benchmarks.

D

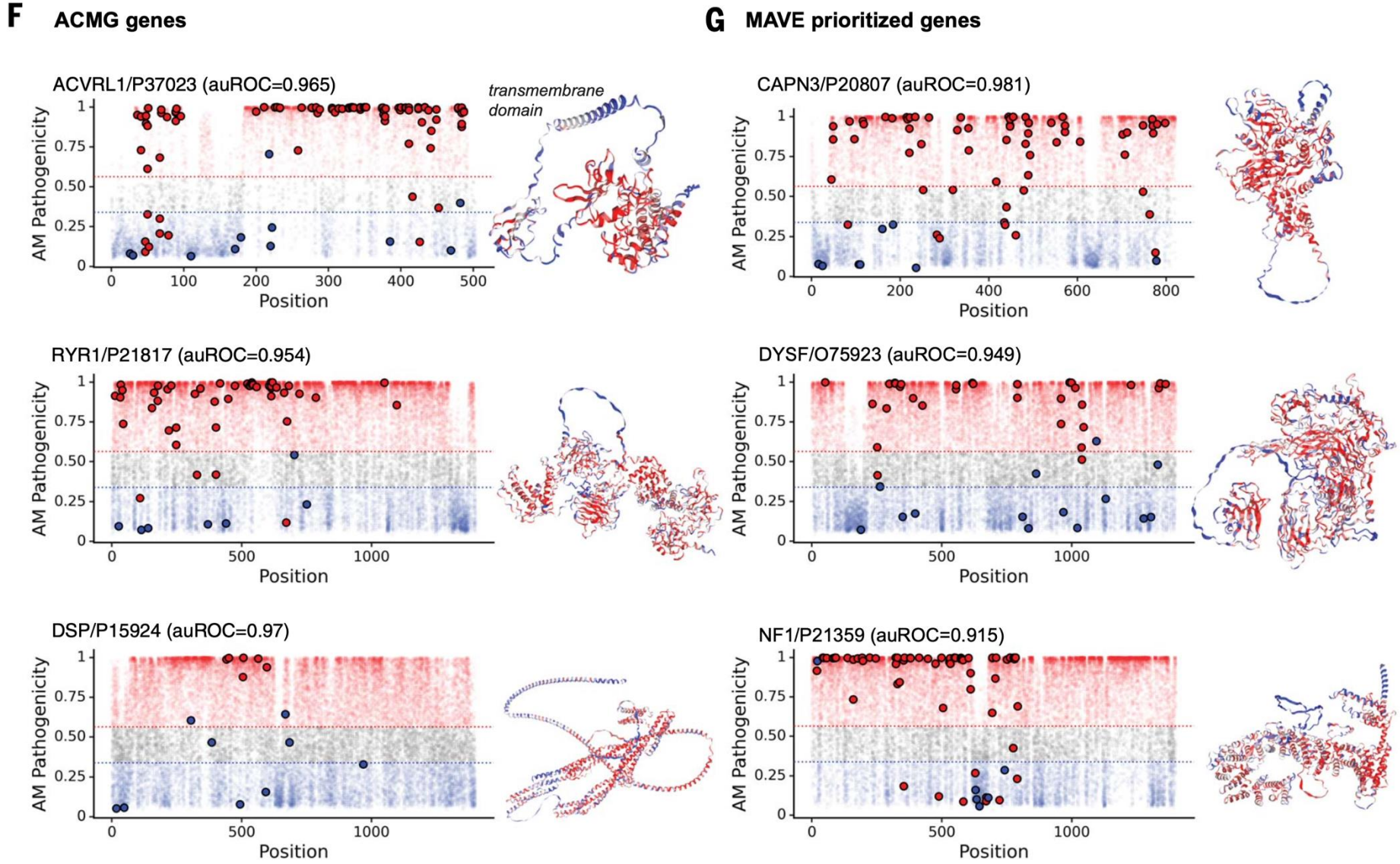


E

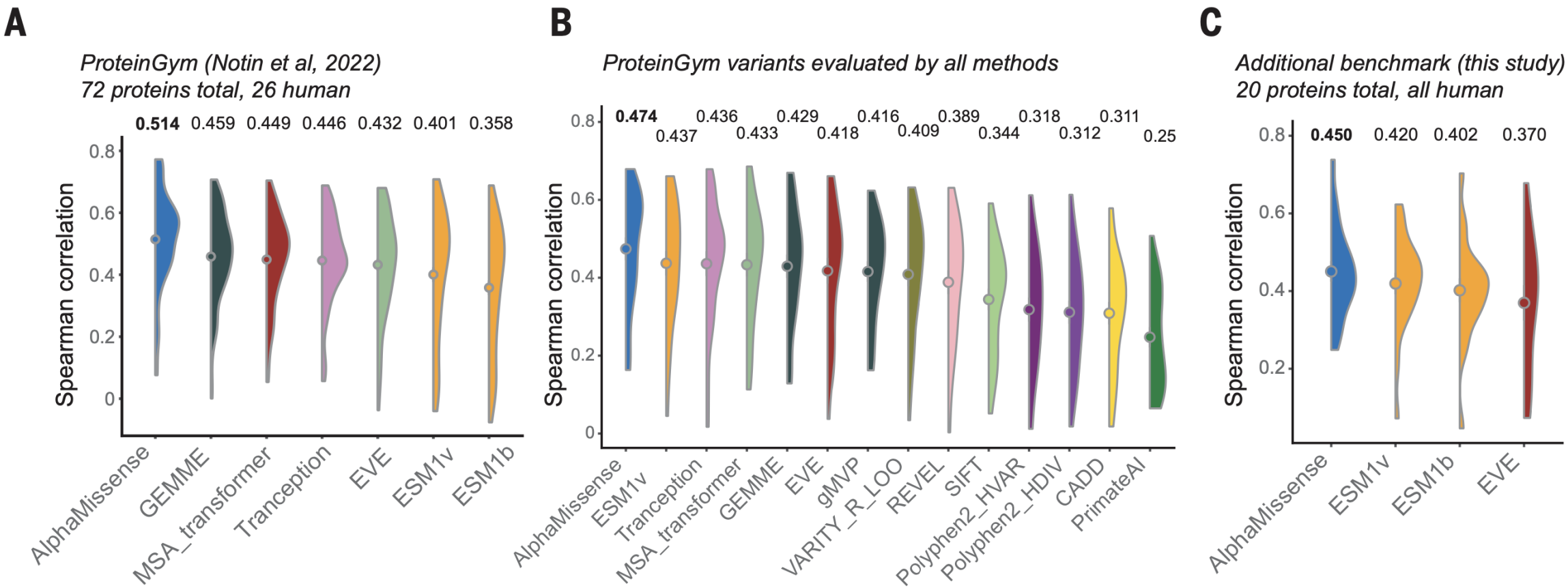


* Precision = True Positive/(All Positive)

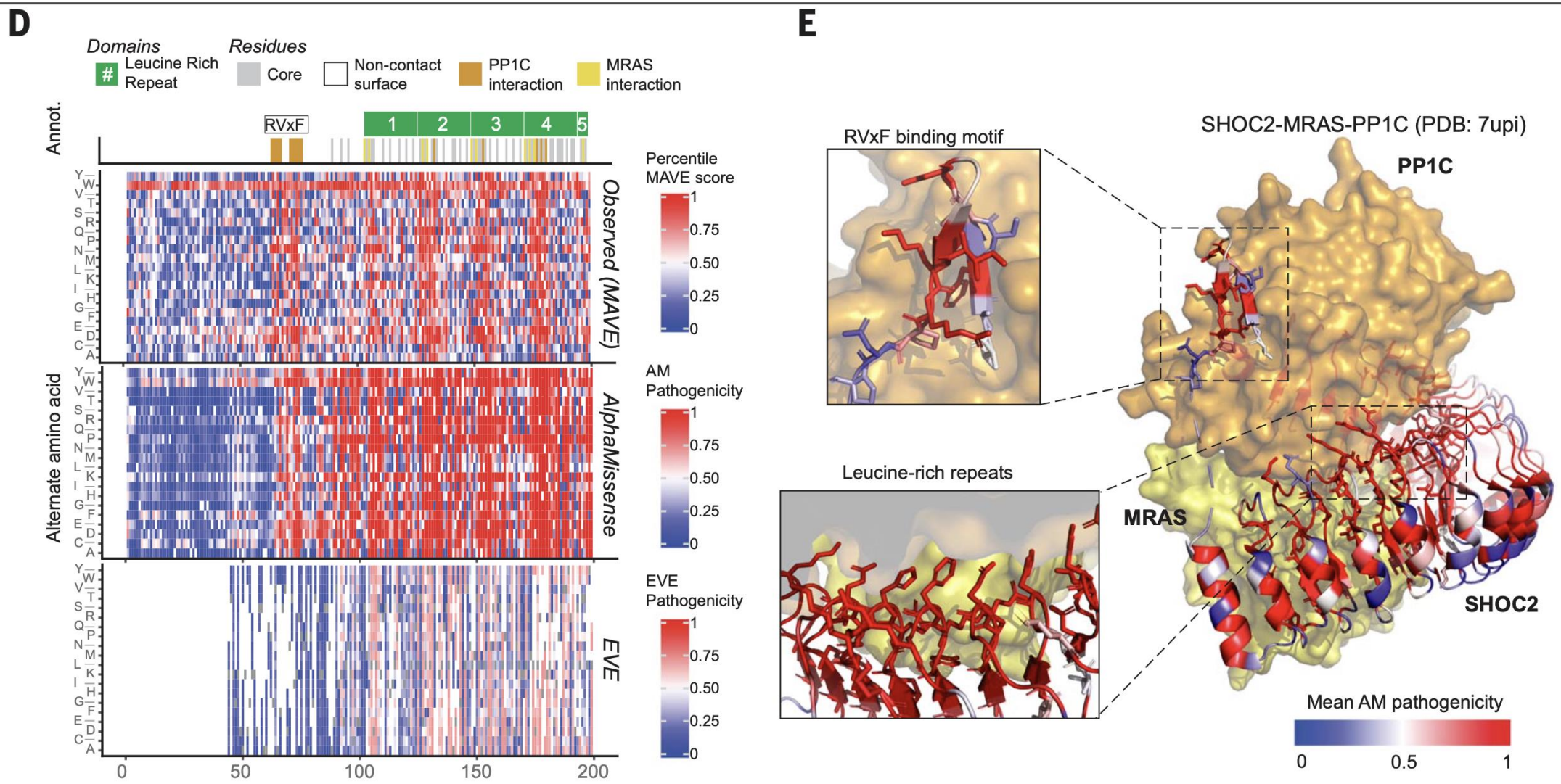
Performance of AlphaMissense on clinically curated classification benchmarks.



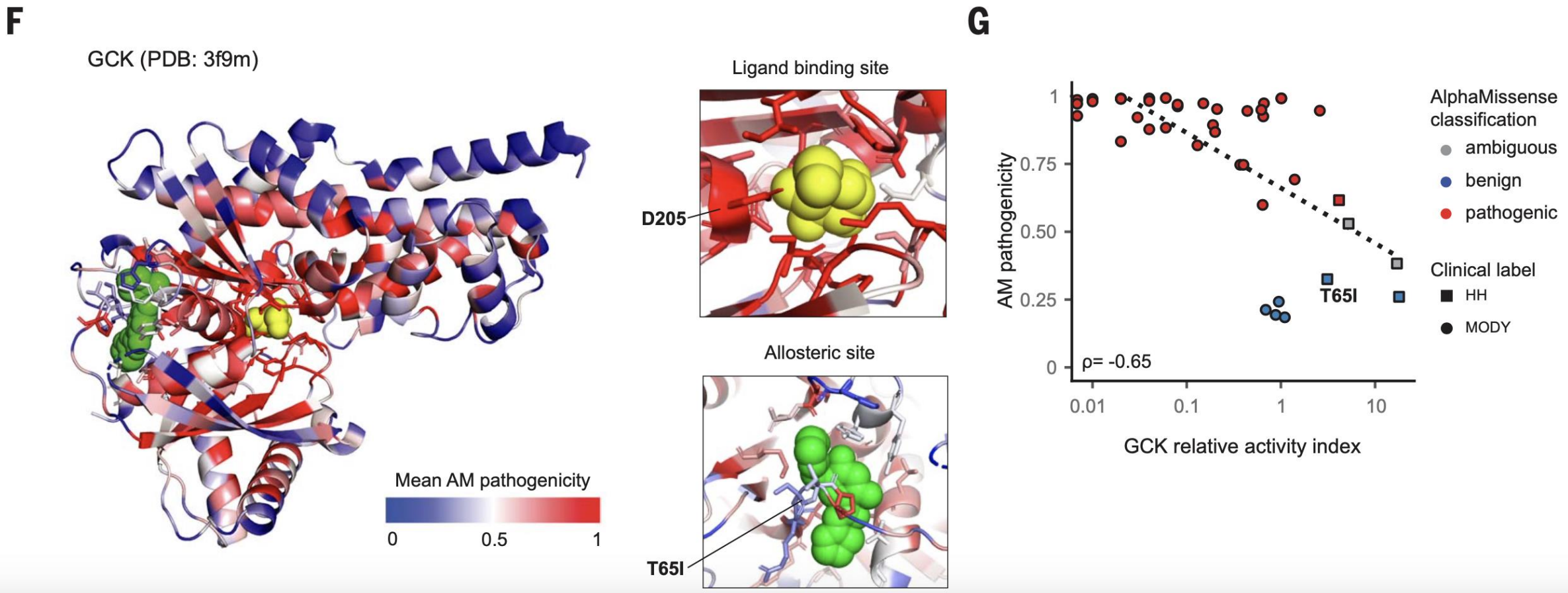
AlphaMissense achieves state-of-the-art agreement with multi-plexed assays of variant effect.



AlphaMissense achieves state-of-the-art agreement with multi-plexed assays of variant effect.



AlphaMissense achieves state-of-the-art agreement with multi-plexed assays of variant effect.



Blatant Self-advertisement:

The Journal of Clinical Investigation

RESEARCH ARTICLE

Comprehensive functional characterization of *SGCB* coding variants predicts pathogenicity in limb-girdle muscular dystrophy type R4/2E

Chengcheng Li,[1] Jackson Wilborn,[2] Sara Pittman,[1] Jil Daw,[1] Jorge Alonso-Pérez,[3] Jordi Díaz-Manera,[4] Conrad C. Weihl,[1] and Gabe Haller[1,2,5]

[1]Department of Neurology and [2]Department of Neurosurgery, Washington University School of Medicine, St. Louis, Missouri, USA. [3]Neuromuscular Disease Unit, Neurology Department, Hospital Universitario Nuestra Señora de Candelaria, Fundación Canaria Instituto de Investigación Sanitaria de Canarias, Tenerife, Spain. [4]John Walton Muscular Dystrophy Research Center, Newcastle University, Newcastle Upon Tyne, United Kingdom. [5]Department of Genetics, Washington University School of Medicine, St. Louis, Missouri, USA.

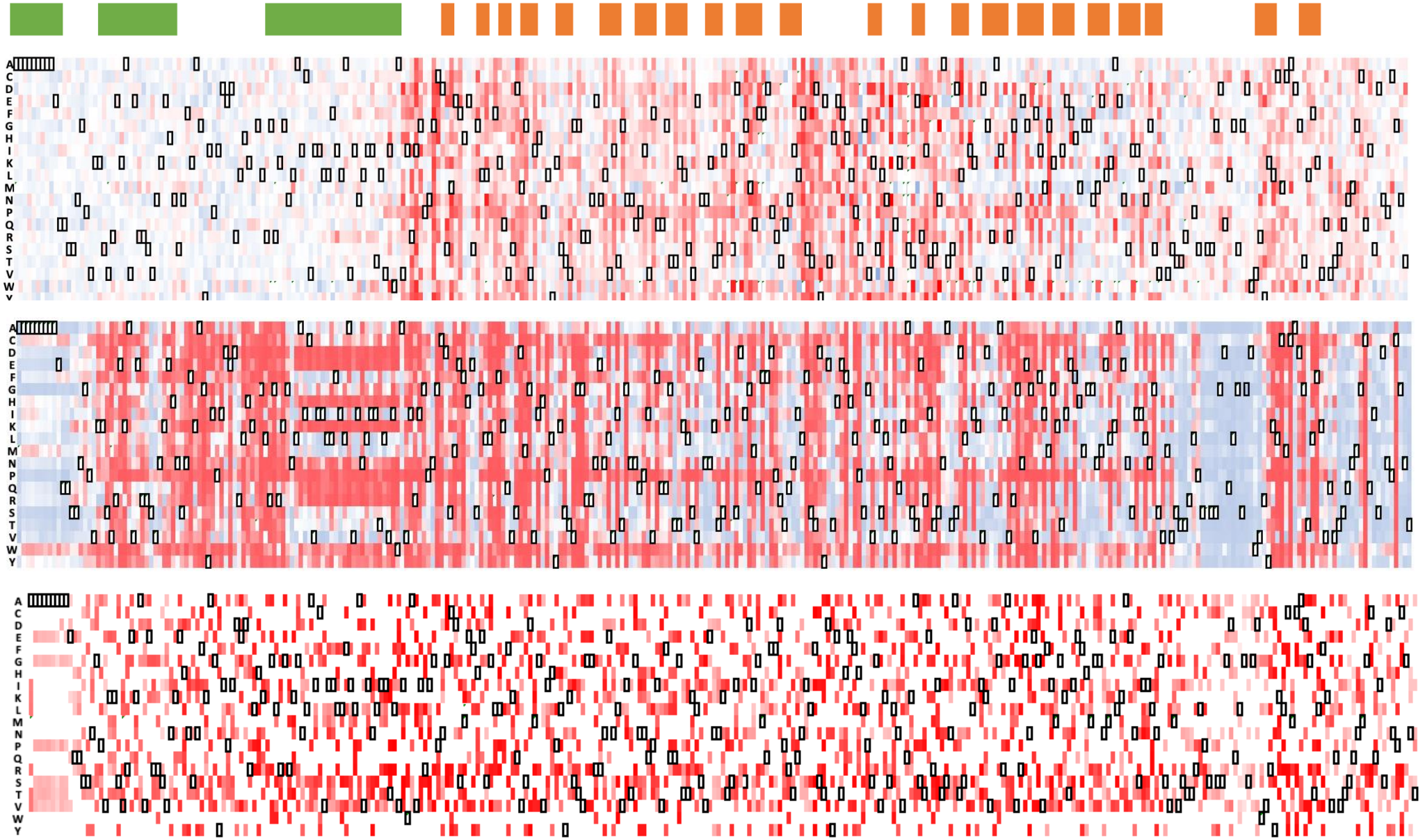
Comparison of SGCB DMS Functional Score Predictors

Alpha-helix
Beta-sheet

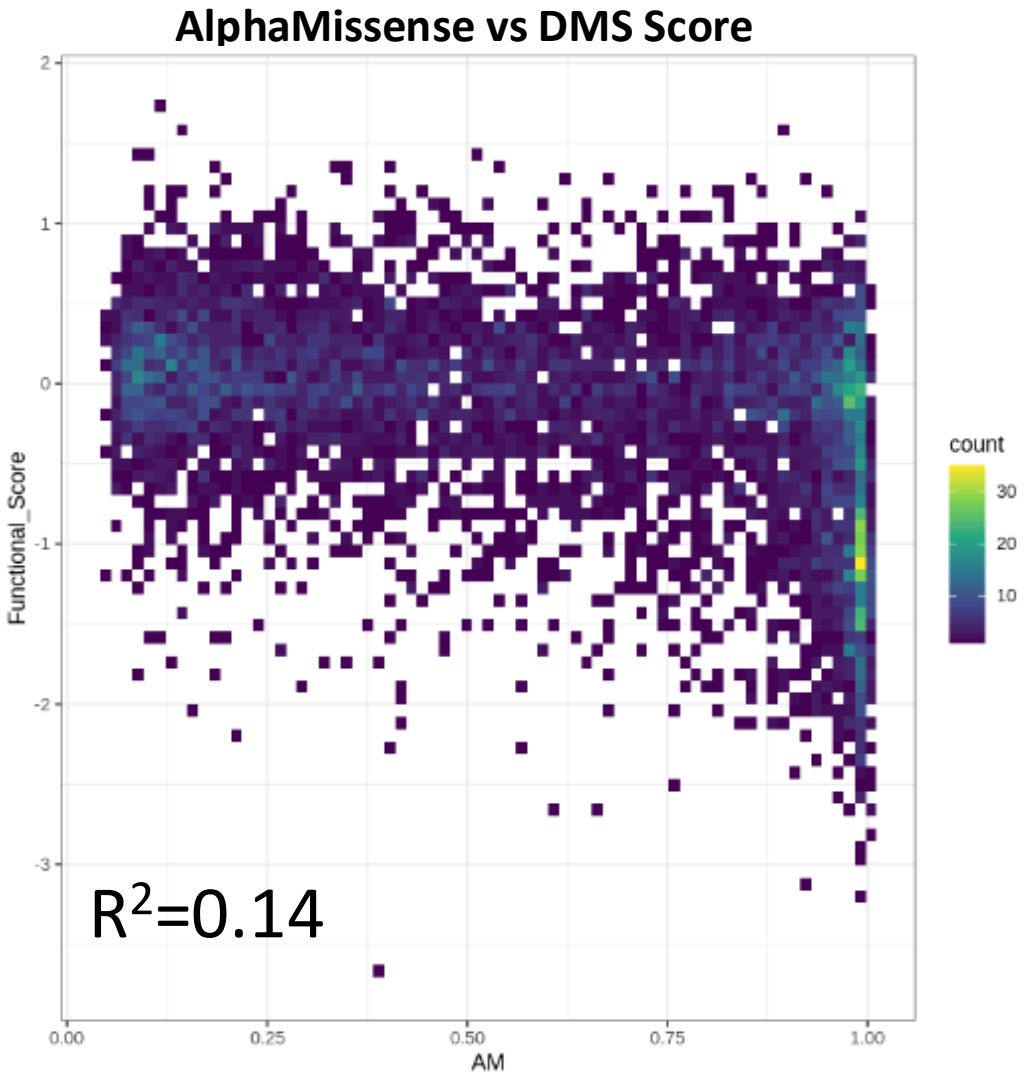
DMS Score

Alpha Missense

REVEL



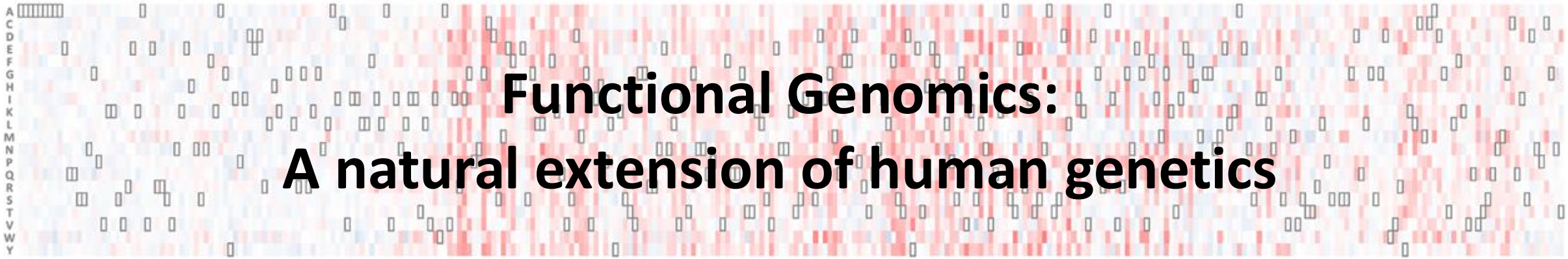
Comparison of SGCB DMS Functional Score Predictors



Conclusion: AM has high pathogenic variant sensitivity but mediocre benign variant accuracy

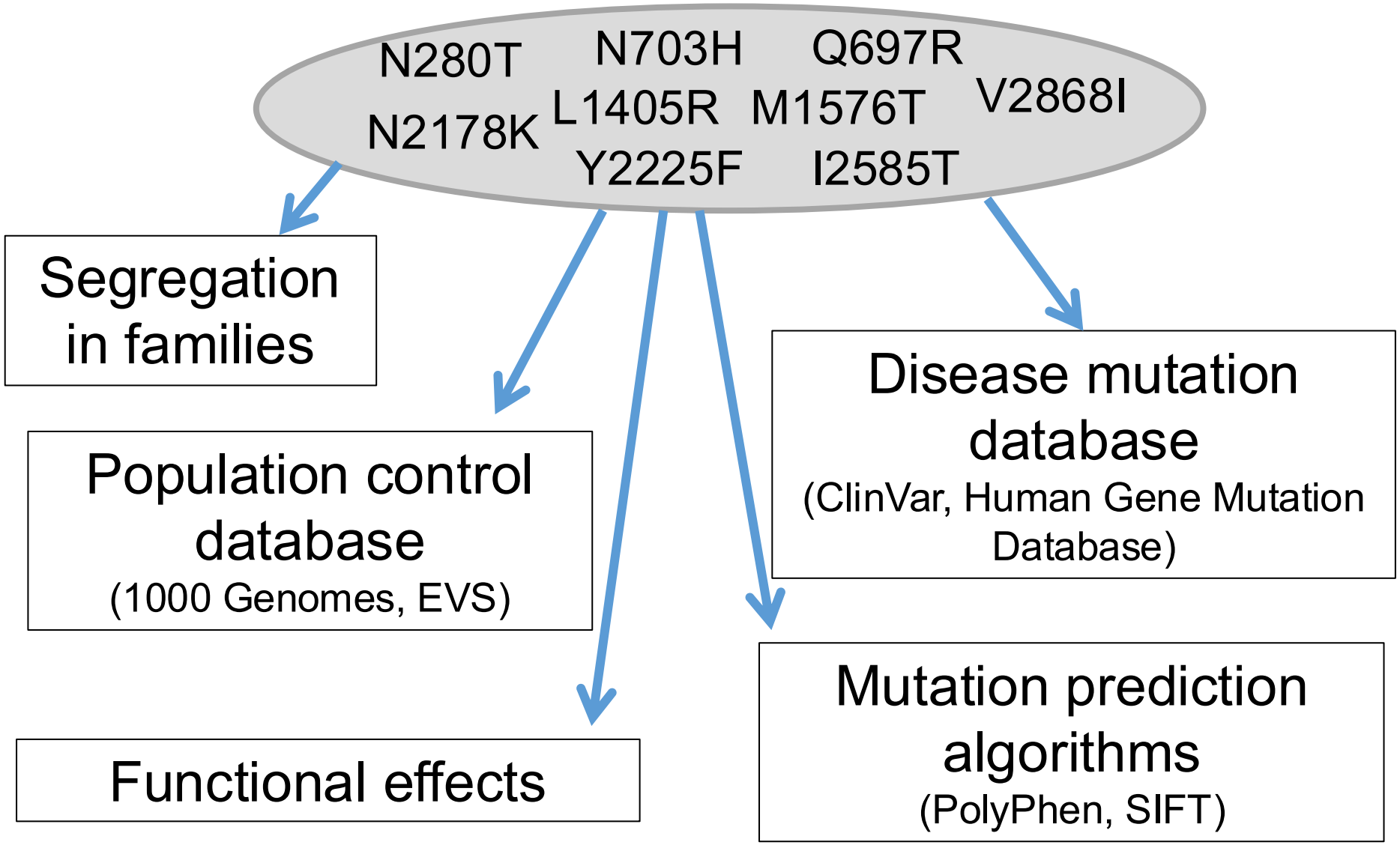
Limitations

- Not great at predicting pathogenicity of variants in disordered regions
- Lacks specificity for benign variants
  - Quite a high rate of predicted pathogenicity (30%?)
- Does it really use structure? Removing structure doesn’t harm the prediction too much.
- Still somehow cheating in predicting ClinVar variants?
  - Is clinVar highly biased toward having common variants as the only benign variants? … Answer: Yes.
  - Does AM model variant frequency… Yes.



Functional Genomics:
A natural extension of human genetics

Predicting Pathogenic vs. benign?

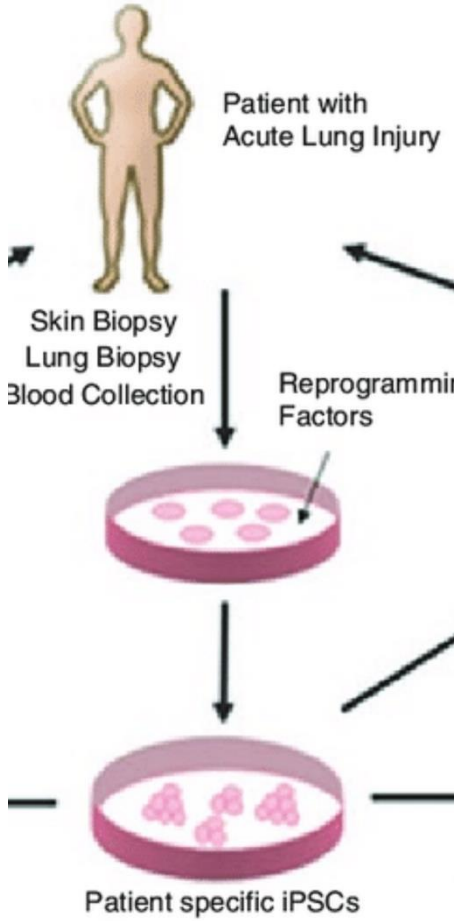


ACMG and AMP Guidelines for interpretation of of sequence variants: Richards et al., 2015

How to functionally test a candidate disease variant?

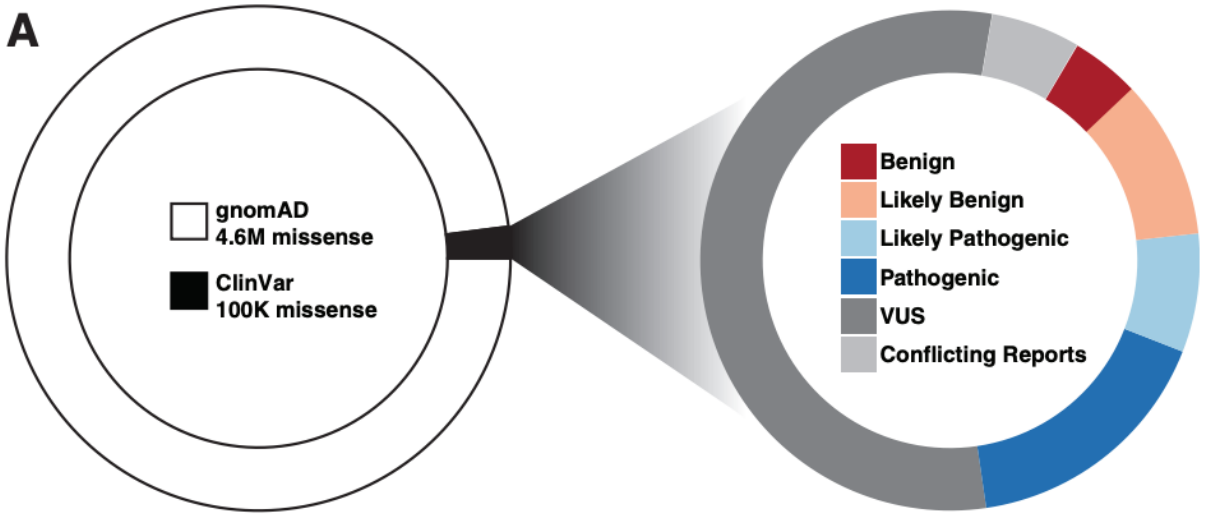


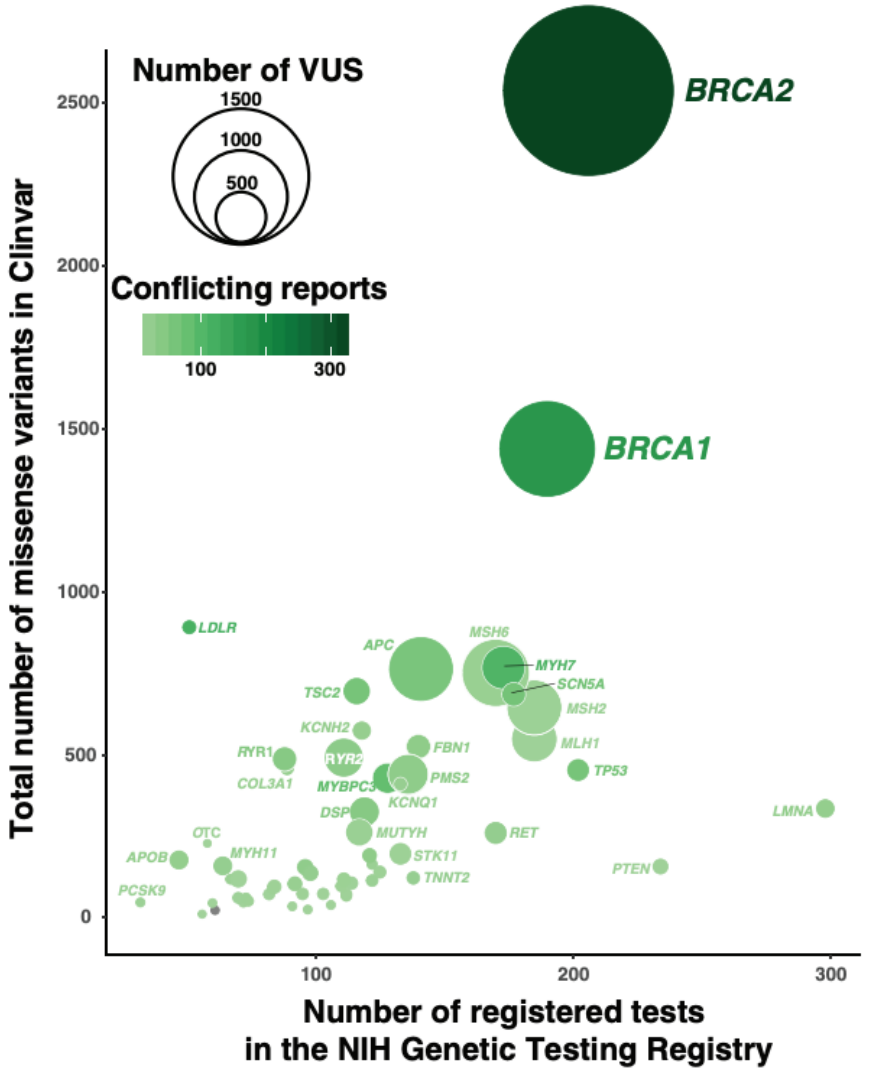




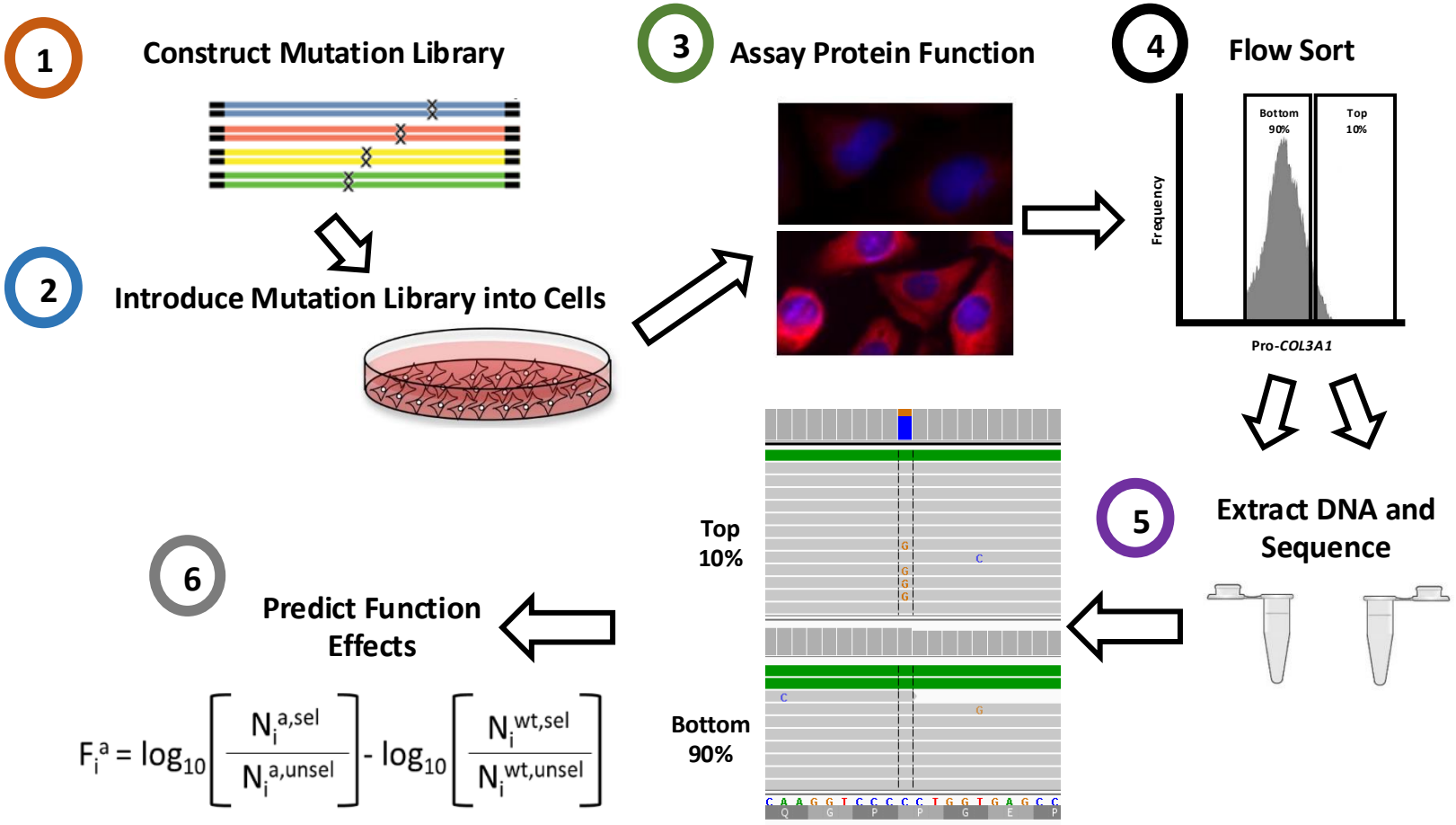
One at a time variant testing x 3,000,000,000 bases x $10,000 x 4 years = forever

The VUS problem: 99% of Vs are VUSs



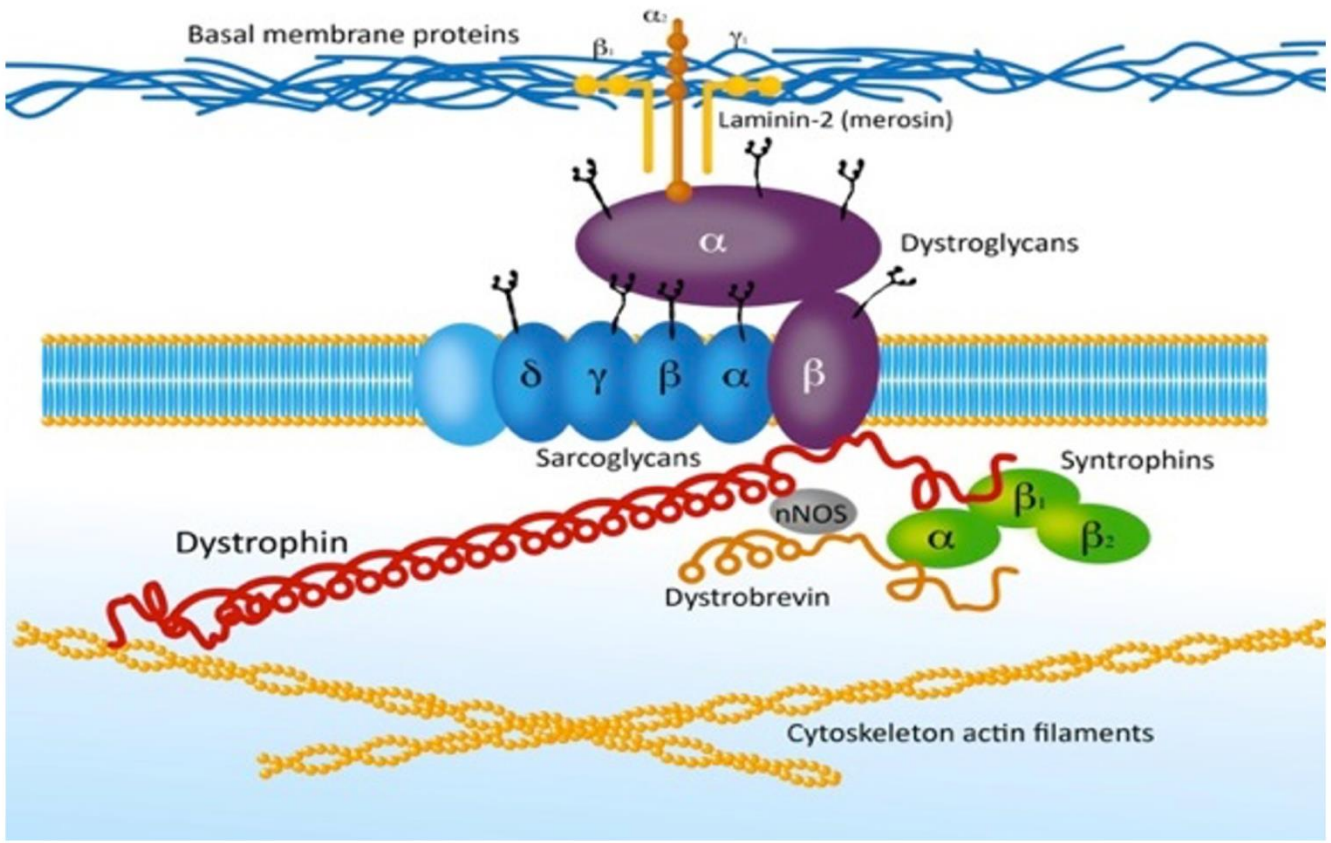


High-throughput Functional Screening of Protein Coding Genetic Variation



Deep Mutational Scan (DMS) of *Sarcoglycans* that cause Limb-Girdle Muscular Dystrophy (LGMD)

SGC Genes encode the sarcoglycans and cause Limb-Girdle Muscular Dystrophy (LGMD)



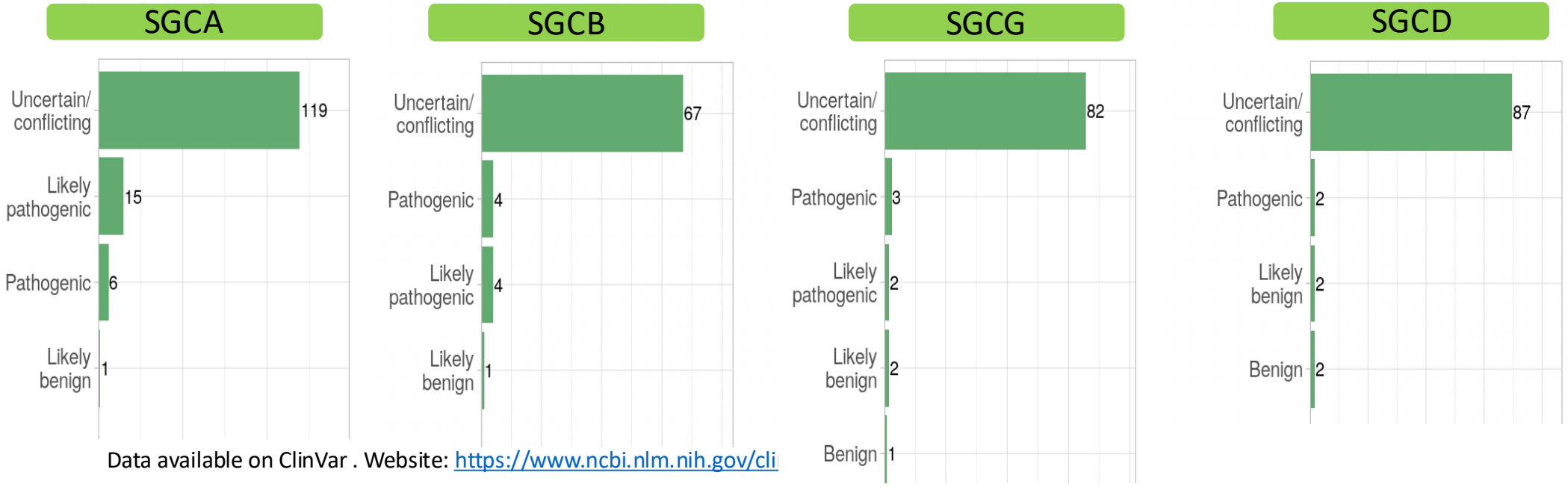
LGMD2C	*SGCG*[22]
LGMD2D	*SGCA*[22]
LGMD2E	*SGCB*[22]
LGMD2F	*SGCD*[22]



. Kirschner, J. & Lochmüller, H. *Sarcoglycanopathies. Handbook of Clinical Neurology* (2011).
doi:10.1016/B978-0-08-045031-5.00003-7

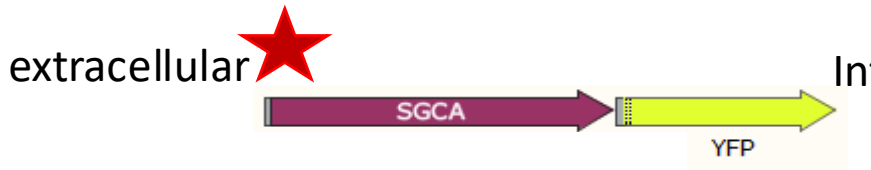
ClinVar Annotations of SGC gene variants are almost entirely VUS

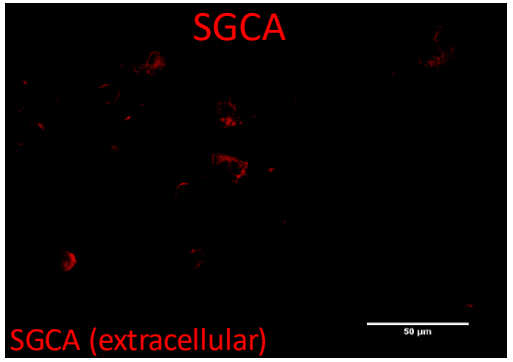
	GRCh37: Chr17:48244740 *GRCh38:* Chr17:50167379			muscular dystrophy type 2D				
3.	NM_000023.4(**SGCA**):c.52C>G (p.Leu18Val) *GRCh37:* Chr17:48244743 *GRCh38:* Chr17:50167382	**SGCA**	L18V	Autosomal recessive limb-girdle muscular dystrophy type 2D	Uncertain significance (Oct 10, 2016)	criteria provided	single submitter	VCV000471334

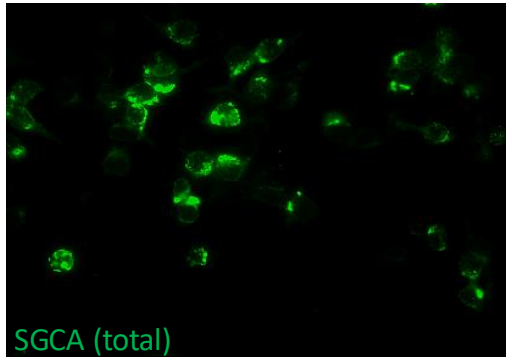


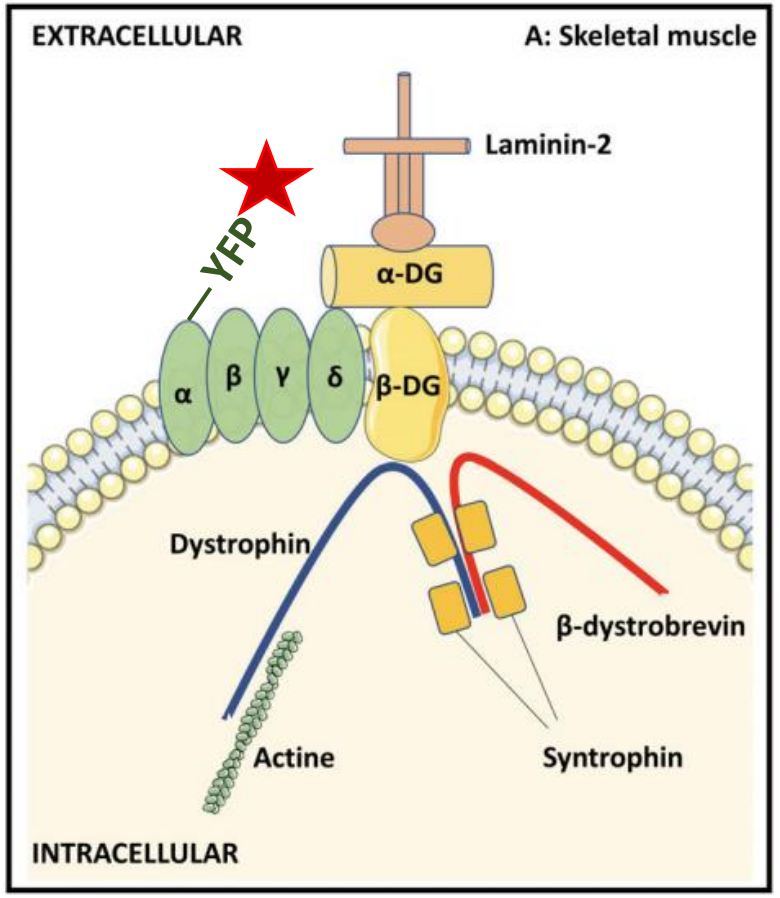
Data available on ClinVar . Website: https://www.ncbi.nlm.nih.gov/cli

Function assay: Cell-surface expression

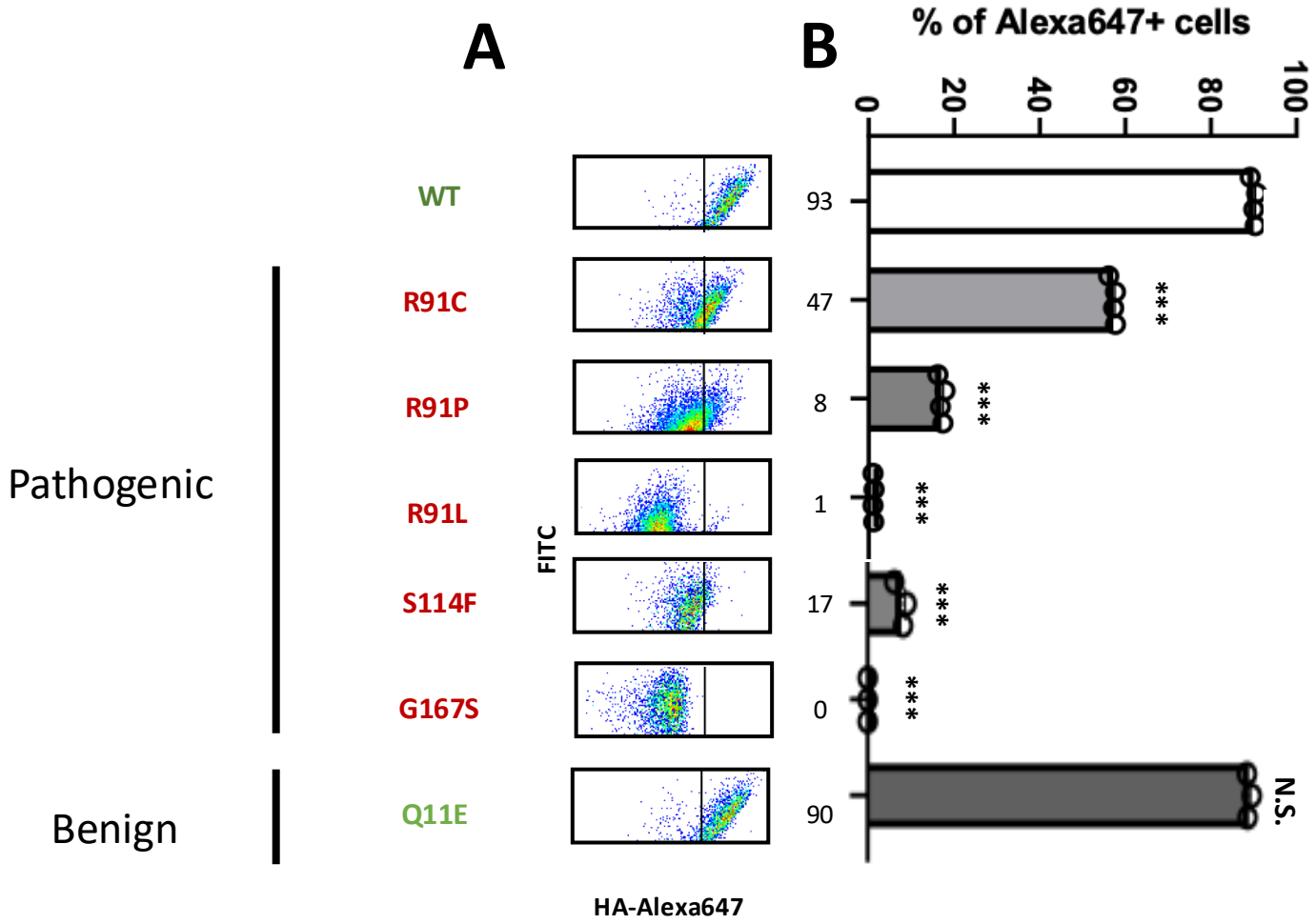








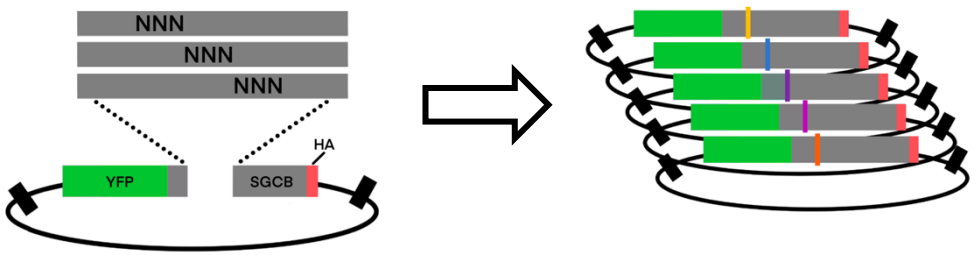
SGCB cell-surface expression separates pathogenic and benign variants



Workflow for Deep Mutational Scanning of *SGCB*

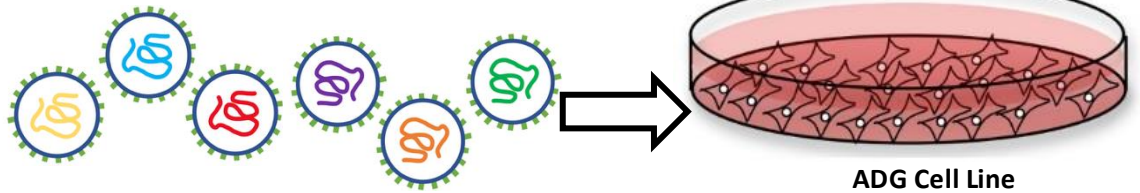
1

Introduce NNN Mutation Libraries into Lentiviral Vector



2

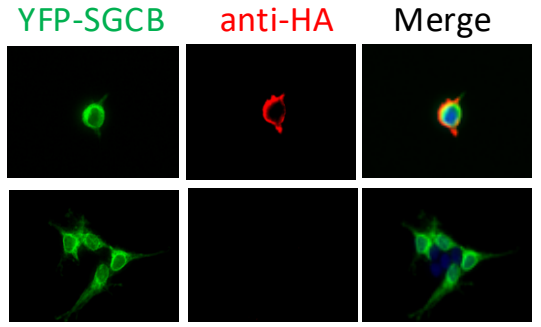
Lentiviral Overexpression



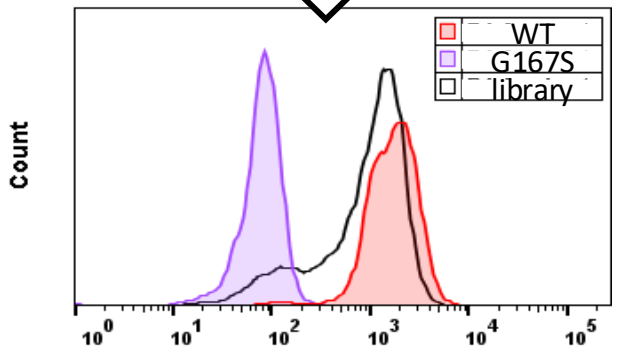
ADG Cell Line
MOI = 0.1

3

Assay Protein Function

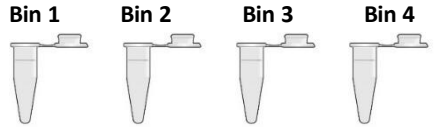


4



HA-Alexa-647

Extract DNA and Sequence



5

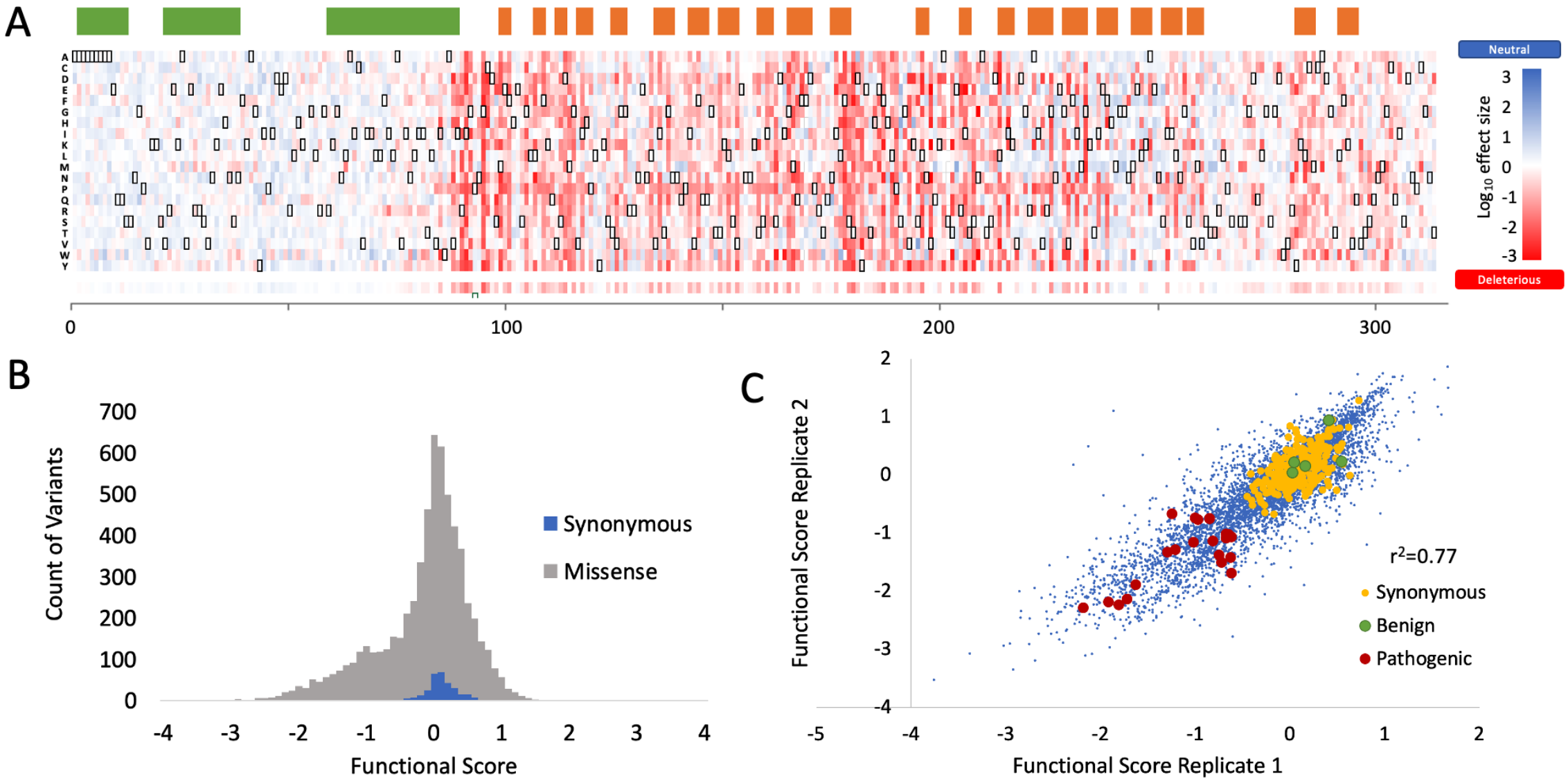


6

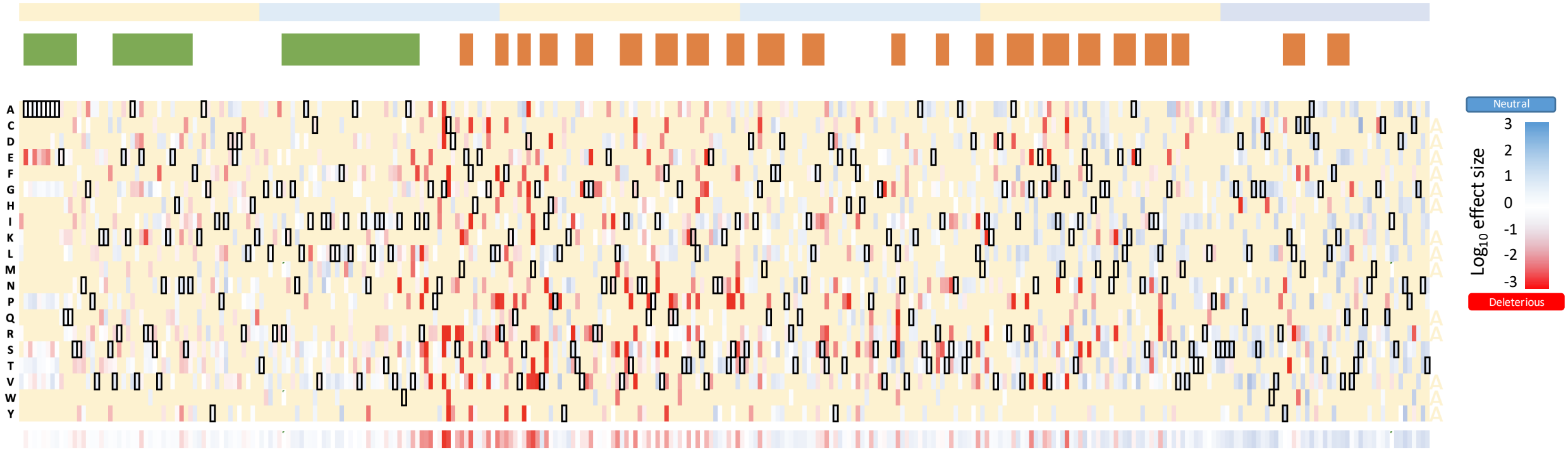
Calculate Function Effects

		Replicate 1				Replicate 2				
Change	Codon	HA1	HA2	HA3	HA4	HA1	HA2	HA3	HA4	Score
G167S	TCC	0.87	0.04	0.07	0.02	0.53	0.16	0.12	0.19	-1
G167S	TCT	0.87	0.1	0.02	0.01	0.78	0.1	0.05	0.07	-1.45
G167S	AGT	0.89	0.04	0.05	0.02	0.94	0.02	0.02	0.02	-1.65
R227R	CGA	0.24	0.25	0.23	0.28	0.25	0.24	0.26	0.25	0.05
R227R	CGC	0.25	0.24	0.23	0.28	0.25	0.25	0.26	0.24	0
R227R	CGG	0.28	0.24	0.23	0.25	0.22	0.23	0.28	0.26	0.05

Functional Effect Map of *SGCB*

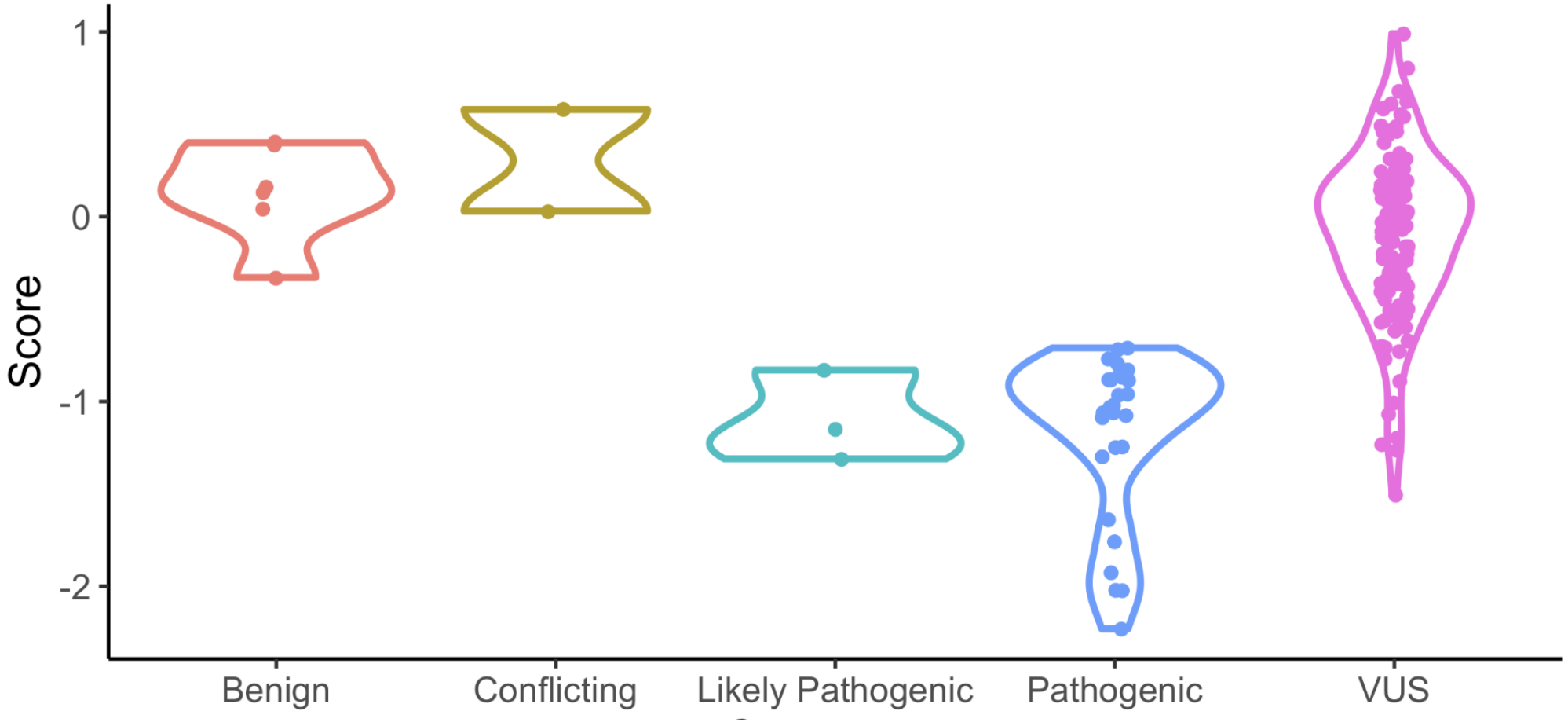


Functional Effect Map of *SGCB (SNVs Only)*

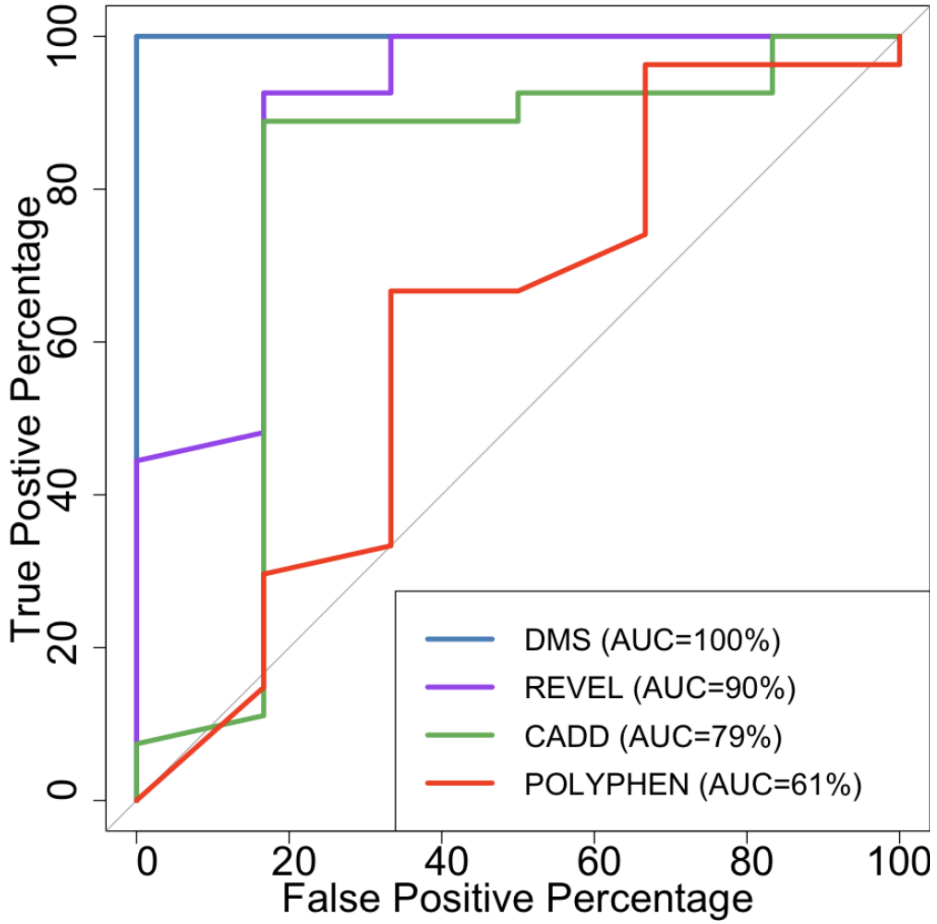


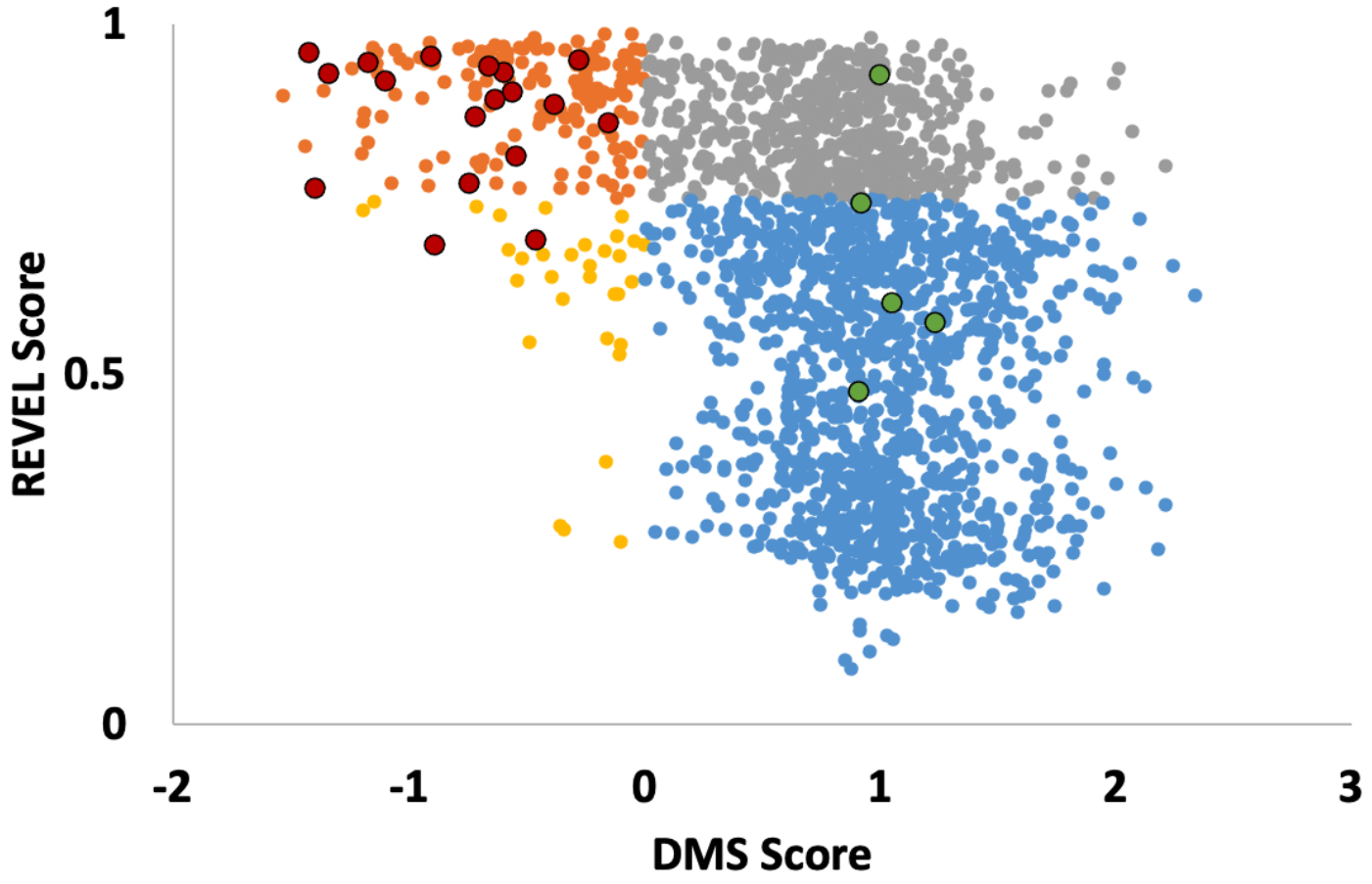
Without creating all possible amino acid changes, you don't obtain a full appreciation for the protein structure or the importance of different domains

DMS Scores corroborate ClinVar Pathogenicity

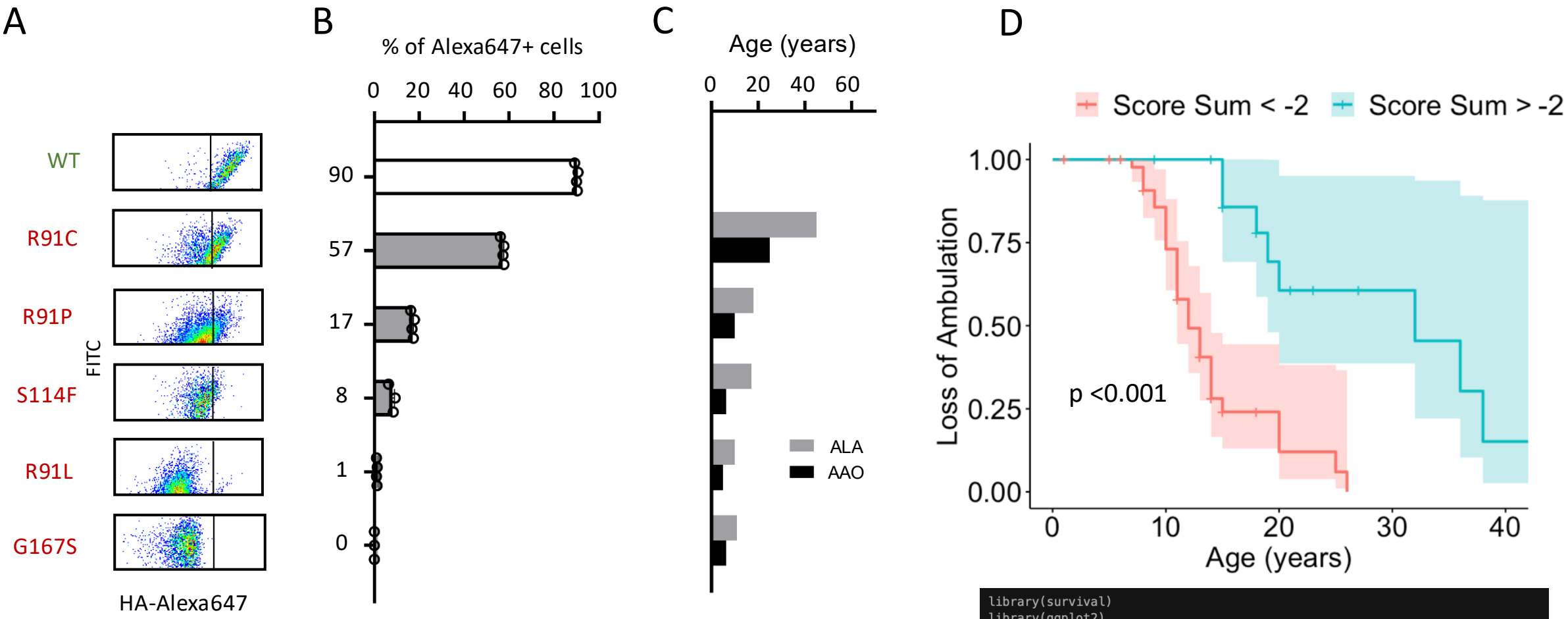


DMS Scores Outperform Variant Prediction Algorithms



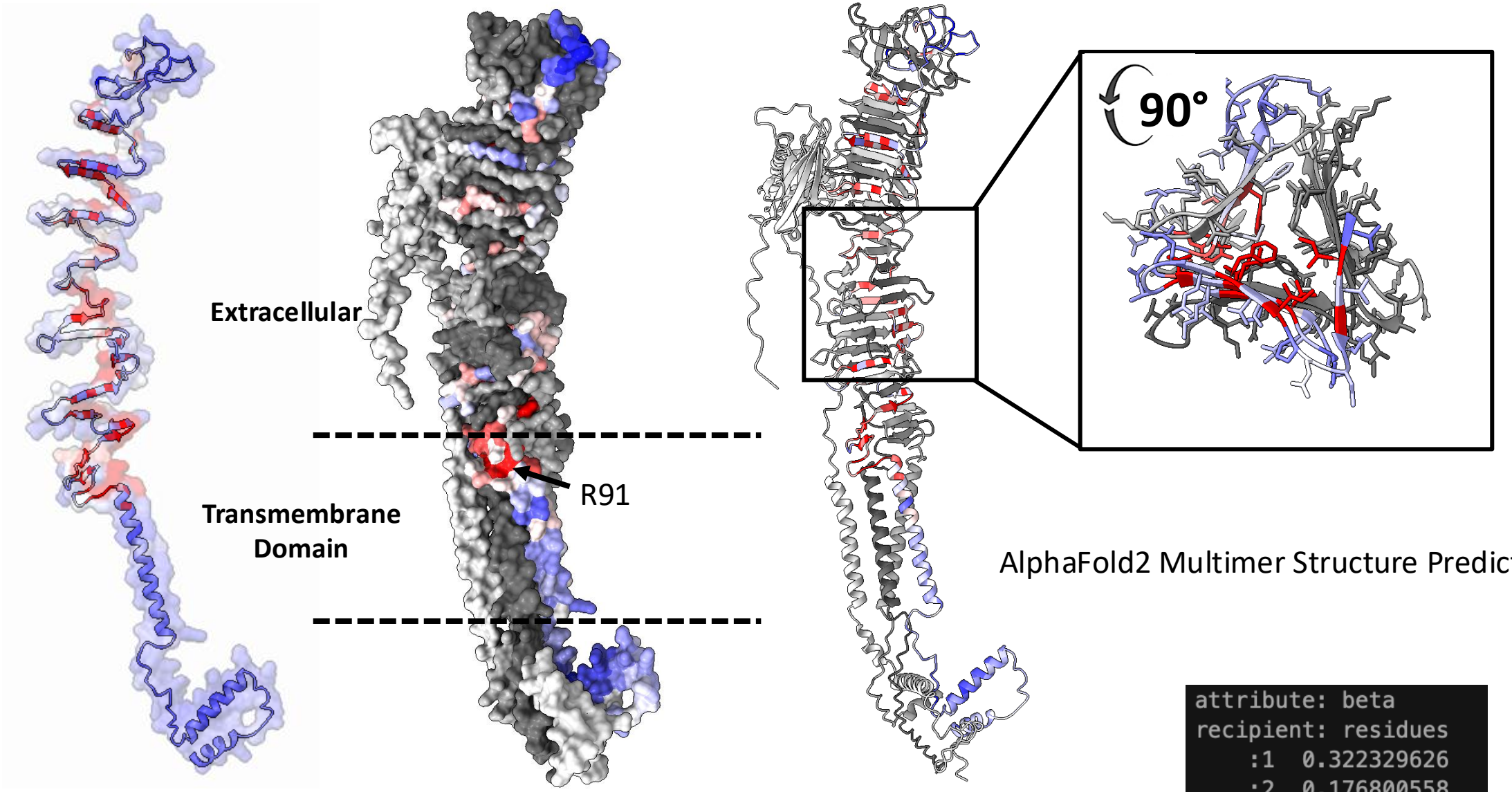


Variant Function Predicts Disease Severity




```
library(survival)
library(ggplot2)
library(survminer)
library(labeling)
survfile=read.csv("/content/forR_obex_survival.csv", header=T)
fit <- survfit(Surv(age, surgery) ~ obexdichot, data = survfile)
ggsurvplot(fit, data = survfile, risk.table=FALSE,pval=TRUE,conf.int=TRUE)
```

Structural insights from *SGCB* deep mutational scanning



AlphaFold2 Multimer Structure Predictions

```
attribute: beta
recipient: residues
    :1  0.322329626
    :2  0.176800558
    :3  0.165839922
```

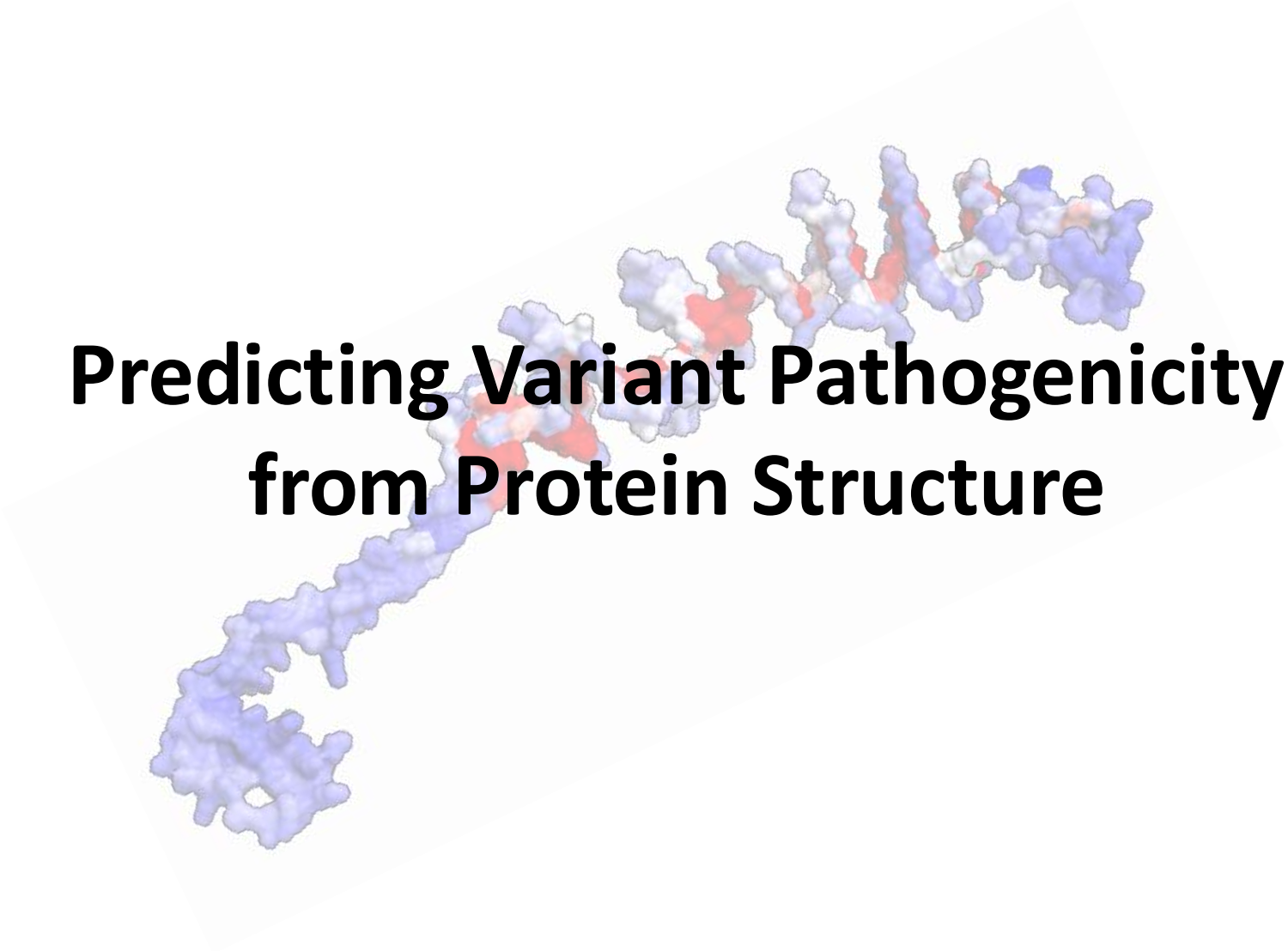
In ChimeraX: color byattr beta /C palette -0.2,red:0,white:0.2,blue



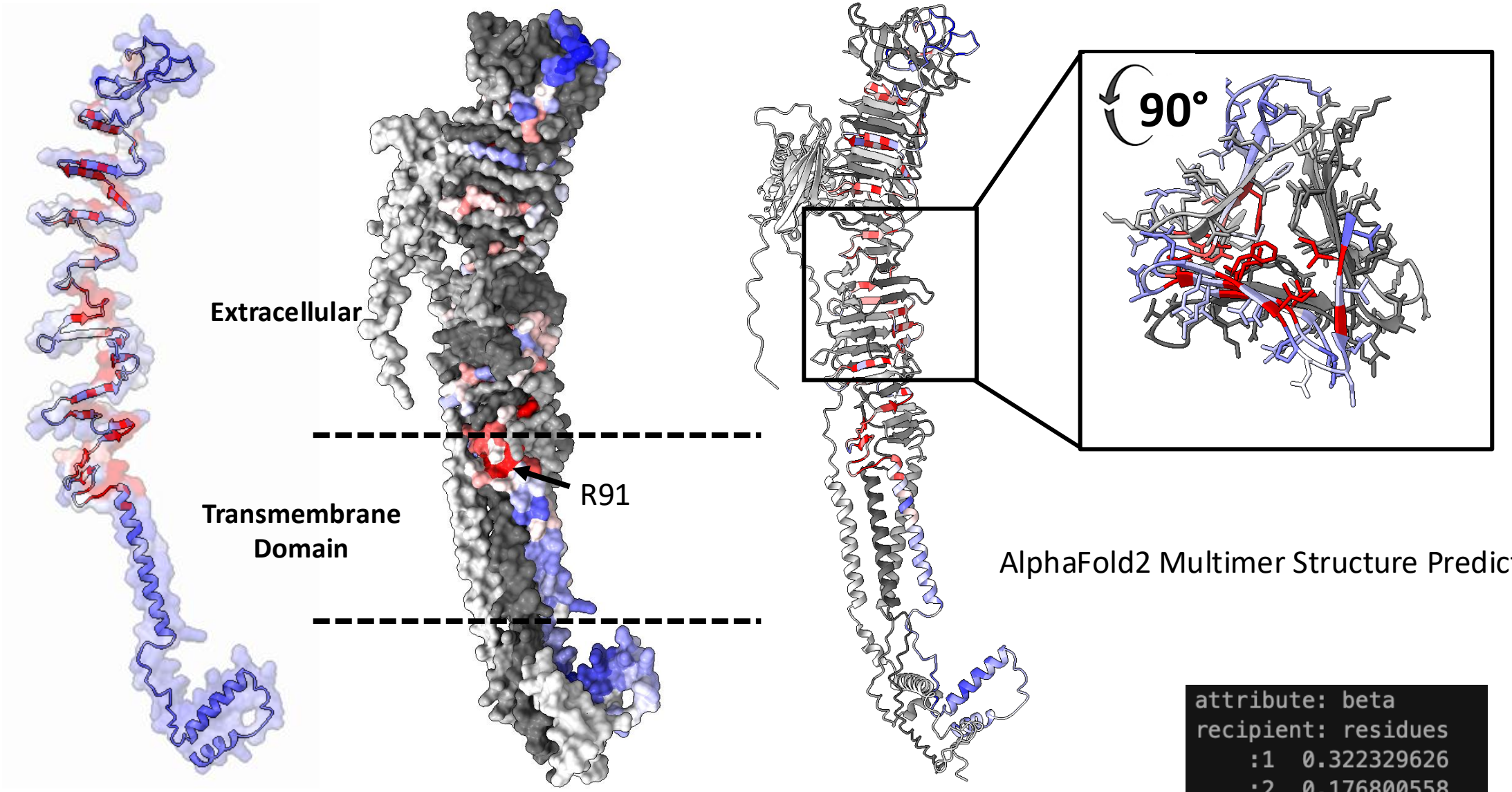
Summary

- Proper membrane localization of β-sarcoglycan requires the co-expression of its partners, α-, γ- and δ-sarcoglycan.
- Mutations in *SGCB* leads to a secondary reduction of membrane localization of **α**-sarcoglycan.
- DMS scores for *SGCB*:
  - perfectly predict pathogenicity of known variants
  - outperform all current prediction algorithms
  - quantitatively measure the severity of the variant and predict disease severity
  - map important intermolecular interactions and corroborate AlphaFold2 multimer structure predictions

Predicting Variant Pathogenicity from Protein Structure



Structural insights from *SGCB* deep mutational scanning

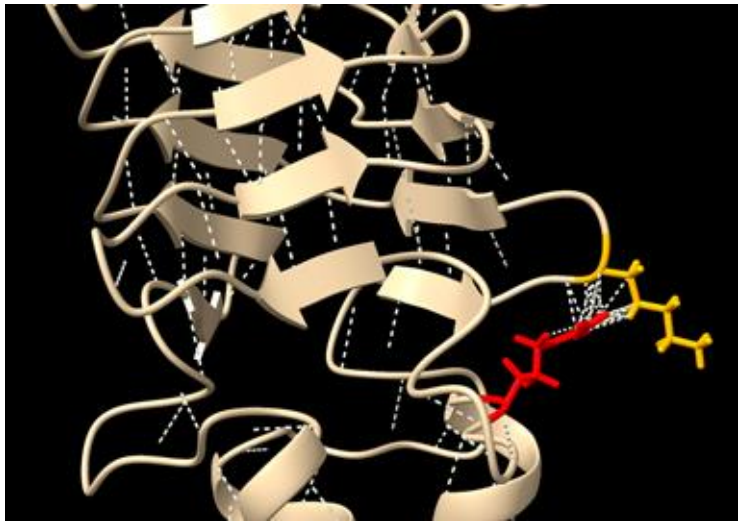


AlphaFold2 Multimer Structure Predictions

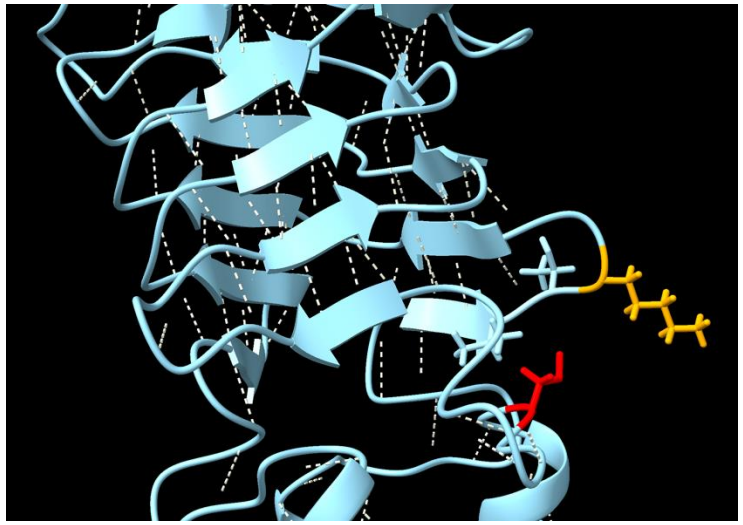
```
attribute: beta
recipient: residues
    :1  0.322329626
    :2  0.176800558
    :3  0.165839922
```

In ChimeraX: color byattr beta /C palette -0.2,red:0,white:0.2,blue

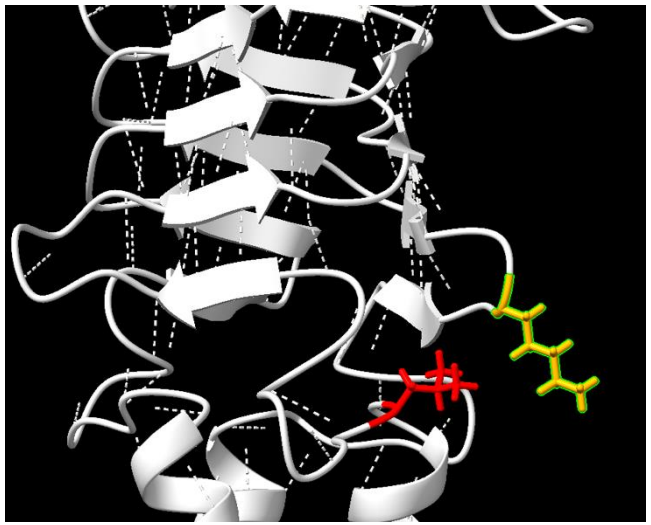
Pathogenic Missense Variants Strongly Affect Protein-Protein Interactions



WT



R91C

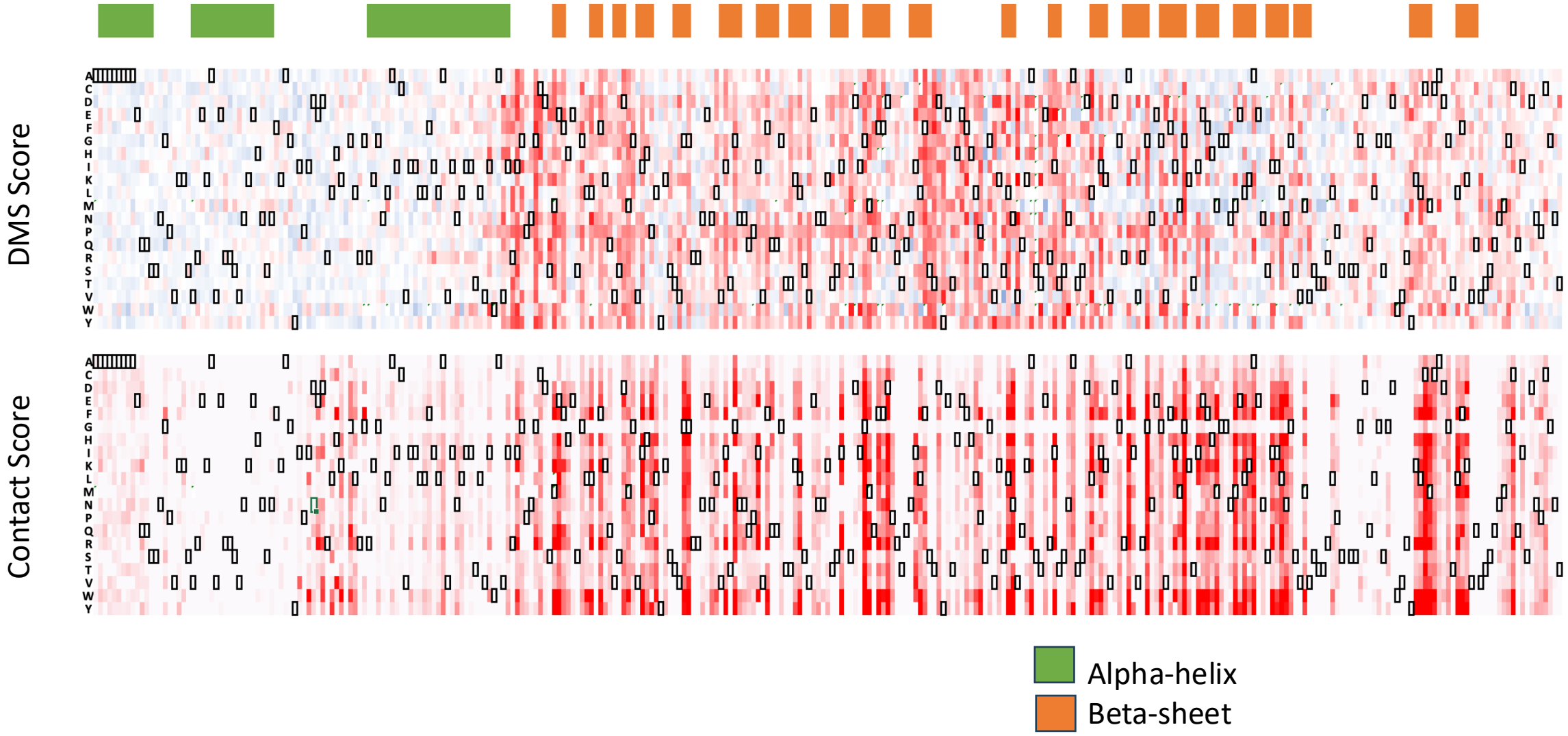


R91L

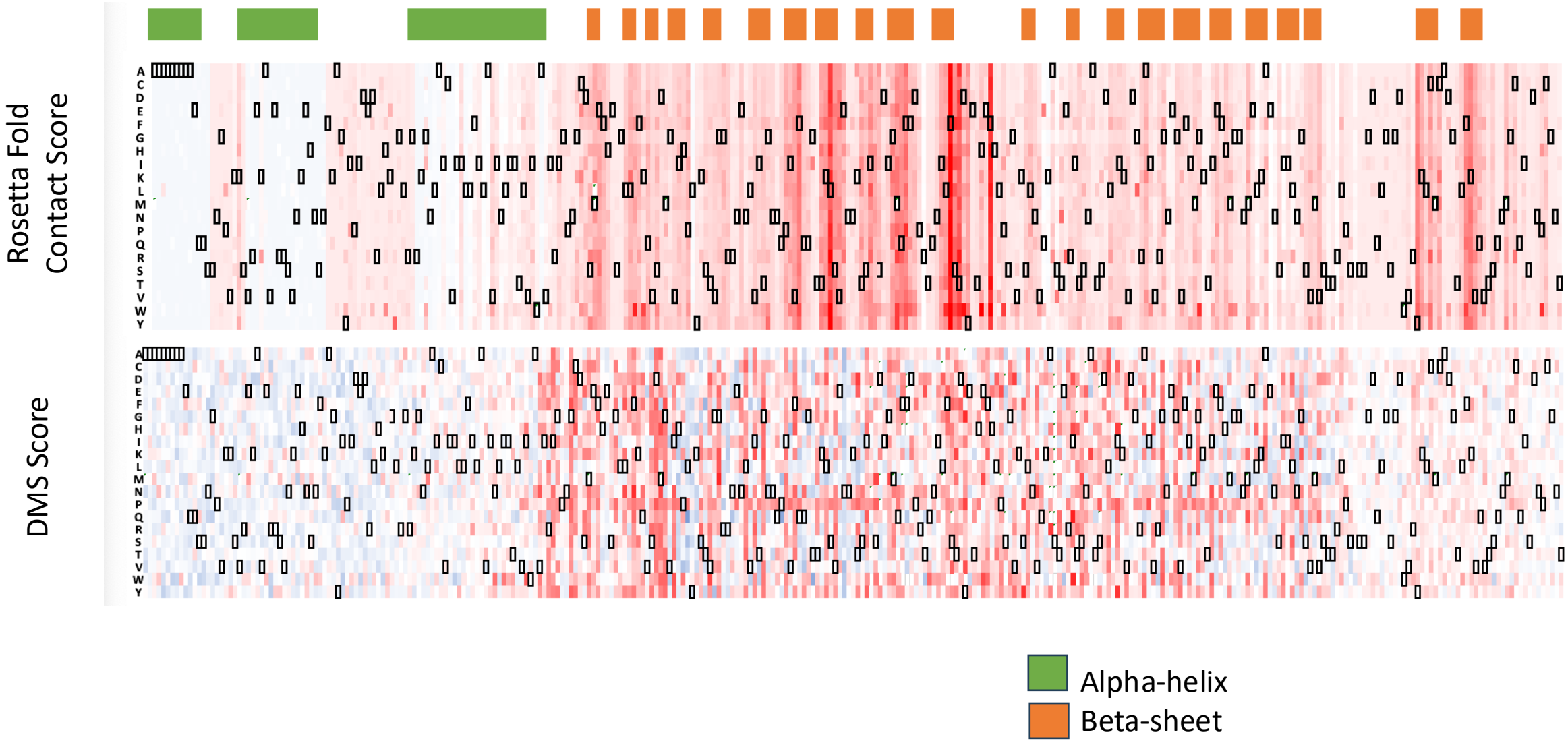
Factor 1: How many contacts are there between WT AA1 and any other AA

Factor 2: How many contacts are there between Mutant AA1 and any other AA

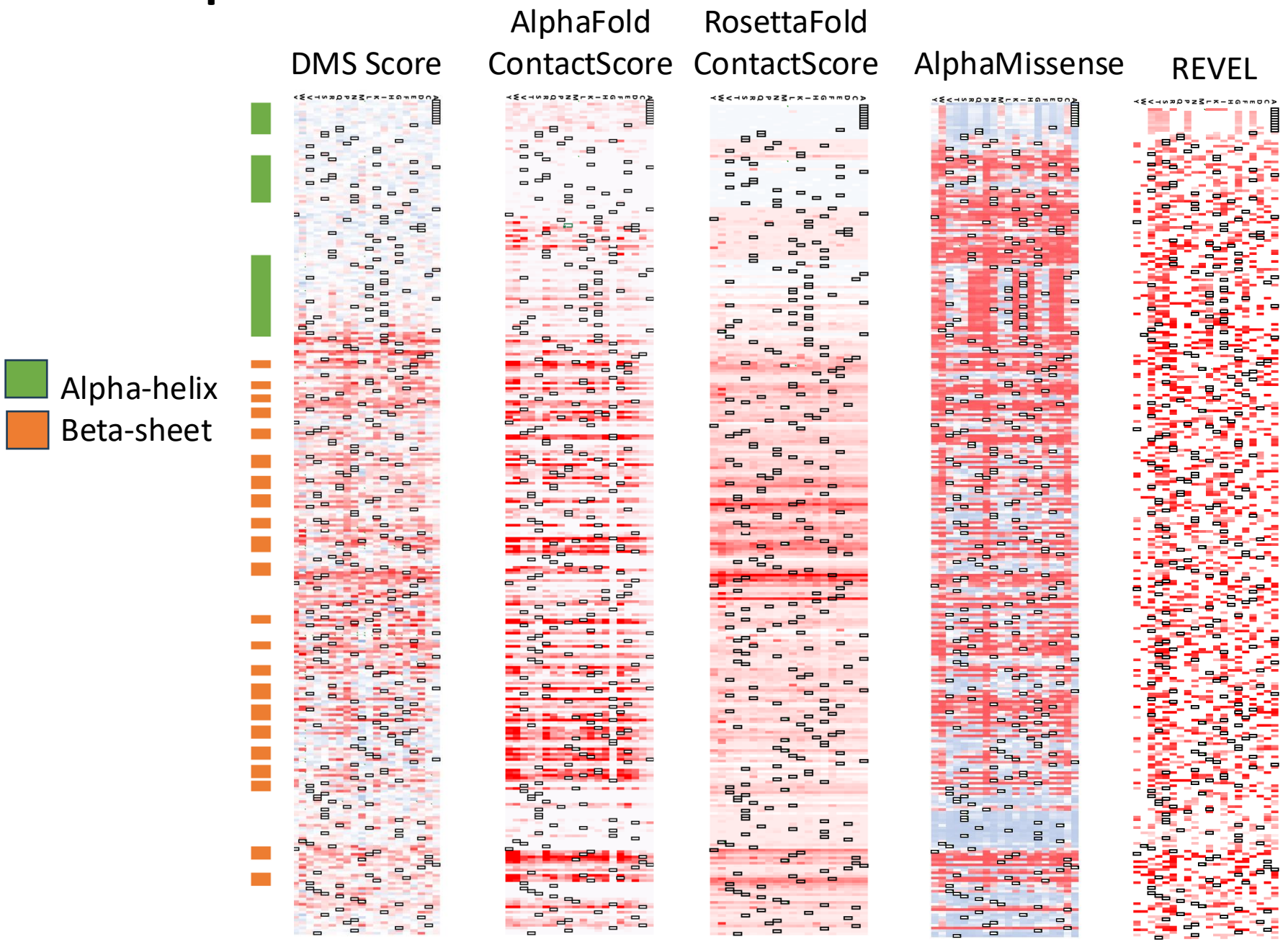
AlphaFold Contact Score predicts SGCB DMS Functional Score



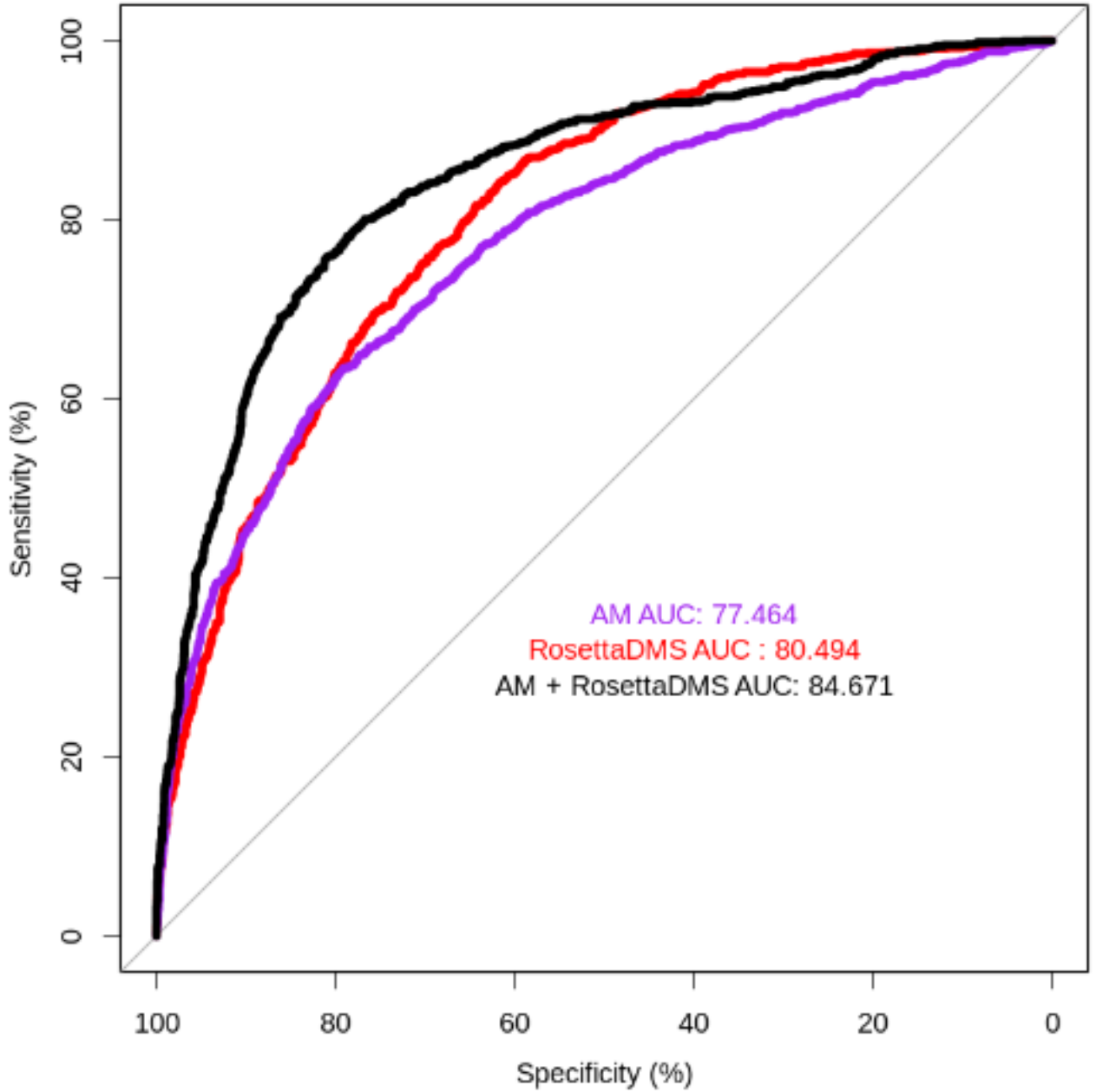
RosettaFold Contact Score predicts SGCB DMS Functional Score



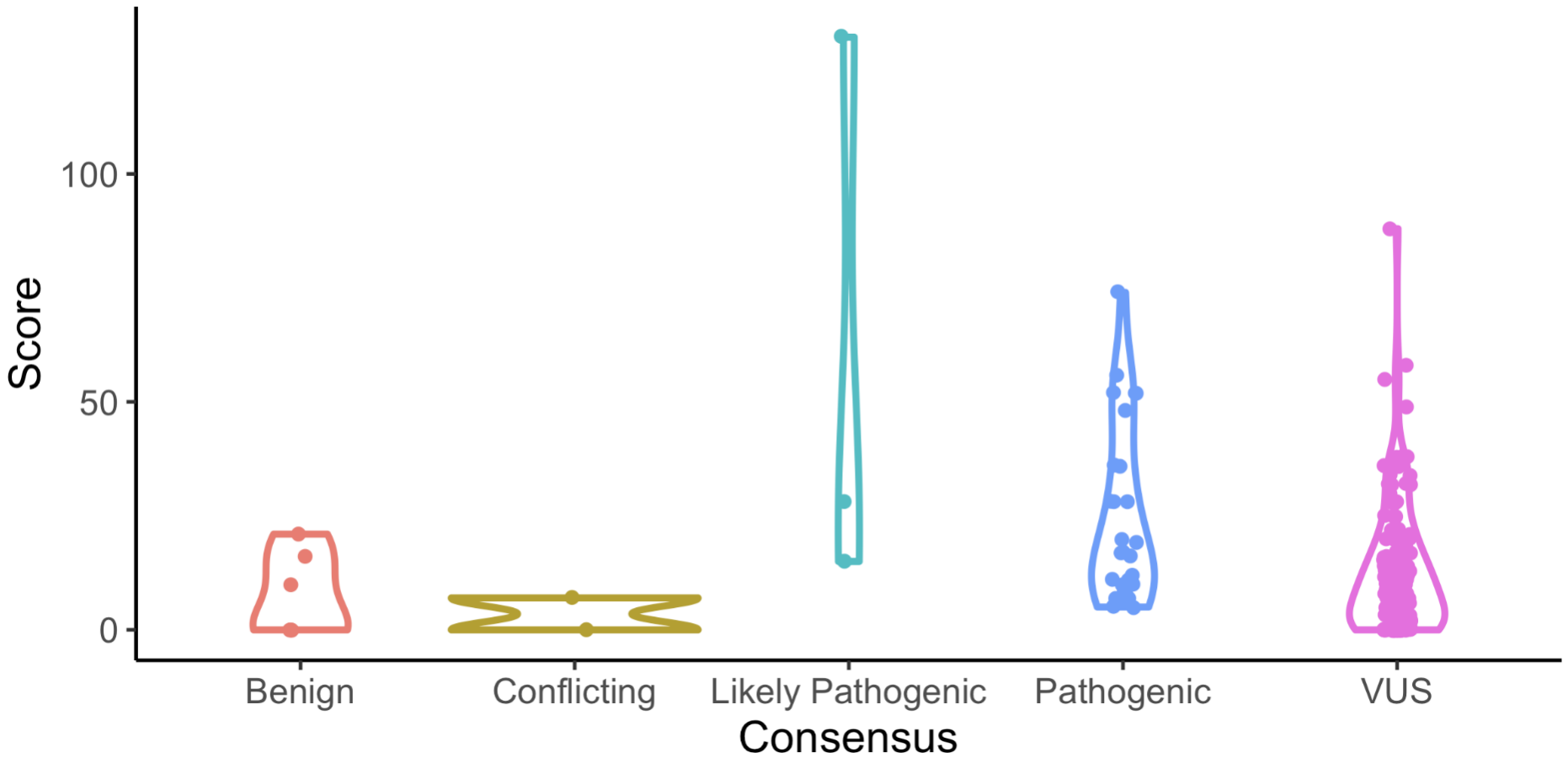
Comparison of SGCB DMS Functional Score Predictors



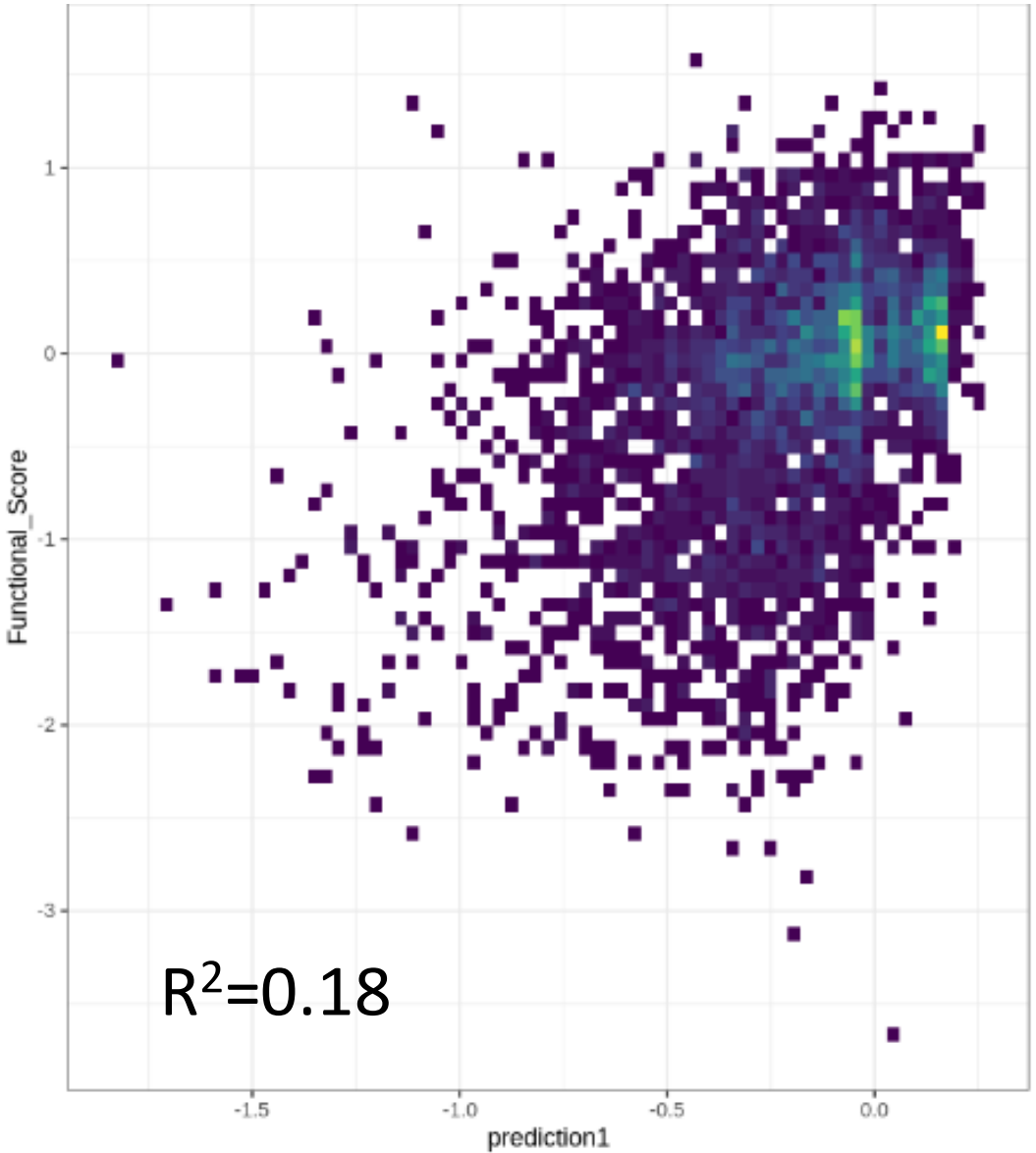
RosettaDMS predicts SGCB DMS Functional Score



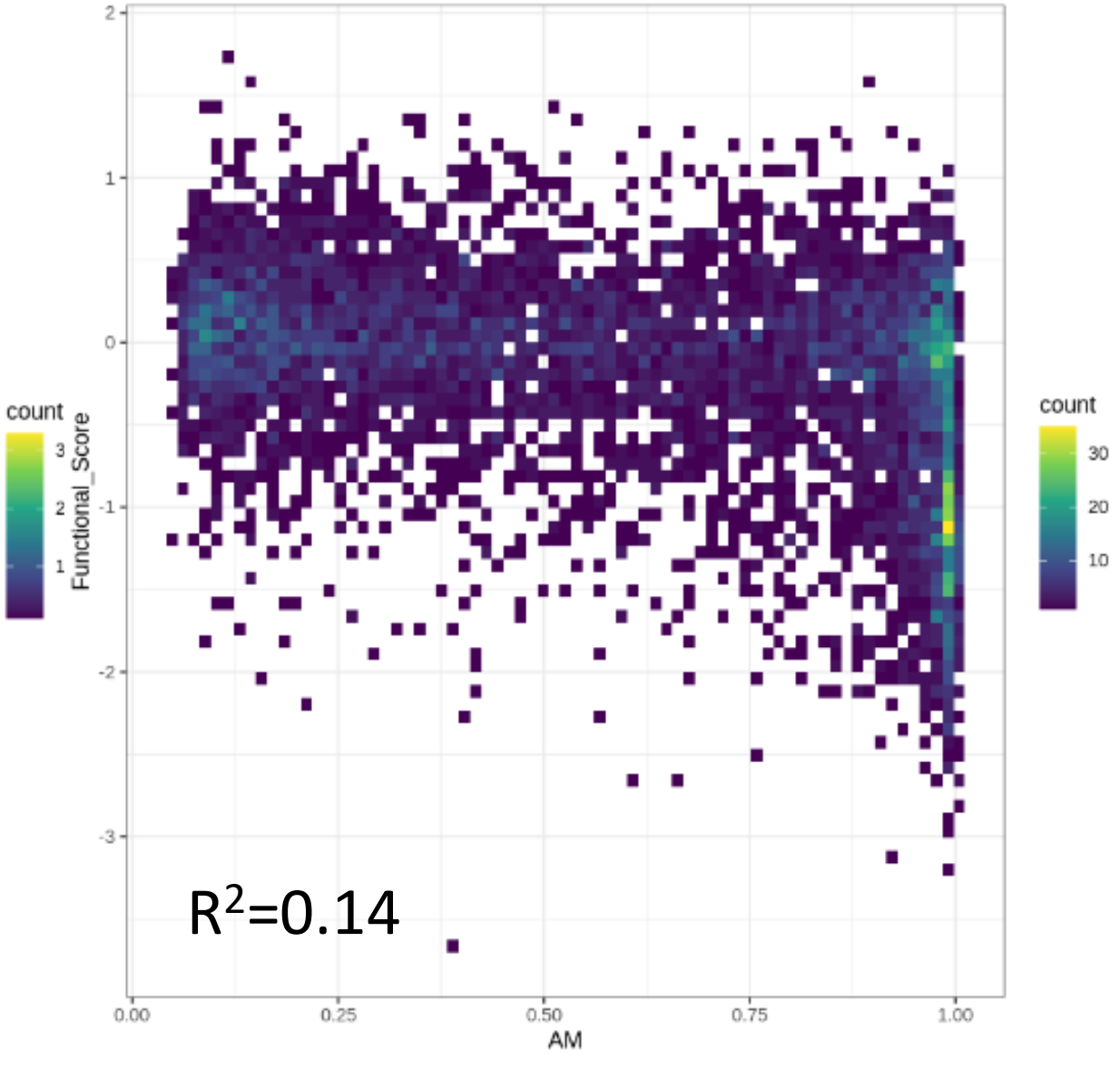
AlphaFold Contact Score predicts SGCB ClinVar Pathogenicity



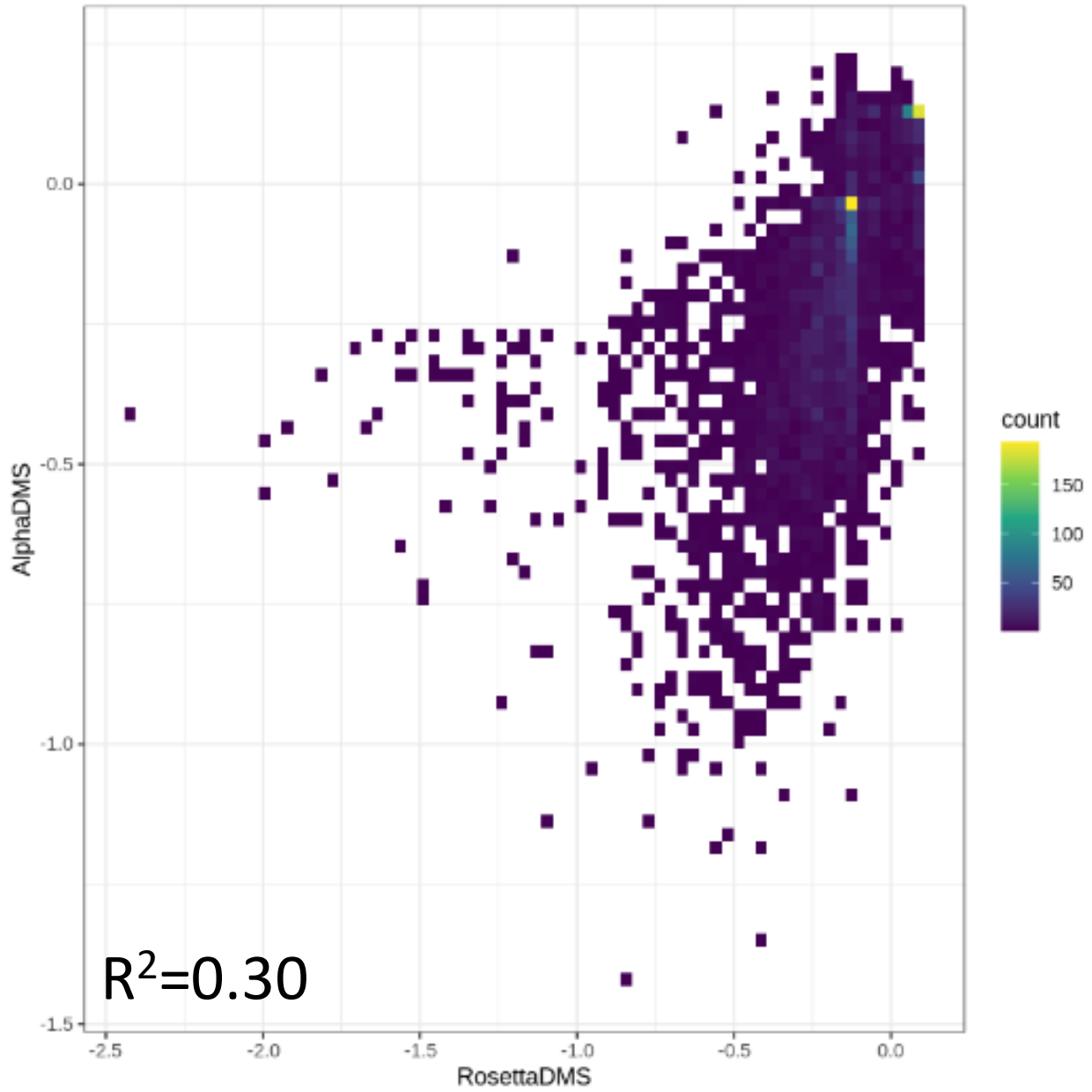
RosettaFold Contact Score vs DMS Score



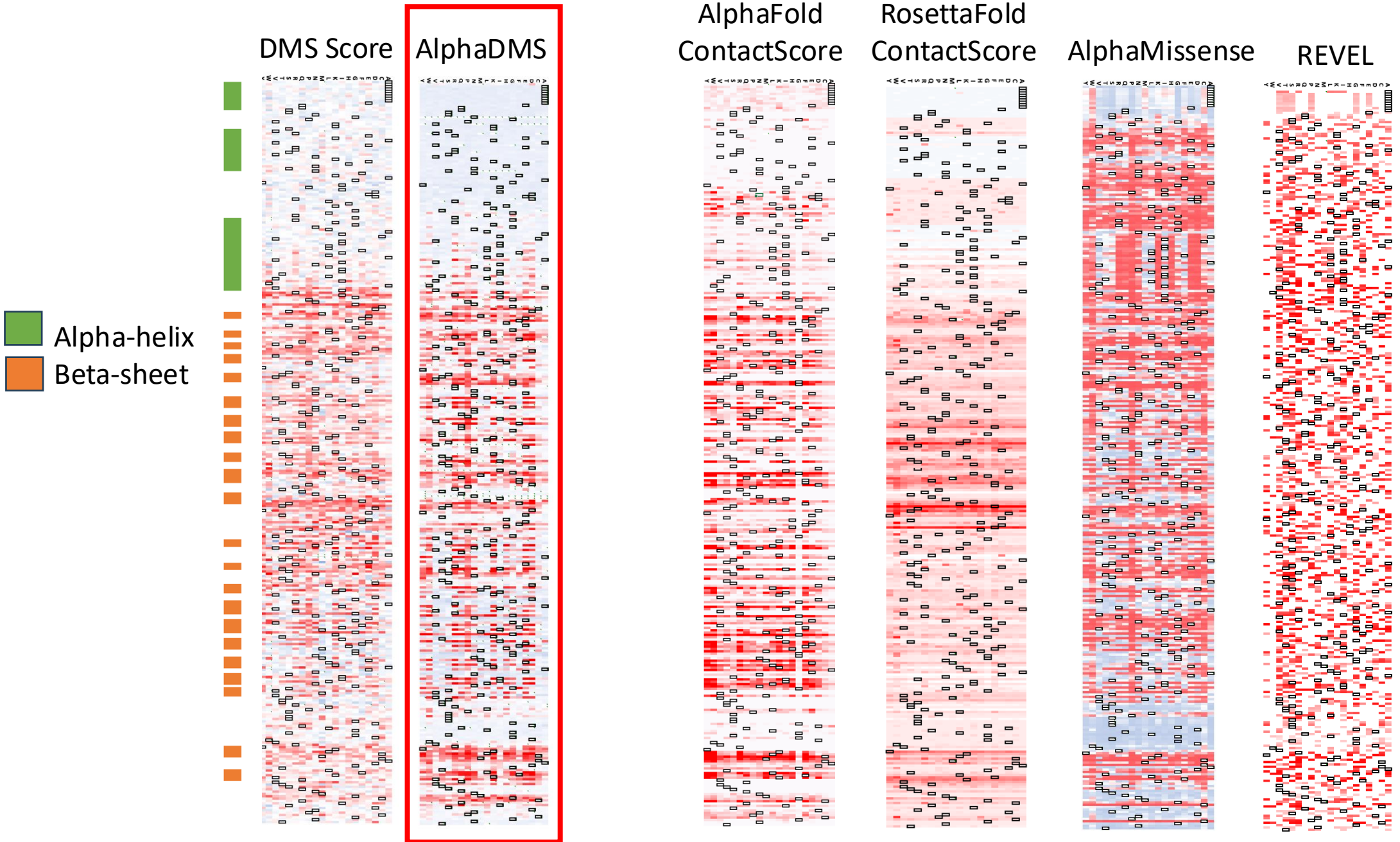
AlphaMissense vs DMS Score



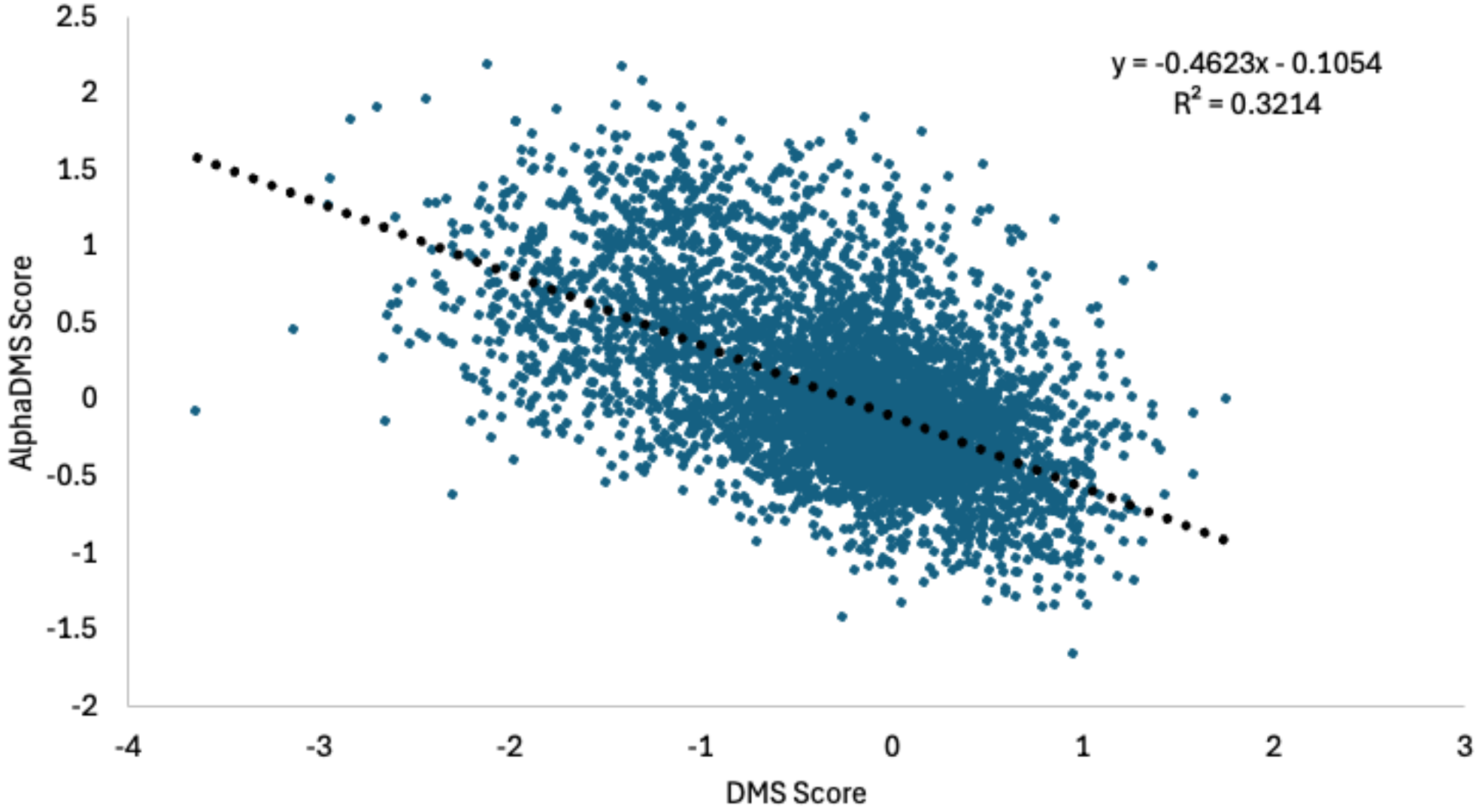
RosettaDMS and AlphaDMS are correlated but not perfectly



Comparison of SGCB DMS Functional Score Predictors



AlphaDMS vs DMS Score



Overall Conclusions

- Deep Mutational Scanning (DMS) is an indispensable tool for the assessment of variant effects.
- Assay development for specific genes is often the most difficult step in the process
- Important considerations:
  - Are the results useful? Clinically? Better understand the protein? Better understand proteins?
  - Are the results reliable? Reproducible and correlated with disease status if applicable?
  - Are multiple assays required to assess all possible gene disruptions (GOF vs LOF, etc)

Acknowledgements

Haller Lab

Ashley Quiggle, PhD
Erin Crosby
Arjun Nair
Geng Wang
Alya Bajwa

Gurnett Lab

Christina Gurnett
Kevin McCall
Momchil Nikolov

University of Utah
Doug Brockmeyer

Icahn School of Medicine
Alison Goate

Washington University in St. Louis
Brooke Sadler
David Limbrick
Jenn Strahle
Jarod Roland
Carlos Cruchaga
John Budde
Lila Solnica-Krezel
Diane Sepich

Funding

R01-AR081901 - NIH NIAMS
P01-NS13113101 - NIH NINDS
CDI-FR-2019-816 - Children's Discover Institute
CDI-II-2022-1015 - Children's Discovery Institute
P50-AG005681 - WU ADRC
P23-00451 - Hydrocephalus Association





National Institute of Arthritis and Musculoskeletal and Skin Diseases





U.S. Department of Health and Human Services

