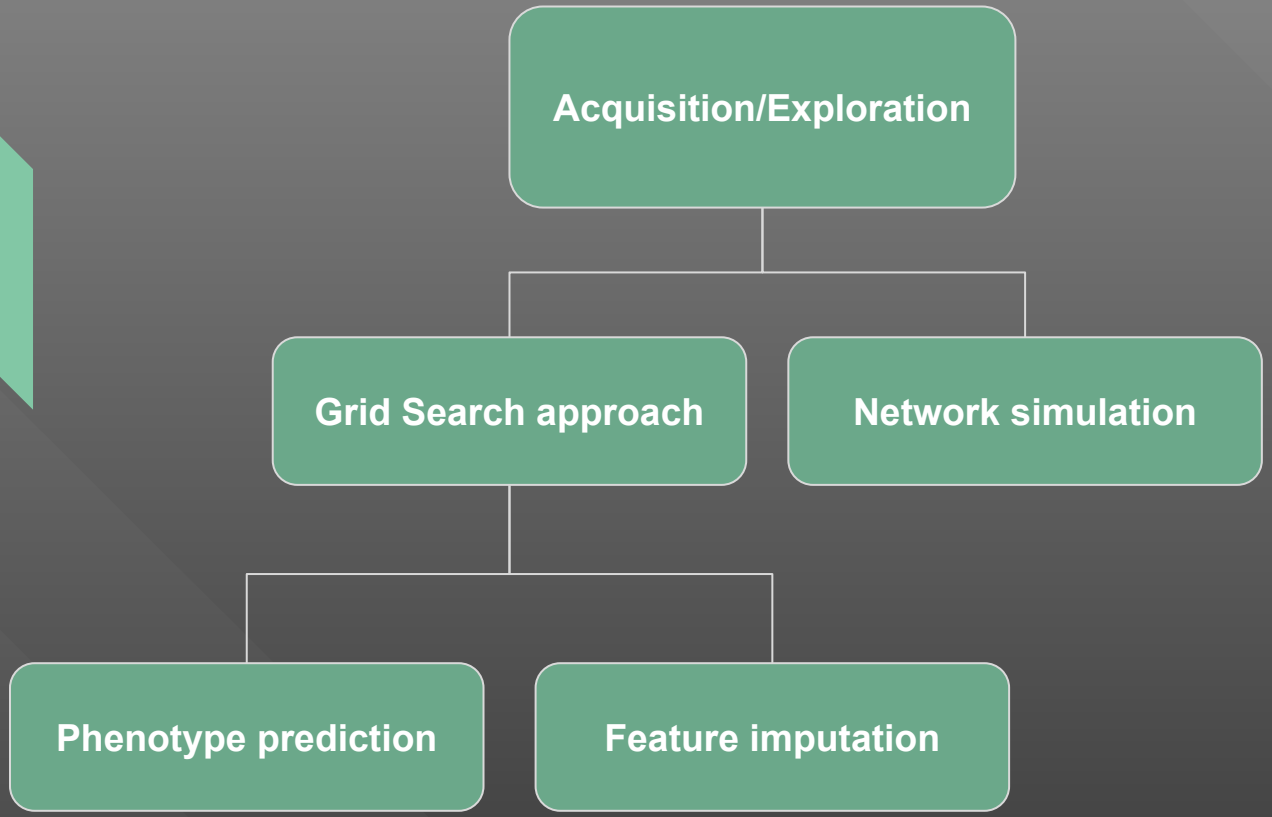
Network tour of mice genetics.

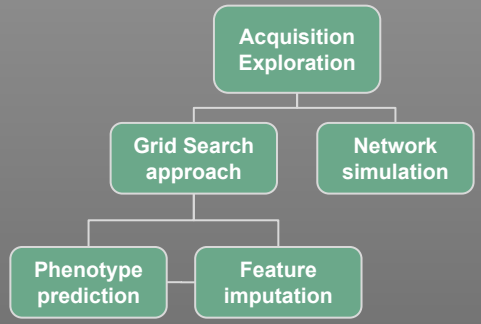
An Approach to Genetic and Expression based Phenotype Prediction on Mouse

Gianni Giusto | Yann Mentha | Raphaël Reis | Lucas Zweili

Pipeline


Acquisition/Exploration
Grid Search approach
Network simulation
Phenotype prediction
Feature imputation

Introduction
Background Informations



Gigi - Intro (~30sec-1 minute)

Gianni - Acquisition / Exploration (2 minutes)

Yann - Grid Search

Raph - Network simulation

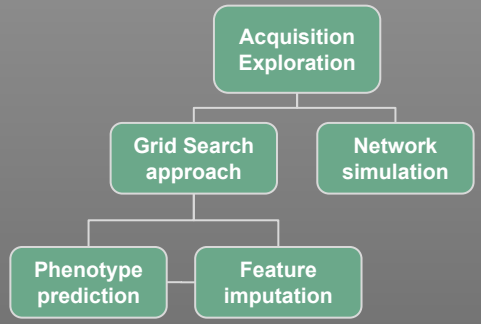
Yannos - Results from Yann - Exploitation Phenotype prediction (Harmonic, LReg, tikho Class)

Raph - Feature Imputation (Baseline & Tikhonov)

Lucas - Features Imputation (Logistic Regression and GNN) / Results

X - Conclusion / Discussion

Introduction

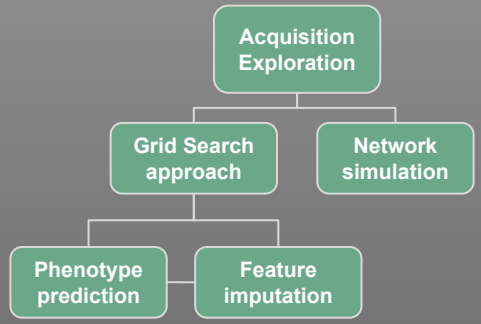


Background and motivation

- Systems genetics approaches such as GWAS or PheWAS are suitable tool to examine different phenotypes
- Graph methods are gaining popularity due to their flexibility and high representative power

-

Introduction



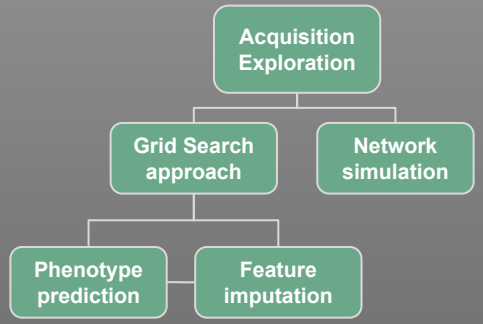
Background and motivation

- Systems genetics approaches such as GWAS or PheWAS are suitable tool to examine different phenotypes
- Graph methods are gaining popularity due to their flexibility and high representative power

Goal

- Take advantage of the network structure of the data to impute missing values
- Build a classifier able to predict the *phenotype* based on the underlying *genotype*

Introduction



Background and motivation

- Systems genetics approaches such as GWAS or PheWAS are suitable tool to examine different phenotypes
- Graph methods are gaining popularity due to their flexibility and high representative power

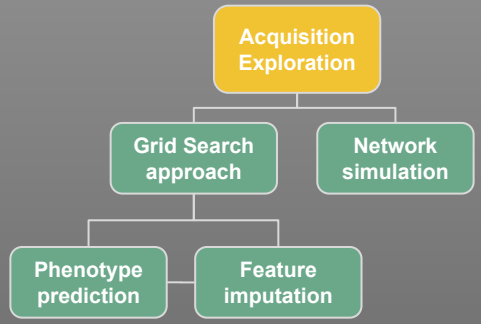
Goal

- Take advantage of the network structure of the data to impute missing values
- Build a classifier able to predict the *phenotype* based on the underlying *genotype*

Challenges

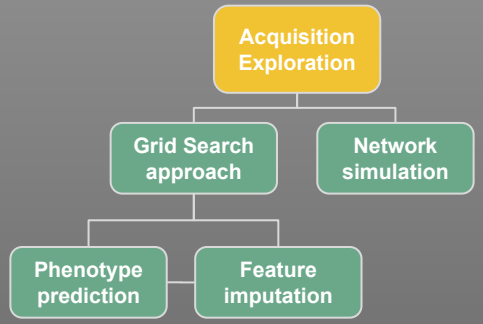
- Deal with the large amount of missing values (~50%)

Acquisition



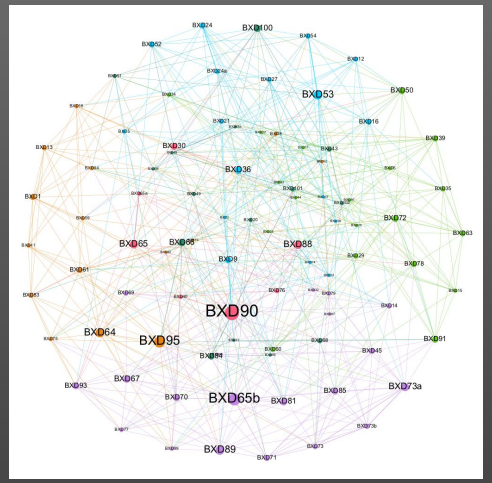
- 3 types of file
  - *genetic* file
  - *phenotype* file
  - *expression* file (about 71'000 genes)
- Network build based on the genetic file (no missing information)
  - Nodes represent mouse strains
  - Edges represent 'similarities' between mouse strains
- Lots of missing values in expression files

Exploration



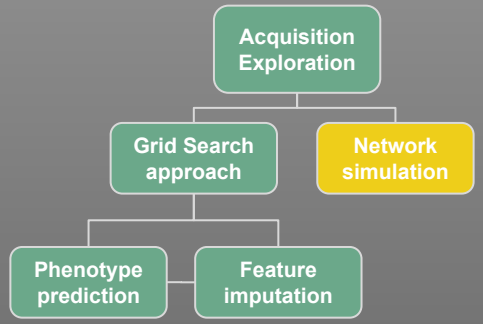
- *Cosine* similarity measurement
- Gaussian kernel (σ=0.53, ε=0.27)
- Property: connected network

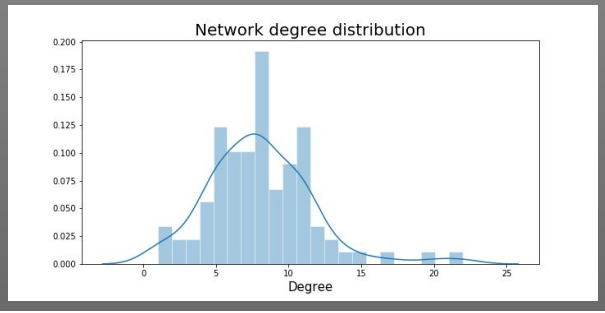
Genetic Graph	
Number of nodes	93
Number of edges	373
Graph density	8.72%
Average degree	8.02
Nb of connected components	1
Diameter of the network	6
Avg clustering coefficient	0.26

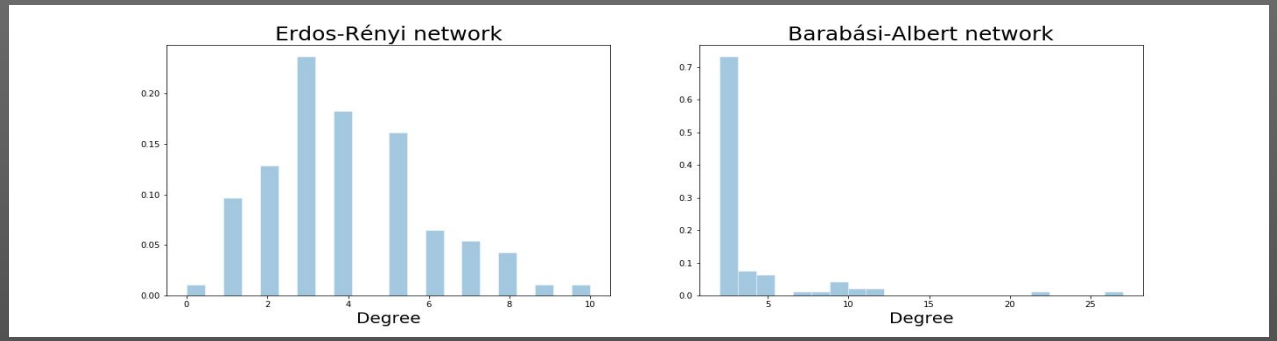


- Modularity maximization
- Low avg. clustering
- High diameter
- No small world property

Network Simulation



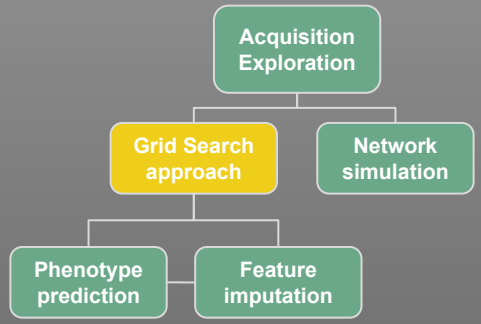




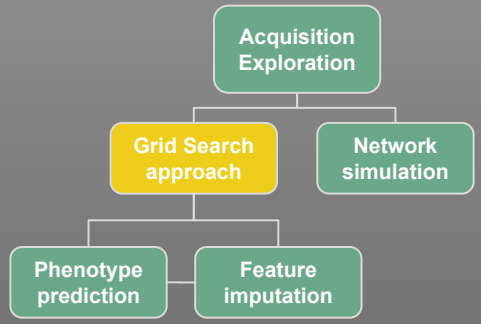
- SNP* random inheritance suggests random network structure
- Degree distribution closer to random than scale-free network

SNP*: Single nucleotide polymorphisme

Grid Search Approach

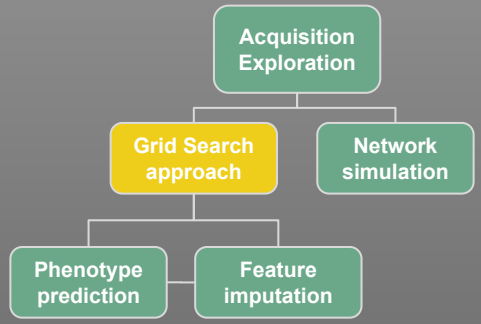


Grid Search Approach



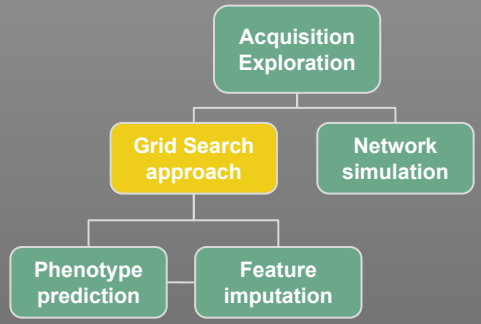
- Many parameters (sigma, epsilon, distance metric,tau etc)

Grid Search Approach



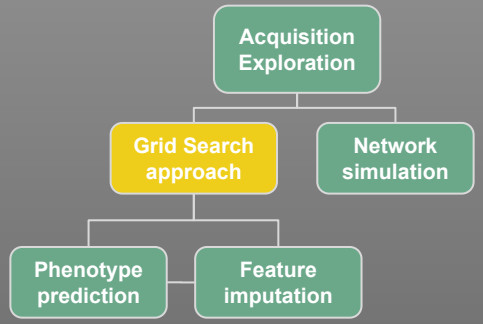
- Many parameters (sigma, epsilon, distance metric,tau etc)
- Sparse dataset (extensive amount of nan values)

Grid Search Approach



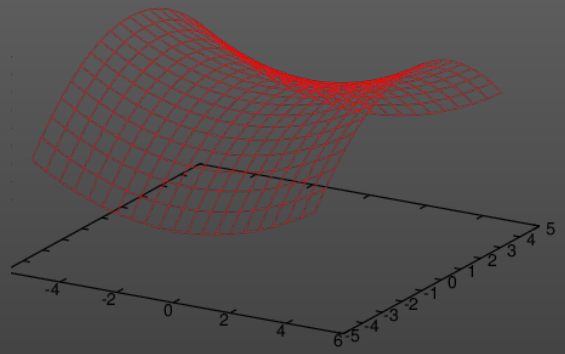
- Many parameters (sigma, epsilon, distance metric,tau etc)
- Sparse dataset (extensive amount of nan values)
- Choice between genetic and expressive datasets to build a network

Grid Search Approach

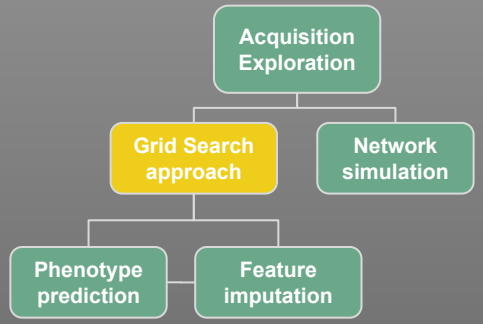


- Many parameters (sigma, epsilon, distance metric,tau etc)
- Sparse dataset (extensive amount of nan values)
- Choice between genetic and expressive datasets to build a network

Solution: GridSearch approach!



Grid Search Approach

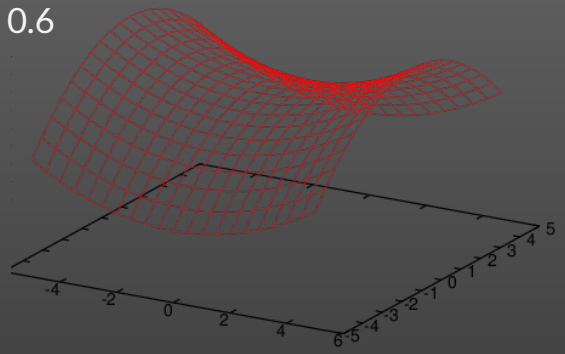


- Many parameters (sigma, epsilon, distance metric,tau etc)
- Sparse dataset (extensive amount of nan values)
- Choice between genetic and expressive datasets to build a network

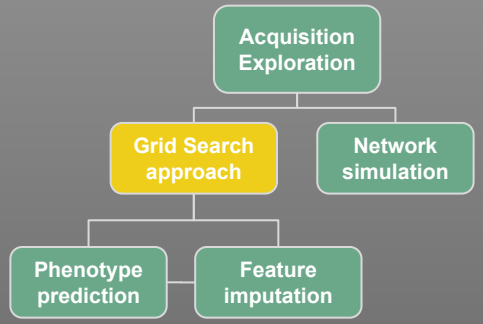
Solution: GridSearch approach!

- Implementation of a CV method on graphs dealing with nan values

Result: list of 12 continuous and discrete genes reaching an accuracy > 0.6



Grid Search Approach

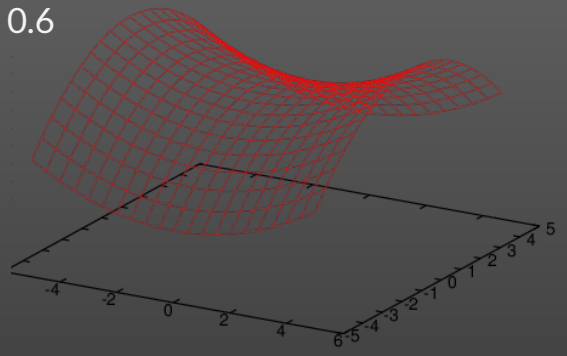


- Many parameters (sigma, epsilon, distance metric,tau etc)
- Sparse dataset (extensive amount of nan values)
- Choice between genetic and expressive datasets to build a network

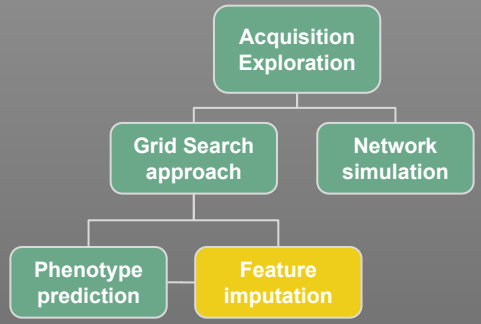
Solution: GridSearch approach!

- Implementation of a CV method on graphs dealing with nan values

Result: list of 12 continuous and discrete genes reaching an accuracy > 0.6
R2 > 0.1



Exploitation - Feature Imputation

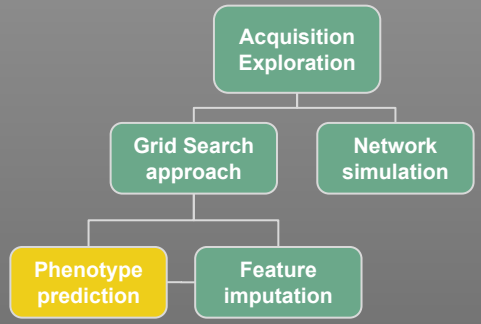


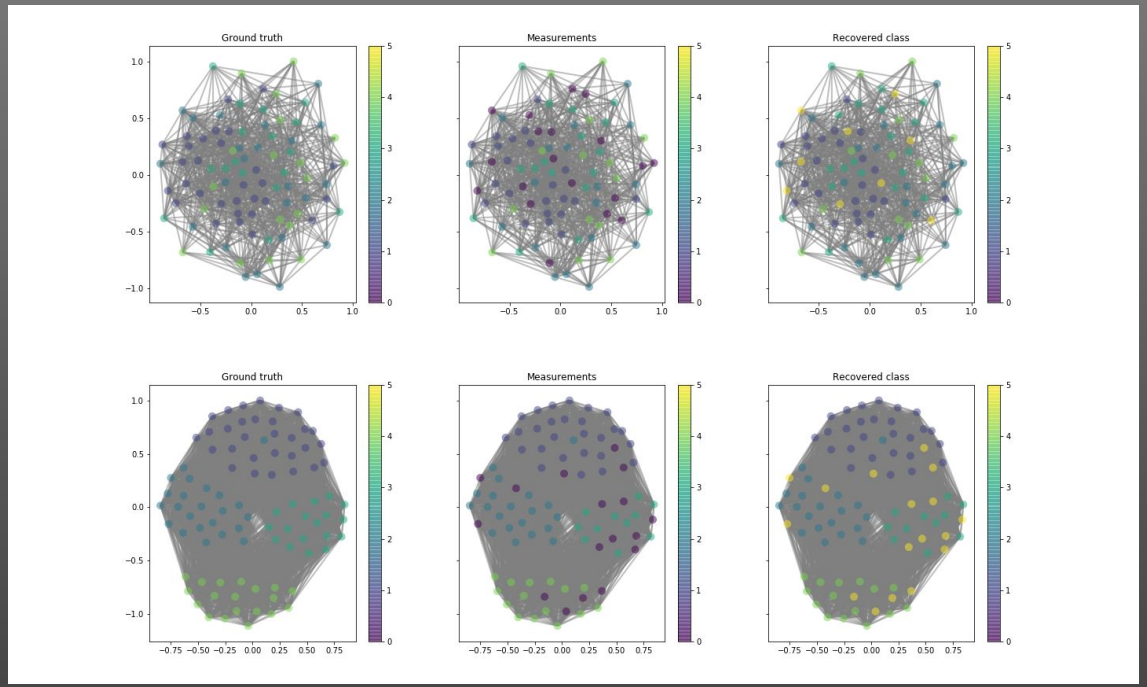
1. Build a baseline : Mean imputation
   a. Pros:
      i. Does not change the sample mean of the feature
      ii. Make sense for univariate analysis
   b. Cons:
      i. Attenuate correlation between imputed variables
      ii. Becomes problematic for multivariate analysis
2. Tikhonov regression imputation
   a. Pros:
      i. Use same feature from other mouse to impute data (unlike standard regression)
      ii. Use graph similarity to impute data
   b. Cons:
      i. No error term include in the estimation (fit exactly the model without residual variance)

3. Missing data imputation by signal filtering (no relevant results)
4. Linear regression (to do)
5. Stochastic regression (to do)

Exploitation - Phenotype Prediction

- Harmonic function[1]
- Logistic regression and RFE
- Tikhonov Classification





[1] Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. "Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions". en. In: (), p. 8.

Results

Model	Graph Based	Features	FS method	Imputation Method	Acc.	MCC
Logistic Regression	no	Genetic (7324 / 2)	RFE	None	0.60 / **0.98**	0.47 / **0.99**
Tikhonov Classification	Genetic (7324 / 2)	no	RFE	None	0.52 / **0.98**	0.40 / 0.98
Harmonic function	Genetic (7324)	no	None	None	0.56	0.39

Genetic data only

Results

Model	Graph Based	Features	FS method	Imputation Method	Acc.	MCC	
Logistic Regression	no	Genetic (7324 / 2)	RFE	None	0.60 / **0.98**	0.47 / **0.99**	Genetic data only
Tikhonov Classification	Genetic (7324 / 2)	no	RFE	None	0.52 / **0.98**	0.40 / 0.98	
Harmonic function	Genetic (7324)	no	None	None	0.56	0.39	
Logistic Regression	no	Express. (30)	MI / CHI2 / RF	Mean	**0.78**	**0.704**	Expression data only
Logistic Regression	no	Express. (30)	MI / CHI2 / RF	Tikhonov	0.74	0.632	

Results

Model	Graph Based	Features	FS method	Imputation Method	Acc.	MCC	
Logistic Regression	no	Genetic (7324 / 2)	RFE	None	0.60 / **0.98**	0.47 / **0.99**	Genetic data only
Tikhonov Classification	Genetic (7324 / 2)	no	RFE	None	0.52 / **0.98**	0.40 / 0.98	
Harmonic function	Genetic (7324)	no	None	None	0.56	0.39	
Logistic Regression	no	Express. (30)	MI / CHI2 / RF	Mean	**0.78**	**0.704**	Expression data only
Logistic Regression	no	Express. (30)	MI / CHI2 / RF	Tikhonov	0.74	0.632	
Graph Neural Network	Genetic (7324)	Express. (30)	MI / CHI2 / RF	Mean	0.609	0.418	Genetic data + Expression data
Graph Neural Network	Genetic (7324)	Express. (30)	MI / CHI2 / RF	Tikhonov	**0.687**	**0.537**	

Discussion & Conclusion

Improvement & Futur work:

Discussion & Conclusion

Improvement & Futur work:

- Acquisition of more data

Discussion & Conclusion

Improvement & Futur work:

- Acquisition of more data
- Regression for a continuous phenotype label instead of classification on discrete label

Discussion & Conclusion

Improvement & Futur work:

- Acquisition of more data
- Regression for a continuous phenotype label instead of classification on discrete label
- Skin color for instance is regulated by >11 loci in the human genome

THANK YOU !

Gianni Giusto | Yann Mentha | Raphaël Reis | Lucas Zweili

Questions ?

Gianni Giusto | Yann Mentha | Raphaël Reis | Lucas Zweili