

Iterative Random Forest Based High Performance Computing Methods Applied to Biological Systems and Human Health

A Dissertation Presented for the
Doctor of Philosophy
Degree

The University of Tennessee, Knoxville

Angelica M. Walker

May 2022

© by Angelica M. Walker, 2022
All Rights Reserved.

To my parents, thank you for everything

Para mis padres, gracias por todo

Acknowledgments

I would like to express my deepest gratitudes towards all of the people who have helped me along the way.

First I would like to thank my fellow lab mates Dr. Ashley Cliff, Jonathon Romero, Dr. David Kainer, Dr. Jared Streich, Matthew Lane, Anna Furches, Doug Hyatt, Dr. Kyle A. Sullivan, and Dr. Hari Chhetri. Thank you for your support and advice in putting this dissertation together. You all truly helped me through the day to day work. You all have lent an ear when I needed to discuss, eyes when I needed my work checked, answered my never ending questions, and so much more that I could never repay you. This document and the science within it would not have been possible without your efforts.

I would next like to thank my doctoral committee Dr. Daniel Jacobson, Dr. David Kainer, Dr. Michela Taufer, and Dr. Michael Langston. Thank you for making time for me in your busy schedules and providing invaluable feedback with this dissertation.

I would like to thank my family for being so supportive of my work, even though they didn't understand a word I said when I talked about it. Thank you for listening to me drone on and on with the boring stuff.

I would like to thank all of the friends who have supported me in my academic career, especially while I was stumbling through my bachelors. There's too many of you to list, but I am so appreciative of every one of you, especially Felipe Camelo.

I would like to thank the instructors who have inspired me to pursue a career in STEM, predominantly Mr. David Smith and Mr. Terry Barchfeld. Thank you for making calculus and physics seem so exciting and approachable all those years ago.

I would also like to thank my dog Riddle, who sat next to me as I typed every line of code and every word of this document as I worked from home throughout the pandemic. I'm glad that she never talked back in all the hours alone we spent together.

Finally, I would like to extend my sincere gratitude towards my fiancé, life partner, and love of my life Alexander J. Kukay. Thank you for understanding and experiencing the highs and lows of completing our doctorates together. Thank you for all of your feedback, support, and keeping me whole. This truly would not have been possible without your support. I'm certainly going to miss our tiny home office that we shared in Knoxville as we move on to bigger things in the future.

Funding provided by The BioEnergy Science Center (BESC) and The Center for Bioenergy Innovation (CBI). The U.S. Department of Energy Bioenergy Research Centers are supported by the Office of Biological and Environmental Research in the DOE Office of Science. This work was funded by the Genomic Science Program of the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research (BER) as part of the Secure Ecosystems Engineering and Design research program in the Secure Biosystems Design Scientific Focus Area (SFA). This research used resources of the Oak Ridge Leadership Computing Facility and the Compute and Data Environment for Science at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

Abstract

As technology improves, the field of biology has increasingly utilized high performance computing techniques to analyze big data and provide insights into biological systems. A reproducible, efficient, and effective method is required to analyze these large datasets of varying types into interpretable results. Iterative Random Forest (iRF) is an explainable supervised learner that makes few assumptions about the relationships between variables and is able to capture complex interactions that are common in biological systems. This forest based learner is the basis of iRF-Leave One Out Prediction (iRF-LOOP), an algorithm that uses a matrix of data to produce all-to-all predictive networks. This dissertation includes a validation of the improved performance of iRF over the industry standard of Random Forest, using synthetic and empirical data from various organisms. Additionally, this dissertation includes the use of iRF to create a predictive model of COVID-19 outcomes using environmental features at the county level in the U.S. This dissertation also includes a whole systems biology study in which an improved iRF-LOOP pre-processing pipeline Divide-Test-Integrate is used to produce new gene-to-gene predictive expression networks for a multiplex network study of the model organism *Saccharomyces cerevisiae* using seed genes of interest from *Septoria musiva*.

Table of Contents

1	Background	1
1.1	Introduction	2
1.2	High Performance Computing	3
1.2.1	Supercomputing in Biology	4
1.2.2	Parallel Computing and Supercomputer Architectures	4
1.2.3	Available Computing Clusters	5
1.3	Network Biology	5
1.3.1	Graph Theory in Biology	5
1.3.2	Network Inference Tools	6
1.3.3	Multiplex Networks	9
1.4	Tree-based methods	10
1.4.1	Decision Trees and Ensemble Methods	10
1.4.2	Random Forest	12
1.4.3	Iterative Random Forest	13
1.5	Data Cleaning and Pre-processing	13
1.5.1	Correlated Features	14
1.5.2	Batch Correction	14
1.5.3	Outliers	15
1.6	Post-Processing Networks	16
1.6.1	Network Evaluation	17
1.6.2	Network Thresholding	21
1.6.3	Network Clustering	22

1.7	Systems Biology	24
1.7.1	Model Organisms	25
1.7.2	Layer Types	26
1.8	Relevant Biological Techniques	28
1.8.1	Genome-Wide Association Study	28
1.8.2	Microarray Expression Datasets	29
1.9	Overview of COVID-19	29
1.9.1	Comorbidities and Environmental Predictors of COVID-19	30
1.9.2	Measurements of Severity	31
1.9.3	Issues Surrounding COVID-19 Studies	31
1.10	Conclusion	32
2	Evaluating the Performance of Random Forest and Iterative Random Forest-Based Methods when Applied to Gene Expression Data	33
2.1	Abstract	34
2.2	Introduction	34
2.3	Methods	37
2.3.1	RF-LOOP and iRF-LOOP	38
2.3.2	RF-subLOOP and iRF-subLOOP	38
2.3.3	Dream Challenge Datasets	40
2.3.4	<i>A. thaliana</i> and <i>P. trichocarpa</i> PEN inference	41
2.3.5	Performance Evaluation	41
2.3.6	Random Intersect Trees (RIT)	45
2.3.7	Computational Resources	46
2.4	Results and Discussion	46
2.4.1	Replacing RF with iRF Improves Overall DREAM Scores	46
2.4.2	iRF-LOOP on <i>A. thaliana</i> and <i>P. trichocarpa</i> Improves Early nDCG and AUnDCG	48
2.4.3	iRF Boosts Important Edges and Improves Feature Selection	48

2.4.4	PEN Edges and RIT Sets Capture Complex Biological Relationships in <i>A. thaliana</i>	53
2.4.5	Replacing RF with iRF Captures More Members of Metabolic Reaction Pathways	56
2.5	Conclusion	59
3	A Predictive Model of COVID-19 County Level Case Fatality Rate	61
3.1	Abstract	62
3.2	Introduction	62
3.3	Methods	64
3.3.1	Environmental Data and Data Pre-Processing	64
3.3.2	COVID-19 Y Vector	66
3.3.3	iRF Predictive Models	67
3.3.4	iRF-LOOP Network Creation	69
3.3.5	Computational Resources	70
3.4	Results and Discussion	70
3.4.1	Climate and Demographic Features are Important in Predicting Overall CFR	70
3.4.2	Climate Features are Important in Predicting Monthly CFR	72
3.4.3	Vitamin D Absorption is Related to Seasonal Features and the Southern United States	76
3.5	Conclusion	78
4	Applying Divide-Test-Integrate To Gene Expression Data Improves the Performance of iRF-LOOP	80
4.1	Abstract	81
4.2	Introduction	81
4.3	Methods	82
4.3.1	Network Inference	83
4.3.2	Divide-Test-Integrate	83
4.3.3	Dividing the Samples	85

4.3.4	Data Preprocessing	87
4.3.5	Performance Evaluation	87
4.3.6	Gene Ontology Semantic Similarity Network	90
4.3.7	Computational Resources	91
4.4	Results and Discussion	91
4.4.1	Supervised Division of Samples in DTI with iRF-LOOP Provides an Improvement Over Traditional iRF-LOOP	91
4.4.2	Unsupervised Division of Samples in DTI with iRF-LOOP Provides an Improvement Over Traditional iRF-LOOP	92
4.4.3	Unsupervised Division of Samples Kept Samples from Same Experi- ment within Same Cluster	95
4.4.4	iRF-LOOP with DTI Captures More Specific Gene-to-Gene Associations	96
4.5	Conclusion	99
5	Exploring <i>Septoria musiva</i> Gene Relationships Through a Multiplex Network of the Model Organism <i>Saccharomyces cerevisiae</i>	100
5.1	Abstract	101
5.2	Introduction	101
5.3	Methods	102
5.3.1	Random Walk with Restart	102
5.3.2	RWRtools	103
5.3.3	Geneset Refinement using Interacting Networks (GRIN)	104
5.3.4	Determination of Seed Genes	105
5.3.5	Biological Layers within the Multiplex	107
5.3.6	Gene Ontology Enrichment	111
5.3.7	Computational Resources	111
5.4	Results and Discussion	112
5.4.1	Multiplex of <i>S. cerevisiae</i> Genes is High Quality	112
5.4.2	Seed and GRIN Retained <i>S. cerevisiae</i> Genes are Trait Specific . . .	115
5.4.3	Genes Retained by GRIN May be Related to Fungal Infection	115

5.5 Conclusion	117
Conclusion	119
Bibliography	121
Appendices	144
A Chapter 2 Supplementary Information	145
B Chapter 3 Supplementary Information	154
C Chapter 4 Supplementary Information	164
D Chapter 5 Supplementary Information	168
Vita	177

List of Tables

2.1	This table depicts the overall scores for the DREAM4 <i>In Silico Size 100 Multifactorial</i> and DREAM5 <i>Network Inference</i> networks using the DREAM scoring system. The iRF-based algorithms outperform all three RF-based algorithms. This table also shows that simply increasing the number of trees in an RF-based model to match the total number of trees used in an iRF-based model does not account for the overall score increase seen in iRF-LOOP. Raw AUPR and AUROC values as well as their p-values can be found in Supplementary Tables 1-4.	47
2.2	This table depicts the top five ranked edges from the PEN created using iRF-LOOP on <i>A. thaliana</i> expression data. None of the edges were true positive edges in the gold standard network, but have some existing biological relationship. This table shows that PENs produced with iRF-LOOP capture relationships beyond simple gene regulation.	54
2.3	This table depicts a sample of three highly prevalent sets determined using RIT on the resulting paths from iRF-LOOP on <i>A. thaliana</i> expression data. All three of these sets have been found to have some known biological relationship. None of these sets were found in the gold standard network, while some were found in the iRF-LOOP PEN as edges with low importance scores. This shows that RIT combined with iRF-LOOP can be used to discover or validate gene to gene relationships in sets.	55

5.1	This table contains the number of unique SNPs that were retained in each GWAS method and <i>S. musiva</i> trait. The FDR P-value threshold for the SNPs was < 0.4	106
5.2	This table depicts the number of unique SNPs retained after combining and thresholding the GWAS results for each trait, as well as the number of unique <i>S. musiva</i> genes and unique <i>S. cerevisiae</i> genes after ortholog mapping. . . .	106
5.3	This table depicts the total number of unique edges and unique genes (or nodes) for each layer included within the multiplex.	108
1	This table contains the AUPR values for the networks within the DREAM4 and DREAM5 challenges.	145
2	This table contains the AUROC values for the networks within the DREAM4 and DREAM5 challenges.	146
3	This table contains the AUPR p-values for the networks within the DREAM4 and DREAM5 challenges.	147
4	This table contains the AUROC p-values for the networks within the DREAM4 and DREAM5 challenges.	148
5	This table depicts a sample fourth order gene set that were discovered using RIT on the resulting paths from iRF-LOOP on <i>A. thaliana</i> expression data. This set had a prevalence of 0.331. This set contains a mix of both known and unknown gene to gene relationships.	152
6	This table depicts a sample fifth order gene set that were discovered using RIT on the resulting paths from iRF-LOOP on <i>A. thaliana</i> expression data. This set had a prevalence of 0.153. This set contains a mix of both known and unknown gene to gene relationships.	153
7	Static climate features details.	155
8	Monthly climate features details.	156
9	Census features details (table one of four).	157
10	Census features details (table two of four).	158
11	Census features details (table three of four).	159
12	Census features details (table four of four).	160

13	Policy features details (table one of two).	161
14	Policy features details (table two of two).	162
15	Number of samples in each iRF model and number of features in each class used.	163
16	Number of samples and genes in each group of samples in <i>S. cerevisiae</i> . (Table one of two)	164
17	Number of samples and genes in each group of samples in <i>S. cerevisiae</i> . (Table two of two)	165
18	Number of samples and genes in each group of samples in <i>A. thaliana</i>	166
19	GO Enrichment for GRIN retained <i>S. cerevisiae</i> genes for the trait number of cankers.	169
20	GO Enrichment for GRIN retained <i>S. cerevisiae</i> genes for the trait cankers per centimeter (table one of two).	170
21	GO Enrichment for GRIN retained <i>S. cerevisiae</i> genes for the trait cankers per centimeter (table two of two).	171
22	GO Enrichment for GRIN retained <i>S. cerevisiae</i> genes for the trait disease severity.	172
23	GO Enrichment for GRIN retained <i>S. cerevisiae</i> genes for the trait plant height (table one of two).	173
24	GO Enrichment for GRIN retained <i>S. cerevisiae</i> genes for the trait plant height (table two of two).	174
25	GO Enrichment for GRIN retained <i>S. cerevisiae</i> genes for all four traits (table one of two).	175
26	GO Enrichment for GRIN retained <i>S. cerevisiae</i> genes for all four traits (table two of two).	176

List of Figures

1.1	This figure depicts a binary decision tree. The root node contains all five samples, and the samples are split into two child nodes based on feature one. The two child nodes are then split into two more child nodes to create the full tree. In this scenario, the most important feature is feature one since it is picked first to split the data, and the least important feature is feature three since it is never selected.	11
1.2	This figure depicts the confusion matrix (a) that is used to score the quality of resulting networks from network inference tools. The gold standard network contains true (1) and false (0) edges, and the predicted network contains predicted truth (1) and predicted false (0) edges. The confusion matrix determines whether the predicted edge is a false positive, true positive, true negative, or false positive. The Precision-Recall curve (b) and Receiver Operating Characteristic curve (c) are determined using the values from the confusion matrix, and used to visually determine the performance of the networks being scored.	18

2.1	This diagram shows the process of creating a transcription factor-to-all other genes GRN using the iRF-subLOOP algorithm. The overall algorithm contains two types of iRF runs. The first run type (a) creates a TF-to-NonTF network by partitioning the expression matrix into TF and NonTF parts. The TFs are then used as the features predicting each of the individual genes in the NonTF matrix. The second run type (b) creates a TF-to-TF network, similar to iRF-LOOP but instead of all genes being used as the input features, only the TFs are used. The resulting networks from (a) and (b) are merged to create a TF-to-All GRN.	39
2.2	This figure depicts the nDCG scores for both <i>A. thaliana</i> and <i>P. trichocarpa</i> for both GENIE3 and iRF-LOOP as k , the number of edges scored, increases. The maximum k for each organism is equal to the size of the gold standard network. For both <i>A. thaliana</i> and <i>P. trichocarpa</i> , the early nDCG is higher in iRF-LOOP than GENIE3, and the overall AUnDCG for both organisms is also higher in the iRF-based algorithm. For <i>A. thaliana</i> , AUnDCG for iRF-LOOP and GENIE3 were 91.897 and 52.694 respectively. For <i>P. trichocarpa</i> , AUnDCG for iRF-LOOP and GENIE3 were 18.672 and 7.972 respectively. This suggests that iRF-LOOP outperforms GENIE3.	49
2.3	In this figure, the blue line depicts the truly observed signal-to-noise ratio for each network as the number of iterations in iRF increases. As the iterations increase, a number of edges are dropped from consideration due to their importance scores equaling zero. To confirm that the improvement in signal-to-noise ratio is not due to simply thresholding the networks, the RF network is thresholded to match the network size of the iRF network at each iteration and is depicted as the orange dashed line. For both <i>A. thaliana</i> and <i>P. trichocarpa</i> , the observed signal-to-noise ratio in iRF is greater than both what is expected from random and if the RF network is simply thresholded. This shows that the unsupervised thresholding from iRF provides an improvement over a simple manual thresholding.	50

2.4	This figure shows the true positive edges, depicted as a square, for the top ranked 100 edges and 500 edges across each iteration for <i>A. thaliana</i> and <i>P. trichocarpa</i> respectively. For iteration 1 in <i>A. thaliana</i> , the first true positive edge is not discovered until rank 325. For iteration 1 in <i>P. trichocarpa</i> , the first true positive edge is not discovered until rank 965. The shifting of the true positive edges towards earlier rankings indicates the true positive edges moving up in the ranks of the final edge list and are more likely to be retained when thresholded.	52
2.5	This figure depicts the compounds, RXNs, and genes included in the pentose phosphate pathway in <i>A. thaliana</i> , and the genes that are recovered by GENIE3 and iRF-LOOP. The resulting PEN from GENIE3 was thresholded to match the same size as the iRF-LOOP PEN. A) Steps in the pathway encompassing the three RXNs according to AraCyc. B) All of the possible edges between the known genes in neighboring steps in the pathway. C) The edges recovered by iRF-LOOP. D) The edges recovered by GENIE3.	57
2.6	This figure depicts the compounds, and a selection of RXNs and genes that are included in the phenylpropanoid biosynthesis pathway in <i>P. trichocarpa</i> . This figure also depicts the genes that are recovered by GENIE3 and iRF-LOOP in the EC 6.2.1.12 and EC 2.3.1.133 RXNs within the phenylpropanoid biosynthesis pathway. For comparison, in this figure the GENIE3 PEN is thresholded to match the same network size as the PEN created by iRF-LOOP. A) The entire phenylpropanoid pathway in <i>P. trichocarpa</i> is shown with the two RXNs of interest shown in blue. B) All of the theoretically possible edges between the known genes associated with the two RXNs. C) The genes and edges recovered by iRF-LOOP. D) The genes and edges recovered by GENIE3.	58

3.1	This figure depicts the different categories of CFR Y vectors, as well as the seven individual iRF models that contain different groupings of features. The CFR vectors include the overall CFR between March 2020 and December 2020, the combined monthly CFR vectors that include the monthly CFR measurements for every county merged into one vector, and the ten individual monthly CFR vectors. The types of iRF models were separated into two categories, one which compared the models that only contain the class of interest, and one that compared the models that exclude one class or set of features at a time. The latter category also includes an all features included model to determine whether the model performs better or worse with the exclusion of the class of interest. In total, 48 iRF models were used in k -fold cross validation to produce comparable MAE scores and the top important features.	68
3.2	This figure contains (a) violin plots of MAE scores of the seven iRF models and (b) the top ten features for the model containing all features all predicting overall CFR between March 2020 and December 2020. The climate alone model had the best MAE compared to the census and policy alone models, while the climate excluded model performed the best when compared to the census excluded and policy excluded model. The climate excluded model also performed similarly to the all features included model. This implies while the climate features are required for a highly accurate model, the inclusion of the census features and policy features together also produces an accurate model. The top important features showcase some climatic seasonal features, but many of the features were demographic, capturing vulnerable populations, such as those in older age groups and those that are most likely to be protected by policy features such as banning visitors in nursing homes.	71

3.3	This figure contains (a) violin plots of MAE scores of the seven iRF models and (b) the top ten features for the model containing all features all predicting the monthly CFR between March 2020 and December 2020. The climate alone model had the best MAE compared to the census and policy alone model, while the climate excluded model performed the worst when compared to the all features model and the models where one feature set was excluded, reiterating that the climate features produce the most accurate models. Within the top important features include many climatic seasonal features, suggesting the the most important factors when predicting patient survivability is the time of year in which they contract infection.	73
3.4	This figure depicts the MAE scores for the iRF models predicting the windowed monthly CFR rates as they change over time. In (a), of the models containing climate alone, census alone, and policy alone the best performing was the climate alone model while the worst performing was the policy alone model. In (b), of the models containing all the features, climate excluded, census excluded, and policy excluded, the best performing model was the census excluded model while the remaining models performed similarly. Thus, the climate features are critical in creating accurate predictions for the windowed CFR iRF models.	75
3.5	This figure focuses on the relationships of vitamin D UV absorption light range in April with other features included in this analysis. The network in (a) shows that many of these features are seasonal climate features and features that are related to the southern region of the U.S. The map in (b) displays the feature of interest, showing that this light range is strongest in the southern states of Alabama, Arizona, California, Louisiana, Mississippi, and Texas. This suggests that this feature that is highly predictive of CFR is highly related to the seasonal climate and surrounding environment of this region.	77

4.1	This figure depicts the Divide-Test-Integrate (DTI) method, in which microarray samples are divided into like groups, the network inference algorithm is applied to each group to produce intermediate networks, and the intermediate networks are integrated into the final gene-to-gene network. This method reduces the bias of housekeeping genes that is common in the traditional method of concatenating all the samples into one matrix. The intermediate networks may also be tested to determine the quality of the network and determine if they are of a high enough quality to integrate into the final network.	84
4.2	This figure depicts the method of using the Bray-Curtis dissimilarity measure to evaluate whether samples for the same experiment clustered together after applying iRF-LOOP on the samples and MCL clustering the resulting network. If two clusters did not share samples from the same experiment, the Bray-Curtis value between the two clusters would be zero. If the samples from the same experiment were evenly divided among the two clusters, the Bray-Curtis value would be zero.	88
4.3	Figure 4.3a depicts the relationship between the quality of the divided group and the number of samples included in the X matrix. The bottom five groups were not included in the integrated final PEN, this cutoff is depicted by the dotted red line. These groups also contained less than 200 samples. Figure 4.3b depicts the nDCG curves for the three gene-to-gene networks. The PCC co-expression network had an AUnDCG score of 474.956, the traditional iRF-LOOP PEN had an AUnDCG score of 1222.944, and the iRF-LOOP with DTI PEN had an AUnDCG score of 1665.545. iRF-LOOP with DTI had the highest AUnDCG score and is the best performing network.	93

- 4.4 Figure 4.4a depicts the relationship between the quality of the divided group and the number of samples included in the X matrix. The bottom scoring group was not included in the final integrated PEN, this cutoff is depicted by the dotted red line. Like *S. cerevisiae*, this group also contained less than 200 samples. Figure 4.4b depicts the nDCG curves for the three gene-to-gene networks. The PCC co-expression network had an AUnDCG score of 30.021, while the traditional iRF-LOOP PEN had an AUnDCG score of 128.722, and the iRF-LOOP with DTI PEN had an AUnDCG score of 169.776. iRF-LOOP with DTI had the highest AUnDCG score and is the best performing network. 94
- 4.5 This figure depicts the Bray-Curtis dissimilarity measures between the 13 clusters of *A. thaliana* samples. All group pairs but one contain a dissimilarity measure greater than 0.8. This shows that the samples from the same experiment tend to cluster together, suggesting that the computationally intensive unsupervised division of samples may not be as efficient as a less computational intensive supervised method, such as text mining the abstracts for biological tags. 97
- 4.6 This figure depicts the GO semantic similarity edges, the traditional iRF-LOOP edges, and the iRF-LOOP with DTI edges between six *A. thaliana* genes. The GO semantic similarity edges are those that contain a weight above 0.5, which are associations that are more specific with the biological system. For DTI to reduce the bias of housekeeping genes, the iRF-LOOP with DTI PEN would capture more of these specific GO associations, which is evident in the network on the right. Of the top 10,000 edges from each PEN, traditional iRF-LOOP captures only 269 of these specific GO semantic similarity associations, while iRF-LOOP with DTI captures 396 of these edges. 98

5.1	This figure represents all of the edges that exist in the multiplex between the 33 genes with the GO annotation GO:0006887. This network is connected, meaning that there does not exist any edge that can be removed to result in two disconnected networks. This depicts how sufficiently the multiplex captures known biological truth between groups of genes.	113
5.2	This figure depicts some of the metrics of RWR Cross Validation. The ROC curves show how well RWR ranks the left-out genes for each of the five sets of genes with the same GO annotation. It performed considerably better than the random classifier. With PR however this observation is not as obvious except for in the early Recall and for the set of GO annotations with the fewest number of genes, GO:0006887 with 33 genes.	114
5.3	These venn diagrams showcase the lack of overlap of genes between sets that are related to different traits. After GRIN filtering, the only overlap occurs in number of cankers and cankers per centimeter, which is likely due to the fact that these two traits are highly correlated.	116
1	This figure depicts the nDCG scores for all of the DREAM4 and DREAM5 networks for both GENIE3 and iRF-LOOP as the number of edges scored, k , increases. The maximum k for each network depends on the number of true positive values in the corresponding gold standard network, ranging from 176 to 249 edges for the DREAM4 networks and 2,066 to 4,012 edges for the DREAM5 networks. For all networks except the DREAM5 Network 4, the AUnDCG for iRF-LOOP is higher than the AUnDCG for GENIE3, this suggests that iRF-LOOP outperforms GENIE3.	149
2	This figure depicts the true signal-to-noise ratio for each iteration of iRF-LOOP for the DREAM challenges as the blue line. The orange line is the corresponding GENIE3 network thresholded to match the same number of edges as the iRF-LOOP network, used to confirm that the unsupervised thresholding contributes to the improvement in signal-to-noise ratio. Unlike the larger empirical networks shown in Figure 2.3, these DREAM networks are too small to make a considerable difference when thresholding the networks.	150

3	This figure depicts the Precision-Recall curves and the corresponding AUPR values for each iteration in the networks used in this analysis. For many of the synthetic DREAM networks, the improvement in AUPR plateaus after 2 or 3 iterations, which may be due to the size of the networks. This may suggest that iterations improve RF, but larger networks are needed to verify this. The Precision-Recall curves themselves show minimal improvement after the second iteration for the empirical networks. However, the AUPR value increases as the number of iterations increases, even after the second iteration. Thus this shows that the addition of iterations to RF improves the AUPR values for the two empirical data sets used in this study.	151
4	This figure depicts the AUnDCG score for the integrated final PEN in DTI iRF-LOOP for varying values of D in the WS formula. The optimal value for D in both organisms is $D = 1$	167
5	This figure shows how the threshold for the STRING PPI network was decided. The elbow visibly at the experimental score threshold of 800 shows that the edges over 800 are of higher quality, and should be included in the multiplex network.	168

Nomenclature

AUnDCG	Area Under the normalized Discounted Cumulative Gain curve
AUPR	Area Under the Precision-Recall curve
AUROC	Area Under the Receiver Operating Characteristic curve
BECA	Batch Effect-Correction Algorithm
CART	Classification and Regression Tree
CFR	Case Fatality Ratio
CPU	Computational Processing Unit
DCG	Discounted Cumulative Gain
DREAM	Dialogue for Reverse Engineering Assessment and Methods
DTI	Divide-Test-Integrate
FPR	False Positive Rate
GENIE3	GENe Network Inference with Ensemble of trees
geTMM	gene length corrected Trimmed Mean of M-value
GO	Gene Ontology
GPU	Graphical Processing Unit
GRIN	Gene Refinement using Interacting Networks

GRN	Gene Regulatory Network
GWAS	Genome-Wide Association Study
HCS	Highly Connected Subgraphs
HPC	High Performance Computing
IDCG	Ideal Discounted Cumulative Gain
IFR	Infection Fatality Ratio
iRF	iterative Random Forest
iRF-LOOP	iterative Random Forest-Leave One Out Prediction
KEGG	Kyoto Encyclopedia of Genes and Genomes
LOE	Lines of Evidence
LOOP	Leave-One-Out-Prediction
MAE	Mean Absolute Error
MCL	Markov Clustering algorithm
MPI	Message Passing Interface
nDCG	normalized Discounted Cumulative Gain
OLCF	Oak Ridge Leadership Computing Facility
PCC	Pearson's Correlation Coefficient
PEN	Predictive Expression Network
PPI	Protein-Protein Interaction Network
RF	Random Forest
RF-LOOP	Random Forest-Leave One Out Prediction

RIT	Random Intersect Trees
RWR	Random Walk with Restart
SGD	Saccharomyces Genome Database
SNP	Single Nucleotide Polymorphism
STRING	Search Tool for the Retrieval of Interacting Genes
TAIR	The Arabidopsis Information Resource
TPR	True Positive Rate
WS	Weighted Sum
X-AI	Explainable Artificial-Intelligence

Chapter 1

Background

1.1 Introduction

Since the turn of the century, the field of biology has increasingly utilized high-throughput measurement techniques to provide insights into biological systems. These systems contain various types of 'omics data, including the genome, transcriptome, proteome, metabolome, microbiome and others that create the "ome-ome." Reproducible, efficient, and effective methods are required to integrate and analyze these varying large datasets into interpretable results. Thus, High Performance Computing (HPC) is becoming more common in the field of biology in order to analyze these larger datasets.

A common method to explore new biological relationships is to represent these relationships as networks, where the biological features, such as genes or proteins, are represented as nodes and the edges between nodes represent the presence of the relationship between the two features. There are multiple tools that can be used to infer these networks based on observations across multiple samples, such as correlation-based methods like Pearson's Correlation Coefficient and Random Forest-based methods like GENE Network Inference with Ensemble of trees (GENIE3) [84]. A collection of individual networks derived from different types of information can be used to create a multiplex network that can be traversed with network exploration algorithms such as Random Walk with Restart [172].

The algorithm that will be the focus of this dissertation is iterative Random Forest (iRF) [9]. The iRF algorithm is a tree-based method that has many applications. It can function as a network inference tool through iRF-LOOP [27], and can also predict unknown values of the chosen dependent variable to produce accuracy metrics through k -fold cross validation.

Before a network inference tool can be applied to a biological dataset, there are pre-processing steps that must be taken to prepare the data appropriately. For example, highly correlated features must be removed, batch effects from different experiments must be corrected, and outliers must be handled appropriately. Furthermore, once the networks have been created, they must be post-processed before use in downstream analyses. The quality of the network is determined using a known gold standard network, or a known network of experimentally validated edges. Usually, thresholds are applied to the edges of a network before further use to capture the highest quality edges. Network clustering may also be used

to extract topologies from a network that indicate strong interactions among sets of features or nodes.

Some network layers are inferred with the use of similarity metrics, such as co-expression networks based on gene expression data analyzed using a correlation tool, predictive expression networks using machine learning tools, or phylogenetic networks using sequence similarity. Other network types are based on direct experimental evidence such as protein-protein interaction networks, or curated information from the literature such as metabolic networks, co-citation networks, or gene regulatory networks. Some organisms are studied as model organisms, thus providing a wealth of knowledge in the field of biology. Biological knowledge obtained from these model organisms can be applied to other organisms that cannot be studied as easily due to being difficult to cultivate or being unethical to study experimentally, such as humans.

Provided in this review is an overview of two relevant biological techniques that are used in this dissertation, the Genome-Wide Association Study and the production of microarray expression datasets. This review also contains an overview of COVID-19, including the known comorbidities and environmental predictors of COVID-19 and the methods used to measure the severity of disease within a population.

1.2 High Performance Computing

The field of biology is constantly improving the tools and techniques used to measure data, which in turn forces researchers to improve the ways in which these datasets are analyzed. As data becomes larger, it becomes increasingly difficult to analyze the data by hand or using spreadsheets and the need for more sophisticated computational approaches becomes apparent. In order to use these tools, a basic understanding of High Performance Computing (HPC) and compute system architecture is necessary. Outlined below is a basic overview of supercomputing and its role in biology.

1.2.1 Supercomputing in Biology

As the instruments used to capture data in biological experiments improve, the resulting datasets become too large to analyze using traditional methods. While supercomputing techniques have been popular to analyze biological data since the 1980's [185], the recent data explosion in the field of biology has forced researchers to use supercomputers in every day research. This data explosion began in mid 2000's with the dramatic reduction in the cost of DNA and RNA sequencing (genomes, transcriptomes, microbiomes) as well as improvements in chromatography-coupled mass-spectrometry (proteomes and metabolomes) [123]. HPC systems and big data analysis techniques are increasingly prevalent in various sub-fields in biology, such as drug discovery [188], molecular dynamics [148], and statistical phylogenetics [7].

1.2.2 Parallel Computing and Supercomputer Architectures

Parallel computing is a type of computing where a large task is broken into separate tasks that are run concurrently on separate processors on a single compute system. Some tasks are simple to parallelize, often called an *embarrassingly parallel problem*, while others require more overhead to distribute tasks across the different threads and nodes within a compute system. The process of task distribution across nodes has been standardized as Message Passing Interface (MPI) [26], and contains two official implementations, MPICH [65] and OpenMPI [54].

Supercomputers, also known as High Performance Computers, are commonly used in big data studies due to the ability to parallelize large computational problems into smaller more manageable subsets. In general, HPC systems contain hundreds to thousands of nodes that contain either Computational Processing Units (CPUs), Graphical Processing Units (GPUs), or a combination of the two. CPUs have one or multiple cores that contain multiple threads that handle simple arithmetic and input/output streams. GPUs, originally used to produce images for displays on consumer devices, have recently become popular in parallel computing. The increase in use is due to their efficiency in mathematical calculations in a smaller, more compact space that requires less infrastructure than a CPU.

1.2.3 Available Computing Clusters

There are two HPC computing clusters that will be used extensively in this dissertation, Summit and Andes. Both are located at Oak Ridge National Laboratory, as a part of the Oak Ridge Leadership Computing Facility. Summit is an IBM system that contains 4,600 nodes, each with two POWER9 processors and six NVIDIA Tesla V100 GPUs. Each POWER9 processor contains 22 cores that contains 4 threads each for a total of 176 threads per node. Each node contains 512 GB of DDR4 memory. Andes contains two partitions, one with 704 CPU compute nodes and one with 9 GPU nodes. Each compute node contains two AMD EPYC 7302 CPUs per node with 16 cores per CPU for a total of 32 cores per node and 256 GB of memory per node. Each GPU node contains two Intel Xeon E5-2695 GPUs with 14 cores per GPU for a total of 28 cores per node and 1 TB of memory per node.

1.3 Network Biology

Networks or graphs are commonly used to depict biological interactions within a biological system. There are many tools that can be used to infer these networks using measured biological data. The final networks may then be combined into a multi-layer network that can be explored to find new biological relationships and explain previously unknown interactions.

1.3.1 Graph Theory in Biology

In graph theory, a graph contains vertices, or nodes, that are connected by edges [14]. These edges can be undirected or directed. Directed edges describe an orientation of the edge, in which one node is influencing another node. These directed edges are depicted with an arrow from the source node to the sink node. These edges may be unweighted or weighted. Unweighted graphs depict an edge between two nodes, where the "weight" is either one (exists) or zero (does not exist). Weighted graphs have an associated weight, which depicts how strong the directed or undirected relationship is between the two nodes. These weights are typically between zero and one, where one is the strongest the relationship can be. In biology, these graphs, also known as networks, can be used to depict biological functions.

For example, these networks can depict different types of interactions between two nodes, such as physical protein-protein interactions or gene regulatory relationships.

1.3.2 Network Inference Tools

There are many tools used in biology that can be used to infer various biological networks. These tools use a matrix where the rows are samples and the columns are features to create the final network. As technologies in biology have improved, the number of features included in the model has increased to hundreds of thousands or even millions of features. Improving these network inference tools to boost variable elimination, also known as feature selection [21], is crucial to creating high quality networks. Some categories of network inference tools include correlation, linear regression, bayesian, mutual information, and Random Forest-based methods.

Correlation

Correlation calculations between two features are commonly used as a means to create biological networks. Correlations are simple to calculate, and in some cases are intuitive in creating networks such as gene co-expression networks. One example of a correlation metric is Pearson’s Correlation Coefficient (PCC) [119]. The PCC between two features is defined as

$$r_{xy} = \frac{cov(x, y)}{\sqrt{var(x)var(y)}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

where x is feature one, y is feature two, and n is the number of samples [67]. Another example of a correlation tool is Spearman’s Correlation Coefficient (SCC) [119]. Unlike PCC which calculates the linear correlation between two features, SCC calculates the monotonic relationship between two variables. The SCC between two features is defined as

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the rank of the observation for feature one and the rank of the observation for feature two and n is the number of observations [33]. There are some issues that come with using these correlation-based methods to infer networks in biology. One is that the correlation is calculated only between two features. Many aspects of biology tend to have higher order interactions, as it is unlikely that two genes interact only with each other. Additionally these methods do not scale well as datasets become larger, since the correlation must be calculated $n^2 - n$ times, where n is the number of features.

Linear Regression

Multiple sparse linear regression approaches have been used to infer biological networks. Unlike correlation approaches where only the relationship between two features are considered, linear regression approaches uses multiple features as multiple regressors to determine the relationship to the dependent variable or feature of interest. The generalized form of this equations can be described as:

$$Y = B_1X_1 + B_2X_2 + \dots + B_nX_n + A$$

where Y is the dependent variable or the feature of interest, B_n are the coefficients that then becomes the weight of the edges in the final network, X_n are the independent features or variables, and A is constant coefficient. Some examples of these methods include Least Absolute Shrinkage and Selection Operator (LASSO) [165], Least Angle Regression (LARS) [40], and Trustful Inference of Gene REgulation using Stability Selection (TIGRESS) [73]. The downside of using linear regression techniques is that many interactions between biological variables are not linear. While the assumption that these interactions are linear can create accurate networks, some edges may be lost.

Bayesian Networks

A Bayesian network represents the conditional independence between variables and is represented as a directed acyclic graph [53]. However, Bayesian networks can be limiting since it does not allow cycles, which is common in biological systems. This can be remedied

by using dynamical Bayesian networks, but requires time series data [137]. Some tools that produce dynamical Bayesian networks and Bayesian networks include Max-Min Parents and Children (MMPC) [168], HITON-PC [4], and Bayesian Network Inference with Java Objects (BANJO) [71].

Mutual Information

Generally, mutual information captures the mutual dependence between two features. The mutual information between two features is defined as

$$I(X; Y) = H(X) - H(X|Y)$$

where H is the entropy of the feature [60]. Entropy captures the average level of uncertainty in the feature’s possible outcomes, where zero is maximum certainty. Entropy for feature X is defined as

$$H(X) = - \sum_{i=1}^n P(x_i) \log P(x_i)$$

Some examples of network inference tools that use mutual information include Algorithm for the Reconstruction of Gene Regulatory Networks (ARACNE) [121] and Context Likelihood of Relatedness (CLR) [42]. Mutual information-based tools capture relationships that may not have been possible using other tools like those that use linear regression and Bayesian networks. But like correlation-based tools, only capture relationships between two features in a vacuum.

Random Forest-Based Methods

Random Forest-based methods build decision trees based on all of the independent features available to predict the dependent feature of interest. This tree building step includes all independent features in the data set and is an embarrassingly parallel problem, creating the forest for each dependent variable that can easily be optimized computationally. There are many benefits to using tree-based methods. For one, all of the features are considered in the model. Additionally, these trees capture complex relationships beyond dependencies

and correlation. These are commonly used in many network inference tools, such as GENE Network Inference with Ensemble of trees (GENIE3) [84], iRafNet [138], BackWard Elimination Random Forest (BWERF) [35], and iterative Random Forest-Leave One Out Prediction (iRF-LOOP) [27].

Determining Which Method to Use

There have been many attempts to score these tools and determine the overall best method that is applicable to a variety of situations [119]. It is also possible for some of these tools to be combined to produce a better network [6, 75]. These high quality networks can then be used in downstream analyses to uncover new biological information.

1.3.3 Multiplex Networks

Many network types can be inferred using different tools, which can be explored separately to discover new biological information. However, biological interactions do not occur in isolation in one layer or network. The need to traverse across network types is crucial to understanding new biological information. A group of networks that share nodes and are merged together is called a multi-layer or multiplex network [89]. There are many methods to explore these multiplex networks, some described here include the Lines of Evidence method and various types of network walking algorithms.

The Lines of Evidence (LOE) [182] method uses multiple layers in a multiplex network to determine the relationship between a group of nodes or features of interest and all other nodes or features in the multiplex network. There are variety of LOE scores, including LOE *breadth* and LOE *depth*. For a group of genes of interest where n is the number of layers in the multiplex network, LOE *breadth* and LOE *depth* is defined as

$$LOE_{breadth}(g) = \text{bin}(D_{network_1}(g)) + \dots + \text{bin}(D_{network_n}(g))$$

$$LOE_{depth}(g) = D_{network_1}(g) + \dots + D_{network_n}(g)$$

$$\text{bin}(D(g)) = \begin{cases} 1 & \text{if } D(g) \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

where g is a gene not in the group of genes of interest and $D(g)$ is the degree or number of edges between g and the genes of interest. The higher the LOE scores, the more likely the gene is related to the group of genes of interest within the biological system.

Network walking is also a method to traverse across multiplex networks. In a random walk, a walker is placed at a node and randomly walks to neighboring nodes using the edges as the path between nodes. The weights associated with the edges are used as random probability for the walker traversing that edge, or as a toll to prevent the walker from using that path [28]. There are variations of the simple random walk algorithm, such as Random Walk with Restart (RWR) [172]. RWR uses random walk but includes a random probability that the walker is removed from its path and brought back to the start gene. This creates a normal distribution of visited nodes and produces a more accurate ranking of nodes that are related to the start node.

1.4 Tree-based methods

Tree-based methods are used to not only infer networks but also for feature selection. Decision trees on their own tend to be poor learners, but when combined in ensemble methods they become a powerful approach as they are non-parametric, can be optimized for speed in fit and prediction, have the ability to be parallelized, are explainable and interpretable, and produce accurate predictions.

1.4.1 Decision Trees and Ensemble Methods

A decision tree contains a root node containing all samples that is split into child nodes based on decisions or thresholds made by the model, separating the samples into smaller groups [Figure 1.1]. The goal of decision trees is to increase the variance explained of the dependent variable by splitting the samples into smaller groups. Meaning, the model looks

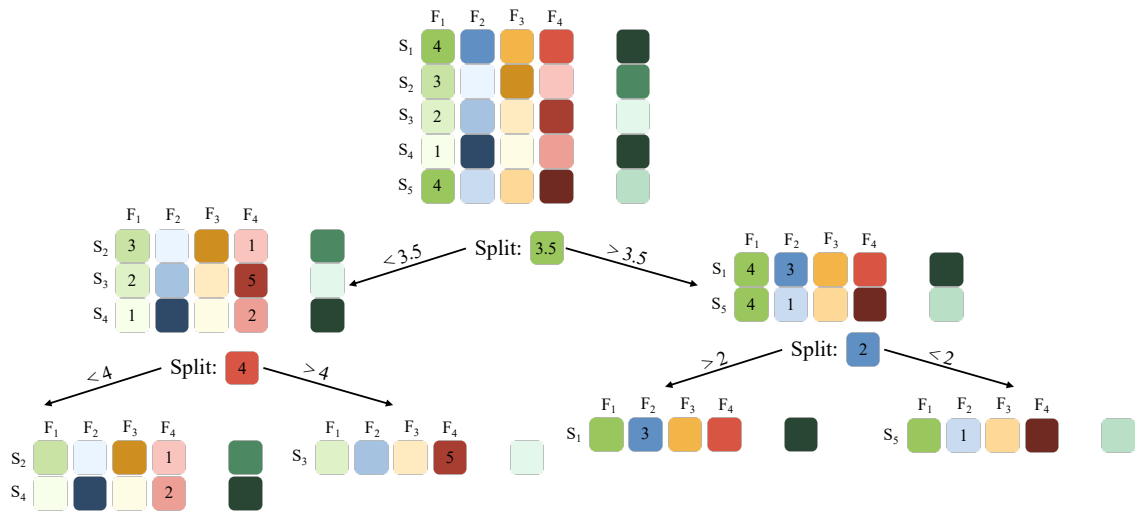


Figure 1.1: This figure depicts a binary decision tree. The root node contains all five samples, and the samples are split into two child nodes based on feature one. The two child nodes are then split into two more child nodes to create the full tree. In this scenario, the most important feature is feature one since it is picked first to split the data, and the least important feature is feature three since it is never selected.

at all possible splits of the samples and chooses the independent variable and threshold of that variable that divides the samples into child nodes that have as little variance as possible in the dependent variable. The classic decision tree used in many machine learning models is the Classification and Regression Tree (CART) [18]. CARTs can fit either continuous dependent variables (regression trees) or discrete dependent variables (classification trees).

Decision trees on their own tend to overfit the data and produce poor models. Ensemble methods use multiple decision trees to improve the model and prevent overfitting. One method is gradient boosted trees, which incrementally adds new trees, but the subsequent tree minimizes the error of the previous tree [52]. Two examples of this method are XGBoost [23] and AdaBoost [51]. The second category of ensemble methods is bootstrap aggregated or bagged decision trees, where for every tree a selection of independent features are used per tree rather than all variables. Two examples of this method are Rotation Forests [144] and Random Forests [17].

1.4.2 Random Forest

Random Forest (RF) [17] is a supervised learner that makes few assumptions about the relationships between variables and captures complex interactions between variables that are common in biological systems. It is explainable such that the importance score, the level importance of each dependent feature in predicting the independent feature, can be obtained. These importance scores can be used to rank and choose maximally important features for feature selection. RF additionally performs well in situations where there are fewer samples than variables, which is common in many biological experiments [15]. RF has proven to be useful in a variety of computational biology and clinical health applications, such as inferring networks of transcripts and proteins from transcriptomic and proteomic data [124], or predicting a binary dependent variable describing patient mortality based on a known set of features [111].

1.4.3 Iterative Random Forest

Iterative Random Forest (iRF) [9] provides an improvement to the RF base learner. In iRF, a standard RF is run where every feature is given equal weighting in the randomised feature sampling process. The feature importance scores from the forest are then used as weights in the feature sampling process in a new random forest. This process continues for a set number of iterations. At each iteration, some features have their importance reduced to zero and are effectively eliminated from the model, while other features have their importances boosted.

Both RF and iRF may be used in Leave One Out Prediction (LOOP), an algorithm that uses a matrix of data to produce all-to-all predictive networks. RF-LOOP has previously been implemented in GENIE3. GENIE3 has been considered the industry standard since winning the Sage Bionetworks Dialogue for Reverse Engineering Assessment and Methods (DREAM) Challenges 4 and 5, in which GENIE3 was the best network inference tool [84, 119]. iRF-LOOP has shown some potential in providing improvements and new information over a correlation-based co-expression network [27], and it is possible that iRF-LOOP can replace RF-LOOP in many contexts. Furthermore, this iterative process also improves the robustness of path analysis algorithms, such as Random Intersection Trees (RIT) [150], which determine sets of features that jointly affect the dependent variable [9]. RIT can be used to predict sets of variables that interact together. In the context of gene expression data, applying RIT to each iRF model has the potential to identify regulatory influences by sets of genes that form complex conditional relationships.

1.5 Data Cleaning and Pre-processing

An important step in the network inference pipeline is data cleaning and pre-processing before the chosen algorithm is applied to the data. The following precautions must be taken for Random Forest-based models to ensure that the resulting predictions are the most accurate. This includes removing features that are highly correlated with each other, correcting for batch effects between different experiments, and detecting and handling

potential outliers in the dataset that can alter the final predictions of the Random Forest-based models.

1.5.1 Correlated Features

When two features are highly correlated with each other, the overall accuracy of Random Forest models decreases [131]. This is due to the importance of two highly correlated features being split between the two when the decision tree chooses the important feature, resulting in half of the importance score that should be allocated to the given feature. The simplest method to amend this is to perform a simple all-to-all correlation calculation, such as PCC, between all features, select a representative feature of the highly correlated features, and remove the remaining features.

1.5.2 Batch Correction

Batch effects are differences between batches of samples or across experiments that are not based on biology. These batch effects have the ability to affect the results of a data analysis pipeline and need to be corrected [105]. Since the amount of data produced in the field of biology is increasing, batch effect mitigation is becoming a more important topic [64]. To correct for these batch effects, Batch Effect-Correction Algorithms (BECAs) are used. There are various classes of BECAs, including linear models, empirical Bayes, and factor-based algorithms [64].

Linear Models

The first class of BECAs are those that use linear models to correct for batch effects. This category includes mean scaling, median scaling, and zero centering [133, 149]. These methods are simple to implement, require little computational power, and remove most variance between samples that are due to batch effects. However, if the groups are unbalanced these methods are not as effective [133].

Empirical Bayes

The second class of BECAs are those that use Bayesian statistics to infer the batch effects of all biological features within a batch. The most popular Empirical Bayes BECA and the most commonly used BECA overall is ComBat [88, 129]. These methods tend to provide an improvement over traditional linear models [64].

Factor-Based

The third class of BECAs are those that do not require the user to specify known batch differences, rather the effect is estimated based on the entire matrix of data. There are two popular BECAs within this class, Surrogate Variable Analysis (SVA) [106], and Removed Unwanted Variation (RUV) [56]. Unlike the linear models and empirical Bayes BECAs, the factor-based BECAs can be used on larger datasets and are useful when the batch effect is not known.

1.5.3 Outliers

An outlier is generally a sample or value that deviates significantly from the rest of the samples or values in a dataset. There are commonly two types of outliers. One is a sample that is abnormal and differs from the other samples, and another is one sample that is mislabeled or miscalculated and belongs with a different set of samples [152]. Outlier correction tools must be able to determine when it is appropriate to remove an outlier. If a sample or value is determined to be an outlier, but is a factual measurement, it is possible that removing the outlier may lead to losing important biological information. There are two components to outlier correction, one of which is identifying the outlier, and the other is determining how to handle the outlier once it has been identified.

Outlier Identification

In a normally distributed dataset, the significance cutoff to determine an outlier is generally accepted to be two to three standard deviations away from the mean [109]. This means that the deviated value lies outside of the remaining 95% to 99.87% of the data. However

many measurements in biology are not normally distributed and there must be a more robust method of determining which sample or value is an outlier. One example is Robust Principal Component Analysis (rPCA) [24], which is similar to classical PCA with the exception of being able to identify the outliers themselves. rPCA unfortunately does not have the ability to determine whether the outlier is due to a mislabeling or due to a biological abnormality. Another example of outlier detection is Fast Outlier Sample Detection (FOSD) [195] which, unlike rPCA, can determine whether a value is simply mislabeled. Knowing that an outlier value is simply mislabeled allows for straight forward outlier handling by relabeling the value with the correct label.

Outlier Handling

Once an outlier value is identified, the outlier must either be removed from the data set or replaced with a more appropriate value. Removing a value from a dataset requires either removing the entire sample from the matrix, or filling the outlier value with a null or zero value. Removing the sample from the matrix may remove valuable biological information that exists in the non-outlier features. Filling the value with a null value is not possible for some algorithms such as Random Forest that requires a numeric value for each measurement in the matrix. Filling the outlier value with a zero is also ill advised, since a zero value may be meaningful in the biological measurement and may impact the final resulting prediction. The remaining option for outlier values is to infer what the outlier value should have been if it was not abnormal. For example, if the dataset is normally distributed and the outlier value fell beyond the two standard deviation cut off, the outlier value can be shifted to the numeric value of the cut off.

1.6 Post-Processing Networks

Once a network has been produced by a network inference tool, the network can contain millions of edges and a mix of signal and noise. The network must be post-processed before use either for biological analysis or in other downstream processes. If multiple networks are created and only one can be used, there must be a scoring system in place to determine

which network is of the highest quality. Additionally, the networks may need to be reduced in size to be more manageable in downstream processes, such as RWR. Networks can also be clustered to discover new biological information that is not obvious from simply exploring the network.

1.6.1 Network Evaluation

It is possible that multiple networks are created but only one can be used in a biological analysis or in a downstream process. In order to determine which network to use, the performance of each network must be evaluated to determine which network is the highest quality. Using the best network is crucial, the network of highest quality will capture more truth and signal than noise and false edges, and the biological results will be the most accurate.

Precision-Recall and Receiver Operating Characteristic

A confusion matrix, also known as an error matrix or a contingency table, shown in Figure 1.2a, is used to evaluate the performance of a network [135]. Using the established gold standard network of known truth (1's) and known false (0's) values and the predicted truth (1's) and predicted false (0's) values in the final network, many metrics can be calculated using the confusion matrix. These metrics include the following [32].

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{True Positive Rate} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{False Positive Rate} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$

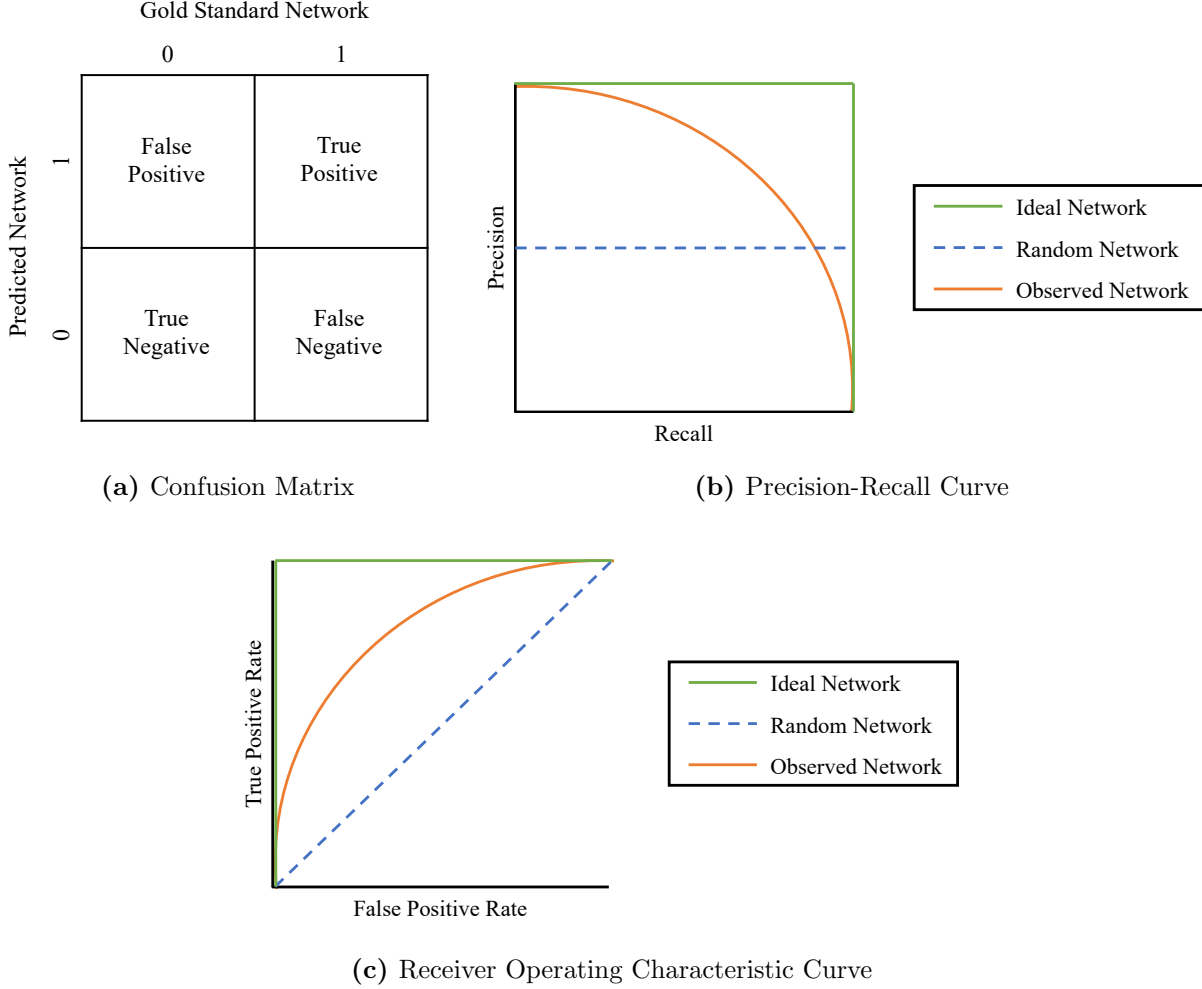


Figure 1.2: This figure depicts the confusion matrix (a) that is used to score the quality of resulting networks from network inference tools. The gold standard network contains true (1) and false (0) edges, and the predicted network contains predicted truth (1) and predicted false (0) edges. The confusion matrix determines whether the predicted edge is a false positive, true positive, true negative, or false positive. The Precision-Recall curve (b) and Receiver Operating Characteristic curve (c) are determined using the values from the confusion matrix, and used to visually determine the performance of the networks being scored.

Precision is the proportion of true positive (correctly identified) values obtained from all values that the model determined to be true. Recall is the proportion of true positive (correctly identified) values obtained from all true values in the gold standard network. The True Positive Rate (TPR) is the proportion of true positive (correctly identified) values obtained out of all of the true values and false negatives in the gold standard network. The False Positive Rate (FPR) is the proportion of false positive (incorrectly identified) values obtained out of all of the false positive and true negatives in the gold standard network.

Precision, Recall, TPR, and FPR is calculated only once for a two class classification problem. To score a network of edges where the weights of the edges vary continuously, the method to use the confusion matrix is to rank the edges and then select different thresholds to manually classify the edges as predicted truth and predicted false edges. Precision and Recall evaluated at these different thresholds is then be plotted to create a Precision-Recall curve, shown in Figure 1.2b. Similarly, TPR and FPR evaluated at these different thresholds can also be plotted to create a Receiver Operating Characteristic curve, shown in Figure 1.2c. The area under these two curves may be calculated, represented as the Area Under the Precision-Recall curve (AUPR) and the Area Under the Receiver Operating Characteristic curve (AUROC). These values vary between zero and one, if the AUPR or AUROC is closer to one the network that is being scored is of higher quality. A perfect network will have an AUPR and AUROC value of one. When deciding whether a single network is best to use in a downstream analysis, the general rule of thumb is that an AUROC value that is less than 0.5 is worse than random chance. For AUPR, this value is also 0.5 for a balanced dataset, but for unbalanced datasets is the ratio of true positive to true negative values [147]. When deciding between multiple networks to determine which is best, the one with the highest AUPR score, AUROC score, or both should be used. For cases where there are significantly more negative than positive values in the gold standard network, it is advised to use AUPR over AUROC, since AUPR does not take true negatives into consideration when performing calculations, but AUROC does [32].

Discounted Cumulative Gain

Discounted Cumulative Gain (DCG) is a scoring metric originally designed to score the relevancy of documents that are ranked by a search engine [86]. DCG can be adapted to rank edges in a network. DCG is defined as

$$DCG_{network,k} = \sum_{i=1}^k \frac{x}{\log_2(i+1)}$$
$$x = \begin{cases} 1 & \text{if edge is a true positive} \\ 0 & \text{if edge is a false positive} \end{cases}$$

where k is the number of edges in the network. Consider a network that contains five edges, where the edges are ranked from highest weighted edge to lowest weighted edge. Let's assume the first three edges were true positives, where they existed as true values in the gold standard network. The last two edges were false positives, where the edges were false values in the gold standard network or were not included in the gold standard network. The DCG of this network would be

$$DCG_{network,5} = \frac{1}{\log_2(2)} + \frac{1}{\log_2(3)} + \frac{1}{\log_2(4)} + \frac{0}{\log_2(5)} + \frac{0}{\log_2(6)}$$
$$DCG_{network,5} = 1 + 0.63092975357 + 0.5 = 2.13$$

Unlike AUPR and AUROC, DCG takes into consideration the ranking of the edges themselves, and assigns a decay parameter in the denominator as the true positives are found later in the ranking of the network. Additionally, DCG does penalize the network for discovering a false positive edge at a high ranking. This is beneficial for use when the complete gold standard network is unknown, which is common for uses in biology.

While DCG can be used between networks to determine which network is the best to use, it is not as informative to determine whether a network is of high quality on its own. In this case it may be better to use the normalized Discounted Cumulative Gain (nDCG) [86]. nDCG is the DCG normalized by the ideal DCG (IDCG) of the gold standard network. This is defined as

$$IDCG_{gold\ standard,k} = \sum_{i=1}^k \frac{1}{\log_2(i+1)}$$

$$nDCG_{network,k} = \frac{DCG_{network,k}}{IDCG_{gold\ standard,k}}$$

Looking back at the previous example of the network of five edges, let's assume that the gold standard network contains only five true values. The nDCG of the scored network would be

$$IDCG_{gold\ standard,5} = \sum_{i=1}^5 \frac{1}{\log_2(i+1)} = 2.95$$

$$nDCG_{network,5} = \frac{DCG_{network,5}}{IDCG_{gold\ standard,5}} = \frac{2.13}{2.95} = 0.72$$

The singular nDCG score ranges between zero and one.

1.6.2 Network Thresholding

Many network inference tools create unwieldy networks. Thousands of features used to produce all-to-all directed networks may contain millions of edges. Not only will downsizing these networks shorten the computational requirement in later analyses, most of these edges are noise and separating the signal from noise will produce more accurate results. However, there is a fine line between removing noise and removing pertinent information. Once a threshold is determined, edges that are above this threshold are kept, and the remaining edges are assigned a value of zero. Depending on the use of the network, the original edge values may be kept or be converted to a single value of one.

One method of thresholding a network is to choose an arbitrary threshold [50], although this is the least effective and is predominantly used as an initial pass before thresholding further. Another method of thresholding is to rank every edge in the network and choose only the top n percent of edges to retain in the final network [3, 27]. This percentage can range from the top 0.01% of edges to the top 10% of edges, depending on how stringent of a threshold is required and the desired final network size. More complex methods include

artificially inserting a control that is filtered out manually to determine the threshold required [175] and using spectral graph theory-based clustering [136]. This latter method traverses the network and chooses nodes and edges to remove based on the number of "nearly-disconnected" sections of the graph.

1.6.3 Network Clustering

Another post-processing tool for networks is clustering, which groups like nodes or features together, distinctly separated from the remaining nodes. These clusters can be evaluated as a whole, and new biological relationships may be discovered based on membership within a cluster. Clusters have two major requirements, one is homogeneity, or that members within a cluster must be similar to each other or share similar traits. The second requirement is that the members within a cluster differ from all other nodes in other clusters [189]. There are multiple algorithms that currently exist to cluster networks, which are described in detail below.

Graph Theory-Based Methods

Some network clustering algorithms use graph theory properties to create groupings of nodes that are highly related to each other. This includes algorithms such as Highly Connected Subgraphs, Clustering Identification via Connectivity Kernels, and Restricted Neighborhood Search Clustering.

The Highly Connected Subgraphs (HCS) [72] algorithm is based on identifying clusters that are highly connected. Highly connected is defined as a graph with $n \geq 1$ nodes where the edge connectivity, or the minimum number of edges that disconnects the graph, is greater than $\frac{1}{2}n$. HCS requires that the network be previously thresholded, since a complete all-to-all network would be a highly connected network. HCS makes no assumptions on the number of clusters or the structure of the clusters created by the algorithm.

The Clustering Identification via Connectivity Kernels (CLICK) [151] algorithm is an extension of HCS, but uses weighted edges and calculates the edge connectivity based on the

weights as probabilities. Like HCS, CLICK makes no assumptions on the number of clusters or the structure of the clusters created by the algorithm.

The Restricted Neighborhood Search Clustering algorithm (RNSC) [98] is a cost-based search algorithm that separates the nodes into clusters by determining the lowest cost split. First, the nodes are separated randomly into n clusters, and individual nodes are moved at random into separate clusters. The node is moved into the cluster that has the lowest cost. To improve the efficiency of this algorithm, this process is repeated multiple times, and some moves that are counter-intuitive are illegal, such as moving a node back to a cluster that it has already been moved out of. Unlike HCS and CLICK, RNSC requires more memory and is much more computationally expensive. While n must be specified to begin the search algorithm, RNSC makes no assumptions on the number of clusters created by the final iteration of the algorithm.

Squared Error-Based Clustering

Some network clustering algorithms use the squared errors to determine whether a node belongs in a cluster or not. This category includes k -means clustering and the extension of k -means clustering Iterative Self Organizing Data Analysis Technique.

K -means [110] clustering is a well known squared error-based clustering algorithm. First the nodes are separated into k random clusters and the center or centroid of the cluster is determined. Then, each data point is moved to the cluster that minimizes the sum of squared errors or the sum of the squared Euclidean distances between the data point and the centroid. While this clustering algorithm is simple, the computational complexity increases exponentially as number of nodes and cluster size increases.

One drawback of using k -means clustering is that it requires the number of clusters, k , to be defined in order to be used. Iterative Self Organizing Data Analysis Technique (ISODATA) [8, 125] is an extension of k -means clustering that does not require the number of clusters to be predefined. Instead, a threshold is predefined by the user to determine the sum of squared errors between clusters and the standard deviation within clusters. The clusters where the centroids are below the sum of squared errors threshold are merged together, and

clusters where the centroids are above the standard deviation threshold are separated into new clusters.

Fuzzy Clustering

Fuzzy clustering, also known as soft clustering, groups nodes in clusters but nodes can have membership in more than one cluster [192]. Fuzzy c -means clustering (FCM) [39] is similar to the hard clustering method k -means clustering, but differs since single nodes do not have membership to one cluster. Rather the node has a likelihood value between zero and one that the node has membership to a given cluster.

Markov Clustering

The Markov Clustering algorithm (MCL) [174] uses a combination of random walks and Markov chains [55] to determine the clusters within a graph. If two nodes are connected by an edge, there is a probability that a walker starting at node A will travel to node B. If two nodes are not connected, then this probability is zero. MCL uses these probabilities in conjunction with the weights of the edges themselves to determine the flow of a graph and which nodes belong in a cluster together. While MCL does not require the number of clusters to be predefined, some parameters do need to be user defined such as the inflation value that can change the structure of the resulting clusters.

1.7 Systems Biology

The previous methods for network inference as well as the pre- and post-processing of these networks can all be applied to systems biology studies to explore new functions and relationships within the system. Model organisms have been used as the basis of these studies since the late 19th century and have a wealth of information available to researchers hoping to draw conclusions to their organism of interest. These model organisms contain vast databases of information that contain networks from multiple different biological measurements and relationship types, from interactions between proteins to networks based on expression profiles of genes throughout the system.

1.7.1 Model Organisms

Model organisms are select species that are widely studied where discoveries in one organism can be applied to other organisms [108]. The first widely known and most commonly studied model organism is the bacteria *Escherichia coli*. *E. coli* studies began as early as the late 19th century. *E. coli* was targeted as the first model organism since it was commonly found in the human digestive tract and was easy to acquire, hardy, and non-pathogenic [12]. The use of *E. coli* as a model organism has brought forward significant biological advancements, from the discovery of the genetic code [30], to the production of artificial insulin to treat diabetes [62], and use in vaccine development [58]. Another example of a commonly used model organism is *Saccharomyces cerevisiae*, a fungi otherwise known as baker’s yeast, which like *E. coli* became crucial in understanding many different biological functions and relationships [132]. Other eukaryotic model organisms include *Schizosaccharomyces pombe* or fission yeast [190] and *Chlamydomonas reinhardtii* or green algae [69]. Two common plant model organisms include *Arabidopsis thaliana* or thale cress [184] and *Populus trichocarpa* or the black cottonwood poplar tree [41]. In the animal kingdom, common model organisms include the nematode *Caenorhabditis elegans* [122], the fruit fly *Drosophila melanogaster* [143], and the mouse *Mus musculus* [139].

Saccharomyces cerevisiae

Saccharomyces cerevisiae is a single cell fungus that has been crucial to human culture by being the basis of bread, wine, and beer since ancient civilization. *S. cerevisiae* became a model organism due to its abundance, its quick doubling time at room temperature, reproduction by both mitosis and meiosis, and the high number of orthologs between *S. cerevisiae* genes and human genes (about 23%) [49]. *S. cerevisiae* was the first eukaryotic genome to be sequenced in 1996 [63]. The genome contains 16 chromosomes, 6,275 genes, and 12 million base pairs [113].

The genome is regularly curated by the Saccharomyces Genome Database (SGD) [25]. The SGD produces publicly available datasets such as hundreds of microarray expression datasets and the reference genome in its entirety. Other databases include YeastCyc

[94] which produces metabolic pathways, the Comprehensive Yeast Genome Database [66] that contains protein-protein interaction data, YEASTRACT [162] that contains literature curated transcription factor to gene relationships, and YeastNet v3 [97] which provides multiple inferred gene-to-gene networks such as gene interaction and phylogenetic networks.

Arabidopsis thaliana

Arabidopsis thaliana is a flowering plant that tends to grow in hardy conditions and is considered a weed. *A. thaliana* grows quickly, is easy to cultivate, is small in size, and is related to many other plant species which makes it an ideal model organism [155]. *A. thaliana* was the first plant genome to be sequenced in 2000 [95]. The genome contains 5 chromosomes, 27,655 coding genes, 5,178 noncoding genes, and 119 million base pairs [159].

The genome is regularly curated by The Arabidopsis Information Resource (TAIR) [159]. Like the SGD for *S. cerevisiae*, TAIR produces multiple publicly available datasets such as the genome in its entirety and polymorphism data. Other databases include AraCyc [128] for metabolic pathways, 1001 Genomes [181] which contains expression data and genotypic variation, and AraNet [103] that produces inferred gene-to-gene networks.

1.7.2 Layer Types

Gene-to-gene networks are commonly used to understand biological relationships between genes in a biological system. One network, or layer, will often only capture gene-to-gene relationships using one method. An understanding of the different layer types, what they capture, and how to create these layers is crucial to deciding which layers are appropriate to use in downstream analyses.

Co-Citation Network

A co-citation network is a network that infers the connection of two nodes based on an analysis of multiple published papers and whether they are cited together frequently. Typically these co-citation network inference algorithms mine abstracts on publicly available

databases like PubMed. An example of such an algorithm is Co-Citer [140]. These networks tend to miss biological information that is currently unknown.

Protein-Protein Interaction Network

A Protein-Protein Interaction network (PPI) is a network that connects two proteins based on whether there is a known interaction that occurs between them. Proteins often physically interact in groups or complexes, thus these networks, used in conjunction with other networks, can lead to new relationships between proteins and genes outside of the interactome [68]. Some commonly used PPI databases include the Search Tool for the Retrieval of Interacting Genes database (STRING) [160] and the Kyoto Encyclopedia of Genes and Genomes database (KEGG) [90]. While PPIs are quite useful, they do not capture all known protein protein interactions in a biological system. Even for a model organism like *S. cerevisiae*, only 50% of the interactome has been discovered [70].

Phylogenetic Network

A phylogenetic network is a network that connects two nodes if they have similar phylogenetic profiles, thus having some evolutionary relationship. Phylogenetic networks capture reticulate evolution, which includes horizontal gene transfer (when genetic material transfers from one lineage to another), homoplasy (when two organs or structures evolve separately in two different species but have similar functions), and hybridization (when two organisms of separate species reproduce) [118].

Metabolic Network

A metabolic network captures the relationship between proteins that share a compound within a metabolic pathway. Metabolism captures the biochemical and physiological processes that occur within the cell. There are many databases to obtain metabolic datasets from, including KEGG and BioCyc [94].

Expression-Based Networks

There are a multitude of gene-to-gene networks that are constructed based on gene expression datasets. One example of this is a Gene Regulatory Network (GRN) that captures biological processes that are based on gene regulation. GRNs are directed networks that connect transcription factors to all other genes. Gene regulation occurs in many cellular processes and understanding this relationship between the regulatory gene and the product is instrumental in understanding these processes [93]. Another example is a Predictive Expression Network (PEN), which is an extension of a GRN in that it contains directed edges from one gene to another gene, but captures relationships beyond directed regulation [27]. Another network type based on expression is a co-expression network, which measures which genes are expressed together in a biological system [173]. Some examples of algorithms to create these networks include GENIE3 to create GRNs, iRF-LOOP to create PENs, and PCC to create co-expression networks.

1.8 Relevant Biological Techniques

A basic understanding of some biological techniques is required to process the resulting data. In particular, data from Genome-Wide Association Studies and microarray expression datasets is used extensively in this dissertation. A brief overview of these two methods is described below.

1.8.1 Genome-Wide Association Study

A Genome-Wide Association Study (GWAS) is a study on the single nucleotide polymorphisms (SNPs) within the individuals of a population and its relationship with phenotypes that are expressed by those same individuals. A SNP is a single base pair in a genome that has been changed to a different base pair. The phenotype is the observed trait of the individual. GWAS can be used to determine if a specific SNP or variant is associated with a phenotype such as a specific human disease [116]. An example of a statistical test that determines this relationship is the χ^2 test in PLINK [120]. One negative aspect of

using GWAS tools is that a large and diverse population size is necessary in order to obtain accurate results.

The population structure – also known as the relatedness among individuals within a population – must be accounted for to prevent false positive associations within a GWAS [158]. There are two types of relatedness, ancestral relatedness and cryptic relatedness. Ancestral relatedness occurs when two individuals have a common ancestor, while cryptic relatedness occurs when the relationship between two individuals is unknown, and is possible in very large sample sizes [158]. Linear mixed models are commonly used to correct for population structure, by incorporating the unmodeled factors into the model [158]. EMMA [92] and Fast-LMM [112] are two GWAS methods that utilize these linear mixed models.

A single-locus GWAS is one that evaluates each SNP independently for the association to the phenotype, while a multi-locus GWAS evaluates multiple SNPs at once [19]. Multi-locus GWAS quickly becomes computationally intensive since the number of SNPs may be in the hundreds of thousands or millions, since SNPs are not often filtered before use [19]. An example single-locus GWAS method is GEMMA [194] and an example multi-locus method is mrMLM [180].

1.8.2 Microarray Expression Datasets

A microarray expression dataset is a matrix of gene expression where the columns measure the expression itself and the rows are the samples of tissues or cells. This matrix is obtained using a chip containing an array of oligonucleotides that, via hybridization, measure the amount of complementary mRNAs in a sample to quantify the expression of each of the genes. Since the first known use of microarrays in the 1980's [22], microarray experiments became increasingly more popular due to the number of genes it could measure at once and quickly became a staple in scientific research [13].

1.9 Overview of COVID-19

This dissertation contains one chapter that analyzes the effects of a patient's surrounding environment on the rate of survival of COVID-19. A fundamental understanding of the

COVID-19 pandemic is required to truly grasp the effects of the environment on COVID-19 outcomes. The Coronavirus disease 2019 or COVID-19 is caused by the SARS-CoV-2 coronavirus that was first discovered in December of 2019 in Wuhan, China [101]. This disease caused varying symptoms, and death in some cases, and some survivors face symptoms for months after initial infection, often called long covid. The effects of long covid are still largely unknown. The COVID-19 pandemic has resulted in nearly 5.5 million deaths worldwide as of January 1st, 2022 [38]. As of early 2022, there are four vaccines that have been approved for use worldwide as protection against COVID-19, including the Pfizer-BioNTech, Moderna, Janssen or Johnson & Johnson, and Oxford-AstraZeneca vaccines. However, as much of the world remains unvaccinated, new variants of the original strain of SARS-CoV-2 have come about threatening overall immunity of those that are fully vaccinated and leaving the door open for new more deadly variants to potentially infect the population.

1.9.1 Comorbidities and Environmental Predictors of COVID-19

There are multiple comorbidities that hinder the survival of COVID-19 of an individual patient. These include diseases such as chronic kidney disease, chronic obstructive pulmonary disease (COPD), diabetes, hypertension, and cardiovascular disease but also demographic features of the patient such as being male and of older age [43, 177]. There are some environmental predictors that may influence COVID-19 survival, one of which is seasonality, or what time of year a patient becomes sick [114]. It is possible that this is due to viral load being higher in the winter months due to people moving indoors, or due to the vitamin D absorption from solar radiation that lessens in the winter months [5]. Another environmental predictor may be the air quality of urban areas. Since SARS-CoV-2 is a respiratory virus, low air quality and high air pollution that is found in urban environments may play a role in the survivability of patients that contract COVID-19 [141].

1.9.2 Measurements of Severity

In order to guide policy making to prevent the spread of COVID-19, there must be a clear measurement of severity of the disease. One method of estimating the severity of disease is Case Fatality Ratio (CFR). The equation of CFR is defined as [59]

$$\text{CFR} = \frac{\text{Cumulative Deaths}}{\text{Cumulative Deaths} + \text{Cumulative Recoveries}}$$

In smaller populations, such as those aboard the February 2020 Diamond Princess cruise ship, it is simple to determine the number of recoveries and deaths that occur [145]. However, in large populations Infection Fatality Ratio (IFR) must be used instead. While the equation remains the same, there is a difference in the estimation of recoveries and deaths in large populations. One example of an IFR estimation, Sero-excess IFR, uses the number of excess deaths in the population as the number of cumulative deaths in the IFR equation [115]. These measures however rely on efficient COVID-19 testing within a community. For example, if only the sickest patients in hospitals are being tested for the disease, the CFR will be biased heavily.

Another measure of infection in a community is the Test Positivity Rate. This is defined as

$$\text{TPR} = \frac{\text{Positive Tests in a Given Time Period}}{\text{Tests Administered in a Given Time Period}}$$

The TPR of a given population may signal potential problems that policy makers and health care workers can expect in the future. For example, a high TPR may signal a surge and a lack of available tests within a community. TPR is highly correlated with ICU occupation with a delay of about twelve days, thus TPR can be used to predict the resources required for hospitals to give care to all who need it [45].

1.9.3 Issues Surrounding COVID-19 Studies

As the pandemic continues, there are some issues that have arisen that need to be taken into consideration as COVID-19 studies are conducted. One of which is that the new variants

of the original alpha strain of the SARS-CoV-2 virus have different symptoms and disease severity associated with it. While it can be determined when a particular variant becomes dominant in a community, this relies on sufficient data collection from the community itself.

Another issue to take into consideration is the lack of reporting on cases and deaths of COVID-19. Some countries simply do not have the infrastructure to test their population and produce detailed databases of the data from their citizens. Even if the countries do produce these datasets, it may only contain statistics for the entire country rather than a specific region. In the U.S., the finest granularity for daily cases and deaths are at the county level, supplied by the Johns Hopkins University Center for Systems Science and Engineering 2019 Novel Coronavirus Visual Dashboard [38]. Daily testing data at the county level is also available from the U.S. Covid Atlas provided by the University of Chicago Center for Spatial Data Science [100]. However, the public release of this information has become a point of contention in some states. In mid 2021, some states have stopped updating their COVID-19 dashboards and in early 2022 some states such as Tennessee have moved to reporting their cases and deaths on a weekly basis rather than a daily basis [10, 163].

1.10 Conclusion

While the wealth of knowledge in the field of biology increases, the need for high performance computing tools to analyze these large datasets becomes increasingly more necessary. iRF can be used to create multiple networks that can then be merged into a multiplex network that can be traversed to explore new biological relationships. iRF can also be used to determine important predictors of a dependent variable.

This dissertation focuses on iterative Random Forest (iRF) and its uses throughout different biological systems. The first chapter validates the use of iRF over traditional Random Forest. The second chapter uses iRF in the context of human health, to predict the outcomes of COVID-19 patients using environmental factors. The third chapter introduces an improved pre-processing pipeline of microarray data to produce Predictive Expression Networks (PENs). This improved technique is demonstrated in chapter four in a systems biology study using *Septoria musiva* and *Saccharomyces cerevisiae*.

Chapter 2

Evaluating the Performance of Random Forest and Iterative Random Forest-Based Methods when Applied to Gene Expression Data

2.1 Abstract

Gene-to-gene networks, such as Gene Regulatory Networks (GRN) and Predictive Expression Networks (PEN) capture relationships between genes and are beneficial for use in downstream biological analyses. There exists multiple network inference tools to produce these gene-to-gene networks from matrices of gene expression data. Random Forest-Leave One Out Prediction (RF-LOOP) is a method that has been shown to be efficient at producing these gene-to-gene networks, frequently known as GENE Network Inference with Ensemble of trees (GENIE3). Here we show that iterative Random Forest-Leave One Out Prediction (iRF-LOOP) produces higher quality networks. We use both synthetic and empirical networks from the Dialogue for Reverse Engineering Assessment and Methods (DREAM) Challenges by Sage Bionetworks, as well as two additional empirical networks created from *Arabidopsis thaliana* and *Populus trichocarpa* expression data.

2.2 Introduction

Gene Regulatory Networks (GRNs) are crucial to understanding the biological system since they show the regulatory relationships between transcription factors and target genes. A GRN can be predicted from an input gene expression matrix (a data set with samples as rows and genes as columns measuring expression levels of each gene) using a variety of algorithms such as Weight Gene Correlation Network Analysis (WGCNA) [102] and BackWard Elimination Random Forest (BWERF) [35]. A conceptual expansion on the GRN is the Predictive Expression Network (PEN). While typical GRNs tend to be focused on transcription factor-to-gene relationships, PENs can capture expression-based relationships beyond regulation by transcription factors. Unlike GRNs, algorithms that produce PENs utilize all genes in the input data, creating an all genes-to-all other genes network. This all-to-all network provides additional information not found in a standard GRN, since these non-transcription factor edges can reveal additional biological relationships, such as genes interacting together in a biosynthetic pathway, or genes that encode proteins that form larger complexes. However, due to the increased number of input genes involved, inferring

a PEN can be computationally expensive and challenging to implement on larger datasets. The most common approach, typified by WGCNA, is to calculate a pairwise correlation between all genes and then threshold the results. While this is computationally efficient for large datasets, the simplistic model only accurately reflects relationships where one gene influences another gene in a linear manner and absent of influence from any other genes. In reality, it is possible that a gene’s expression is jointly influenced by many other genes, possibly in a non-linear manner, and a better model should account for this.

Random Forest (RF) [17] is a supervised learner that makes few assumptions about the relationships between variables and is able to capture complex interactions between variables that are common in biological systems. It is explainable, such that the importance score of each feature in the model can be obtained and ranked. RF provides a good base learner for the purposes of gene regulatory inference. One may fit a RF model to predict the expression of a target gene by using the expression of all other genes across all samples as model features. The result is a bipartite network with directional edges between predictor genes (features) with high importance in the model, and the regulatory target gene. This approach can be extended with RF-Leave One Out Prediction (RF-LOOP), which performs this process for every gene in the gene expression matrix. If there are n genes in the expression matrix then n RF models are produced. Each gene is used as the dependent variable once, resulting in n bipartite networks. When the bipartite networks from all n models are merged, it produces an all-to-all PEN. RF-subLOOP is a modification of RF-LOOP, in which the features used in each individual RF model are limited to a subset of all possible features (e.g. known transcription factors), while the dependent variables selected are the potential regulatory target genes. This produces a traditional GRN containing only TF-to-target edges.

Several implementations of RF-LOOP and RF-subLOOP exist, the best known being GENE Network Inference with Ensemble of trees (GENIE3) [84]. GENIE3 was the overall winner of the DREAM4 and DREAM5 competitions, proving the robustness of the RF-LOOP approach to inferring gene-to-gene connectivity based on expression data [84, 119].

Iterative Random Forest (iRF) [9] provides a potential improvement to the RF base learner used by GENIE3 and several other GRN inference methods. In iRF, a standard Random Forest is initially run where every feature is given equal weighting in the randomized

feature sampling process. The feature importance scores from the forest are then used as weights in the feature sampling process in a new random forest. This process continues for a set number of iterations. At each iteration, some features may have their importance reduced to zero and are effectively eliminated from the model, while other features have their importance boosted. For expression network inference, iRF provides improvements that would not be achieved by simply increasing the number of trees in a single random forest, such as removing spurious edges from the network completely (when their importance is zero) and boosting important edges in the final edge ranking list. Furthermore, this iterative process also improves the robustness of downstream path analysis algorithms [9], such as Random Intersection Trees (RIT) [150], which determine sets of features that jointly affect the dependent variable. In the context of gene expression data, applying RIT to each iRF model has the potential to identify regulatory influences by sets of genes that form complex conditional relationships.

The core of both RF and iRF is the decision tree. A decision tree has a root node of samples that continually splits into child nodes based on a threshold that separates the samples into groups until a stopping criterion is met. In this case, a binary decision tree is used, which splits the samples into two child nodes to increase the variance explained of the dependent variable. The decision tree evaluates all possible splits based on the features in the model, and chooses the feature and split that divides the set of samples into two child sets that each has as little variance as possible. Decision trees tend to overfit the data, but this issue can be countered by using an ensemble of decision trees such as RF.

In a Random Forest, each tree only uses a sample of the data (rows) randomly selected with replacement, while only a subset of all features is assessed for each split. K or $Mtry$ is the number of features evaluated at each split, which is fixed for the entire forest. In this analysis we use $K = \sqrt{n}$ where n is the total number of features. At the completion of the forest building, importance scores are aggregated across all trees to produce a ranked list of important independent features in predicting the dependent variable. These feature important scores dictate how important the individual feature was in creating the decision tree.

Here we compare the performance of RF and iRF in producing GRNs and PENs from gene expression data by comparing the networks resulting from using RF-LOOP (via GENIE3) or iRF-LOOP and iRF-subLOOP or RF-subLOOP (via GENIE3). Using both synthetic and experimental data, we find that iRF outperforms RF on various metrics, producing more accurate predictions, smaller networks with improved signal-to-noise ratio, and higher quality top ranked edges.

We have included an in depth analysis of the top ranked PEN edges produced by iRF-LOOP on *Arabidopsis thaliana* gene expression data as well as the top sets of interacting genes produced by RIT on the resulting iRF tree paths. This analysis shows that the predictive power available using iRF can reveal interactions beyond TF-driven interactions. Additionally, we have included an analysis comparing the resulting biological pathways in PENs produced by GENIE3 and iRF-LOOP on *A. thaliana* and *Populus trichocarpa* expression data. Overall, this study shows that iRF provides meaningful biological information that would not have been obtained using a RF-based method.

2.3 Methods

In order to compare iRF-LOOP versus GENIE3, we used multiple synthetic datasets and two empirical datasets for evaluation. Synthetic datasets were provided by Sage Bionetworks in the form of their Dialogue for Reverse Engineering Assessment and Methods (DREAM) Challenges, in particular DREAM4 and DREAM5 challenges for inferring PENs and GRNs from population-scale expression data.

We also compare iRF-LOOP and GENIE3 in creating PENs from real gene expression data available for *A. thaliana* from 1001 Genomes [96, 181] and for *P. trichocarpa* from the NCBI SRA database [169]. The GRNs and PENs generated for each dataset were evaluated against a gold standard network, which is a known network of confirmed true positive and true negative edges expected in the inferred network.

2.3.1 RF-LOOP and iRF-LOOP

Both RF-LOOP and iRF-LOOP use an input matrix of features measured in a population of samples to produce all-to-all predictive networks. For example, for an input matrix of n gene expression features, RF-LOOP and iRF-LOOP will produce n models where the expression of $n - 1$ genes are predicting the expression of the i^{th} gene, and $i \in \{1...n\}$. The importance scores for the i^{th} model determine the strength of the edges from the predictor genes to the i^{th} dependent gene. However, importance scores across each of the n models are not comparable, so they must be normalized before they are merged into one final network. The network that results from applying RF-LOOP or iRF-LOOP to a gene expression matrix is a Predictive Expression network (PEN). We used five iterations as the default number of iterations for iRF-LOOP and iRF-subLOOP.

The latest version of GENIE3 [2] is used throughout this analysis as an implementation of RF-LOOP, in which the importance scores for each generated random forest are normalized by the total variance of the dependent variable (Y vector) for that forest. In contrast, in the High Performance Computing (HPC) implementation of iRF used for iRF-LOOP, the importance scores of each generated iterative random forest are normalized by the sum of the importance scores in that iRF model. This HPC version of iRF is parallelized across compute nodes and is capable of running large datasets of hundreds of thousands of features [27].

2.3.2 RF-subLOOP and iRF-subLOOP

For cases in which a known list of regulatory genes exists, a transcription factor-to-all other genes network (GRN) can be created instead of an all-to-all network (PEN) using the RF-subLOOP algorithm. This requires subsetting the genes used as independent features to only the known regulatory genes [Figure 2.1]. RF-subLOOP is implemented in GENIE3 software by using the `regulators` argument, denoting the regulators to be used as predictors. This use is recommended by the authors due to the computational complexity of performing RF-LOOP for large datasets. Our iRF-LOOP implementation, however, does not have this same capability, so in order to implement the subLOOP functionality we used two separate

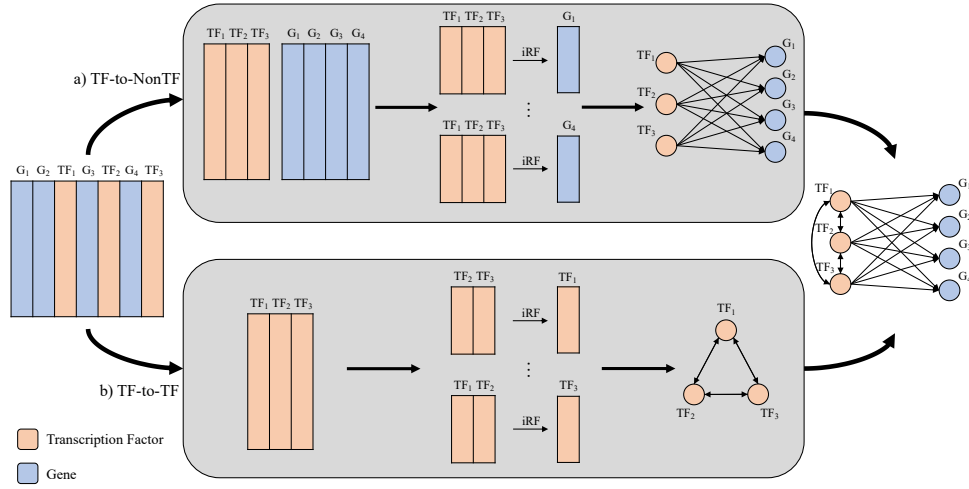


Figure 2.1: This diagram shows the process of creating a transcription factor-to-all other genes GRN using the iRF-subLOOP algorithm. The overall algorithm contains two types of iRF runs. The first run type (a) creates a TF-to-NonTF network by partitioning the expression matrix into TF and NonTF parts. The TFs are then used as the features predicting each of the individual genes in the NonTF matrix. The second run type (b) creates a TF-to-TF network, similar to iRF-LOOP but instead of all genes being used as the input features, only the TFs are used. The resulting networks from (a) and (b) are merged to create a TF-to-All GRN.

types of iRF runs [Figure 2.1]. The first run (a) separates all the feature columns into two matrices, one of all transcription factors and one of all other non-transcription factor genes. For every gene in the non-transcription factor matrix, the given gene is used as the dependent variable and the entire matrix of transcription factors is used as model features. This produces a TF-to-NonTF network. The second run (b) is similar to iRF-LOOP, but only the transcription factors are used as model features and as dependent variables. This produces a TF-to-TF network. Combining these two networks creates an overall TF-to-All network. This separation is crucial - if iRF-LOOP is run on all genes instead and the NonTF-to-All edges are simply dropped from the final network, the resulting paths of the trees would include these unwanted features and impact the final importance scores, making it incomparable to the GENIE3 implementation of RF-subLOOP.

2.3.3 Dream Challenge Datasets

The DREAM4 *In Silico Size 100 Multifactorial* sub challenge provides five small synthetic expression data sets of 100 samples and 100 genes to predict five networks [84]. GENIE3 and iRF-LOOP were used to create PENs from each of the DREAM4 expression datasets. For DREAM4 datasets, GENIE3 used 1000 trees for each RF model, while iRF-LOOP used 1000 trees for five iterations, yielding 5000 trees total for each iRF model. GENIE3 was also run with 5000 trees per RF model to ensure that any improvement shown in iRF-LOOP was due to the iterative process, not the increase in trees.

The DREAM5 *Network Inference* challenge provides one synthetic and two experimental expression data sets, of a varying number of samples and genes, to predict three networks [119]. Network 1, the synthetic dataset, contains 1,643 total genes and 805 samples. Networks 3 and 4, empirical datasets from *E. coli* and *S. cerevisiae*, contain 4,511 total genes and 805 samples and 5,950 total genes and 536 samples respectively. In each of these datasets, a subset of genes were flagged as known transcription factors (195, 334, and 333 transcription factors for networks 1, 3, and 4 respectively) and the expectation of the DREAM5 competition was that TF-to-target GRNs would be inferred for each dataset. Therefore, RF-subLOOP (via GENIE3 with the known transcription factors used as regulators) and iRF-subLOOP were used to create GRNs from the DREAM5 expression

datasets. For DREAM5, GENIE3 was run with 1000 trees only, while iRF-subLOOP was run with 1000 trees at each iteration for five iterations.

2.3.4 *A. thaliana* and *P. trichocarpa* PEN inference

We compared the performance of GENIE3 and iRF-LOOP for inferring PENs from real *A. thaliana* and *P. trichocarpa* gene expression data from leaf tissues. For *A. thaliana*, SRA data were obtained from Kawakatsu, et al. [96], totaling 6,584 FASTQ files from 728 samples and 38,186 genes. STAR [37] was used to map reads to the TAIR 10.1 *A. thaliana* reference genome. 15 samples and 186 corresponding runs were removed from the resulting raw counts matrix since the corresponding runs did not have a GSM sample ID and could not be tied to genotypes. Raw counts per gene were summed for each sample. Two more samples, GSM2135743 and GSM2136308, were removed due to low overall counts. Genes with less than 50 reads in at least 10% of samples were then removed, removing 18,199 genes. Raw counts were converted to gene length corrected Trimmed Mean of M-value (geTMM) [153]. The resulting X matrix for iRF-LOOP contained 711 samples and 19,987 genes. For *P. trichocarpa*, a similar process was applied. The RNAseq data was obtained from the NCBI SRA database (SRA numbers: SRP097016-SRP097036; www.ncbi.nlm.nih.gov/sra). Reads were aligned with STAR, using the *Populus trichocarpa* v.3.0 reference genome [169], totaling 470 samples and 41,335 genes. Genes with less than 50 reads in 10% or more of the samples were removed. The rest were converted to geTMM, resulting in a genotype-transcript matrix with 15,205 genes across 470 samples. For both *A. thaliana* and *P. trichocarpa*, GENIE3 was run with 1000 trees and iRF-LOOP was run with 1000 trees at each iteration for five iterations.

2.3.5 Performance Evaluation

The competing models were evaluated by scoring their ability to generate PENs and GRNs that contain correct (gold) edges using several metrics. For the synthetic DREAM4 challenges, gold standard networks are provided and, because these data are simulated, the gold edges contain 100% of the truth with total accuracy. The gold standard networks

provided for the five DREAM4 datasets contained both true positives and true negatives, where 1 denoted a positive edge, and 0 denoted a negative edge, for all 9,900 possible PEN edges that can be inferred from all 100 unique input genes. The number of true edges (1's) per network ranged from 176 to 249, meaning that each gold network is heavily imbalanced towards negative edges.

For the DREAM5 challenges, the gold standard networks and expression datasets are also provided with some caveats. These gold standard networks are not complete with 100% accuracy, and for the two empirical gold standard networks only the high confidence regulatory interactions are included. The DREAM5 datasets were considerably bigger than the DREAM4 challenge datasets. The synthetic dataset (DREAM5 Network 1) contained 1,643 unique genes and transcription factors. The gold network contains 178 unique transcription factors connecting to 1,565 target genes. Since the DREAM5 goal was to infer a GRN rather than a PEN, the gold network contains 278,392 TF-to-target edges rather than all-to-all edges, of which 4,012 are positives (1's) and 274,380 are negatives (0's). The empirical *E. coli* gold network (DREAM5 Network 3) contains 152,280 edges connecting 141 unique transcription factors to 1,081 target genes. 2,066 of these edges are positives and 150,214 of these edges are negatives. The positive edges only include interactions in which there existed "strong evidence" in RegulonDB [57, 119]. The empirical *S. cerevisiae* gold network (DREAM5 Network 4) contained 227,202 edges connecting 114 transcription factors to 1,994 target genes. 3,940 of these edges are positives and 223,262 of these edges are negatives. The positive edges are based on an analysis of ChIP data and TF binding motifs with a stringent threshold to only consider edges with high confidence [82, 117, 119].

For evaluating the *A. thaliana* and *P. trichocarpa* PENs, we assembled gold standard networks using only verified true positive edges of known TF-to-gene and gene-to-gene relationships. The gold standard network for *A. thaliana* was manually constructed using two different sets of edges: a literature curated transcription factor to target network [87] and a gene to gene network constructed from AraCyc [128] reactions where two genes are connected if they share a common metabolite substrate or product. Using only the transcription factor to target edges as the gold network would neglect to score non-TF driven relationships that could be observed in a PEN inferred from all genes. Gold edges that linked to genes

that were not in the original expression matrix were subsequently dropped from the gold network. The gold TF-to-gene network contained 231 unique transcription factors, 469 unique target genes for a total of 948 directed edges, and the gene to gene network contained 2,383 unique genes and 17,715 directed edges. Thus, the overall *A. thaliana* PEN gold standard network contained 18,663 directed edges between 2,864 unique genes. The gold network for *P. trichocarpa* was constructed similarly. The literature curated transcription to target network from *A. thaliana* was mapped to *P. trichocarpa* using orthologs, and the gene to gene network was constructed from PoplarCyc reactions via the Plant Metabolic Network [74] similar to *A. thaliana*. Again, edges contained only genes that were found in the input expression matrix. The transcription to target network contained 96 unique transcription factors, 163 unique target genes for a total of 379 directed edges. The gene to gene network contained 1505 unique genes and 8889 directed edges. The overall *P. trichocarpa* PEN gold standard network contained 9268 directed edges for 1690 unique genes.

It is worth noting that these real-world gold standard networks are incomplete since they do not contain any true negative edges, are missing many known non-metabolic and non-TF driven relationships, are missing hitherto unknown biological relationships that may indeed be evident in the input data, and probably contain several biologically false positive relationships.

Sage Bionetworks provides scoring methods for both DREAM4 (as a python package) and DREAM5 (as a MATLAB script) that calculates the Area Under the Precision-Recall curve (AUPR), the Area Under the Receiver Operator Characteristic curve (AUROC), and p-values for AUPR and AUROC. The p-values represent the probability that the AUPR or AUROC that is achieved is better than a random permutation of the network edges that have been submitted to the official DREAM challenge. For DREAM4, all of the 9,900 network edges are used in the random permutation for each of the five networks, but in DREAM5 only the network edges of the submitted challenge networks are used. In the final scoring for DREAM4 all submitted edges are scored. For DREAM5 only the top ranked 100,000 edges are scored and the remaining edges are assumed to occur at random.

In order to summarize scores across all five DREAM4 networks, Sage Bionetworks provides an overall score, defined as

$$overall\ score = -0.5 \log_{10}(p_1 p_2)$$

where p_1 is the geometric mean of the five AURC p-values and p_2 is the geometric mean of the five AUROC p-values. The submissions to the official DREAM4 challenge are ranked based on the highest overall score. Similarly, for the DREAM5 networks, the overall score is defined as

$$AUPR\ score = \frac{1}{3} \sum_{i=1}^3 -\log_{10}(p_{AUPR,i})$$

$$AUROC\ score = \frac{1}{3} \sum_{i=1}^3 -\log_{10}(p_{AUROC,i})$$

$$overall\ score = \frac{AUPR\ score + AUROC\ score}{2}$$

where $p_{AUPR,i}$ is the AUPR p-value for network i , and $p_{AUROC,i}$ is the AUROC p-value for network i . Like DREAM4, the submissions to the official DREAM5 challenge are ranked based on the highest overall score. Although Sage Bionetworks has provided leaderboards for these two challenges, GENIE3 was rerun on these data sets to verify results with the updated codebase.

Implementing AUPR and AUROC comes with some caveats. AUPR and AUROC will penalize a false positive edge that does not appear in the gold standard network, but it is possible for edges to be ranked highly due to an uncharacterized yet true relationship between the two genes that is present in the dataset at hand. Since a 100% complete gold standard network is not possible in real empirical datasets, AUPR and AUROC are not ideal to use in these scenarios. Additionally, precision and recall at threshold k do not take into consideration the ranking of each edge. For example, it is possible at threshold k two networks can have the same number of true edges and false edges corresponding to the same precision and recall values, but the rank order of these edges can be wildly different. In addition to calculating the AUPR and AUROC of the empirical datasets, normalized Discounted Cumulative Gain (nDCG) [86] was calculated for every edge from one to k ,

where k is the number of true positive values in the gold standard network. This is defined as follows:

$$nDCG_{network,k} = \frac{DCG_{network,k}}{IDCG_{gold\ standard,k}}$$

$$DCG_{network,k} = \sum_{i=1}^k \frac{x}{\log_2(i+1)}$$

$$x = \begin{cases} 1 & \text{if edge is a true positive} \\ 0 & \text{if edge is a false positive} \end{cases}$$

$$IDCG_{gold\ standard,k} = \sum_{i=1}^k \frac{1}{\log_2(i+1)}$$

Using nDCG accounts for how early the true positive edges are found. If a true positive edge is found at a higher rank in the network than another true positive edge, the higher rank edge boosts the score more than the lower ranked true positive edge. This yields a better score if the true positive edges are found earlier in the ranking versus spread out across the whole list. Usually, the resulting networks are large and are thresholded before use, thus capturing true positive edges earlier in the ranking is necessary to retain them after thresholding.

To emphasize the biological significance of iterative RF in network prediction, we included an analysis of two select biological pathways from AraCyc and PlantCyc for *A. thaliana* and *P. trichocarpa* respectively, produced by GENIE3 and iRF-LOOP.

2.3.6 Random Intersect Trees (RIT)

The algorithm RIT [150] works by mining the node-split pathways in the forests resulting from iRF to find sets of features that occur consecutively along the pathways more than expected by chance. This suggests that the model uses those features in conjunction with each other in a potentially non-linear manner. RIT is able to efficiently discover feature interactions of any order if their joint importance is high enough. We ran RIT on the

resulting forests from the *A. thaliana* iRF-LOOP and identified a sample of highly prevalent sets of interacting genes.

2.3.7 Computational Resources

The computational resource used to run both the HPC implementation of iRF-LOOP [27] and GENIE3 was Summit. Summit is a supercomputer at Oak Ridge Leadership Computing Facility (OLCF) located at Oak Ridge National Laboratory. Summit is an IBM system that contains 4,608 nodes, each with two POWER9 processors and six NVIDIA Tesla V100 GPUs. Each POWER9 processor contains 22 cores that contains 4 threads each for a total of 176 threads per node. Each node contains 512 GB of DDR4 memory.

2.4 Results and Discussion

2.4.1 Replacing RF with iRF Improves Overall DREAM Scores

For the synthetic DREAM4 overall score metric, iRF-LOOP using 1000 trees per iteration (40.521) outperformed RF-LOOP as implemented by either the original GENIE3 or by the updated GENIE3 code [Table 2.1]. This was the case whether GENIE3 was run with 1000 trees, or with 5000 trees to match the total number of trees produced by iRF-LOOP. This suggests that the addition of iterations of random forest improves the overall quality of PENs inferred from these datasets, and that the increase in overall score observed for iRF-LOOP is not simply due to the increase in overall number of trees from 1000 to 5000.

The DREAM5 results showed a similar, but more dramatic improvement in the overall DREAM score metric when using iRF in place of RF for inferring GRNs with the subLOOP algorithm. iRF-subLOOP obtained a score over 50% greater than the best performing GENIE3 score [Table 2.1]. For synthetic Network 1 data, iRF-subLOOP produced a GRN with a p-value of $3.8\text{e-}269$, while GENIE3 produced a GRN with a p-value of $6.1\text{e-}121$ [Supplementary Table 3]. This resulted in an overall AUPR score of 91.829 for iRF-subLOOP and 47.097 for GENIE3. However, for the empirical datasets (DREAM5 Network 3 and DREAM5 Network 4) the raw AUPR and AUROC values for GRNs generated by

Table 2.1: This table depicts the overall scores for the DREAM4 *In Silico Size 100 Multifactorial* and DREAM5 *Network Inference* networks using the DREAM scoring system. The iRF-based algorithms outperform all three RF-based algorithms. This table also shows that simply increasing the number of trees in an RF-based model to match the total number of trees used in an iRF-based model does not account for the overall score increase seen in iRF-LOOP. Raw AUPR and AUROC values as well as their p-values can be found in Supplementary Tables 1-4.

DREAM Challenge	Base Learner	Algorithm	Number of Trees Per Iteration	Overall Score
DREAM4	RF	GENIE3 (original)	1000	37.428
DREAM4	RF	GENIE3 (new)	1000	39.375
DREAM4	RF	GENIE3 (new)	5000	39.446
DREAM4	iRF	iRF-LOOP	1000	40.521
DREAM5	RF	GENIE3 (original)	1000	40.279
DREAM5	RF	GENIE3 (new)	1000	43.329
DREAM5	iRF	iRF-subLOOP	1000	65.466

GENIE3 and iRF-subLOOP were highly similar [Supplementary Tables 1 and 2]. Thus the improvement of iRF-subLOOP over GENIE3 was primarily driven by the AUPR p-value component of the DREAM scoring metric for the synthetic DREAM5 Network 1 dataset.

2.4.2 iRF-LOOP on *A. thaliana* and *P. trichocarpa* Improves Early nDCG and AUnDCG

Since the DREAM scoring system does not adequately capture the rankings of the edges within the networks, applying nDCG to these DREAM networks and observing the changes over iterations may be more beneficial to determining whether iterations in random forests provides an improvement for all types of network prediction.

The nDCG scores for both GENIE3 and iRF-LOOP were calculated at every value of k from zero to the number of true positive values in the gold standard network, 18,663 for *A. thaliana* and 9,268 for *P. trichocarpa*. For the DREAM4 networks, this ranges from 176 to 249 edges. For the DREAM5 networks, this ranges from 2,066 to 4,012 edges. When we scored the DREAM4 and DREAM5 networks with nDCG, we found that the AUnDCG was greater in iRF-LOOP than GENIE3 for all expression networks except for DREAM5 Network 4 [Supplemental Figure 1]. For *A. thaliana*, iRF-LOOP obtained an AUnDCG of 91.897, which is a 1.75 fold increase over GENIE3’s AUnDCG of 52.694 [Figure 2.2]. The improvement in AUnDCG when using iRF-LOOP instead of GENIE3 was even more pronounced in the *P. trichocarpa* networks where iRF-LOOP scored a 18.672, a 2.5 fold increase over GENIE3. Thus, iRF-LOOP is more likely to rank true positive edges more highly than GENIE3. Furthermore, iRF-LOOP tends to have a higher early nDCG, i.e., the nDCG for the higher ranked edges of the network, than GENIE3.

2.4.3 iRF Boosts Important Edges and Improves Feature Selection

Adding iterations to RF-LOOP provides two major advantages. First is feature selection, which comes from the feature culling performed at each iteration. In Figure 2.3, the *A. thaliana* and *P. trichocarpa* networks both see a two-fold reduction in the resulting network

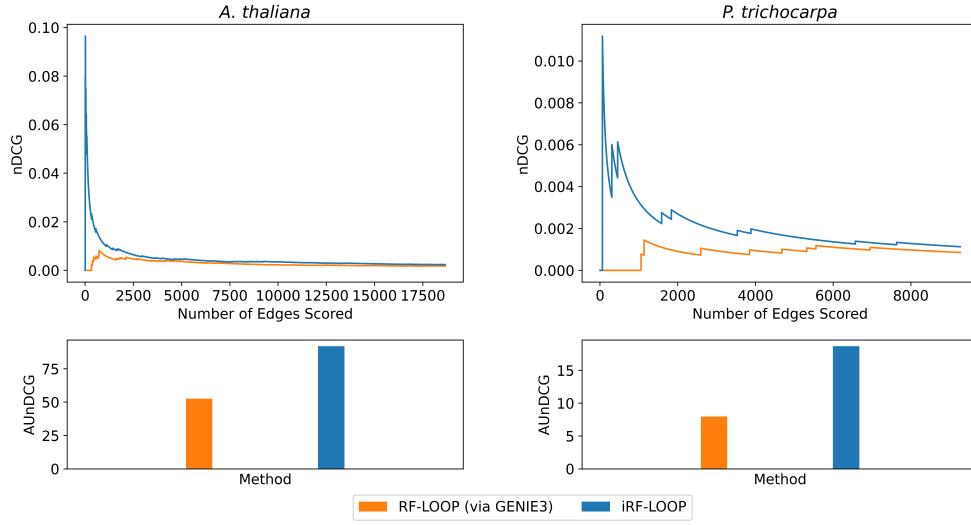


Figure 2.2: This figure depicts the nDCG scores for both *A. thaliana* and *P. trichocarpa* for both GENIE3 and iRF-LOOP as k , the number of edges scored, increases. The maximum k for each organism is equal to the size of the gold standard network. For both *A. thaliana* and *P. trichocarpa*, the early nDCG is higher in iRF-LOOP than GENIE3, and the overall AUnDCG for both organisms is also higher in the iRF-based algorithm. For *A. thaliana*, AUnDCG for iRF-LOOP and GENIE3 were 91.897 and 52.694 respectively. For *P. trichocarpa*, AUnDCG for iRF-LOOP and GENIE3 were 18.672 and 7.972 respectively. This suggests that iRF-LOOP outperforms GENIE3.

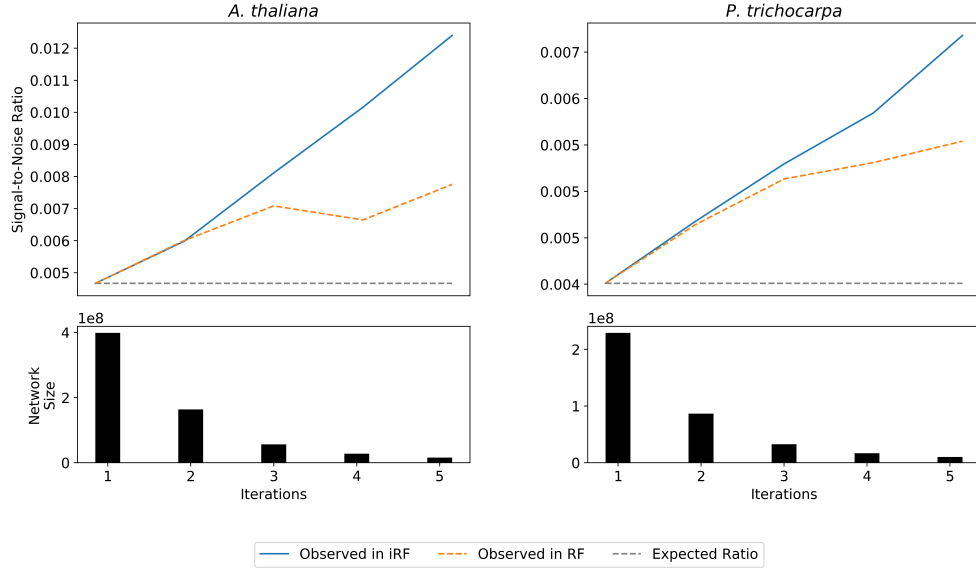


Figure 2.3: In this figure, the blue line depicts the truly observed signal-to-noise ratio for each network as the number of iterations in iRF increases. As the iterations increase, a number of edges are dropped from consideration due to their importance scores equaling zero. To confirm that the improvement in signal-to-noise ratio is not due to simply thresholding the networks, the RF network is thresholded to match the network size of the iRF network at each iteration and is depicted as the orange dashed line. For both *A. thaliana* and *P. trichocarpa*, the observed signal-to-noise ratio in iRF is greater than both what is expected from random and if the RF network is simply thresholded. This shows that the unsupervised thresholding from iRF provides an improvement over a simple manual thresholding.

sizes at each iteration as certain genes are deemed to have zero importance in the model predicting the target gene. The networks from the DREAM4 and DREAM5 challenges also see a decrease in network sizes, but with the smaller data sets used in this challenge this effect is not as extreme [Supplementary Figure 2]. With this decrease in network size comes an increase in the signal-to-noise ratio, where the percentage of edges kept that are true positive edges increases over iterations. This culling of noisy edges is a form of unsupervised thresholding. When comparing the effectiveness of this unsupervised thresholding in iRF-LOOP to manually removing the same number of lowest ranked edges from the GENIE3 network, it performs best in larger networks and data sets, such as the *A. thaliana* and *P. trichocarpa* networks [Figure 2.3].

The second major advantage of the iterative approach to RF is the boosting of important edges after each iteration. Since resulting networks are typically thresholded for use, it is imperative that true positive edges are ranked higher in the network than false edges. In Figure 2.4, the true positive edges within the first 100 and 500 ranked edges and their ranking per iteration are shown for *A. thaliana* and *P. trichocarpa* respectively (where a rank of 1 is the best possible). The improvement in ranking of true positive edges can clearly be seen for *A. thaliana*. There are no true positives in the first 100 edges after one iteration and the first true positive edge was not found until rank 325. After the second iteration, two true positives were found in the top 100 ranks, and as iterations increase these true positives shift up in the rankings and more true positives move into the top 100 edges. A similar pattern can be observed in *P. trichocarpa*, where there are no true positives in the first 500 edges after one iteration and the first true positive edge was found at rank 965. As more iterations are added, the true positives shift up the ranks, boosting the true edges to the higher rankings in the final network. This boosting of important edges can also be seen in the Precision-Recall curves, since the AUPR values increase over iterations for the *A. thaliana* and *P. trichocarpa* networks [Supplementary Figure 3]. While this effect is not as obvious for some of the smaller networks in the DREAM challenges, there is an overall trend of AUPR increasing as the number of iterations increases [Supplementary Figure 3].

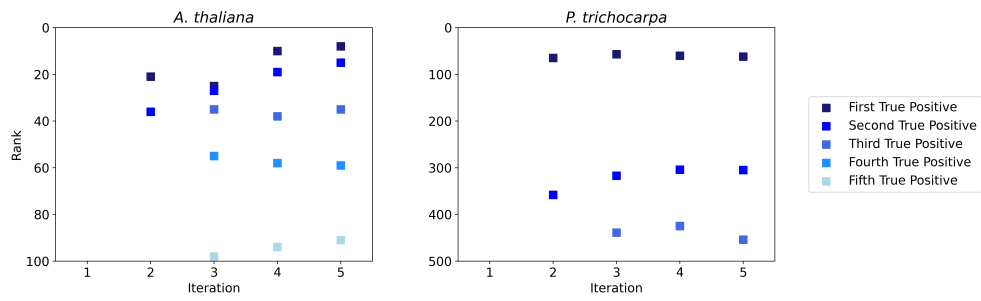


Figure 2.4: This figure shows the true positive edges, depicted as a square, for the top ranked 100 edges and 500 edges across each iteration for *A. thaliana* and *P. trichocarpa* respectively. For iteration 1 in *A. thaliana*, the first true positive edge is not discovered until rank 325. For iteration 1 in *P. trichocarpa*, the first true positive edge is not discovered until rank 965. The shifting of the true positive edges towards earlier rankings indicates the true positive edges moving up in the ranks of the final edge list and are more likely to be retained when thresholded.

2.4.4 PEN Edges and RIT Sets Capture Complex Biological Relationships in *A. thaliana*

Table 2.2 contains the top five ranked edges from the iRF-LOOP network on *A. thaliana*. None of these five edges are found in the gold standard network. However, many of these edges describe a relationship that is not as simple as gene-to-gene or regulator-to-target. For example, third ranked edge show connections between extensin genes and are associated with functions in the cell wall [146]. The edges ranked second and fourth represent the relationship between two genes with two different directionalities, and both are involved in the Aminoacyl-tRNA biosynthesis pathway [91]. The edge ranked fifth shows connections between the two genes CruA (AT5G44120) and CruC (AT4G28520), which both contribute to hexamer formation via an intermediary gene CruB. In the pathway CruA regulates CruB which regulates CruC [77, 81]. Of the five top edges that were not found in the gold network, all five of these edges had some sort of biological relationship shown in the literature. It is apparent that not only do PENs capture regulatory information in a biological system, but also capture other complex biological relationships.

Table 2.3 contains a sample of three highly prevalent sets from the RIT analysis on the resulting trees from the iRF-LOOP *A. thaliana* run. These three sets were chosen based on the three highest prevalence scores, which is dependent on the target gene and thus not comparable across iRF models. Duplicate sets were removed. None of the three sets existed as edges in the gold standard network. All three sets existed as edges in the iRF-LOOP network, both as Gene A to Gene B and Gene B to Gene A, though some had low importance scores. RIT was able to identify all three highly prevalent sets of genes as having some known biological relationship, such as existing in the same pathway or having the same classification. This shows that RIT combined with iRF-LOOP can yield new information that would not have been possible previously, and sheds light on higher order sets of interacting genes. Supplementary Tables 5 and 6 contains two samples of higher order sets, one of size four and one of size five, where known biological relationships were recovered among the sets.

Table 2.2: This table depicts the top five ranked edges from the PEN created using iRF-LOOP on *A. thaliana* expression data. None of the edges were true positive edges in the gold standard network, but have some existing biological relationship. This table shows that PENs produced with iRF-LOOP capture relationships beyond simple gene regulation.

Rank	Importance Score	Start Gene — End Gene ID	Start Gene Function	End Gene Function	Relationship
1	0.970	AT5G05430 — AT5G05420	RNA-binding protein — FKBP-like peptidyl-prolyl cis-trans isomerase family protein		Endosperm specific genes [34]
2	0.962	ArthCt101 — ArthCt092	Transfer RNA — Transfer RNA		Noncoding transfer RNAs, both in Aminoacyl-tRNA biosynthesis pathway [91]
3	0.959	AT3G28550 — AT1G23720	Proline rich extensin like family protein — Proline rich extensin like family protein		EXT genes in cell wall [146]
4	0.949	ArthCt092 — ArthCt101	Transfer RNA — Transfer RNA		Noncoding transfer RNAs, both in Aminoacyl-tRNA biosynthesis pathway [91]
5	0.941	AT5G44120 — AT4G28520	RmlC-like cupins superfamily protein — Cruciferin 3		Seed storage proteins that are down and up regulated together, both contribute to hexamer formation [77, 81]

Table 2.3: This table depicts a sample of three highly prevalent sets determined using RIT on the resulting paths from iRF-LOOP on *A. thaliana* expression data. All three of these sets have been found to have some known biological relationship. None of these sets were found in the gold standard network, while some were found in the iRF-LOOP PEN as edges with low importance scores. This shows that RIT combined with iRF-LOOP can be used to discover or validate gene to gene relationships in sets.

Prev- alence	Gene ID	Function	Target	Target Function	Relation- ship
0.966	AT3G44630 — AT3G44400	Both disease resistance protein (TIR- NBS-LRR class) family	AT3G44480	Encodes a TIR- NB-LRR R-protein RPP1 that confers resistance to Per- onospora parasitica	All three encode disease resistant proteins, regulation is linked in some ex- periments [161]
0.964	AT3G28550 — AT3G54590	Proline- rich extensin- like family protein — Hydroxy- proline- rich glyco- protein	AT3G54580	Proline- rich extensin- like family protein	EXT genes in cell wall [146]
0.957	AT3G44480 — AT3G44630	Both disease resistance protein (TIR- NBS-LRR class) family	AT3G44400	Disease resistance protein (TIR- NBS-LRR class) family	All three encode disease resistant proteins

2.4.5 Replacing RF with iRF Captures More Members of Metabolic Reaction Pathways

To compare GENIE3 and iRF-LOOP on the *A. thaliana* and *P. trichocarpa* expression data in this section, the GENIE3 PEN is thresholded to match the size of the iRF-LOOP PEN which was automatically reduced in size due to the culling of edges of zero importance per iteration. For *A. thaliana* this is 15,719,197 edges and for *P. trichocarpa* this is 10,180,484 edges.

There are 1,996 unique RXN-to-RXN pathways found in the gold standard AraCyc gene-to-gene edges. 535 of these pathways were discovered using GENIE3 on the *A. thaliana* expression data, comprising 1,076 edges in the network. For iRF-LOOP 688 RXN-to-RXN pathways were discovered comprising 1,776 edges in the network. This increase shows that the true positive edges were elevated in the rankings. Additionally, the average percentage of edges discovered for each RXN-to-RXN pathway in GENIE3 was 8.279%, where the percentage in iRF-LOOP is 12.878%. Figure 2.5 shows this improvement in edges discovered in a pathway in *A. thaliana*. The pentose phosphate pathway is shown in Figure 2.5A, containing 2 RXN-to-RXN pathways, EC 1.1.1.49 to EC 3.1.1.31 and EC 3.1.1.31 to EC 1.1.1.44. Figure 2.5B shows the gold edges expected in this pathway, and Figure 2.5C shows the edges discovered using iRF-LOOP compared to Figure 2.5D which shows the edges discovered using GENIE3. While not all edges from the gold standard pathway are discovered in iRF-LOOP, it is possible that iRF-LOOP is discovering the most important edges. Additionally, iRF-LOOP discovers more edges than GENIE3 does.

There are 1,075 unique RXN-to-RXN pathways found in the gold standard PlantCyc gene-to-gene edges used for the *P. trichocarpa* analysis. Similar to the *A. thaliana* pathways, GENIE3 discovered 244 RXN-to-RXN pathways using 523 edges, while iRF-LOOP discovered 269 RXN-to-RXN pathways using 625 edges, again showing the boosting of true positive edges. The average percentage of edges discovered for each pathway also increased from GENIE3 to iRF-LOOP, from 5.270% to 6.357%. Figure 2.6 contains the Phenylpropanoid biosynthesis pathway, an important pathway used in bioenergy research. Since only a little over half of the total possible pathways were discovered using iRF-LOOP,

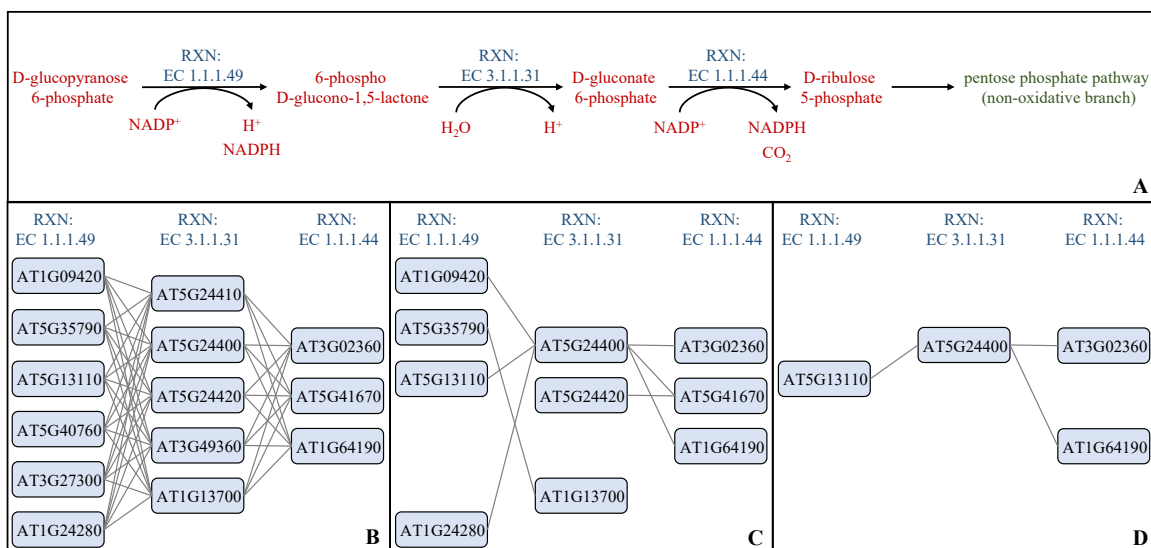


Figure 2.5: This figure depicts the compounds, RXNs, and genes included in the pentose phosphate pathway in *A. thaliana*, and the genes that are recovered by GENIE3 and iRF-LOOP. The resulting PEN from GENIE3 was thresholded to match the same size as the iRF-LOOP PEN. A) Steps in the pathway encompassing the three RXNs according to AraCyc. B) All of the possible edges between the known genes in neighboring steps in the pathway. C) The edges recovered by iRF-LOOP. D) The edges recovered by GENIE3.

Figure 2.6B only shows the single RXN-to-RXN pathway discovered in iRF-LOOP, EC 6.2.1.12 to EC 2.3.1.133. Figure 2.6C shows the edges discovered by iRF-LOOP and Figure 2.6D shows the edges discovered by GENIE3. Again, not all edges are discovered in iRF-LOOP, but it is possible that only the most important edges are discovered, and it is clear that iRF-LOOP discovers more edges than GENIE3.

2.5 Conclusion

The use of whole networks in biology is unwieldy, an all-to-all directed network of tens of thousands of genes quickly becomes hundreds of millions of edges. Typically either only the top k edges or top n percent of edges are selected to be used in further studies or computational analysis. The value to use for thresholding is often difficult to discern, the goal being capturing high quality edges and culling noise [136]. To determine whether to use a network created by one method over another, the chosen network must have the correct edges found higher in the ranking of all edges than the network not chosen. The better algorithm must boost the true edges up in the ranking so that they are retained after thresholding. Culling noisy and true negative edges from the network entirely helps to ameliorate the thresholding problem.

The *A. thaliana* and *P. trichocarpa* empirical networks were many orders of magnitude larger than the DREAM challenge networks, thus the effect of this unsupervised thresholding can be observed. At the end of each iteration in the empirical networks around half of the edges were dropped entirely from the network, removing noisy edges from the final network. When comparing the final network after five iterations of iRF to the network resulting from one iteration of RF, a higher number of total RXN pathways were discovered in the iRF network, as well as a higher percentage of the individual RXN pathways being discovered. This process of discovering pathways is comparable to nDCG, where iRF ranks the true positive edges higher in the network than RF.

PENs serve a different purpose than GRNs. GRNs focus on primarily regulatory relationships while PENs capture directional gene to gene relationships other than the direct regulation of one gene on another. RIT as applied to iRF tree paths can be used to create

new layers in addition to PENs and GRNs and captures unique set relationships beyond gene to gene that are likely quite common in biological systems. PENs and other non-regulatory layers can provide expanded functional content for graph learning algorithms and other downstream applications such as Random Walk with Restart and other lines of evidence methods [166].

iRF can be used successfully in contexts beyond gene expression analysis. RF-LOOP tools such as GENIE3 have been a benchmark for other network inference problems, such as networks of transcriptomic and proteomic data [124]. iRF can be used beyond building networks, since RF has also shown to be effective in classification problems [31]. iRF contains a multitude of advantages that could replace the use of RF in biology and beyond.

Chapter 3

A Predictive Model of COVID-19 County Level Case Fatality Rate

3.1 Abstract

The COVID-19 pandemic has caused significant damage across the globe, causing millions of deaths worldwide and impacting health care systems, economies, and political systems for decades to come. While many studies have been conducted to provide a mechanistic and clinical understanding of COVID-19 infection, few studies have captured the relationship between the patient’s immediate environment and their likelihood of surviving the initial infection. Here we use iterative Random Forest (iRF) to predict the Case Fatality Ratio (CFR) at the county level in the United States using demographic and socioeconomic descriptors as well as the surrounding climatic attributes and public health measures that were put in place for each county. We show that of these classes of county level descriptors, the most important class of features were those that related to the climate of the county and captured the changing of the seasons in the United States. This information can be used to better understand which areas of the country will see increased levels of CFR and where resources could be allocated to lessen the severity of disease.

3.2 Introduction

The Coronavirus disease 2019 or COVID-19 is caused by the respiratory virus SARS-CoV-2 and was first discovered in Wuhan, China in December of 2019 [101]. Since then, the ongoing COVID-19 pandemic has had a significant impact on the world’s healthcare systems and economies. COVID-19 symptoms range from mild cold and flu-like symptoms to severe symptoms that can result in death. Even for those patients who survive the initial COVID-19 infection, some individuals continue to experience symptoms for months which is often called long covid. The COVID-19 pandemic has resulted in nearly 5.5 million deaths worldwide as of January 1, 2022 [38]. While there are four vaccines that are approved for use worldwide—Pfizer-BioNTech, Moderna, Janssen or Johnson & Johnson, and Oxford-AstraZeneca vaccines—the moral dilemma of how these vaccines are distributed has become an ongoing debate and many countries still have sub-optimal vaccination rates.

The resulting vulnerable reservoir has allowed for continued, large scale circulation of SARS-CoV-2 and the concomitant mutation and likely recombination that has led to new variants of concern. Additionally, the rise of these new variants threaten the existing immunity within a community, whether it was obtained through natural infection or through immunization.

Many studies have been performed to capture the comorbidities that are known to impact the survival of COVID-19 infection. This includes diseases such as chronic obstructive pulmonary disease (COPD), malignancy, diabetes, hypertension, chronic kidney disease, cerebrovascular disease, and cardiovascular disease [43, 99]. Other comorbidities related to the demographic makeup of the patient have been found to hinder the survival of COVID-19 infection, such as being male and of older age [43]. Other studies have found that there exist environmental predictors that impact survival of COVID-19 infection. Seasonality, or the time of year the patient contracts the infection, influences survival due to the higher viral load in the winter months as people move indoors more often and congregate for the holidays [114]. Interestingly, low air quality and high air pollution found in urban environments has also been found to impact survivability of COVID-19 infections [141].

Before the three COVID-19 vaccines became available via the Food and Drug Administration Emergency Use Authorization at the end of 2020, a total of 353,000 individuals in the United States had died from COVID-19 [38]. In mid 2021, vaccination rates stagnated and, according to the Centers for Disease Control and Prevention, 21.8% of the U.S. population over the age of 5 still remains unvaccinated as of January 1, 2022 [20]. While many studies have focused on the patient’s individual demographic makeup and medical history when determining survivability of COVID-19, it is important to also analyze the effect of an individual’s environment on the likelihood of surviving COVID-19 if contracted. The finest granularity for daily cases and deaths in the U.S. is at the county level, supplied by the Johns Hopkins University Center for Systems Science and Engineering 2019 Novel Coronavirus Visual Dashboard [38]. However, the public release of this information has become a point of contention and some states have stopped updating their COVID-19 dashboards entirely [10]. The release of vaccines at the end of 2020 has improved survivability of COVID-19 considerably, however vaccine data has not been released to the public with the fine granularity that is found in the cases and deaths. Thus, this study focuses only on the cases

and deaths observed in 2020. As such, these results are relevant to the large number of individuals who are still unvaccinated both in the U.S. and around the world.

We used an explainable-artificial intelligence (X-AI) tool, iterative Random Forest (iRF) [9, 27], to determine which environmental factors were the most important in the predictive model of COVID-19 patient outcomes. Environmental factors are various measurements of the patient’s surroundings, including but not limited to specific climate or weather related measurements and the socioeconomic makeup of the community. Throughout this study we have found that climatic factors, especially those related to seasonality, were more important than COVID-19 related policy, demographic, and socioeconomic factors. This analysis also included an exploration into which environmental factors are related to the highly important climate factors.

3.3 Methods

In this analysis, we used iRF models to determine which environmental county level measurements, or features, are most important in predicting a county’s Case Fatality Ratio, or COVID-19 outcomes. The X-AI tool iRF was used for feature engineering, as well as to measure accuracy of the fit of the model to the data provided. iRF in this case produced two outputs, a list of ranked features based on level of importance in splitting the data set within the decision trees, and Mean Absolute Error accuracy scores that depict how accurate the model predicts the Case Fatality Ratio. iRF was also used in the algorithm iRF-Leave One Out Prediction (iRF-LOOP) [27] to create all-to-all connected networks of the county features. Thus, while iRF on its own identifies important features in predicting COVID-19 outcomes, the all-to-all network produced by iRF-LOOP was used to explore feature-to-feature relationships between the important features and other features within the model.

3.3.1 Environmental Data and Data Pre-Processing

The patient’s environment was captured using three classes of features, the county’s climate, demographic and socioeconomic makeup, and the COVID-19 policies and public health

measures that were in place in the county. In order to boost the model accuracy, the data was pre-processed to remove any highly correlated features.

Climate Features

The climate features included two groups, static measurements and dynamic monthly measurements. The monthly measures were comprised of non-static features such as monthly average solar radiation and monthly average cloud cover. This category included 12 separate measurement types for a total of 144 features. The 41 static features included features such as percent forest cover and soil acidity. All told, this was a total of 185 climate and weather related features, mapped to 3,233 counties in the United States [46, 47, 48, 85, 167, 186, 196, 197]. Details for these features can be found in Supplementary Tables 7 and 8.

Details for the climate data and pre-processing of this data was provided and completed by Jared Streich.

Census Features

The census features were comprised of measurements from the 2019 American Community Survey produced by the United States Census Bureau [170]. These features included demographic, socioeconomic, and housing features such as the Gini Index of Income Inequality and breakdowns of populations based on race, ethnicity, age, and gender. These 1606 features were captured using the R package `tidycensus` [176]. Many of these features were normalized in order to reduce bias from more densely populated counties. Two additional features were calculated, the population density of people per square mile and the ratio of water area to land area. In total, 1608 demographic and socioeconomic features were mapped to 3,220 counties in the United States. Details for these features can be found in Supplementary Tables 9-12.

COVID-19 Policy and Public Health Features

The COVID-19 policy and public health related features contained a variety of features such as whether religious gatherings were exempt from stay at home orders and the percentage of

days that evictions were halted [164, 171]. These features were originally state level features that were converted to county level. In total there were 48 such features for 3,142 counties. Two of the features captured mask usage, both the chance that a single individual was wearing a mask and an estimate of the number of people wearing a mask in a given county. Ten of the features were binary, such as whether the face mask mandate was enforced by a fine and if liquor stores remained open. The remaining 36 non-binary features that have start and stop dates were converted to the percentage of days during the given time period where policy was in place, such as the percentage of days that the county had a stay at home order. Details for these features can be found in Supplementary Tables 13 and 14.

Correlation Analysis

Highly correlated features may negatively impact the accuracy of the iRF model, since the importance of one association is split across two correlated features, decreasing their overall importance of this association. One step in the data pre-processing to account for this phenomenon was to select one of the highly correlated features to be a representative of the other features, while the other features were dropped. Pearson’s Correlation Coefficient was calculated between all features for that particular category of predictive models. Feature pairs with a correlation coefficient of greater than or equal to 0.90 or less than or equal to -0.90 were considered highly correlated. Details on the number of features retained in each model can be found in Supplementary Table 15.

3.3.2 COVID-19 Y Vector

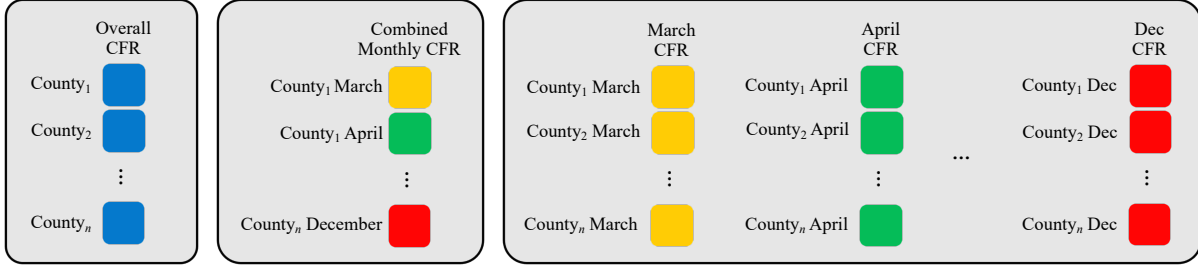
The Case Fatality Ratio (CFR) was used as the dependent feature to determine the impact of COVID-19 on the county. There are two versions of this vector that was used, one which captured the CFR between March 2020 and December 2020, and one which captured the CFR at each month by using the first and last days of the month as the start and end dates. The number of confirmed cases and confirmed deaths was collected from the Johns Hopkins University COVID-19 Data Repository [38]. Any county with less than 10 cases was removed from this analysis. The CFR was calculated as the proportion of deaths to cases within the

time frame specified. The overall CFR between March 2020 and December 2020 was defined for 3,226 counties. In addition, the number of counties where monthly CFR was defined ranges from 796 counties in March to 3,180 counties in December.

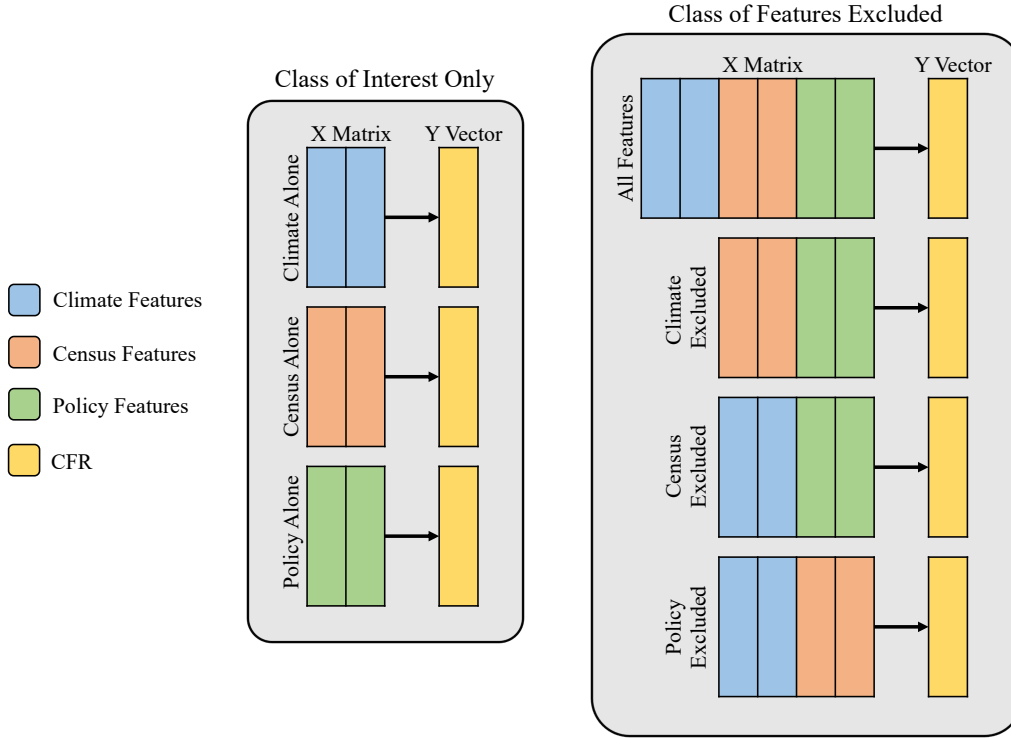
3.3.3 iRF Predictive Models

Three categories of iRF predictive models were used in this study. The first category was a single iRF model predicting the CFR from March 2020 to December 2020, where the features used in the model captured the environment over the course of the year. This category of iRF models showcased what environmental features were the most important in predicting the outcome of those who became ill at any point throughout the pandemic. The second category was a single iRF model predicting the monthly CFR from March 2020 to December 2020, where the features used in the month captured the environment in each month. This category of iRF models captured seasonal effects or the effects of features that change over time, such as the COVID-19 policy features and the seasonal climate. The final category of iRF models included ten separate monthly iRF models, where each model contained the samples for the month predicting only the CFR for that month. This category of iRF models captured whether a feature or a group of features has a larger impact in one month versus another, and determined whether the model has better accuracy at different times of the year. These three categories of CFR Y vectors are showcased in Figure 3.1a.

Within each category of iRF models, there was a set of seven individual iRF models in which each model contains different groupings of features. All seven of the models predicted the CFR feature. The first three models were the climate features only, census features only, and the COVID-19 policy and public health features only. This tested the accuracy of the groups of features in predicting the CFR on their own. The fourth iRF model contained all of the environmental, climate, census, and policy features to predict the CFR. This model was used to compare accuracy scores to the other models in which a group of features was removed. The final three iRF models contained combinations of the three feature groups, census and policy only, climate and policy only, and climate and census only. This leave-one-set-out analysis was compared to the all features model to determine the overall impact



(a) The three categories of CFR Y Vectors.



(b) The seven types of iRF Models.

Figure 3.1: This figure depicts the different categories of CFR Y vectors, as well as the seven individual iRF models that contain different groupings of features. The CFR vectors include the overall CFR between March 2020 and December 2020, the combined monthly CFR vectors that include the monthly CFR measurements for every county merged into one vector, and the ten individual monthly CFR vectors. The types of iRF models were separated into two categories, one which compared the models that only contain the class of interest, and one that compared the models that exclude one class or set of features at a time. The latter category also includes an all features included model to determine whether the model performs better or worse with the exclusion of the class of interest. In total, 48 iRF models were used in k -fold cross validation to produce comparable MAE scores and the top important features.

on the accuracy of the set of features removed. These different iRF model types are shown in Figure 3.1b. In total, 84 iRF predictive models were included in this analysis.

iRF k -fold cross validation was used to obtain accuracy scores for each of the iRF predictive models. For this analysis, 5-fold cross validation was repeated ten times for a total of 50 models. In each model, the data was split with 80% of the data used as training and 20% of the data used as testing. The accuracy of the model was determined using Mean Absolute Error (MAE). The MAE accuracy scores was then compared across the iRF models that use different sets of features. Comparing the iRF models that contain only one set of features (climate alone, census alone, and policy alone) will shed light to which class of features is the most accurate at predicting COVID-19 outcomes. Comparing the models that contain only two of the three sets will determine which sets yield the highest accuracy when combined together. Comparing these models to the model that contains all three features will determine which set is least important in determining COVID-19 outcomes.

3.3.4 iRF-LOOP Network Creation

iRF-Leave One Out Prediction (iRF-LOOP) [27] is used to create all-to-all predictive networks to determine which features are related to other features. This relationship goes beyond the type of correlation that is commonly used in Pearson Correlation-based networks, and can capture relationships that may not be as simple as a linear relationship. While the iRF predictive models described above depict which features are most important in predicting COVID-19 CFR at the county level, an all-to-all network produced by iRF-LOOP will show which of the important features are highly related and whether a single feature in particular is a proxy for a different feature.

iRF-LOOP was run on the merged X matrix of the overall March 2020 to December 2020 climate, census, and COVID-19 policy features with the highly Pearson correlated features removed. With 1,461 features and 3,112 samples, or counties, the resulting iRF-LOOP all-to-all network contained 1,608,516 non-zero edges. This network was thresholded to the top 0.5% of non-zero edges for the network exploration section of this analysis.

3.3.5 Computational Resources

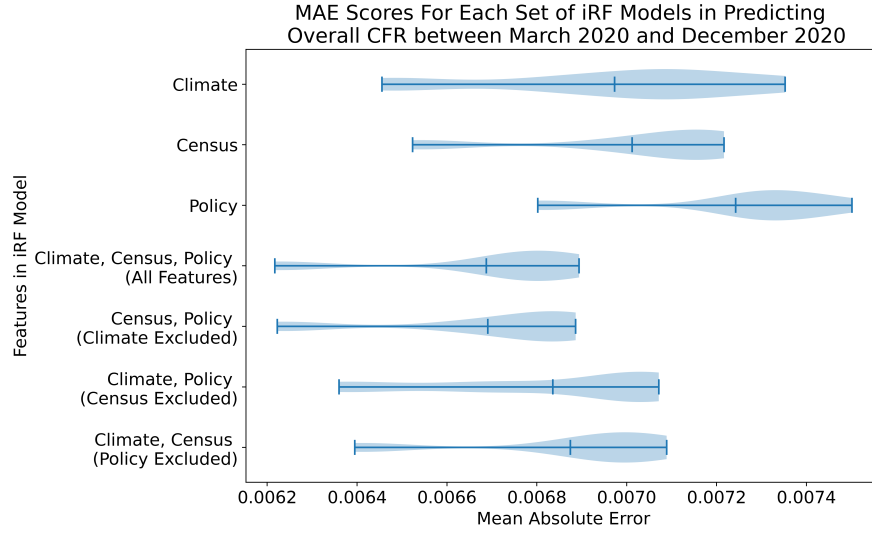
The computational resource used to run iRF on these datasets was the Summit supercomputer located at Oak Ridge National Laboratory in the Oak Ridge Leadership Computing Facility (OLCF). Summit is an IBM system that contains 4,600 nodes. Each node contains two POWER9 processors and six NVIDIA Tesla V100 GPUs. Each POWER9 processors contain 22 cores with 4 threads each for a total of 176 threads per node and 512 GB of DDR4 memory.

3.4 Results and Discussion

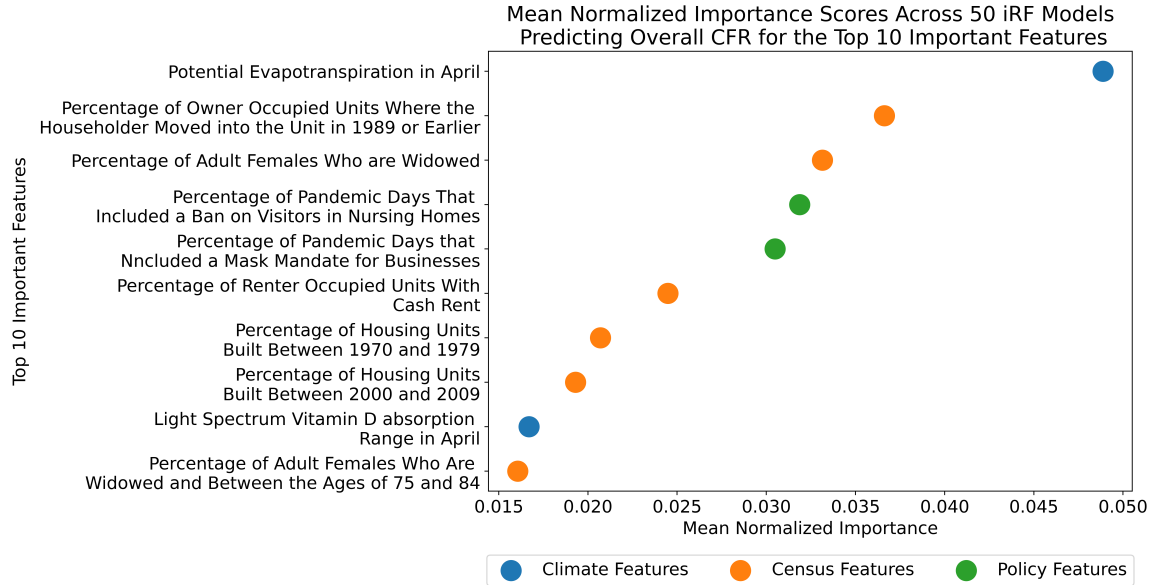
3.4.1 Climate and Demographic Features are Important in Predicting Overall CFR

When comparing the MAE scores for the climate alone, census alone, and policy alone iRF models that are predicting the overall CFR from March 2020 to December 2020, the best performing model was the climate alone model with an average MAE of $6.973\text{E-}3$ while the worst was the policy alone model with an average MAE of $7.243\text{E-}3$ [Figure 3.2a]. This may suggest that the surrounding climate of the county in which the patient resides may impact their COVID-19 outcome. Additionally, this may suggest that COVID-19 related policies had a smaller effect, but since policies were relaxed considerably before the fall surge in 2020 this may not be accurate, and may change when the iRF model is split into windowed monthly models. All of the models that excluded one feature set out of the iRF model performed worse than the model that contained all of the feature sets. However, the iRF model that excludes the climate features performed similarly to the model that contains all of the features, with average MAE scores of $6.688\text{E-}3$ for the model of all features and $6.691\text{E-}3$ for the model excluding the climate features [Figure 3.2a]. This may suggest that climate features are not as important as the census and policy features combined. However, this changes when the iRF model is split into windowed monthly models (see below).

For the iRF model with climate, census, and policy features predicting the overall CFR for March 2020 to December 2020, the top ten important features are depicted in Figure



(a) MAE scores for the iRF models predicting overall CFR.



(b) Top ten important features for the iRF model predicting overall CFR.

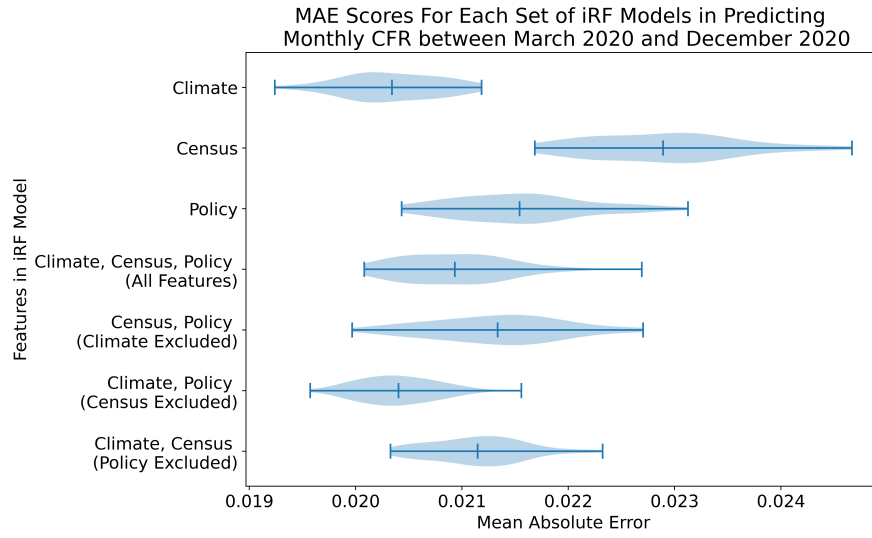
Figure 3.2: This figure contains (a) violin plots of MAE scores of the seven iRF models and (b) the top ten features for the model containing all features all predicting overall CFR between March 2020 and December 2020. The climate alone model had the best MAE compared to the census and policy alone models, while the climate excluded model performed the best when compared to the census excluded and policy excluded model. The climate excluded model also performed similarly to the all features included model. This implies while the climate features are required for a highly accurate model, the inclusion of the census features and policy features together also produces an accurate model. The top important features showcase some climatic seasonal features, but many of the features were demographic, capturing vulnerable populations, such as those in older age groups and those that are most likely to be protected by policy features such as banning visitors in nursing homes.

3.2b. The top important feature, potential evapotranspiration in April, was one of the highly correlated features that was chosen as a representative feature for 27 different features, such as the average temperature, the average minimum temperature, and average maximum temperature for many different months throughout the year. This could be capturing the seasonality of COVID-19, suggesting that outcomes of survival are influenced by the time of year that the patient becomes sick [114]. Many of these features may be capturing the counties that contain a high percentage of individuals over the age of 50, such as the percentage of home owners that moved into their home 1989 or earlier and the percentage of adult widowed females. Also among the top features includes the vitamin D absorption range of UV light in April, which suggests that vitamin D could affect outcomes in COVID-19 patients [5]. Two policy features were included in the top 10 features, one in which visitors were banned from nursing homes and one in which masks were mandated in public spaces. This may suggest that these policy measures were indeed protecting the more vulnerable populations.

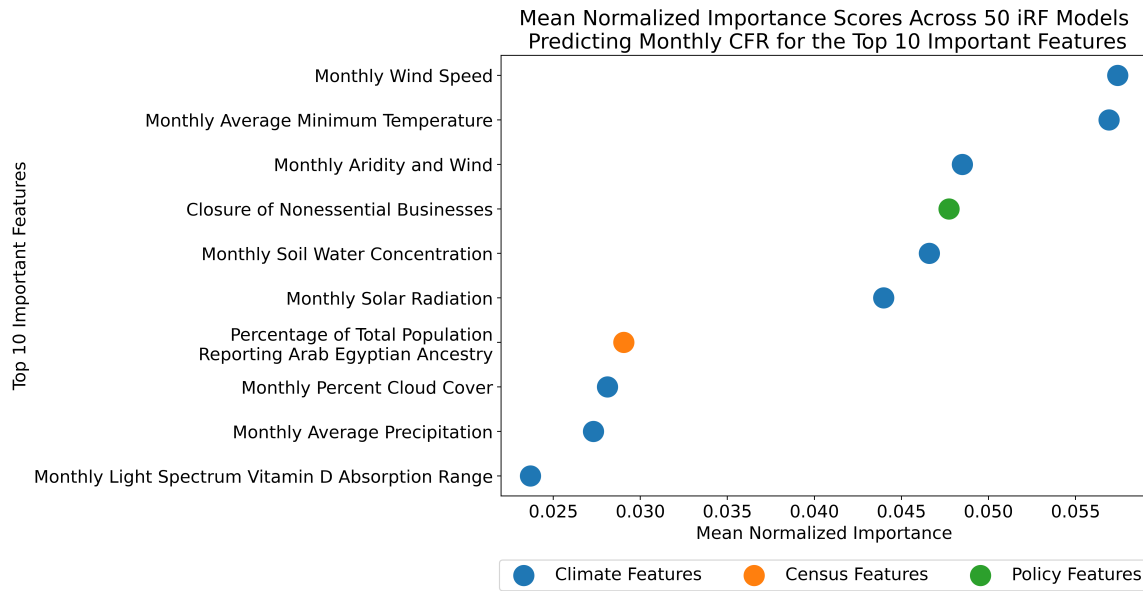
Thus, when predicting the overall CFR, there is evidence to conclude that climate features produce highly accurate iRF models and is important in determining the important environmental features in predicting COVID-19 outcomes. While the demographic and socioeconomic census features are top important features, these mostly captured the vulnerable populations and may change as the model is split into monthly windows.

3.4.2 Climate Features are Important in Predicting Monthly CFR

When comparing the MAE scores for the climate alone, census alone, and policy alone iRF models that are predicting the monthly CFR between March 2020 to December 2020, the best performing model was again the climate alone model with an average MAE of $2.034E-2$ while the worst was the census alone model with an average MAE of $2.289E-2$ [Figure 3.3a]. This reinforces the finding that the surrounding climate of the patient is crucial in predicting COVID-19 outcomes. Interestingly, the census alone model performing the worst implies that the demographics and socioeconomic status of the county in which the patient resides has little impact on the COVID-19 patient outcomes. This is supported by the fact that the climate and policy iRF model that excludes the census features outperforms all



(a) MAE scores for the iRF models predicting combined monthly CFR.



(b) Top ten important features for the iRF model predicting monthly CFR.

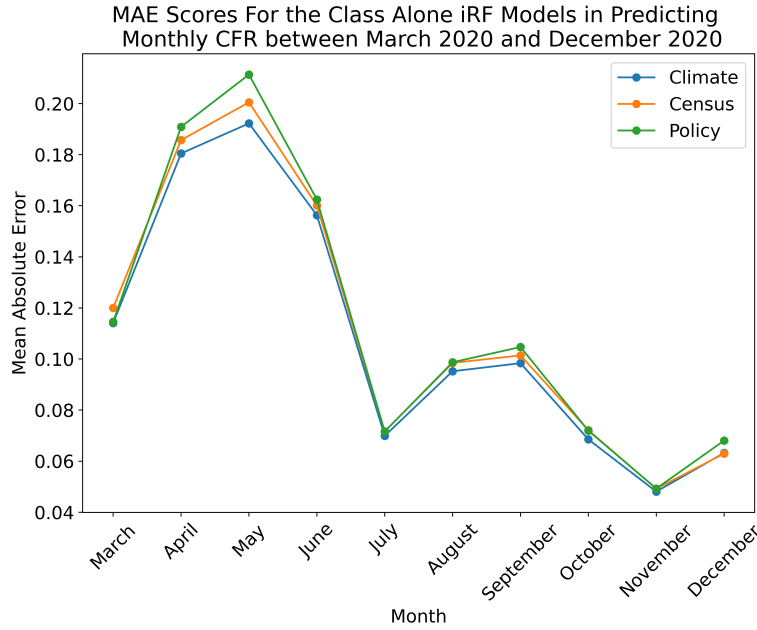
Figure 3.3: This figure contains (a) violin plots of MAE scores of the seven iRF models and (b) the top ten features for the model containing all features all predicting the monthly CFR between March 2020 and December 2020. The climate alone model had the best MAE compared to the census and policy alone model, while the climate excluded model performed the worst when compared to the all features model and the models where one feature set was excluded, reiterating that the climate features produce the most accurate models. Within the top important features include many climatic seasonal features, suggesting the the most important factors when predicting patient survivability is the time of year in which they contract infection.

of the models that excluded one feature as well as the iRF model that included all of the features, with an average MAE of $2.040\text{E-}2$ [Figure 3.3a]. Additionally, the iRF model that excludes the climate features performs the worst with an average MAE of $2.134\text{E-}2$ which implies again that climate is the most important set of features when predicting COVID-19 outcomes.

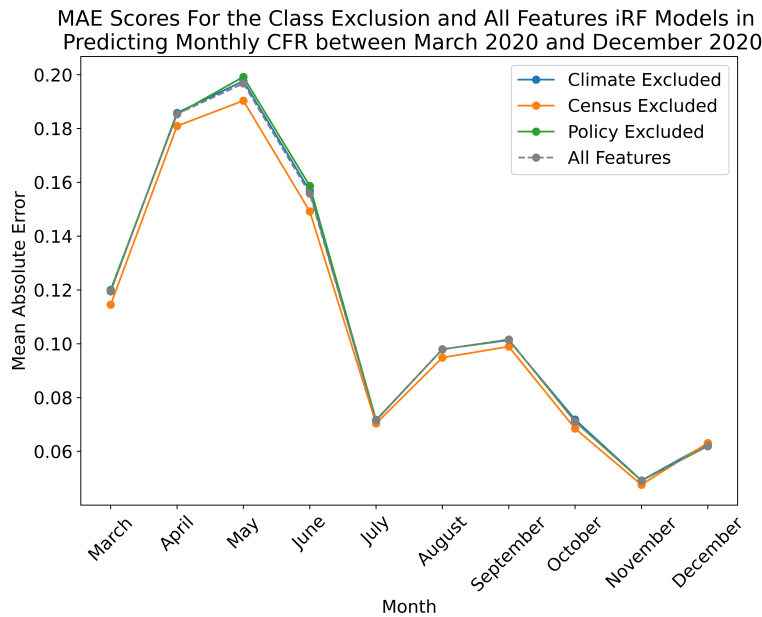
Figure 3.3b depicts the top ten important features from the iRF model that includes all three features sets predicting the monthly CFR for March 2020 through December 2020. Unlike the overall CFR iRF model, eight of the top ten features were climate related features. Many of these features are also seasonal, such as monthly average minimum temperature and monthly average precipitation. This again implies that COVID-19 outcomes may be seasonal. Additionally, another important climate feature was the monthly vitamin D absorption range and other solar related features such as monthly percent cloud cover and monthly solar radiation. This reinforces the finding that vitamin D is affecting COVID-19 outcomes.

In addition to the monthly CFR Y vector contained in one iRF model, the average MAE scores for the single monthly models are available in Figure 3.4 to view the change in accuracy over time as the monthly window changes. Consistently across all months, the best performing iRF model of those that contain only the feature set of interest was the climate alone model, while the worst was the policy alone model 3.4a. Like the iRF model that contains all of the monthly CFR values, the iRF model that excludes census outperforms the climate excluded, policy excluded, and all features models 3.4a. This showcases once again that climate is the most important set of features when predicting COVID-19 outcomes, but that climate and policy in combination is also beneficial in predicting these outcomes. Additionally, the predictions become more accurate in the latter half of the year, likely due to the fact that there are more county samples in the model as the SARS-CoV-2 virus spread across the country.

The overall trend that is observable in all three of these categories of iRF models—predicting overall CFR, monthly CFR, and windowed monthly CFR—is that climate was the most accurate and thus the most important set of features when predicting COVID-19 outcomes at the county level. While climate alone was shown to produce more accurate predictions, there is evidence to conclude that climate in combination with the policy features



(a) MAE scores predicting monthly CFR in the climate alone, census alone, and policy alone iRF models.



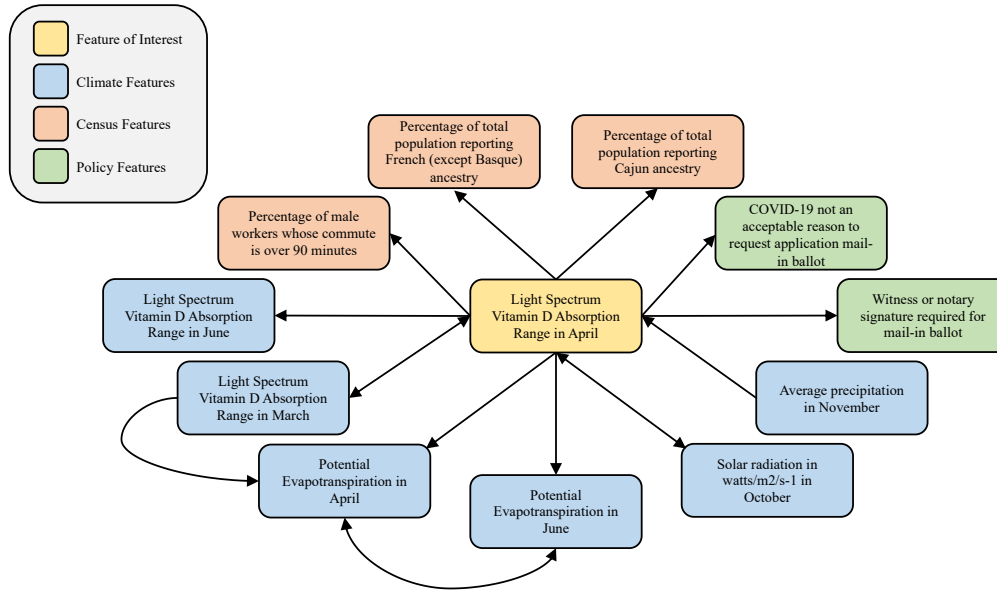
(b) MAE scores predicting monthly CFR in the all features, climate excluded, census excluded, and policy excluded iRF models.

Figure 3.4: This figure depicts the MAE scores for the iRF models predicting the windowed monthly CFR rates as they change over time. In (a), of the models containing climate alone, census alone, and policy alone the best performing was the climate alone model while the worst performing was the policy alone model. In (b), of the models containing all the features, climate excluded, census excluded, and policy excluded, the best performing model was the census excluded model while the remaining models performed similarly. Thus, the climate features are critical in creating accurate predictions for the windowed CFR iRF models.

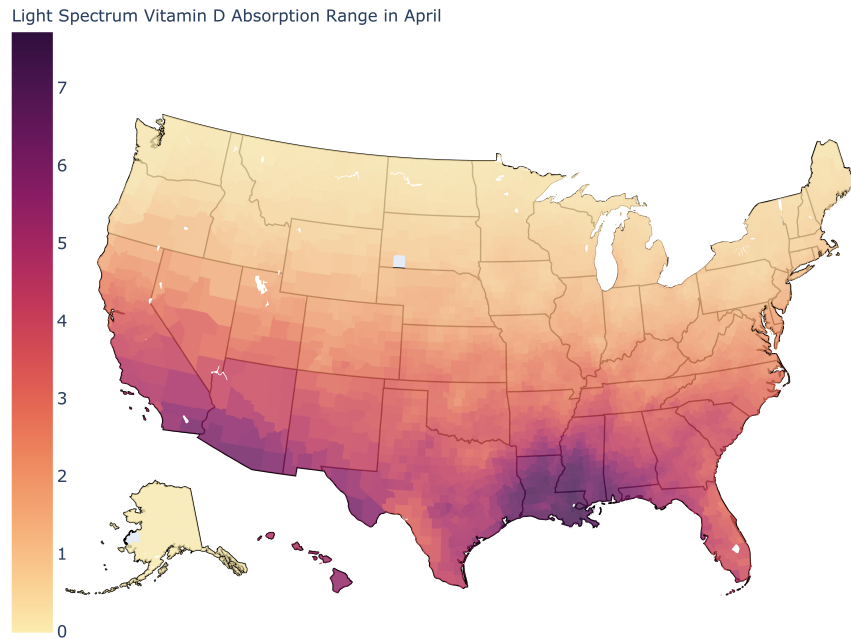
also produces highly accurate iRF models and is important in determining the important environmental features in predicting COVID-19 outcomes.

3.4.3 Vitamin D Absorption is Related to Seasonal Features and the Southern United States

Included in this analysis is an exploration into the all-to-all network that is created using iRF-LOOP on the environmental features. The vitamin D UV absorption range in April was a top important feature for the iRF model predicting the overall CFR, and the monthly vitamin D UV absorption range was a top important feature for the iRF model predicting the monthly CFR. This feature was explored in the iRF-LOOP all-to-all network to determine if it is a proxy for another feature or if it is highly related to other features. Figure 3.5a displays the other features that contain an edge connected to the vitamin D absorption feature of interest. Six of these features are climate related, in particular are seasonal features or are related to solar radiation such as potential evapotranspiration and precipitation. Two of these features are policy features that are associated with Southern states, in particular mail-in voting restrictions that were put in place for the 2020 Presidential Election. Three of these features were climate features, one in which males have commutes over 90 minutes and two features that capture the Cajun/French population in Louisiana. This commute feature captures rural areas common in the Southern United States, while the Cajun/French populations that are found in Louisiana are also hot spots for the vitamin D absorption range in April, shown in Figure 3.5b. Importantly, using iRF alone in this analysis does not reveal whether the vitamin D absorption feature is predicting an increase in CFR or a decrease in CFR for a particular county. Thus, we cannot assume whether this feature and its related features are causing an increase in survivability or a decrease in survivability. It is possible that vitamin D absorption is decreasing the likelihood of death since vitamin D has been shown to be impactful for survival of COVID-19, but it is also possible that this feature is capturing the survivability in counties that have predominantly ignored the precautions of the CDC and other public health authorities [183].



(a) Sub-network pulled from iRF-LOOP all-to-all network that depicts the features related to vitamin D absorption light range in April.



(b) County level map of vitamin D absorption light range in April.

Figure 3.5: This figure focuses on the relationships of vitamin D UV absorption light range in April with other features included in this analysis. The network in (a) shows that many of these features are seasonal climate features and features that are related to the southern region of the U.S. The map in (b) displays the feature of interest, showing that this light range is strongest in the southern states of Alabama, Arizona, California, Louisiana, Mississippi, and Texas. This suggests that this feature that is highly predictive of CFR is highly related to the seasonal climate and surrounding environment of this region.

3.5 Conclusion

In this study we have presented different classes of iRF predictive models that showcase the performance of environmental features in predicting the COVID-19 Case Fatality Ratio for counties in the United States. The models predicting the overall CFR for all of 2020 suggests that COVID-19 related policies and public health measures had a lesser effect than the surrounding climate, demographic, and socioeconomic factors within the individual’s county. This may be inaccurate due to the wide variability of the policy implementations throughout the course of the pandemic. However, the models using windowed monthly measurements of CFR show that climate, particularly climate in combination with policy related features, provide the most accurate predictions. This may be capturing the seasonality of COVID-19 infections and portray the risk of infection in vulnerable populations across the U.S. as individuals move indoors and congregate for the holidays but also the effect of available hospital beds decreasing and as a result quality of care declining.

The algorithm iRF also functions as a feature engineering tool, providing a list of features that are the most important in the creation of the model. In analyzing the top ten important features in the iRF models predicting both overall CFR and the windowed monthly CFR we have found that the top features are a mix of known and unknown features that impact survivability of COVID-19. These known features include demographic features that capture the elderly populations that were devastated in the early stages of the pandemic. Others include policy features that slow the spread of COVID-19, such as mask mandates and bans in nursing homes. Included in these important features were many climate features that change seasonally, such as temperature and solar radiation, indicating once again that seasonality may be a factor in COVID-19 survivability.

We have shown that one important feature that was captured in multiple models was the light spectrum absorption range of vitamin D, which is known to impact COVID-19 outcomes [5]. An in-depth analysis of this feature was provided in this study, using iRF-LOOP to produce an all-to-all predictive network of features and explore the related environmental features. Not only was vitamin D indicative of seasonal features, but was also shown to be related to the Southern region of the United States, in particular Alabama, Arizona,

California, Louisiana, Mississippi, and Texas. In this study we do not focus on the direction of the relationship between these features and an increase in Case Fatality Ratio, thus it is unknown if these highly related features are causing or preventing an increase in deaths.

When health care providers administer life saving care to those who are infected with SARS-CoV-2, it is difficult to correct for the effects of the patients environment. However, this important information can be used to create not only life saving policy decisions but also supply chain and employment decisions within the health care system to anticipate when CFR may peak. Additionally, this approach can be applied to a variety of infectious diseases as well as other diseases and disorders in order to understand the effects of one's environment on their health.

Chapter 4

Applying Divide-Test-Integrate To Gene Expression Data Improves the Performance of iRF-LOOP

4.1 Abstract

Network inference algorithms are often applied to gene expression data to produce gene-to-gene networks. These networks provide a wealth of information to mine for new biological interactions between genes. While many studies use network inference algorithms such as Pearson’s Correlation Coefficient and iterative Random Forest-Leave One Out Prediction (iRF-LOOP) to produce these networks, few have done so by combining batch effect correction and Divide-Test-Integrate (DTI) to capture more accurate and specific associations. In this study, we use batch correction (via ComBat) and DTI with iRF-LOOP to produce Predictive Expression Networks from *Saccharomyces cerevisiae* and *Arabidopsis thaliana* microarray expression data. We show that the networks produced using DTI and iRF-LOOP reduce the bias of housekeeping genes and produce higher quality networks. These high quality networks may be used in other downstream analyses to discover new biological interactions, and this method of applying DTI to iRF-LOOP may be used with a variety of biological datasets.

4.2 Introduction

Gene-to-gene relationships are often characterized as networks, where nodes represent the genes and the edges represent the presence of an interaction between these genes. These networks aid in understanding the biological system as a whole. An example of such a network is the co-expression network, which depicts whether two genes are co-expressed together. Many of these networks are produced using correlation-based network inference tools, such as Pearson’s Correlation Coefficient (PCC). Another example is a Gene Regulatory Network (GRN) which captures transcription factor-to-gene relationships in which one gene is regulating the expression of another. An expansion on the GRN is the Predictive Expression Network (PEN), which aims to capture predictive relationships among all genes and thus beyond just transcription factor-to-gene. Many of these networks are used in downstream processes, such as Random Walk with Restart [166]. Thus, it is desirable for these networks to be of the highest quality possible to ensure accurate results. Many

of these networks must be thresholded to a manageable size before use, thus biologically truthful edges must be ranked highly to be retained after thresholding.

Expression data typically contains a matrix of gene expression information where the columns measure the expression of a specific gene and the rows are samples of either cells or tissues. RNA-seq and microarray expression data are two examples of these data types. Microarray experiments became common in scientific research at the turn of the century, predominantly due to the ability to measure the expression of thousands of genes for one sample [13]. Traditionally, these microarray experiments are concatenated together to form one matrix, and this matrix is used in the network inference algorithms to produce the gene-to-gene network. Unfortunately, this method produces a bias towards housekeeping genes that are consistently expressed throughout a variety of experiments.

In this study, we compare the Divide-Test-Integrate (DTI) [104] method to the traditional concatenation method in combination with batch correction in microarray gene expression data using the network inference tool iterative Random Forest-Leave One Out Prediction (iRF-LOOP) [27]. We present two methods to divide the samples into their individual groups, one supervised method and one unsupervised method based on Markov clustering. We also validate that the DTI method reduces the bias towards the consistently expressed housekeeping genes by intersecting the top ranked genes with highly ranked gene-to-gene edges in the Gene Ontology semantic similarity network.

4.3 Methods

In this analysis, we used microarray gene expression data from hundreds of experiments in *S. cerevisiae* and *A. thaliana* to determine if the DTI method provided an improvement in the creation of gene-to-gene PENs. The network inference tool used in this analysis is the Explainable-Artificial Intelligence (X-AI) tool iRF-LOOP. The resulting PENs from the traditional iRF-LOOP method of concatenating the samples from multiple experiments into one matrix and the method using iRF-LOOP with DTI were compared to determine if the additional use of computational resources provided a benefit in the inference of gene-to-gene networks based on gene expression data.

4.3.1 Network Inference

There are a variety of algorithms that have been used in the past to create gene-to-gene networks from gene expression data. Typically, these datasets require a matrix where rows are samples and columns are genes. The most commonly used network prediction algorithm is PCC, and has been used for the YeastNet Version 3 co-expression network creation [97].

In this analysis, we used iRF-LOOP to produce the gene-to-gene PENs. iRF-LOOP is a Random Forest-based method that selects one gene to be the dependent variable, creates an iRF model with the remaining genes as independent variables, and repeats the process for all genes in the matrix. Each iRF model produces feature importance scores for each gene that is then converted into a network of all independent genes predicting the dependent gene with an associated normalized importance score or edge weight. The important scores for each iRF model are merged to create the final all genes-to-all genes directed PEN.

4.3.2 Divide-Test-Integrate

The standard method of merging the expression data from different experiments into one matrix followed by running the network inference algorithm is heavily biased towards house-keeping genes that are frequently expressed together in different scenarios or experiments. The DTI method has been used by YeastNet since Version 2 [104] to account for these issues, grouping these experiments into like groups and using PCC to create co-expression gene-to-gene networks.

In DTI, instead of combining multiple experiments and samples into a single expression vector for each gene to be evaluated with a network inference algorithm, samples of similar types are separated into groups to be evaluated individually. The network inference algorithm produces an intermediate network for each group, then the intermediate networks are merged together to produce the final network. Separating the samples into like groups encourages the discovery of relationships that would've otherwise been missed. This method is depicted in Figure 4.1.

For each n grouping of samples, the smaller subset of microarray experiments was merged and then the network inference tool was run. The n total gene-to-gene intermediate networks

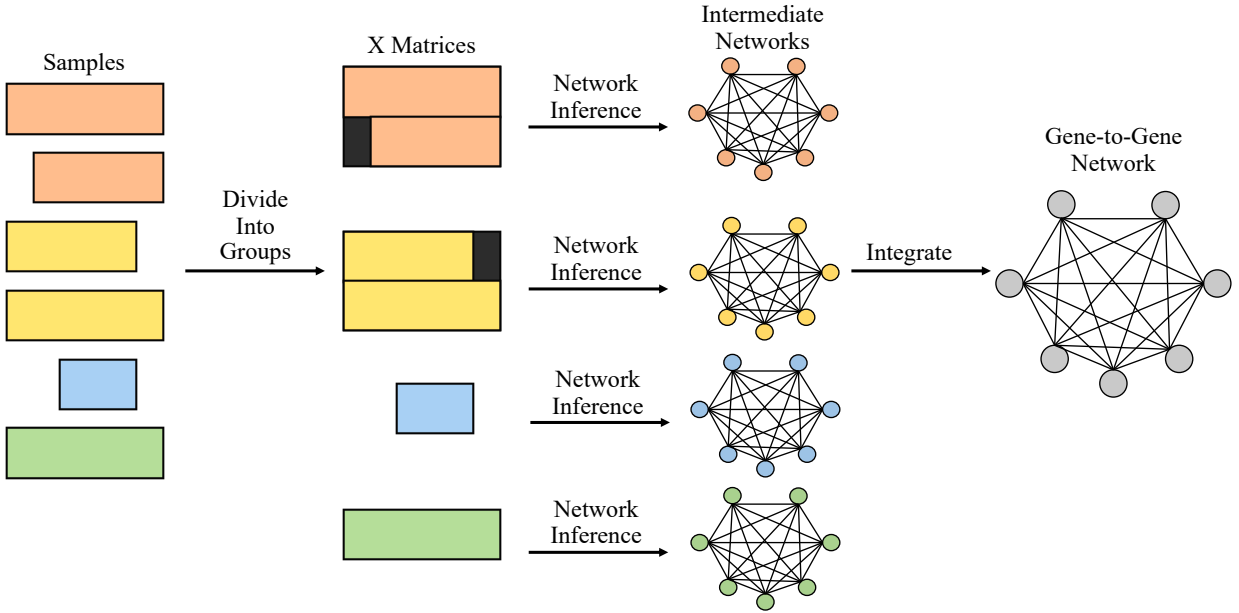


Figure 4.1: This figure depicts the Divide-Test-Integrate (DTI) method, in which microarray samples are divided into like groups, the network inference algorithm is applied to each group to produce intermediate networks, and the intermediate networks are integrated into the final gene-to-gene network. This method reduces the bias of housekeeping genes that is common in the traditional method of concatenating all the samples into one matrix. The intermediate networks may also be tested to determine the quality of the network and determine if they are of a high enough quality to integrate into the final network.

are then combined into one final gene-to-gene network using the weighted sum (WS) of each unique edge, defined as:

$$WS = x_0 + \sum_{i=1}^m \frac{x_i}{D \times i}$$

where $x = \{x_0, x_1, \dots, x_m\}$ is the sorted list of correlations between two genes from highest to lowest, m is the number of gene-to-gene networks that the edge appears in, and $D \geq 1$ is the decay rate.

4.3.3 Dividing the Samples

Here we test two methods of separating the samples in the matrix into groups to use in DTI. One method is a supervised method of dividing samples, by dividing the samples based whether the associated experiments measured similar functions. The second method is an unsupervised method of dividing samples using a clustering algorithm, in this case sample similarity was determined with iRF-LOOP using sample vectors and Markov clustering (MCL) [174] was used to determine which samples are similar enough to categorize into the same group.

For the supervised method of grouping samples we used microarray expression data measured from *S. cerevisiae*. The Saccharomyces Genome Database (SGD) [25] contains over 500 different microarray datasets each from different experiments for *S. cerevisiae*. In total, this analysis used 532 experiments, containing 11,323 samples and 7090 genes. Each of the microarray datasets contain one or more biological tags, for example heat shock treatment or osmotic stress. These 56 biological groups were used as a form of supervised sample groupings for the DTI process. In total, the 56 groups contained between 13 and 2,347 samples, and between 3,567 and 6,880 genes. Details on the number of samples and genes in each group may be found in Supplementary Tables 16 and 17.

For the unsupervised method of grouping samples we used microarray expression data measured from *A. thaliana*. This data set contained 6,057 samples and 21,678 genes from 304 experiments [76]. In order to cluster these samples, first all of the samples were placed into one matrix, where the rows are genes and the columns are samples. iRF-LOOP was then

applied to create a sample-to-sample network, rather than the typical gene-to-gene network. This sample-to-sample network was thresholded to the top 0.1% of edges and clustered using MCL with an inflation value of 1.1 which initially created 110 clusters. However, clusters with less than 100 samples were merged together, which resulted in 13 groups that contained 103 to 1,152 samples per group. Each group contained 21,678 genes. Details on the number of samples and genes in each group may be found in Supplementary Table 18.

To determine whether this unsupervised method of grouping samples provided a benefit over simply keeping samples from the same experiment together, the Bray-Curtis measure of dissimilarity [16] was used to determine how diverse the groups of *A. thaliana* samples are. This is defined between two groups as

$$\text{Bray-Curtis dissimilarity} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

$$C_{ij} = \sum_{k=1}^n x$$

$$x = \begin{cases} a & \text{if the two groupings share samples from the given experiment and } a < b \\ b & \text{if the two groupings share samples from the given experiment and } a > b \\ 0 & \text{if the two groupings do not share samples from the given experiment} \end{cases}$$

$$S_i = \text{number of samples in Group } i$$

$$S_j = \text{number of samples in Group } j$$

Where n is the total number of experiments, a is the number of samples from the given experiment that are in Group i , and b is the number of samples given experiment that are in Group j . The Bray-Curtis dissimilarity measure varies between 0 and 1. A value of 0 means that the two groups all share the same experiments and the samples within experiments were evenly divided between the two groups. A value of 1 means that the two groups do not share any samples that came from the same experiment. The annotation of which sample

belonged to which experiment was provided and used to determine the Bray-Curtis value between pairs of clusters of samples [Figure 4.2].

4.3.4 Data Preprocessing

Batch effects are the differences between batches of samples or across experiments that are not based in biology. These batch effects may affect the results of a data analysis pipeline and need to be corrected before use. In this analysis, ComBat [88] was used as the batch correction tool. Thus, before any matrices of merged microarray experiments were used in any part of this analysis, batch correction using ComBat was applied to account for these batch effects. This includes batch correction prior to the iRF-LOOP sample clustering of *A. thaliana* samples. Other data cleaning steps included converting gene names to one naming convention and removing whole samples or genes that were entirely null values.

4.3.5 Performance Evaluation

The resulting networks were scored to determine which network inference method — PCC, traditional iRF-LOOP, or iRF-LOOP with DTI — performed the best at capturing true biological relationships between two genes. These networks were sorted based on their edge weight from highest to lowest weight and scored against a gold standard network. In the case of the PCC network, the absolute value of the edge weights were used in the sorting of edges. The gold standard networks contain edges that represented experimentally validated relationships between the two genes.

For *S. cerevisiae*, this gold standard network consisted of two edge types. The first edge type was a literature curated transcription factor-to-target gene network from YEASTRACT [162] and contained 307 transcription factors, 6,722 target genes, and 201,971 edges. The second edge type was gene-to-gene metabolic network created from YeastCyc [94] reactions, where two genes are connected based on whether they share a compound within a metabolic pathway, and contained 2,205 edges and 701 genes. In total, the gold standard network to score the resulting *S. cerevisiae* gene-to-gene networks contained 204,166 edges and 6,744

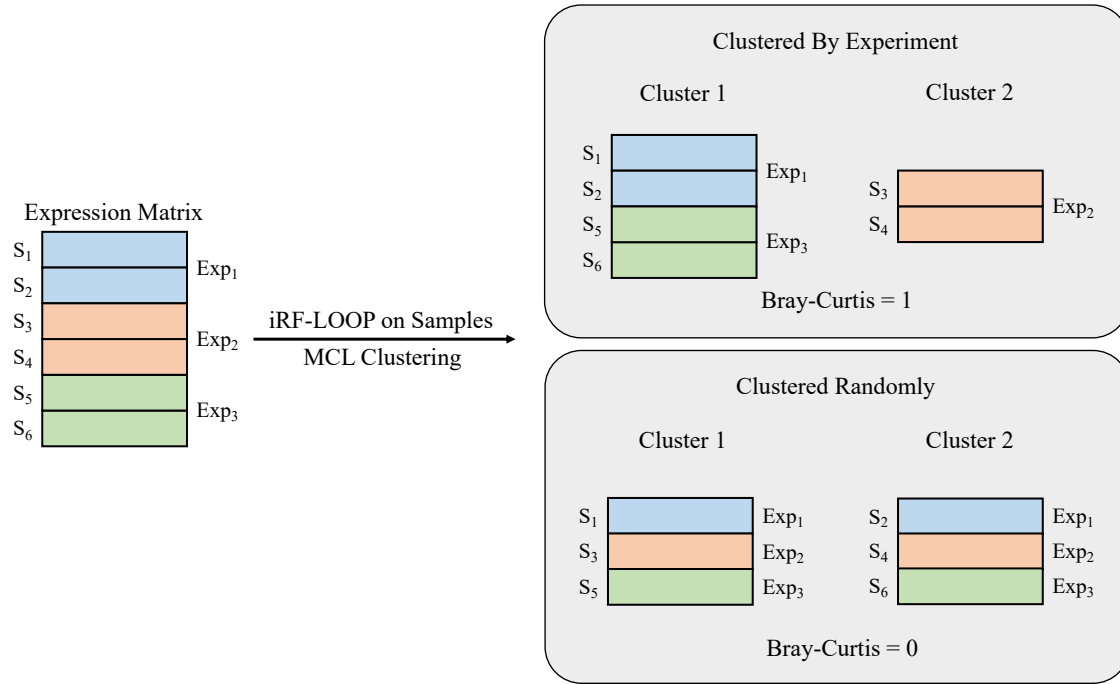


Figure 4.2: This figure depicts the method of using the Bray-Curtis dissimilarity measure to evaluate whether samples for the same experiment clustered together after applying iRF-LOOP on the samples and MCL clustering the resulting network. If two clusters did not share samples from the same experiment, the Bray-Curtis value between the two clusters would be zero. If the samples from the same experiment were evenly divided among the two clusters, the Bray-Curtis value would be zero.

unique genes. Duplicate edges and edges that contained genes not originally present in the microarray expression datasets were removed.

The *A. thaliana* gold standard network was constructed similarly. The literature curated transcription factor-to-target gene network contained 1,236 edges with 288 transcription factors and 552 target genes [87]. The metabolic network from AraCyc [128] contained 19,057 edges and 2,471 genes. Again, duplicate edges and edges with genes not included in the original microarray expression dataset were removed resulting in a 20,293 edges and 3,040 unique genes.

While the gold standard networks used in this analysis are high quality, they are predominantly incomplete. Not every true biological relationship is captured in these networks. Additionally, there are no edges in the gold standard networks that capture whether two genes do not have a relationship. Thus, traditional binary scoring metrics that use true positives, false positives, true negatives, and false negatives in combination cannot be used in this case, such as Precision-Recall curves. Thus, the scoring metric used to determine the performance of the network was the Area Under the normalized Discounted Cumulative Gain curve (AUnDCG). Normalized Discounted Cumulative Gain (nDCG) [86] was calculated at every edge from one to k , where k is the number of true positive values in the gold standard network. This is defined as follows:

$$nDCG_{network,k} = \frac{DCG_{network,k}}{IDCG_{gold\ standard,k}}$$

$$DCG_{network,k} = \sum_{i=1}^k \frac{x}{\log_2(i+1)}$$

$$x = \begin{cases} 1 & \text{if edge is a true positive} \\ 0 & \text{if edge is a false positive} \end{cases}$$

$$IDCG_{gold\ standard,k} = \sum_{i=1}^k \frac{1}{\log_2(i+1)}$$

The X axis of the nDCG curve is the value of k , while the Y axis is the nDCG at k . The area under this curve is used for the AUnDCG. The nDCG curve provides two major insights into the performance of these network inference tools. First, nDCG accounts for the rankings of the truth edges within the sorted network. Typically, these large networks are thresholded to the top n edges or the top $n\%$ edges to be used in downstream analyses. Ideally, the best network inference method would rank the truth edges higher in the sorted edge list to be captured in this thresholding step. Second, the nDCG scoring method does not negatively impact the overall score of the model if a highly ranked edge that may be true and is detected by the model is missing from the gold standard network. This is crucial in cases such as in *A. thaliana* where the gold standard network is more sparse and considerably smaller than the predicted gene-to-gene network.

The AUnDCG scoring was also applied to intermediate PENs resulting from the division of samples into groups within the DTI method. Only the PENs of high quality were included in the WS formula. The intermediate PENs were ranked based on their AUnDCG score, and the bottom 10% of groups were dropped.

4.3.6 Gene Ontology Semantic Similarity Network

The DTI method provides a potential improvement over the traditional method of concatenating the samples by reducing the bias of gene-to-gene interactions of housekeeping genes. To determine the effectiveness of this bias reduction, the resulting *A. thaliana* PENs are intersected with a Gene Ontology (GO) semantic similarity network. This GO semantic similarity network connects two genes based on whether they share a Biological Process term. These terms are hierarchical, some terms are broad and encompass many genes, others are specific and encompass few genes. The associated edge weight between two genes in the GO semantic similar network depend on this hierarchy. The higher the edge weight, the higher the likelihood of these two genes having a specific and rare interaction.

The GO semantic similarity network was created using the R package `GOSemSim` [193]. Only GO annotations with solid evidence codes were used. The codes IEA, ISS, NAS, ND, RCA were excluded. The edge weight was calculated by averaging the GO semantic similarity scores achieved using both the Wang [179] and the Resnik [142] calculations. While

the Resnik similarity is determined by the distance to the common ancestor term, the Wang similarity is a graph-based measure that considers the graph topology in the final score. These edges were filtered to include edges of weight 0.5 or more, to include only edges that likely had a specific association. This GO semantic similarity network contained 696,524 edges and 7,623 unique genes.

GO semantic similarity network created by David Kainer.

4.3.7 Computational Resources

The iRF models used in this analysis was run on the computational resources located at Oak Ridge National Laboratory in the Oak Ridge Leadership Computing Facility (OLCF), primarily the Summit supercomputer. Summit is an IBM system that contains 4,600 nodes. Each node contains two POWER9 processors and six NVIDIA Tesla V100 GPUs. Each POWER94 processor contains 22 cores with 4 threads each for a total of 176 threads per node and 512GB of DDR4 memory. Post-processing of many of these networks was performed on the Andes supercomputer at OLCF. Andes contains two partitions, one with 9 GPU nodes and one with 704 CPU compute nodes. Each CPU node contains two AMD EPYC 7302 CPUs per node with 16 cores per CPU, for 32 cores per node and 256 GB of memory per node. Each GPU node contains two Intel Xeon E5-2695 GPUs with 14 cores per GPU for a total of 28 cores per node and 1 TB of memory per node.

4.4 Results and Discussion

4.4.1 Supervised Division of Samples in DTI with iRF-LOOP Provides an Improvement Over Traditional iRF-LOOP

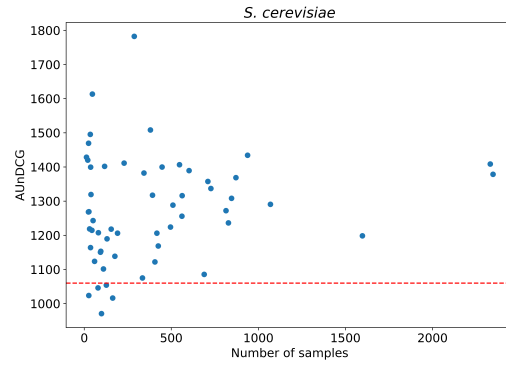
The DTI method applied to the *S. cerevisiae* microarray expression data produced 56 intermediate PENs that were evaluated using AUnDCG to only include the groups of the highest quality in the final PEN. Five intermediate PENs were not included due to their low quality. These experiments included the biological tags cell morphogenesis with an AUnDCG of 1023.735, cellular ion homeostasis with an AUnDCG of 970.620, filamentous

growth with an AUnDCG of 1016.424, metal or metalloid ion stress with an AUnDCG of 1054.243, and protein dephosphorylation with an AUnDCG of 1046.276. Additionally, the number of samples in these groups were small, between 26 and 163 samples. Typically, iRF-LOOP performs poorly when the X matrices have a small number of samples. The relationship between the number of samples in the group and the AUnDCG score may be found in Figure 4.3a. Additionally, the optimal value for D in the WS formula was found to be $D = 1$ [Supplementary Figure 4a]. Thus, the final *S. cerevisiae* PEN using iRF-LOOP with DTI used 51 different groups of samples with $D = 1$.

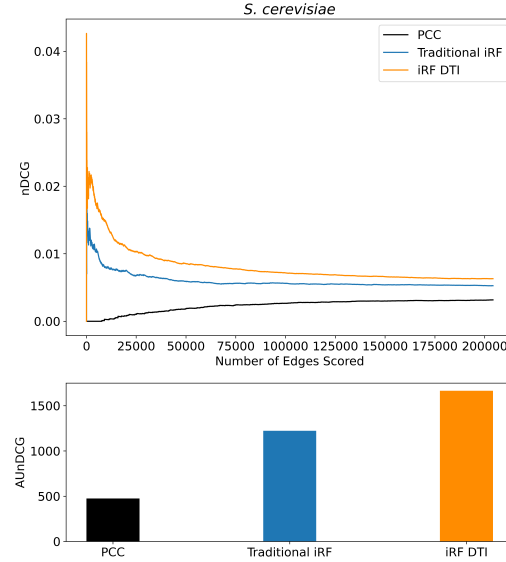
The overall performance of final *S. cerevisiae* PENs from the traditional concatenation method of iRF-LOOP and from iRF-LOOP with DTI as well as the final *S. cerevisiae* co-expression network using PCC were evaluated and the results are shown in Figure 4.3b. The PCC co-expression network had an AUnDCG score of 474.956, while the traditional iRF-LOOP PEN had an AUnDCG score of 1222.944, and the iRF-LOOP with DTI PEN had an AUnDCG score of 1665.545. Across the top 204,166 edges that were scored, the iRF-LOOP with DTI PEN consistently had the best nDCG score, which resulted in the highest AUnDCG score. This suggests that iRF-LOOP with DTI provides a considerable improvement over the traditional concatenation method of iRF-LOOP, PCC and other single matrix network inference methods.

4.4.2 Unsupervised Division of Samples in DTI with iRF-LOOP Provides an Improvement Over Traditional iRF-LOOP

The DTI method was also applied to *A. thaliana* microarray expression data. After iRF-LOOP and MCL clustering, 13 clusters of samples resulted in 13 intermediate PENs that were evaluated using AUnDCG. The intermediate PEN from Group 12 was not included in the final PEN, with an AUnDCG score of 39.568. Like *S. cerevisiae*, this group contained the least samples with only 103 rows in the X matrix. The relationship between the number of samples and the AUnDCG score is shown in Figure 4.4a. The optimal value for D in the WS formula was found to be $D = 1$ [Supplementary Figure 4b]. Thus, the final *A. thaliana* PEN using iRF-LOOP with DTI used 12 different groups of samples with $D = 1$.

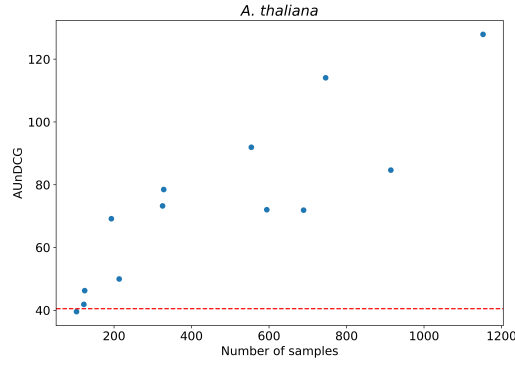


(a) Relationship between AUnDCG scores for the intermediate PENs in *S. cerevisiae* and number of samples in the X matrix.

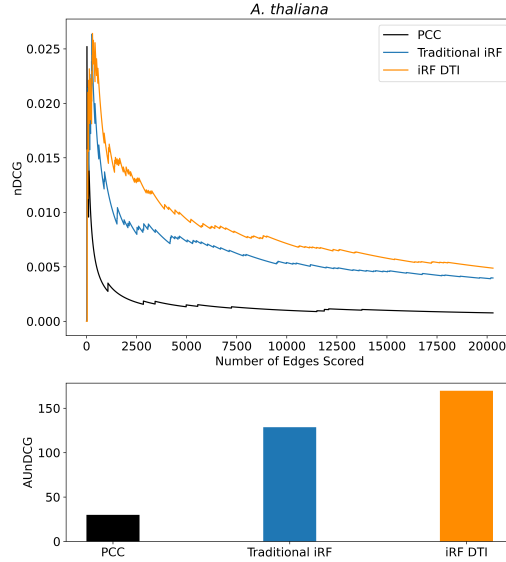


(b) Overall score of the gene-to-gene networks produced by PCC, Traditional iRF-LOOP, and iRF-LOOP with DTI in *S. cerevisiae*.

Figure 4.3: Figure 4.3a depicts the relationship between the quality of the divided group and the number of samples included in the X matrix. The bottom five groups were not included in the integrated final PEN, this cutoff is depicted by the dotted red line. These groups also contained less than 200 samples. Figure 4.3b depicts the nDCG curves for the three gene-to-gene networks. The PCC co-expression network had an AUnDCG score of 474.956, the traditional iRF-LOOP PEN had an AUnDCG score of 1222.944, and the iRF-LOOP with DTI PEN had an AUnDCG score of 1665.545. iRF-LOOP with DTI had the highest AUnDCG score and is the best performing network.



(a) Relationship between AUnDCG scores for the intermediate PENs in *A. thaliana* and number of samples in the X matrix.



(b) Overall score of the gene-to-gene networks produced by PCC, Traditional iRF-LOOP, and iRF-LOOP with DTI in *A. thaliana*.

Figure 4.4: Figure 4.4a depicts the relationship between the quality of the divided group and the number of samples included in the X matrix. The bottom scoring group was not included in the final integrated PEN, this cutoff is depicted by the dotted red line. Like *S. cerevisiae*, this group also contained less than 200 samples. Figure 4.4b depicts the nDCG curves for the three gene-to-gene networks. The PCC co-expression network had an AUnDCG score of 30.021, while the traditional iRF-LOOP PEN had an AUnDCG score of 128.722, and the iRF-LOOP with DTI PEN had an AUnDCG score of 169.776. iRF-LOOP with DTI had the highest AUnDCG score and is the best performing network.

The overall performance of the final *A. thaliana* PENs from the traditional concatenation method of iRF-LOOP and from iRF-LOOP with DTI as well as the final *A. thaliana* co-expression network using PCC were evaluated and the results are shown in Figure 4.4b. The PCC co-expression network had an AUnDCG score of 30.021, while the traditional iRF-LOOP PEN had an AUnDCG score of 128.722, and the iRF-LOOP with DTI PEN had an AUnDCG score of 169.776. These scores are a whole order of magnitude lower than the AUnDCG scores for *S. cerevisiae*. This is likely due to *S. cerevisiae* being a long-studied model organism with a lower number of genes and thus is likely to have with proportionately more truth edges than *A. thaliana*, thus there is simply more truth edges to capture in the top edges. Additionally, the iRF-LOOP with DTI nDCG curve is not consistently above the traditional iRF-LOOP nDCG curve and the PCC nDCG curve [Figure 4.4b]. However, the overall AUnDCG for *A. thaliana* is highest for iRF-LOOP with DTI. Thus, it may be concluded that iRF-LOOP with DTI provides an improvement over traditional iRF-LOOP when the samples are divided in an unsupervised manner overall when thresholded to the top 20,294 edges, however the same conclusion can not be applied to the early nDCG in the first 1,000 edges.

4.4.3 Unsupervised Division of Samples Kept Samples from Same Experiment within Same Cluster

While both the unsupervised and supervised division of samples within DTI provide improvements over the traditional method of concatenating samples into one matrix, the unsupervised division method requires considerably more computational power than the supervised division method. The unsupervised division method requires an iRF-LOOP model where the samples are the features and the genes are the rows. For *A. thaliana*, this iRF-LOOP model contained 6,057 separate iRF models where each model contained 21,678 genes. The Bray-Curtis measure of dissimilarity measures how diverse the clustered samples were, and whether samples from the same experiment tended to cluster together. A value of Bray-Curtis closer to 1 signifies that there is little overlap in experiments between the two clusters, and samples tended to cluster together based on experiment. A value of

Bray-Curtis closer to 0 signifies that there is an overlap in experiments between the two clusters and the samples did not cluster together based on experiment.

The Bray-Curtis dissimilarity measures between all 13 clusters of samples is shown in Figure 4.5. The dissimilarity measure between Group 2 and Group 4 was 0.65, but the remaining dissimilarity measures between groups were greater than 0.8. For many of the pairings, the dissimilarity measure was 1. This indicates that the samples frequently clustered together based on experiment. Thus, while this computationally expensive unsupervised grouping provides a potential division of samples, it is likely just as effective to group based on experiment. While the *A. thaliana* experiments did not have a biological tag that could have been used to group the samples in the supervised manner, a less computationally expensive task could have been to text mine the biological tags from the experiment abstract and the groupings would have been similar to the unsupervised groupings.

4.4.4 iRF-LOOP with DTI Captures More Specific Gene-to-Gene Associations

The cluster of genes used in this analysis come from the GO semantic similarity network, shown in Figure 4.6. These six genes are AT3G09440 (HSP70), AT5G56030 (HSP81-2), AT5G02500 (HSC70-1), AT1G62740 (HOP2), AT3G25230 (ROF1), and AT5G09590 (MTHSC70-2). HSP70, HSP81-2, MTHSC70-2 are heat shock proteins, HSC70-1 encodes a heat shock protein, and HOP2 and ROF1 are potential co-chaperones of HSP90/HSP70 [11]. All six of these genes contain the GO term GO:0050896 response to stimulus, of which *A. thaliana* contains 5,506 genes with this term [1]. Some genes contain more specific associations, such as the two genes HSP70 and MTHSC70-2 with an edge weight of 0.6615. These two genes share the GO term GO:0051085 chaperone cofactor-dependent protein refolding with only 45 other genes, one of which being HSC70-1 [1]. Ideally, for DTI to reduce the bias of housekeeping genes, the iRF-LOOP with DTI PEN would capture more of these specific GO associations. Figure 4.6 contains the edges between these six genes in the top 10,000 edges of the traditional iRF-LOOP PEN and the iRF-LOOP with DTI PEN.

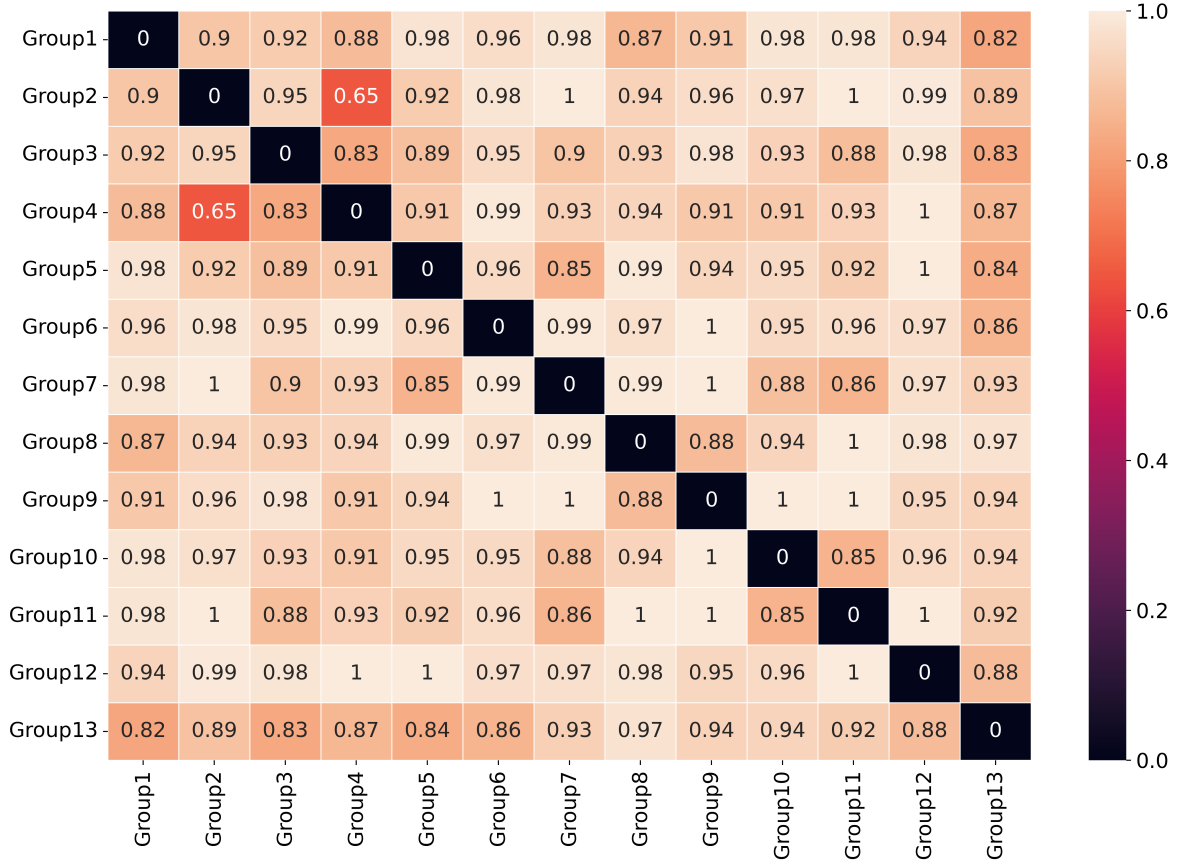


Figure 4.5: This figure depicts the Bray-Curtis dissimilarity measures between the 13 clusters of *A. thaliana* samples. All group pairs but one contain a dissimilarity measure greater than 0.8. This shows that the samples from the same experiment tend to cluster together, suggesting that the computationally intensive unsupervised division of samples may not be as efficient as a less computational intensive supervised method, such as text mining the abstracts for biological tags.

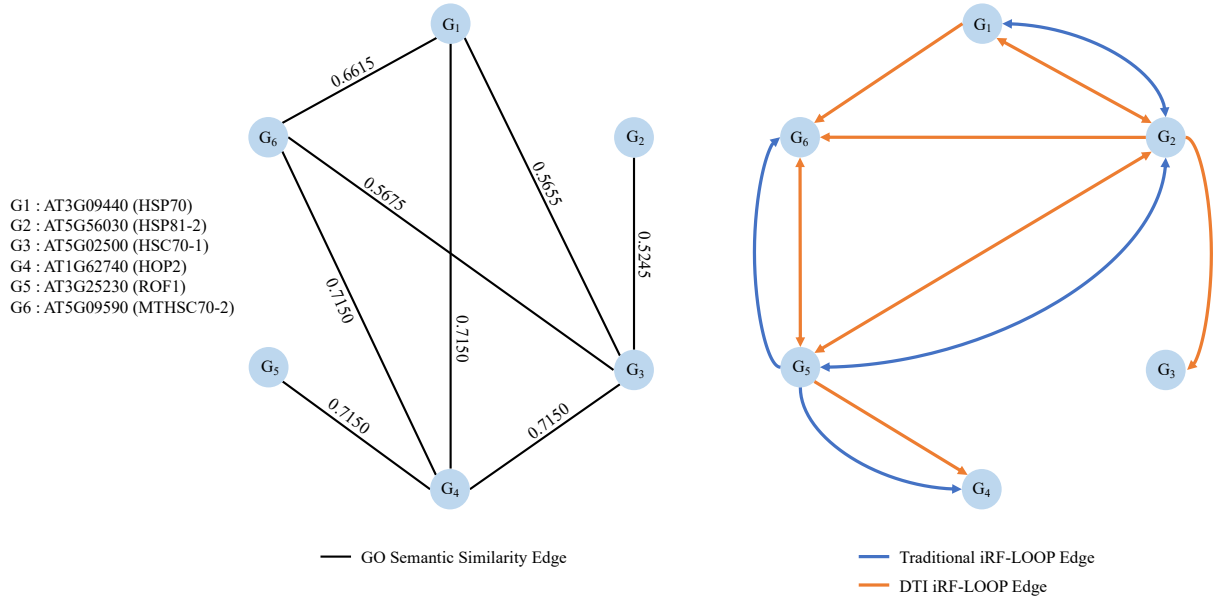


Figure 4.6: This figure depicts the GO semantic similarity edges, the traditional iRF-LOOP edges, and the iRF-LOOP with DTI edges between six *A. thaliana* genes. The GO semantic similarity edges are those that contain a weight above 0.5, which are associations that are more specific with the biological system. For DTI to reduce the bias of housekeeping genes, the iRF-LOOP with DTI PEN would capture more of these specific GO associations, which is evident in the network on the right. Of the top 10,000 edges from each PEN, traditional iRF-LOOP captures only 269 of these specific GO semantic similarity associations, while iRF-LOOP with DTI captures 396 of these edges.

While the traditional iRF-LOOP PEN captures edges between five of the six genes, the iRF-LOOP with DTI PEN captures edges between all six genes and captures more edges between these genes. Of the top 10,000 edges of the iRF-LOOP with DTI PEN, 396 edges contain a GO semantic similarity score of 0.5 or greater. Of the top 10,000 edges of the traditional iRF-LOOP PEN, 269 edges contain a GO semantic similarity score of 0.5 or greater. Thus, the iRF-LOOP with DTI PEN reduces the bias of housekeeping genes that is found in the traditional concatenation iRF-LOOP method.

4.5 Conclusion

In this study we have presented a supervised and unsupervised method of dividing samples in *S. cerevisiae* and *A. thaliana* microarray gene expression data to cluster into groups for DTI. In both organisms, the PEN produced with DTI and iRF-LOOP outperforms the PEN produced using the traditional method of sample concatenation and iRF-LOOP as well as the co-expression network produced using PCC. This suggests that DTI and iRF-LOOP is effective in producing gene-to-gene networks from gene expression data, and may be effective if applied to a variety of biological data.

While the unsupervised division of samples used for the *A. thaliana* expression data produced high quality networks, an in depth analysis shows that the samples tended to cluster together based on experiment. The unsupervised division of samples using iRF-LOOP was computationally expensive. We suggest that if computational resources are limited, it may be equally effective to text mine the given abstracts for the experiments and group the experiments based on biological terms or tags.

We have shown that the PEN produced using DTI and iRF-LOOP captured more specific associations than the traditional iRF-LOOP method by intersecting top ranked PEN edges from both methods with top ranked GO semantic similarity edges. DTI and iRF-LOOP successfully reduced the bias of consistently expressed genes and produced a more unique network of specific associations beyond simply commonly co-expressed genes. The PENs produced using DTI can be used alone or in combination in a variety of downstream analyses and can add depth to a multi-layer system such as a multiplex network [89].

Chapter 5

Exploring *Septoria musiva* Gene Relationships Through a Multiplex Network of the Model Organism *Saccharomyces cerevisiae*

5.1 Abstract

Septoria musiva is a fungal pathogen that causes leaf spot and canker disease in various poplar trees. While biological data for *S. musiva* are sparse, there exists a plethora of resources for the study of *Saccharomyces cerevisiae* since it is one of the most studied eukaryotic model organisms. In this study, we map *S. musiva* genes believed to be involved in fungal infection of *P. trichocarpa* to their orthologs in *S. cerevisiae*. We use the various available *S. cerevisiae* gene-to-gene networks in combination with Random Walk with Restart (RWR) and Gene Refinement using Interacting Networks (GRIN) to produce a multi-layer analyses and determine which genes are most influential in infecting *P. trichocarpa*.

5.2 Introduction

Systems biology does not operate under the assumption that all interactions occur in isolation. Some biological interactions may be represented as a graphical network, in which the nodes are genes, the edges represent the interaction between two genes, and their weights are the strength of that interaction. Compiling multiple biological network types, such as predictive expression networks and protein-protein interactions can yield results that would otherwise not be possible if looking at a single network. The need to traverse across network types is crucial to understanding new biological information. A group of networks with shared nodes that are merged together is called a multi-layer or multiplex network [89]. There exist many methods to explore these multiplex networks, such as Lines of Evidence (LOE) [182] or Random Walk with Restart (RWR) [172].

Sphaerulina musiva, also known as *Septoria musiva*, is a fungal pathogen that causes canker disease in many popular trees. Native to *Populus deltoides*, *S. musiva* infection has also been found in *Populus trichocarpa*, *Populus balsamifera* and *Populus nigra* [44, 107]. The disease presents itself as cankers on the stems and leaves and spots on the leaves that are reddish brown in color. The cankers tend to occur on the green young branches, and can discolor the tissue and even cause death of the tree [44].

The study of the model organism *Saccharomyces cerevisiae*, also known as baker’s yeast, is crucial in understanding many different biological functions and relationships [132]. *S. cerevisiae* was the first eukaryotic genome to be sequenced in 1996 [63], and multiple databases exist to supply the most accurate and in depth datasets to be used in biological research. Unfortunately there are not as many studies performed on *S. musiva* since it is not a model organism, hence many interactions remain undiscovered. It is possible to use *S. musiva* genes mapped to their ortholog *S. cerevisiae* genes, and exploration in these fuller *S. cerevisiae* networks to recover these missed interactions.

This study utilizes the information gained from the plant-host interaction between *P. trichocarpa* and *S. musiva* in combination with the wealth of knowledge provided by the *S. cerevisiae* networks to determine which *S. musiva* genes are the most important in fungal infection. A selection of *S. musiva* genes related to traits such as disease severity and number of cankers were mapped to their orthologs in *S. cerevisiae*, and used as seed genes within an RWR multiplex network of *S. cerevisiae* genes. The RWR-based tools filter the initial seed genes and produce a more refined set of *S. cerevisiae* genes that may be related to *S. musiva* infection in *P. trichocarpa*.

5.3 Methods

5.3.1 Random Walk with Restart

Random Walk with Restart (RWR) is an algorithm that determines the relationship between two nodes within a network. For biological applications multiple layers of networks, such as metabolic networks and co-citation networks, can be combined into one multiplex network, or a merged network of multiple unconnected graphs, to be used in RWR. A random walker, or particle, is placed on the seed node and walks throughout the weighted edge multiplex network. At random, the walker will be removed from its path and placed back on the seed node, restarting its path. The relationships between the seed node and all other nodes discovered are then ranked, giving a ranked list of node to node interactions. The R package RWR-MH or Random Walk with Restart-Multiplex Heterogeneous [172] is the basis of the

analysis used in this study. This study focuses primarily on homogeneous networks, or networks that only contain gene-to-gene edges.

5.3.2 RWRtools

The `RWRtools` R package contains a variety of functions that were used throughout this analysis. In particular, `RWR Make Multiplex` created the `Rdata` object used in other functions, `RWR Cross Validation` was used to validate that the multiplex is of high quality, and `RWR Network Score` determined the weights of the individual layers to be used in downstream analyses.

The `RWRtools` R package is pending publication by David Kainer.

RWR Cross Validation

RWR Cross Validation employs K -fold cross validation on a given set of genes, and ranks each gene in an input set when the selected gene is left out. This method produced metrics such as Precision-Recall and Receiver Operating Characteristic curves for the ranked gene list. Ideally, a high quality multiplex captures the relationship between genes that are known to have a biological interaction. In this case, Gene Ontology (GO) terms were used to represent a subset of genes that are likely to have a biological relationship. Five GO terms were selected with a moderate number of genes to test whether the multiplex successfully captures these specific gene-to-gene relationships. These terms were GO:0006887 (exocytosis), GO:0008213 (protein alkylation), GO:0016050 (vesicle organization), GO:0043543 (protein acylation), GO:0048856 (anatomical structure development), with 33, 43, 70, 61, 57 genes respectively. In this analysis, $K = 5$.

RWR Network Score

RWR Network Score produced a series of metrics given a single network or layer and a gold standard network, or a network of known truth edges. One metric is the score per network edge, or the score of the individual edges within the layer. This metric may be compared across layers of varying sizes since it is normalized by the number of edges in the given layer.

RWR Network Score was applied to every layer within the multiplex. The score per network edge was used as the τ parameter in other functions, or the probability that the particle begins its random walk in that layer.

The gold standard network used in this case is a GO semantic similarity network. This network connected two genes based on whether they share a Biological Process term, and the edge weight measured how specific and rare the interaction between these two genes are. The GO semantic similarity network was created using the R package `GOSemSim` [193]. Only GO annotations with solid evidence codes were used. The codes IEA, ISS, NAS, ND, RCA were excluded. The edge weight was calculated by averaging the GO semantic similarity scores achieved using both the Wang [179] and the Resnik [142] calculations. While the Resnik similarity is determined by the distance to the common ancestor term, the Wang similarity is a graph-based measure that considers the graph topology in the final score. These edges were filtered to include edges of weight 0.5 or more, to include only edges that likely had a specific association. This GO semantic similarity network contained 696,524 edges and 7,623 unique genes.

GO semantic similarity network created by David Kainer.

5.3.3 Geneset Refinement using Interacting Networks (GRIN)

Geneset Refinement using Interacting Networks (GRIN) builds on the `RWRtools` R package, using the multiplex network and a set of genes of interest. GRIN separates the set of genes into retained and removed genes. GRIN operates on the assumption that the input gene set contains true and false positives, and uses the Mann-Whitney U test in combination with the connectivity of the multiplex to retain the true positives and remove the false positives. The input gene set in this analysis were genes that were expected to influence *S. musiva* infection in *P. trichocarpa* based on SNPs obtained from the GWAS performed on *S. musiva*. The retained genes from GRIN were used in downstream analyses and were assumed to be the final set of *S. cerevisiae* genes whose orthologs in *S. musiva* impact colonization.

The GRIN R package is pending publication by Kyle A. Sullivan.

5.3.4 Determination of Seed Genes

Five *P. trichocarpa* genotypes were selected (GW-11026, GW-9823, BESC-400, BESC-234, CMBF-28-1) for inoculation of *S. musiva*. 112 isolates of *S. musiva* were used, for a total of 560 isolate-genotype combinations. Of each combination, four replicates were included, except for isolates MN-1, MN-14, and MN-6 in which eight replicates were included for a total of 2,300 samples. For each sample, four traits were measured, the height of the plant, the number of cankers on the plant, the number of cankers per centimeter of height, and disease severity measured on a scale from one to five.

Inoculation of S. musiva in P. trichocarpa and data collection performed by Jared LeBoldus.

Five Genome-Wide Association Study (GWAS) methods were used in this analysis. This includes GCTA [191] and the four models within GAPIT [178] including FarmCPU, MLMM, MLM, and Blink. 20 analyses were run for each GWAS method, one for each combination of measured trait and genotype. The *S. musiva* SNPs with an FDR corrected P-value of 0.4 or better were retained, resulting in 204, 83, 60, and 17 unique SNPs for the traits number of cankers, cankers per centimeter, disease severity, and height respectively, for a total of 340 unique SNPs. The number of unique SNPs for each GWAS method and trait combination is shown in Table 5.1.

The GWAS was performed by Hari Chhetri.

The *S. musiva* SNPs obtained from the GWAS were then mapped to *S. musiva* genes. The coding DNA sequences and their positions were obtained from Mycocosm by the Joint Genome Institute [36, 134]. The SNP was mapped to either the gene in which it was positioned within, or the nearest genes to the left and to the right of the SNP. In total, the 340 unique SNPs were mapped to 393 *S. musiva* genes. The number of unique *S. musiva* genes for each trait is shown in Table 5.2

The 393 unique *S. musiva* genes must be mapped to *S. cerevisiae* genes for use in the multiplex. Fungi Ensembl [80] produces both high and low confidence orthologs for mapping between *S. musiva* and *S. cerevisiae*. Of the 393 unique *S. musiva* genes, 141 orthologs were found in the Fungi Ensembl mapping. Additionally, InParanoid [156] was used to

Table 5.1: This table contains the number of unique SNPs that were retained in each GWAS method and *S. musiva* trait. The FDR P-value threshold for the SNPs was < 0.4

Trait	GCTA	FarmCPU	MLMM	MLM	Blink	Total Unique SNPs
Number of Cankers	70	135	54	1	28	204
Cankers per cm	16	38	34	1	23	83
Disease Severity	1	51	15	0	59	60
Height	0	0	0	0	17	17
Total Unique SNPs	79	210	100	1	120	340

Table 5.2: This table depicts the number of unique SNPs retained after combining and thresholding the GWAS results for each trait, as well as the number of unique *S. musiva* genes and unique *S. cerevisiae* genes after ortholog mapping.

Trait	SNPs	<i>S. musiva</i> Genes	<i>S. cerevisiae</i> Genes
Number of Cankers	204	248	118
Cankers per cm	83	118	53
Disease Severity	60	77	30
Height	17	25	26
Total Unique	340	393	196

produce orthologs between the two organisms. InParanoid is a BLAST-based method of assigning orthologs. The protein FASTA files for the two organisms were provided by NCBI, assembly accessions GCF_000146045.2 for *S. cerevisiae* and GCF_000320565.1 for *S. musiva*. InParanoid identified 144 unique orthologs. Between these two methods, 196 unique *S. cerevisiae* genes were mapped from *S. musiva*. The division of unique SNPs, *S. musiva* genes, and *S. cerevisiae* genes for each trait is available in Table 5.2. Since height can be related to fungal infection but not as significantly as the other three traits, the traits were kept separate for each run using GRIN as well as combined for the fifth run. Additionally, the cankers per centimeter trait is highly correlated with height and number of cankers, since the value of cankers per centimeter was calculated for each plant using its height and the number of cankers.

5.3.5 Biological Layers within the Multiplex

The layers included in the multiplex are all *S. cerevisiae* gene-to-gene networks, creating a homogeneous multiplex. The number of edges and unique genes within each layer is depicted in Table 5.3.

Co-Citation Network

The co-citation network contained gene-to-gene edges that represent whether it is inferred that the two genes have a connection based on an analysis of tens of thousands of abstracts available on PubMed. Genes that appear in the same publication tend to be co-functional, and may capture new information that is missed in the other networks. This network contains 82,319 edges for 4,345 genes. This network is provided by YeastNet v3 [97]. Every edge in this network was normalized by the highest weighted edge.

Gene Interaction Network

The gene interaction network contained 149,498 edges for 4,365 genes. The gene interaction network captures whether genes are co-functional, and was calculated using Pearson's Correlation Coefficient (PCC) on three synthetic genetic array (SGA) datasets and five

Table 5.3: This table depicts the total number of unique edges and unique genes (or nodes) for each layer included within the multiplex.

Layer	Number of Edges	Number of Genes
Co-citation	82,319	4,346
Gene Interaction	149,498	4,365
Metabolic	2,205	701
Phenotypic	20,096	1,011
Phylogenetic	54,496	2,463
Predictive Expression	40,406	6,825
Protein-Protein Interaction	215,261	5,669
Transcription Factor-to-Gene	201,965	6,731
Total Unique	766,246	7,128

epistatic miniarray (E-MAP) datasets. The eight networks were merged into a single network for the final gene interaction network. This network is provided by YeastNet v3 [97]. Every edge in this network was normalized by the highest weighted edge.

Metabolic Network

The metabolic network contained edges that represented genes that were connected based on whether they share a compound within a metabolic pathway. This was produced using YeastCyc [94] reactions and contained 2,205 edges for 701 genes. Every edge in this network had an edge weight of one.

Phenotypic Network

The phenotypic network contained edges that represented how similar two genes were based on their associated phenotype. This network was created using the phenotype annotations and hierarchy provided by the Saccharomyces Genome Database (SGD) [25]. Each edge value was calculated using the Resnik Information Content Similarity between the two genes [127]. While this measure is commonly used for Gene Ontology (GO) term interactions, this formula was applied to phenotype term interactions in this case. The phenotypic network contained 20,095,664 edges for 6,550 genes. This network was sorted and thresholded to the top 0.1% of edges to include 20,096 edges for 1,011 unique genes and was normalized by the highest weight edge.

Phylogenetic Network

The phylogenetic network contained 54,496 edges and 2,463 genes, which linked two genes if they contained similar phylogenetic profiles. For each domain of life, Archaea, Bacteria, and Eukaryota, PBLAST was first run on all *S. cerevisiae* protein sequences against all of the genomes in the domain. The phylogenetic profile matrices of the blast hit scores was then used to calculate the similarity between profiles, using mutual information scores. The final phylogenetic network was created by combining the three separate networks. This network is provided by YeastNet v3 [97]. Every edge in this network was normalized by the highest weighted edge.

Predictive Expression Network (PEN)

Predictive Expression Networks (PENs) are gene to gene networks that define relationships beyond regulatory interactions. The PEN used in this analysis was created using iterative Random Forest-Leave One Out Prediction (iRF-LOOP) in combination with the Divide-Test-Integrate data pre-processing method and ComBat experimental batch correction on microarray expression data provided by the SGD [25, 27, 88, 104]. This PEN contained 40,405,658 directed edges for 7,090 unique genes, and was sorted and thresholded to the top 0.1% of edges. The final PEN contained 40,406 edges for 6,825 genes and was normalized by the highest weighted edge.

Protein-Protein Interaction (PPI) Network

There exists two PPI networks provided by YeastNet v3, one that is literature curated and one that is curated using high throughput experiments. The literature curated PPI network contained 54,421 edges and 5,293 genes, while the high throughput PPI network contained 141,347 edges and 5,487 genes. The literature curated PPI network combined information from multiple PPI databases, including BioGRID [157], the Comprehensive Yeast Genome Database [66], the IntAct Molecular Interaction Database [79], and the Database of Interacting Proteins [187]. The high throughput PPI network used eleven different published experiments that used yeast-2-hybrid, protein-fragment complementation assay, and/or affinity purification followed by mass spectrometry to infer the PPI links. These eleven separate networks were combined into the final high throughput PPI network.

In addition to the PPI networks provided by YeastNet v3, there exists a literature curated PPI network of physical links from the STRING database [160] with 480,760 edges for 5,966 genes. The STRING database was not included in the literature curated PPI network from YeastNet v3 and contains a different selection of databases, such as KEGG [90] and Gene Ontology [29] to create the final network. This network was thresholded to only include edges where the experimental score was greater than 800. This threshold was determined using the RWR Network Score function within the `RWRtools` R package to score varying thresholds, in which the score per network edge dropped considerably at the experimental

score threshold of 800 [Supplementary Figure 5]. This network contained 58,016 edges for 1,954 genes.

The three PPI networks were merged, duplicate edges were dropped, and the resulting edge weight was set to one. The final merged PPI network contained 215,261 edges and 5,669 genes.

Transcription Factor-to-Gene Network

The transcription factor-to-gene network, which contains 307 transcription factors, 6,726 target genes, and 201,973 edges from YEASTRACT [162]. This network was literature curated and contained only high quality edges. Every edge in this network had an edge weight of one.

5.3.6 Gene Ontology Enrichment

The resulting retained genes from GRIN were then analyzed using the GO enrichment analysis pipeline through PANTHER [126]. The GO terms that were retained after the GO enrichment analyses were those with an FDR adjusted P-value less than 0.05.

5.3.7 Computational Resources

The predominant computing cluster used throughout this analyses was Andes, located at Oak Ridge National Laboratory as a part of the Oak Ridge Leadership Computing Facility. Andes contains two partitions, one with 704 CPU compute nodes and one with 9 GPU nodes. Each compute node contains two AMD EPYC 7302 CPUs per node with 16 cores per CPU for a total of 32 cores per node and 256 GB of memory per node. Each GPU node contains two Intel Xeon E5-2695 GPUs with 14 cores per GPU for a total of 28 cores per node and 1 TB of memory per node.

5.4 Results and Discussion

5.4.1 Multiplex of *S. cerevisiae* Genes is High Quality

The multiplex network that contained all eight *S. cerevisiae* gene-to-gene layers was scored using five sets of genes that were annotated with the GO terms GO:0006887, GO:0008213, GO:0016050, GO:0043543, and GO:0048856. Five sub-networks were created, one for each GO term, where the nodes were the genes in each set, and the nodes were connected if an edge existed in the multiplex. All five of these sub-networks were connected graphs, meaning there does not exist a single node that could be removed to separate the sub-network into two separate networks. The sub-network for the set of genes in GO:0006887 is depicted in Figure 5.1. The connectivity of these sub-networks showcases the quality of the multiplex network, since these known biological interactions between all of the genes in each set of genes are captured within the variety of layers.

The RWR Cross Validation function evaluates the multiplex network as a whole, using these sets of genes and producing the RWR rank of the left-out genes. This produced True Positive Rates, False Positive Rates or Recall, and Precision values that were plotted to produce Precision-Recall and Receiver Operating Characteristic Curves for each set of genes. The AUPR and AUROC values were also calculated for each set of genes. The AUPR values were 0.570, 0.241, 0.279, 0.180, and 0.098 and the AUROC values were 0.974, 0.841, 0.948, 0.872, and 0.749 for GO:0006887, GO:0008213, GO:0016050, GO:0043543, and GO:0048856 respectively. The average AUPR and average AUROC value was 0.274 and 0.877 respectively. The PR and ROC curves are depicted in Figure 5.2. The AUROC values and the corresponding curves show that the multiplex captures the these more rare and interconnected gene relationships better than random chance. The average AUROC value of 0.877 is also considerably higher than the random classifier value of 0.5. Since the random classifier value for AUPR is the ratio of true positive values to true negative values, these values are 4.630E-3, 6.033E-3, 9.820E-3, 8.558E-3, and 7.997E-3, for the set containing genes from GO:0006887, GO:0008213, GO:0016050, GO:0043543, and GO:0048856 respectively [147]. All five of the AUPR values are greater than their respective random classifier value, thus performing better than random chance. Thus, it appears that the multiplex performs

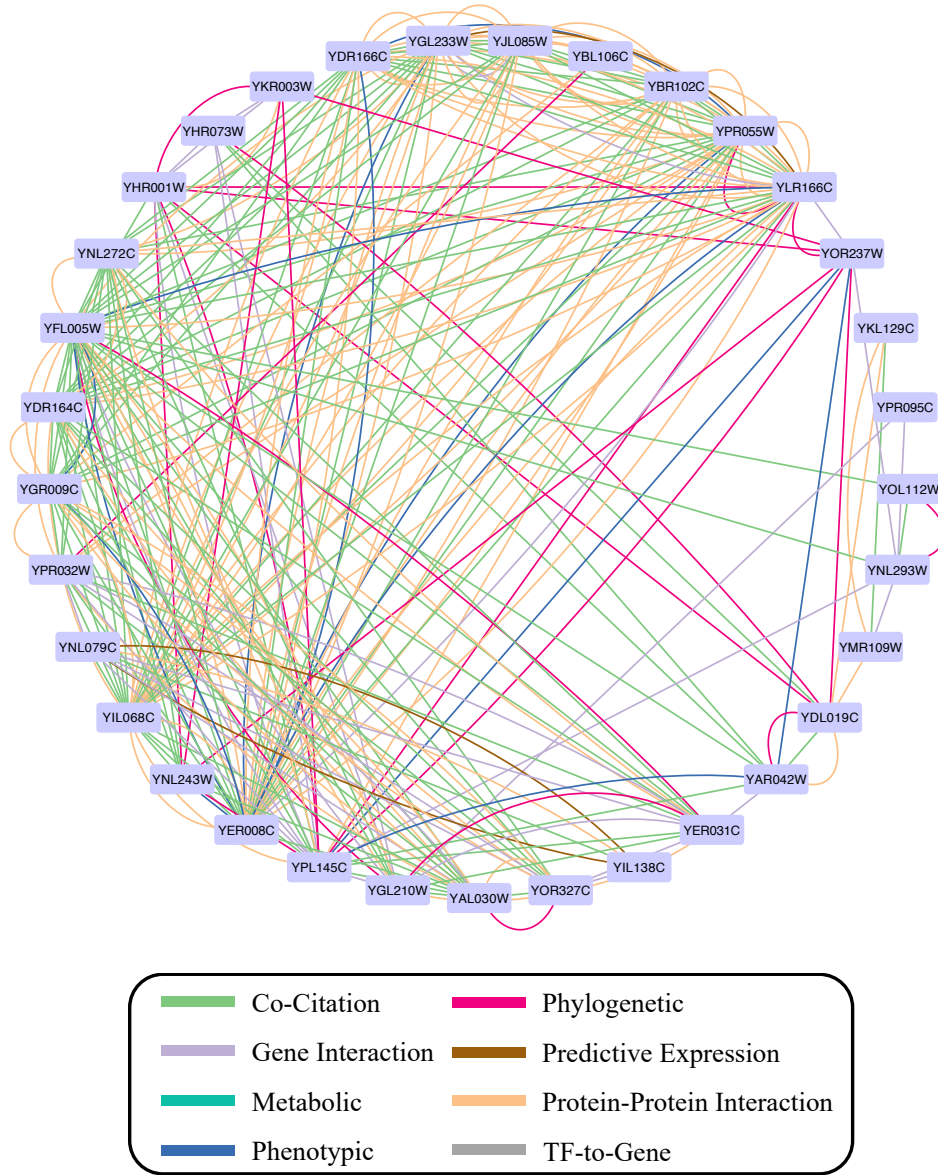
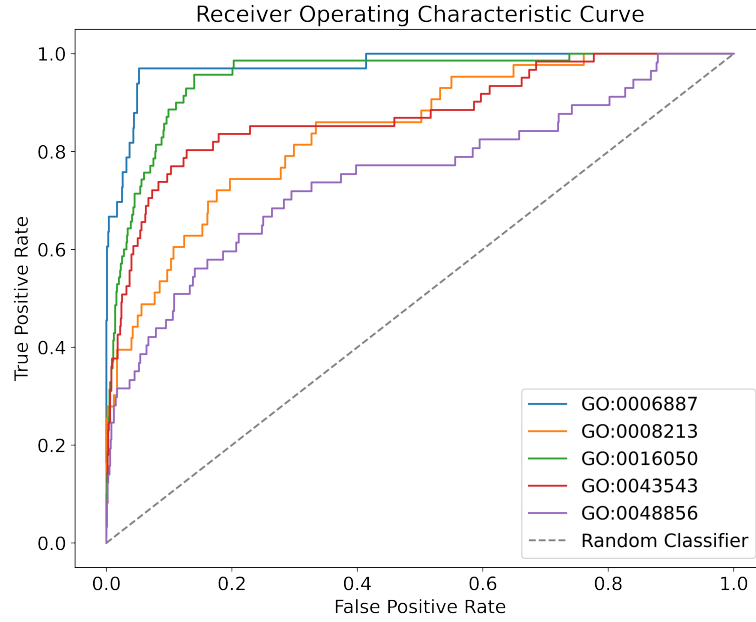
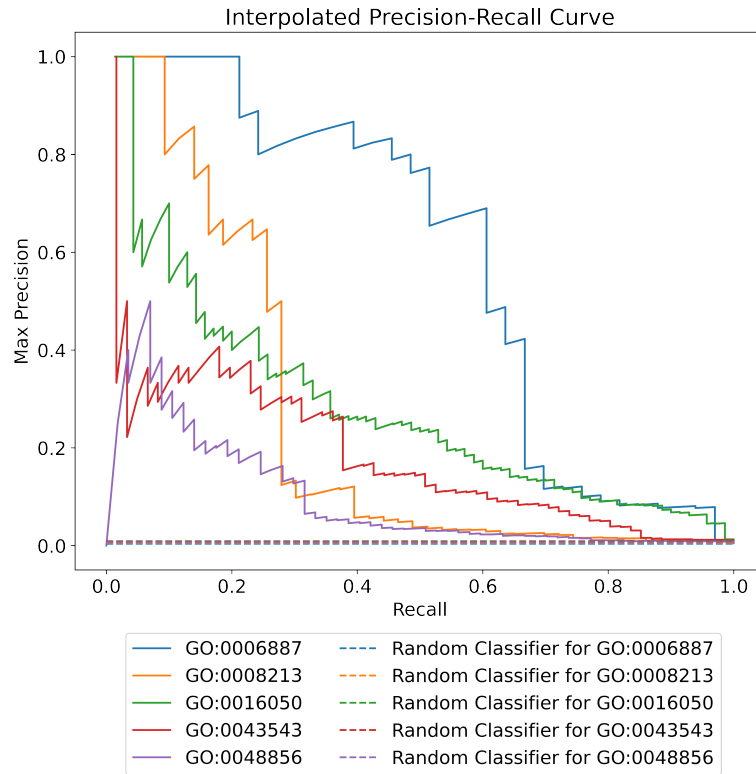


Figure 5.1: This figure represents all of the edges that exist in the multiplex between the 33 genes with the GO annotation GO:0006887. This network is connected, meaning that there does not exist any edge that can be removed to result in two disconnected networks. This depicts how sufficiently the multiplex captures known biological truth between groups of genes.



(a) ROC Curves from RWR Cross Validation



(b) PR Curves from RWR Cross Validation

Figure 5.2: This figure depicts some of the metrics of RWR Cross Validation. Both of these curves show how well RWR ranks the left-out genes for each of the five sets of genes with the same GO annotation. All five curves in both ROC and PR performed considerably better than the random classifier.

better at capturing the more rare gene-to-gene interactions than the ones seen in the larger gene sets.

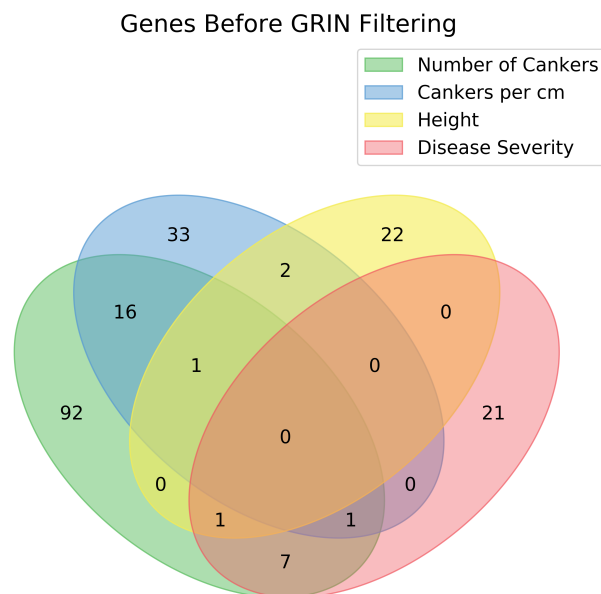
5.4.2 Seed and GRIN Retained *S. cerevisiae* Genes are Trait Specific

The four trait specific seed gene list before GRIN filtering had very little overlap across sets [Figure 5.3a]. The overlaps of note include the disease severity and number of cankers traits sharing seven genes and number of cankers and cankers per centimeter sharing sixteen genes. However, the latter may be due to the fact that cankers per centimeter is correlated with number of cankers. After GRIN filtering, the sixteen shared genes between number of cankers and cankers per centimeter dropped to seven genes [Figure 5.3b].

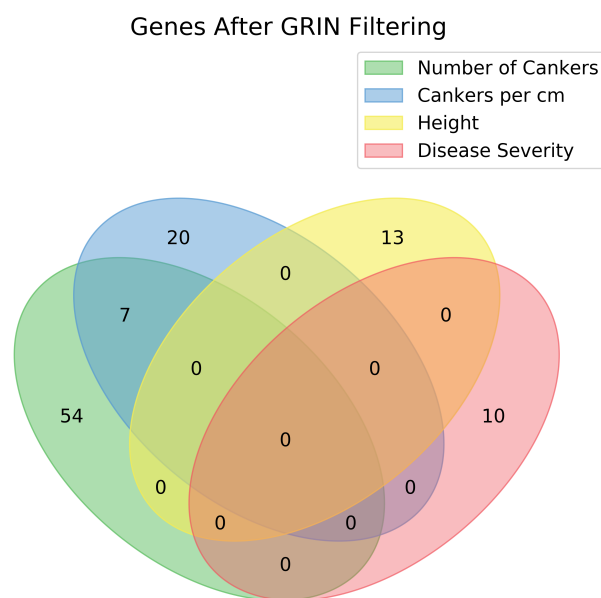
The seven genes that overlapped between disease severity and number of cankers were YER173W, YNL031C, YOR219C, YAL067C, YLL063C, YFL041W, YGL048C, YJR040W, and YNL321W. However, these genes were not enriched for any GO terms, and these seven genes were removed from either the number of cankers set of genes, the disease severity set of genes, or both after GRIN filtering. After GRIN filtering, no other sets of genes overlapped across traits. Thus, the *S. cerevisiae* genes that were retained after GRIN filtering were trait specific. The retained sets of genes were of size 61, 27, 13, and 10 for the traits number of cankers, cankers per centimeter, height, and disease severity respectively. The GRIN filtering that contained all 196 unique genes retained 46 genes. The GRIN filtering halved the number of seed genes, allowing for a smaller set of genes to analyze and gain biological insights from.

5.4.3 Genes Retained by GRIN May be Related to Fungal Infection

The GO enrichment analysis was performed on the retained gene sets from the original four sets of seed genes separated by trait as well as the combined set of all genes. The tables of enrichment results from PANTHER are attached in Supplementary Tables 19-26. Some



(a) Genes originally mapped from *S. musiva* to *S. cerevisiae* and used as seed genes in GRIN



(b) Genes retained after GRIN filtering

Figure 5.3: These venn diagrams showcase the lack of overlap of genes between sets that are related to different traits. After GRIN filtering, the only overlap occurs in number of cankers and cankers per centimeter, which is likely due to the fact that these two traits are highly correlated.

of the GO terms that are enriched in the resulting retained gene sets may be related to *S. musiva* fungal infection in *P. trichocarpa*.

Thiamine appears in multiple GO terms in the enrichment for genes for the trait number of cankers and cankers per centimeter, many of them with fold enrichment greater than 100. Thiamine diphosphate, also known as vitamin B1, is a cofactor involved in numerous metabolic pathways. Previous investigations have shown that in the rust fungus *Uromyces fabae* thiamine biosynthesis is active during the parasitic growth stage of plant infection [154].

Phytochelatins biosynthetic process and phytochelatins metabolic process are two GO terms with fold enrichment greater than 100 for the disease severity related gene set. Interestingly phytochelatins has been shown to protect plants, in particular *Arabidopsis thaliana*, against fungal infection [78]. While this does not explain why phytochelatins would be crucial in *S. musiva*, it is likely that disease severity is jointly contributed by biological processes occurring in both organisms.

5.5 Conclusion

In this study we have presented the compilation of various *S. cerevisiae* gene-to-gene networks and used these in a multiplex with RWR-based tools to refine suspected genes that influence fungal infection of *S. musiva* in *P. trichocarpa* to a finer set of potential true positives. We have shown that the multiplex used was of high quality, accurately capturing true and rare gene-to-gene relationships. We have shown that the genes tend to be trait specific. Additionally, we have shown that some of the genes that were retained may be related to the host-pathogen relationship.

One caveat to this study is the loss of information that occurs when mapping genes from one organism to another. Unfortunately, there is not enough known truth about the gene-to-gene relationships within *S. musiva* to perform an analysis such as this. Additionally, *S. cerevisiae* and *S. musiva* are entirely different organisms, and while these results may be meaningful it is possible that some truth found in *S. cerevisiae* is not true in *S. musiva*.

The genes found in this analysis may be beneficial in a future study when analyzing the colonization efficacy of *S. musiva* in *P. trichocarpa* and other poplar trees.

Conclusion

Iterative Random Forest (iRF) has been validated here to outperform Random Forest in the prediction of gene-to-gene expression based networks. Predominantly, iRF-Leave One Out Prediction (iRF-LOOP) produces Predictive Expression Networks (PENs) that perform an unsupervised culling of noisy edges while boosting the true biological interactions higher in the sorted edge list to be retained after thresholding. While iRF-LOOP is shown here to be successful in producing gene-to-gene expression based networks, it is possible that iRF-LOOP can be applied to a variety of data types to produce directed all-to-all networks.

Included here is an analysis into the environmental factors surrounding a patient that influences their survivability of COVID-19 infection. Utilizing iRF models with different classes of features, climatic features, demographic and socioeconomic features, and COVID-19 related policy features predicting the Case Fatality Ratio (CFR) at the county level, we have shown that COVID-19 survivability is dependent on the climatic features, predominantly those that capture the seasonality of the region such as the light spectrum absorption range of vitamin D. This knowledge may be used to predict the CFR within other regions and among the unvaccinated community.

The Divide-Test-Integrate (DTI) method was applied here to microarray gene expression data to produce higher quality PENs in *Saccharomyces cerevisiae* and *Arabidopsis thaliana*. Traditional network inference algorithms typically use all samples from many experiments in one matrix, while the DTI method separates samples into like groups, performs the network inference to produce the intermediate network, and integrates the intermediate networks into a final, high quality network. Here we show the PENs produced using DTI and iRF-LOOP produce higher quality networks that reduce the bias of consistently expressed housekeeping genes.

The high quality *S. cerevisiae* PEN produced above was also used in a multiplex network containing a variety of biological layers to mine for *Septoria musiva* genes that influence fungal infection of *Populus trichocarpa*. The multiplex itself was of high quality, and effectively captured rare interactions between genes. We identified *S. musiva* genes that were specific to the observed trait, and filtered these genes using Gene Refinement using Interacting Networks to those that were most likely involved in fungal infection.

Bibliography

- [1] (2021). The gene ontology resource: enriching a gold mine. *Nucleic acids research*, 49(D1):D325–D334. [96](#)
- [2] Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., et al. (2017). Scenic: single-cell regulatory network inference and clustering. *Nature methods*, 14(11):1083–1086. [38](#)
- [3] Ala, U., Piro, R. M., Grassi, E., Damasco, C., Silengo, L., Oti, M., Provero, P., and Di Cunto, F. (2008). Prediction of human disease genes by human-mouse conserved coexpression analysis. *PLoS computational biology*, 4(3):e1000043. [21](#)
- [4] Aliferis, C. F., Tsamardinos, I., and Statnikov, A. (2003). Hiton: a novel markov blanket algorithm for optimal variable selection. In *AMIA annual symposium proceedings*, volume 2003, page 21. American Medical Informatics Association. [8](#)
- [5] Alipio, M. (2020). Vitamin d supplementation could possibly improve clinical outcomes of patients infected with coronavirus-2019 (covid-19). *Available at SSRN*, 3571484. [30](#), [72](#), [78](#)
- [6] Ashoor, A. S. and Naji, A. A. K. K. (2021). Integrated network of pearson correlation coefficients, kendall tau, spearman rho and its impact on disease, health indicators and mortality ratings in babil province. In *Proceedings of First International Conference on Mathematical Modeling and Computational Science: ICMACS 2020*, pages 83–96. Springer. [9](#)
- [7] Ayres, D. L., Darling, A., Zwickl, D. J., Beerli, P., Holder, M. T., Lewis, P. O., Huelsenbeck, J. P., Ronquist, F., Swofford, D. L., Cummings, M. P., et al. (2012). Beagle: an application programming interface and high-performance computing library for statistical phylogenetics. *Systematic biology*, 61(1):170–173. [4](#)
- [8] Ball, G. H. and Hall, D. J. (1967). A clustering technique for summarizing multivariate data. *Behavioral science*, 12(2):153–155. [23](#)

- [9] Basu, S., Kumbier, K., Brown, J. B., and Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, 115(8):1943–1948. [2](#), [13](#), [35](#), [36](#), [64](#)
- [10] Bauer, W. (2021). Many states are not reporting the latest covid-19 numbers. [32](#), [63](#)
- [11] Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., and Huala, E. (2015). The arabidopsis information resource: making and mining the ”gold standard” annotated reference plant genome. *genesis*, 53(8):474–485. [96](#)
- [12] Blount, Z. D. (2015). The natural history of model organisms: The unexhausted potential of e. coli. *Elife*, 4:e05826. [25](#)
- [13] Bolón-Canedo, V., Sánchez-Marono, N., Alonso-Betanzos, A., Benítez, J. M., and Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. *Information sciences*, 282:111–135. [29](#), [82](#)
- [14] Bondy, J. A., Murty, U. S. R., et al. (1976). *Graph theory with applications*, volume 290. Macmillan London. [5](#)
- [15] Boulesteix, A.-L., Janitza, S., Kruppa, J., and König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6):493–507. [12](#)
- [16] Bray, J. R. and Curtis, J. T. (1957). An ordination of the upland forest communities of southern wisconsin. *Ecological monographs*, 27(4):326–349. [86](#)
- [17] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32. [12](#), [35](#)
- [18] Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Classification and regression trees. [12](#)
- [19] Bush, W. S. and Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS computational biology*, 8(12):e1002822. [29](#)

- [20] Centers for Disease Control and Prevention (2022). *COVID-19 Vaccinations in the United States, Jurisdiction*. Accessed: January 20, 2022. [63](#)
- [21] Chandrashekar, G. and Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28. [6](#)
- [22] Chang, T.-W. (1983). Binding of cells to matrixes of distinct antibodies coated on solid surface. *Journal of immunological methods*, 65(1-2):217–223. [29](#)
- [23] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4. [12](#)
- [24] Chen, X., Zhang, B., Wang, T., Bonni, A., and Zhao, G. (2020). Robust principal component analysis for accurate outlier sample detection in rna-seq data. *BMC bioinformatics*, 21(1):1–20. [16](#)
- [25] Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., et al. (2012). Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic acids research*, 40(D1):D700–D705. [25](#), [85](#), [109](#), [110](#)
- [26] Clarke, L., Glendinning, I., and Hempel, R. (1994). The mpi message passing interface standard. In *Programming environments for massively parallel distributed systems*, pages 213–218. Springer. [4](#)
- [27] Cliff, A., Romero, J., Kainer, D., Walker, A., Furches, A., and Jacobson, D. (2019). A high-performance computing implementation of iterative random forest for the creation of predictive expression networks. *Genes*, 10(12):996. [2](#), [9](#), [13](#), [21](#), [28](#), [38](#), [46](#), [64](#), [69](#), [82](#), [110](#)
- [28] Codling, E. A., Plank, M. J., and Benhamou, S. (2008). Random walk models in biology. *Journal of the Royal society interface*, 5(25):813–834. [10](#)
- [29] Consortium, G. O. (2017). Expansion of the gene ontology knowledgebase and resources. *Nucleic acids research*, 45(D1):D331–D338. [110](#)

- [30] Crick, F., Barnett, L., Brenner, S., Watts-Tobin, R. J., et al. (1961). General nature of the genetic code for proteins. [25](#)
- [31] Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., and Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, 88(11):2783–2792. [60](#)
- [32] Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. [17](#), [19](#)
- [33] de Winter, J. C., Gosling, S. D., and Potter, J. (2016). Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological methods*, 21(3):273. [7](#)
- [34] Dekkers, B. J., Pearce, S., van Bolderen-Veldkamp, R., Marshall, A., Widera, P., Gilbert, J., Drost, H.-G., Bassel, G. W., Müller, K., King, J. R., et al. (2013). Transcriptional dynamics of two seed compartments with opposing roles in arabidopsis seed germination. *Plant physiology*, 163(1):205–215. [54](#)
- [35] Deng, W., Zhang, K., Busov, V., and Wei, H. (2017). Recursive random forest algorithm for constructing multilayered hierarchical gene regulatory networks that govern biological pathways. *PloS one*, 12(2):e0171532. [9](#), [34](#)
- [36] Dhillon, B., Feau, N., Aerts, A. L., Beauseigle, S., Bernier, L., Copeland, A., Foster, A., Gill, N., Henrissat, B., Herath, P., et al. (2015). Horizontal gene transfer and gene dosage drives adaptation to wood colonization in a tree pathogen. *Proceedings of the National Academy of Sciences*, 112(11):3451–3456. [105](#)
- [37] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21. [41](#)
- [38] Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534. [30](#), [32](#), [62](#), [63](#), [66](#)

- [39] Dunn, J. C. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. [24](#)
- [40] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499. [7](#)
- [41] Ellis, B., Jansson, S., Strauss, S. H., and Tuskan, G. A. (2010). Why and how populus became a "model tree". In *Genetics and genomics of Populus*, pages 3–14. Springer. [25](#)
- [42] Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., and Gardner, T. S. (2007). Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, 5(1):e8. [8](#)
- [43] Fang, X., Li, S., Yu, H., Wang, P., Zhang, Y., Chen, Z., Li, Y., Cheng, L., Li, W., Jia, H., et al. (2020). Epidemiological, comorbidity factors with severity and prognosis of covid-19: a systematic review and meta-analysis. *Aging (Albany NY)*, 12(13):12493. [30](#), [63](#)
- [44] Feau, N., Mottet, M.-J., Périnet, P., Hamelin, R. C., and Bernier, L. (2010). Recent advances related to poplar leaf spot and canker caused by septoria musiva: Minireview/minisynthèse. *Canadian Journal of Plant Pathology*, 32(2):122–134. [101](#)
- [45] Fenga, L. and Gaspari, M. (2021). Predictive capacity of covid-19 test positivity rate. *Sensors*, 21(7):2435. [31](#)
- [46] Fick, S. E. and Hijmans, R. J. (2017). Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *International journal of climatology*, 37(12):4302–4315. [65](#)
- [47] Fischer, G., Nachtergaele, F., Prieler, S., Van Velthuisen, H., Verelst, L., and Wiberg, D. (2008). Global agro-ecological zones assessment for agriculture (gaez 2008). *IIASA, Laxenburg, Austria and FAO, Rome, Italy*, 10. [65](#)
- [48] Fischer, G., Nachtergaele, F. O., Prieler, S., Teixeira, E., Toth, G., van Velthuisen, H., Verelst, L., and Wiberg, D. (2012). *Global Agro-ecological Zones (GAEZ v3.0) Model*

- Documentation*. International Institute for Applied Systems Analysis (IIASA) and the Food and Agriculture Organization of the United Nations (FAO), Laxenburg, Austria and Rome, Italy. [65](#), [155](#)
- [49] Foury, F. (1997). Human genetic diseases: a cross-talk between man and yeast. *Gene*, 195(1):1–10. [25](#)
- [50] Freeman, T. C., Goldovsky, L., Brosch, M., Van Dongen, S., Mazière, P., Grocock, R. J., Freilich, S., Thornton, J., and Enright, A. J. (2007). Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS computational biology*, 3(10):e206. [21](#)
- [51] Freund, Y., Schapire, R. E., et al. (1996). Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer. [12](#)
- [52] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232. [12](#)
- [53] Friedman, N., Linial, M., Nachman, I., and Pe’er, D. (2000). Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620. [7](#)
- [54] Gabriel, E., Fagg, G. E., Bosilca, G., Angskun, T., Dongarra, J. J., Squyres, J. M., Sahay, V., Kambadur, P., Barrett, B., Lumsdaine, A., et al. (2004). Open mpi: Goals, concept, and design of a next generation mpi implementation. In *European Parallel Virtual Machine/Message Passing Interface Users’ Group Meeting*, pages 97–104. Springer. [4](#)
- [55] Gagniuc, P. A. (2017). *Markov chains: from theory to implementation and experimentation*. John Wiley & Sons. [24](#)
- [56] Gagnon-Bartsch, J. A. and Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552. [15](#)
- [57] Gama-Castro, S., Salgado, H., Peralta-Gil, M., Santos-Zavaleta, A., Muniz-Rascado, L., Solano-Lira, H., Jimenez-Jacinto, V., Weiss, V., Garcia-Sotelo, J. S., Lopez-Fuentes, A., et al. (2010). Regulondb version 7.0: transcriptional regulation of escherichia coli k-12

- integrated within genetic sensory response units (gensor units). *Nucleic acids research*, 39(suppl_1):D98–D105. [42](#)
- [58] Gentschev, I., Dietrich, G., and Goebel, W. (2002). The e. coli α -hemolysin secretion system and its use in vaccine development. *Trends in microbiology*, 10(1):39–45. [25](#)
- [59] Ghani, A. C., Donnelly, C. A., Cox, D. R., Griffin, J., Fraser, C., Lam, T., Ho, L., Chan, W., Anderson, R., Hedley, A., et al. (2005). Methods for estimating the case fatality ratio for a novel, emerging infectious disease. *American journal of epidemiology*, 162(5):479–486. [31](#)
- [60] Gierlichs, B., Batina, L., Tuyls, P., and Preneel, B. (2008). Mutual information analysis. In *International Workshop on Cryptographic Hardware and Embedded Systems*, pages 426–442. Springer. [8](#)
- [61] Giuntoli, B., Lee, S. C., Licausi, F., Kosmacz, M., Oosumi, T., van Dongen, J. T., Bailey-Serres, J., and Perata, P. (2014). A trihelix dna binding protein counterbalances hypoxia-responsive transcriptional activation in arabidopsis. *PLoS biology*, 12(9):e1001950. [152](#)
- [62] Goeddel, D. V., Kleid, D. G., Bolivar, F., Heyneker, H. L., Yansura, D. G., Crea, R., Hirose, T., Kraszewski, A., Itakura, K., and Riggs, A. D. (1979). Expression in escherichia coli of chemically synthesized genes for human insulin. *Proceedings of the National Academy of Sciences*, 76(1):106–110. [25](#)
- [63] Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 genes. *Science*, 274(5287):546–567. [25](#), [102](#)
- [64] Goh, W. W. B., Wang, W., and Wong, L. (2017). Why batch effects matter in omics data, and how to avoid them. *Trends in biotechnology*, 35(6):498–507. [14](#), [15](#)
- [65] Gropp, W., Lusk, E., Doss, N., and Skjellum, A. (1996). A high-performance, portable implementation of the mpi message passing interface standard. *Parallel computing*, 22(6):789–828. [4](#)

- [66] Güldener, U., Münsterkötter, M., Kastenmüller, G., Strack, N., van Helden, J., Lemer, C., Richelles, J., Wodak, S. J., García-Martínez, J., Pérez-Ortín, J. E., et al. (2005). Cygd: the comprehensive yeast genome database. *Nucleic acids research*, 33(suppl_1):D364–D368. [26](#), [110](#)
- [67] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182. [6](#)
- [68] Hakes, L., Pinney, J. W., Robertson, D. L., and Lovell, S. C. (2008). Protein-protein interaction networks and biology-what’s the connection? *Nature biotechnology*, 26(1):69–72. [27](#)
- [69] Harris, E. H. (2001). Chlamydomonas as a model organism. *Annual review of plant biology*, 52(1):363–406. [25](#)
- [70] Hart, G. T., Ramani, A. K., and Marcotte, E. M. (2006). How complete are current yeast and human protein-interaction networks? *Genome biology*, 7(11):1–9. [27](#)
- [71] Hartemink, A. et al. (2005). Banjo: Bayesian network inference with java objects. [8](#)
- [72] Hartuv, E. and Shamir, R. (2000). A clustering algorithm based on graph connectivity. *Information processing letters*, 76(4-6):175–181. [22](#)
- [73] Haury, A.-C., Mordelet, F., Vera-Licona, P., and Vert, J.-P. (2012). Tigress: trustful inference of gene regulation using stability selection. *BMC systems biology*, 6(1):1–17. [7](#)
- [74] Hawkins, C., Ginzburg, D., Zhao, K., Dwyer, W., Xue, B., Xu, A., Rice, S., Cole, B., Paley, S., Karp, P., et al. (2021). Plant metabolic network 15: A resource of genome-wide metabolism databases for 126 plants and algae. *Journal of Integrative Plant Biology*. [43](#)
- [75] He, C.-c. and Gao, X.-g. (2018). Structure learning of bayesian networks based on the lars-mmpe ordering search method. In *2018 37th Chinese Control Conference (CCC)*, pages 9000–9006. IEEE. [9](#)

- [76] He, F., Yoo, S., Wang, D., Kumari, S., Gerstein, M., Ware, D., and Maslov, S. (2016). Large-scale atlas of microarray data reveals the distinct expression landscape of different tissues in arabidopsis. *The Plant Journal*, 86(6):472–480. [85](#)
- [77] Hegedus, D. D., Coutu, C., Harrington, M., Hope, B., Gerbrandt, K., and Nikolov, I. (2015). Multiple internal sorting determinants can contribute to the trafficking of cruciferin to protein storage vacuoles. *Plant molecular biology*, 88(1-2):3–20. [53](#), [54](#)
- [78] Hématy, K., Lim, M., Cherk, C., Piślewska-Bednarek, M., Sanchez-Rodriguez, C., Stein, M., Fuchs, R., Klapprodt, C., Lipka, V., Molina, A., et al. (2020). Moonlighting function of phytochelatase1 in extracellular defense against fungal pathogens. *Plant Physiology*, 182(4):1920–1932. [117](#)
- [79] Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., et al. (2004). Intact: an open source molecular interaction database. *Nucleic acids research*, 32(suppl_1):D452–D455. [110](#)
- [80] Howe, K. L., Contreras-Moreira, B., De Silva, N., Maslen, G., Akanni, W., Allen, J., Alvarez-Jarreta, J., Barba, M., Bolser, D. M., Cambell, L., et al. (2020). Ensembl genomes 2020?enabling non-vertebrate genomic research. *Nucleic acids research*, 48(D1):D689–D695. [105](#)
- [81] Hu, Y., Zhou, L., Yang, Y., Zhang, W., Chen, Z., Li, X., Qian, Q., Kong, F., Li, Y., Liu, X., et al. (2021). The gibberellin signaling negative regulator rga-like3 promotes seed storage protein accumulation. *Plant Physiology*, 185(4):1697–1707. [53](#), [54](#)
- [82] Hu, Z., Killion, P. J., and Iyer, V. R. (2007). Genetic reconstruction of a functional transcriptional regulatory network. *Nature genetics*, 39(5):683–687. [42](#)
- [83] Huh, S. U. (2021). New function of hypoxia-responsive unknown protein in enhanced resistance to biotic stress. *Plant Signaling & Behavior*, 16(3):1868131. [152](#)

- [84] Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):e12776. [2](#), [9](#), [13](#), [35](#), [40](#)
- [85] IGBP-DIS (1998). Soildata (v. 0). a program for creating global soil-property databases. *IGBP Global Soils Data Task*. [65](#)
- [86] Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446. [20](#), [44](#), [89](#)
- [87] Jin, J., He, K., Tang, X., Li, Z., Lv, L., Zhao, Y., Luo, J., and Gao, G. (2015). An arabidopsis transcriptional regulatory map reveals distinct functional and evolutionary features of novel transcription factors. *Molecular biology and evolution*, 32(7):1767–1773. [42](#), [89](#)
- [88] Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127. [15](#), [87](#), [110](#)
- [89] Kanawati, R. (2015). Multiplex network mining: A brief survey. *IEEE Intell. Informatics Bull.*, 16(1):24–27. [9](#), [99](#), [101](#)
- [90] Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1):D353–D361. [27](#), [110](#)
- [91] Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30. [53](#), [54](#)
- [92] Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723. [29](#)
- [93] Karlebach, G. and Shamir, R. (2008). Modelling and analysis of gene regulatory networks. *Nature reviews Molecular cell biology*, 9(10):770–780. [28](#)

- [94] Karp, P. D., Billington, R., Caspi, R., Fulcher, C. A., Latendresse, M., Kothari, A., Keseler, I. M., Krummenacker, M., Midford, P. E., Ong, Q., et al. (2019). The biocyc collection of microbial genomes and metabolic pathways. *Briefings in bioinformatics*, 20(4):1085–1093. [26](#), [27](#), [87](#), [109](#)
- [95] Kaul, S., Koo, H. L., Jenkins, J., Rizzo, M., Rooney, T., Tallon, L. J., Feldblyum, T., Nierman, W., Benito, M.-I., Lin, X., et al. (2000). Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *nature*, 408(6814):796–815. [26](#)
- [96] Kawakatsu, T., Huang, S.-s. C., Jupe, F., Sasaki, E., Schmitz, R. J., Urich, M. A., Castanon, R., Nery, J. R., Barragan, C., He, Y., et al. (2016). Epigenomic diversity in a global collection of arabidopsis thaliana accessions. *Cell*, 166(2):492–505. [37](#), [41](#)
- [97] Kim, H., Shin, J., Kim, E., Kim, H., Hwang, S., Shim, J. E., and Lee, I. (2014). Yeastnet v3: a public database of data-specific and integrated functional gene networks for saccharomyces cerevisiae. *Nucleic acids research*, 42(D1):D731–D736. [26](#), [83](#), [107](#), [109](#)
- [98] King, A. D., Pržulj, N., and Jurisica, I. (2004). Protein complex prediction via cost-based clustering. *Bioinformatics*, 20(17):3013–3020. [23](#)
- [99] King Jr, J. T., Yoon, J. S., Rentsch, C. T., Tate, J. P., Park, L. S., Kidwai-Khan, F., Skanderson, M., Hauser, R. G., Jacobson, D. A., Erdos, J., et al. (2020). Development and validation of a 30-day mortality index based on pre-existing medical administrative data from 13,323 covid-19 patients: The veterans health administration covid-19 (vaco) index. *PLoS One*, 15(11):e0241825. [63](#)
- [100] Kolak, M., Li, X., Lin, Q., Wang, R., Menghaney, M., Yang, S., and Anguiano Jr, V. (2021). The us covid atlas: A dynamic cyberinfrastructure surveillance system for interactive exploration of the pandemic. *Transactions in GIS*, 25(4):1741–1765. [32](#)
- [101] Kumar, A., Singh, R., Kaur, J., Pandey, S., Sharma, V., Thakur, L., Sati, S., Mani, S., Asthana, S., Sharma, T. K., et al. (2021). Wuhan to world: the covid-19 pandemic. *Frontiers in Cellular and Infection Microbiology*, 11:242. [30](#), [62](#)

- [102] Langfelder, P. and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):1–13. [34](#)
- [103] Lee, I., Ambaru, B., Thakkar, P., Marcotte, E. M., and Rhee, S. Y. (2010). Rational association of genes with traits using a genome-scale gene network for arabidopsis thaliana. *Nature biotechnology*, 28(2):149–156. [26](#)
- [104] Lee, I., Li, Z., and Marcotte, E. M. (2007). An improved, bias-reduced probabilistic functional gene network of baker’s yeast, *saccharomyces cerevisiae*. *PloS one*, 2(10):e988. [82](#), [83](#), [110](#)
- [105] Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739. [14](#)
- [106] Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161. [15](#)
- [107] Lenz, R. R., Louie, K. B., Søndreli, K. L., Galanie, S. S., Chen, J.-G., Muchero, W., Bowen, B. P., Northen, T. R., and LeBoldus, J. M. (2021). Metabolomic patterns of septoria canker resistant and susceptible populus trichocarpa genotypes 24 hours postinoculation. *Phytopathology®*, 111(11):2052–2066. [101](#)
- [108] Leonelli, S. and Ankeny, R. A. (2013). What makes a model organism? *Endeavour*, 37(4):209–212. [25](#)
- [109] Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology*, 49(4):764–766. [15](#)
- [110] Likas, A., Vlassis, N., and Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461. [23](#)

- [111] Lin, K., Hu, Y., and Kong, G. (2019). Predicting in-hospital mortality of patients with acute kidney injury in the icu using random forest model. *International journal of medical informatics*, 125:55–61. [12](#)
- [112] Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833–835. [29](#)
- [113] Liu, W., Li, L., Ye, H., Chen, H., Shen, W., Zhong, Y., Tian, T., and He, H. (2017). From *saccharomyces cerevisiae* to human: The important gene co-expression modules. *Biomedical reports*, 7(2):153–158. [25](#)
- [114] Liu, X., Huang, J., Li, C., Zhao, Y., Wang, D., Huang, Z., and Yang, K. (2021). The role of seasonality in the spread of covid-19 pandemic. *Environmental research*, 195:110874. [30](#), [63](#), [72](#)
- [115] Luo, G., Zhang, X., Zheng, H., and He, D. (2021). Infection fatality ratio and case fatality ratio of covid-19. *International Journal of Infectious Diseases*, 113:43–46. [31](#)
- [116] MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic acids research*, 45(D1):D896–D901. [28](#)
- [117] MacIsaac, K. D., Wang, T., Gordon, D. B., Gifford, D. K., Stormo, G. D., and Fraenkel, E. (2006). An improved map of conserved regulatory sites for *saccharomyces cerevisiae*. *BMC bioinformatics*, 7(1):1–14. [42](#)
- [118] Makarenkov, V., Kevorkov, D., and Legendre, P. (2006). Phylogenetic network construction approaches. *Applied mycology and biotechnology*, 6:61–97. [27](#)
- [119] Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Kellis, M., Collins, J. J., and Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804. [6](#), [9](#), [13](#), [35](#), [40](#), [42](#)

- [120] Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., and Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International journal of methods in psychiatric research*, 27(2):e1608. [28](#)
- [121] Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006). Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC bioinformatics*, volume 7, pages 1–15. Springer. [8](#)
- [122] Marsh, E. K. and May, R. C. (2012). *Caenorhabditis elegans*, a model organism for investigating immunity. *Applied and environmental microbiology*, 78(7):2075–2081. [25](#)
- [123] Marx, V. (2013). The big challenges of big data. *Nature*, 498(7453):255–260. [4](#)
- [124] McClure, R. S., Wendler, J. P., Adkins, J. N., Swanstrom, J., Baric, R., Kaiser, B. L. D., Oxford, K. L., Waters, K. M., and McDermott, J. E. (2019). Unified feature association networks through integration of transcriptomic and proteomic data. *PLoS computational biology*, 15(9):e1007241. [12](#), [60](#)
- [125] Memarsadeghi, N., Mount, D. M., Netanyahu, N. S., and Le Moigne, J. (2007). A fast implementation of the isodata clustering algorithm. *International Journal of Computational Geometry & Applications*, 17(01):71–103. [23](#)
- [126] Mi, H., Muruganujan, A., Ebert, D., Huang, X., and Thomas, P. D. (2019). Panther version 14: more genomes, a new panther go-slim and improvements in enrichment analysis tools. *Nucleic acids research*, 47(D1):D419–D426. [111](#)
- [127] Mistry, M. and Pavlidis, P. (2008). Gene ontology term overlap as a measure of gene functional similarity. *BMC bioinformatics*, 9(1):1–11. [109](#)
- [128] Mueller, L. A., Zhang, P., and Rhee, S. Y. (2003). Aracyc: a biochemical pathway database for arabidopsis. *Plant physiology*, 132(2):453–460. [26](#), [42](#), [89](#)

- [129] Müller, C., Schillert, A., Röthmeier, C., Trégouët, D.-A., Proust, C., Binder, H., Pfeiffer, N., Beutel, M., Lackner, K. J., Schnabel, R. B., et al. (2016). Removing batch effects from longitudinal gene expression-quantile normalization plus combat as best approach for microarray transcriptome data. *PloS one*, 11(6):e0156594. [15](#)
- [130] Mustroph, A., Lee, S. C., Oosumi, T., Zanetti, M. E., Yang, H., Ma, K., Yaghoubi-Masihi, A., Fukao, T., and Bailey-Serres, J. (2010). Cross-kingdom comparison of transcriptomic adjustments to low-oxygen stress highlights conserved and plant-specific responses. *Plant Physiology*, 152(3):1484–1500. [152](#)
- [131] Nicodemus, K. K. and Malley, J. D. (2009). Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics*, 25(15):1884–1890. [14](#)
- [132] Nielsen, J. (2019). Yeast systems biology: model organism and cell factory. *Biotechnology journal*, 14(9):1800421. [25](#), [102](#)
- [133] Nygaard, V., Rødland, E. A., and Hovig, E. (2016). Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, 17(1):29–39. [14](#)
- [134] Ohm, R. A., Feau, N., Henrissat, B., Schoch, C. L., Horwitz, B. A., Barry, K. W., Condon, B. J., Copeland, A. C., Dhillon, B., Glaser, F., et al. (2012). Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen dothideomycetes fungi. *PLoS pathogens*, 8(12):e1003037. [105](#)
- [135] Patro, V. M. and Patra, M. R. (2014). Augmenting weighted average with confusion matrix to enhance classification accuracy. *Transactions on Machine Learning and Artificial Intelligence*, 2(4):77–91. [17](#)
- [136] Perkins, A. D. and Langston, M. A. (2009). Threshold selection in gene co-expression networks using spectral graph theory techniques. In *BMC bioinformatics*, volume 10, pages 1–11. BioMed Central. [22](#), [59](#)

- [137] Perrin, B.-E., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J., and d’Alche Buc, F. (2003). Gene networks inference using dynamic bayesian networks. *Bioinformatics*, 19(suppl_2):ii138–ii148. [8](#)
- [138] Petralia, F., Wang, P., Yang, J., and Tu, Z. (2015). Integrative random forest for gene regulatory network inference. *Bioinformatics*, 31(12):i197–i205. [9](#)
- [139] Phifer-Rixey, M. and Nachman, M. W. (2015). The natural history of model organisms: Insights into mammalian biology from the wild house mouse *mus musculus*. *Elife*, 4:e05959. [25](#)
- [140] Qiao, N., Huang, Y., Naveed, H., Green, C. D., and Han, J.-D. J. (2013). Cociter: an efficient tool to infer gene function by assessing the significance of literature co-citation. *PloS one*, 8(9):e74074. [27](#)
- [141] Rahimi, N. R., Fouladi-Fard, R., Aali, R., Shahryari, A., Rezaali, M., Ghafouri, Y., Ghalhari, M. R., Asadi-Ghalhari, M., Farzinnia, B., Gea, O. C., et al. (2021). Bidirectional association between covid-19 and the environment: a systematic review. *Environmental Research*, 194:110692. [30](#), [63](#)
- [142] Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of artificial intelligence research*, 11:95–130. [90](#), [104](#)
- [143] Roberts, D. B. (2006). *Drosophila melanogaster*: the model organism. *Entomologia experimentalis et applicata*, 121(2):93–103. [25](#)
- [144] Rodriguez, J. J., Kuncheva, L. I., and Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *IEEE transactions on pattern analysis and machine intelligence*, 28(10):1619–1630. [12](#)
- [145] Russell, T. W., Hellewell, J., Jarvis, C. I., Van Zandvoort, K., Abbott, S., Ratnayake, R., Flasche, S., Eggo, R. M., Edmunds, W. J., Kucharski, A. J., et al. (2020). Estimating the infection and case fatality ratio for coronavirus disease (covid-19) using

- age-adjusted data from the outbreak on the diamond princess cruise ship, february 2020. *Eurosurveillance*, 25(12):2000256. [31](#)
- [146] Saha, P., Ray, T., Tang, Y., Dutta, I., Evangelous, N. R., Kieliszewski, M. J., Chen, Y., and Cannon, M. C. (2013). Self-rescue of an extensin mutant reveals alternative gene expression programs and candidate proteins for new cell wall assembly in a rabidopsis. *The Plant Journal*, 75(1):104–116. [53](#), [54](#), [55](#)
- [147] Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432. [19](#), [112](#)
- [148] Sanbonmatsu, K. and Tung, C.-S. (2007). High performance computing in biology: multimillion atom simulations of nanoscale systems. *Journal of structural biology*, 157(3):470–480. [4](#)
- [149] Schuyler, R. P., Jackson, C., Garcia-Perez, J. E., Baxter, R. M., Ogolla, S., Rochford, R., Ghosh, D., Rudra, P., and Hsieh, E. W. (2019). Minimizing batch effects in mass cytometry data. *Frontiers in immunology*, 10:2367. [14](#)
- [150] Shah, R. D. and Meinshausen, N. (2014). Random intersection trees. *The Journal of Machine Learning Research*, 15(1):629–654. [13](#), [36](#), [45](#)
- [151] Sharan, R. and Shamir, R. (2000). Click: a clustering algorithm with applications to gene expression analysis. In *Proc Int Conf Intell Syst Mol Biol*, volume 8, page 16. [22](#)
- [152] Shieh, A. D. and Hung, Y. S. (2009). Detecting outlier samples in microarray data. *Statistical applications in genetics and molecular biology*, 8(1). [15](#)
- [153] Smid, M., van den Braak, R. R. C., van de Werken, H. J., van Riet, J., van Galen, A., de Weerd, V., van der Vlugt-Daane, M., Bril, S. I., Lalmahomed, Z. S., Kloosterman, W. P., et al. (2018). Gene length corrected trimmed mean of m-values (getmm) processing of rna-seq data performs similarly in intersample analyses while improving intrasample comparisons. *BMC bioinformatics*, 19(1):1–13. [41](#)

- [154] Sohn, J., Voegelé, R. T., Mendgen, K., and Hahn, M. (2000). High level activation of vitamin b1 biosynthesis genes in haustoria of the rust fungus *Uromyces fabae*. *Molecular plant-microbe interactions*, 13(6):629–636. [117](#)
- [155] Somerville, C. and Koornneef, M. (2002). A fortunate choice: the history of arabidopsis as a model plant. *Nature Reviews Genetics*, 3(11):883–889. [26](#)
- [156] Sonnhammer, E. L. and Östlund, G. (2015). Inparanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic acids research*, 43(D1):D234–D239. [105](#)
- [157] Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl_1):D535–D539. [110](#)
- [158] Sul, J. H., Martin, L. S., and Eskin, E. (2018). Population structure in genetic studies: Confounding factors and mixed models. *PLoS genetics*, 14(12):e1007309. [29](#)
- [159] Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Müller, R., Ploetz, L., et al. (2007). The arabidopsis information resource (tair): gene structure and function annotation. *Nucleic acids research*, 36(suppl_1):D1009–D1014. [26](#)
- [160] Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., et al. (2019). String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, 47(D1):D607–D613. [27](#), [110](#)
- [161] Tan, X., Meyers, B. C., Kozik, A., West, M. A., Morgante, M., St Clair, D. A., Bent, A. F., and Michelmore, R. W. (2007). Global expression analysis of nucleotide binding site-leucine rich repeat-encoding and related genes in arabidopsis. *BMC plant biology*, 7(1):1–20. [55](#)
- [162] Teixeira, M. C., Monteiro, P. T., Palma, M., Costa, C., Godinho, C. P., Pais, P., Cavaleiro, M., Antunes, M., Lemos, A., Pedreira, T., et al. (2018). Yeabstract: an

- upgraded database for the analysis of transcription regulatory networks in *saccharomyces cerevisiae*. *Nucleic acids research*, 46(D1):D348–D353. [26](#), [87](#), [111](#)
- [163] The Associated Press (2022). Tennessee moving from daily to weekly covid-19 data updates. [32](#)
- [164] The New York Times (2020). *nytimes/covid-19-data/mask-use*. Accessed: July 28 2020. [66](#)
- [165] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288. [7](#)
- [166] Tong, H., Faloutsos, C., and Pan, J.-Y. (2008). Random walk with restart: fast solutions and applications. *Knowledge and Information Systems*, 14(3):327–346. [60](#), [81](#)
- [167] Trabucco, A. and Zomer, R. (2010). High-resolution global soil-water balance explicit for climate–standard vegetation and soil conditions. *CGIAR Consortium for Spatial Information*. [65](#)
- [168] Tsamardinos, I., Aliferis, C. F., and Statnikov, A. (2003). Time and sample efficient discovery of markov blankets and direct causal relations. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 673–678. [8](#)
- [169] Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., et al. (2006). The genome of black cottonwood, *populus trichocarpa* (torr. & gray). *science*, 313(5793):1596–1604. [37](#), [41](#)
- [170] United States Census Bureau (2019). *American Community Survey: Data Profiles*. Accessed: Mar. 11 2021. [65](#)
- [171] U.S. Department of Health & Human Services (2021). *COVID-19 State and County Policy Orders*. Accessed: March 12 2021. [66](#)

- [172] Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., Odelin, G., Levy, N., Cau, P., Remy, E., and Baudot, A. (2019). Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics*, 35(3):497–505. [2](#), [10](#), [101](#), [102](#)
- [173] Van Dam, S., Vosa, U., van der Graaf, A., Franke, L., and de Magalhaes, J. P. (2018). Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in bioinformatics*, 19(4):575–592. [28](#)
- [174] Van Dongen, S. M. (2000). *Graph clustering by flow simulation*. PhD thesis. [24](#), [85](#)
- [175] Voy, B. H., Scharff, J. A., Perkins, A. D., Saxton, A. M., Borate, B., Chesler, E. J., Branstetter, L. K., and Langston, M. A. (2006). Extracting gene networks for low-dose radiation using graph theoretical algorithms. *PLoS computational biology*, 2(7):e89. [22](#)
- [176] Walker, K. and Herman, M. (2021). *tidycensus: Load US Census Boundary and Attribute Data as ‘tidyverse’ and ‘sf’-Ready Data Frames*. R package version 0.11.4. [65](#)
- [177] Wang, B., Li, R., Lu, Z., and Huang, Y. (2020). Does comorbidity increase the risk of patients with covid-19: evidence from meta-analysis. *Aging (Albany NY)*, 12(7):6049. [30](#)
- [178] Wang, J. and Zhang, Z. (2021). Gapit version 3: boosting power and accuracy for genomic association and prediction. *Genomics, proteomics & bioinformatics*. [105](#)
- [179] Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., and Chen, C.-F. (2007). A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281. [90](#), [104](#)
- [180] Wang, S.-B., Feng, J.-Y., Ren, W.-L., Huang, B., Zhou, L., Wen, Y.-J., Zhang, J., Dunwell, J. M., Xu, S., and Zhang, Y.-M. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Scientific reports*, 6(1):1–10. [29](#)
- [181] Weigel, D. and Mott, R. (2009). The 1001 genomes project for arabidopsis thaliana. *Genome biology*, 10(5):1–5. [26](#), [37](#)

- [182] Weighill, D., Jones, P., Shah, M., Ranjan, P., Muchero, W., Schmutz, J., Sreedasyam, A., Macaya-Sanz, D., Sykes, R., Zhao, N., et al. (2018). Pleiotropic and epistatic network-based discovery: integrated networks for target gene discovery. *Frontiers in Energy Research*, 6:30. [9](#), [101](#)
- [183] Wells, C. K. (2021). Louisiana will soon be one of the few southern states with a mask mandate still in effect. [76](#)
- [184] Wienkoop, S., Baginsky, S., and Weckwerth, W. (2010). Arabidopsis thaliana as a model organism for plant proteome research. *Journal of proteomics*, 73(11):2239–2248. [25](#)
- [185] Wilcox, G. L., Quiocho, F. A., Levinthal, C., and Harvey^o, S. C. (1987). Minnesota conference on supercomputing in biology: Proteins, nucleic acids, and water. *Journal of Computer-Aided Molecular Design*, 1:271–281. [4](#)
- [186] Wilson, A. M. and Jetz, W. (2016). Remotely sensed high-resolution global cloud dynamics for predicting ecosystem and biodiversity distributions. *PLoS biology*, 14(3):e1002415. [65](#)
- [187] Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000). Dip: the database of interacting proteins. *Nucleic acids research*, 28(1):289–291. [110](#)
- [188] Xu, J. and Ye, J. (2020). Perspectives on supercomputing and artificial intelligence applications in drug discovery. *Supercomputing Frontiers and Innovations*, 7(3). [4](#)
- [189] Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678. [22](#)
- [190] Yanagida, M. (2002). The model unicellular eukaryote, schizosaccharomyces pombe. *Genome biology*, 3(3):1–4. [25](#)
- [191] Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82. [105](#)

- [192] Yang, M.-S. (1993). A survey of fuzzy clustering. *Mathematical and Computer modelling*, 18(11):1–16. [24](#)
- [193] Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). Gosemsim: an r package for measuring semantic similarity among go terms and gene products. *Bioinformatics*, 26(7):976–978. [90](#), [104](#)
- [194] Zhou, X. and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods*, 11(4):407–409. [29](#)
- [195] Zhou, Y., Xing, C., Shen, W., Sun, Y., Wu, J., and Zhou, X. (2011). A fast algorithm for outlier detection in microarray. In *International Conference on Computer Science, Environment, Ecoinformatics, and Education*, pages 513–519. Springer. [16](#)
- [196] Zomer, R. J., Bossio, D. A., Trabucco, A., Yuanjie, L., Gupta, D. C., and Singh, V. P. (2007). *Trees and water: smallholder agroforestry on irrigated lands in Northern India*, volume 122. IWMI. [65](#)
- [197] Zomer, R. J., Trabucco, A., Bossio, D. A., and Verchot, L. V. (2008). Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. *Agriculture, ecosystems & environment*, 126(1-2):67–80. [65](#)

Appendices

A Chapter 2 Supplementary Information

Table 1: This table contains the AUPR values for the networks within the DREAM4 and DREAM5 challenges.

DREAM Chal- lenge	Base Learner	Algorithm	Number of Trees Per Itera- tion	Net 1	Net 2	Net 3	Net 4	Net 5
DREAM4	RF	GENIE3 (Official Scores)	1000	0.154	0.155	0.231	0.208	0.197
DREAM4	RF	GENIE3	1000	0.165	0.158	0.237	0.218	0.213
DREAM4	RF	GENIE3	5000	0.161	0.158	0.249	0.220	0.213
DREAM4	iRF	iRF- LOOP	1000	0.177	0.150	0.267	0.248	0.253
DREAM5	RF	GENIE3 (Official Scores)	1000	0.291	-	0.093	0.021	-
DREAM5	RF	GENIE3	1000	0.303	-	0.095	0.020	-
DREAM5	iRF	iRF- subLOOP	1000	0.387	-	0.070	0.020	-

Table 2: This table contains the AUROC values for the networks within the DREAM4 and DREAM5 challenges.

DREAM Chal- lenge	Base Learner	Algorithm	Number of Trees Per Itera- tion	Net 1	Net 2	Net 3	Net 4	Net 5
DREAM4	RF	GENIE3 (Official Scores)	1000	0.745	0.733	0.775	0.791	0.798
DREAM4	RF	GENIE3	1000	0.755	0.738	0.786	0.802	0.801
DREAM4	RF	GENIE3	5000	0.758	0.739	0.784	0.800	0.805
DREAM4	iRF	iRF- LOOP	1000	0.743	0.711	0.775	0.794	0.792
DREAM5	RF	GENIE3 (Official Scores)	1000	0.815	-	0.617	0.518	-
DREAM5	RF	GENIE3	1000	0.816	-	0.617	0.516	-
DREAM5	iRF	iRF- subLOOP	1000	0.816	-	0.614	0.516	-

Table 3: This table contains the AUPR p-values for the networks within the DREAM4 and DREAM5 challenges.

DREAM Chal- lenge	Base Algorithm Lear- ner	Number of Trees Per Itera- tion	Net 1	Net 2	Net 3	Net 4	Net 5	P- Value
DREAM4 RF	GENIE3 (Official Scores)	1000	3.3e- 34	7.9e- 54	1.8e- 54	5.5e- 47	4.6e- 44	1.0e- 46
DREAM4 RF	GENIE3	1000	7.7e- 37	3.5e- 55	4.8e- 56	1.3e- 49	6.8e- 48	4.1e- 49
DREAM4 RF	GENIE3	5000	5.5e- 36	4.8e- 55	9.3e- 57	5.7e- 50	8.5e- 48	4.1e- 49
DREAM4 iRF	iRF- LOOP	1000	9.5e- 40	1.5e- 51	8.3e- 64	5.6e- 57	1.2e- 57	6.1e- 54
DREAM5 RF	GENIE3 (Official Scores)	1000	1.60e- 104	-	5.15e- 20	1.58e- 01	-	41.295
DREAM5 RF	GENIE3	1000	6.18e- 121	-	3.54e- 21	2.34e- 01	-	47.097
DREAM5 iRF	iRF- subLOOP	1000	3.81e- 269	-	1.32e- 07	6.473e- 01	-	91.829

Table 4: This table contains the AUROC p-values for the networks within the DREAM4 and DREAM5 challenges.

DREAM Chal- lenge	Base Algorithm Lear- ner	Number of Trees Per Itera- tion	Net 1	Net 2	Net 3	Net 4	Net 5	P- Value
DREAM4 RF	GENIE3 (Official Scores)	1000	3.3e- 18	1.1e- 28	9.7e- 34	6.7e- 33	1.9e- 34	1.4e- 29
DREAM4 RF	GENIE3	1000	3.1e- 19	7.1e- 30	2.6e- 36	5.2e- 35	5.2e- 35	4.4e- 31
DREAM4 RF	GENIE3	5000	1.3e- 19	4.1e- 30	8.4e- 36	1.2e- 34	5.8e- 36	3.1e- 31
DREAM4 iRF	iRF- LOOP	1000	6.6e- 18	3.0e- 24	1.1e- 33	1.3e- 33	2.7e- 33	1.5e- 28
DREAM5 RF	GENIE3 (Official Scores)	1000	3.06e- 106	-	5.00e- 11	1.06e- 02	-	39.263
DREAM5 RF	GENIE3	1000	1.19e- 107	-	3.65e- 11	4.78e- 02	-	39.561
DREAM5 iRF	iRF- subLOOP	1000	6.39e- 108	-	1.75e- 09	4.40e- 02	-	39.103

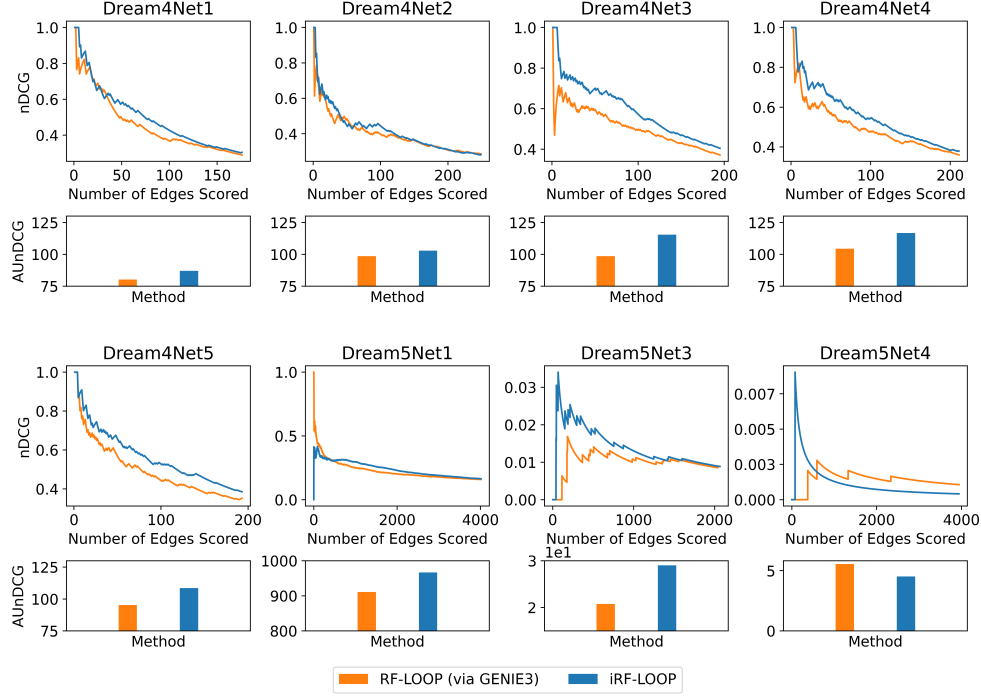


Figure 1: This figure depicts the nDCG scores for all of the DREAM4 and DREAM5 networks for both GENIE3 and iRF-LOOP as the number of edges scored, k , increases. The maximum k for each network depends on the number of true positive values in the corresponding gold standard network, ranging from 176 to 249 edges for the DREAM4 networks and 2,066 to 4,012 edges for the DREAM5 networks. For all networks except the DREAM5 Network 4, the AUnDCG for iRF-LOOP is higher than the AUnDCG for GENIE3, this suggests that iRF-LOOP outperforms GENIE3.

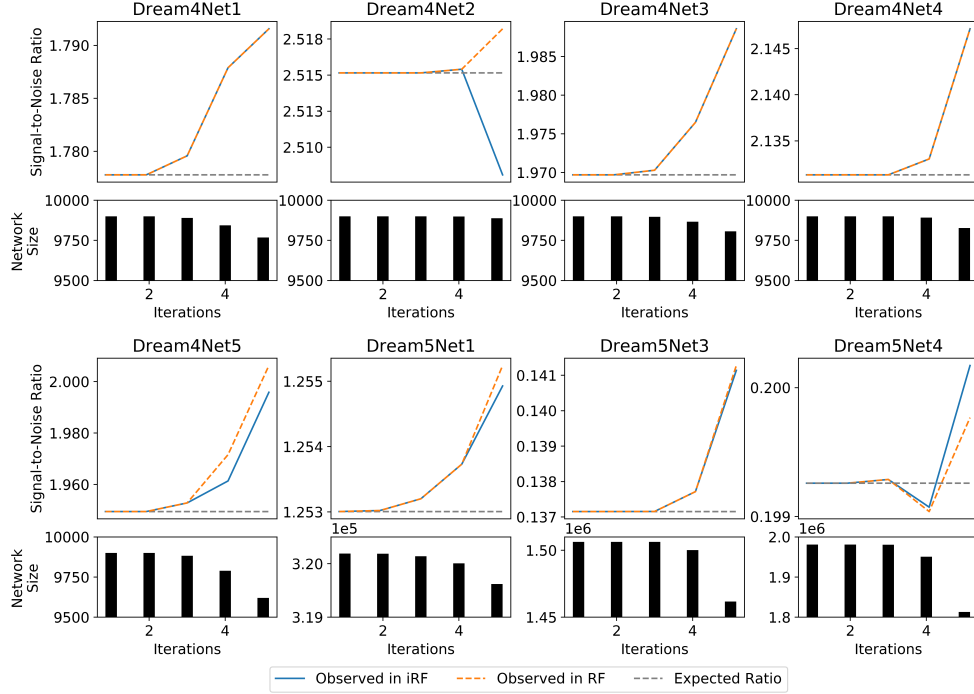


Figure 2: This figure depicts the true signal-to-noise ratio for each iteration of iRF-LOOP for the DREAM challenges as the blue line. The orange line is the corresponding GENIE3 network thresholded to match the same number of edges as the iRF-LOOP network, used to confirm that the unsupervised thresholding contributes to the improvement in signal-to-noise ratio. Unlike the larger empirical networks shown in Figure 2.3, these DREAM networks are too small to make a considerable difference when thresholding the networks.

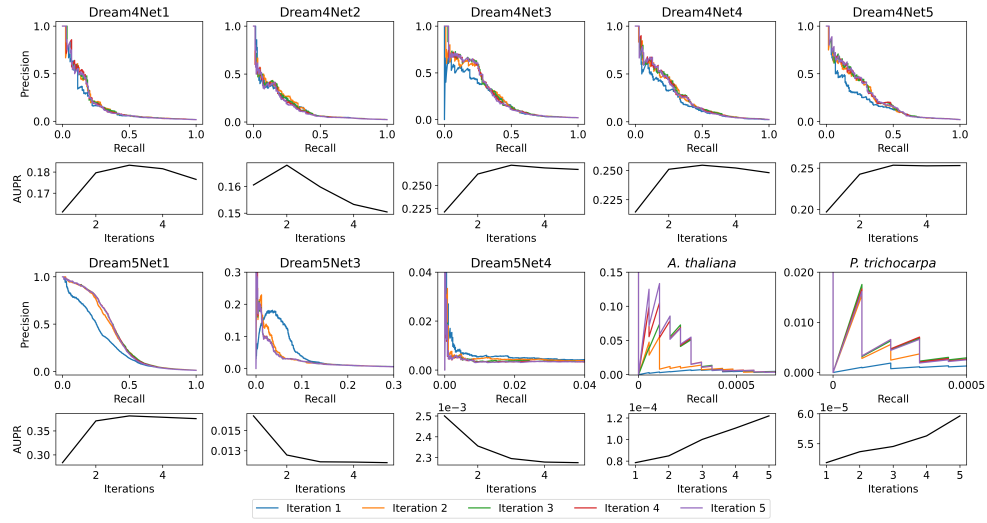


Figure 3: This figure depicts the Precision-Recall curves and the corresponding AUPR values for each iteration in the networks used in this analysis. For many of the synthetic DREAM networks, the improvement in AUPR plateaus after 2 or 3 iterations, which may be due to the size of the networks. This may suggest that iterations improve RF, but larger networks are needed to verify this. The Precision-Recall curves themselves show minimal improvement after the second iteration for the empirical networks. However, the AUPR value increases as the number of iterations increases, even after the second iteration. Thus this shows that the addition of iterations to RF improves the AUPR values for the two empirical data sets used in this study.

Table 5: This table depicts a sample fourth order gene set that were discovered using RIT on the resulting paths from iRF-LOOP on *A. thaliana* expression data. This set had a prevalence of 0.331. This set contains a mix of both known and unknown gene to gene relationships.

Gene ID	Function	Target	Target Function	Relationship
AT3G10040	Hypoxia	AT3G27220	Hypoxia	HUP43, HUP29,
—	Response		Response	and HUP6 are
AT5G39890	Attenuator		Unknown	all Hypoxia
—	1 (HRA1)		Protein 6	- Responsive
AT1G12805	— Plant		(HUP6)	Unknown
—	Cysteine			Proteins
AT5G15120	Oxidase 2			(HUPs). HRA1
	(PCO2),			interacts with
	Hypoxia			RAP2.12,
	Response			RAP2.12
	Unknown			binds to PCO1
	Protein 43			promoters in
	(HUP43)			the Hypoxia-
—				Responsive
	Nucleotide			Promoter
	binding			Element
	protein			regions, which
— Plant				includes HUP6.
	Cysteine			RAP2.12 also
	Oxidase 1			binds to HUP29
	(PCO1),			and HUP43
	Hypoxia			[130, 61, 83].
	Response			
	Unknown			
	Protein 29			
	(HUP29)			

Table 6: This table depicts a sample fifth order gene set that were discovered using RIT on the resulting paths from iRF-LOOP on *A. thaliana* expression data. This set had a prevalence of 0.153. This set contains a mix of both known and unknown gene to gene relationships.

Gene ID	Function	Target	Target Function	Relation- ship
AT4G3860	Ubiquitin-	AT1G70320	Ubiquitin-	UPL1,
—	Protein Ligase		Protein	UPL2,
AT3G50370	3 (UPL3) —		Ligase 2	UPL3
—	Hypothetical		(UPL2)	are all
AT1G80070	protein —			ubiquitin
—	Encodes a factor			protein
AT3G02260	that influences			ligases
—	pre-mRNA			
AT1G55860	splicing and			
	is required			
	for embryonic			
	development			
	— Calossin-like			
	protein required			
	for polar auxin			
	transport —			
	Ubiquitin-			
	Protein Ligase 1			
	(UPL1)			

B Chapter 3 Supplementary Information

Table 7: Static climate features details.

Description	Calculation
Actual evapotranspiration	Average actual Evapo-transpiration, 50 year average
Global elevation	None
Irrigated cultivated land via GMIA	See [48] for details
Irrigated cultivated land	See [48] for details
Rain-fed cultivated land	See [48] for details
Percent desert scrubland	See [48] for details
Excess salt concentration	See [48] for details
Percent forest based on tree cover	See [48] for details
Slope east (high resolution)	See [48] for details
Slope north (high resolution)	See [48] for details
Slope north (low resolution)	See [48] for details
Undefinable slope/dense topographical changes (high resolution)	See [48] for details
Undefinable slope/dense topographical changes (low resolution)	See [48] for details
Slope west (high resolution)	See [48] for details
Slope west (low resolution)	See [48] for details
FAO's global median elevation estimate	See [48] for details
Number of 3 arc second grid cells that belong to the land mask and fall into 5 minutes grid cells	See [48] for details
Percent grassland based on tree cover	See [48] for details
Barren/very sparsely vegetated land	See [48] for details
Rooting conditions based on soil compaction and particle size	See [48] for details
Oxygenation estimates of soil	See [48] for details
Soil salt concentration estimation	See [48] for details
Soil characteristics (soil texture, soil organic carbon, soil pH, total exchangeable bases)	See [48] for details
Soil organic carbon, soil texture, base saturation, cation exchange capacity of soil and of clay fraction	See [48] for details
Soil drainage and soil phases affecting soil drainage	See [48] for details
Soil acidity based on hydrogen released in solution	See [48] for details
Soil textures, bulk density, coarse fragments, vertic soil properties and soil phases affecting root penetration and soil depth and soil volume	See [48] for details
Calcium carbonate and gypsum	See [48] for details
Percent urban cover based on population density	See [48] for details
Percent water availability based on natural and synthetic water mechanisms	See [48] for details

Table 8: Monthly climate features details.

Description	Calculation
Monthly aridity and wind speed	Aridity (scaled 0-1) $\times$ WindSpeed (Scaled 0-1)
Monthly percent cloud cover	Weather station and satellite interpolation
Monthly soil water concentration as a percent of max capacity	Thornthwaite-Mather water balance
Monthly average precipitation	Weather station and satellite interpolation
Monthly solar radiation in watts/ $m^2/s-1$	Weather station and satellite interpolation
Monthly average temperature	Weather station and satellite interpolation
Monthly average maximum temperature	Weather station and satellite interpolation
Monthly average minimum temperature	Weather station and satellite interpolation
Monthly average vapor pressure	Tetens vapor pressure calculation
Monthly wind speed in meters/second	Weather station and satellite interpolation
Monthly potential evapotranspiration	Hargreaves PET Method
Monthly light spectrum vitamin D absorption range	Integral of watts/ $m^2/s-1$ between 280nm and 315 multiplied by fractional ozone cover and cloud density

Table 9: Census features details (table one of four).

Category	Number of Features	Calculation
Total population	1	None
Sex by age by ethnicity and race	252	Normalized by total population
Ancestry reported	108	Normalized by total population
Nativity and citizenship status	5	Normalized by total population
Period of naturalization	7	Normalized by total population
Geographical mobility in the past year by age	90	Normalized by total population
Means of transportation to work by sex	32	Normalized by total population of each sex
Whether place of work is in state, in county, or neither by sex	8	Normalized by total population of each sex
Commute length in minutes by sex	24	Normalized by total population of each sex
Number of workers in household by vehicles available	20	Normalized by number of workers
Whether a household received Supplemental Security Income (SSI), cash public assistance income, or food stamps/SNAP in the past 12 months	12	Normalized by total population
Breakdown of relationships between householder and child (own child, grandchild, etc.)	7	Normalized by total population
Breakdown of whether adults live with relatives, spouse, alone, or with non-relatives by age	24	Normalized by total population of each age group
Children who live with their grandparents by age	3	Normalized by total population
Whether or not a household contains relatives or non-relatives	2	Normalized by total households
Breakdown of marital status by age by sex	184	Normalized by total population of each sex
Breakdown of woman who have given birth in the past 12 months by marital status and age	18	Normalized by total population
Breakdown of school enrollment by level of schooling, type of schooling, and sex	46	Normalized by total population of each sex
Disability status by sex and age	24	Normalized by total population of each sex

Table 10: Census features details (table two of four).

Category	Number of Features	Calculation
Hearing difficulty status by sex and age	24	Normalized by total population of each sex
Vision difficulty status by sex and age	24	Normalized by total population of each sex
Cognitive difficulty status by sex and age	20	Normalized by total population of each sex
Ambulatory difficulty status by sex and age	20	Normalized by total population of each sex
Self-care difficulty status by sex and age	20	Normalized by total population of each sex
Independent living difficulty status by sex and age	16	Normalized by total population of each sex
Breakdown of household income in the last 12 months in USD	16	Normalized by total households
Gini Index of income inequality	1	None
Veteran status of the civilian population by sex by age	26	Normalized by total population of each sex
Breakdown of veterans vs. non-veterans by educational attainment	8	Normalized by either the number of veterans or number of non-veterans
Breakdown of work status by age	28	Normalized by total population of each age bracket
Total number of housing units	1	None
Household occupancy status (occupied vs. vacant)	2	Normalized by total number of housing units
Household tenure status (owner vs. renter)	2	Normalized by total number of housing units
Household vacancy status (for rent, for sale, sold, etc)	7	Normalized by total number of housing units
Whether the current resident of the vacant household resides elsewhere	2	Normalized by total number of housing units
Race of householder	9	Normalized by total number of housing units
Tenure by age of householder	18	Normalized by either the total number of owner occupied or renter occupied housing units
Tenure by household size	14	Normalized by either the total number of owner occupied or renter occupied housing units

Table 11: Census features details (table three of four).

Category	Number of Features	Calculation
Tenure by educational attainment	8	Normalized by either the total number of owner occupied or renter occupied housing units
Tenure by occupants per room	10	Normalized by either the total number of owner occupied or renter occupied housing units
Tenure by plumbing facilities by occupants per room	4	Normalized by either the total number of owner occupied or renter occupied housing units
Median number of rooms in a household	1	None
Tenure by number of rooms	18	Normalized by either the total number of owner occupied or renter occupied housing units
Median number of rooms by tenure	2	None
Breakdown of households by number of units in structure	10	Normalized by total number of housing units
Mortgage status by age of householder	16	Normalized by either the total number of owner occupied or renter occupied housing units
Median gross rent in USD	1	None
Tenure by units in structure	20	Normalized by either the total number of owner occupied or renter occupied housing units
Breakdown of year structure was built	10	Normalized by total number of housing units
Median year structure was built	1	None
Breakdown of year structure was built by tenure	20	Normalized by either the total number of owner occupied or renter occupied housing units
Median year structure was built by tenure	2	None
Year householder moved into the unit by tenure	12	Normalized by either the total number of owner occupied or renter occupied housing units
Median year householder moved into the unit by tenure	2	None
Breakdown of heating fuel used in households	9	Normalized by total number of housing units

Table 12: Census features details (table four of four).

Category	Number of Features	Calculation
Breakdown of number of bedrooms	6	Normalized by total number of housing units
Breakdown of number of bedrooms by tenure	12	Normalized by either the total number of owner occupied or renter occupied housing units
Breakdown of telephone service availability by tenure	16	Normalized by either the total number of owner occupied or renter occupied housing units
Breakdown of vehicles available by tenure	12	Normalized by either the total number of owner occupied or renter occupied housing units
Breakdown of kitchen facilities present by tenure	4	Normalized by either the total number of owner occupied or renter occupied housing units
Whether meals were included in rent by age	8	Normalized by total population in each age bracket
Breakdown of gross rent of units in USD	26	Normalized by total number of renter occupied units
Whether utilities are included in rent	2	Normalized by total number of renter occupied units
Breakdown of gross rent as a percentage of household income in the past 12 months	10	Normalized by total number of renter occupied units
Breakdown of home value in USD	26	Normalized by total housing units
Mortgage status of owner occupied units	7	Normalized by total owner occupied units
Monthly owner costs in USD by mortgage status	36	Normalized by total owner occupied units
Monthly housing costs in USD	16	Normalized by total occupied units
Health insurance coverage status by sex and by age	36	Normalized by sex and age
Breakdown of health insurance coverage type by age	61	Normalized by total population of each age bracket
Breakdown of computer presence and internet availability by age	15	Normalized by total population of each age bracket
Breakdown of Bachelor's degrees types for the adult population	5	Normalized by total population over 25 years old
Breakdown of language spoken at home for the population over 5 years old	37	Normalized by total population over 5 years old

Table 13: Policy features details (table one of two).

Description	Binary or nonbinary
Religious gatherings exempt from stay at home and safer at home orders	Binary
Mask mandate enforced by fines	Binary
Mask mandate enforced by criminal charge or citation	Binary
No legal enforcement of mask mandate	Binary
Attempt by state government to prevent local government from implementing mask mandate	Binary
Alcohol/liquor stores open	Binary
Firearms sellers open	Binary
COVID-19 not an acceptable reason to request application mail-in ballot	Binary
Witness or notary signature required for mail-in ballot	Binary
Automatic mail-in ballot system in response for COVID-19	Binary
Stay at home order	Non-binary
Safer at home order	Non-binary
Closure of K-12 schools	Non-binary
Closure of day cares	Non-binary
Banned visitors from nursing homes	Non-binary
Closed nonessential businesses	Non-binary
Closed restaurants	Non-binary
Closed gyms	Non-binary
Closed movie theaters	Non-binary
Closed bars	Non-binary
Closed casinos	Non-binary
Closed bars for a second time	Non-binary
Closed movie theaters for a second time	Non-binary

Table 14: Policy features details (table two of two).

Description	Binary or nonbinary
Closed gyms for a second time	Non-binary
Closed restaurants for a second time	Non-binary
Closed casinos for a second time	Non-binary
Public mask mandate	Non-binary
Business mask mandate	Non-binary
Quarantine mandate for some travelers	Non-binary
Quarantine mandate for all travelers	Non-binary
Overall eviction moratorium	Non-binary
Second overall eviction moratorium	Non-binary
Third overall eviction moratorium	Non-binary
Freeze initiation of evictions	Non-binary
Second freeze initiation of evictions	Non-binary
Suspend court hearings for evictions	Non-binary
Second suspend court hearings for evictions	Non-binary
Freeze enforcement of evictions	Non-binary
Second freeze enforcement of evictions	Non-binary
COVID-19 hardship policy	Non-binary
Second COVID-19 hardship policy	Non-binary
CARES act pleading	Non-binary
Open with CDC guidance	Non-binary
Utilities shutoff moratorium	Non-binary
Utilities re-connection	Non-binary
Stopped personal visitation in state prisons	Non-binary

Table 15: Number of samples in each iRF model and number of features in each class used.

Model	Climate Features	Census Features	Policy Features	Total Features (Columns)	Samples (Rows)
Overall	63	1351	47	1461	3112
Monthly	29	1316	46	1391	23870
March	29	1119	47	1195	793
April	27	1255	45	1327	1769
May	29	1272	47	1348	1860
June	28	1289	46	1363	2123
July	31	1320	46	1397	2637
August	31	1304	44	1379	2718
September	30	1321	43	1394	2833
October	29	1328	44	1401	3001
November	29	1340	45	1414	3073
December	28	1339	46	1413	3063
Total Possible	185/53 ¹	1608	48	1841/1709 ¹	—

¹ Features in overall model/Features in monthly models

C Chapter 4 Supplementary Information

Table 16: Number of samples and genes in each group of samples in *S. cerevisiae*. (Table one of two)

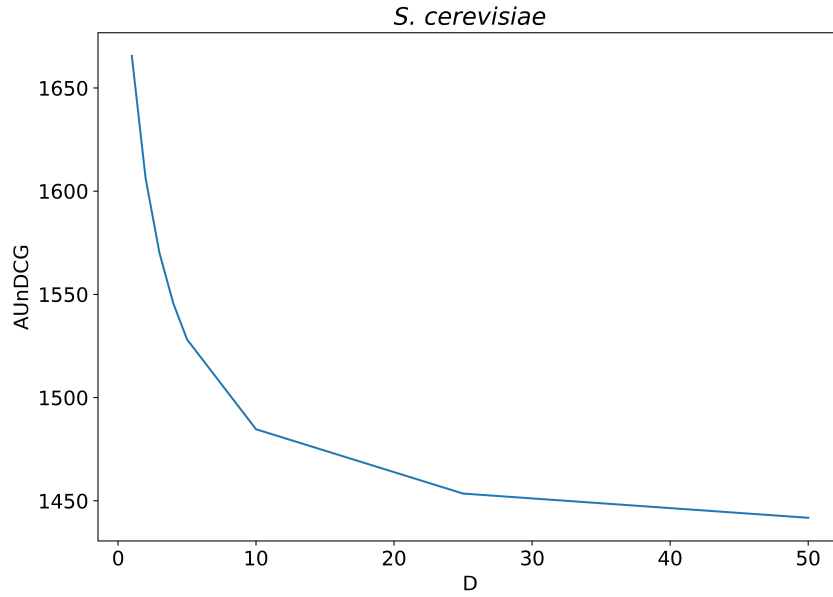
Biological Tag	Number of Samples	Number of Genes
Amino Acid Metabolism	39	5052
Amino Acid Utilization	495	6166
Carbon Utilization	2331	6576
Cell Aging	51	4815
Cell Cycle Regulation	343	5724
Cell Morphogenesis	26	5998
Cellular Ion Homeostasis	99	6222
Chemical Stimulus	2347	6687
Chromatin Organization	602	6430
Cofactor Metabolism	44	4869
Diauxic Shift	110	6212
Disease	47	4729
DNA Damage Stimulus	509	6460
DNA Replication Recombination and Repair	37	5704
Evolution	710	6370
Fermentation	846	6463
Filamentous Growth	163	6258
Flocculation	25	4314
Genetic Interaction	287	5581
Genome Variation	20	5008
Heat Shock	938	6314
Histone Modification	417	6607
Lipid Metabolism	81	6314
mRNA Processing	176	6296
Mating	117	6167
Metabolism	13	5615
Metal or Metalloid Ion Stress	127	6437
Mitotic Cell Cycle	425	6069

Table 17: Number of samples and genes in each group of samples in *S. cerevisiae*. (Table two of two)

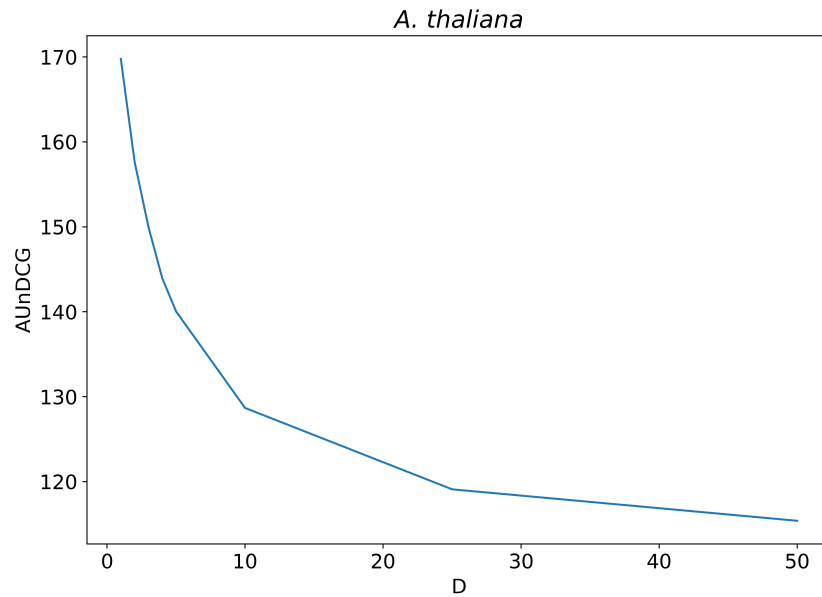
Biological Tag	Number of Samples	Number of Genes
Nitrogen Utilization	871	6414
Nutrient Utilization	727	6217
Organelles Biogenesis Structure and Function	24	3567
Osmotic Stress	828	6880
Oxidative Stress	1069	6874
Oxygen Level Alteration	547	6451
Phosphorus Utilization	561	6461
Ploidy	155	5678
Protein Dephosphorylation	79	6229
Protein Modification	35	4577
Protein Phosphorylation	562	6521
Protein Trafficking Localization and Degradation	27	5655
Proteolysis	59	5862
QTLs	131	6167
Radiation	95	6183
Respiration	814	6361
Response to Unfolded Protein	380	6329
RNA Catabolism	406	6211
Sporulation	334	6242
Starvation	447	6256
Stationary Phase Entry	191	6335
Stationary Phase Maintenance	689	5710
Sulfur Utilization	392	6382
Synthetic Biology	31	5581
Transcription	1597	6639
Transcriptional Regulation	229	4981
Translational Regulation	36	4767
Ubiquitin or ULP Modification	93	6236

Table 18: Number of samples and genes in each group of samples in *A. thaliana*.

Group Label	Number of Samples	Number of Genes
Group1	1152	21678
Group2	914	21678
Group3	689	21678
Group4	594	21678
Group5	554	21678
Group6	328	21678
Group7	325	21678
Group8	213	21678
Group9	193	21678
Group10	124	21678
Group11	122	21678
Group12	103	21678
Group13	746	21678



(a) AUnDCG for the integrated *S. cerevisiae* PEN as D varies.



(b) AUnDCG for the integrated *A. thaliana* PEN as D varies.

Figure 4: This figure depicts the AUnDCG score for the integrated final PEN in DTI iRF-LOOP for varying values of D in the WS formula. The optimal value for D in both organisms is $D = 1$.

D Chapter 5 Supplementary Information

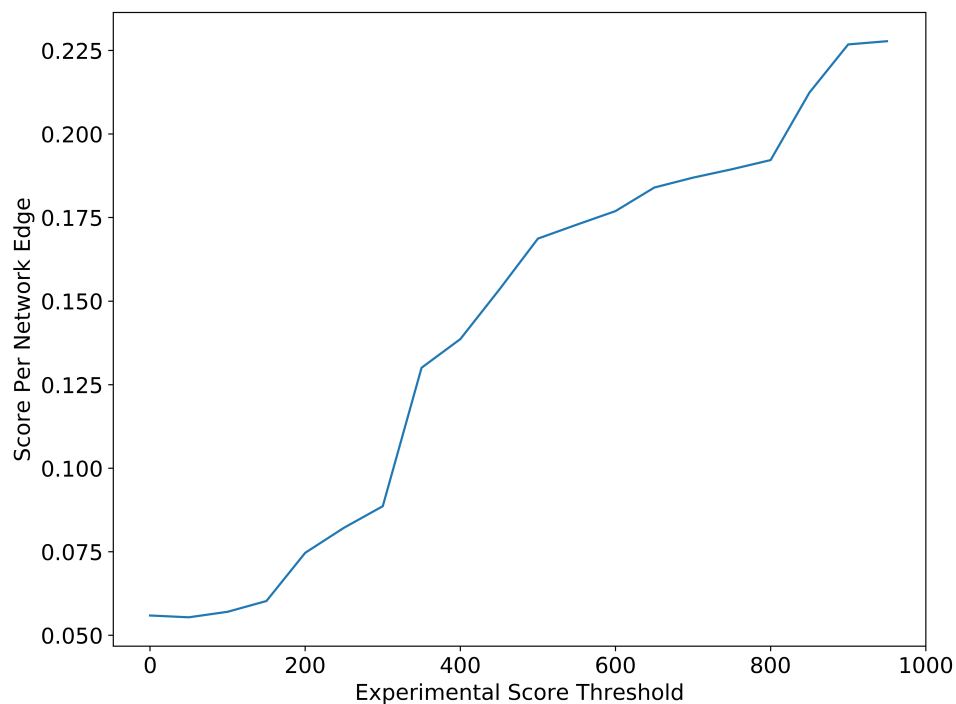


Figure 5: This figure shows how the threshold for the STRING PPI network was decided. The elbow visibly at the experimental score threshold of 800 shows that the edges over 800 are of higher quality, and should be included in the multiplex network.

Table 19: GO Enrichment for GRIN retained *S. cerevisiae* genes for the trait number of cankers.

GO Biological Process		<i>S. cerevisiae</i> Genes	<i>S. musiva</i> Retained Genes	Expected	Fold Enrich- ment	FDR P-Value
thiamine biosynthetic process (GO:0009229)	diphosphate	8	4	.08	49.59	2.16E-02
thiamine metabolic process (GO:0042357)	diphosphate	8	4	.08	49.59	1.08E-02
thiamine biosynthetic process (GO:0009228)		12	4	.12	33.06	1.54E-02
thiamine metabolic process (GO:0006772)		13	4	.13	30.52	1.66E-02
thiamine-containing compound biosynthetic process (GO:0042724)		13	4	.13	30.52	1.43E-02
primary biosynthetic process (GO:0034309)	alcohol	14	4	.14	28.34	1.59E-02
thiamine-containing compound metabolic process (GO:0042723)		15	4	.15	26.45	1.60E-02
alcohol biosynthetic process (GO:0046165)		70	7	.71	9.92	1.66E-02
organic hydroxy compound biosynthetic process (GO:1901617)		92	7	.93	7.55	2.36E-02
alcohol metabolic process (GO:0006066)		115	8	1.16	6.90	1.48E-02
small molecule biosynthetic process (GO:0044283)		312	12	3.15	3.81	2.86E-02
small molecule metabolic process (GO:0044281)		712	20	7.18	2.79	1.90E-02

Table 20: GO Enrichment for GRIN retained *S. cerevisiae* genes for the trait cankers per centimeter (table one of two).

GO Biological Process		<i>S. cerevisiae</i> Genes	<i>S. musiva</i> Retained Genes	Expected	Fold Enrich- ment	FDR P-Value
thiamine biosynthetic process (GO:0009229)	diphosphate	8	4	.04	> 100	7.68E-04
thiamine metabolic process (GO:0042357)	diphosphate	8	4	.04	> 100	3.84E-04
thiamine biosynthetic process (GO:0009228)		12	4	.05	74.69	9.30E-04
thiamine metabolic process (GO:0006772)		13	4	.06	68.95	9.09E-04
thiamine-containing compound biosynthetic process (GO:0042724)		13	4	.06	68.95	7.27E-04
primary alcohol biosynthetic process (GO:0034309)		14	4	.06	64.02	7.77E-04
cellular response to salt stress (GO:0071472)		11	3	.05	61.11	1.09E-02
thiamine-containing compound metabolic process (GO:0042723)		15	4	.07	59.75	8.41E-04
phospholipid translocation (GO:0045332)		15	3	.07	44.81	1.96E-02
lipid translocation (GO:0034204)		16	3	.07	42.01	2.07E-02
primary alcohol metabolic process (GO:0034308)		23	4	.10	38.97	3.25E-03
regulation of membrane lipid distribution (GO:0097035)		23	3	.10	29.23	3.76E-02
response to salt stress (GO:0009651)		25	3	.11	26.89	4.53E-02
pyrimidine-containing compound biosynthetic process (GO:0072528)		34	4	.15	26.36	9.62E-03
pyrimidine-containing compound metabolic process (GO:0072527)		42	4	.19	21.34	1.63E-02

Table 21: GO Enrichment for GRIN retained *S. cerevisiae* genes for the trait cankers per centimeter (table two of two).

GO Biological Process	<i>S. cerevisiae</i> Genes	<i>S. musiva</i> Retained Genes	Expected	Fold Enrich- ment	FDR P-Value
vitamin biosynthetic process (GO:0009110)	51	4	.23	17.57	2.44E-02
water-soluble vitamin biosynthetic process (GO:0042364)	51	4	.23	17.57	2.32E-02
lipid transport (GO:0006869)	69	5	.31	16.24	8.74E-03
alcohol biosynthetic process (GO:0046165)	70	5	.31	16.01	8.41E-03
water-soluble vitamin metabolic process (GO:0006767)	56	4	.25	16.01	3.11E-02
vitamin metabolic process (GO:0006766)	56	4	.25	16.01	2.97E-02
lipid localization (GO:0010876)	77	5	.34	14.55	1.08E-02
cellular response to environmental stimulus (GO:0104004)	62	4	.28	14.46	3.95E-02
cellular response to abiotic stimulus (GO:0071214)	62	4	.28	14.46	3.80E-02
organic hydroxy compound biosynthetic process (GO:1901617)	92	5	.41	12.18	1.96E-02
alcohol metabolic process (GO:0006066)	115	5	.51	9.74	3.52E-02
phosphorus metabolic process (GO:0006793)	661	11	2.95	3.73	2.08E-02

Table 22: GO Enrichment for GRIN retained *S. cerevisiae* genes for the trait disease severity.

GO Biological Process	<i>S. cerevisiae</i> Genes	<i>S. musiva</i> Retained Genes	Expected	Fold Enrich- ment	FDR P-Value
phytochelatin biosynthetic process (GO:0046938)	2	2	.00	> 100	7.55E-02
phytochelatin metabolic process (GO:0046937)	2	2	.00	> 100	3.77E-02
secondary metabolite biosynthetic process (GO:0044550)	2	2	.00	> 100	2.52E-02
zymogen activation (GO:0031638)	2	2	.00	> 100	1.89E-02
secondary metabolic process (GO:0019748)	6	2	.01	> 100	3.51E-02
mitochondrial electron transport, succinate to ubiquinone (GO:0006121)	7	2	.01	> 100	3.76E-02
mitochondrial ATP synthesis coupled electron transport (GO:0042775)	39	3	.06	46.54	3.69E-02
ATP synthesis coupled electron transport (GO:0042773)	39	3	.06	46.54	3.08E-02
aerobic electron transport chain (GO:0019646)	39	3	.06	46.54	2.64E-02
respiratory electron transport chain (GO:0022904)	43	3	.07	42.21	3.04E-02
oxidative phosphorylation (GO:0006119)	45	3	.07	40.33	3.08E-02
electron transport chain (GO:0022900)	51	3	.08	35.59	3.59E-02

Table 23: GO Enrichment for GRIN retained *S. cerevisiae* genes for the trait plant height (table one of two).

GO Biological Process	<i>S. cerevisiae</i> Genes	<i>S. musiva</i> Retained Genes	Expected	Fold Enrich- ment	FDR P-Value
maltose transport (GO:0015768)	3	3	.01	> 100	1.97E-04
mRNA localization result- ing in posttranscriptional regulation of gene expres- sion (GO:0010609)	2	2	.00	> 100	8.16E-03
oligosaccharide transport (GO:0015772)	4	3	.01	> 100	2.76E-04
disaccharide transport (GO:0015766)	4	3	.01	> 100	2.30E-04
negative regulation of Rho protein signal transduction (GO:0035024)	5	2	.01	> 100	1.90E-02
maltose metabolic process (GO:0000023)	14	4	.03	> 100	6.56E-05
regulation of Rho protein signal transduction (GO:0035023)	9	2	.02	> 100	4.24E-02
carbohydrate transmembrane transport (GO:0034219)	31	5	.07	75.06	3.54E-05
glucosamine-containing compound biosynthetic process (GO:1901073)	19	3	.04	73.48	5.42E-03
chitin biosynthetic process (GO:0006031)	19	3	.04	73.48	4.96E-03
aminoglycan biosynthetic process (GO:0006023)	19	3	.04	73.48	4.58E-03
disaccharide metabolic pro- cess (GO:0005984)	27	4	.06	68.95	2.85E-04
amino sugar biosynthetic process (GO:0046349)	21	3	.05	66.48	5.58E-03
oligosaccharide metabolic process (GO:0009311)	30	4	.06	62.05	3.66E-04
carbohydrate transport (GO:0008643)	43	5	.09	54.11	7.93E-05
chitin metabolic process (GO:0006030)	27	3	.06	51.71	8.64E-03

Table 24: GO Enrichment for GRIN retained *S. cerevisiae* genes for the trait plant height (table two of two).

GO Biological Process	<i>S. cerevisiae</i> Genes	<i>S. musiva</i> Retained Genes	Expected	Fold Enrich- ment	FDR P-Value
aminoglycan metabolic process (GO:0006022)	27	3	.06	51.71	8.19E-03
glucosamine-containing compound metabolic process (GO:1901071)	28	3	.06	49.86	8.60E-03
amino sugar metabolic process (GO:0006040)	30	3	.06	46.54	9.92E-03
proton transmembrane transport (GO:1902600)	99	5	.21	23.50	8.89E-04
inorganic cation transmembrane transport (GO:0098662)	164	5	.35	4.19	5.84E-03
cellular carbohydrate metabolic process (GO:0044262)	139	4	.30	13.39	3.56E-02
inorganic ion transmembrane transport (GO:0098660)	180	5	.39	12.93	7.99E-03
cation transmembrane transport (GO:0098655)	207	5	.44	11.24	1.14E-02
ion transmembrane transport (GO:0034220)	260	6	.56	10.74	4.59E-03
cation transport (GO:0006812)	246	5	.53	9.46	2.35E-02
ion transport (GO:0006811)	344	6	.74	8.12	1.01E-02
transmembrane transport (GO:0055085)	459	6	.99	6.08	4.04E-02

Table 25: GO Enrichment for GRIN retained *S. cerevisiae* genes for all four traits (table one of two).

GO Biological Process		<i>S.</i> <i>cerevisiae</i> Genes	<i>S. musiva</i> Retained Genes	Expected	Fold Enrich- ment	FDR P-Value
maltose	transport	3	3	.02	> 100	5.09E-03
(GO:0015768)						
oligosaccharide	transport	4	3	.03	98.64	5.06E-03
(GO:0015772)						
disaccharide	transport	4	3	.03	98.64	4.72E-03
(GO:0015766)						
aminophospholipid	translo-	5	3	.04	78.91	7.05E-03
cation (GO:0140331)						
sterol import	(GO:0035376)	6	3	.05	65.76	9.35E-03
thiamine	diphosphate	8	4	.06	65.76	3.46E-03
biosynthetic	process					
(GO:0009229)						
thiamine	diphosphate	8	4	.06	65.76	2.30E-03
metabolic	process					
(GO:0042357)						
aminophospholipid		7	3	.05	56.37	1.14E-02
transport (GO:0015917)						
thiamine biosynthetic pro-		12	4	.09	43.84	4.97E-03
cess (GO:0009228)						
thiamine metabolic process		13	4	.10	40.47	5.39E-03
(GO:0006772)						
thiamine-containing		13	4	.10	40.47	4.62E-03
compound	biosynthetic					
process (GO:0042724)						
maltose metabolic process		14	4	.11	37.58	4.59E-03
(GO:0000023)						
primary	alcohol	14	4	.11	37.58	4.13E-03
biosynthetic	process					
(GO:0034309)						
thiamine-containing		15	4	.11	35.07	4.73E-03
compound	metabolic					
process (GO:0042723)						
phospholipid translocation		15	4	.11	35.07	4.34E-03
(GO:0045332)						

Table 26: GO Enrichment for GRIN retained *S. cerevisiae* genes for all four traits (table two of two).

GO Biological Process	<i>S. cerevisiae</i> Genes	<i>S. musiva</i> Retained Genes	Expected	Fold Enrich- ment	FDR P-Value
lipid translocation (GO:0034204)	16	4	.12	32.88	4.98E-03
regulation of membrane lipid distribution (GO:0097035)	23	4	.17	22.87	1.19E-02
primary alcohol metabolic process (GO:0034308)	23	4	.17	22.87	1.13E-02
phospholipid transport (GO:0015914)	27	4	.21	19.48	1.80E-02
disaccharide metabolic pro- cess (GO:0005984)	27	4	.21	19.48	1.72E-02
oligosaccharide metabolic process (GO:0009311)	30	4	.23	17.54	2.39E-02
carbohydrate transmembrane transport (GO:0034219)	31	4	.24	16.97	2.58E-02
pyrimidine-containing com- pound biosynthetic process (GO:0072528)	34	4	.26	15.47	3.43E-02
lipid transport (GO:0006869)	69	7	.52	13.34	6.60E-03
lipid localization (GO:0010876)	77	7	.59	11.96	3.27E-03
sulfur compound biosynthetic process (GO:0044272)	76	6	.58	10.38	9.08E-03

Vita

Angelica M. Walker was born to loving parents Marilyn and Jason Walker in Orlando, Florida in 1996. She attended The Johns Hopkins University in Baltimore, Maryland as a first generation college student from 2015 to 2019 to obtain a Bachelors of Science in Computer Science and Molecular and Cellular Biology. In 2019 she moved to Knoxville, Tennessee to attend the University of Tennessee, Knoxville and pursue a Doctorate of Philosophy in Data Science and Engineering with a focus in Health and Biological Systems through the Bredezen Center of Interdisciplinary Research and Graduate Education. She completed her Ph.D. in the Biosciences Division at Oak Ridge National Laboratory under the advising of Dr. Daniel Jacobson.