

CAGI slide pack guidelines

I agree to my slides being used according to the following terms:

Yizhou Yin

- Authors: please include THIS slide at the beginning of your presentation to indicate that you agree with reuse of your slides in CAGI presentations, according to the terms below, and digitally sign by writing your name above.
- To disseminate the CAGI results, I encourage all CAGI participants to remix these CAGI slides for inclusion in their oral and poster presentations
- CAGI participants may use or publish any content of these slides only in compliance with the CAGI Data Use Agreement
- CAGI participants may not cite or publish any content of these slides except with the written permission of the originating author or in the primary CAGI publications
- CAGI participants must acknowledge the original authors of these slides and CAGI 2018. Include the acknowledgement banner across the bottom
- CAGI participants must include the CAGI credits slide in their presentations
- Slides with a red slash (no remix) may not be included in a presentation unless all attendees are registered CAGI participants who signed the data use agreement
- Slides with a red X over the entire slide may not be used in any circumstances
- Slides with a blue slash (embargoed) may not be used unless the embargo is explicitly lifted
- Slides without slash or X may be tweeted unless flagged with the “No tweet” icon.
- CAGI participants using these slides must explain the “No tweet” icon, which means that the slide content should not be disseminated outside the presenting conference hall. If the presentation venue permits pervasive tweeting, you may not include these slides in your talk
- Note for slide authors: we expect all CAGI participants to abide by these restrictions, and will make best effort to ensure they are followed. However, but bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.



Predict the effect of missense mutations on PTEN and TPMT protein stability

Yizhou Yin, Lipika R. Pal, Kunal Kundu, John Moul^t*

IBBR, University of Maryland College Park

CAGI5 Conference, Chicago, IL

5th July, 2018

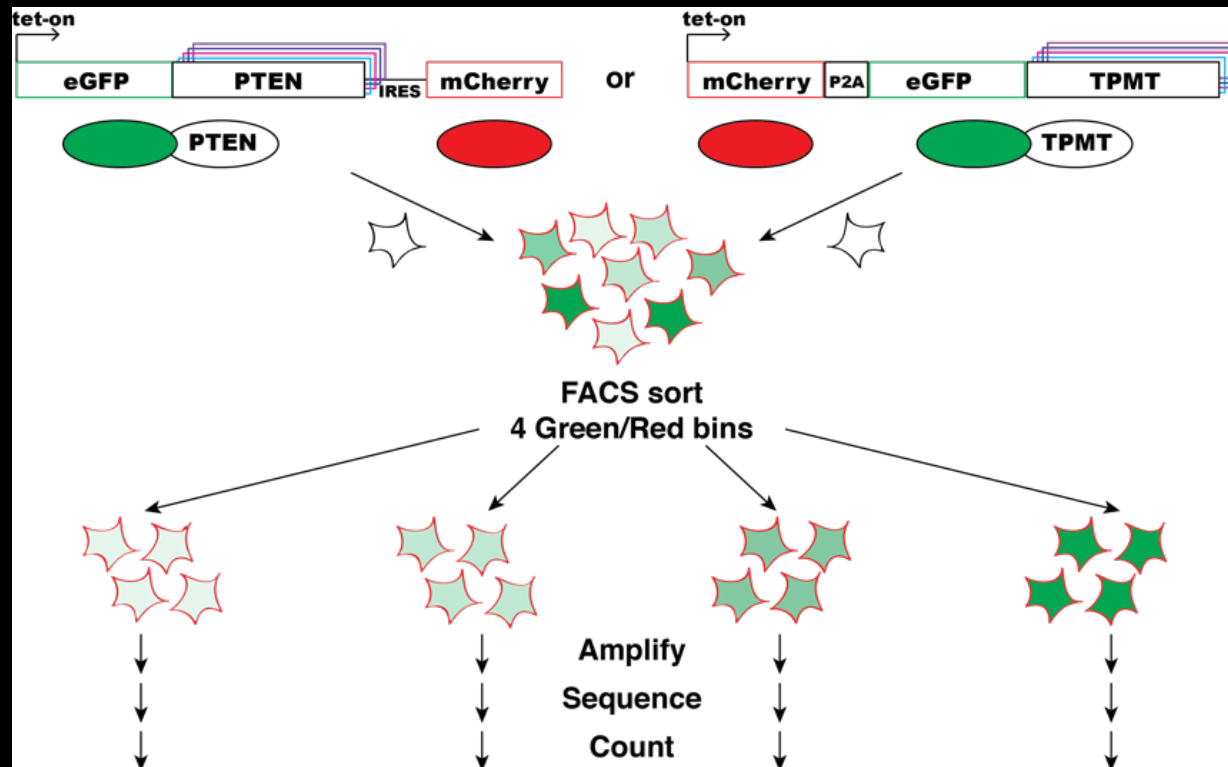


UNIVERSITY OF MARYLAND | NIST
INSTITUTE FOR BIOSCIENCE
& BIOTECHNOLOGY RESEARCH

PTEN and TPMT challenges



Predict variant effect on stability relative to the wild type



<https://genomeinterpretation.org/content/predict-effect-missense-mutations-pten-and-tpmt-protein-stability>

Combining sequence-based methods and structure-based methods

Ensemble method

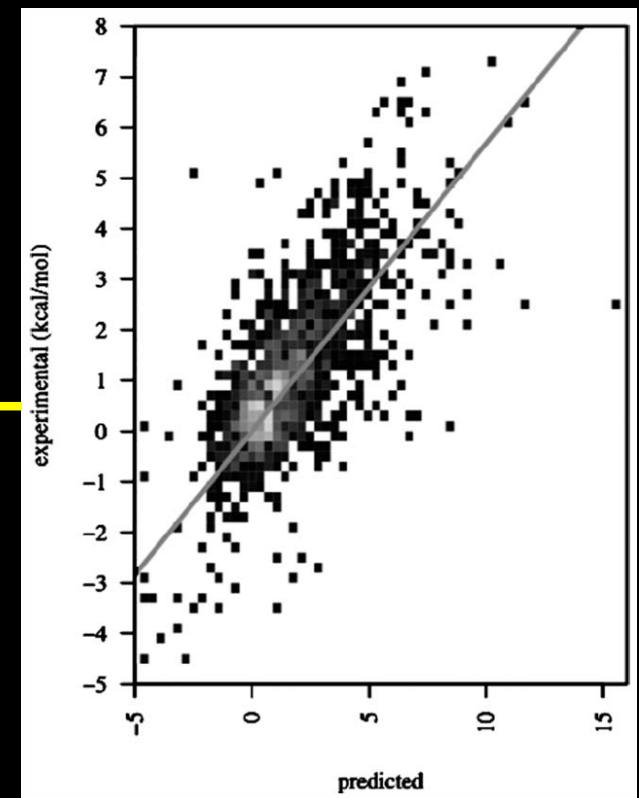
PolyPhen-2
SIFT
SNPs3D Profile
PON-P2
PROVEAN
VEST3
SNAP2

SNPs3D Stability

Structure
based SVM
binary
classifier

Calibration and
Combination

ROSETTA ddg monomer



Kellogg et al., 2010

Submission Plan

PTEN TPMT	Silent	Nonsense	Missense				
			Core/partial	Surface	Unstructured	Missing	*Use Ensemble if
submission1	wild type	zero	Ensemble	Ensemble	Ensemble	0.5	
submission2	wild type	zero	ROSETTA_ddG	ROSETTA_ddG	Ensemble	0.5	
submission3	wild type	zero	SNPs3D_Stability*	SNPs3D_Stability*	Ensemble	0.5	Prediction ≥ 0.6 or binding pocket in TPMT
submission4	wild type	zero	Ensemble	ROSETTA_ddG	Ensemble	0.5	
submission5	wild type	zero	Ensemble	SNPs3D_Stability	Ensemble	0.5	Prediction ≥ 0.6 or binding pocket in TPMT

Positive control

An estimated theoretical limit that only considers experiment noise

1. For each variant, use reported experimental score as the prediction.
2. Given the reported mean and standard deviation, randomly draw **N** samples from the normal distribution, where N equals the reported repeat count.

Marsaglia Polar Method

$$s = x^2 + y^2 < 1 \quad (-1 < x < 1, -1 < y < 1)$$

$$z_1 = x \sqrt{-2\ln(s)/s} \quad z_2 = y \sqrt{-2\ln(s)/s}$$

$$\text{Sample} = z \times SD + \text{Mean}$$

3. Use the sample mean as the answer key
4. Iterate the above steps 100 times

Overall Performance

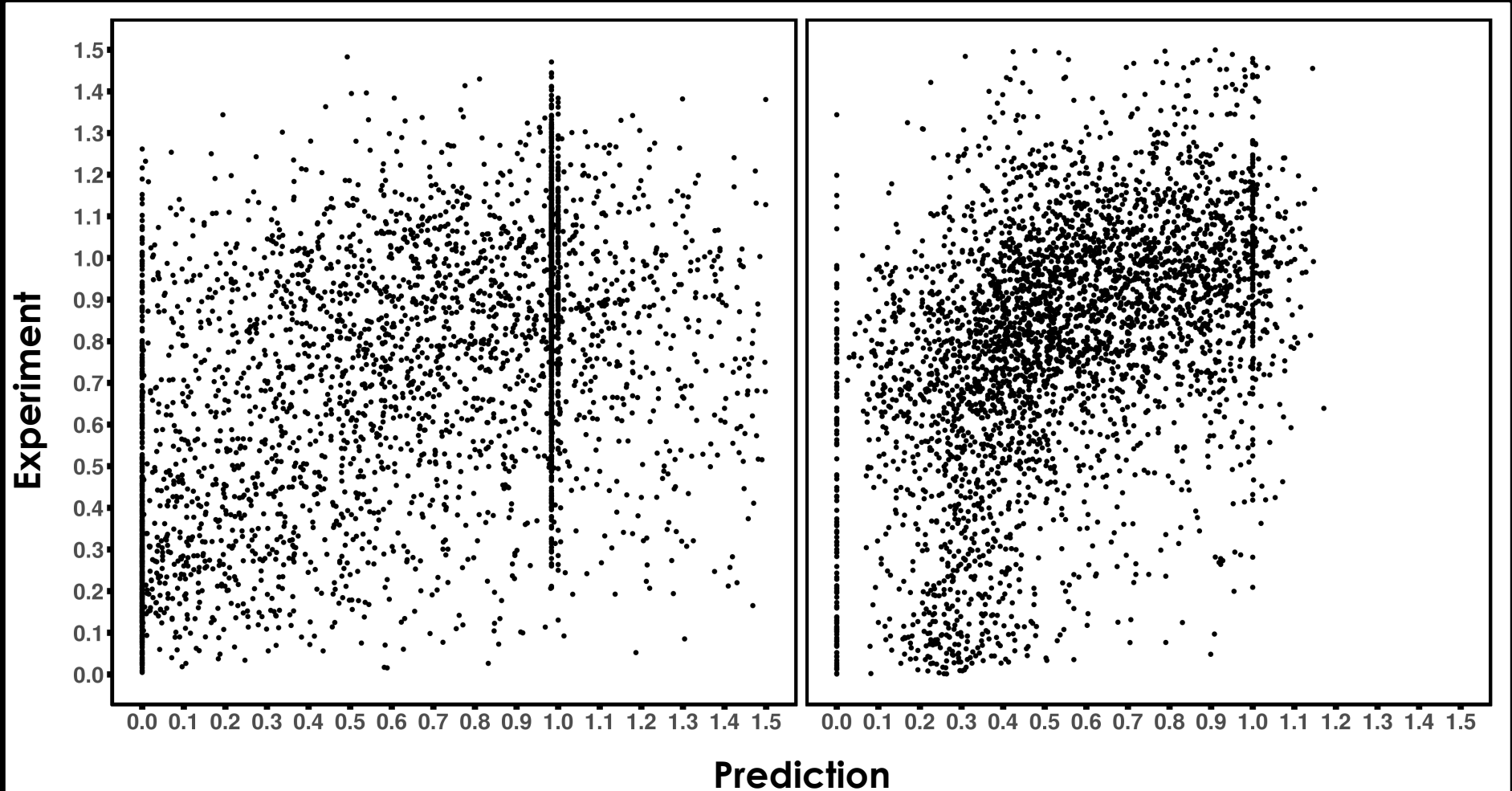
	PTEN						TPMT					
	S1	S2	S3	S4	S5	+	S1	S2	S3	S4	S5	+
RMSE	0.39	0.43	0.40	0.41	0.38	0.12	0.40	0.40	0.43	0.39	0.40	0.13
Pearson's r	0.47	0.49	0.46	0.41	0.46	0.94	0.47	0.45	0.44	0.44	0.46	0.93
Spearman's rho	0.48	0.49	0.44	0.45	0.45	0.94	0.49	0.45	0.44	0.46	0.47	0.91

Variants with negative experiment scores are excluded in all analyses

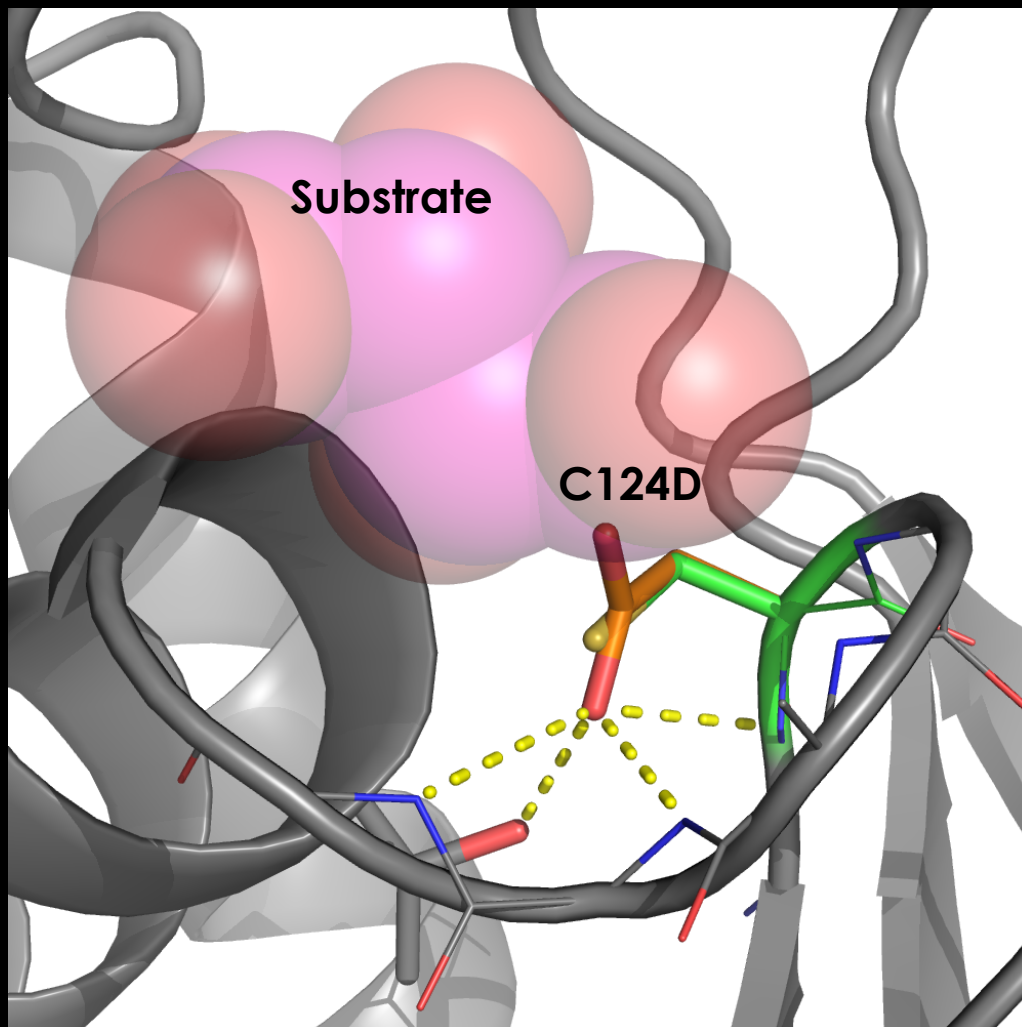
Overall scatter plots

PTEN – Submission2 (ROSETTA ddG)
RMSE 0.43, Pearson's r 0.49,
Spearman's ρ 0.49

TPMT – Submission1 (Ensemble)
RMSE 0.40, Pearson's r 0.47,
Spearman's ρ 0.49



One example of the outliers



PDB: 1D5R

PTEN – C124D

Experimental score: **1.414**

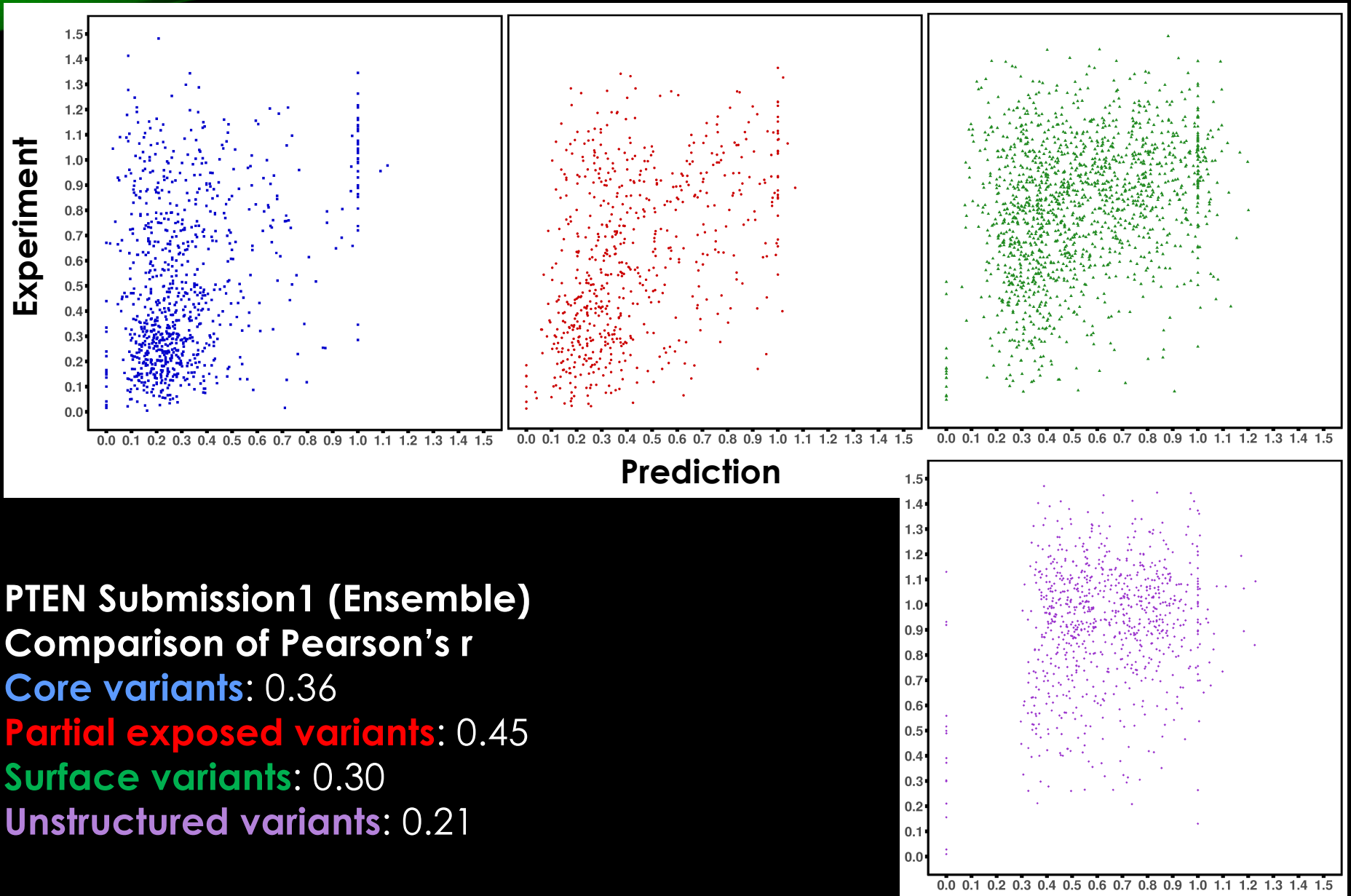
Ensemble score: **0.087**

(All **7** individual methods
predict deleterious)

ROSETTA score: **0.776**

Potentially disrupts substrate
binding, but may not affect
protein stability

Performance on subsets



PTEN Submission1 (Ensemble)

Comparison of Pearson's r

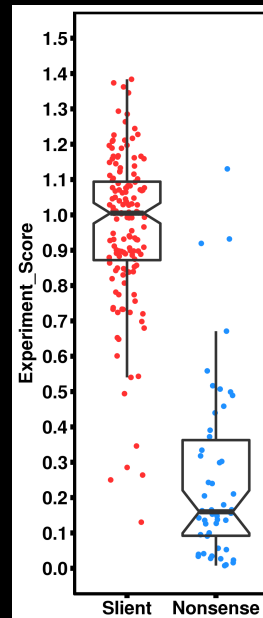
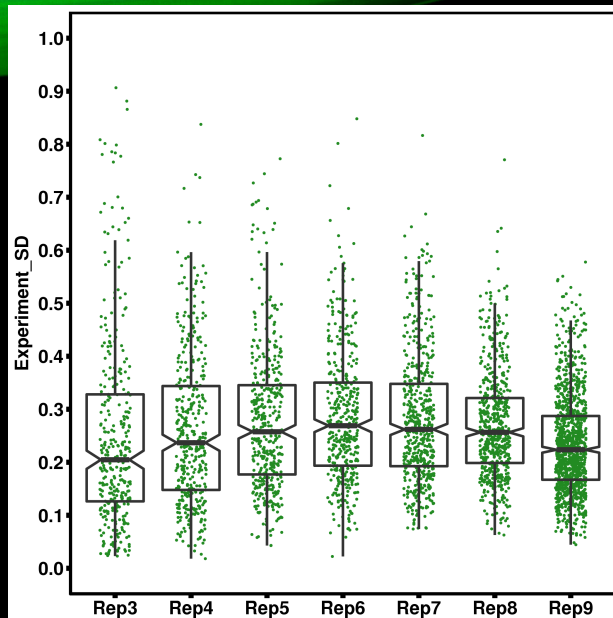
Core variants: 0.36

Partial exposed variants: 0.45

Surface variants: 0.30

Unstructured variants: 0.21

Experiment noise

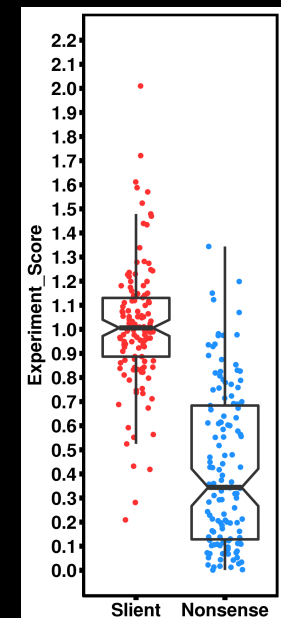
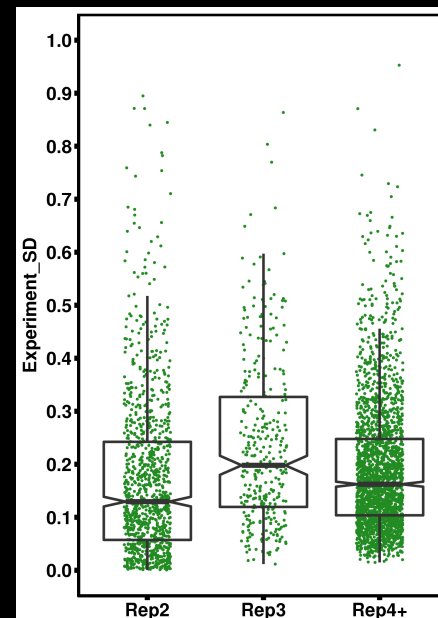


PTEN

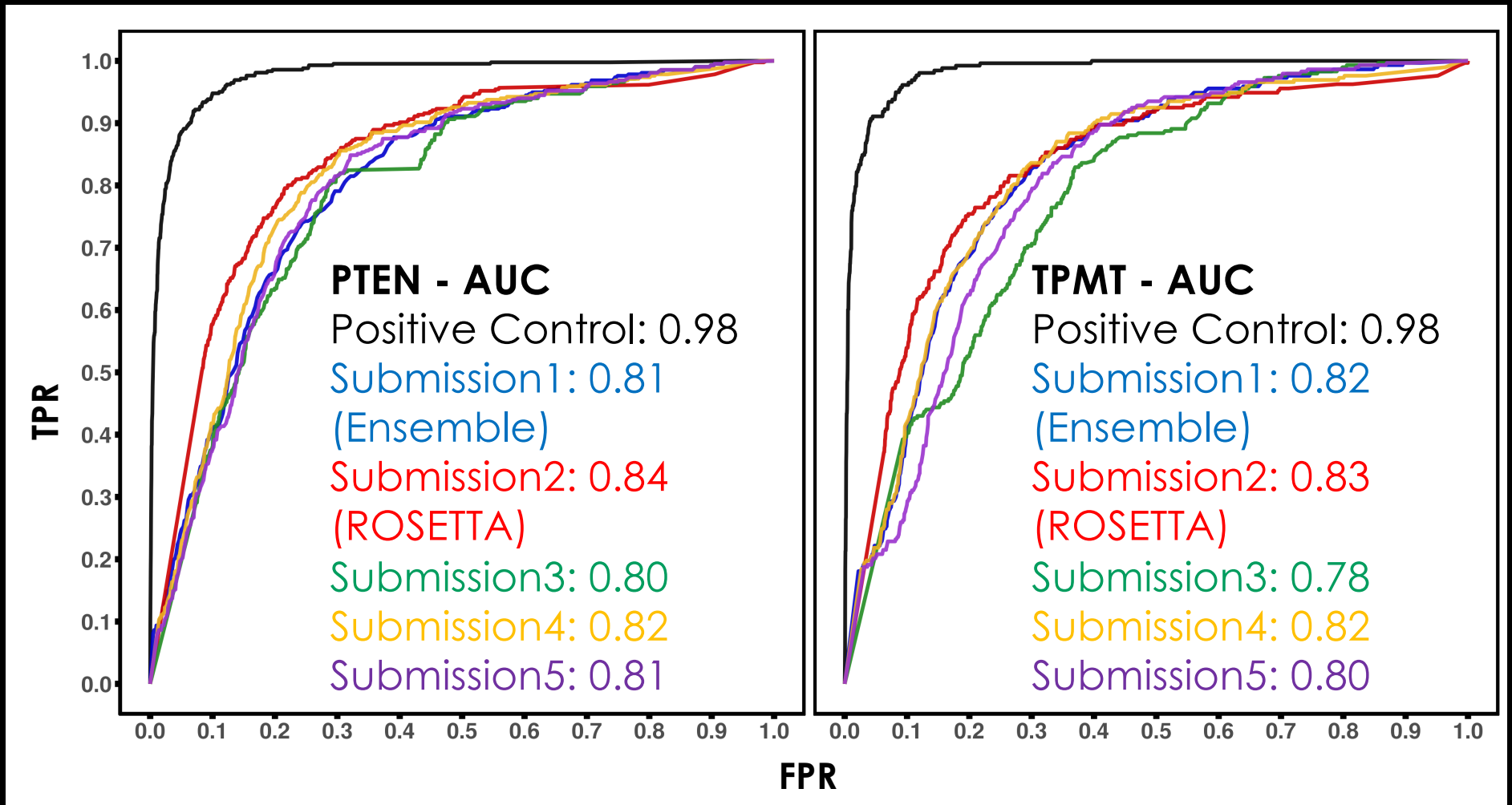
- Subset of 9-repeat variants gives Pearson's r of **0.50**
- 24 (17%) of silent variants have experimental score < 0.8
- 21 (42%) of nonsense variants have experimental score > 0.2

TPMT

- Subset of >3-repeat variants gives Pearson's of **0.54**
- 17 (13%) of silent variants have experimental score < 0.8
- 84 (65%) of nonsense variants have experimental score > 0.2



ROC curve and AUC



Unstable protein cutoff at 0.25

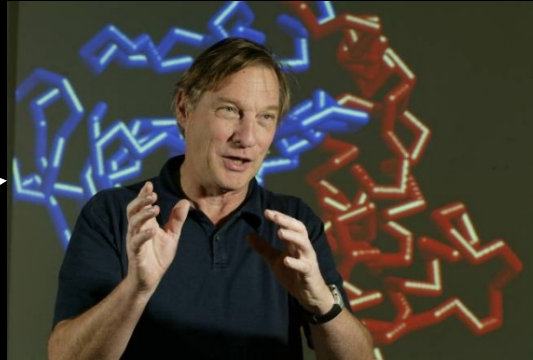
Summary

- ❑ Structure based methods gives prediction as good as sequence based methods on large data sets. The two methods are not highly correlated, asking for better way to combine them.
- ❑ Both methods correctly captured the difference between variant subsets. The unstructured subset is the most difficult one to predict.
- ❑ The theoretical upper limit of the prediction will not be affected much by the experiment noise, as long as there are enough number of repeats.
- ❑ Investigation on the silent variants, nonsense variants, and low repeat variants didn't show substantial impact on the performance.
- ❑ The ROC curve and AUC are more clinically relevant. Both methods achieved a moderately good AUC comparable to the NAGLU challenge.

ACKNOWLEDGEMENTS

Moult Lab

John Moult
Lipika R. Pal
Kunal Kundu



Data Providers

Kenneth Matreyek
Lea Starita
Doug Fowler

Assessor

Yana Bromberg

CAGI

John Moult
Steven E. Brenner
Gaia Andreoletti
Other organizers