SBV Improver Sub-Challenge 3:
Inter-Species Pathway Perturbation Prediction

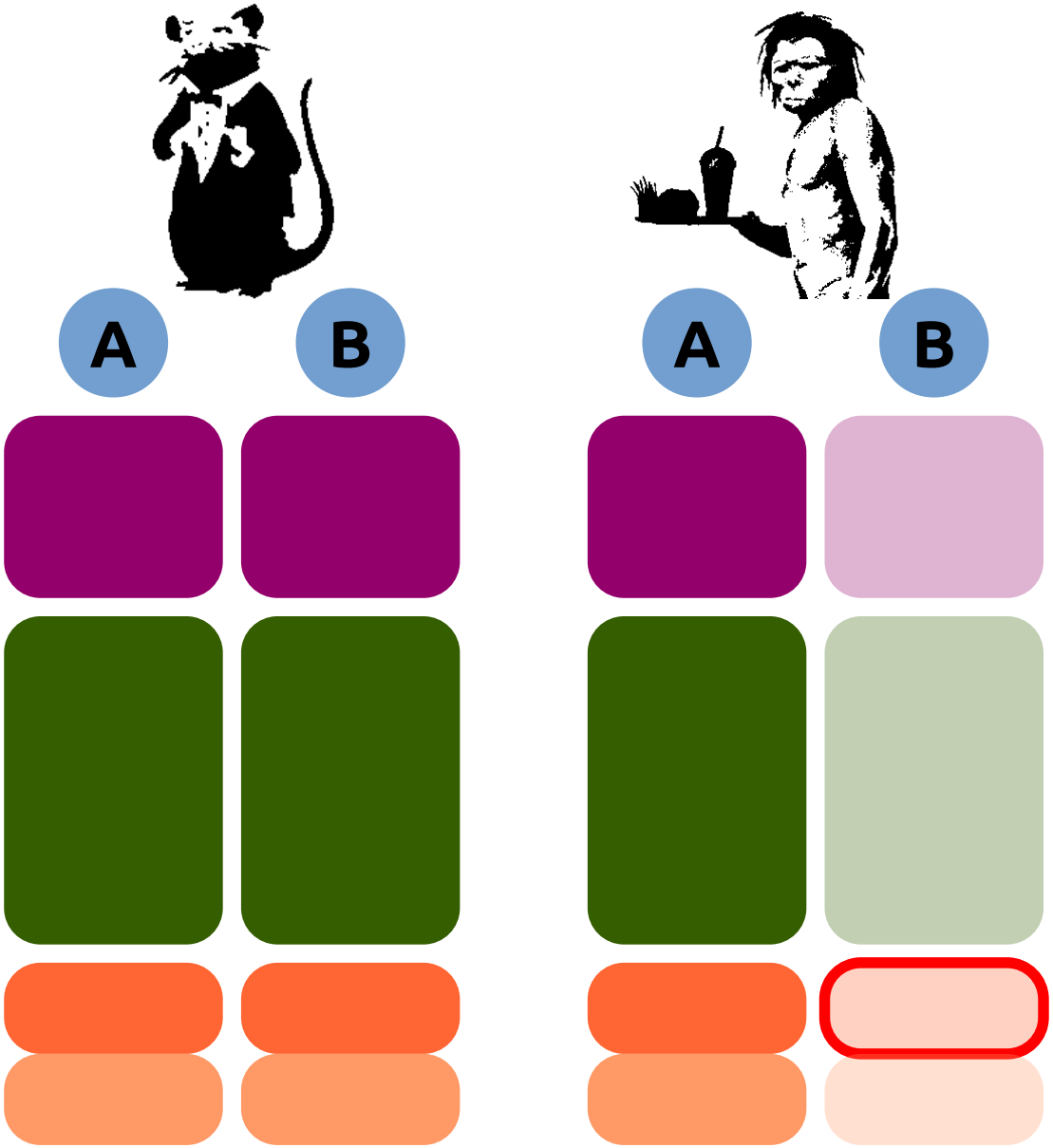
Team Edith's method

Christoph Hafemeister
sbv IMPROVER Symposium 2013

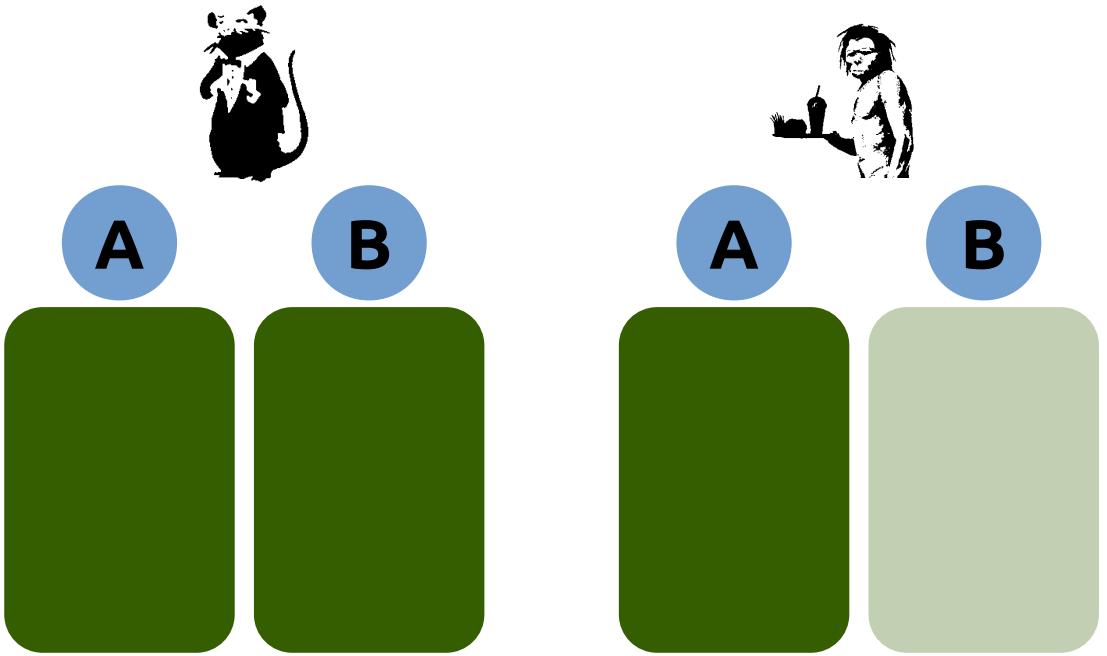
Sub-Challenge 3 Ranking Table

Team	AUPR	Pearson	BAC	Rank
50	0.19	0.59	0.54	1
49	0.12	0.53	0.53	2
133	0.12	0.54	0.54	2
52	0.1	0.52	0.54	4
131	0.11	0.5	0.52	5
105	0.11	0.52	0.51	6
111	0.06	0.41	0.43	7

Challenge Description



Challenge Description

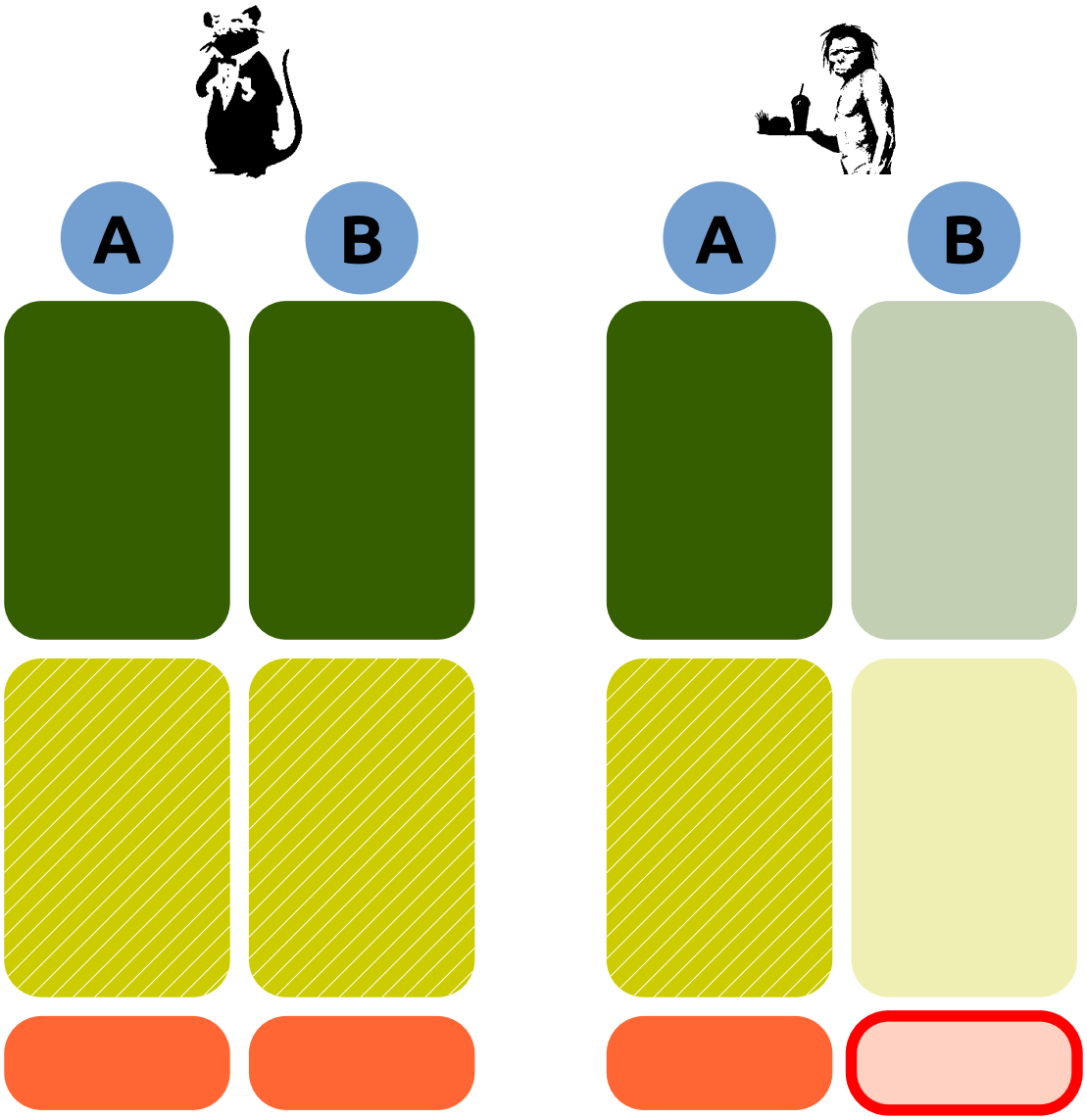


Stimulus Subset

Gene Expression

Gene Set Enrichment FDR

Challenge Description



Stimulus Subset

Gene Expression

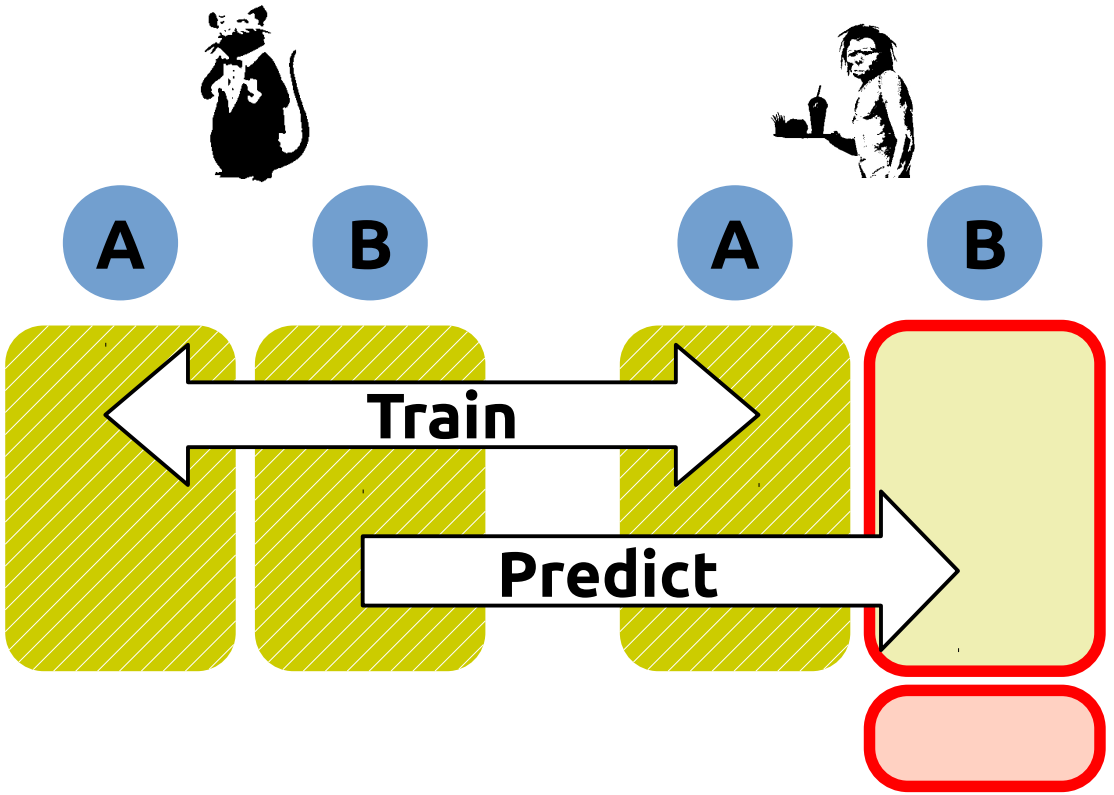
LIMMA

Differential Expression
t-value

GSEA

Gene Set Enrichment FDR

Method Overview



Stimulus Subset

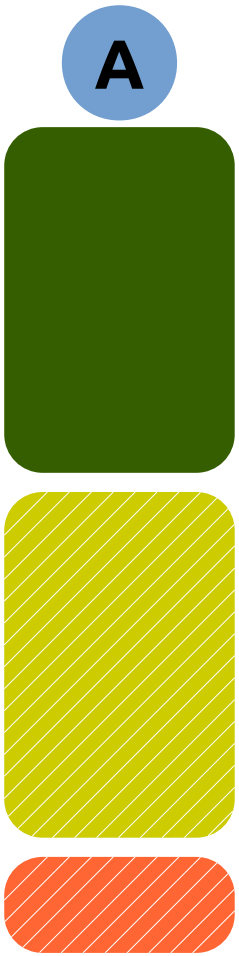
Differential Expression
t-value

GSEA

Gene Set Enrichment FDR

- Operate in t-value space
- Train model using training set
- Predict human test set using rat test set
- Compute human GSE FDR scores

Test LIMMA/GSEA pipeline



Gene Expression

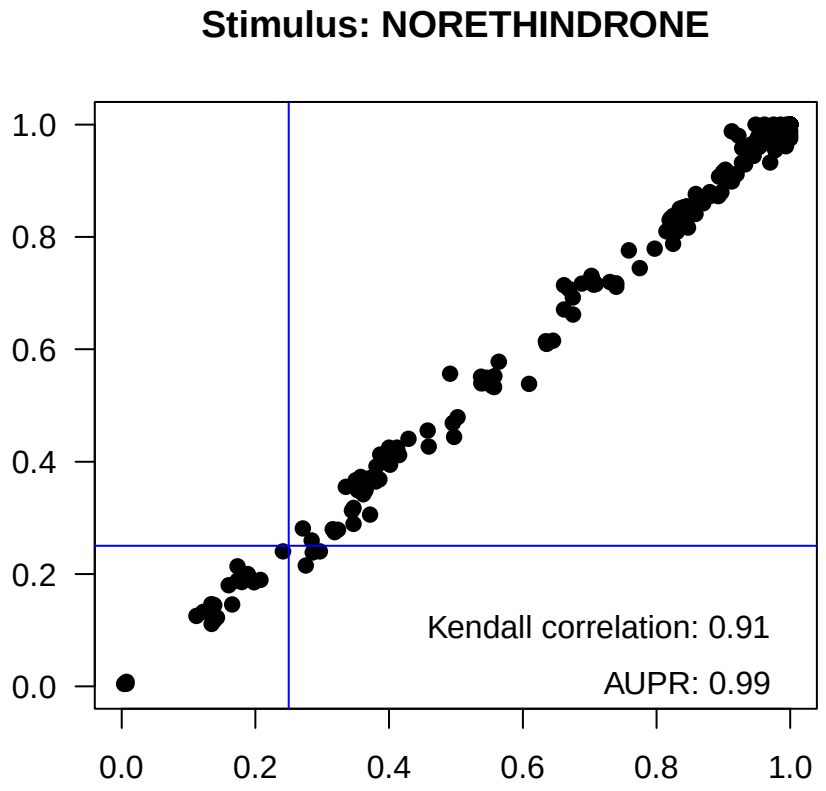
LIMMA

Differential Expression
t-value

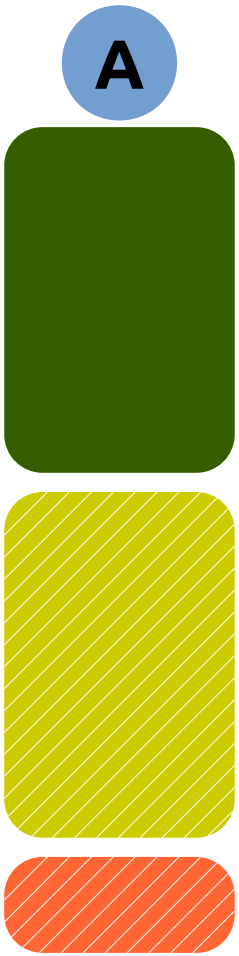
GSEA

Gene Set Enrichment FDR

GSE FDR: computed vs provided



Test LIMMA/GSEA pipeline



Gene Expression

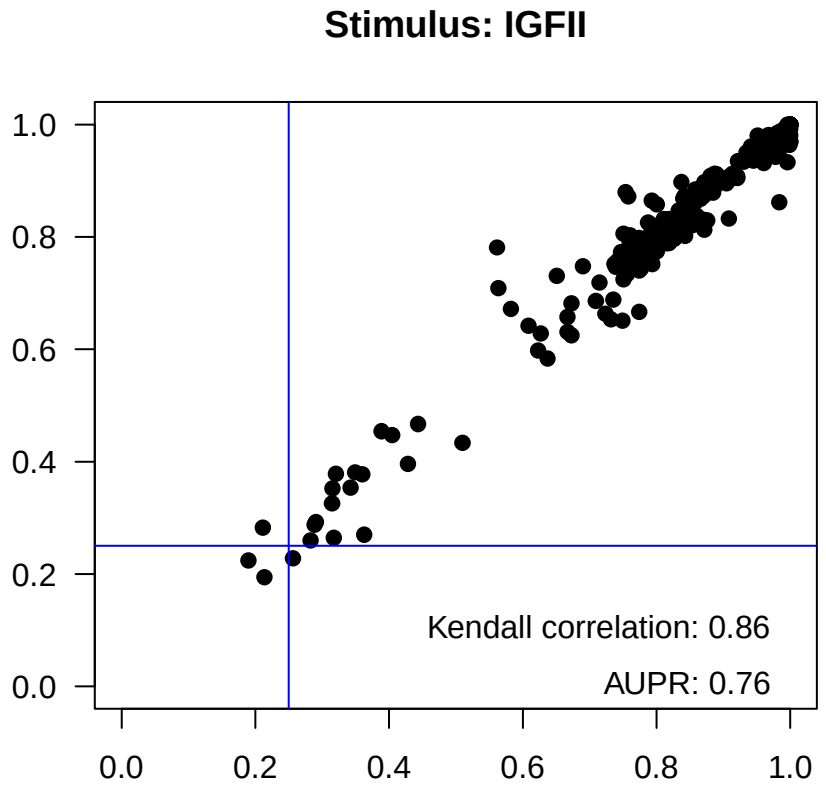
LIMMA

Differential Expression
t-value

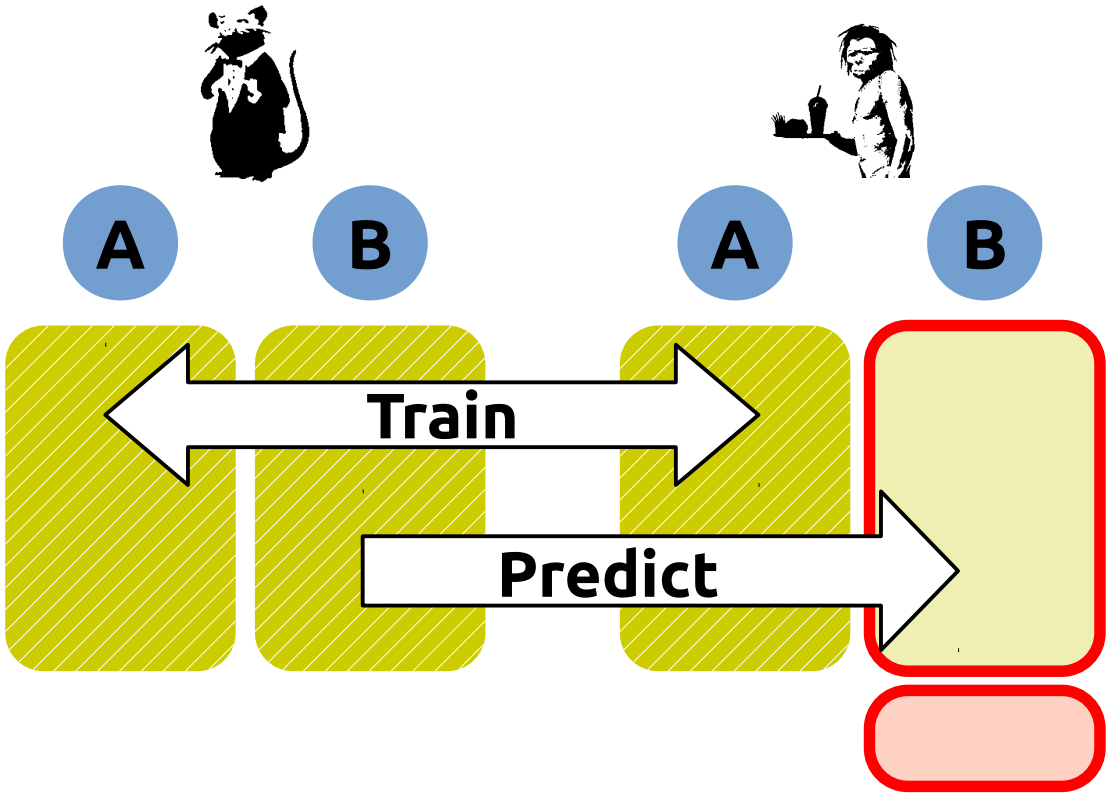
GSEA

Gene Set Enrichment FDR

GSE FDR: computed vs provided



The Model



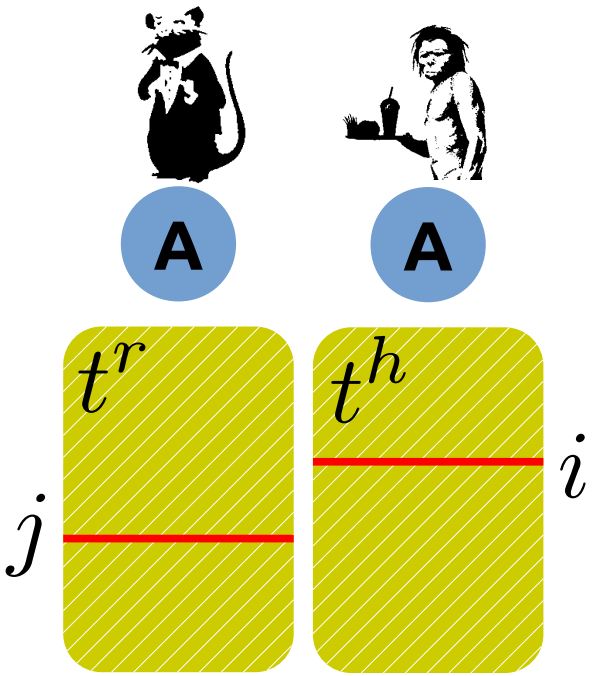
Stimulus Subset

Differential Expression
t-value

GSEA

Gene Set Enrichment FDR

The Model



Model t-values of each human gene as a linear fit of some rat gene t-values

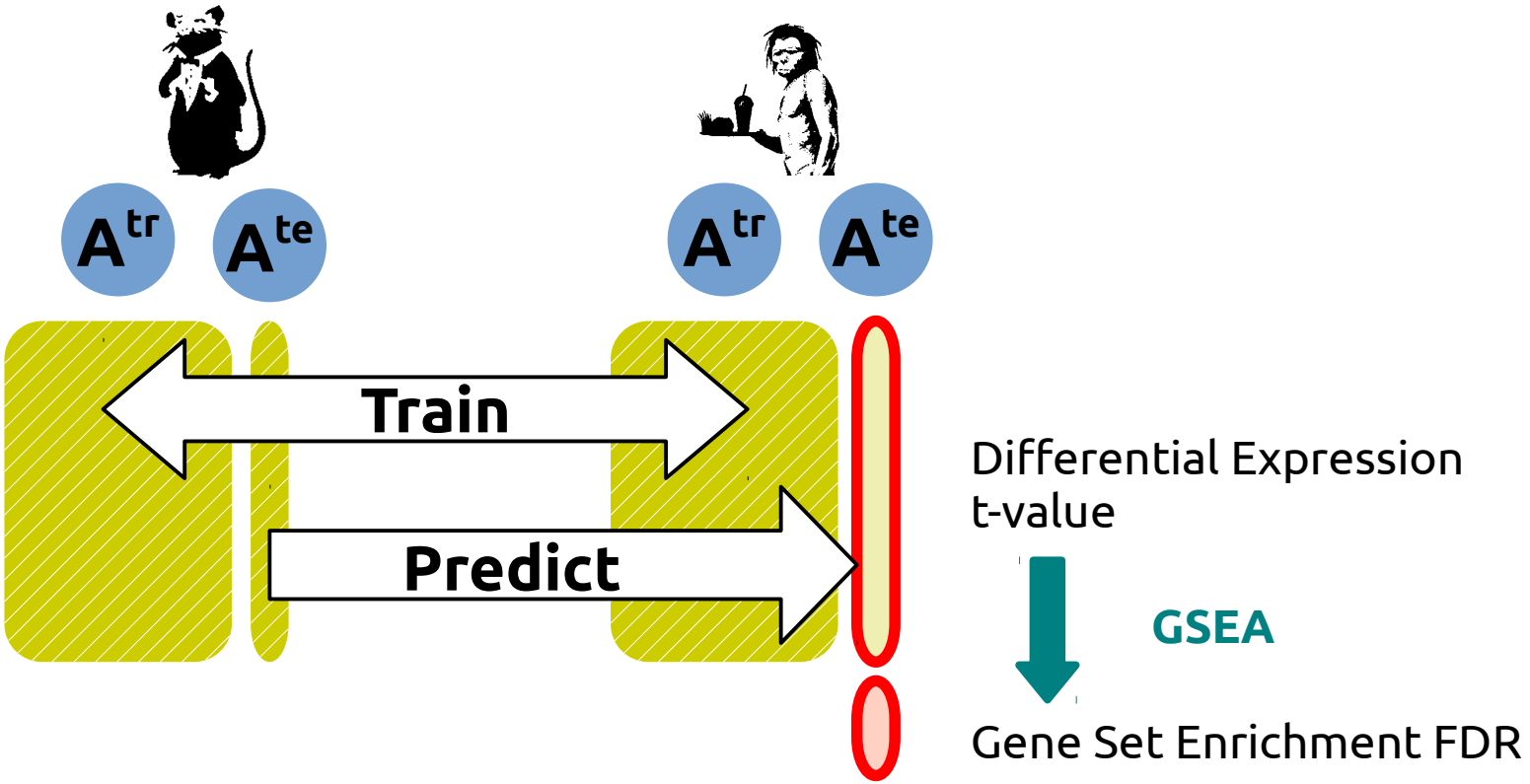
$$t_i^h = \beta t_j^r + \alpha$$

Minimize the sum of squares of the residuals of the linear regression model

$$\min_{j,\alpha,\beta} \sum_{s=1}^{n} (t_{is}^h - \beta t_{js}^r - \alpha)^2$$

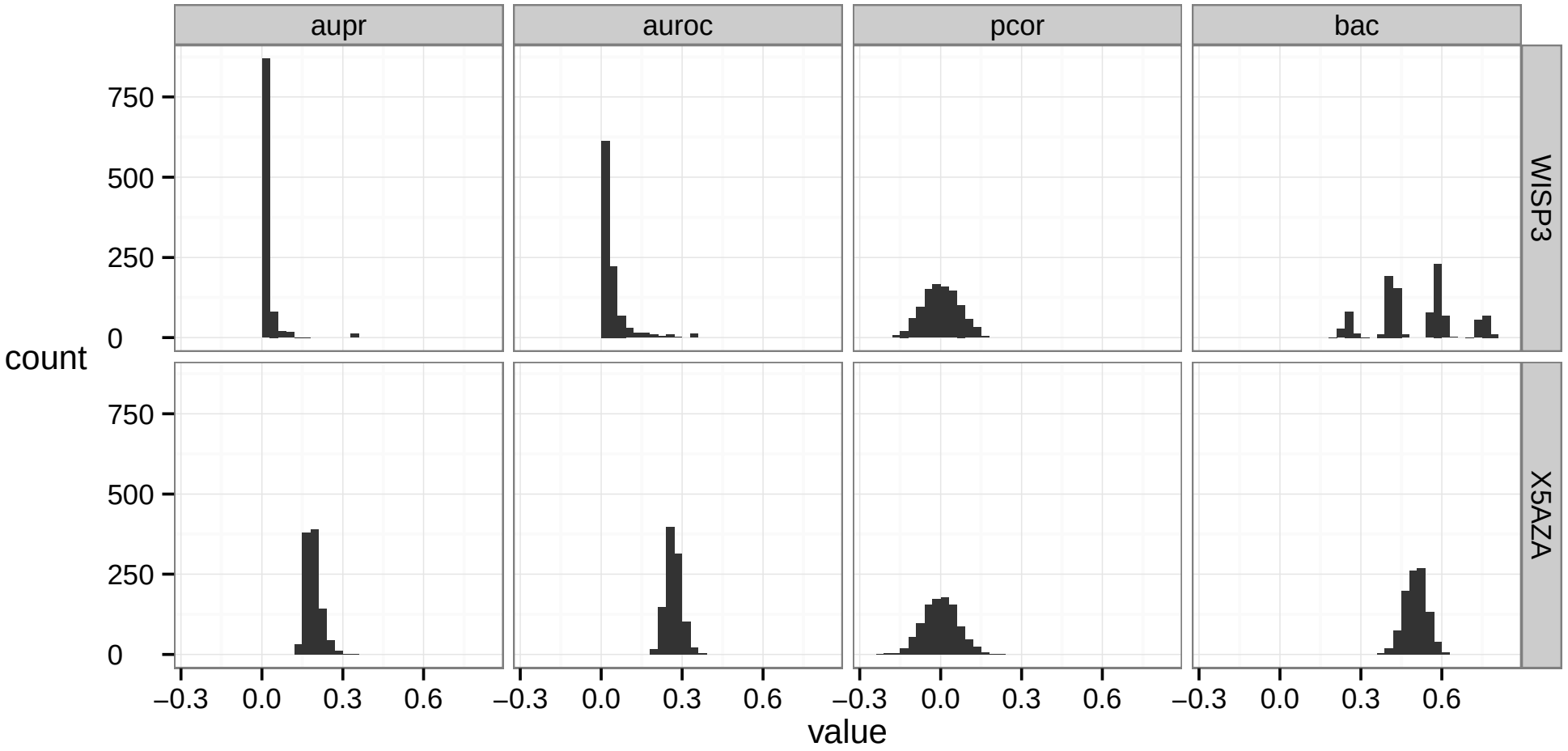
Model Validation

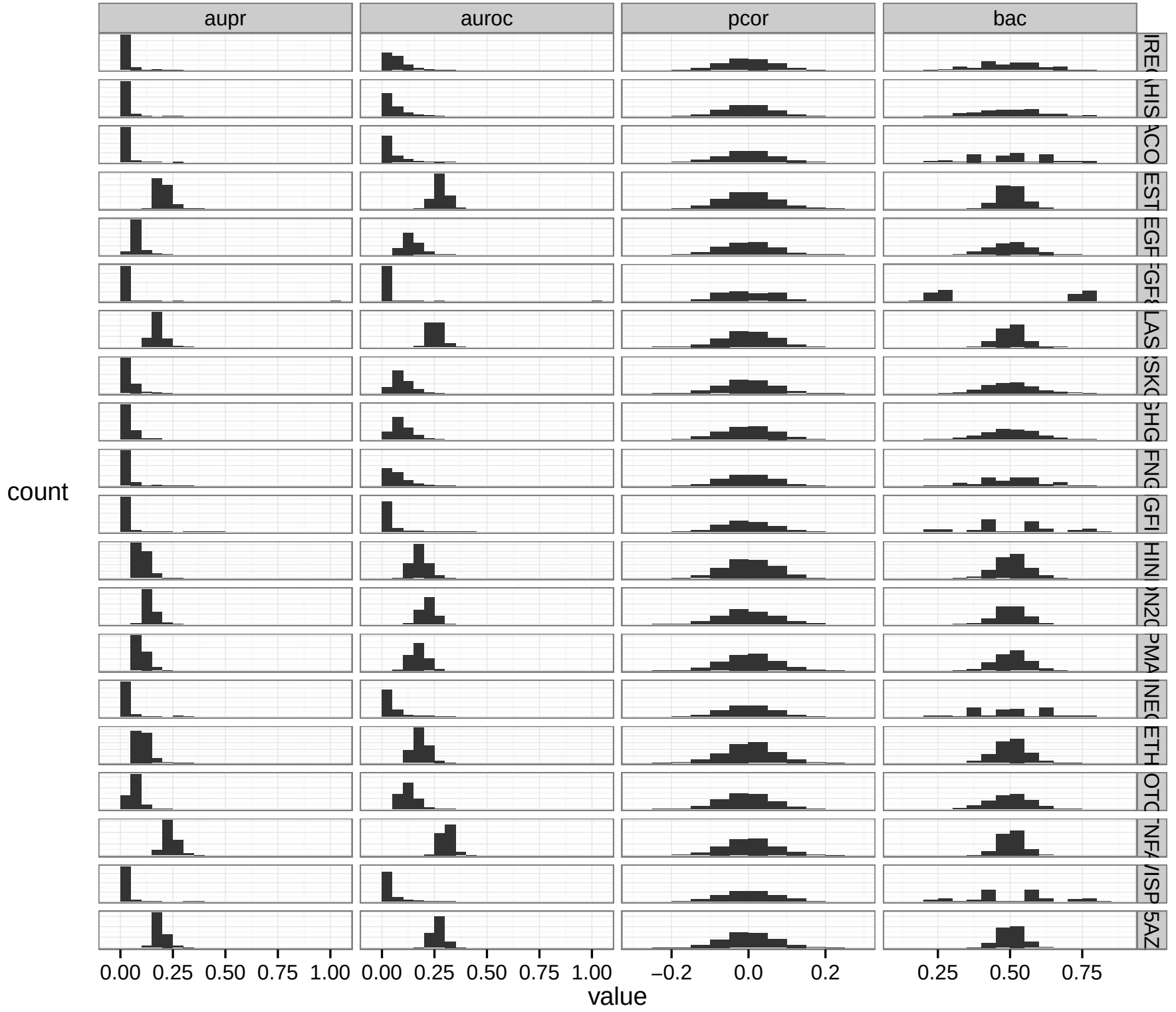
- Validate model using leave-one-out cross-validation
- Use every observation from training set as test set once

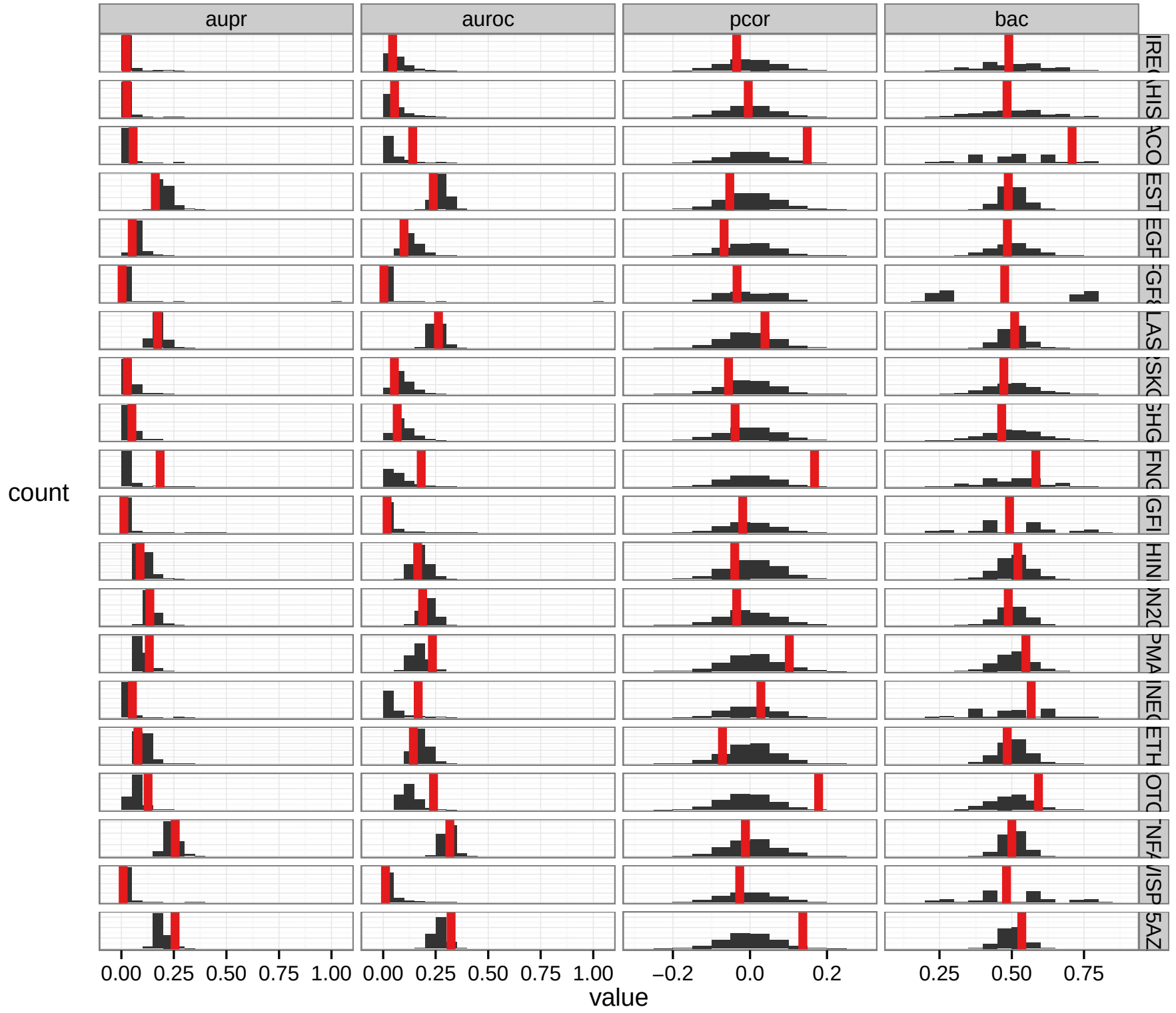


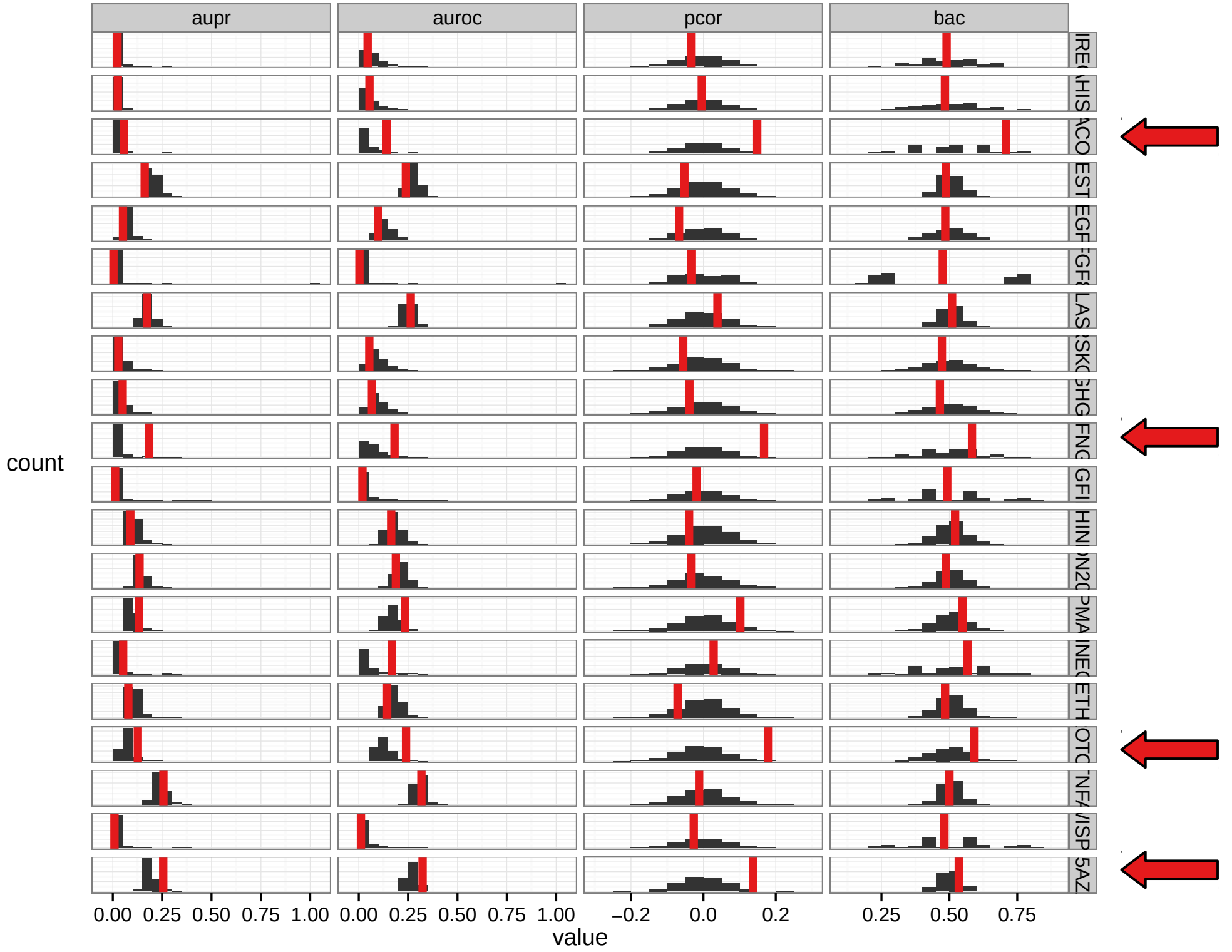
Baseline Results

- Establish a baseline for method development
- Generate 1000 random Gene Set Enrichment FDR matrices
- Use AUPR and AUROC as performance metric




aupr
auroc
pcor
bac
count
value
0.00
0.25
0.50
0.75
1.00
−0.2
0.0
0.2
0.25
0.50
0.75


aupr
auroc
pcor
bac
count
0.00
0.25
0.50
0.75
1.00
−0.2
0.0
0.2
value


aupr
auroc
pcor
bac
count
value
0.00
0.25
0.50
0.75
1.00
−0.2
0.0
0.2

Cross-Validation Results

Results are not so great

Only better than random for four stimuli

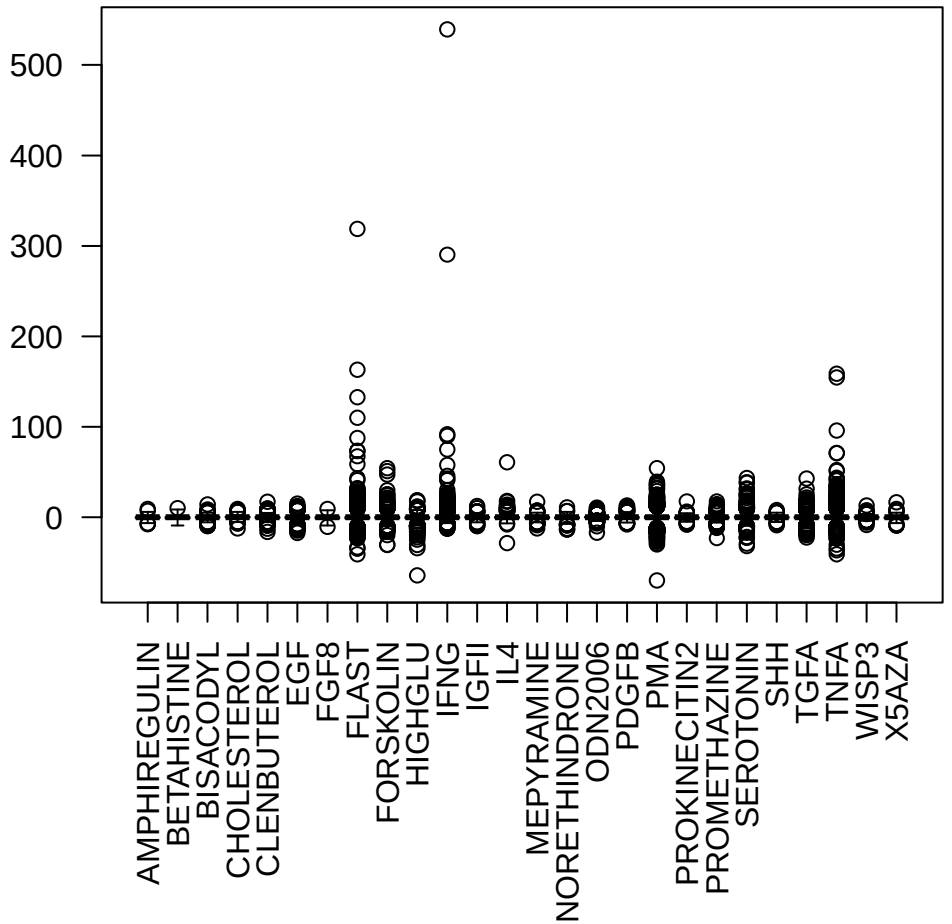
- BISACODYL
- IFNG
- SEROTONIN
- X5AZA

Mean performance

aupr	auroc	pcor	bac
0.095	0.148	0.016	0.518

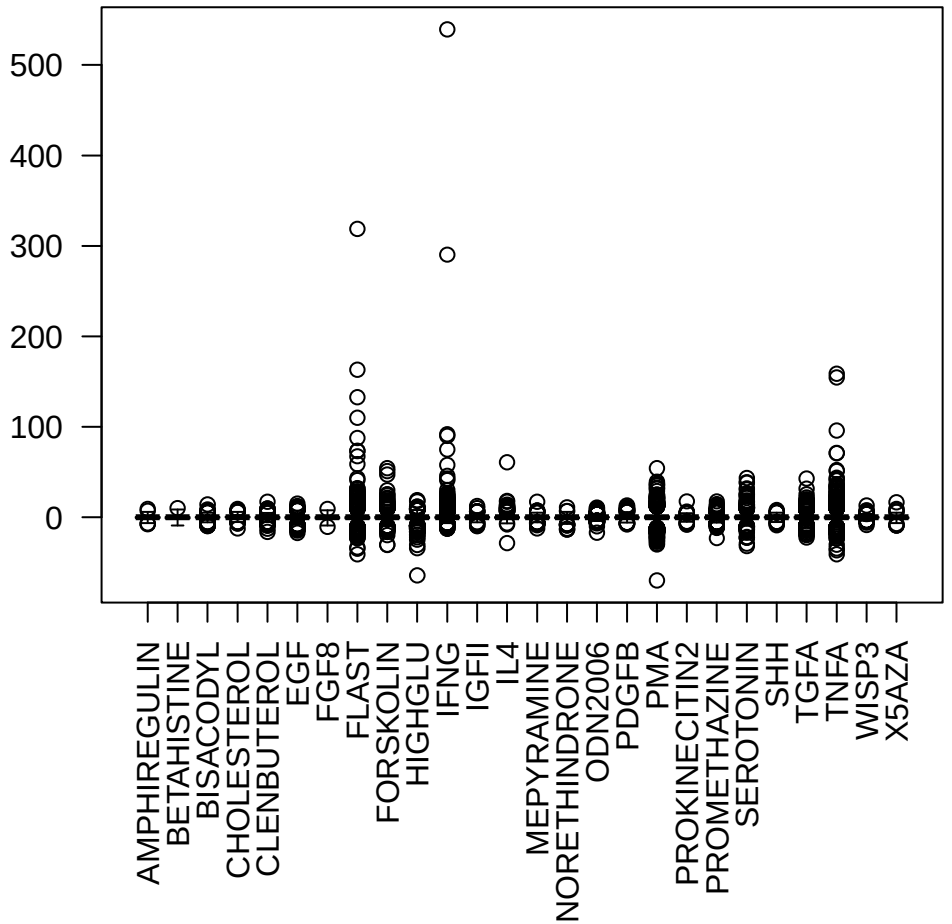
Standardize Rat T-values

Before



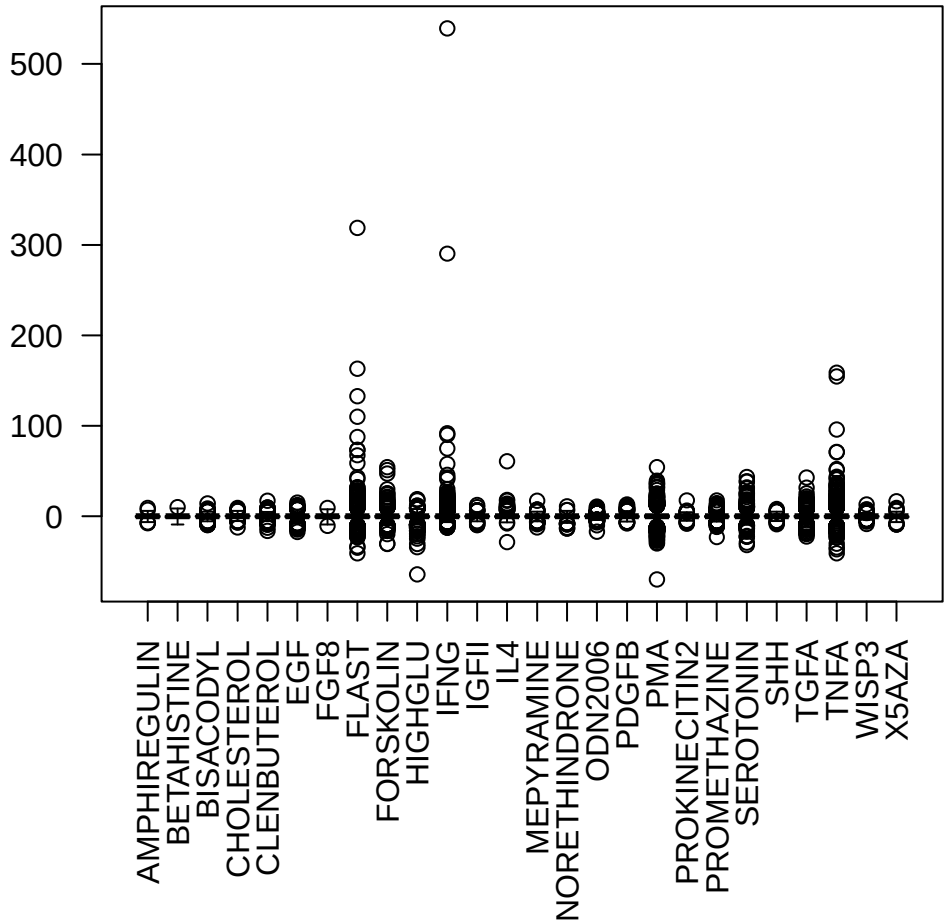
Standardize Rat T-values

Before

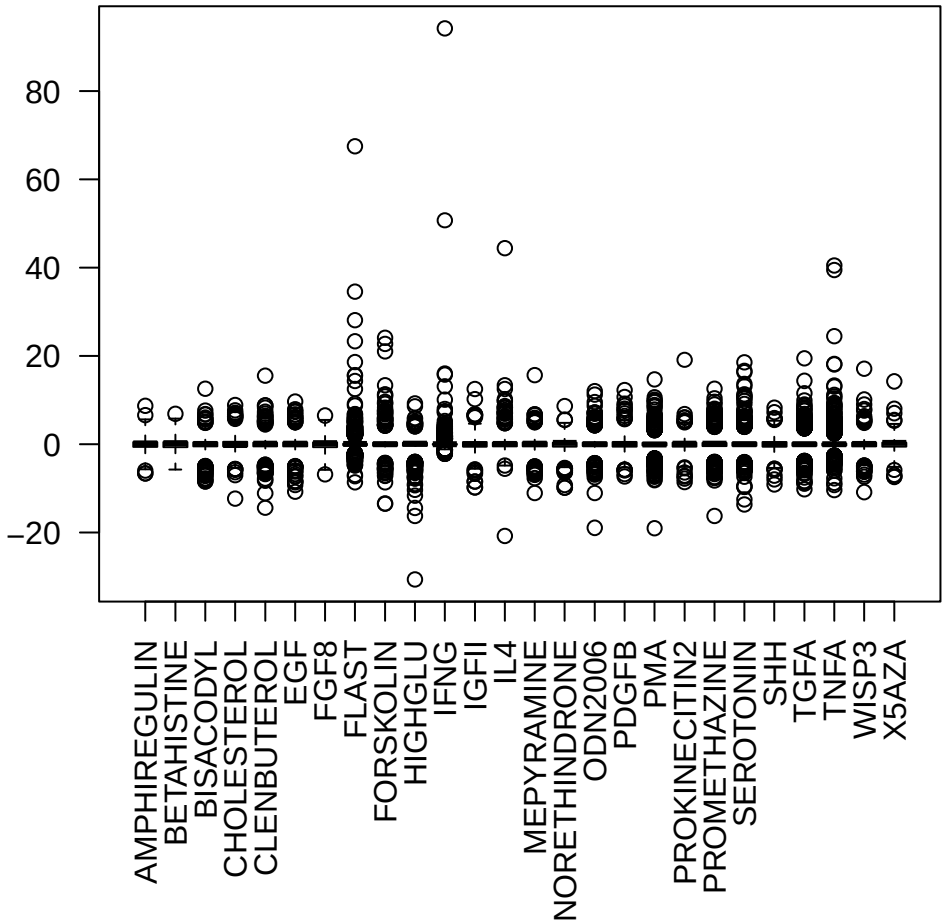


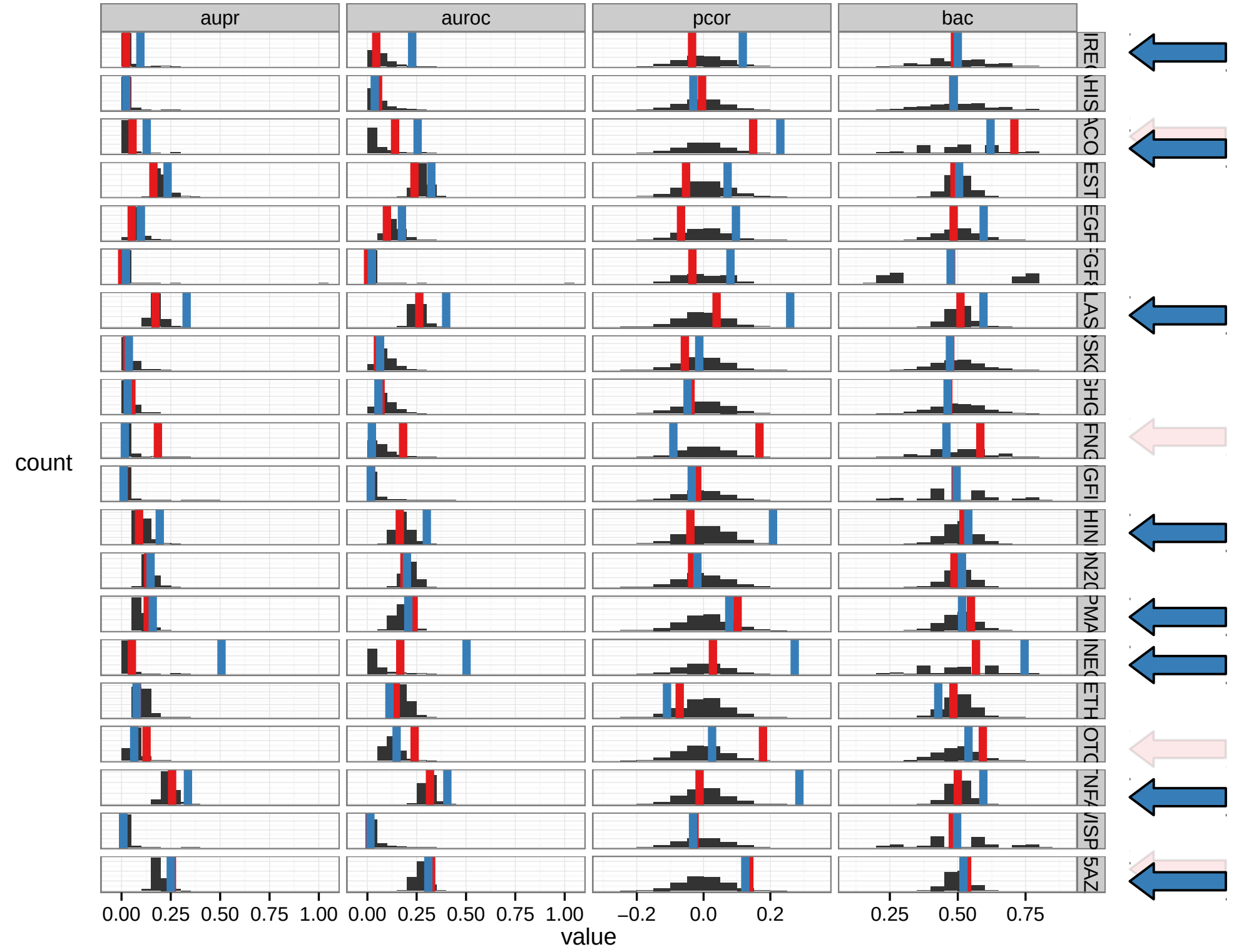
Standardize Rat T-values

Before



After




aupr
auroc
pcor
bac
count
value
0.00
0.25
0.50
0.75
1.00
−0.2
0.0
0.2

Cross-Validation Results

After standardization results are better

Better than random for eight stimuli

- AMPHIREGULIN
- BISACODYL
- FLAST
- NORETHINDRONE
- PMA
- PROKINECITIN2
- TNFA
- X5AZA

Mean performance

	aupr	auroc	pcor	bac
Before	0.095	0.148	0.016	0.518
After Standardization	0.138	0.190	0.074	0.528

Use Stimuli-Response Similarity

Idea: Use similarity between test stimulus and training stimuli

When training the model, care more strongly about predicting t-values for certain stimuli

Approach: Weighted linear regression

Define Weights

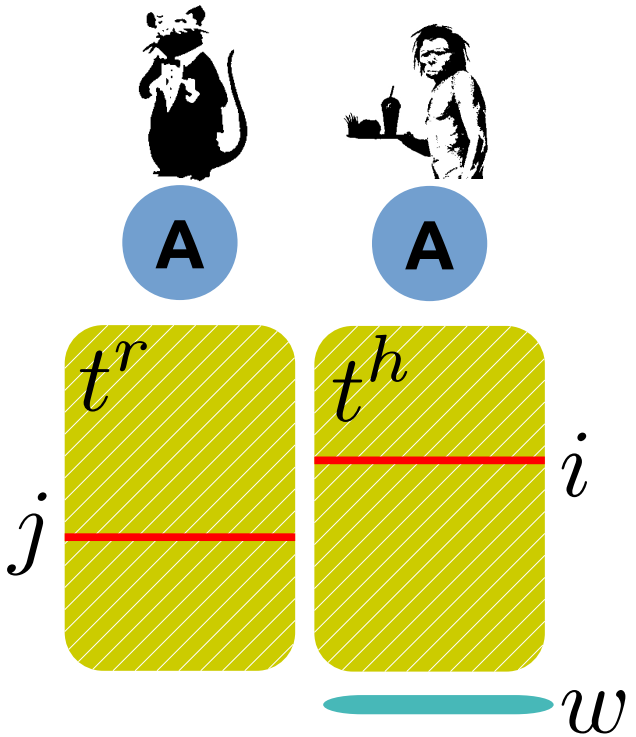
Given a test stimulus, weight vector is based on the number of common genes with high absolute t-values

$$w_s = |\text{top100}^{test} \cap \text{top100}^{train\ s}| + 1$$

genes with top 100 absolute t-values

Weighted Model

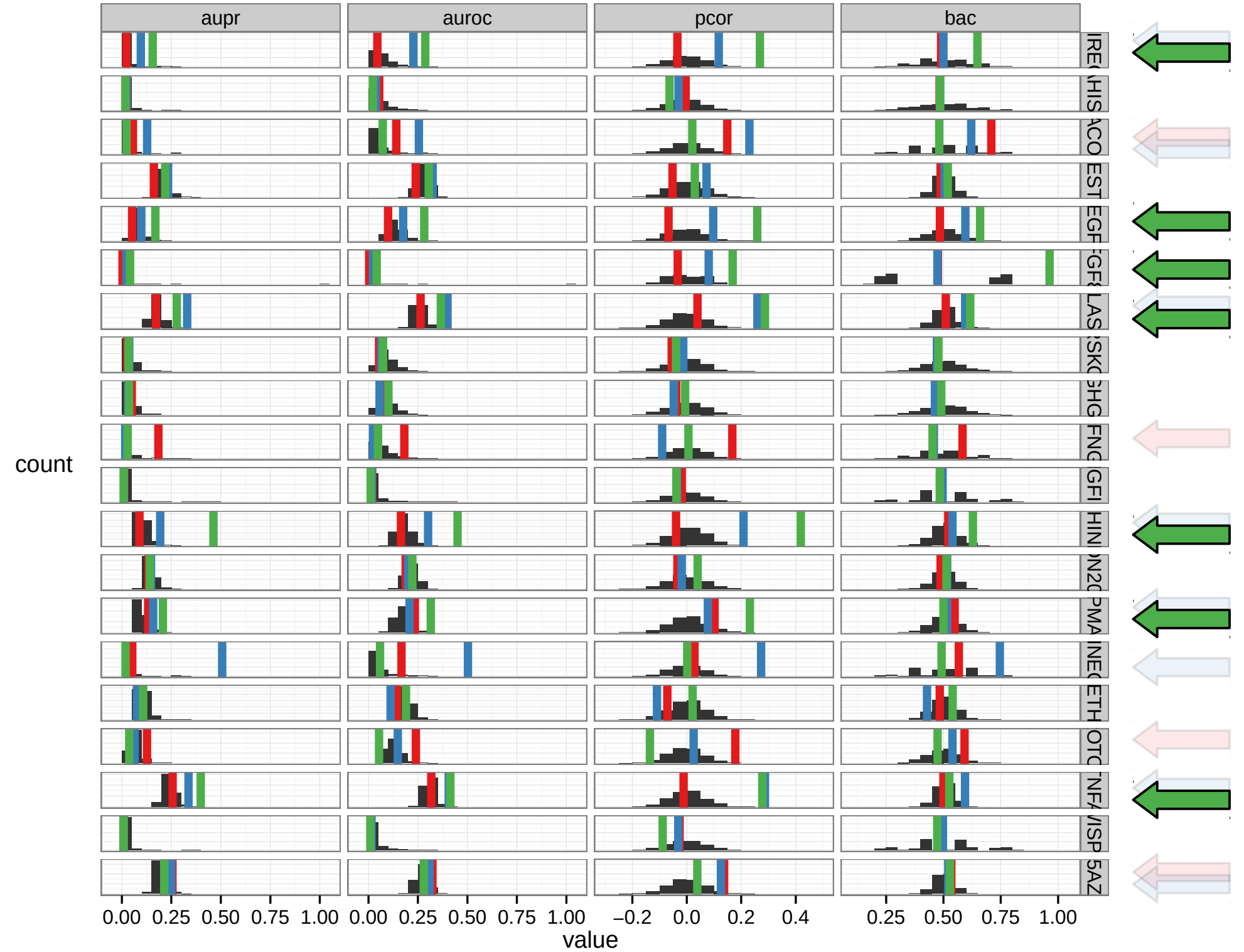
Approach: Weighted linear regression for every human gene **and every test stimulus**



$$t_i^h = \beta t_j^r + \alpha$$

Minimize the **weighted** sum of squares of the residuals of the linear regression model

$$\min_{j,\alpha,\beta} \sum_{s=1}^{n} w_s (t_{is}^h - \beta t_{js}^r - \alpha)^2$$


aupr
auroc
pcor
bac
count
value
0.00
0.25
0.50
0.75
1.00
−0.2
0.0
0.2
0.4

Cross-Validation Results

Better than random for seven stimuli

- AMPHIREGULIN
- EGF
- FGF8
- FLAST
- NORETHINDRONE
- PMA
- TNFA

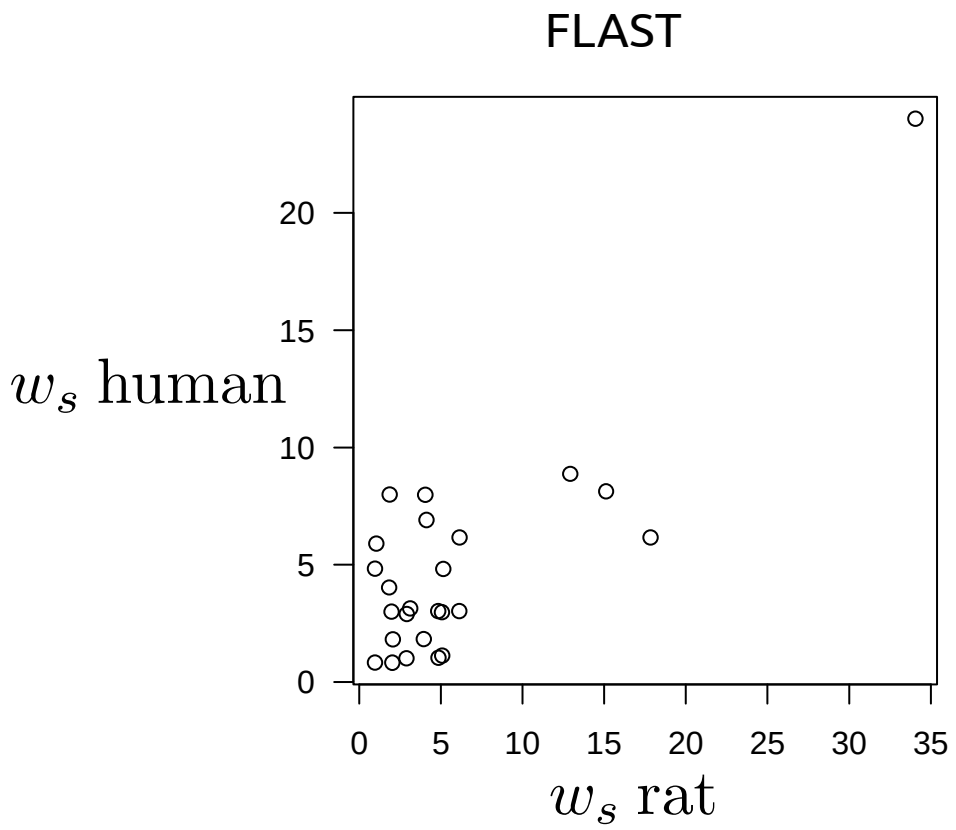
Mean performance

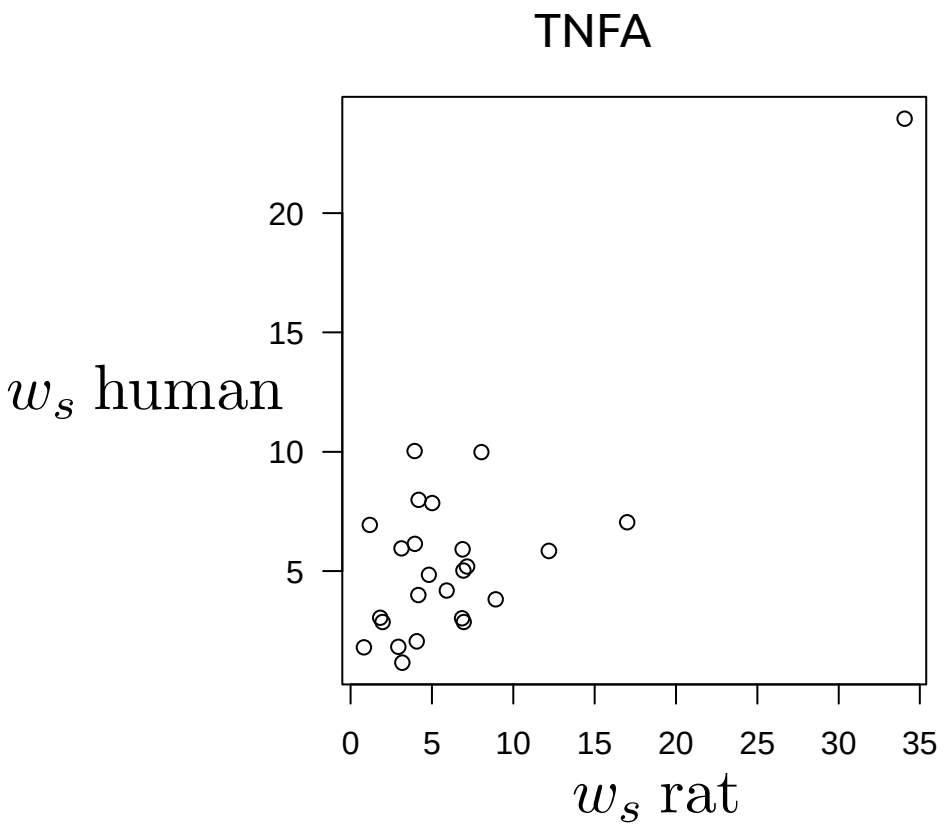
	aupr	auroc	pcor	bac
Before	0.095	0.148	0.016	0.518
After Standardization	0.138	0.190	0.074	0.528
Standardized & Weighted	0.130	0.180	0.085	0.548

Problems of Weighted Model

Model assumption: Stimuli similarities in rat are also true in human

Given training set, this is only true for some stimuli



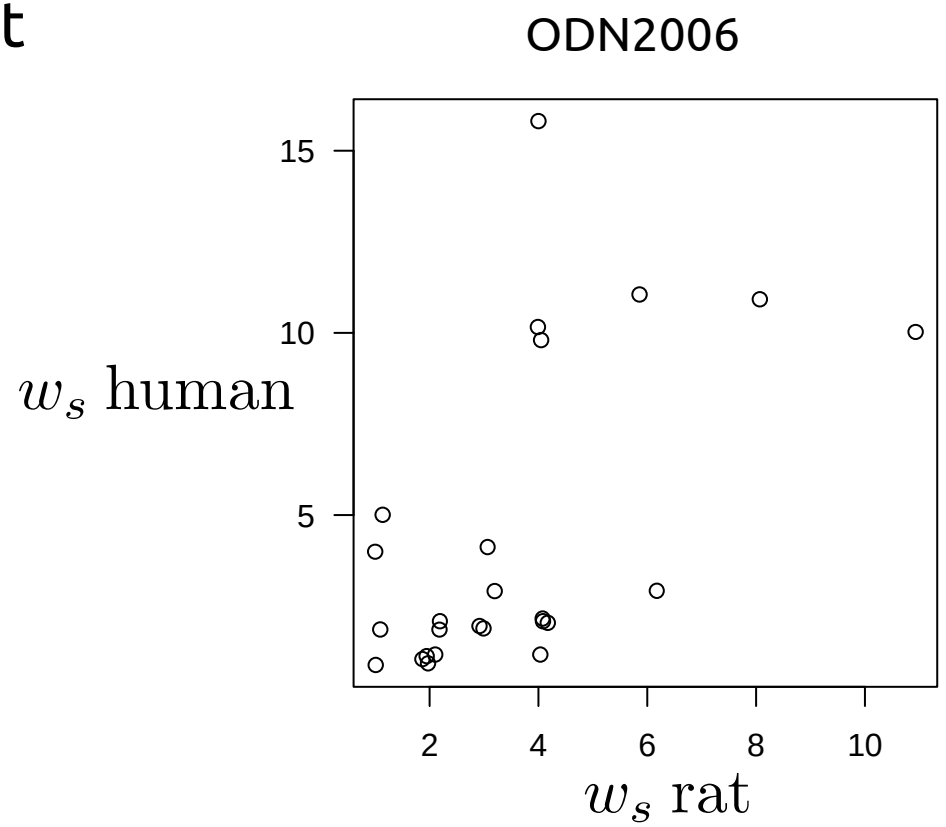


Problems of Weighted Model

Model assumption: Stimuli similarities in rat are also true in human

Given training set, this is only true for some stimuli

Not for most



Summary

- Regression based model operating in t-value space
- Each human gene has some rat gene as predictor
- GSE results better than random for 1/3 of the stimuli during cross-validation

Not covered

- Questionable biological significance
- Interpretation of model parameters

Acknowledgments

My mentor: Rich Bonneau

NEW YORK UNIVERSITY

CENTER FOR GENOMICS
AND SYSTEMS BIOLOGY
NEW YORK UNIVERSITY

Image credits:  &  by Banksy

Availability

Presentation: **goo.gl/rgXLyI**

Code: **goo.gl/Dl7vL8**