Cross-validation for comparing qSIP prediction models trained on same or other groups

Toby Dylan Hocking
toby.hocking@nau.edu
toby.hocking@r-project.org

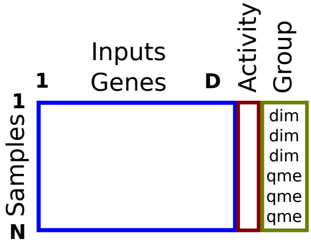
January 11, 2024

Motivation and data for predicting growth from genetics

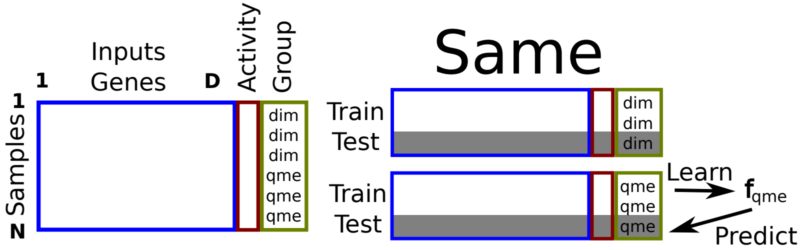
- Goal: look at past experiments to see which genes influence growth.
- Collaboration with Jeff Propster, based on data from Hungate et al, mBio (2021), Stone et al, ISME (2023)
- Seven past experiments with no omics data, but qSIP for total of 188,826 Amplicon Sequence Variants (ASVs), including Arizona elevation gradient (experiment=dim), Quantitative Microbial Ecology (experiment=qme), and others (ant, drp, eag, sro, win).
- picrust2 (Douglas et al, Nature 2020) infers frequency of 8,380 genes in the ASV genome (integer from 0 to 10).
- Can we predict a taxon's qSIP growth/activity from its vector of gene frequencies?

Machine learning predictive analysis of qSIP data

- Inputs/features $\mathbf{x} \in \mathbb{R}^D$ is vector of frequencies for $D = 8380$ genes (range from 0 to 10).
- Output $y \in \mathbb{R}$ is relative activity/growth per day from qSIP (excess atom fraction/EAF normalized by maximum isotope enrichment and incubation length, ranging from 0 to 0.3315).
- Want to learn $f(\mathbf{x}) = y$ (predict growth from genes).
- One hypothesis in these data: can learn f on mixed conifer (MC) controls in experiment=dim (room temp), and accurately predict experiment=qme at temp=15C (or vice versa).
- Question: is this expectation consistent with the data?
- Answer by using 10-fold cross-validation: train on one experiment or other, quantify prediction error on test set.



Inputs
Genes
1
D
Activity
Group
Samples
1
N
dim
dim
dim
qme
qme
qme



Inputs

Genes

1

D

Activity

Group

1

Samples

N

dim
dim
dim
qme
qme
qme

Same

Train

Test

dim
dim
dim

Train

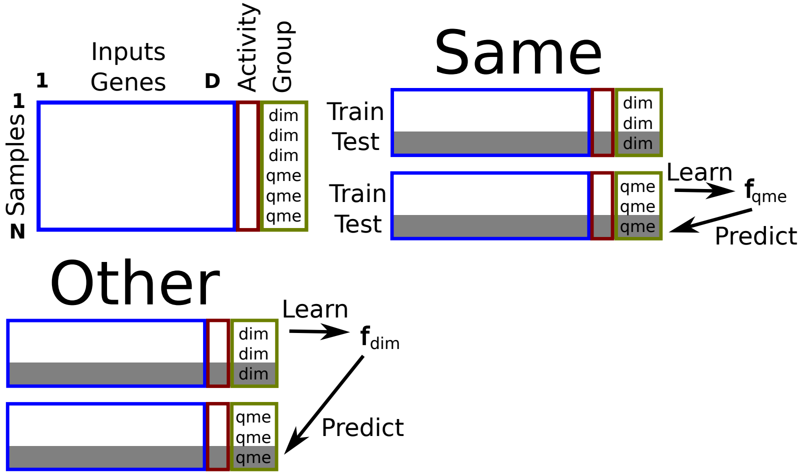
Test

qme
qme
qme

Learn

$\mathbf{f}_{\text{qme}}$

Predict



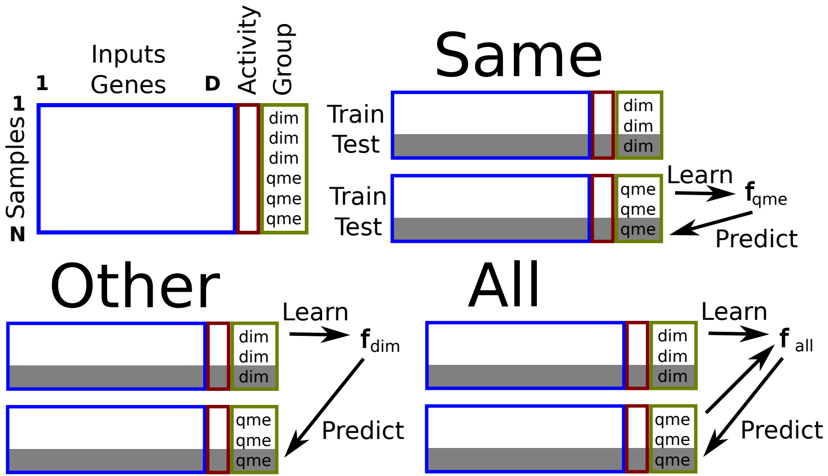
Inputs
Genes
1 D
Activity
Group
Samples
1
N
dim
dim
dim
qme
qme
qme

Same

Train
Test
dim
dim
dim
Train
Test
qme
qme
qme
Learn
f_{qme}
Predict

Other

dim
dim
dim
Learn
f_{dim}
qme
qme
qme
Predict



Inputs
Genes
1
D
Activity
Group
Samples
1
N
dim
dim
dim
qme
qme
qme

Same

Train
Test
dim
dim
dim
Train
Test
qme
qme
qme
Learn
f_{qme}
Predict

Other

dim
dim
dim
Learn
f_{dim}
qme
qme
qme
Predict

All

dim
dim
dim
Learn
f_{all}
qme
qme
qme
Predict

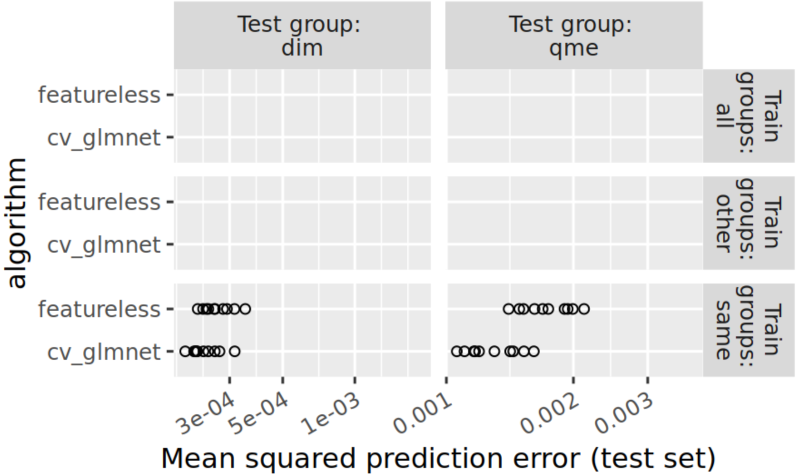
Comparison 1: controls in different experiments

- Data table with $N = 7710$ rows/observations (ASVs), across two experiments dim=3120, qme=4590.
- $D = 8380$ gene features.
- We compare two learning algorithms

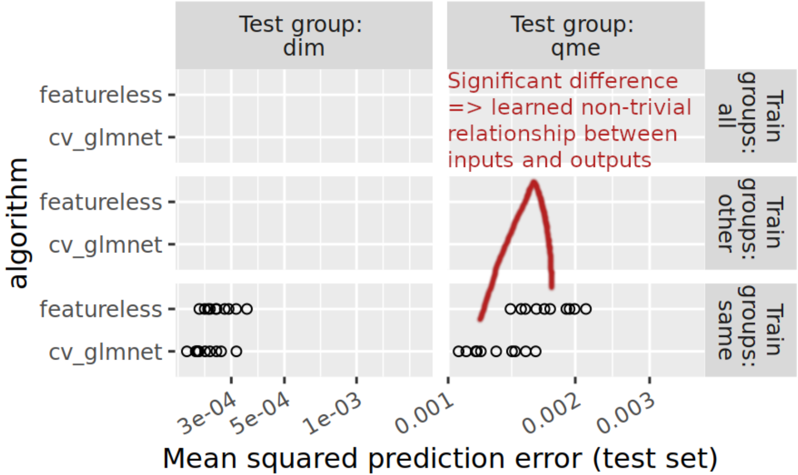
  cv_glmnet: L1 regularized linear model (LASSO), small subset of important genes selected and used for prediction (other un-important genes are not used for prediction).

  featureless ignore all genes/features, and always predict mean output in train set.
- If there is any non-trivial relationship/pattern learned between inputs and outputs, then **linear model should have smaller prediction error than featureless**.
- If patterns are similar in different groups/experiments (dim and qme), then **linear model should have similar prediction error, when trained on other groups/experiments**.

controls between experiments

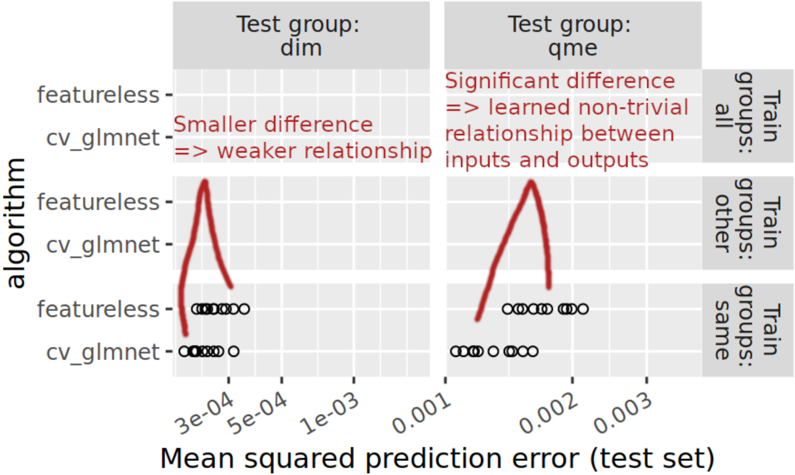


controls between experiments



Significant difference => learned non-trivial relationship between inputs and outputs

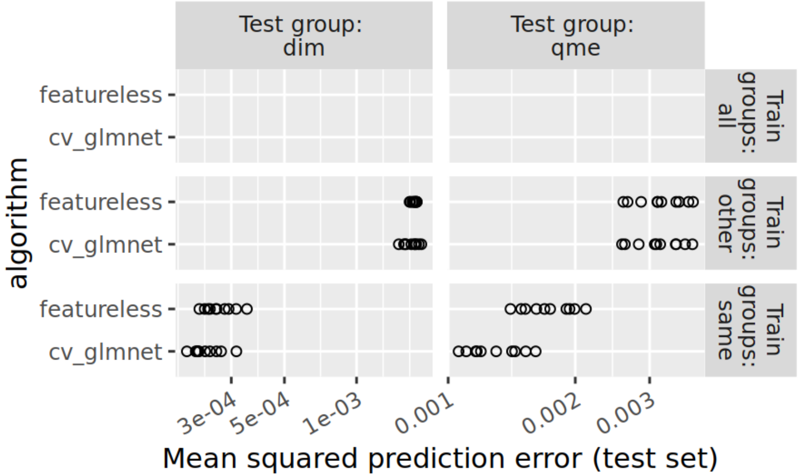
controls between experiments



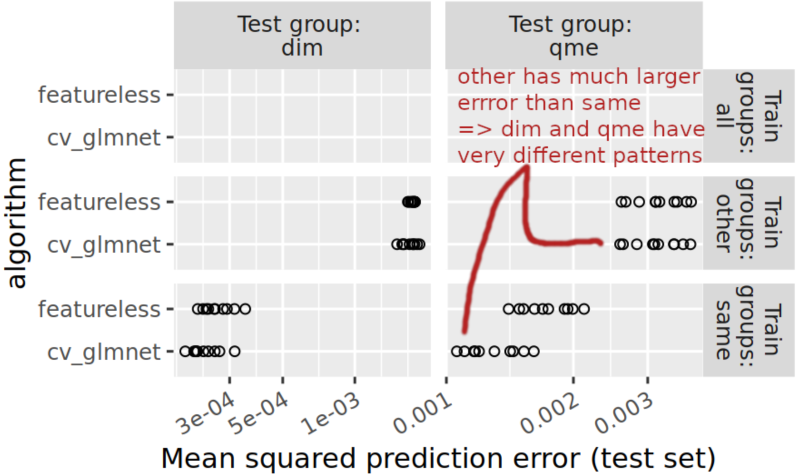
Smaller difference => weaker relationship

Significant difference => learned non-trivial relationship between inputs and outputs

controls between experiments

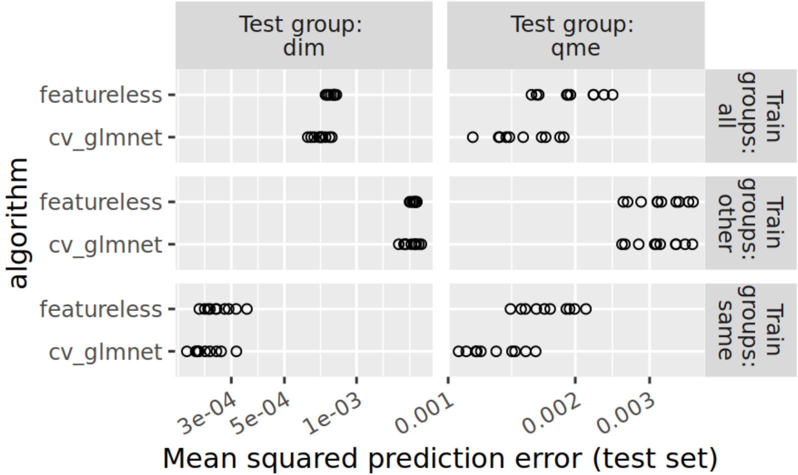


controls between experiments



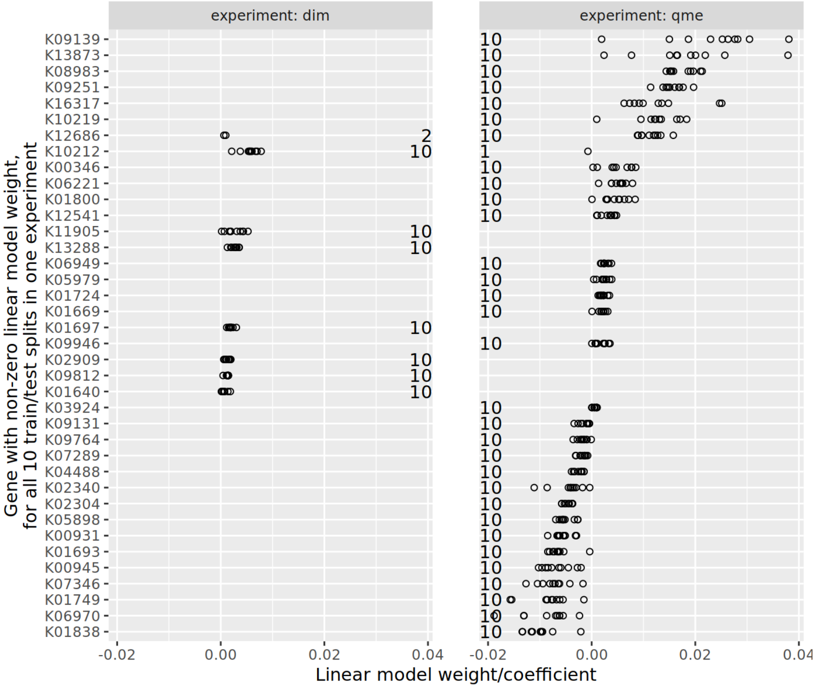
other has much larger errror than same => dim and qme have very different patterns

controls between experiments



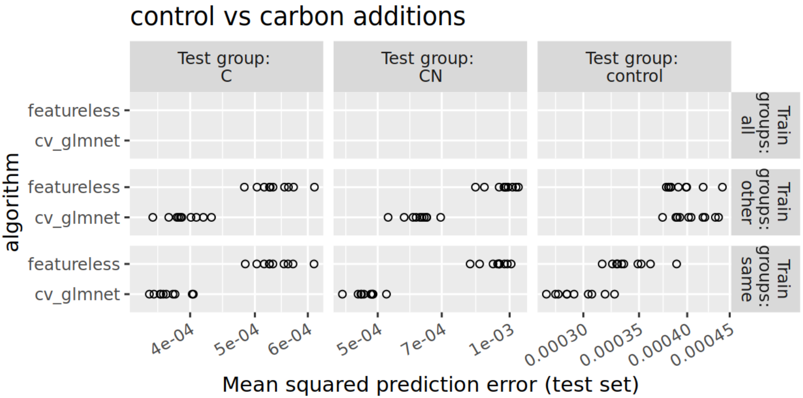
Interpretation of linear model prediction error and weights

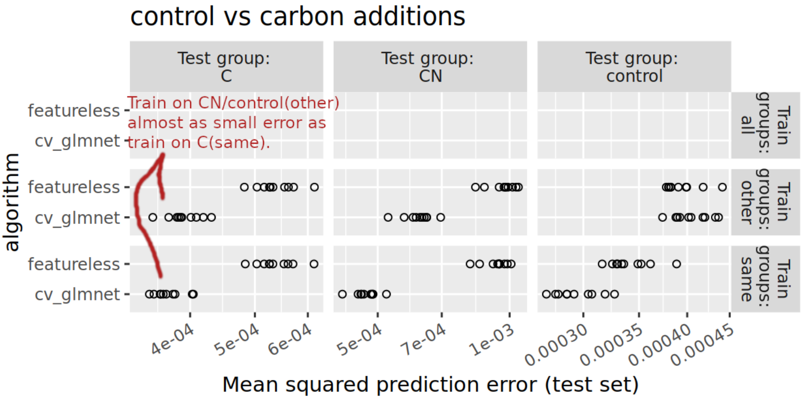
- Hypothesis was: expect we can learn f on mixed conifer (MC) controls in experiment=dim (room temp), and accurately predict experiment=qme at temp=15C (or vice versa).
- Prediction error cross-validation analysis on previous slide is not consistent with that hypothesis.
- So there should be a different prediction function in each experiment. What is the difference?
- The L1 regularized linear model (LASSO) can be interpreted in terms of which genes are important/used for prediction (non-zero weights/coefficients) and others are ignored (weights=0, not used for prediction).
- Compute and plot weights which are non-zero/important in all 10 train/test splits of cross-validation.


experiment: dim
experiment: qme
Gene with non-zero linear model weight,
for all 10 train/test splits in one experiment
K09139
K13873
K08983
K09251
K16317
K10219
K12686
K10212
K00346
K06221
K01800
K12541
K11905
K13288
K06949
K05979
K01724
K01669
K01697
K09946
K02909
K09812
K01640
K03924
K09131
K09764
K07289
K04488
K02340
K02304
K05898
K00931
K01693
K00945
K07346
K01749
K06970
K01838
2
10
10
10
10
10
10
10
10
10
10
10
10
10
10
10
1
10
10
10
10
10
10
10
10
10
10
10
10
10
10
10
10
10
10
10
10
10
10
10
10
10
10
-0.02
0.00
0.02
0.04
-0.02
0.00
0.02
0.04
Linear model weight/coefficient

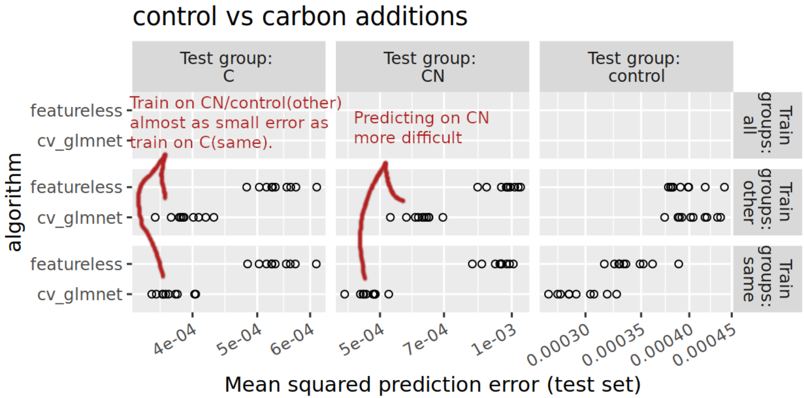
Comparison 2: control versus carbon additions

- $N = 60877$ samples total, in 3 groups/treatments: control=17225, C=23214 (carbon added), CN=20438 (carbon and nitrogen added).
- Same $D = 8380$ gene features.
- Can we train on one group/treatment, and predict accurately on another?


control vs carbon additions
Test group: C
Test group: CN
Test group: control
Train groups: all
Train groups: other
Train groups: same
algorithm
featureless
cv_glmnet
featureless
cv_glmnet
featureless
cv_glmnet
4e-04
5e-04
6e-04
5e-04
7e-04
1e-03
0.00030
0.00035
0.00040
0.00045
Mean squared prediction error (test set)

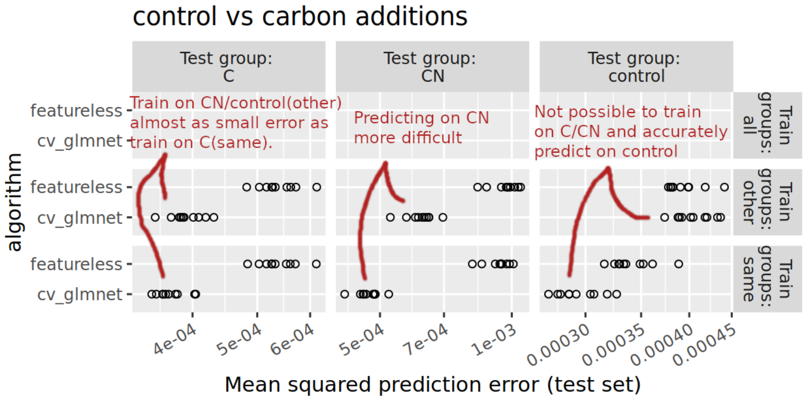


Train on CN/control(other) almost as small error as train on C(same).



Train on CN/control(other) almost as small error as train on C(same).

Predicting on CN more difficult



Train on CN/control(other) almost as small error as train on C(same).

Predicting on CN more difficult

Not possible to train on C/CN and accurately predict on control

Discussion and conclusions

- Often we want to know if we have similar or different patterns in different data groups (train on one experiment/treatment, predict on another).
- Cross-validation can be used to determine the extent to which we can train on one group, and accurately predict on another.
- Machine learning algorithms like L1 regularized linear models (LASSO/cv_glmnet) are additionally interpretable in terms of which features are used for prediction (can be compared between models trained on different groups).
- Free/open-source software available: mlr3resampling R package on CRAN and `https://github.com/tdhock/mlr3resampling`
- Let's collaborate! Contact: toby.hocking@nau.edu, toby.hocking@r-project.org