

CAGI slide pack guidelines

I agree to my slides being used according to the following terms: _____ *Nick Grishin* _____

- 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 2013. 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, bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.



CALM1 challenge assessment

Jing Zhang



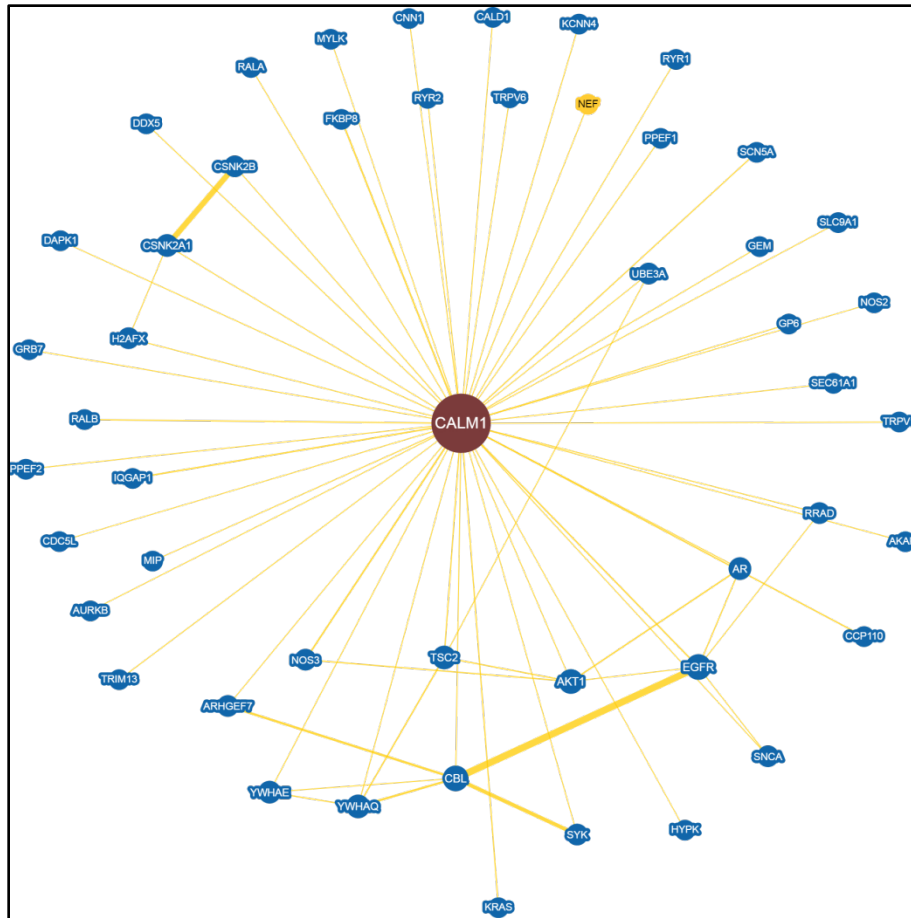
Lisa N. Kinch

Nick V. Grishin

*Departments of Biophysics and Biochemistry,
University of Texas Southwestern Medical Center at Dallas,
Howard Hughes Medical Institute*

Calmodulin Function: Calcium Binding

Functional Residues = Peptide Binding



Physical Interactions in BioGRID:

- 330 high throughput
- 185 low input

Calmodulin Structure

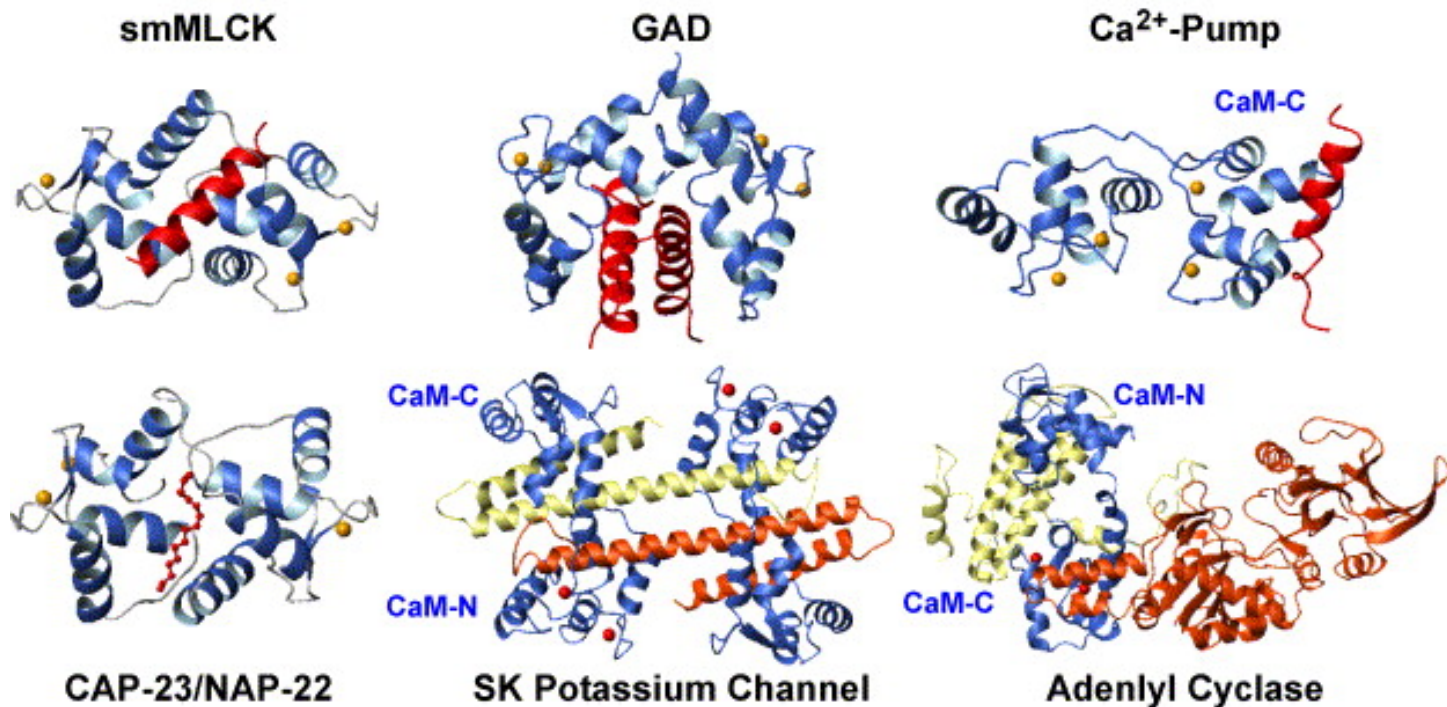


Calmodulin Structures

Structures listed in UniProt

8 EM, 35 NMR, 91 Xray

Binding examples from PMID: 15590057



CALM1 Mutation Datasets

High-accuracy subset of **1813** **single** amino acid variants for which at least three independent barcoded clones are represented, providing internal replicates of the experiment.

4 groups participated

Some groups submitted 2 predictions,
e.g. group1_1 and group1_2

Group ID	Submissions	
Group1	Group1-1	Group1-2
Group2	Group2-1	Group2-1
Gruop3	Gruop3-1	
Group4	Group4-1	Group4-2

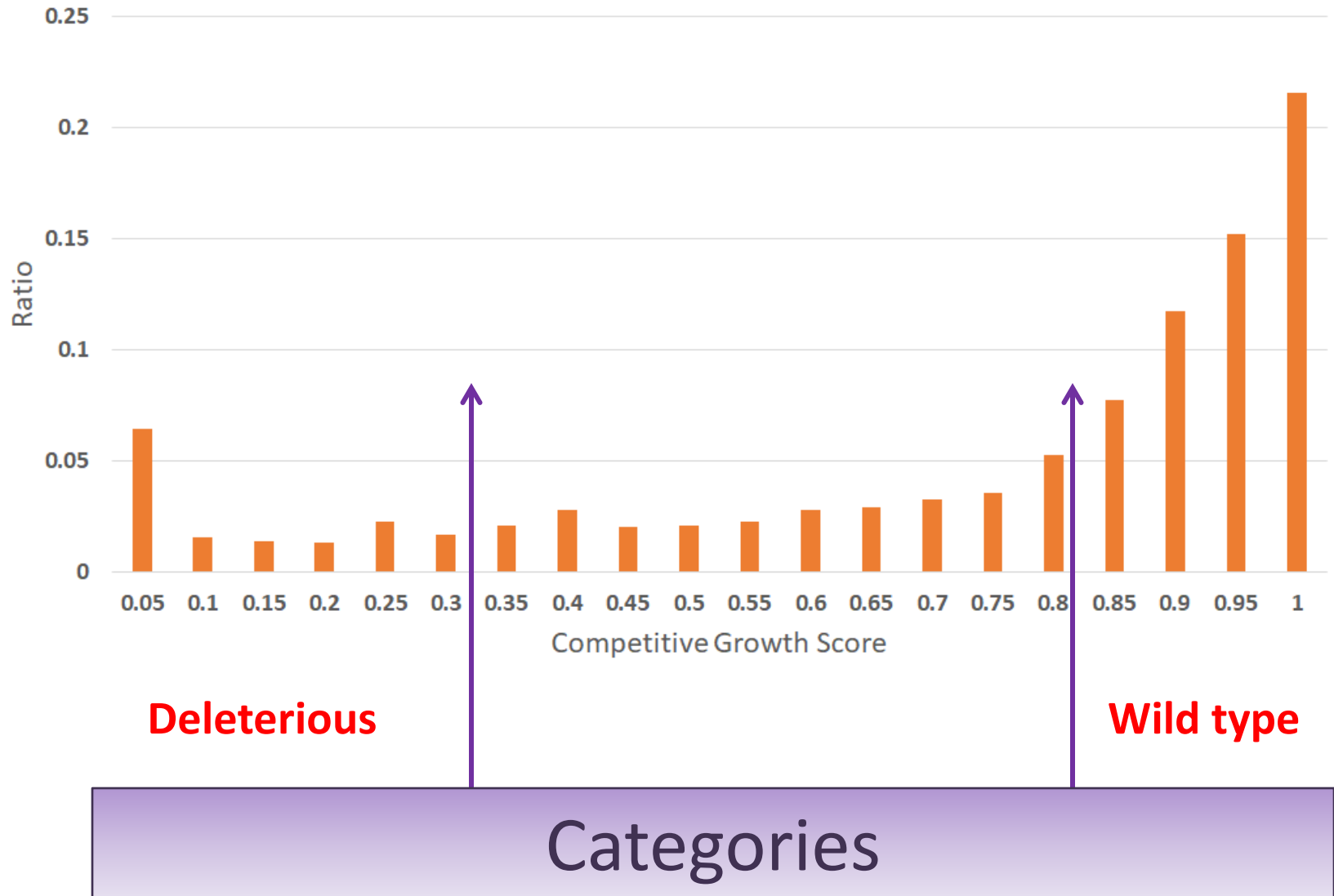
Our tasks:

1. Develop a comprehensive and rigorous **assessment scheme**, with positive and negative controls, and ranks the groups by their performance;

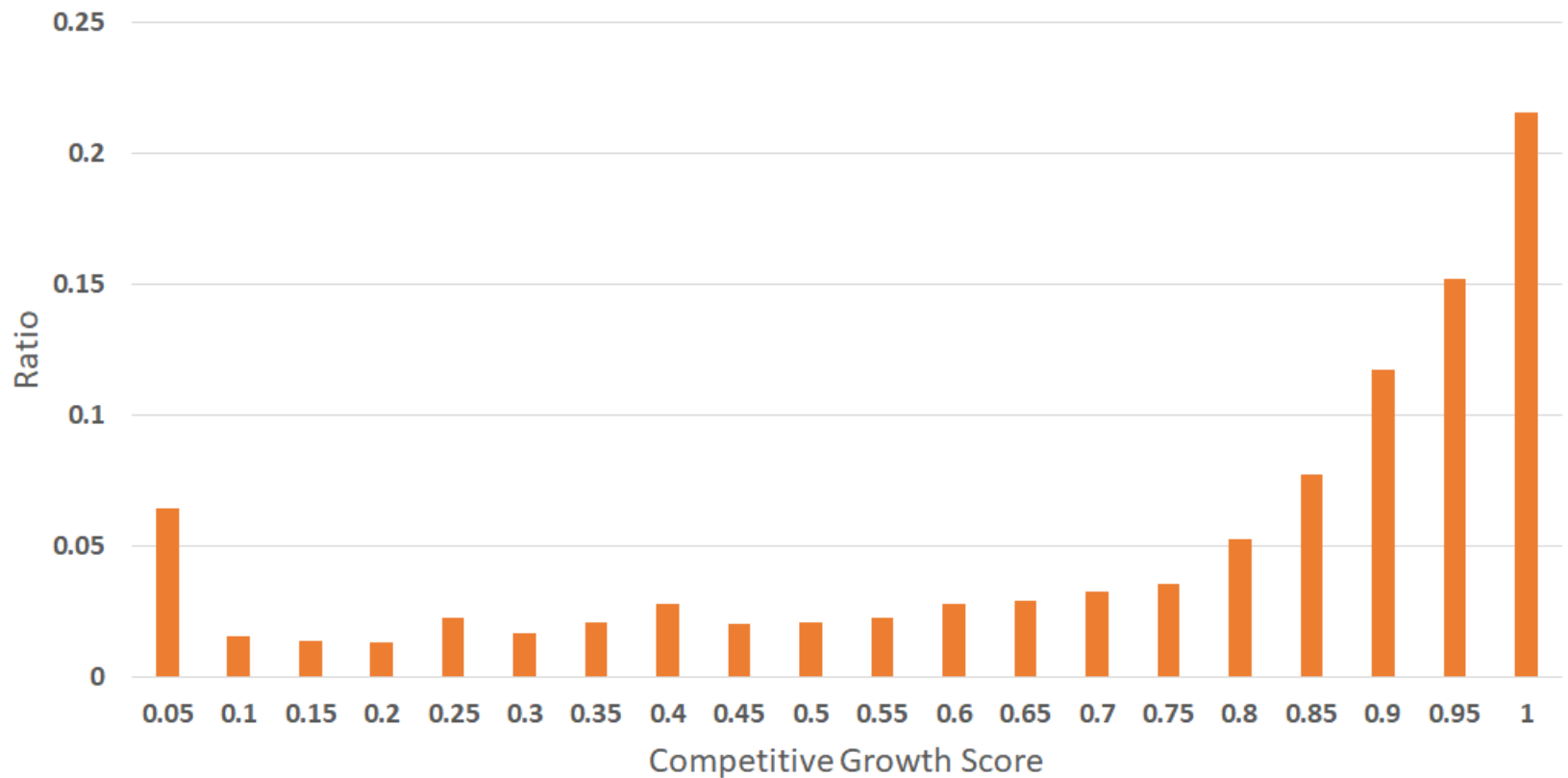
Our tasks:

1. Develop a comprehensive and rigorous **assessment scheme**, with positive and negative controls, and ranks the groups by their performance;
2. Compare predictions for different mutations and see **what works and what does not**.

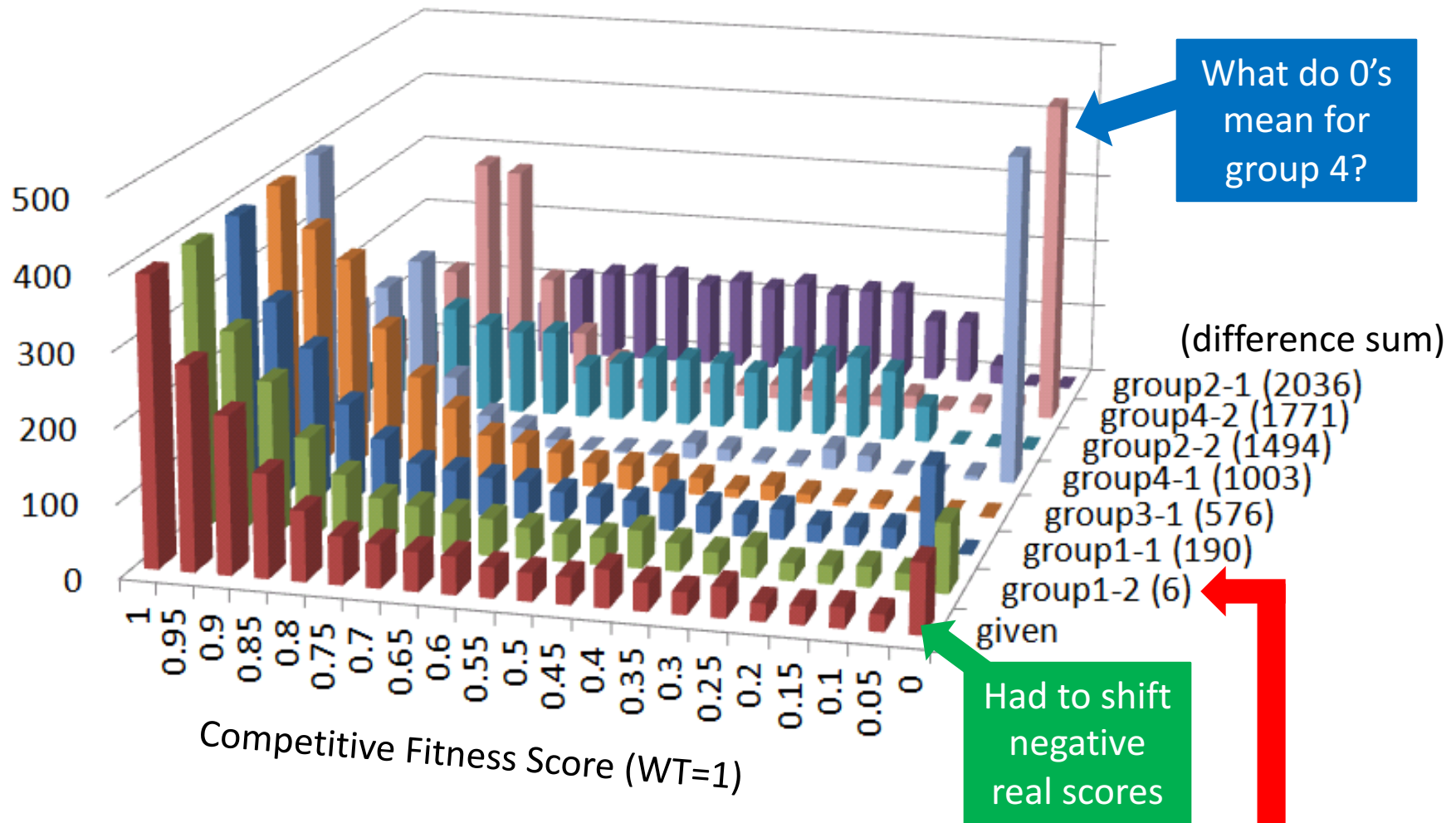
CALM Mutation Datasets



Predictors were given the distribution of experimental values



Given Experimental Distribution



Only one group (1-2) mimicked the given experimental distribution of scores

Considerations

- Only one groups have distribution of predicted scores very similar to the distribution of experimental scores;
- Differences in distributions wouldn't matter for evaluation based on comparing ranks;
- For RMSD-like measures to be meaningful, assessors had to transform prediction scores: **don't let it happen next time!**

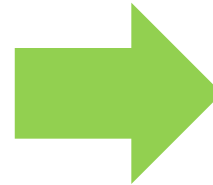
Quantile-like transformation of predicted values to match experimental values

Mutations ranked by
predicted scores

mutation	Pred rank
S56K	1
A165D	2
K45F	3
C89G	4
L214A	5

Experimental values
ranked low-to-high

rank	Exper score
1	0
2	0.12
3	0.21
4	0.34
5	0.48



Assign experimental
value to predicted
mutation

mutation	Pred transformed
S56K	0
A165D	0.12
K45F	0.21
C89G	0.34
L214A	0.48

Performance Evaluation

Since no single score captures all aspects of predictions, we used many scores measuring different properties

1. **Rank-based scores.** These don't depend on the scale of predictions.
2. **Value-based scores** computed on predictions **transformed** by us.
3. **Value-based scores** computed on **original** predictions.

General idea about which scores to use

1. **What effect is caused by a mutation:** deleterious, wild-type.
Performance on a binary classification.
E.g., using ROC curves and areas under them.
2. **Strength of the effect:** comparison of **values** and their **ranks**.
E.g., Kendall tau or RMSD-like measures.

Measures for evaluation

Three groups of measures:

Categories performance:

- ROC (Receiver operating characteristic) of deleterious mutations
- Roc for wild-type mutations
- MCC for deleterious mutations and wild-type mutations
- F1 score

Rank performance:

- Spearman's rank correlation coefficient
- Kendall tau

Value comparison:

- Pearson correlation
- RMSD
- Value distance test

Description of each measure

ROC (Receiver operating characteristic) of deleterious mutations

It is plot of the true positive rate versus the false positive. Mutations with experimental scores less than 0.3 are regarded as deleterious mutations (true). Others are not-deleterious (false) Mutations are ranked by their predicted scores from low to high, that is, from most mutations predicted to be deleterious to least deleterious. The area under the curve was used as measure.

ROC (Receiver operating characteristic) of mild/no effects mutations

It is plot of the true positive rate versus the false positive. Mutations with experimental scores more than 0.8 are regarded as mild/wild mutations (true). Predictions from each group are ranked from high to low that is, from least affecting mutations to most deleterious mutations. The area under the curve was used as measure.

MCC (Matthews correlation coefficient)

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

F1 score

Mutations are classified into 4 categories based on predicted and experimental scores. ≤ 0.3 : highly deleterious, $0.3 \sim 0.8$: moderately deleterious, $0.8 \sim 1.0$: benign. The mutations classified in the same categories by experimental scores and predicted scores are considered correctly predicted.

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad \text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN}$$

Kendall Tau-b

$$\tau_B = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}$$

$$n_0 = \frac{n(n-1)}{2}, \text{ where } n \text{ is data size}$$

n_c = number of concordant (x,y) pairs

n_d = discordant pairs

$$n_1 = \sum_j \frac{t_j(t_j - 1)}{2} \quad (t_j = \text{number } x \text{ values tied at } j\text{th value})$$

$$n_2 = \sum_k \frac{u_k(u_k - 1)}{2} \quad (u_k = \text{number } y \text{ values tied at } k\text{th value})$$

Spearman's rank correlation

$$r_s = \rho_{rg_X, rg_Y} = \frac{\text{cov}(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}}$$

ρ is correlation;

cov: covariance between rank variables

σ : standard deviation of rank variables.

Pearson correlation coefficient

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

cov is covariance of X and Y

σ is standard deviation

RMSD (Root-mean-square deviation)

$$RMSD = \sqrt{\frac{(X_i - X_e)^2}{N}}$$

X_i is predictive scores

X_e is experimental scores

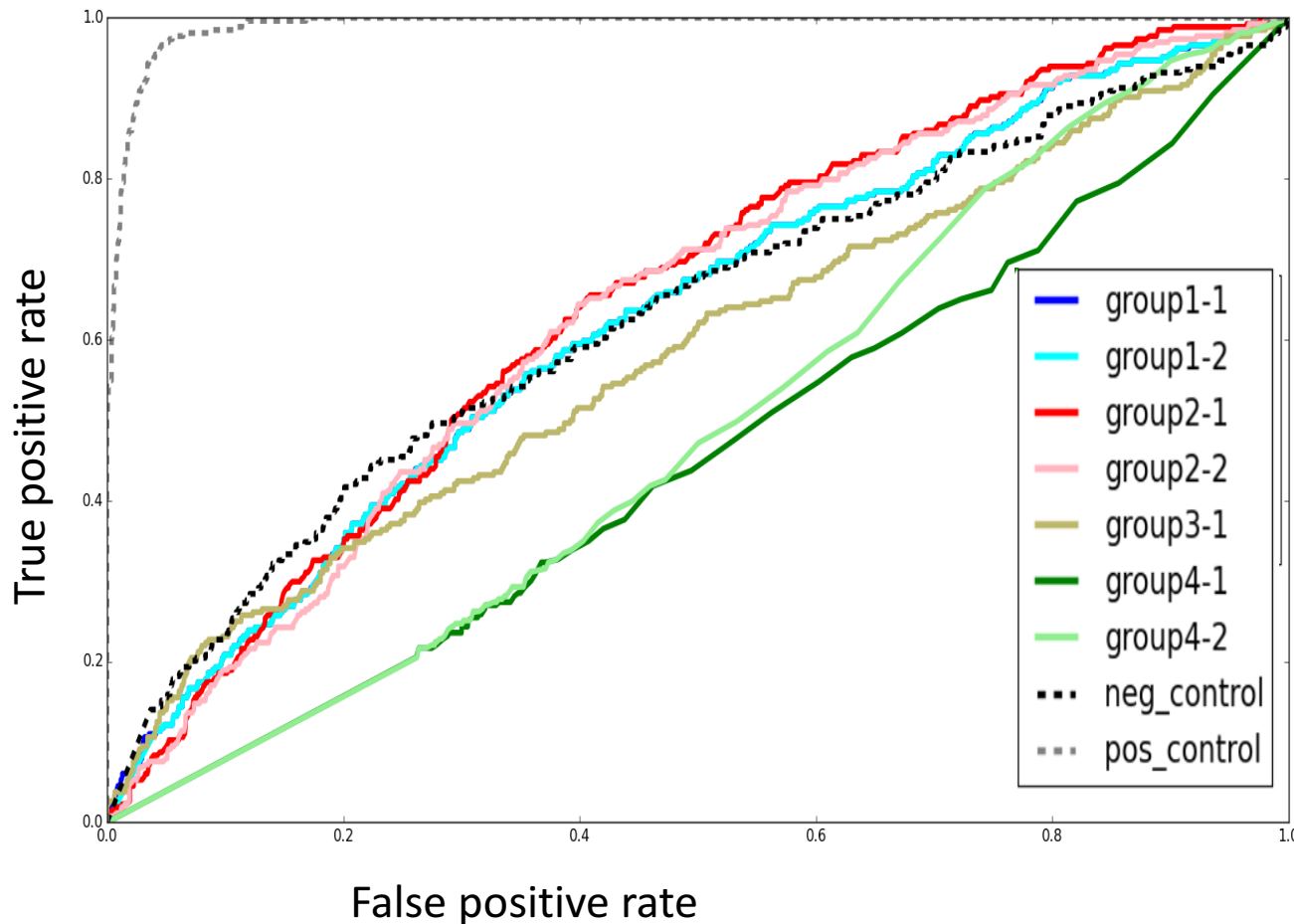
N is the number of mutations

Value difference test

It is the area under the curve of the percentage of mutations against the absolute difference between experimental score and predicted score of mutations.

ROC (deleterious vs. non-deleterious)

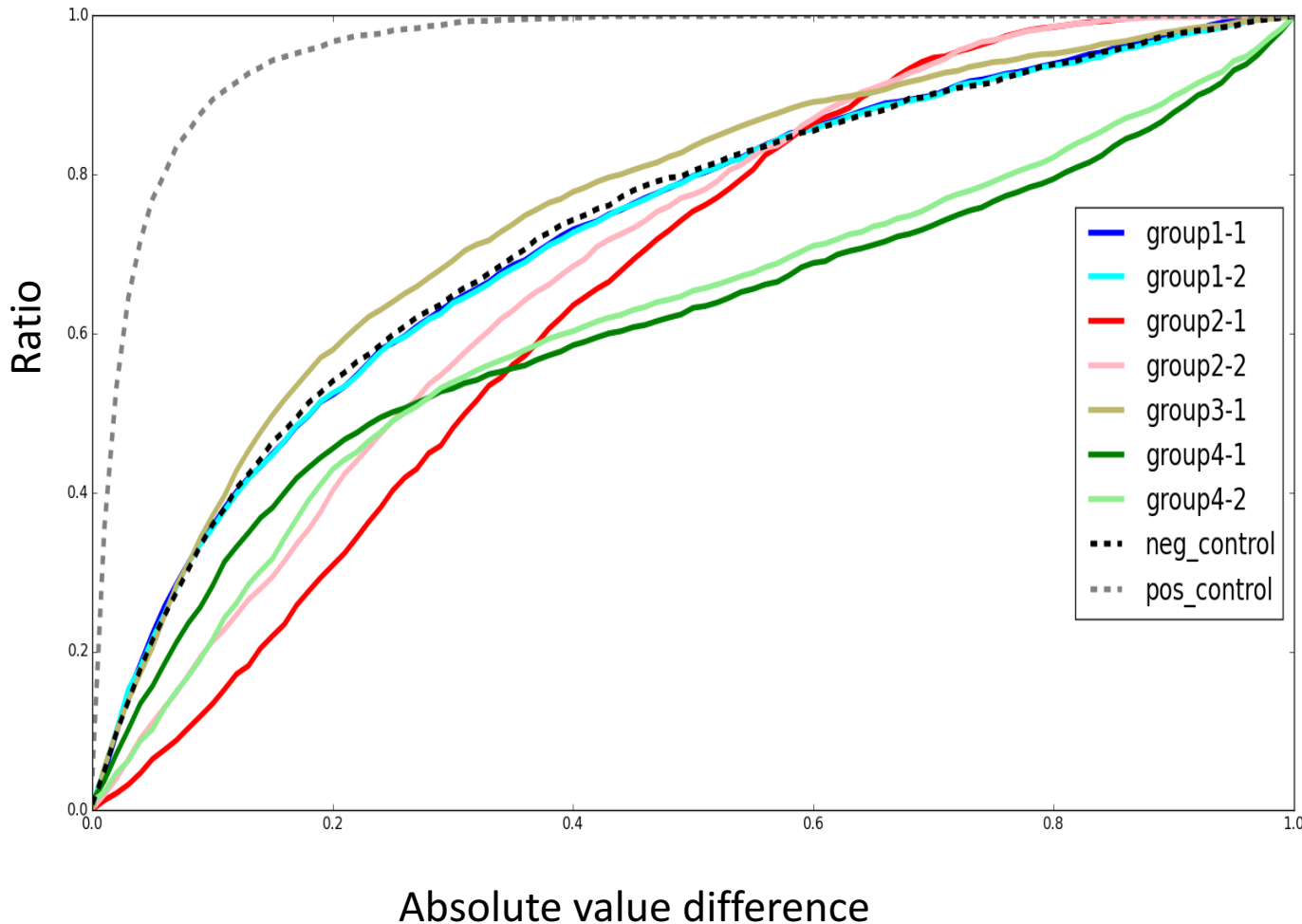
The mutations are ranked by prediction scores from low to high (most deleterious to least deleterious).



group	AUC
pos_control	0.99
group2-1	0.65
group2-2	0.64
neg_control	0.63
group1-1	0.63
gruop1-2	0.63
group3-1	0.58
group4-1	0.45
group4-2	0.49

Value difference test

It is the area under the curve of the percentage of mutations against the absolute difference between experimental score and predicted score of mutations



group	AUC
pos_control	0.95
group3-1	0.75
neg_control	0.73
group1-1	0.73
group1-2	0.73
group2-2	0.69
group2-1	0.66
group4-1	0.61
group4-2	0.61

Groups of scores

Measures were grouped into **three** categories:

(1) Rank-based (don't change with score transformation):

- Spearman's rank correlation coefficient
- Kendall tau
- ROC for deleterious mutations
- ROC for wild-type mutations

Value-based (differ between original and transformed prediction scores) applied to **(2) transformed prediction scores** and **(3) original prediction scores**:

- F1 score for overall categories identification
- MCC (Matthews correlation coefficient) for deleterious/wild-type mutations
- Pearson correlation
- RMSD
- Value distance test

CAGI CALM challenge scores

group_id	Rank based scores				original value based						rescaled value based					
	Tau	spearman	dele roc	wild roc	rmsd	pearson	value diff	mcc_dele	mcc_wild	f1	rmsd	pearson	value diff	mcc_dele	mcc_wild	f1
negative_control	0.15	0.23	0.63	0.61	0.37	0.24	0.73	0.16	0.17	0.5	0.37	0.24	0.73	0.16	0.17	0.5
positive_control	0.85	0.98	0.99	0.98	0.07	0.97	0.95	0.85	0.92	0.92	0.07	0.97	0.96	0.85	0.92	0.92
group1-1	0.15	0.22	0.63	0.61	0.38	0.22	0.73	0.11	0.16	0.48	0.38	0.22	0.73	0.11	0.16	0.48
gruop1-2	0.15	0.22	0.63	0.6	0.38	0.23	0.73	0.11	0.16	0.48	0.38	0.23	0.73	0.11	0.16	0.48
group2-1	0.17	0.25	0.65	0.61	0.4	0.24	0.66	0.13	0.1	0.29	0.38	0.24	0.73	0.11	0.17	0.48
group2-2	0.15	0.23	0.64	0.6	0.37	0.22	0.69	0.1	0.09	0.34	0.38	0.22	0.72	0.09	0.17	0.47
group3-1	0.07	0.108	0.58	0.57	0.35	0.17	0.75	0.09	0.1	0.51	0.4	0.17	0.71	0.14	0.08	0.46
group4-1	-0.02	-0.03	0.45	0.48	0.51	-0.04	0.61	-0.05	-0.02	0.4	0.43	-0.04	0.67	-0.05	-0.03	0.4
group4-2	-0.01	-0.02	0.49	0.49	0.5	-0.03	0.61	-0.03	-0.01	0.34	0.43	-0.03	0.67	-0.05	-0.03	0.4

Combine Scores:

Average Z-scores within each **category**;

Sum for categories in each **subset**;

Categories

group_id	Rank based scores				original value based					rescaled value based					
	Tau	spearman	dele roc	wild roc	rmsd	pearson	value diff	mcc_dele	mcc_wild	f1	rmsd	pearson	value diff	mcc_dele	mcc_wild

Controls!

“Control ought to be stupid.”

V. M. Stepanov

(1) Positive control

To answer the question:

How do errors in experimental measurements affect the scores?

I251K

1.24	+/-	0.32
mean		standard deviation

assume Gaussian distribution

Generate a “perfect” prediction:

for each mutation, generate a Gaussian random variable with the mean and standard deviation set to **experimental values**

Controls!

“Control ought to be stupid.”

V. M. Stepanov

(2) Negative control

To answer the question:

How does a predictor based on amino acid usage in positions of a multiple sequence alignment compares to submitted predictions?

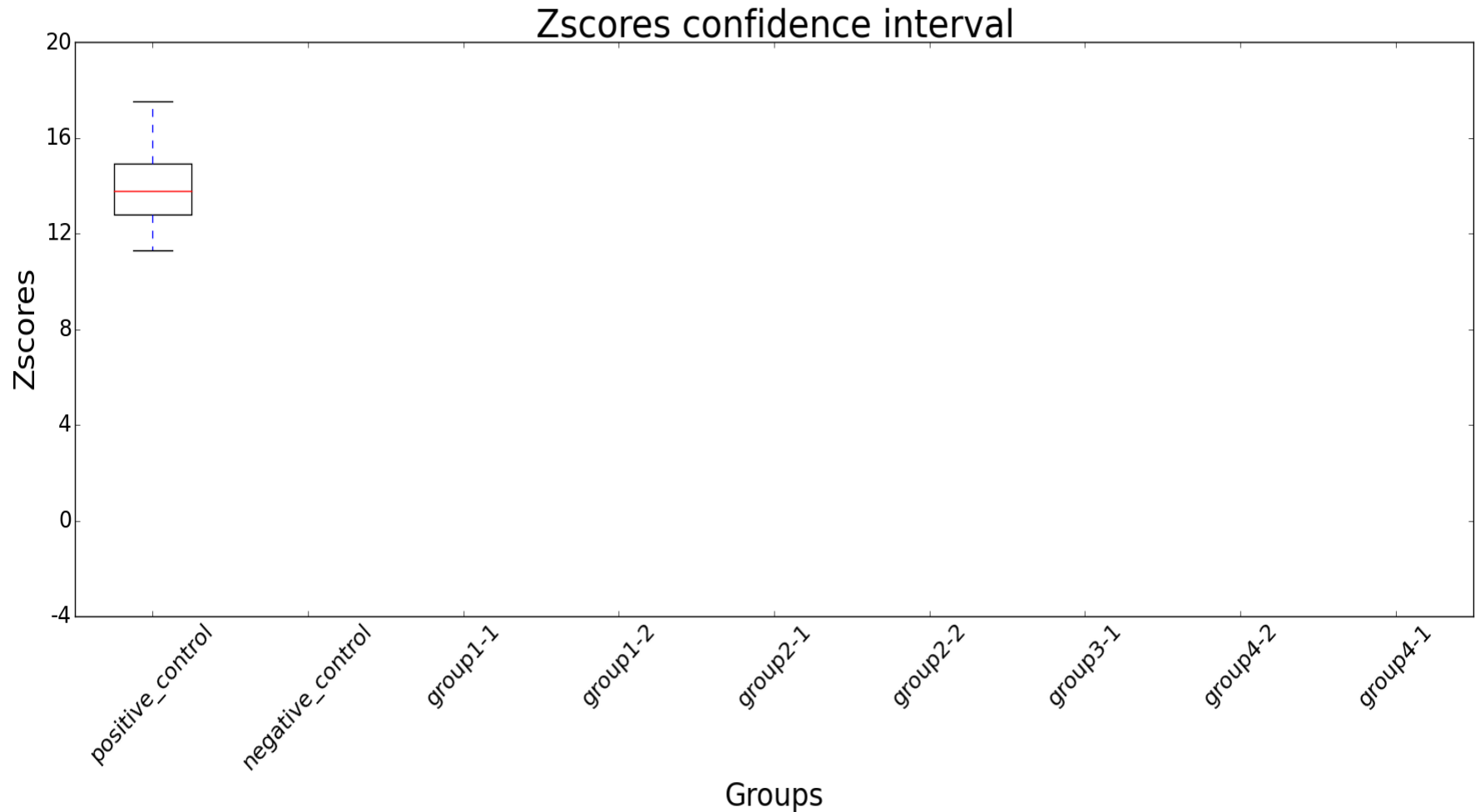
Generate a “dummy” prediction:

for each mutation, compute the score:

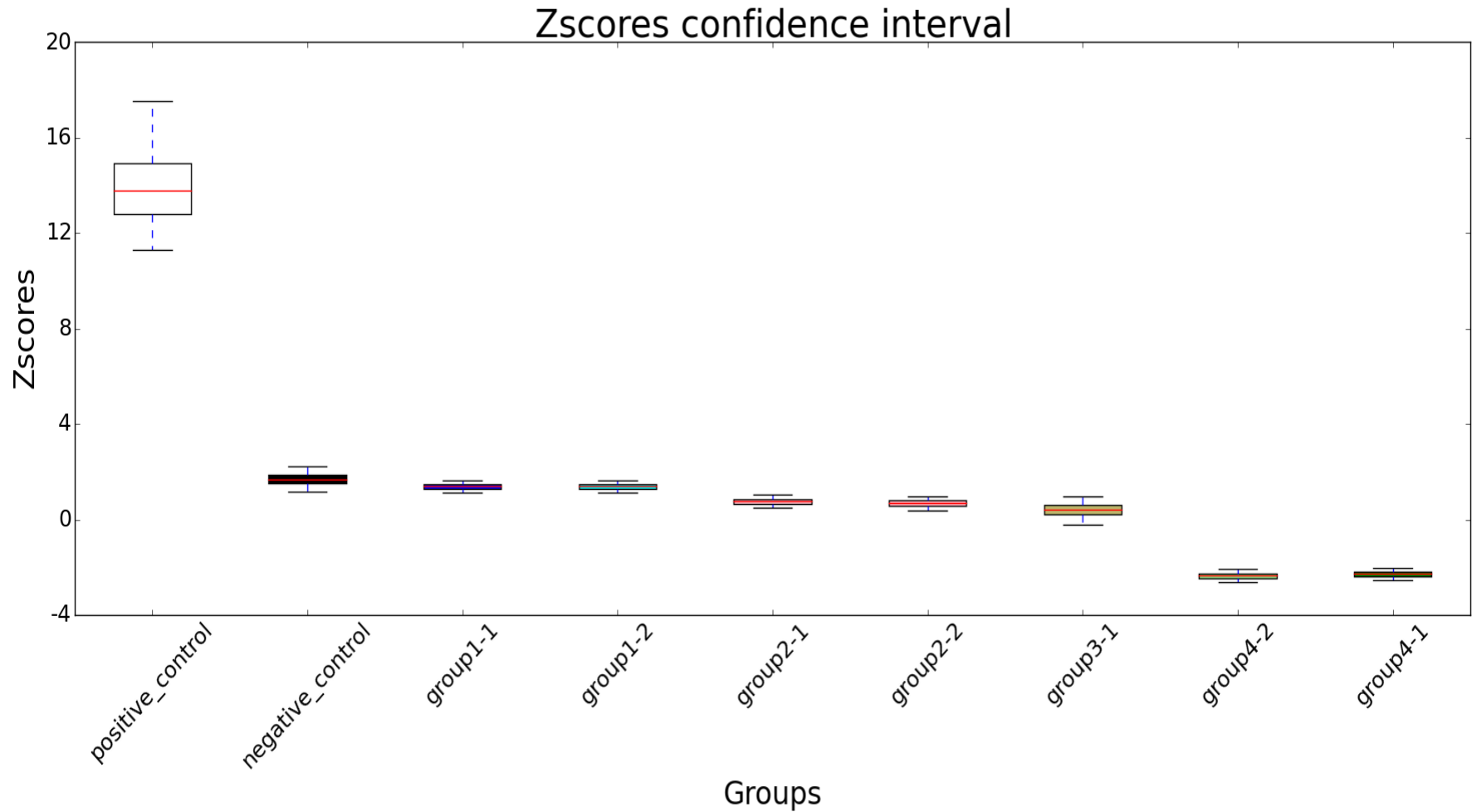
$$\ln \left(\frac{P_m}{q_m} \right) - \ln \left(\frac{P_w}{q_w} \right)$$

P_w , the probability of wild type or P_m mutated residue in multiple sequence alignment; q , the background amino acid frequencies of Robinson and Robinson. If more than one mutation – add these values. Rank mutations by these scores and assign experimental value with that rank (the same as quantile transformation we apply to predictions).

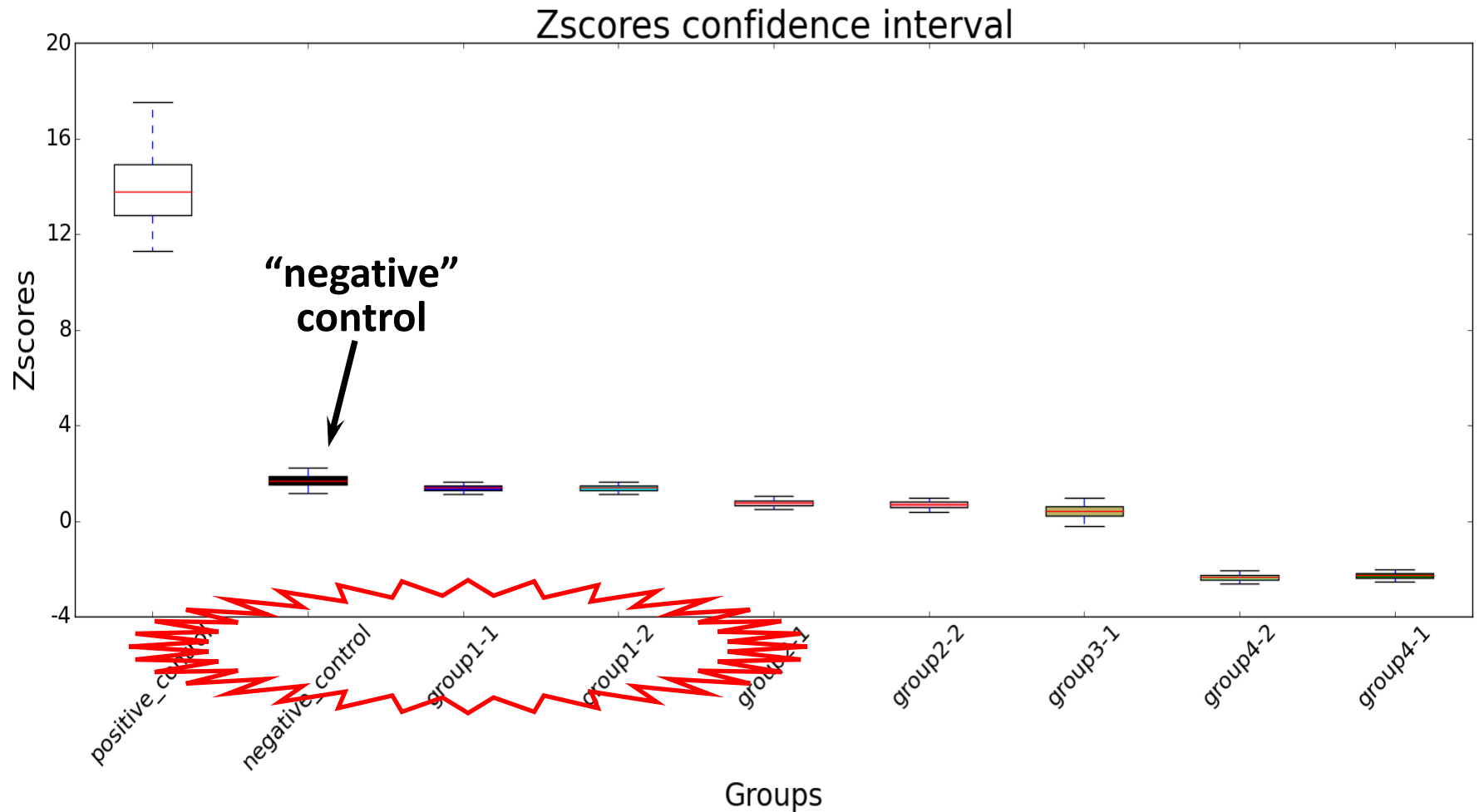
Let's see the results!



Let's see the results!



Let's see the results!



Zscores for the CALM challenge

group	Rank_AVE	Orig_AVE	Tran_AVE	Sum	Rank
positive_control	7.15	5.69	5.55	18.40	na
negative_control	0.70	0.75	0.61	2.06	1
group1-1	0.68	0.57	0.44	1.70	2
gruop1-2	0.64	0.59	0.45	1.68	3
group2-1	0.87	-0.01	0.48	1.34	4
group2-2	0.69	-0.02	0.31	0.98	5
group3-1	-0.13	0.36	0.22	0.46	6
group4-2	-1.27	-0.78	-0.95	-2.99	7
group4-1	-1.49	-0.72	-0.96	-3.16	8

Final ranking for the CALM challenge

group	Sum	Rank
negative_control	2.06	1
group1-1	1.70	2
gruop1-2	1.68	3
group2-1	1.34	4
group2-2	0.98	5
group3-1	0.46	6
group4-2	-2.99	7
group4-1	-3.16	8

Correlation between predictions of groups

		group1-2	group2-1	group2-2	group3-1	group4-1	group4-2	negative control
Sequence-based	group1-1	0.99	0.74	0.73	0.46	0.15	0.15	0.62
	group1-2		0.73	0.72	0.45	0.15	0.15	0.62
Integration of multiple predictors	group2-1			0.98	0.45	0.13	0.13	0.68
	group2-2				0.45	0.13	0.13	0.67
Machine learning for stability	group3-1					-0.06	-0.06	0.5
	group4-1						0.97	0.07
Molecular dynamics	group4-2							0.07

Statistics and confidence

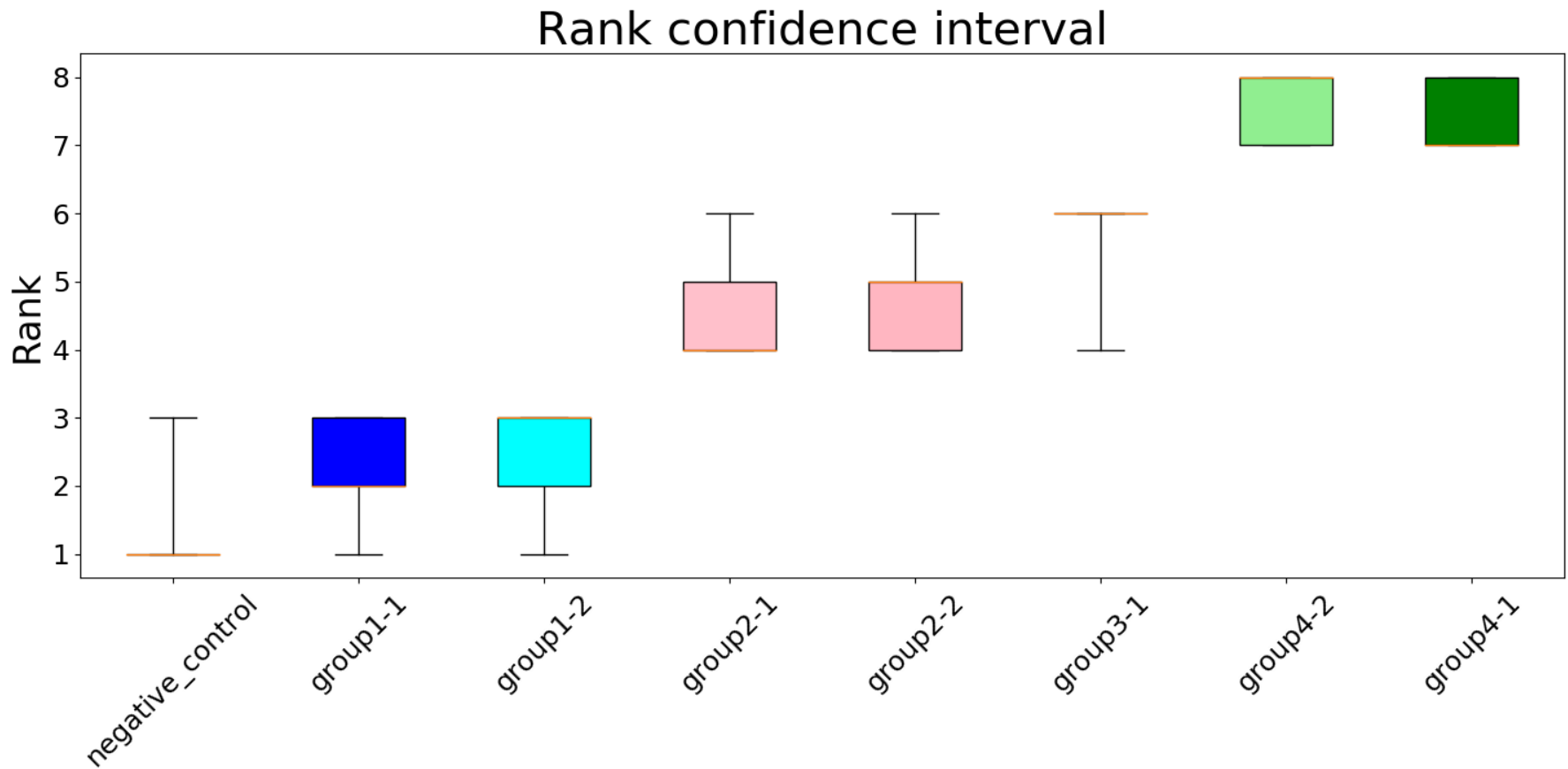
To estimate the confidence of performance of predictors, we simulated **50** experimental datasets assuming the experimental score for each mutation has a distribution

$\text{Guassian}^{\sim}(\text{exp score}, \text{exp stderr})$

and for each dataset, **100** bootstrap replicates on mutations were generated. In total, we have **5000** datasets to compute confidence intervals and head-to-head tests.

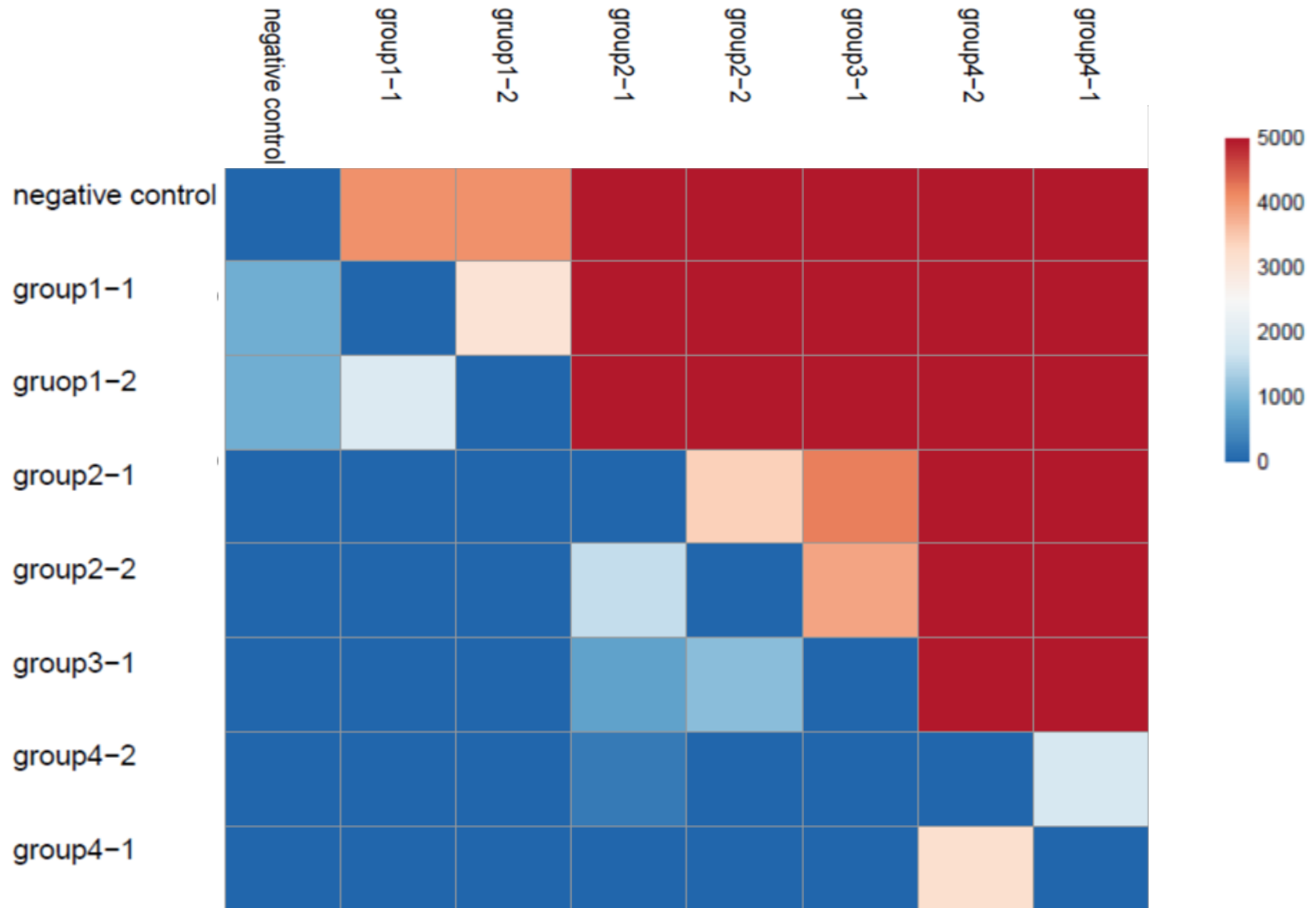
95% confidence intervals of ranks

Combined Scores: Average Z-scores for each Category & Sum them for final scores



Head-to-head test

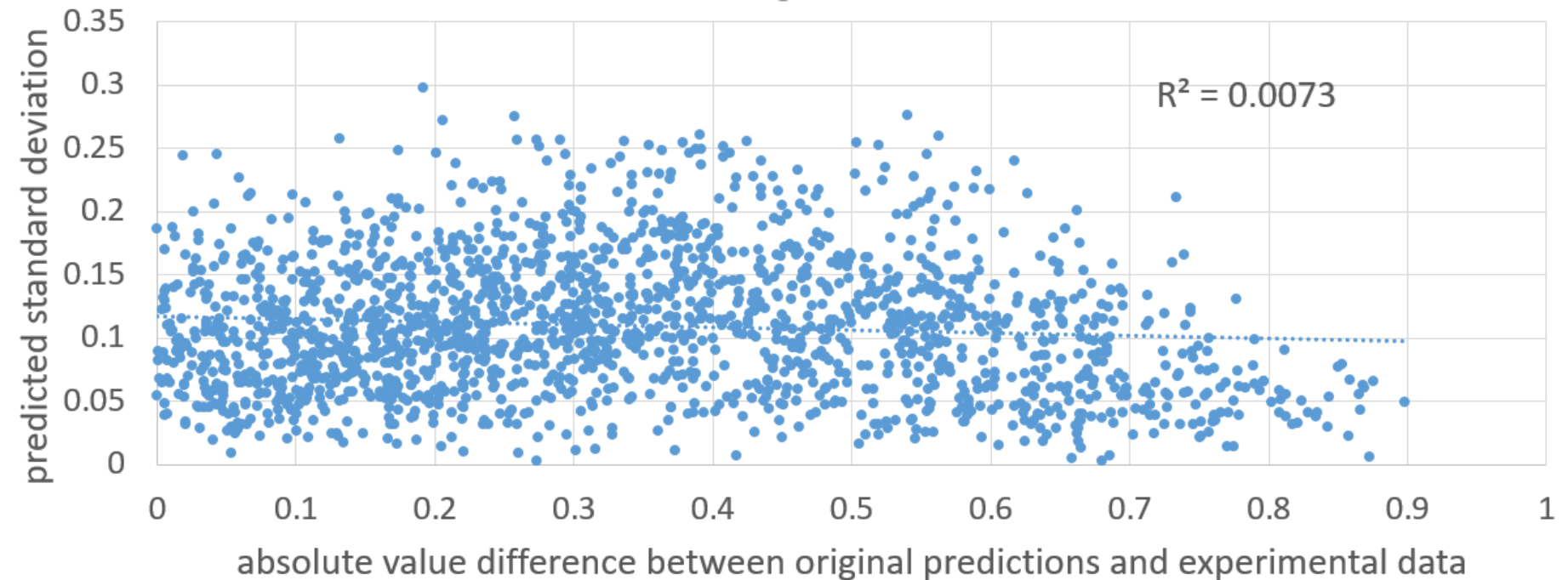
In the 5000 simulated replicates, we compare a pair of groups and calculate the number of replicates one group beats the other



Evaluation on standard deviation prediction

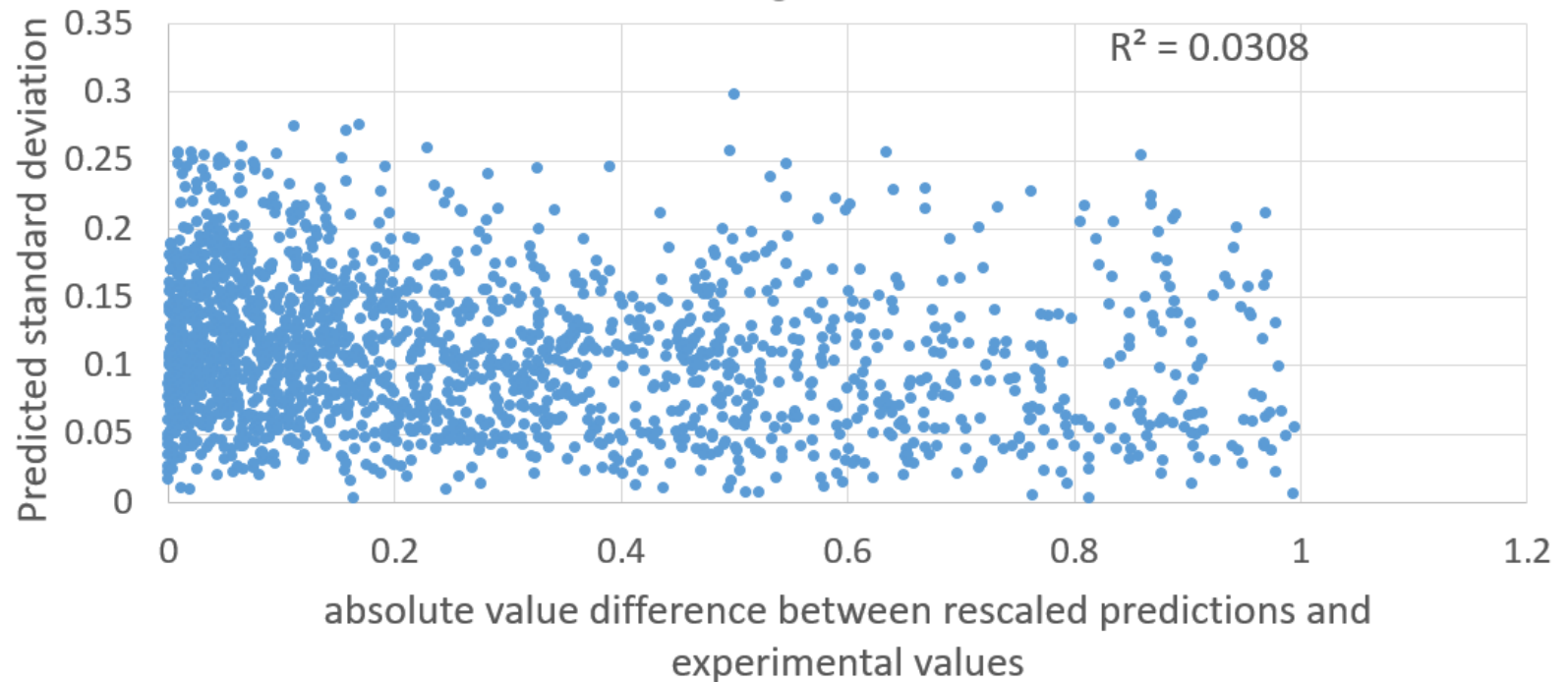
Only **one** group (group2) assigned different standard deviation to each mutation and they didn't perform very well

No correlation



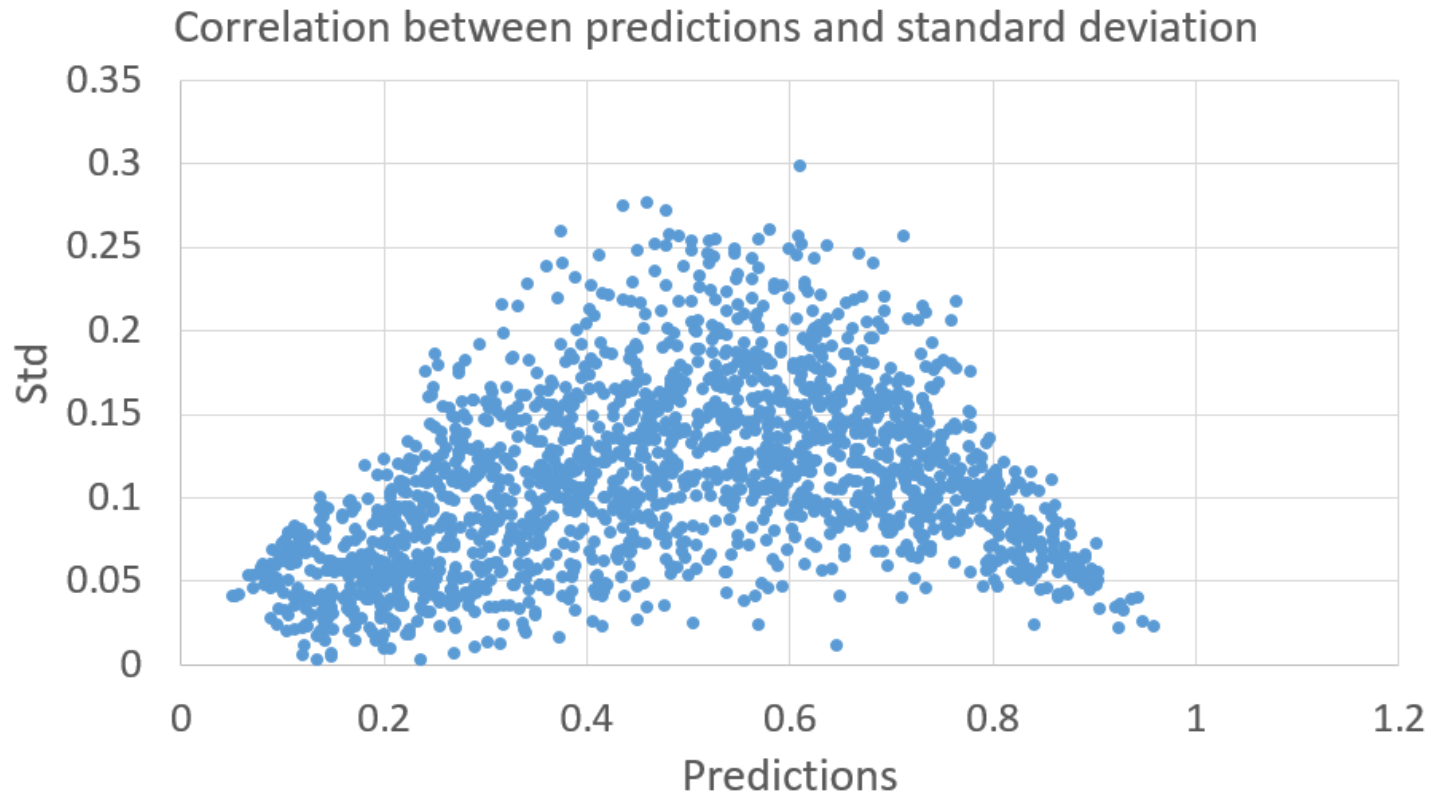
Evaluation on standard deviation prediction

No correlation



Evaluation on standard deviation prediction

Only **one** groups assigned different standard deviation to each mutation and they didn't perform very well



For mutation predicted to be deleterious or benign, the standard deviation is lower the mutations predicted to have intermediate effects.

Conclusions about ranking

Among four groups, not a single one performed significantly better than our “negative” control.

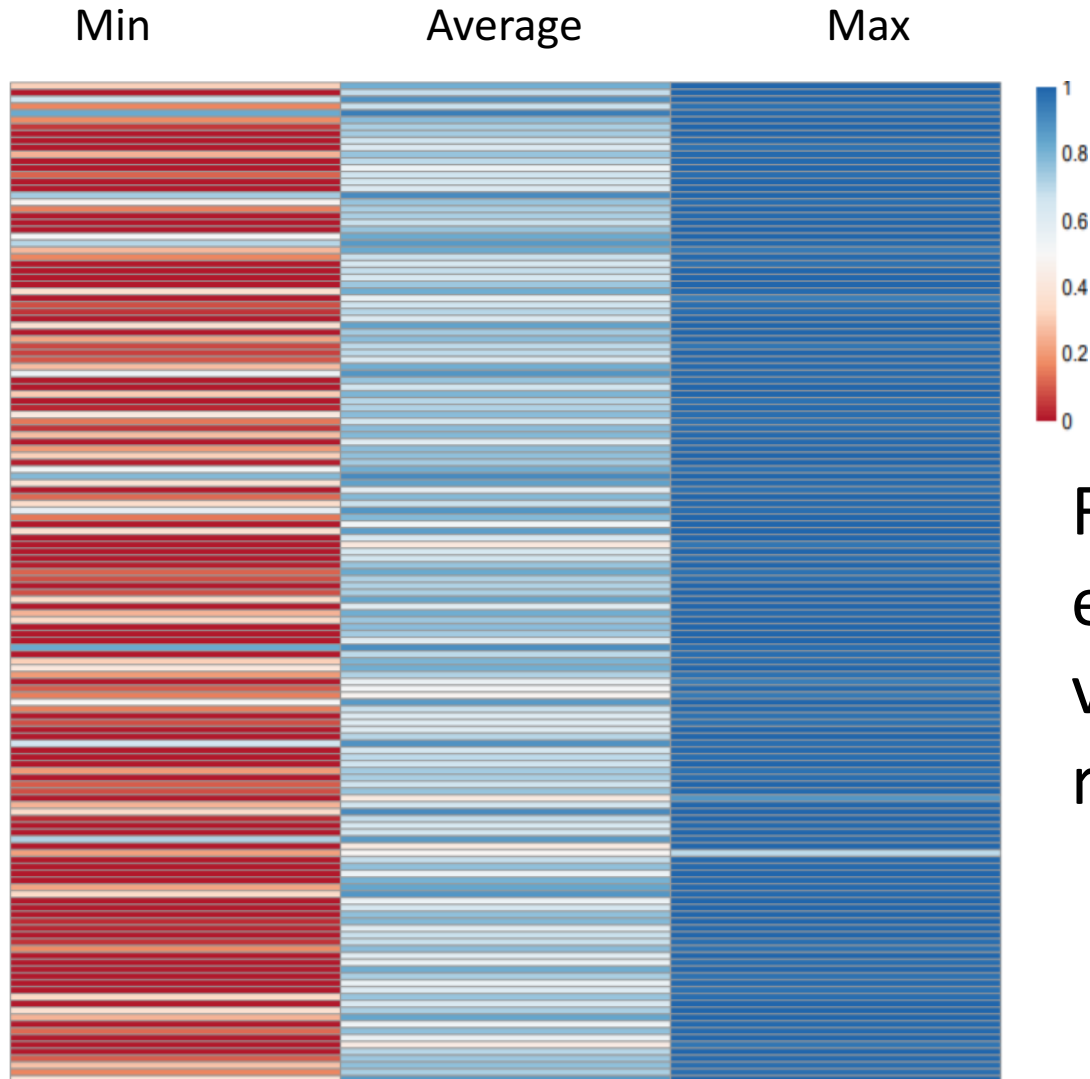
Q1: What contributes to good predictions?

Well predicted mutations

vs.

Badly predicted mutations

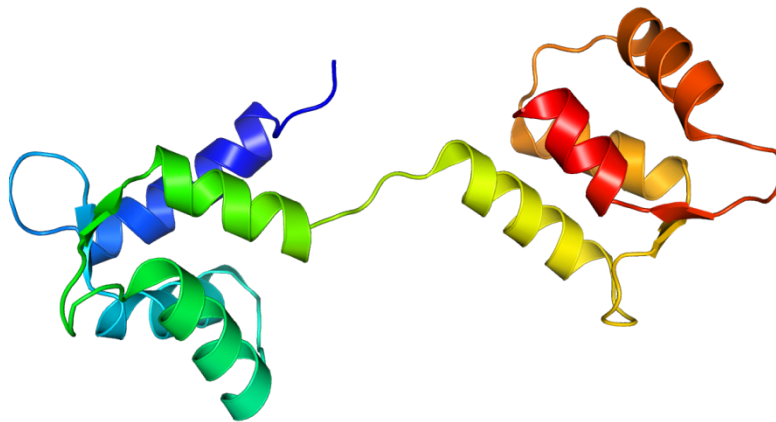
Simple statistics for experimental score at each position



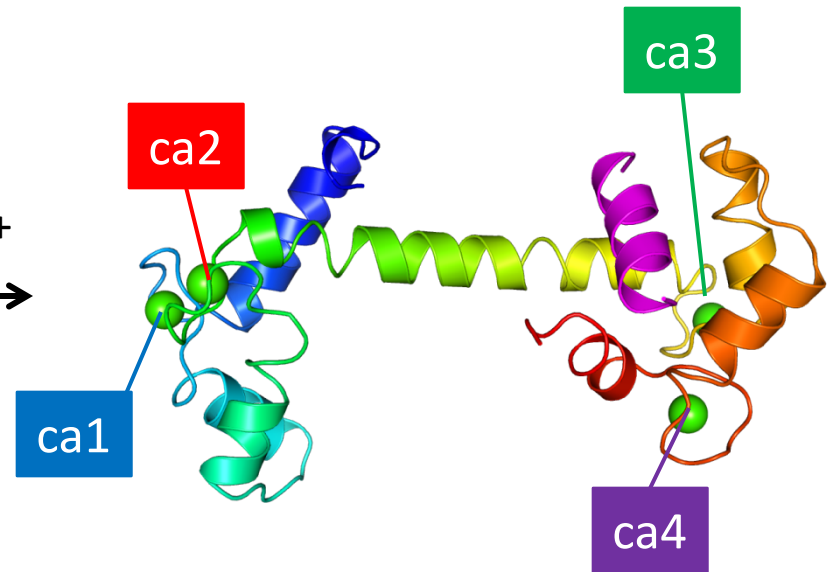
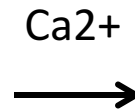
For most positions, the effects of mutations vary based on mutated residues.

Calmodulin Function: Calcium Binding

Functional Residues = Calcium Binding



1ljj: Apo calmodulin



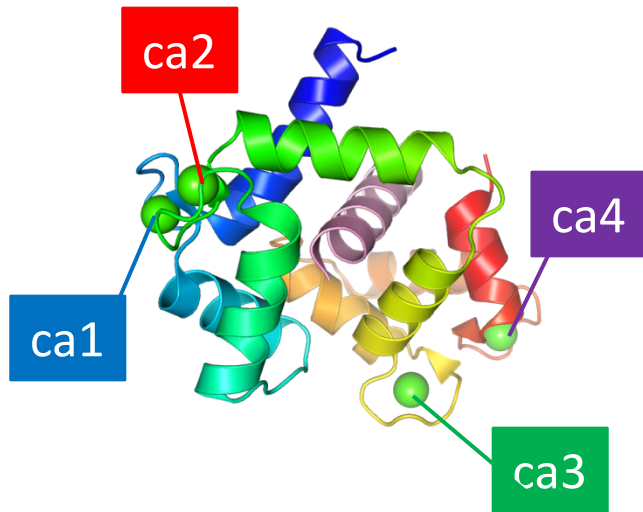
4djc: Calmodulin in complex with the inactivation gate (DIII-IV linker) of the cardiac sodium channel (Na(V)1.5)

Calmodulin Function: Calcium Binding

Functional Residues = Peptide Binding

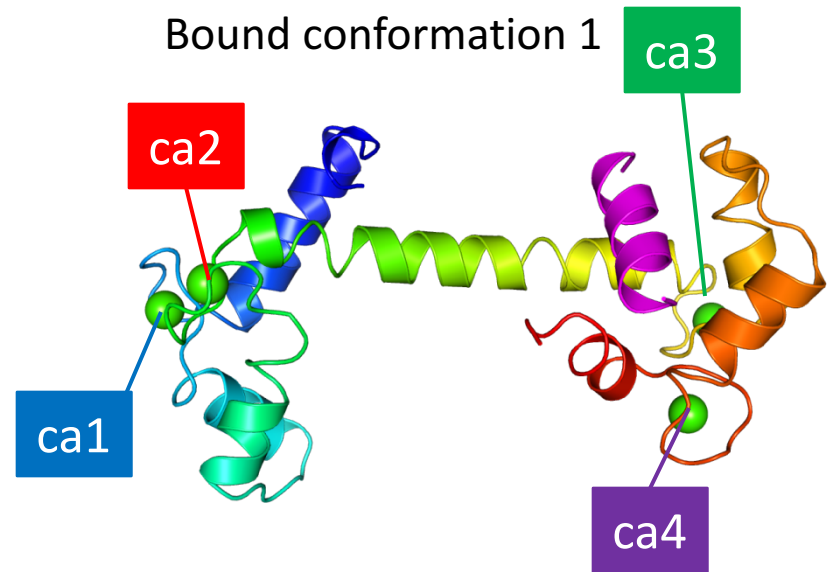
Also several peptide binding modes

Bound conformation 2



2f3y: Calmodulin in complex with the Hydrophobic **IQ Domain** of the Cardiac Ca(v)1.2 Calcium Channel

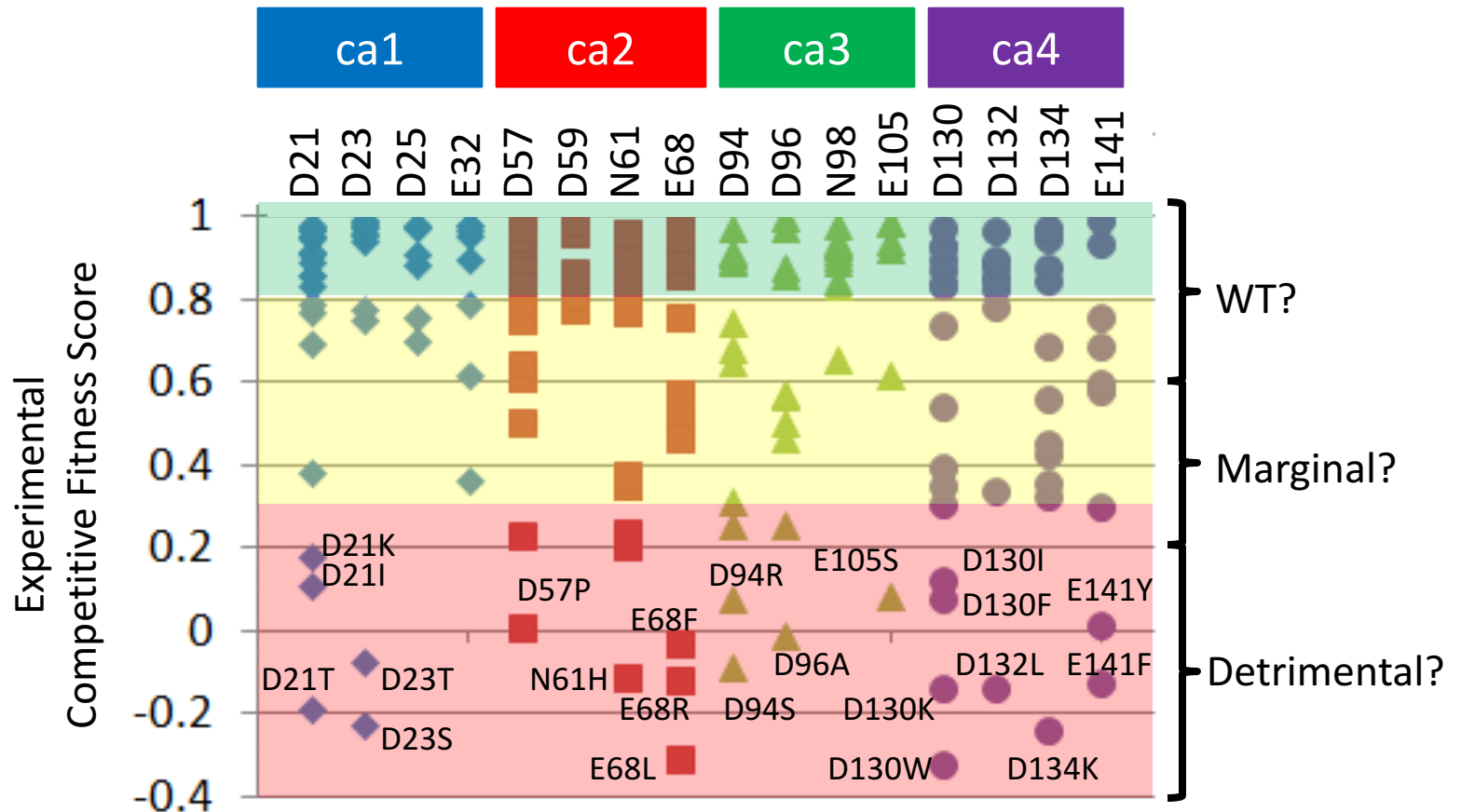
Bound conformation 1



4djc: Calmodulin in complex with the inactivation gate (**DIII-IV linker**) of the cardiac sodium channel (Na(V)1.5)

Experimental Competitive Fitness Scores

Functional Residues: Calcium Binding



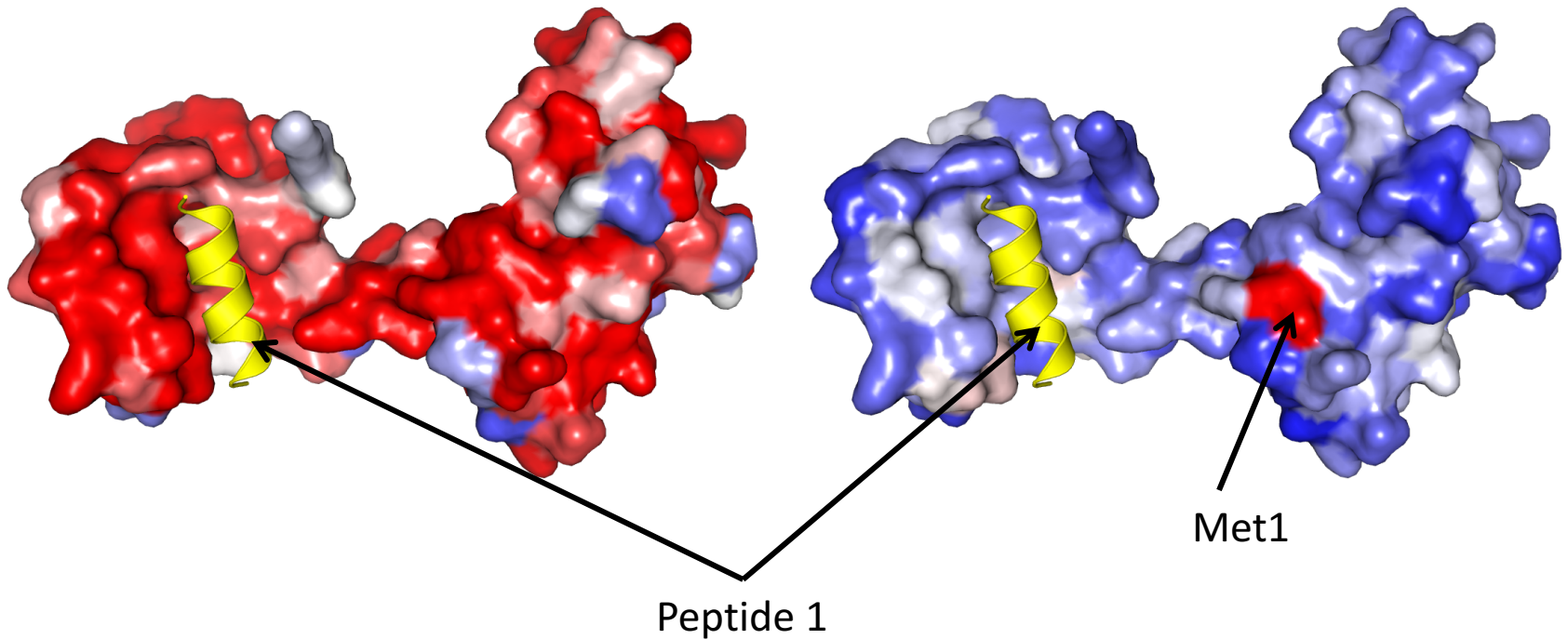
Many WT mutations: Calcium binding residues seem to be redundant

Experimental Competitive Fitness Scores



Colored by **Minimum** Site EFScore

Colored by **Mean** Site EFScore

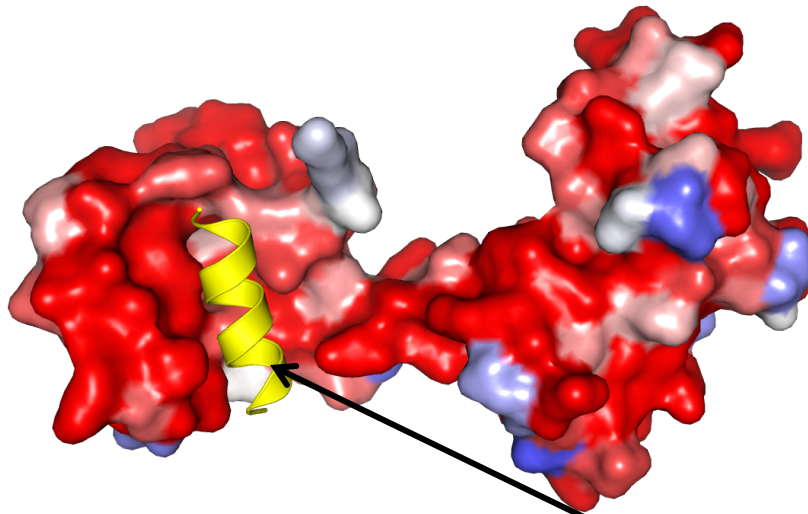


Experimental Competitive Fitness Scores

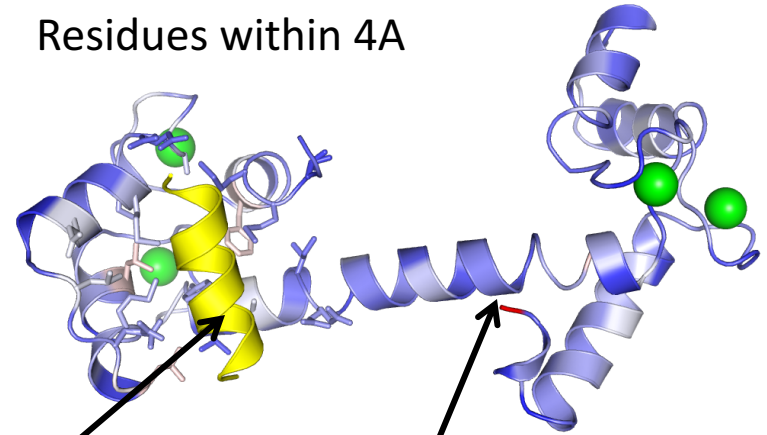


Colored by **Minimum** Site EFScore

Colored by **Mean** Site EFScore



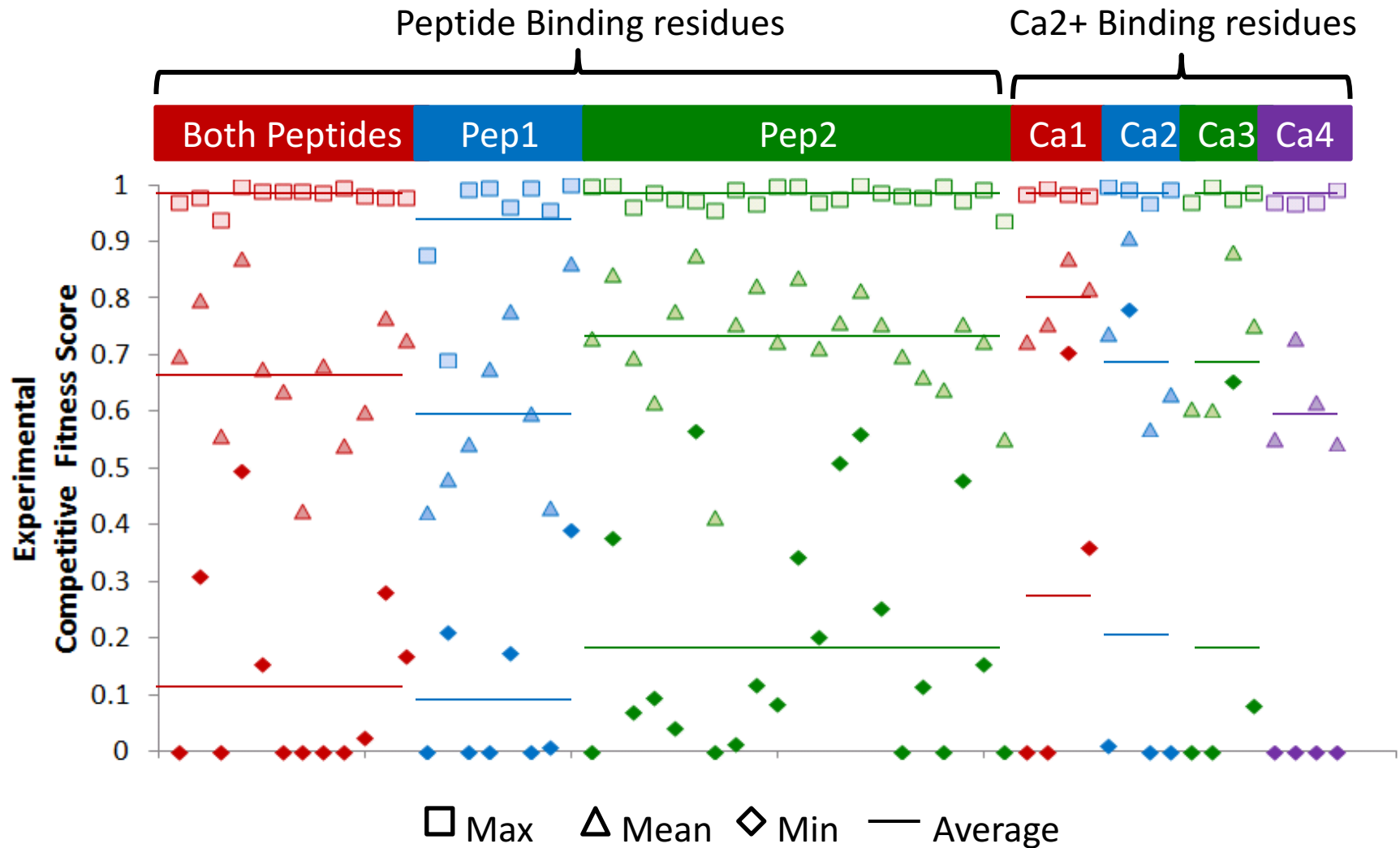
Residues within 4Å



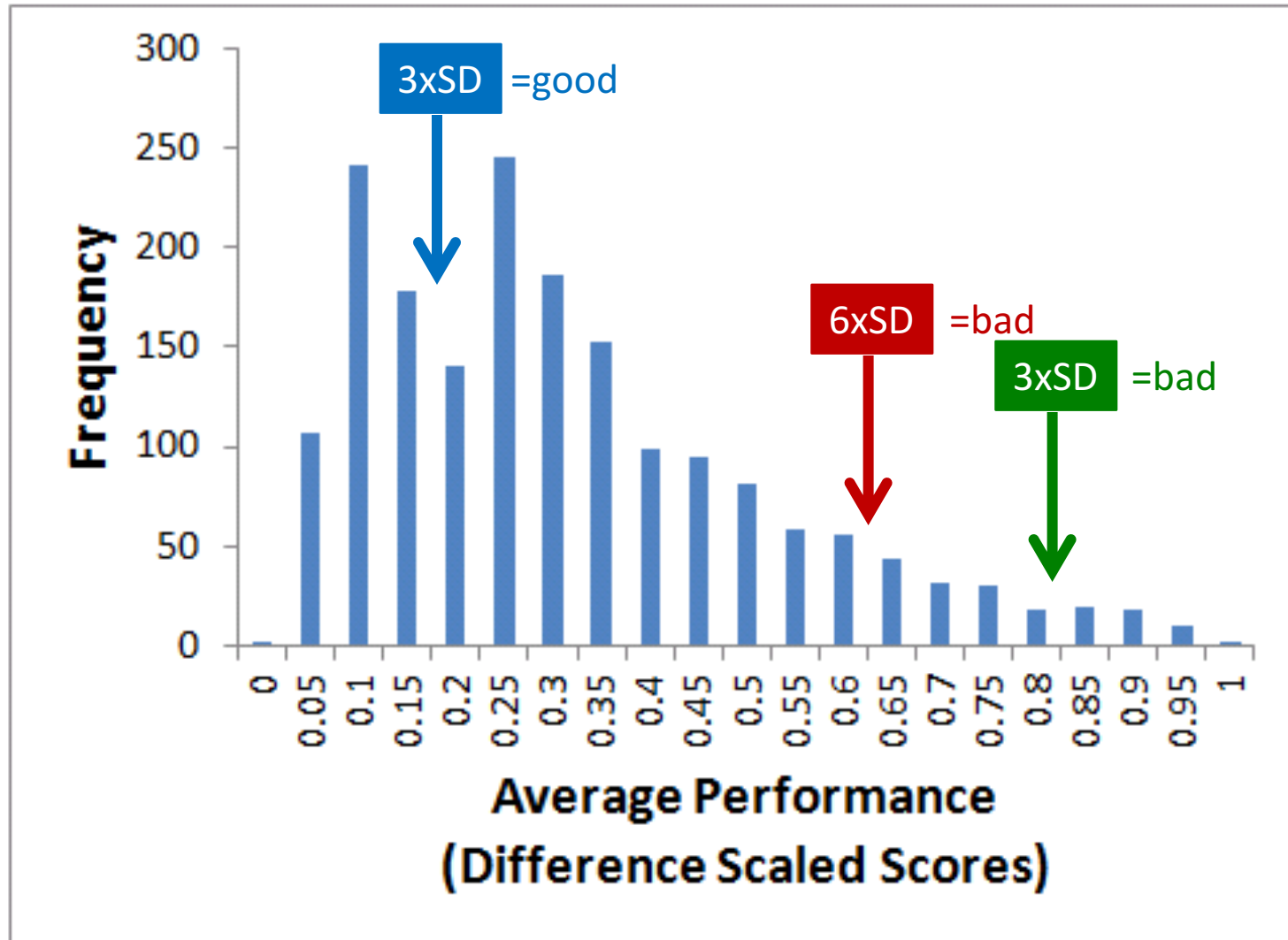
Met1

Peptide 1

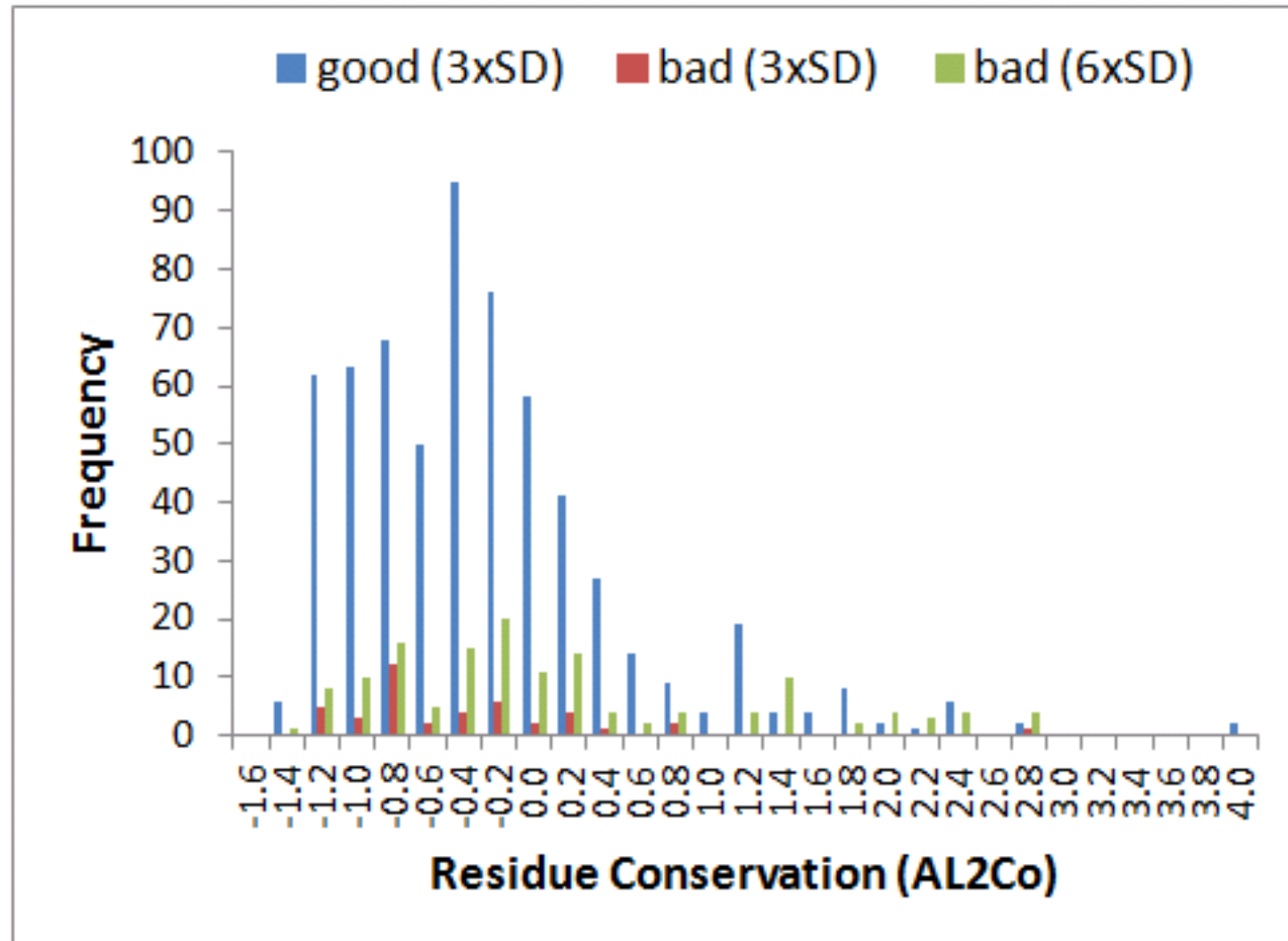
Experimental Competitive Fitness Scores



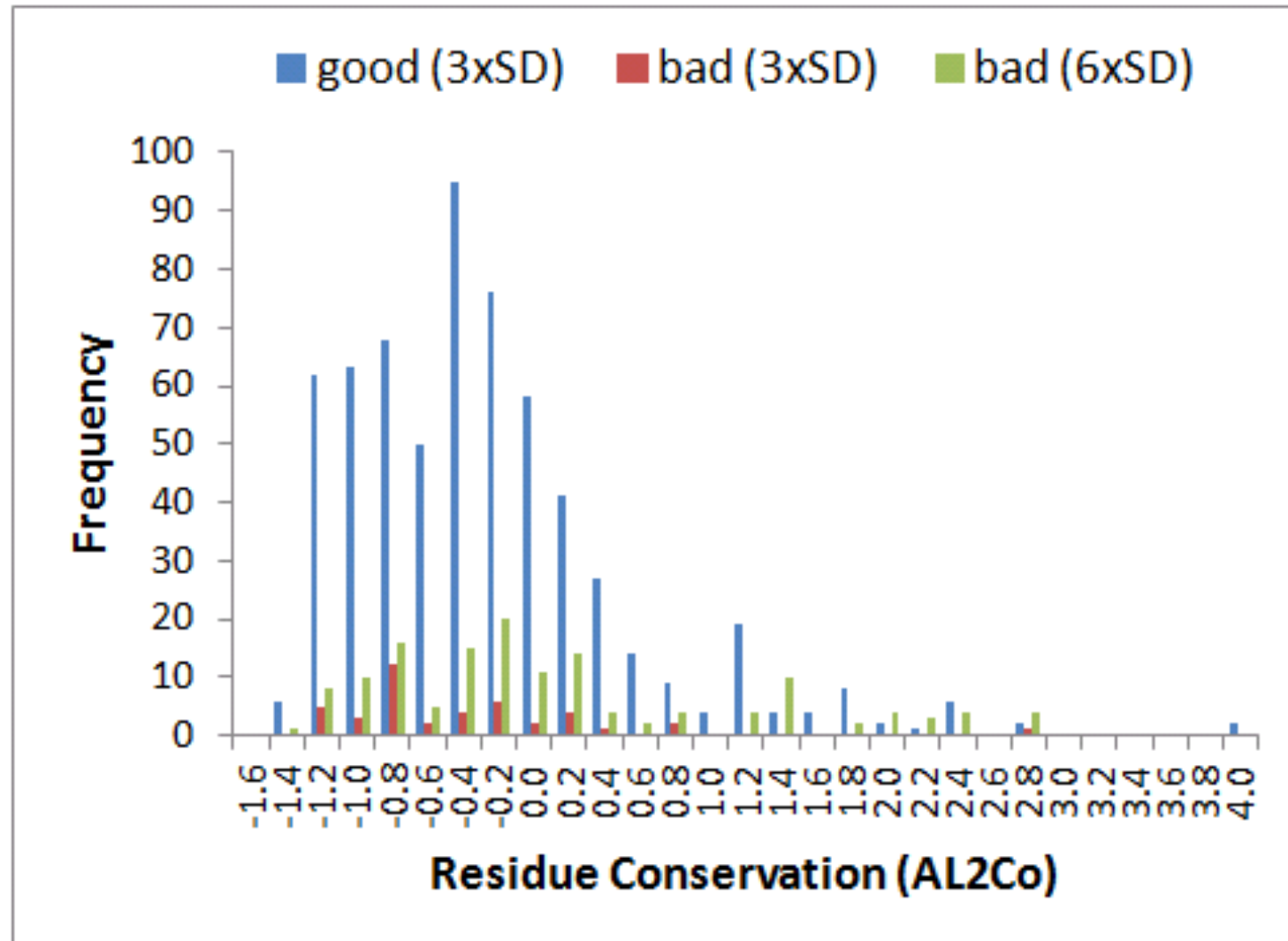
Average Predictor Performance



Average Predictor Performance

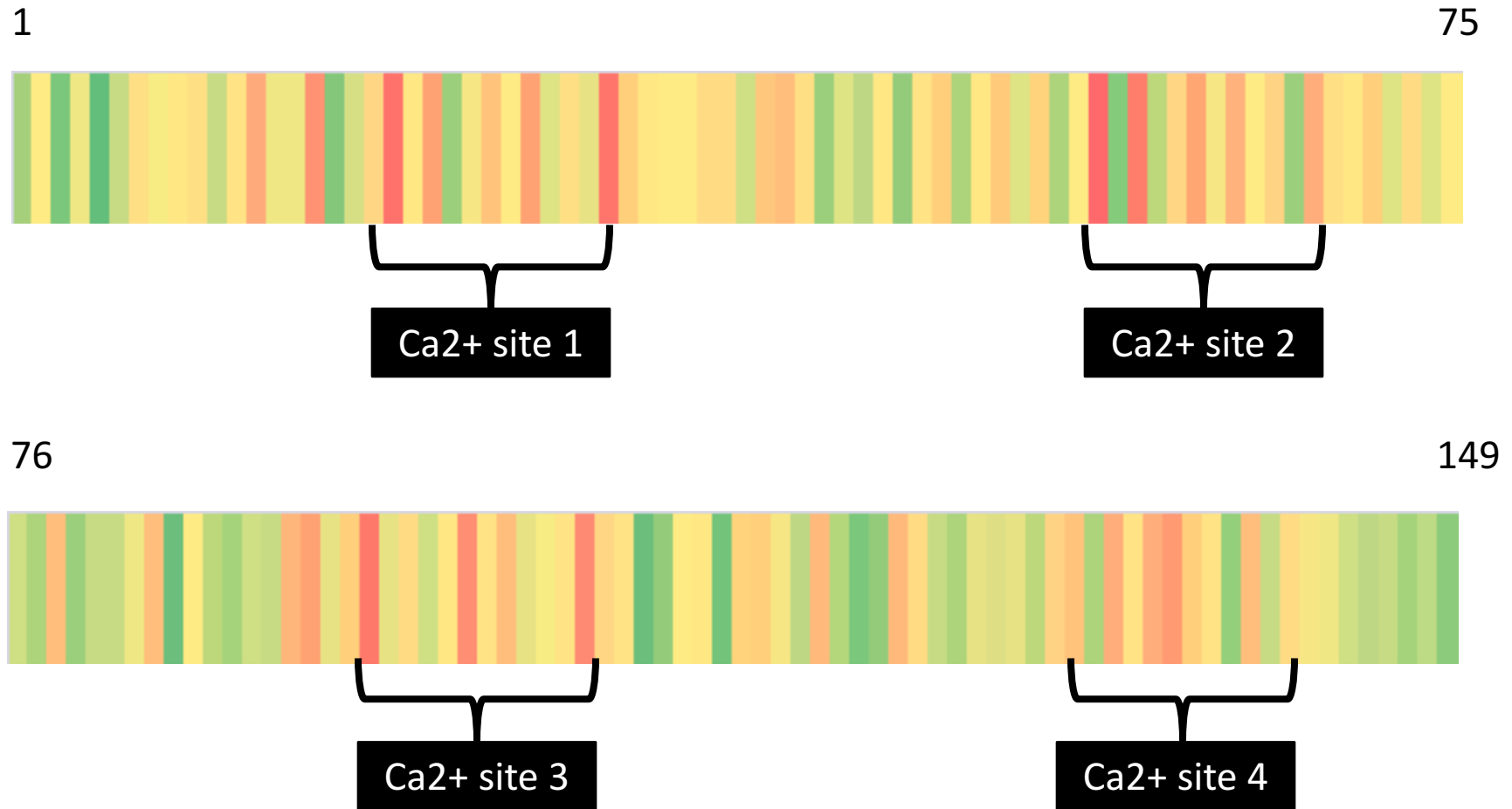


Average Predictor Performance



Predictor Performance: Residue Position

Colored by performance from **green** (good) to **red** (bad)



Predictor Performance: Residue Position

Colored by performance from **green** (good) to **red** (bad)

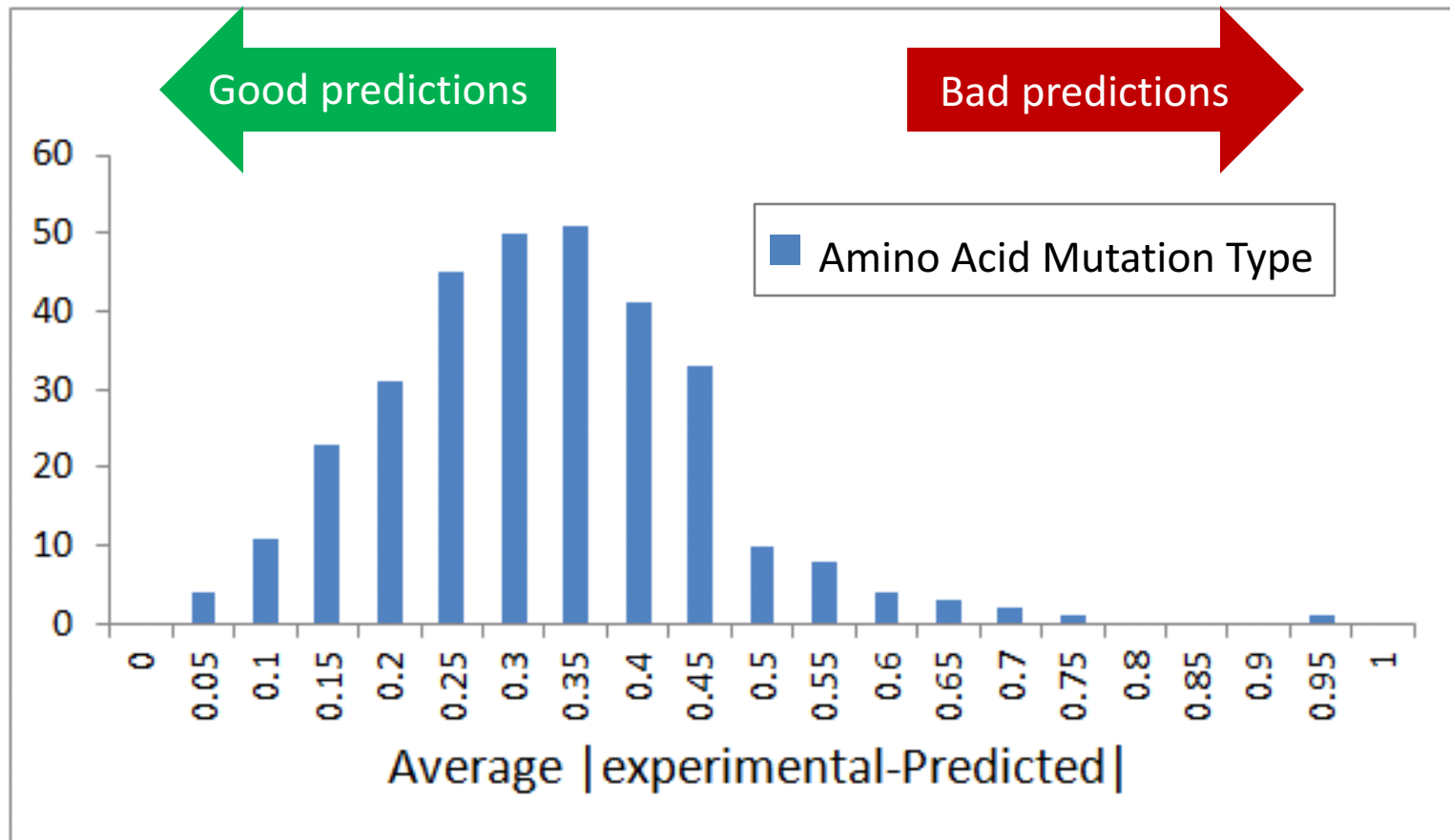


**good predictions are
detrimental**

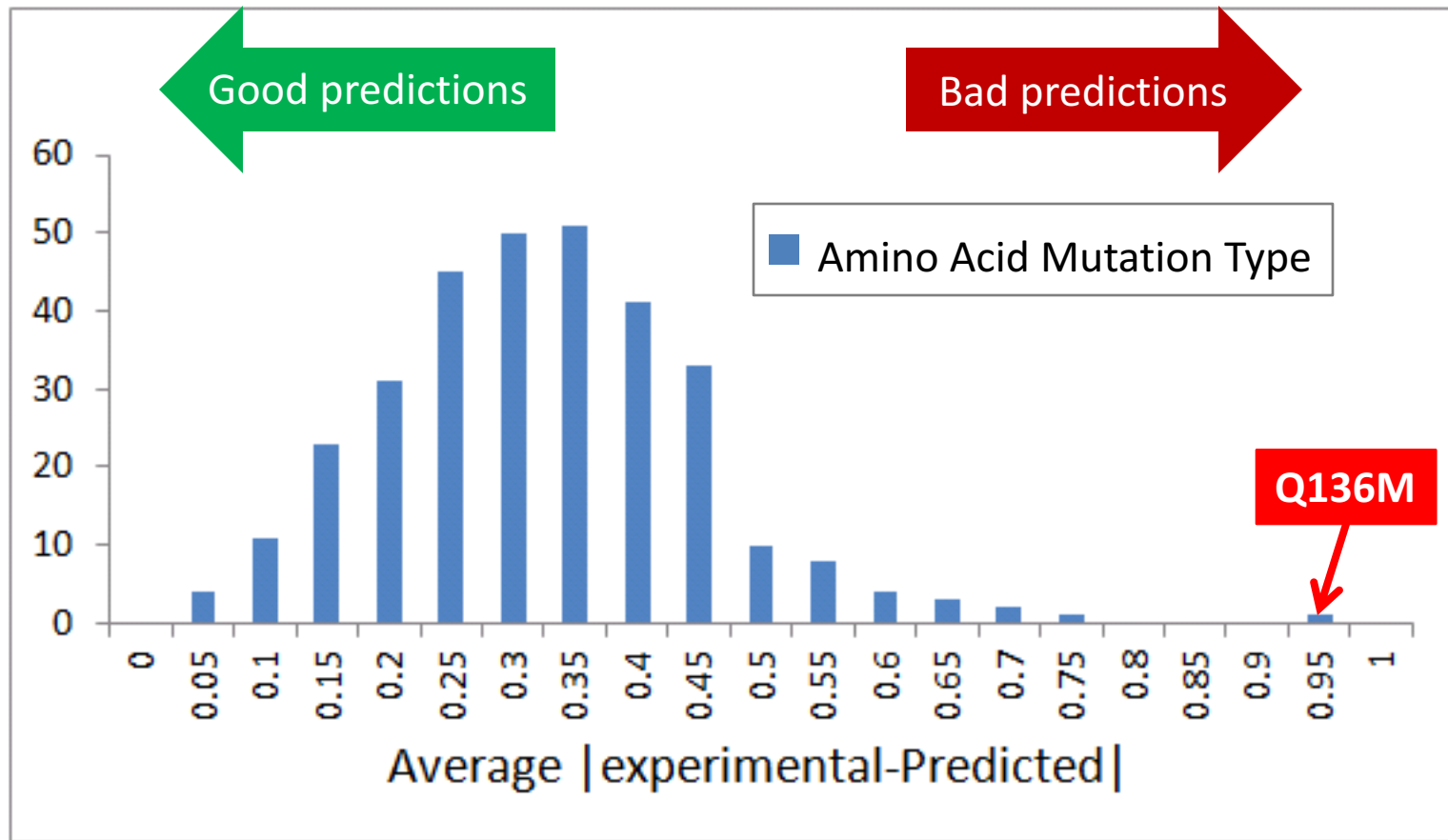
poor predictions are WT
**fitness depends on residue
type**

Mutation	Exp.Score	Performance
D21T	-0.19	0.318517111
D21I	0.109	0.229874254
D21K	0.182	0.142874254
D21A	0.382	0.154697175
D21H	0.694	0.32891146
D21P	0.77	0.456697175
D21Q	0.792	0.475282889
D21C	0.836	0.538482889
D21F	0.859	0.66866146
D21R	0.86	0.64591146
D21N	0.893	0.445459079
D21S	0.913	0.554554317
D21E	0.919	0.442926766
D21Y	0.95	0.69791146
D21W	0.959	0.725947175
D21L	0.966	0.743673365
D21V	0.973	0.770054317
D21G	0.983	0.686863841

Predictor Performance: Residue Type



Predictor Performance: Residue Type



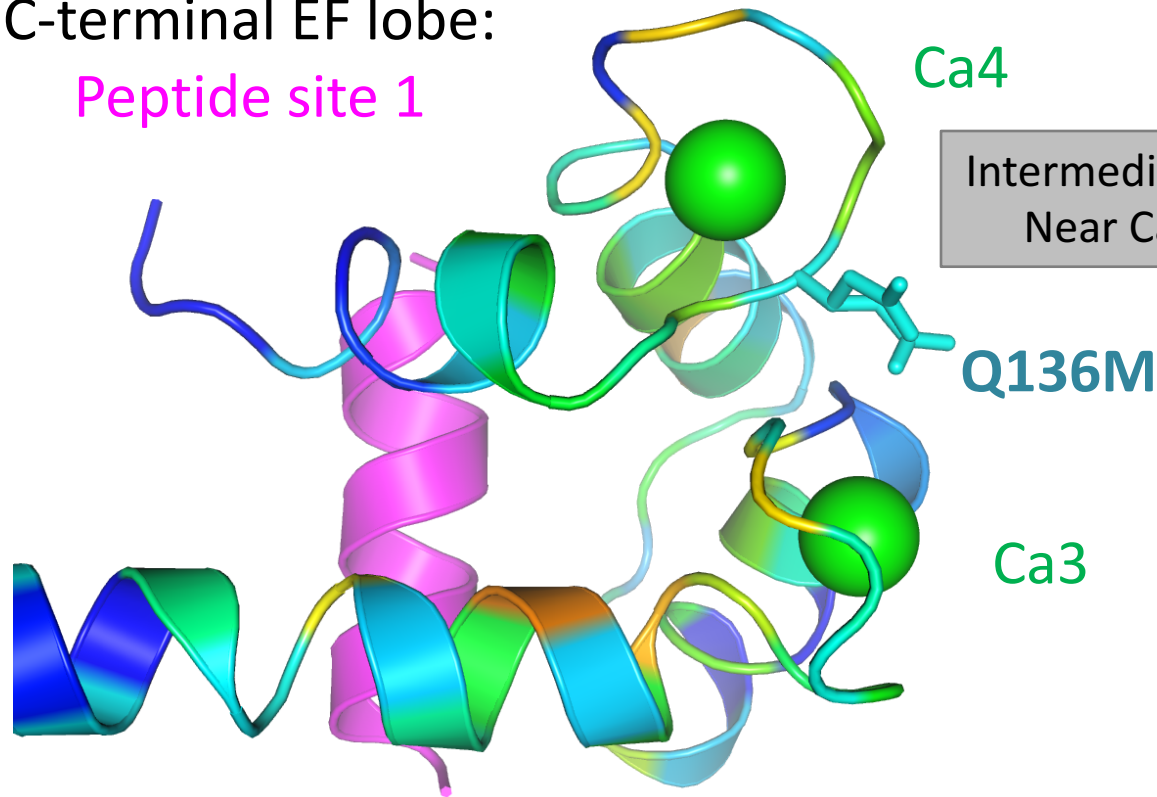
Q136M: Single residue change predicted as WT
experimental value = -0.226

Predictor Performance

CALM1 colored rainbow: blue (variable) to red (conserved)

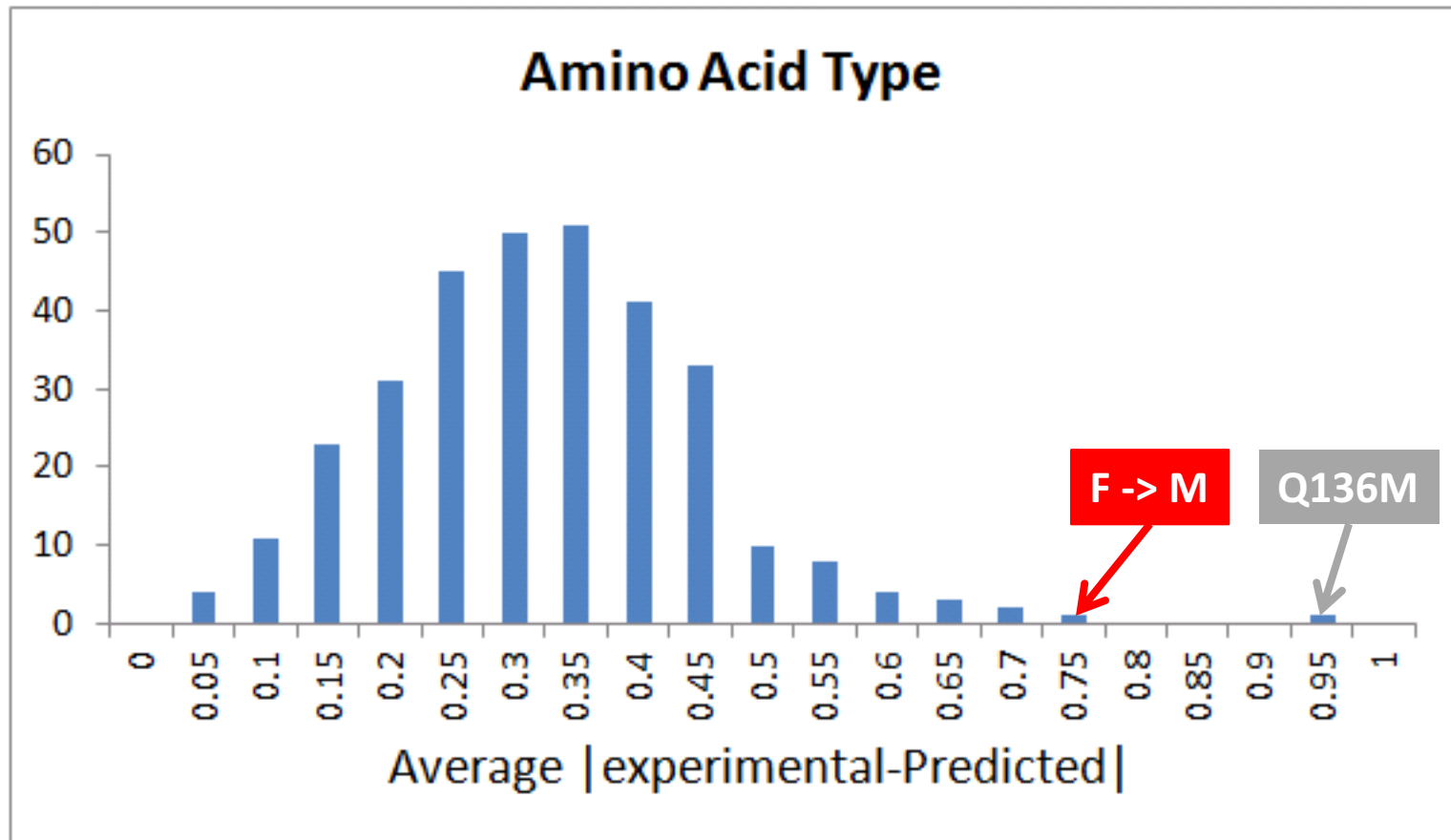
C-terminal EF lobe:

Peptide site 1



Q136M: Single residue change predicted as WT
experimental value = 0

Predictor Performance

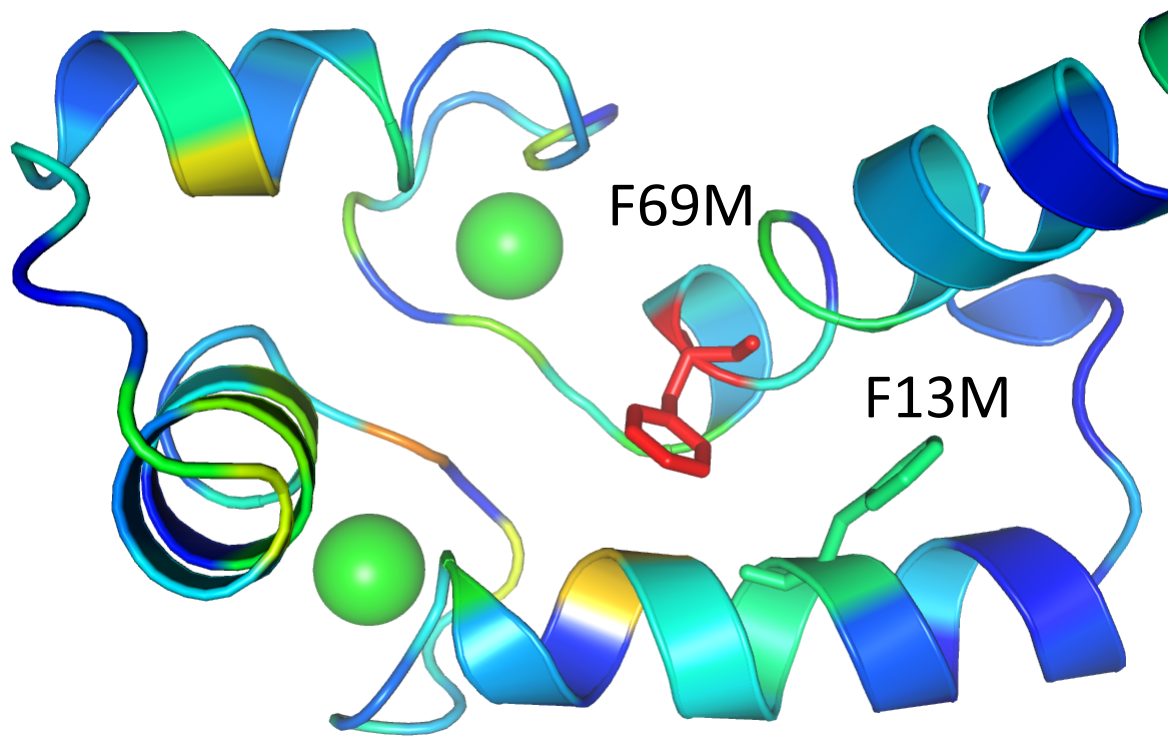


F13M and F69M predicted WT
experimental value = 0

Predictor Performance

CALM1 colored rainbow: blue (variable) to red (conserved)

PDB:4djc; N-lobe *no peptide* in site2



F13M and F69M predicted WT
experimental value = 0

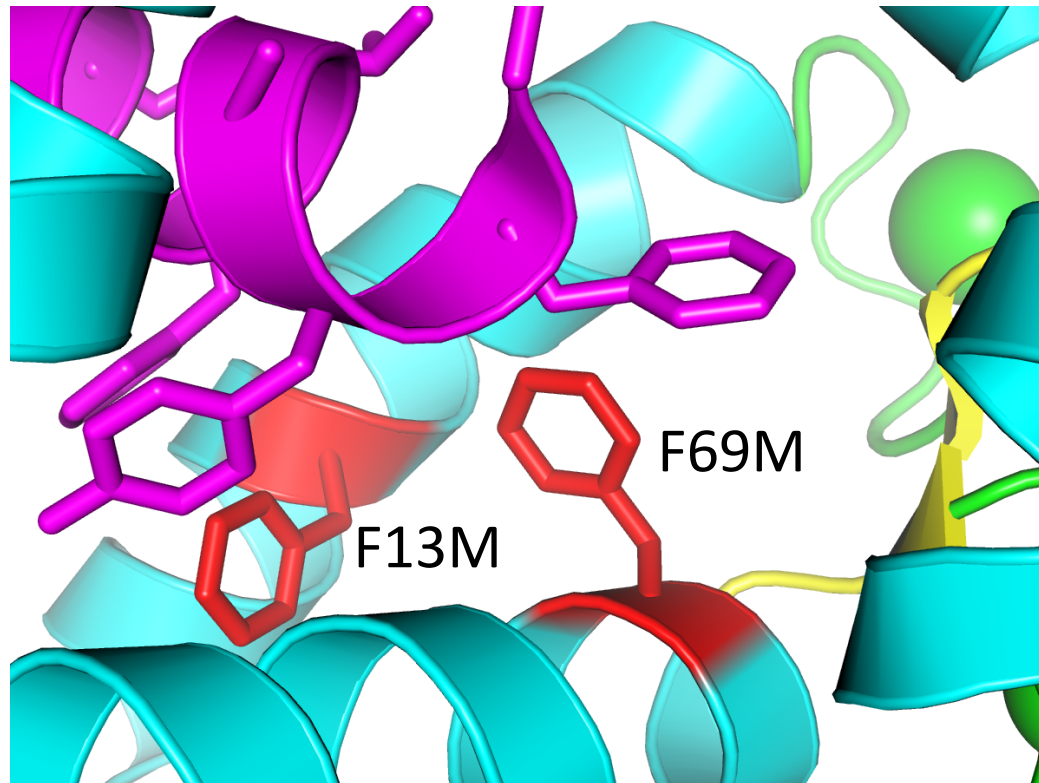
F69M: conserved
residue in peptide
binding site!

Group	pred	scaled
1-1	0.57	0.57
1-2	0.57	0.57
2-1	0.23	0.31
2-2	0.32	0.51
3-1	0.49	0.10
4-1	0.90	0.91
4-2	0.90	0.99

Scores reflect amino
acid exchange values?

Predictor Performance

PDB:2f3y; *alternate conformation* with peptide2



F13M and F69M predicted WT
experimental value = 0

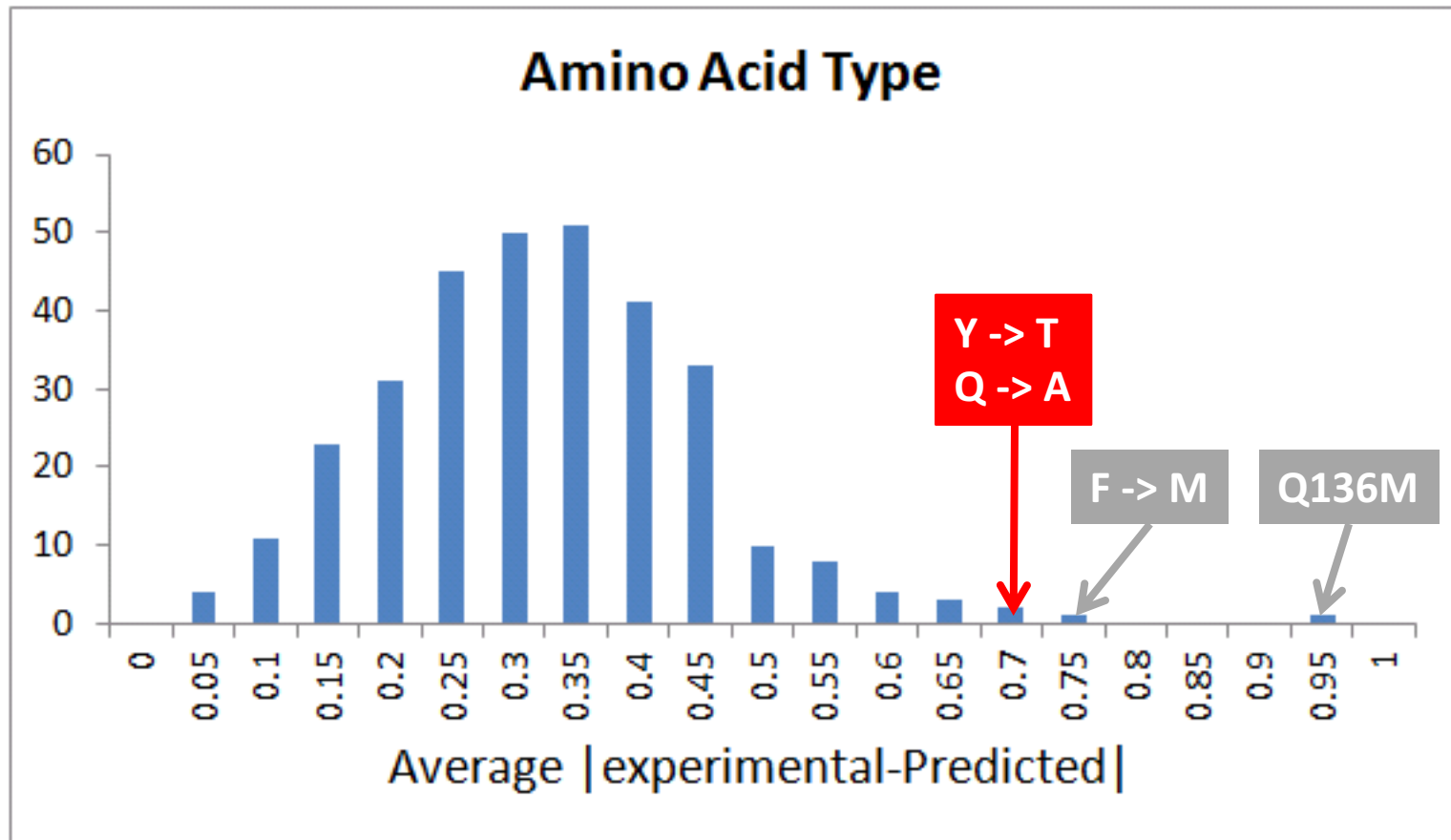
F69M: conserved
residue in peptide
binding site!

Group	pred	scaled
1-1	0.57	0.57
1-2	0.57	0.57
2-1	0.23	0.31
2-2	0.32	0.51
3-1	0.49	0.10
4-1	0.90	0.91
4-2	0.90	0.99

Scores reflect amino
acid exchange values?

**Should consider
peptide binding**

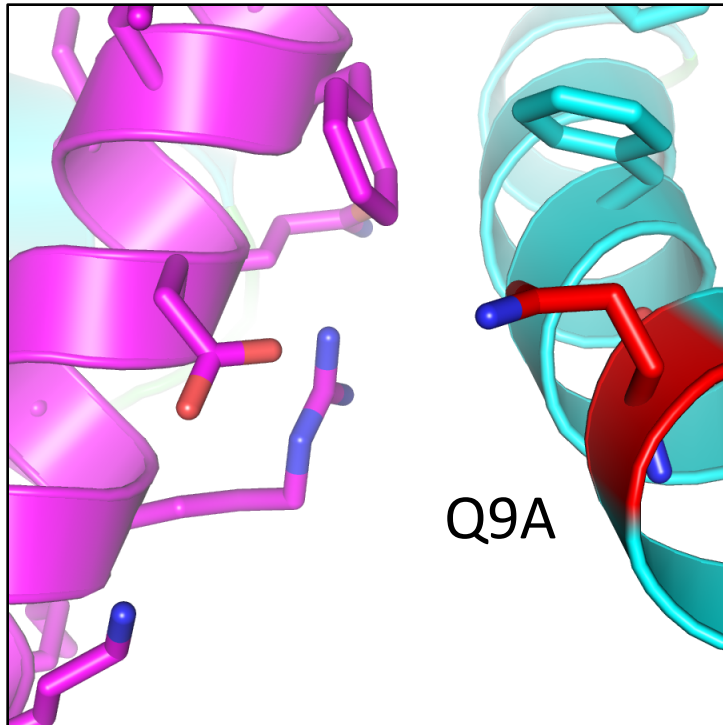
Predictor Performance



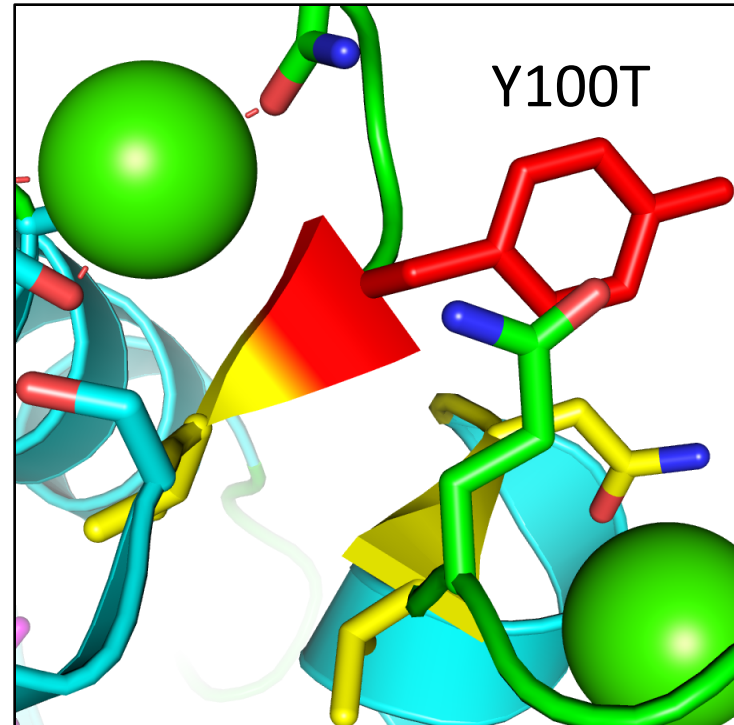
Y100T and Q9A predicted near WT
experimental value = 0

Predictor Performance

Near Peptide2 Binding



Near Ca²⁺ Binding



Y100T and Q9A predicted near WT
experimental value = 0

Predictor Performance

CALM1 Prediction Trends

- Conservations are less meaningful – few invariant positions
- Data coverage is high: need to fine tune amino acid mutation types
- Generally poor predictions of functional sites

Predictor Performance

CALM1 Prediction Difficulties

- Alternate conformations to consider
(over 100 available structures)
- Apparent redundancy in Ca^{2+} binding sites
(binding sites not invariant)
- Variable peptide interactions
(2 sites, conformations, dimerization, etc.)
- Assembly into larger complexes
(5k7l Ca^{2+} channel tetramer)

Prediction Score Distributions

