Cross-validation for training and testing co-occurrence network inference algorithms

Daniel Agyapong

da2343@nau.edu

PhD student

Dr. Toby Dylan Hocking

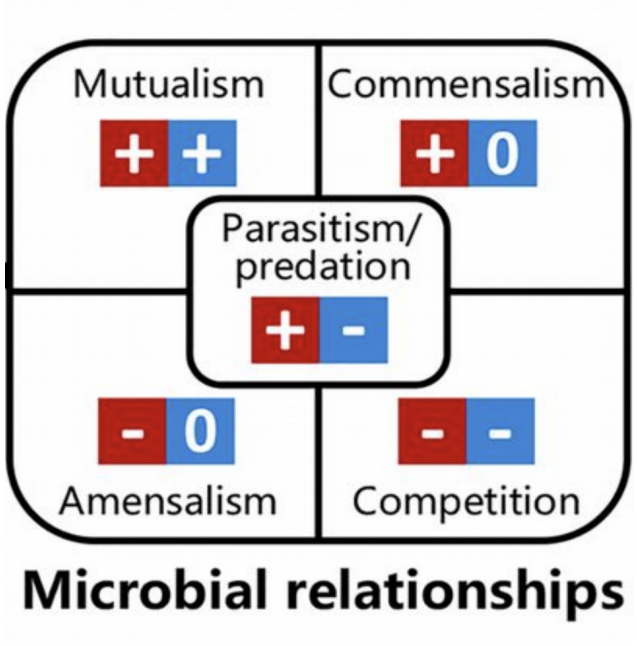
Toby.Hocking@nau.edu

Machine Learning Lab Director

NAU NORTHERN ARIZONA UNIVERSITY

School of Informatics, Computing, and Cyber Systems

Background

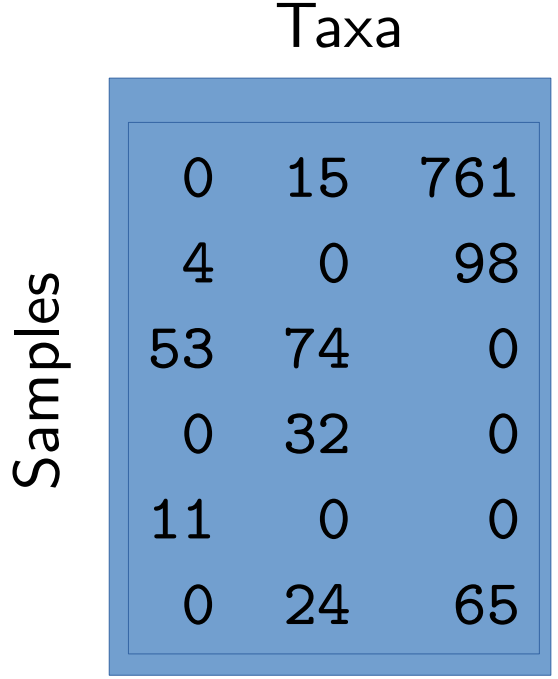


- Micro-organisms form complex ecological interactions :
  - **Mutualism**: Both parties benefit. Mutual cross-feeding.
  - **Parasitism/Predation**: One side benefits whilst the other side loses. Relationships such as predator-prey and host-parasite interactions.
  - **Competition**: Both parties lose. When there is insufficient resources for both organisms, they compete for the limited resources.
  - **Commensalism**: One organism benefits without harming the other.
  - **Amensalism**: One organism is harmed but the other is unaffected.
- Reconstructing microbial ecological networks to represent these interactions would help to understand the complex behaviors in microbial communities.

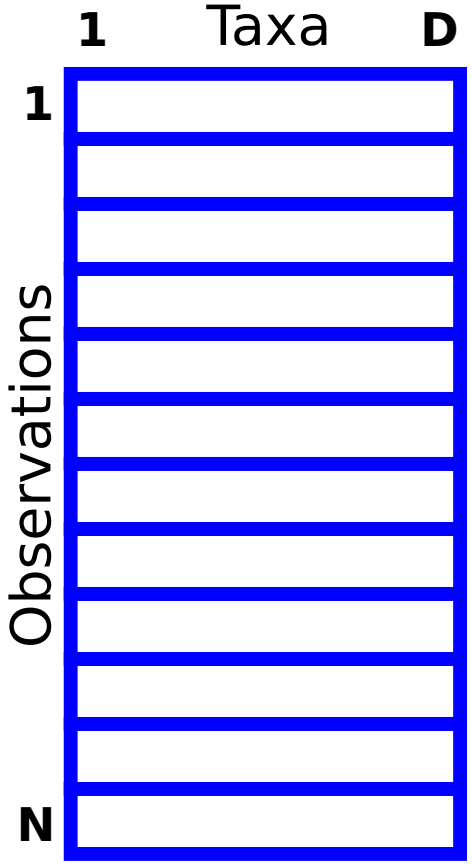
Real Microbiome Abundance Data

Data	Citation	Samples	Taxa
amgut1	https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004226	289	127
amgut2		296	138
hmp216S	https://ibdmdb.org/tunnel/public/summary.html	47	45
hmp2prot		47	43
enterotype	https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061217	280	553
esophagus		3	58
crohns	https://www.mcgill.ca/statisticalgenetics/software	100	5
Baxter_CRC	http://www.raeslab.org/companion/ocean-interactome.html	490	117
glne007		490	338
iOraldat	https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03911-w	86	63

Each data set is a matrix of counts, for example:

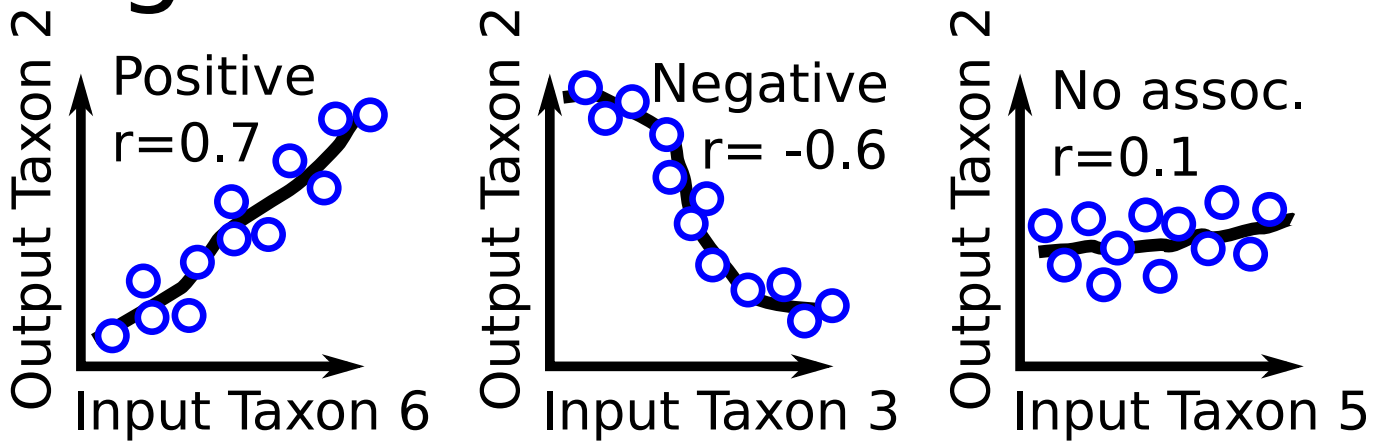


Algorithm learns regression model and co-occurence network

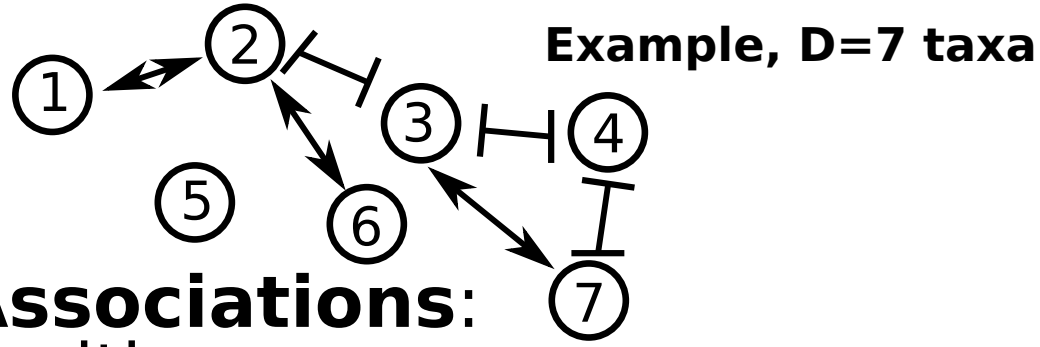


Learn correlations between taxa

Regression model

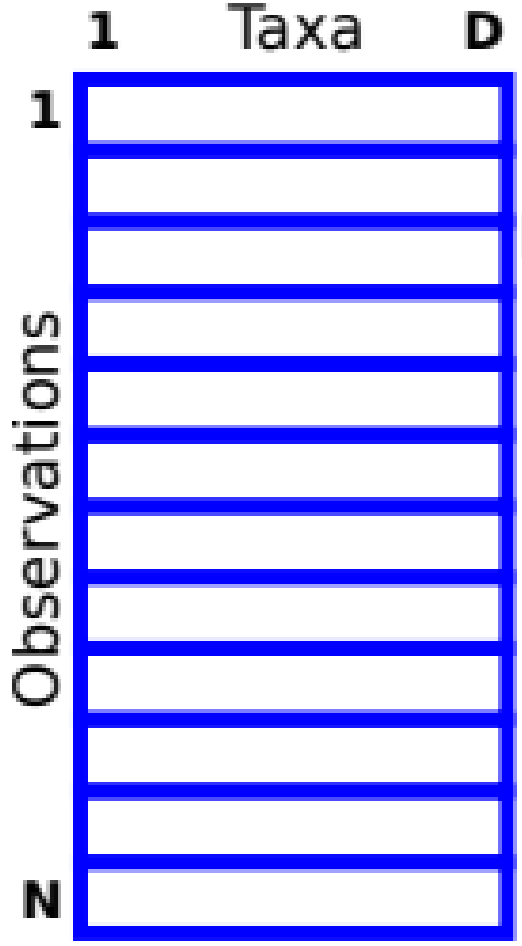


Co-occurence network

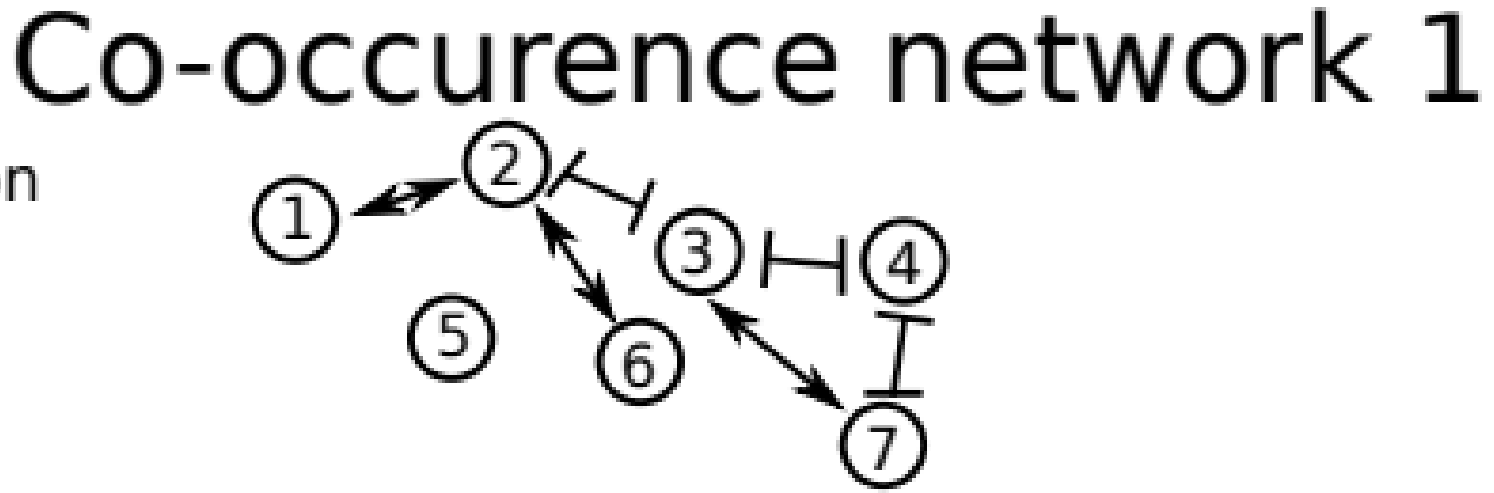


Associations:
Positive ↔
Negative ⊣

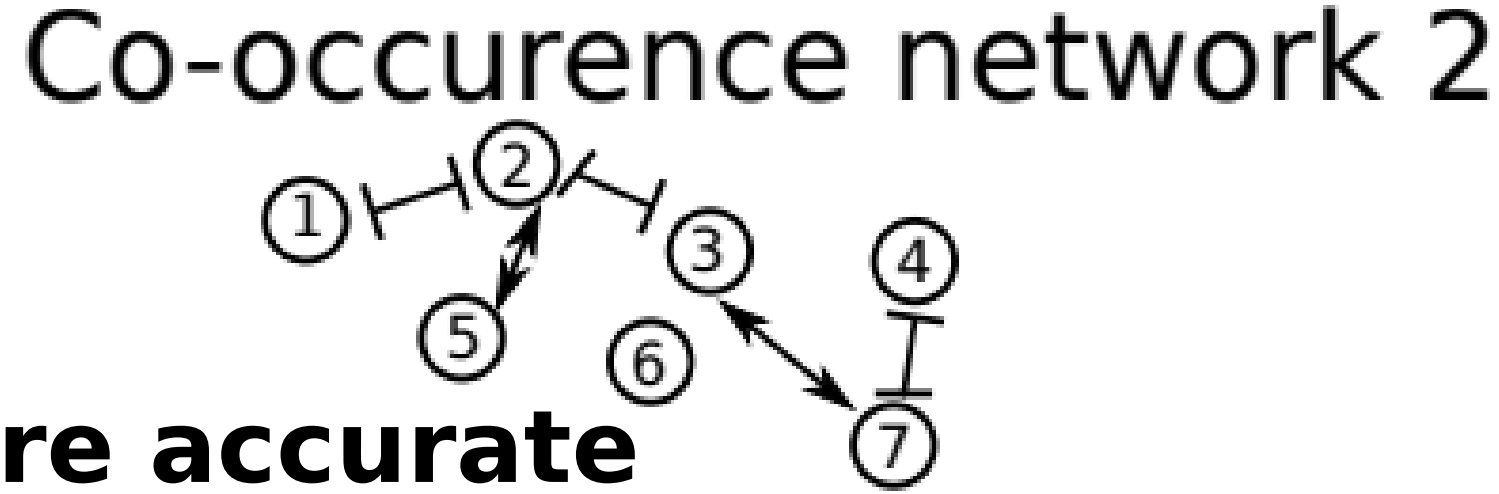
Two algorithms give two different co-occurence networks



Algo 1
Ex: correlation



Algo 2
Ex: Lasso



Which is a more accurate interpretation for these data?

Research Questions

There are many existing algorithms, each with various hyper-parameters which determine the sparsity (number of edges) in network.
- Linear/Pearson correlation: threshold on r2.
- Lasso: degree of L1 regularization.

For a particular real data set, like the ones we will be gathering in this project, how can we automatically learn hyper-parameters? (let the data tell us the “best” threshold, rather than choosing arbitrarily)

And which algorithm is best? (Pearson correlation or Lasso?)

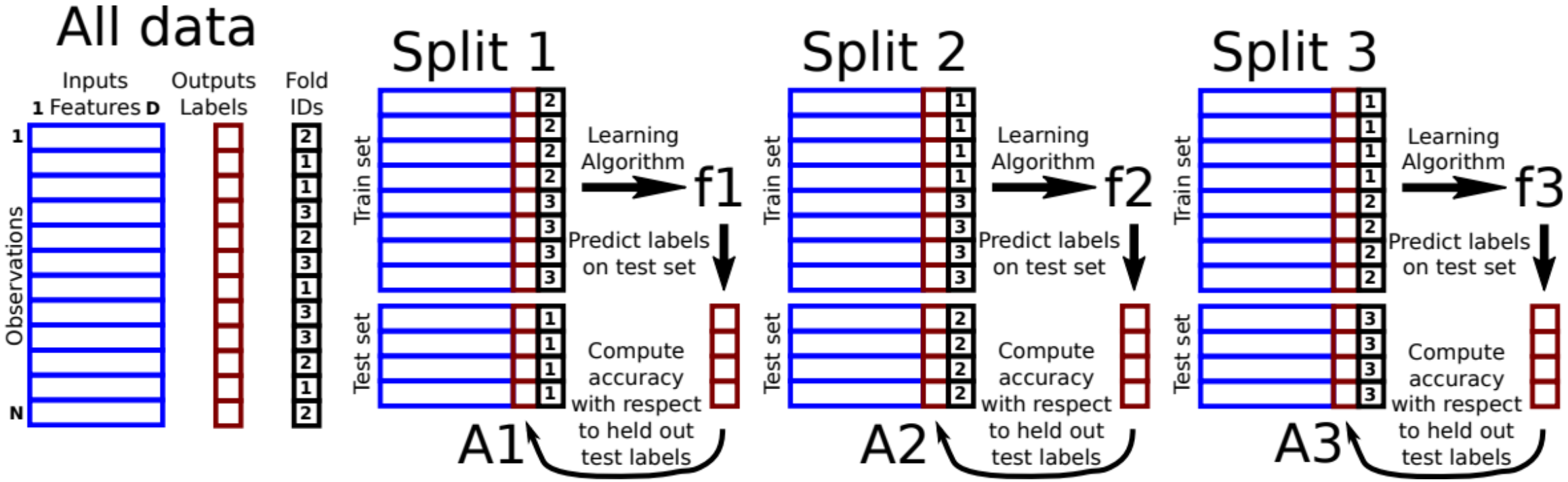
Previous Correlation Based Methods

Method	SparCC (2012)	REBACCA(2015)
Link	https://rdrr.io/github/zdk123/SpiecEasi/man/sparcc.html	https://faculty.wcas.northwestern.edu/hji403/REBACCA.htm
Algorithms Compared	SparCC, Pearson	REBACCA, SparCC, BP, ReBoot
How they compare	Computing the number of true-positives (TP), false-positives (FP), true-negatives (TN) and false-negatives (FN) detected in the Pearson network by treating the SparCC network as the true one.	**Consistency** of correlated pairs identified independently from the three datasets (A correlated pair of OTUs is consistent between two datasets if the pair has the same signs of correlations in both datasets).
Category of Evaluation Type	External data	External data

Previous LASSO Based Methods

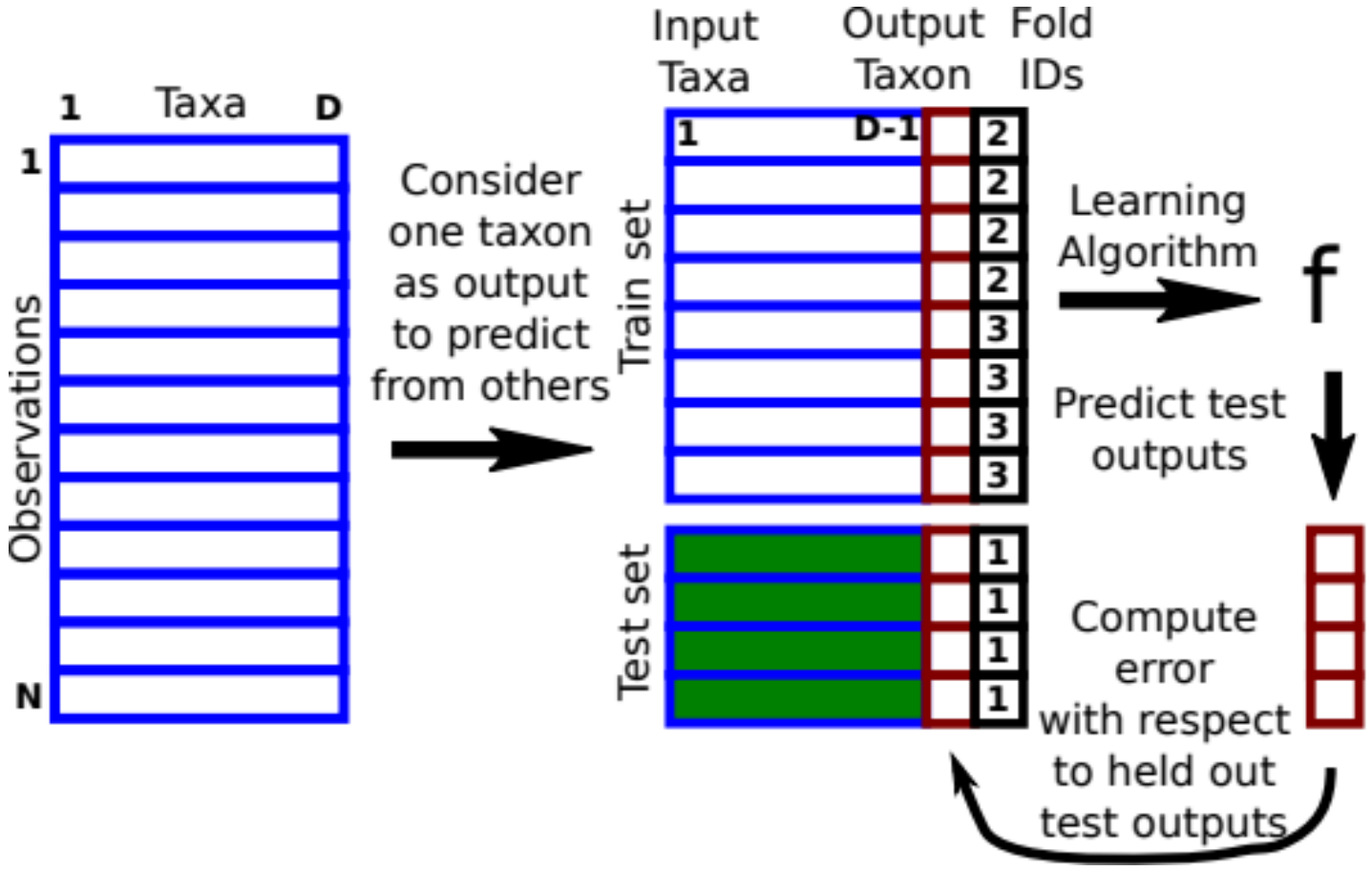
Method	SPIEC-EASI (2015)	CCLasso (2015)
Link	https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004226#pcbi-1004226-g006	https://github.com/huayingfang/CCLasso
Algorithms Compared	SPIEC-EASIE, SparCC, CCREPE	CCLasso, SparCC
How they compare	Consistency between two models by computing Hamming Distance (the difference between the upper triangular part of the two adjacency matrices) between reference and new models.	**Frobenius Accuracy** with respect to estimating correlation matrix from data using half samples (measured by the Frobenius norm distance between the estimated correlation matrices and a reference correlation matrix). **Reproducibility** (measured by the fraction of the same edges shared for the two steps in the first reference network which only the top 1/4 edges is used)
Category of Evaluation Type	External Data (Amgut Dataset)	Sub-sample analysis

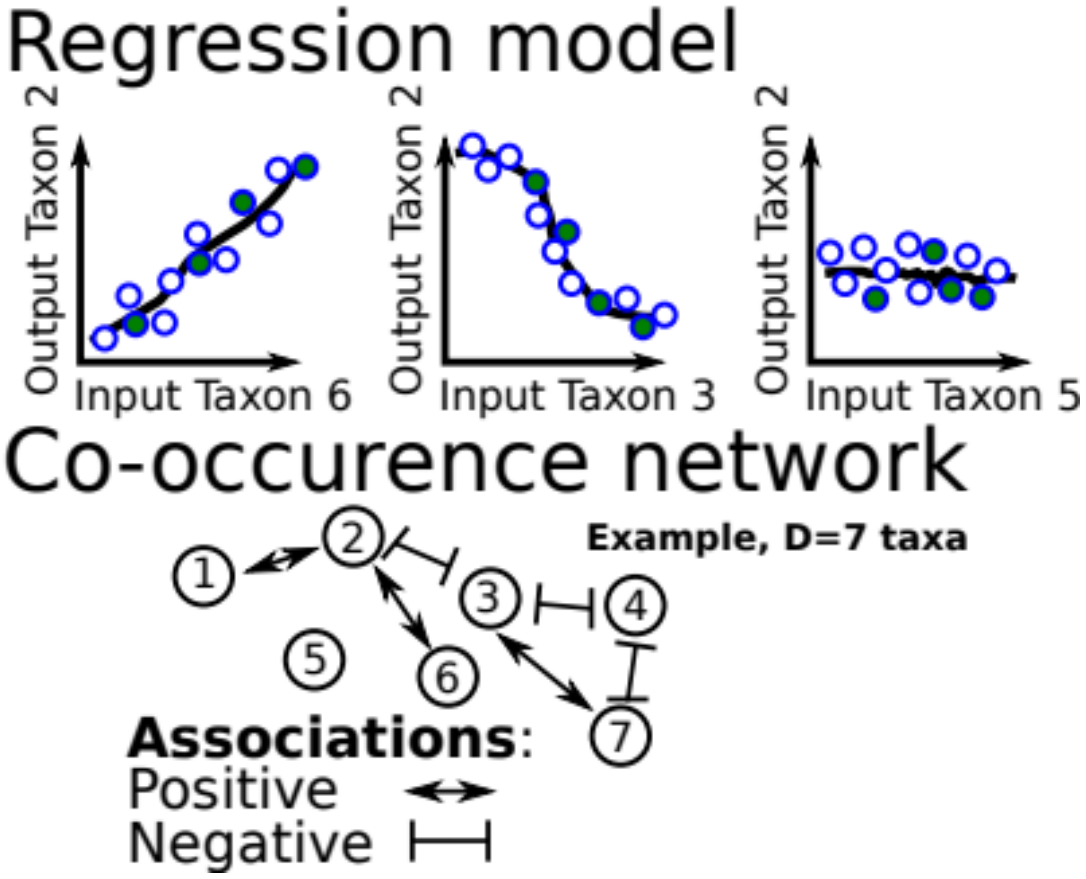
Cross-validation for supervised learning



- K-Fold cross-validation: each observation assigned a fold ID, K=3 means fold IDs between 1 and 3.
- For each fold ID, use corresponding observations as a test set to evaluate generalization ability of learning algorithm (trained on all other observations).

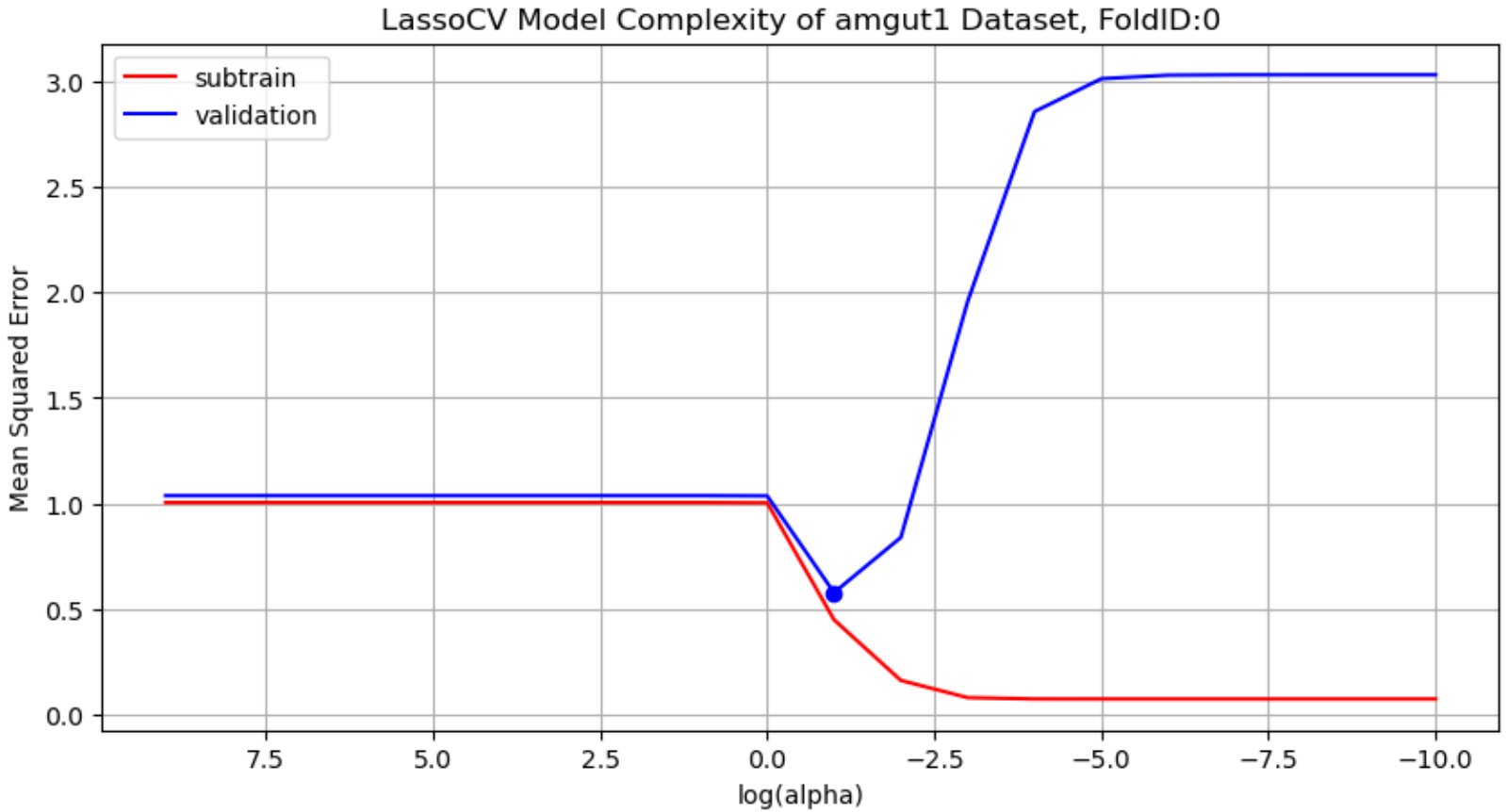
Proposal: cross-validation for training and testing co-occurrence network inference algorithms





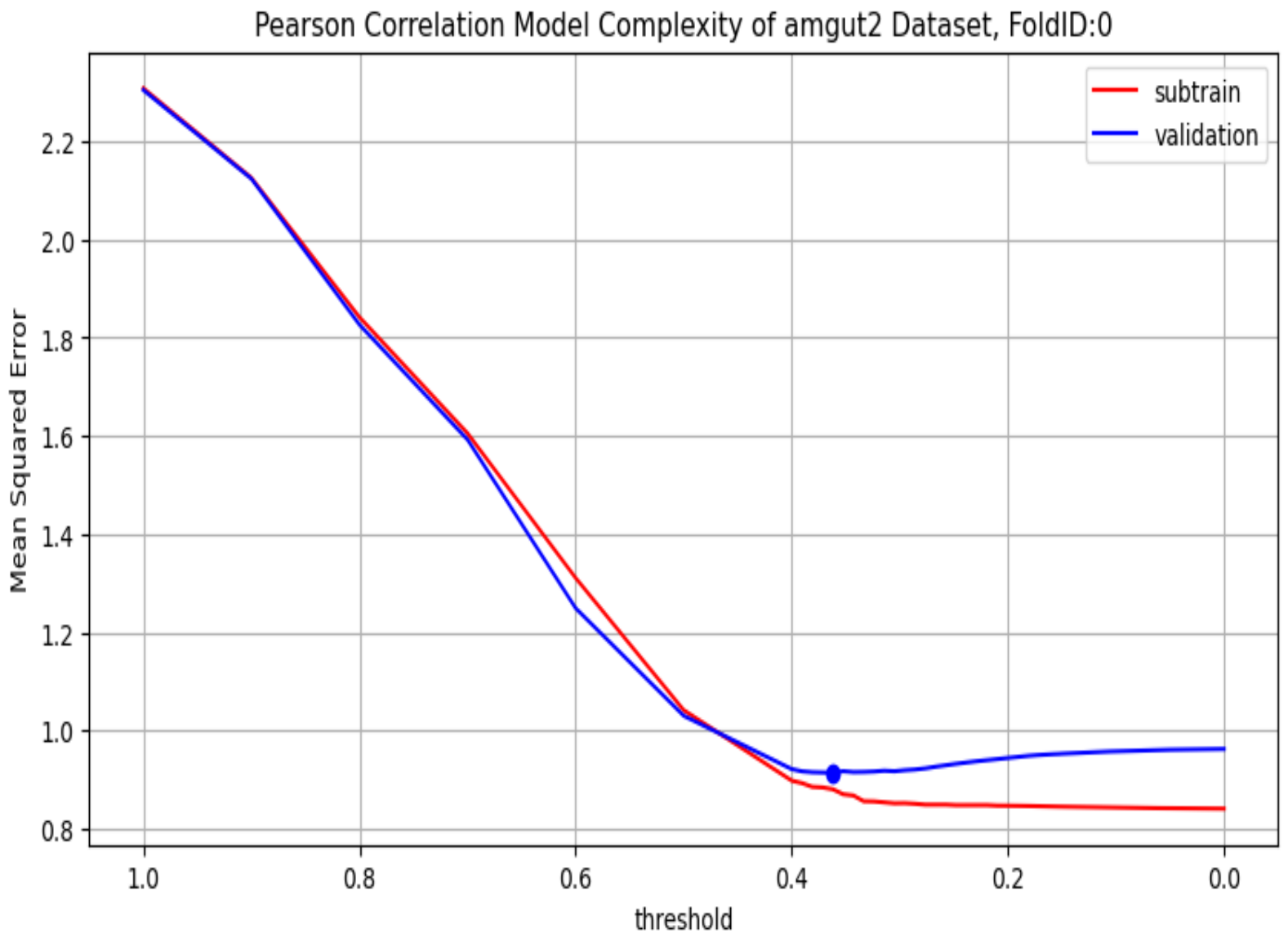
Repeat for each output taxon and fold ID.

Results: Training the Lasso algorithm with cross-validation



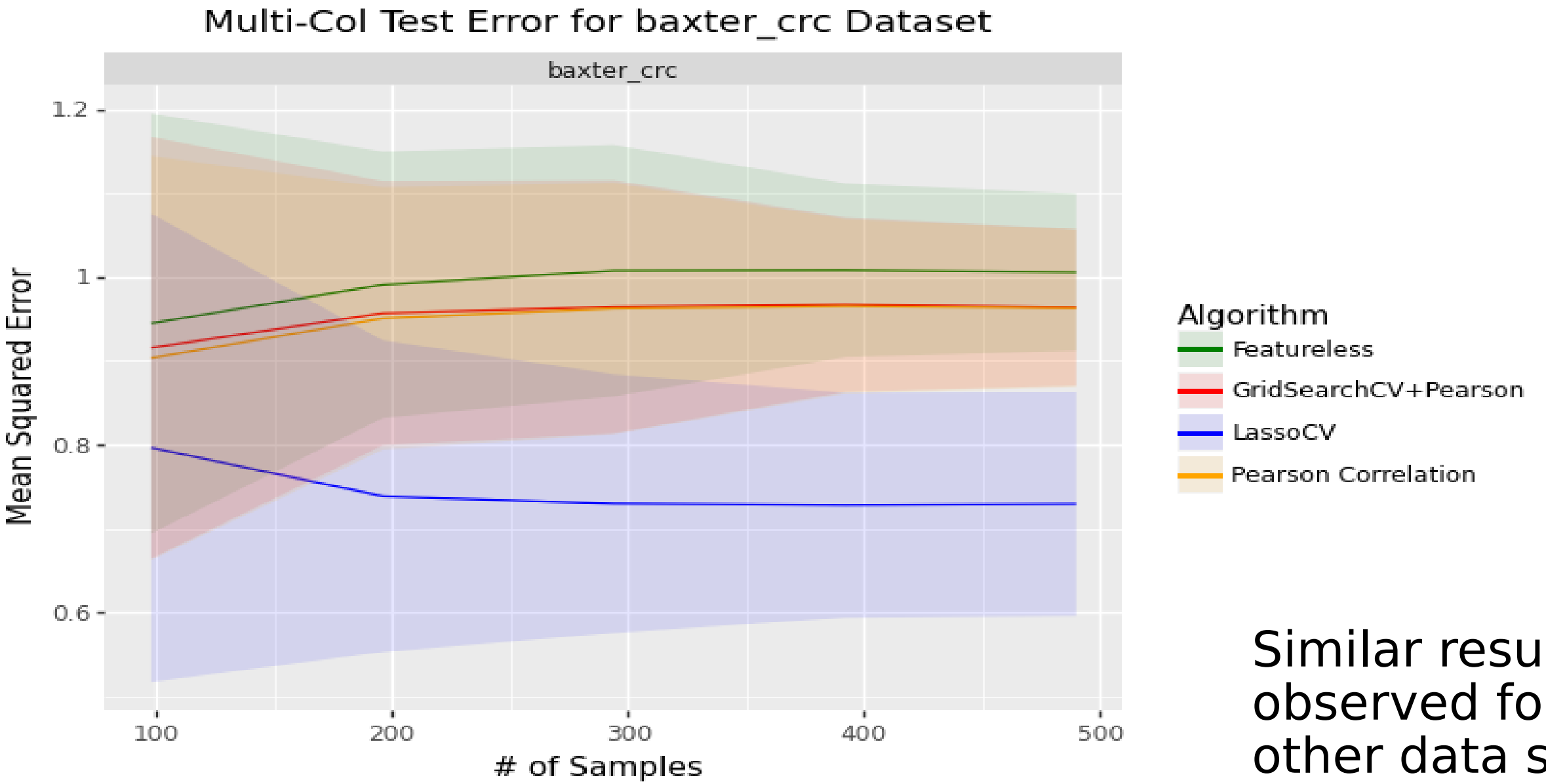
- Train set split into subtrain set (used to learn regression coefficients) and validation set (used to learn model complexity, degree of L1 regularization, which controls sparsity / number of edges in co-occurence network).
- Subtrain error decreases, while the validation error shows expected U shape.
- We select the alpha value (degree of L1 regularization) which minimizes the validation error.

Results: training the Pearson correlation threshold using cross-validation



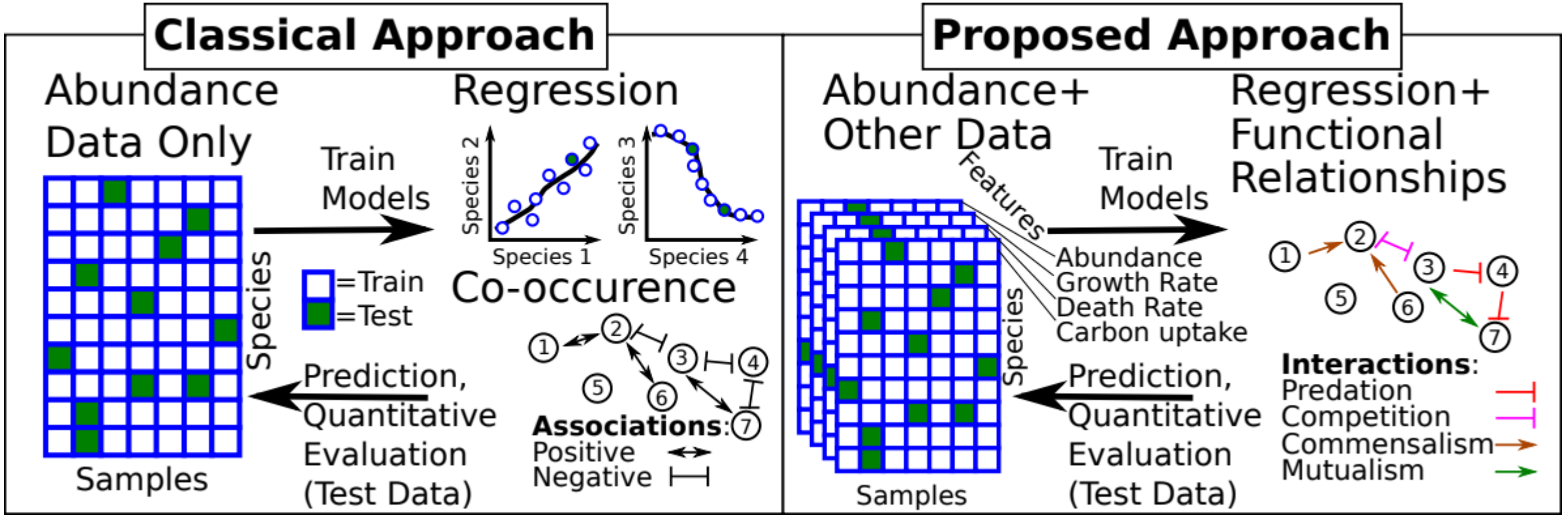
- Subtrain error decreases as the model complexity increases whilst the validation error shows a U shape.
- We select the threshold which gives the minimum validation error, in this example r2=0.35 (any smaller r2 values will have no edge in the co-occurrence network).

Results: algorithms can be compared using test error, and Lasso has smaller error in this particular data set



Similar results observed for other data sets

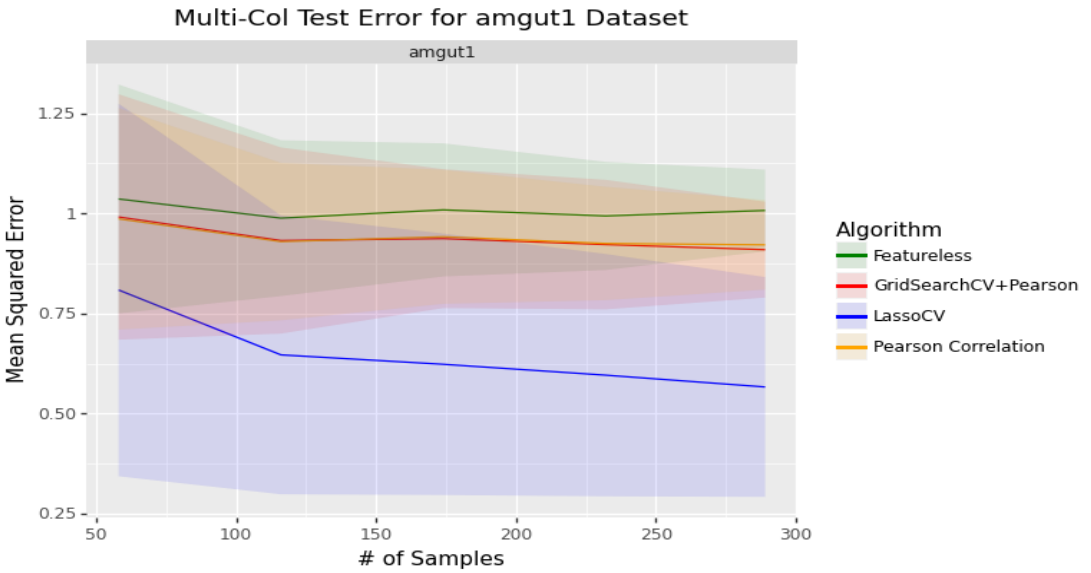
Future work: cross-validation for training and testing interaction network inference algorithms, using several qSIP data features

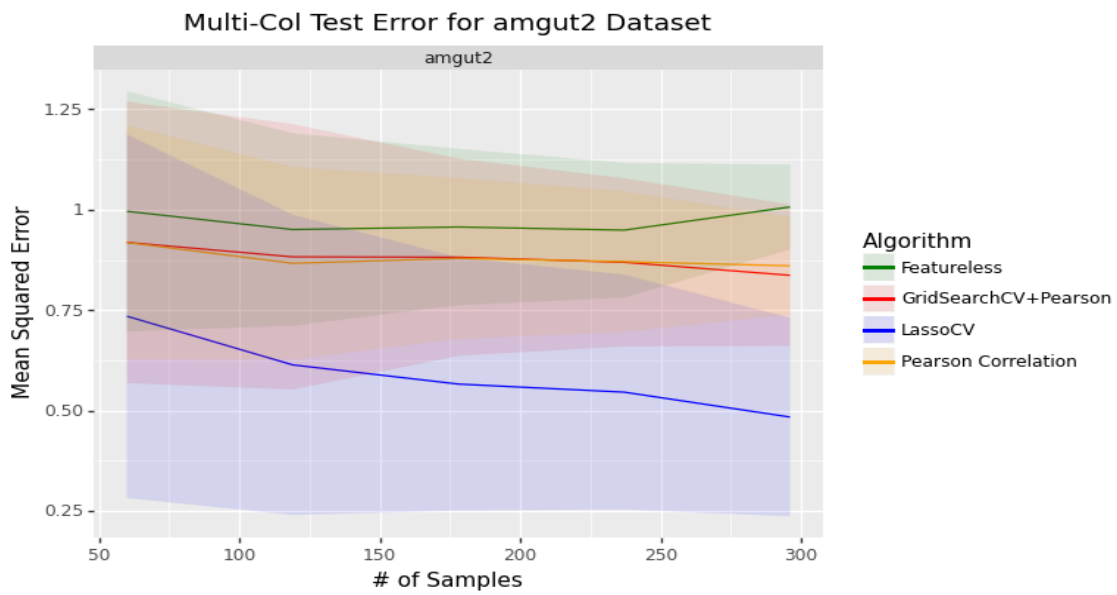


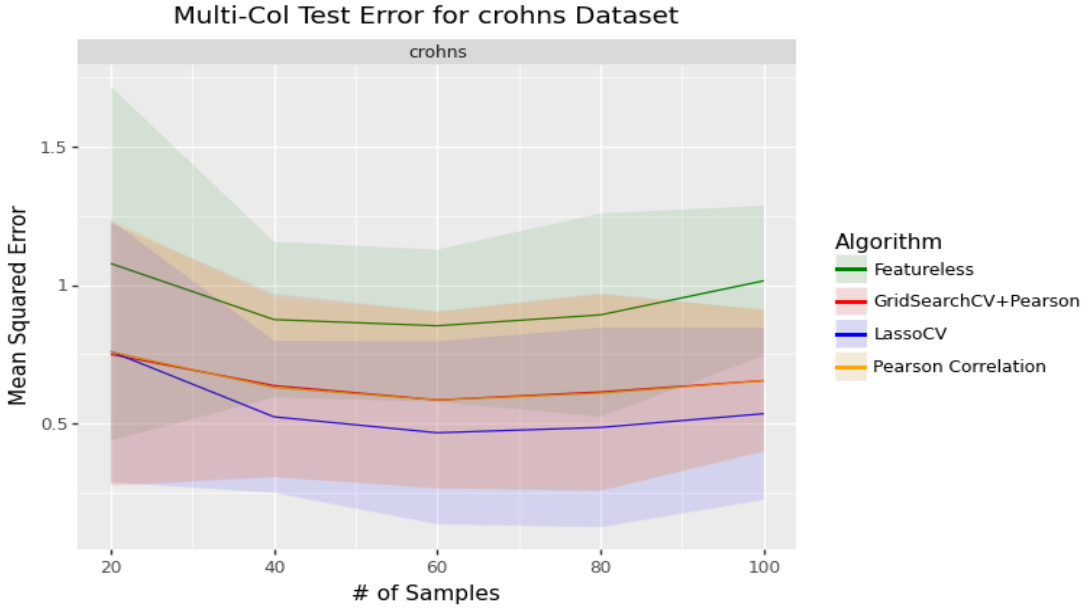
Contact: toby.hocking@nau.edu

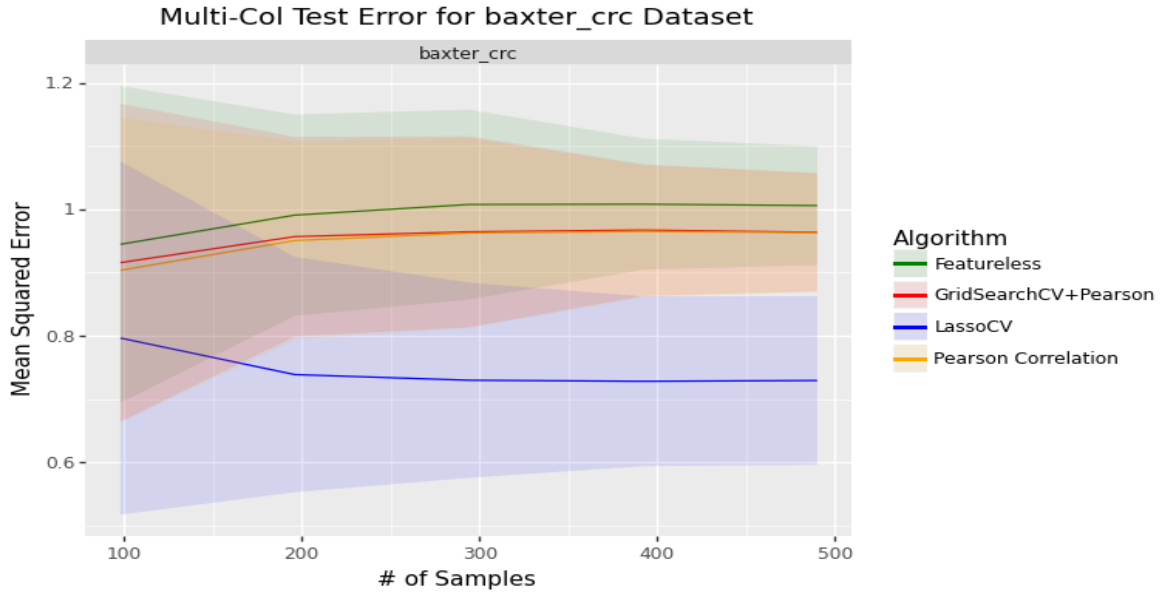
Reproducibility: https://github.com/EngineerDanny/CS685-Microbe-Network-Research

Results: algorithms can be compared using test error, and Lasso has smaller error in four different data sets



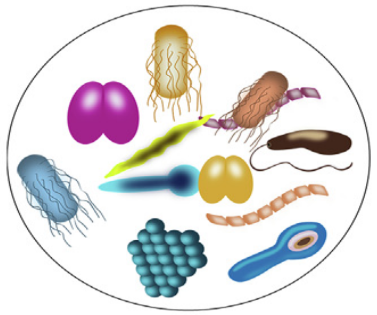




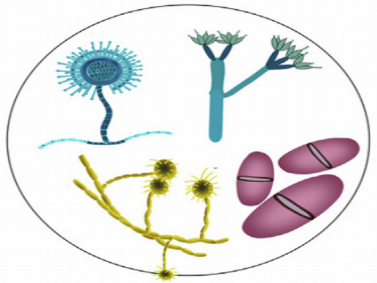


INTRODUCTION

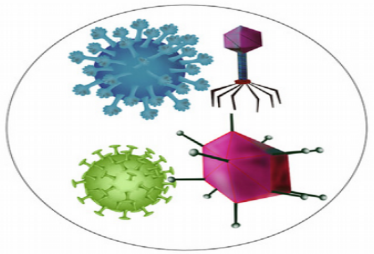
Bacteria



Fungi



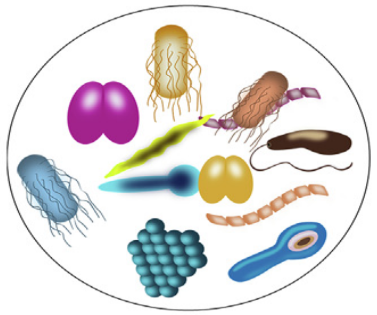
Virome



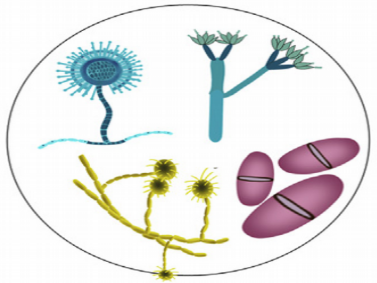
- Microbial communities consist of micro-organisms such as bacteria, virus and fungi.
- Micro-organisms have built robust ecosystems in various environments such as soil, sea water and various human organs.
- Microbiome has been associated with conditions such as obesity, colorectal cancer and inflammatory bowel disease.
- Understanding microbial interactions and relationships may provide great insights in restoring a healthy microbial community.

INTRODUCTION

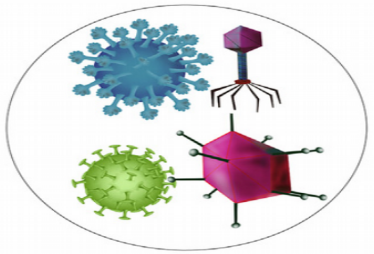
Bacteria



Fungi



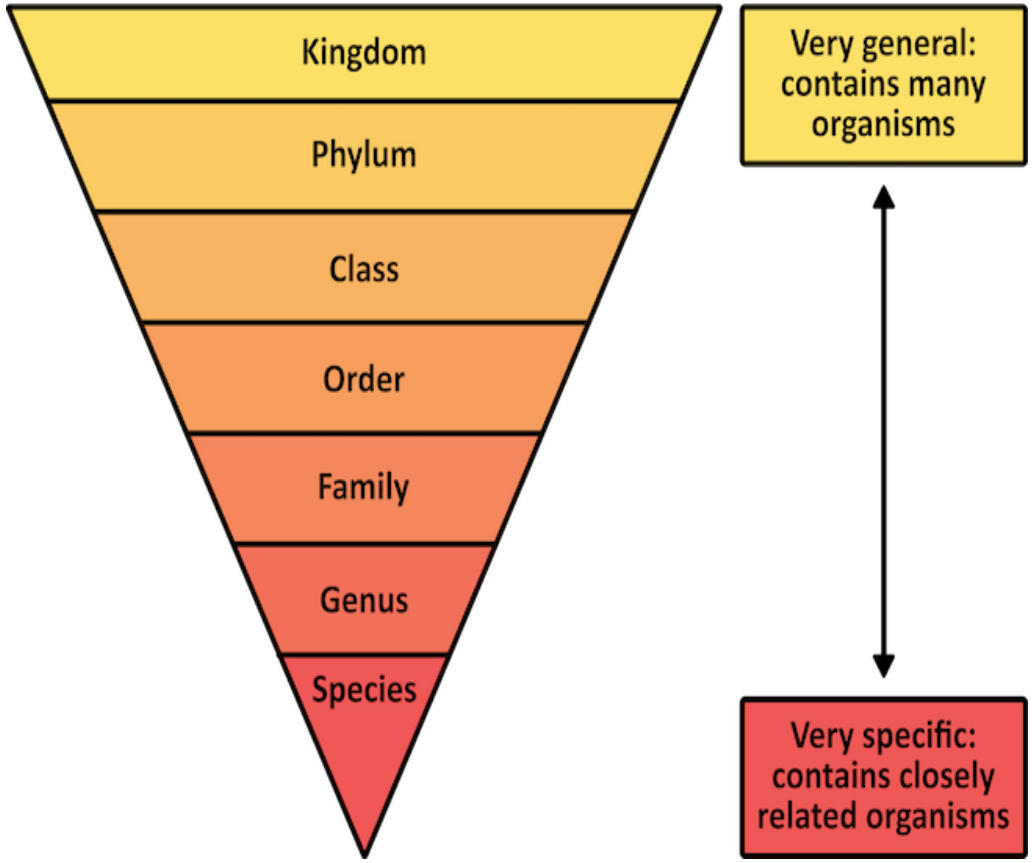
Virome



- Studies have revealed several cases of complex dynamics such as interactions between micro-organisms and their effects on the host.
- Network-based analytical approaches have helped in the study of systems with complex microbial interactions.
- Two of the most popular methods for determining microbial interactions are the Pearson Correlation method and the Least Absolute Shrinkage and Selection Operator (LASSO) method.

BACKGROUND

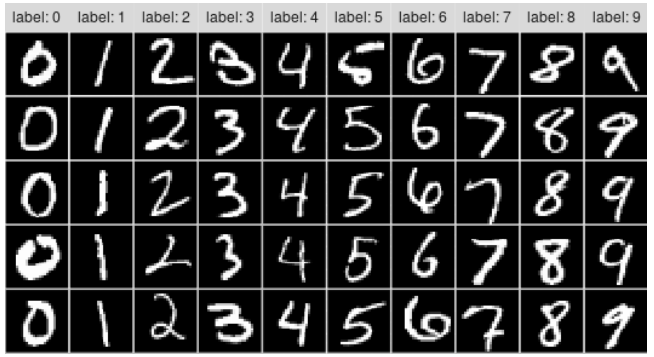
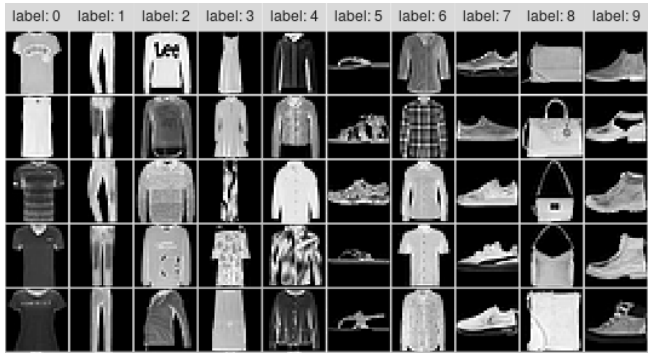
King Philip Can Only Find Green Socks



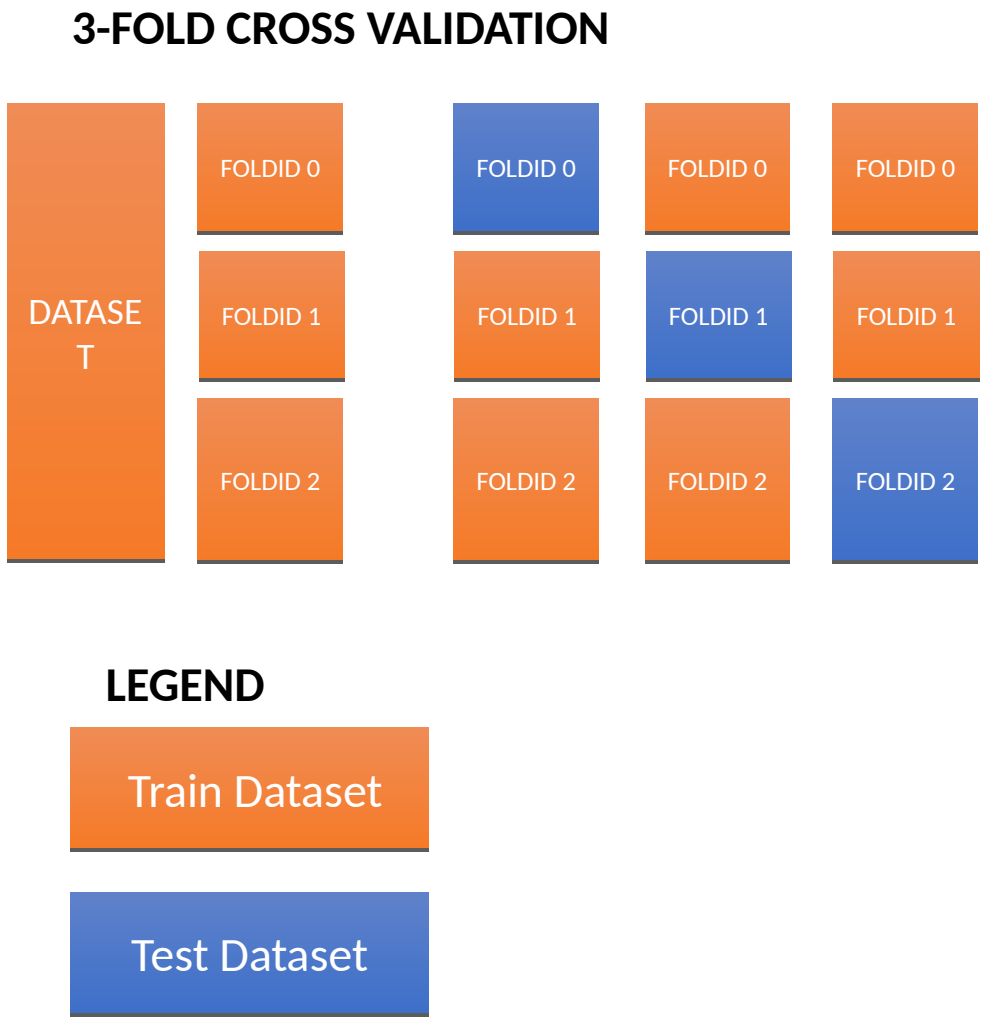
Animal Example	Taxonomic Rank	Bacteria Example
Animalia	**Kingdom**	Bacteria
Chordata	**Phylum**	Proteobacteria
Mammalia	**Class**	Gamma proteobacteria
Primate	**Order**	Vibrionales
Hominidae	**Family**	Vibrionaceae
Homo	**Genus**	Vibrio
sapiens	**Species**	*Vibrio Cholerae*
Human	**Common Name**	Vibrio Cholerae

Ref: https://www.mometrix.com/academy/biological-classification-system

Supervised machine learning algorithm returns a different function, depending on the train data set

Learning Algorithm	Train data	Learned function	Predictions on test data
Learn(		) → g	g( ) = 0 g( ) = 1 g( ) = 1
Learn(		) → h	h( ) = 0 h( ) = 0 h( ) = 1

K-FOLD CROSS VALIDATION ALGORITHM



- K- fold cross validation algorithm is a technique used to train model parameters and compare prediction accuracy
- Learning the parameters of a prediction function and testing it on the same data leads to over-fitting (When a model does very well on testing data but does poorly new or unseen data).
- To avoid over-fitting, it is a common practice in Machine Learning to hold out some of the data as the test set.

PEARSON CORRELATION MODEL

- **Pearson's correlation coefficient** is the standard tool to infer a network through correlation analysis among all pairs of OTU (Operational Taxonomic Unit) samples.
- It is a number that ranges from –1 to 1 and measures the strength and direction of the relationship between two variables.
- Pearson Coefficient Formular :

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

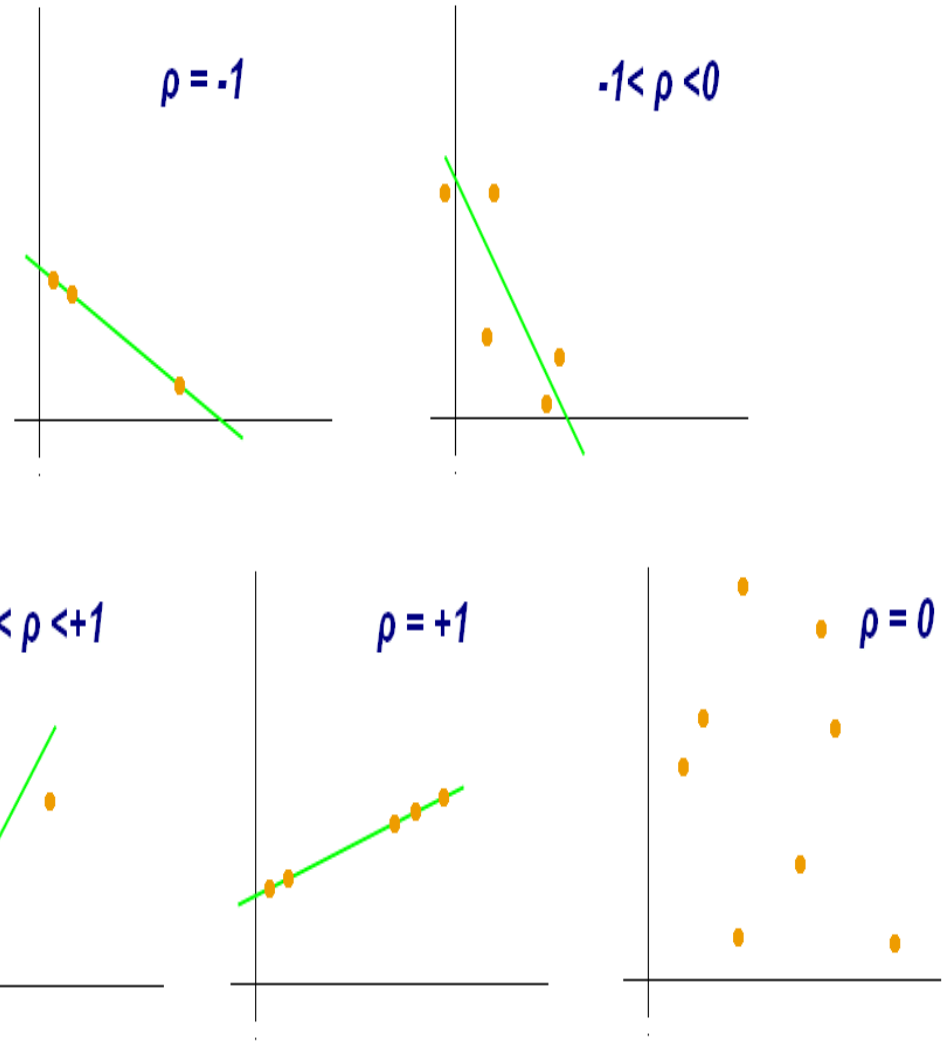
r = Pearson Correlation Coefficient

x_i = x variable samples

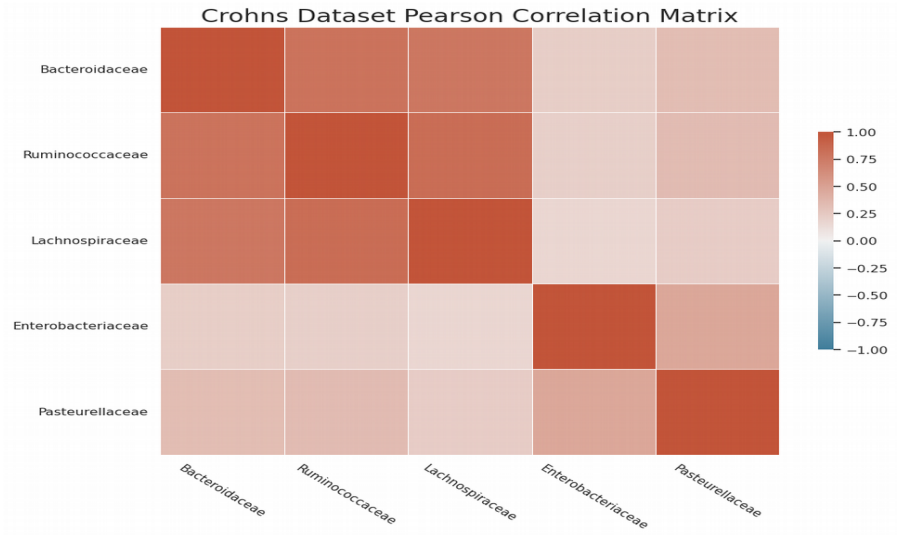
y_i = y variable sample

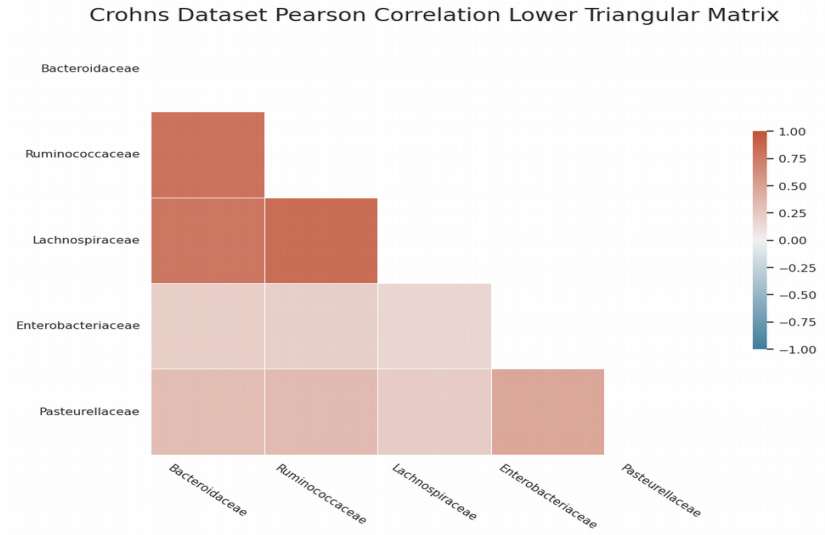
$\bar{x}$ = mean of values in x variable

$\bar{y}$ =mean of values in y variable

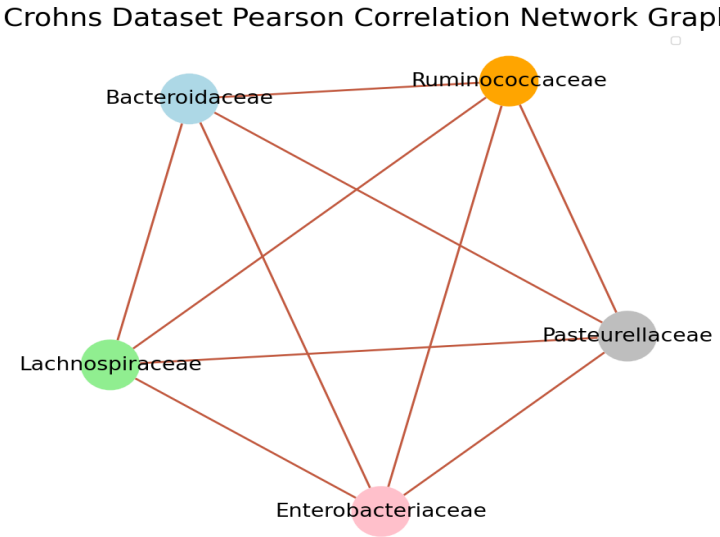


PEARSON CORRELATION MATRIX

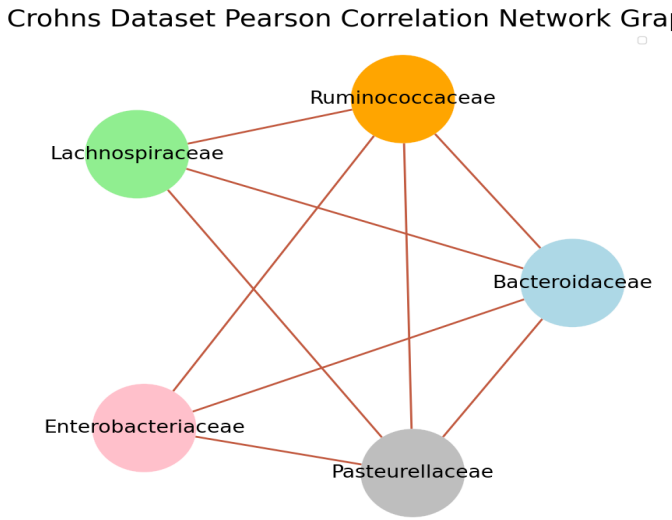




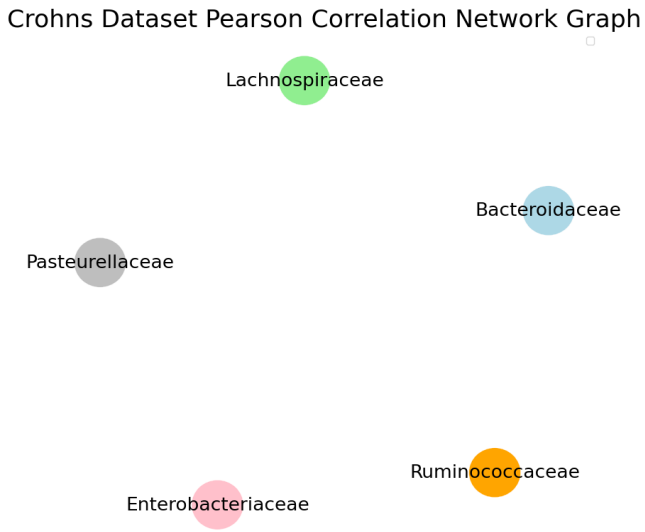
NETWORK GRAPH



Threshold = 0



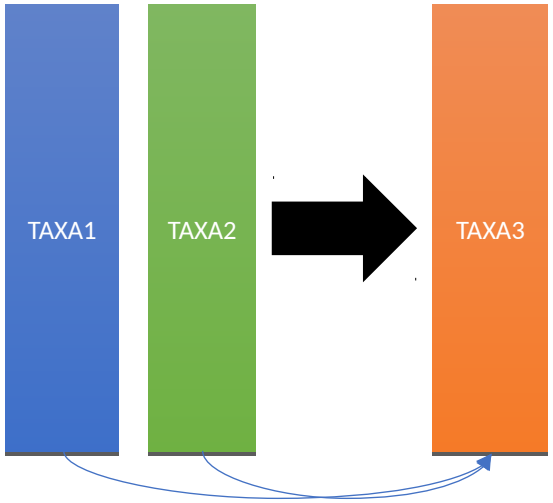
Threshold = 0.2



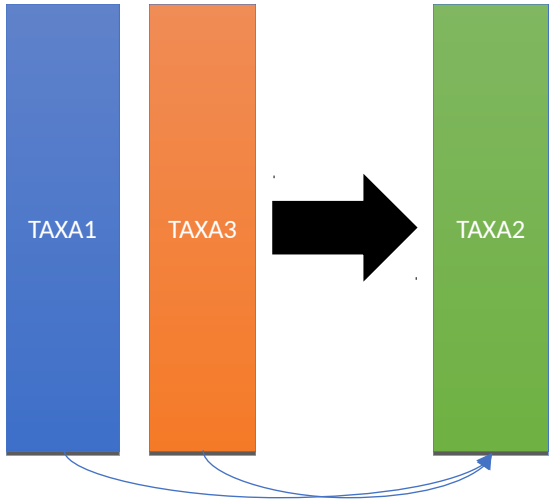
Threshold = 1

MULTI-COLUMN TEST ERROR FOR PEARSON CORRELATION

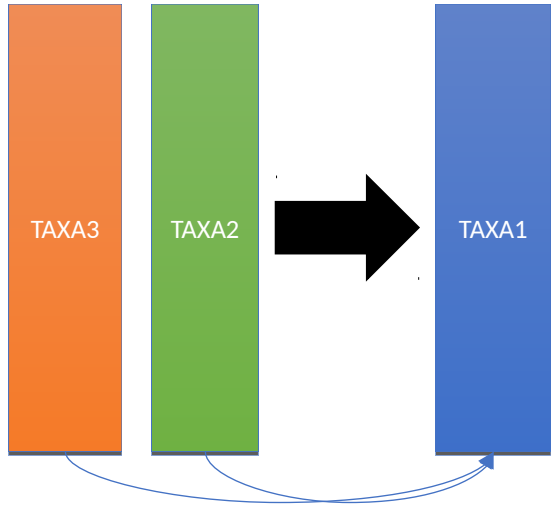
Taxa 3 is predicted using Taxa 1 and Taxa 2



Taxa 2 is predicted using Taxa 1 and Taxa 3



Taxa 1 is predicted using Taxa 3 and Taxa 2



- The mean of each of the taxa predictions are computed resulting in one big taxa prediction column.
- The **Mean Squared Error** of the prediction with respect to the actual taxa labels is computed.
- The average of the **Mean Squared Error** is noted for each number of samples used.

LASSO REGRESSION MODEL

- The LASSO is also known as Least Absolute Shrinkage and Selection Operator. It is a form of linear regression which uses L1 regularization technique and variable selection to increase the accuracy of prediction.

- L1 regularization adds a penalty which causes the regression coefficient of the less contributing variable to shrink to zero or near zero.

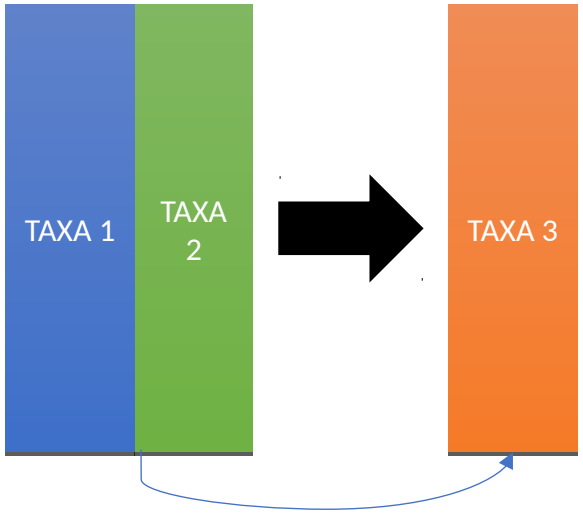
➢ Loss function:

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2+\lambda\sum_{j=1}^{p}\left|\hat{\beta}_j\right|$$

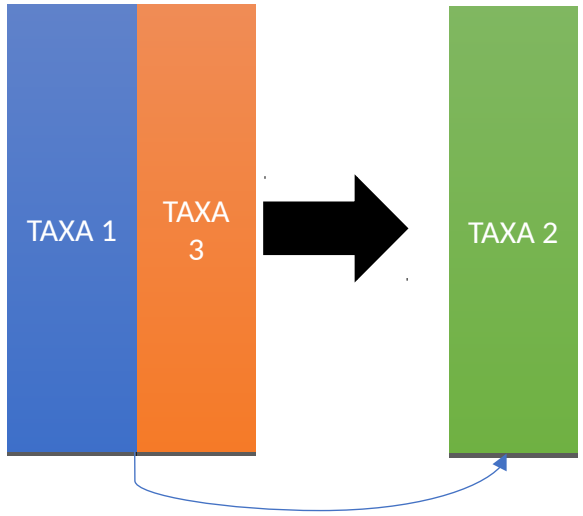
- ƛ is a tuning parameter (amount of shrinkage). When ƛ = 0, no parameters are eliminated.

MULTI-COLUMN TEST ERROR FOR LASSOCV MODEL

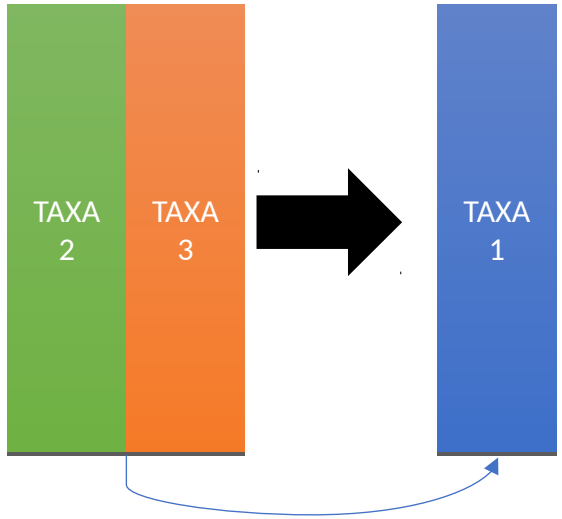
Taxa 3 is predicted using Taxa 1 and Taxa 2



Taxa 2 is predicted using Taxa 1 and Taxa 3



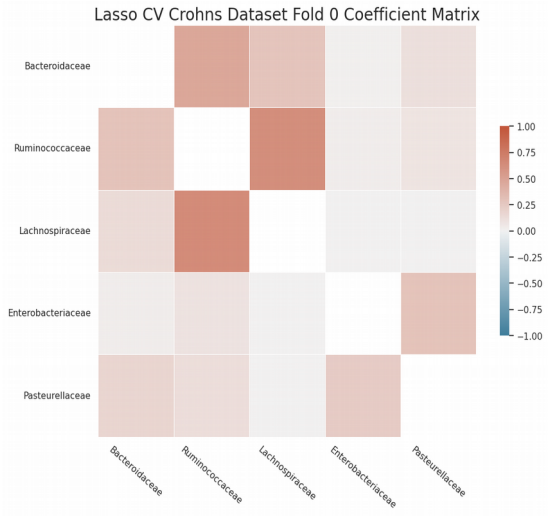
Taxa 1 is predicted using Taxa 2 and Taxa 3

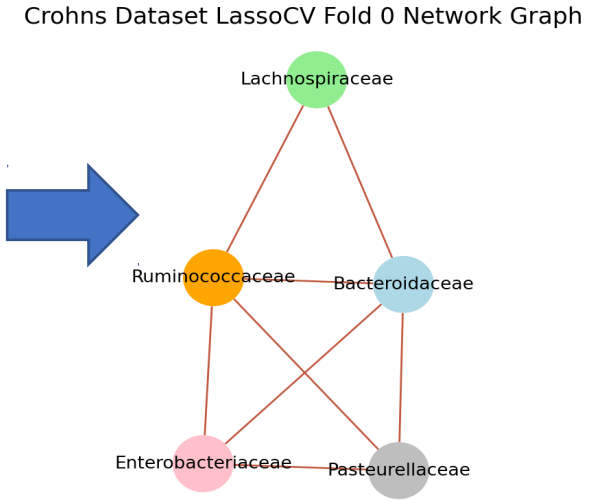


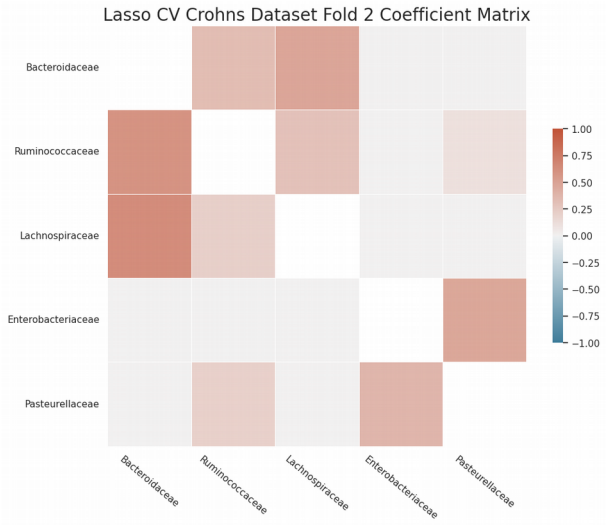
- The **Mean Squared Error** of predicting each of the taxa columns is computed.
- The average of the **Test Error** is recorded.

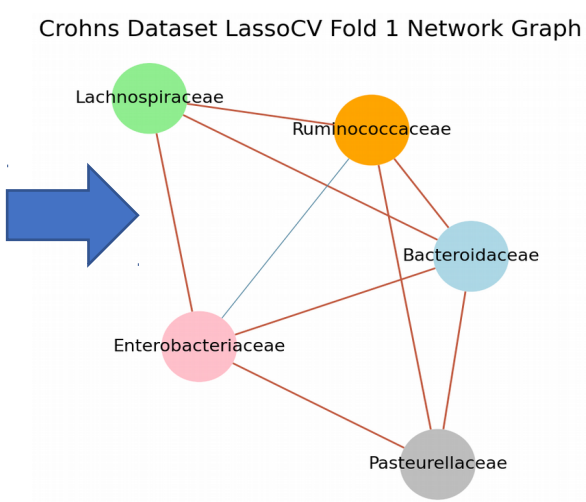
- The experiment is performed for different number of samples in the datasets.
- The coefficients of the model with the optimum **alpha** is recorded.

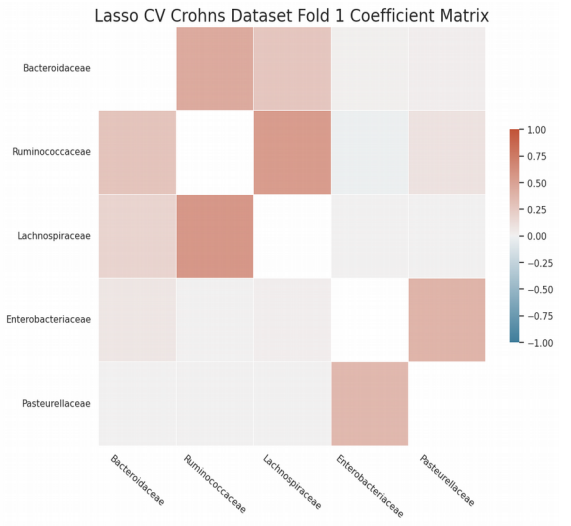
LASSOCV COEFFICIENT MATRIX

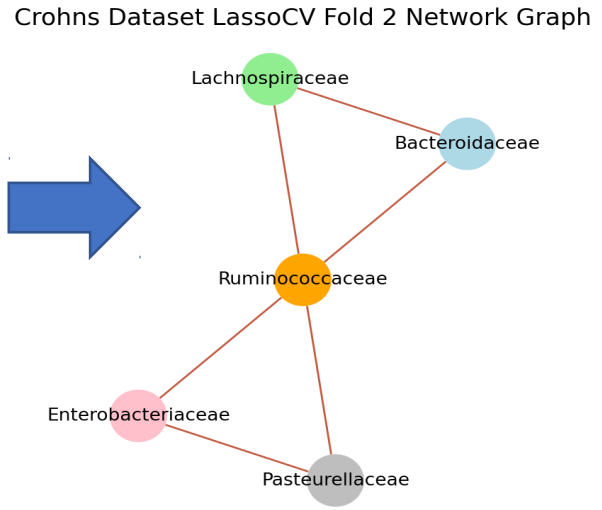








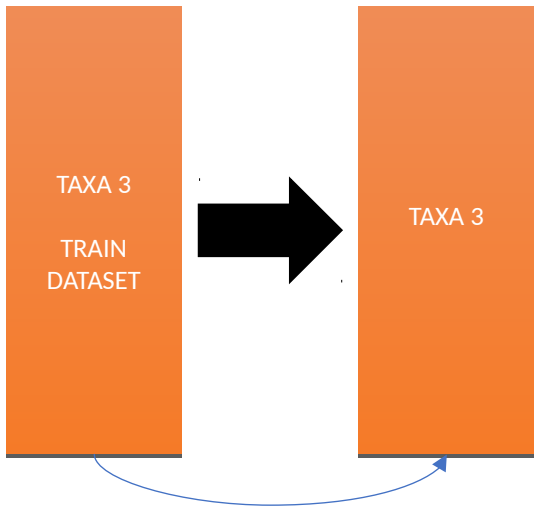




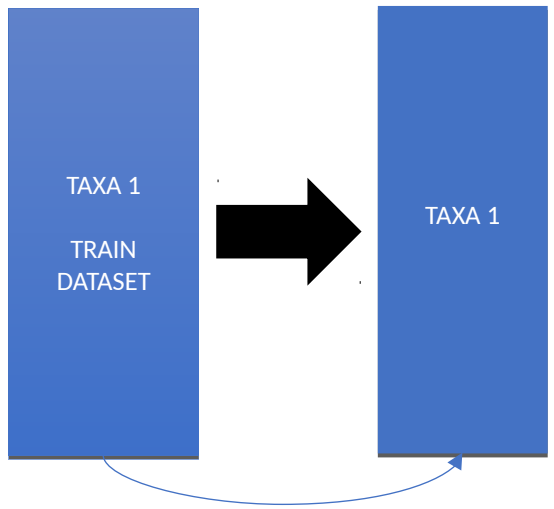
- Each of the FoldID gives a different coefficient matrix hence different network graph.
- All of the network graphs are valid for each fold.

MULTI-COLUMN TEST ERROR FOR FEATURELESS/BASELINE

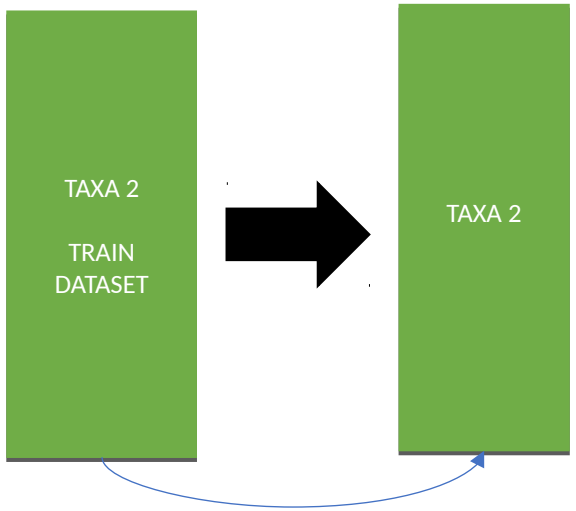
Taxa 3 is predicted using the mean of Taxa 3 train column



Taxa 1 is predicted using the mean of Taxa 1 train column



Taxa 2 is predicted using the mean of Taxa 2 train column



- The mean of each of the taxa predictions are computed resulting in one big taxa prediction column.
- The **Mean Squared Error** of the prediction with respect to the actual taxa labels is computed.
- The average of the **Mean Squared Error** is noted for each number of samples used.

CONCLUSION

It can be inferred from the graph on the previous slides that:

- The test error decreases as the number of samples in the dataset increases.

- The LASSOCV Model does better than the Pearson Correlation Algorithm.

REFERENCES

- https://www.liebertpub.com/doi/10.1089/cmb.2021.0406
- https://smnh.tau.ac.il/en/interactions-among-living-organisms/
- https://scikit-learn.org/stable/modules/cross_validation.html
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7768662/
- https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004226
- https://doi.org/10.1128/mSystems.00124-19
- https://www.thoughtco.com/commensalism-definition-and-examples-4114713
- https://www.sciencedaily.com/releases/2018/05/180515092931.htm

THANK YOU

ANY QUESTIONS?