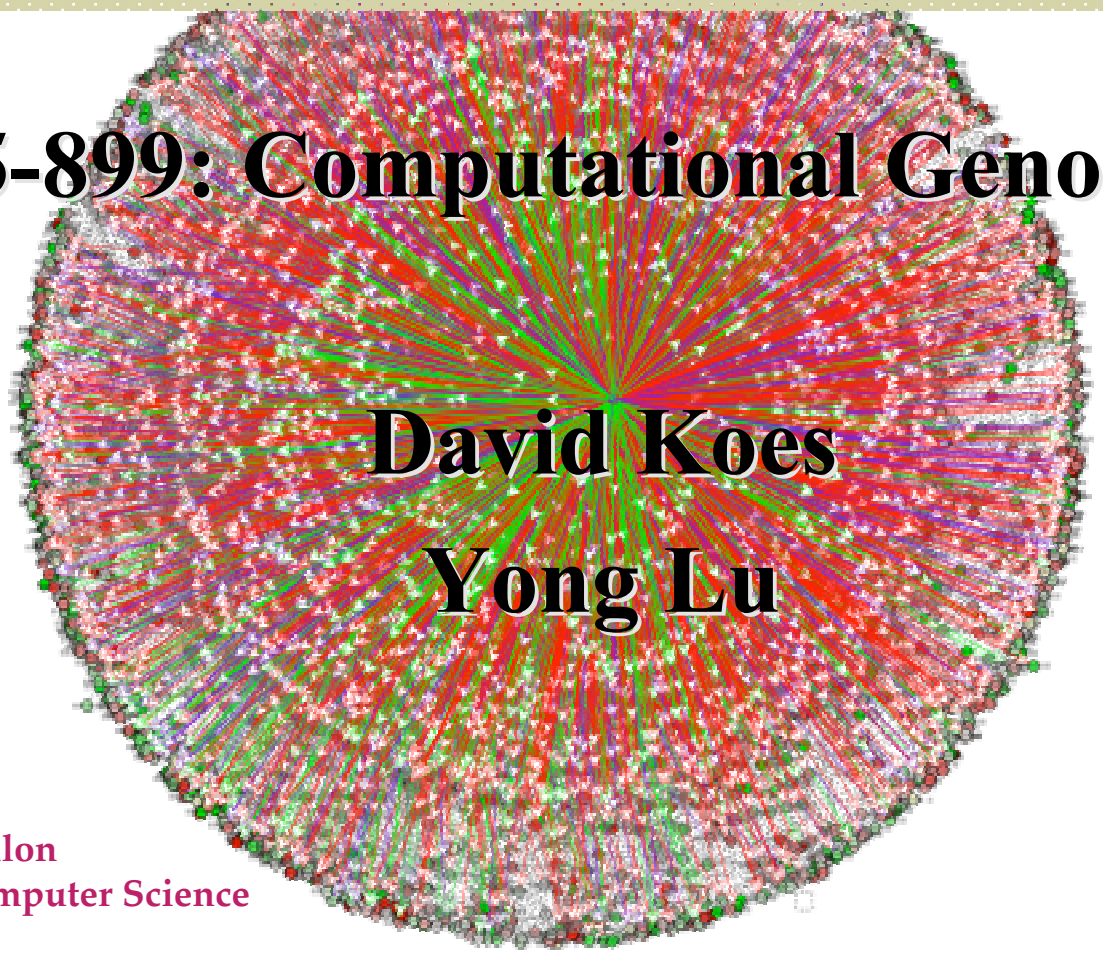
Identifying and Understanding Differential Transcriptor Binding



15-899: Computational Genomics

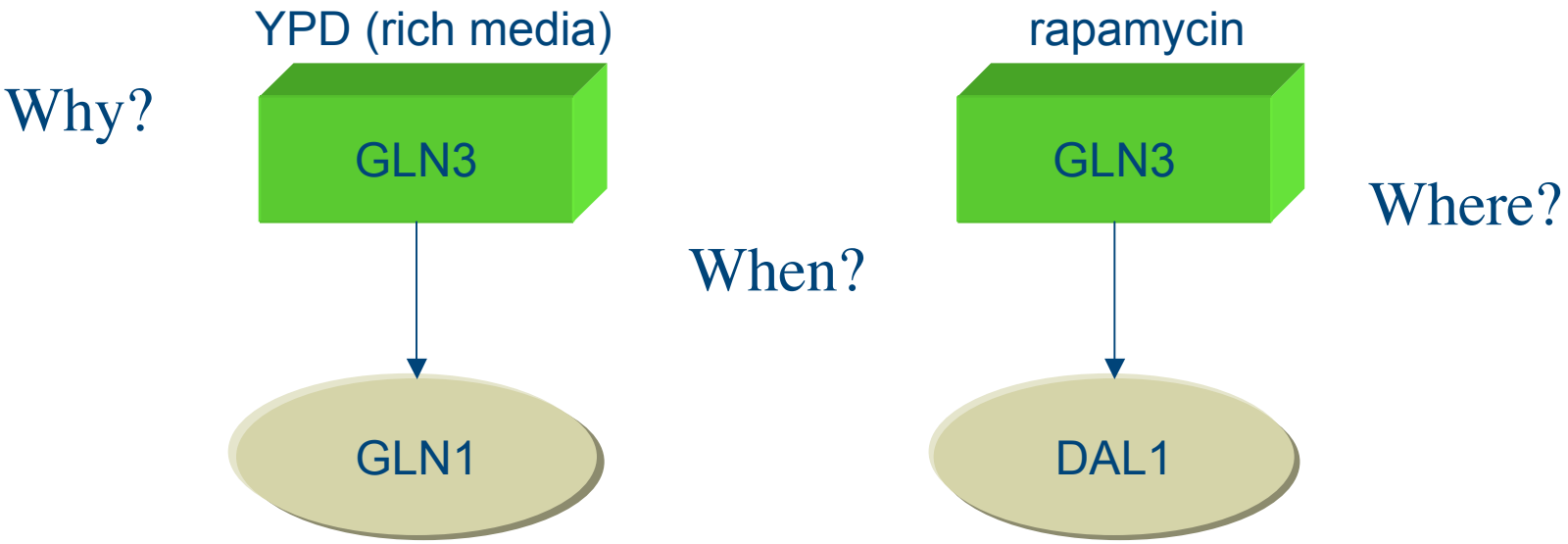
David Koes

Yong Lu

Carnegie Mellon
School of Computer Science

Motivation

- Under different conditions, a transcription factor binds to different genes



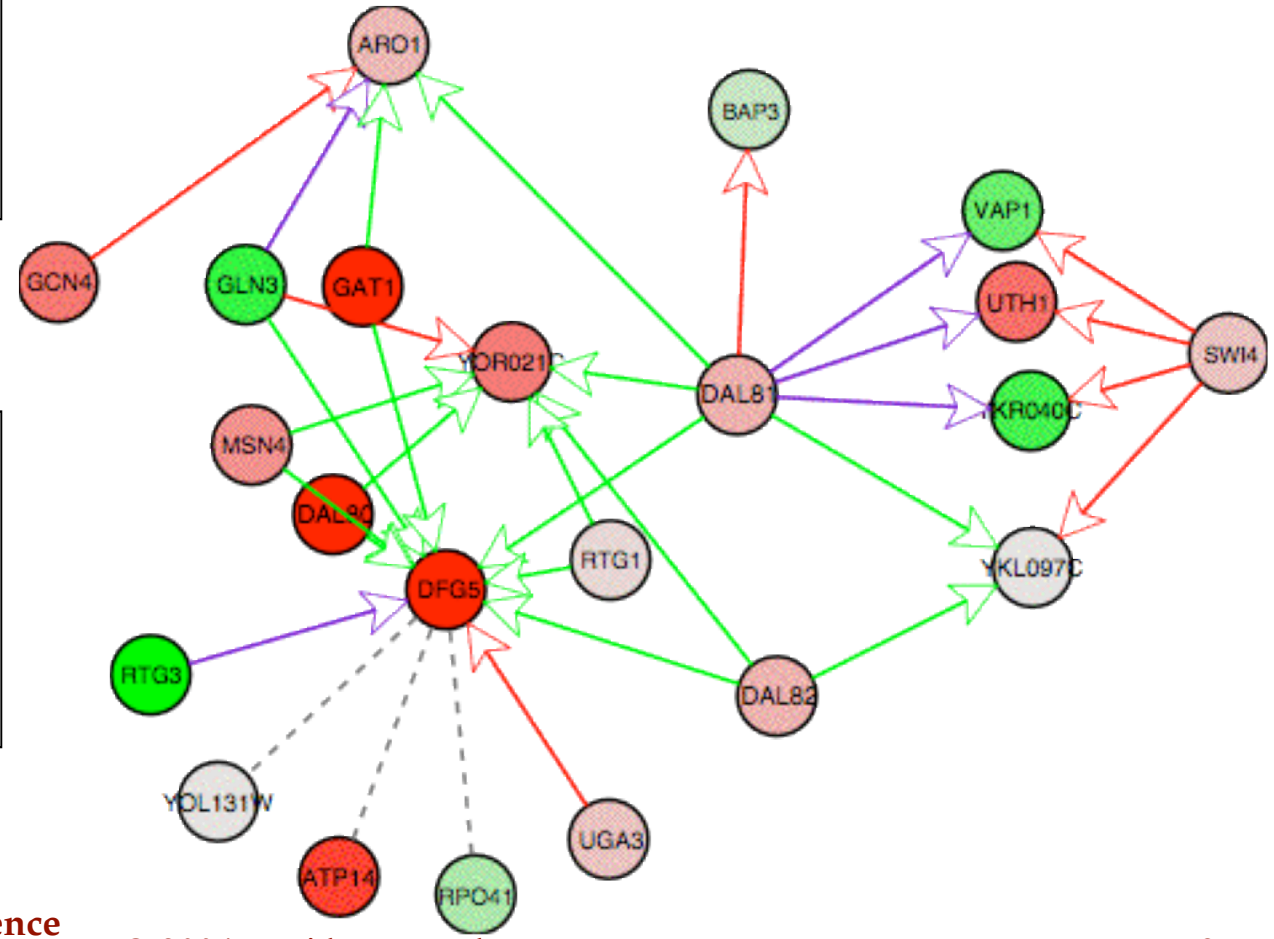
Example: Difference Graph

Nodes

- higher ypd expr
- higher rap expr

Edges

- rap binding only
- ypd binding only
- ypd & rap binding
- protein interaction



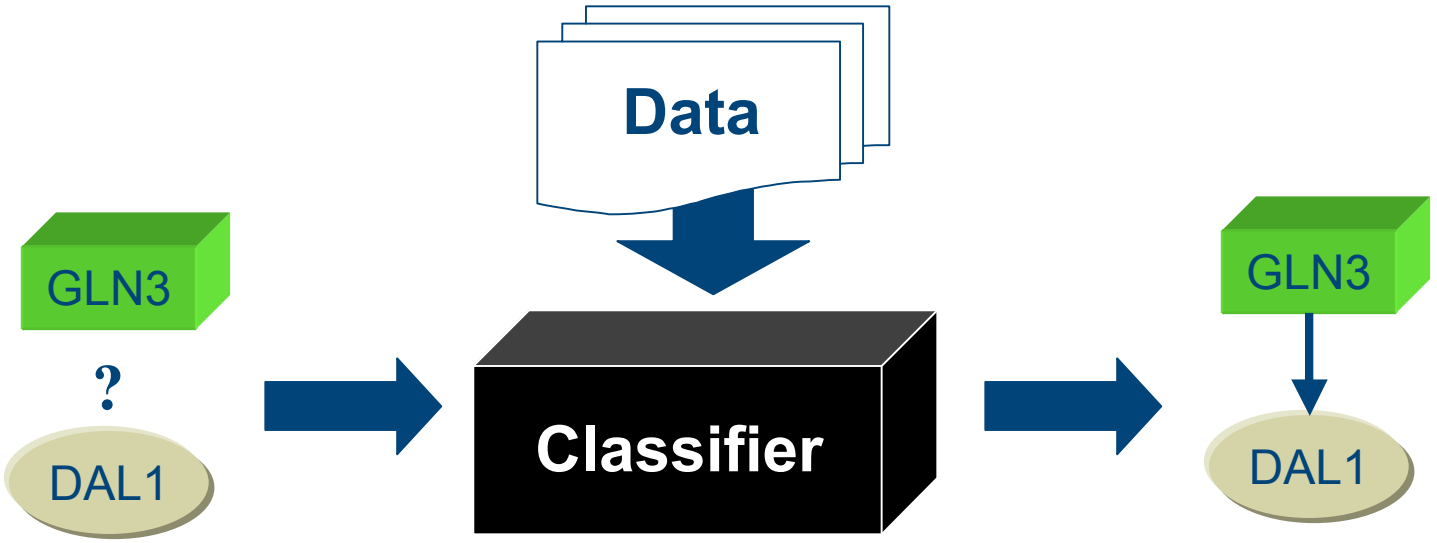
Carnegie Mellon
School of Computer Science

Goals

- Can we fill in the missing values of a binding experiment?
- Can we predict all the values of a binding experiment?
- Can we explain the differences in binding?

Approach: Classification

- Given a transcription factor/gene pair
  - will there be binding under rapamycin?



Outline

- Data Sources
- Feature Selection
- Classifiers
- Results

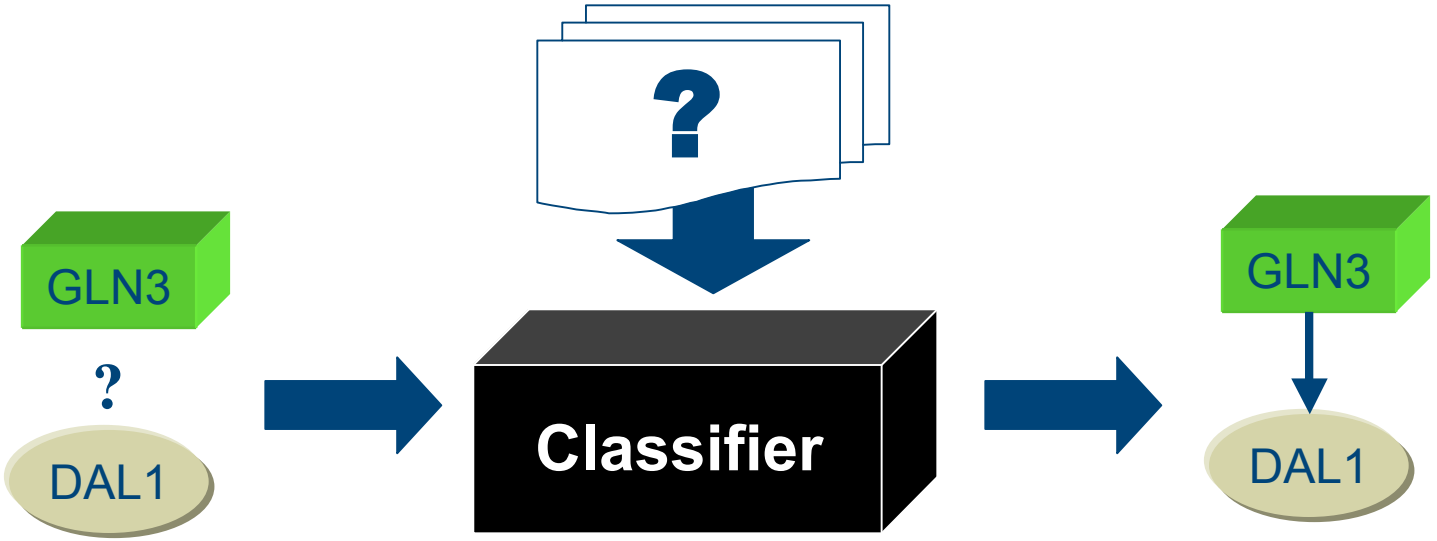
Data: Saccharomyces cerevisiae

- Expression
  - ypd and rapamycin micro-array data
    - http://www-schreiber.chem.harvard.edu/home/protocols/partitioning/
- Binding
  - genome wide location analysis
    - YPD: http://web.wi.mit.edu/young/regulator_network/
    - Rapamycin: http://www.psrg.lcs.mit.edu/Networks/modules.html
- Protein Interaction
  - two-hybrid method
    - http://genome.c.kanazawa-u.ac.jp/Y2H/

Outline

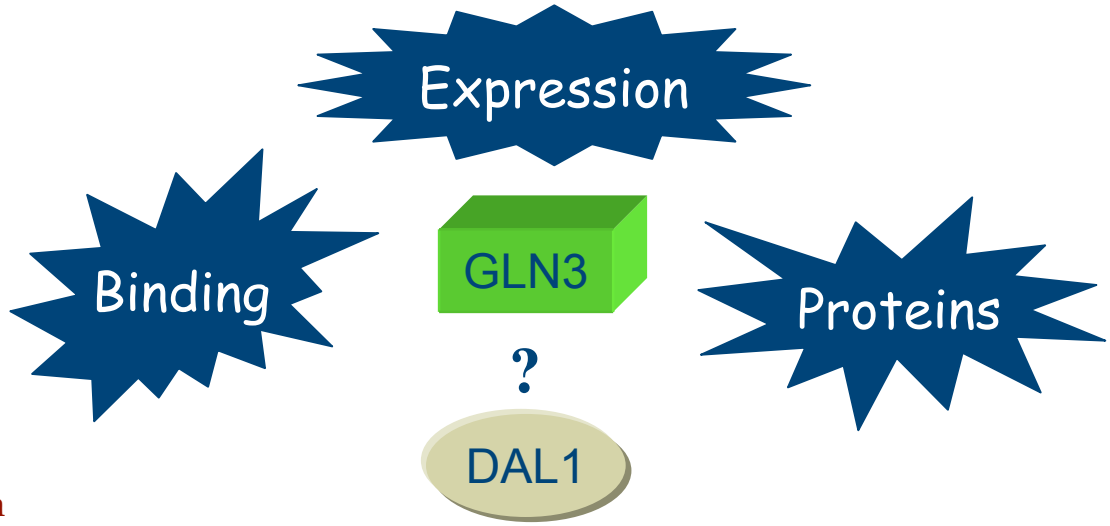
- Data Sources
- Feature Selection
  - sparse and precise
  - dense and aggregate
- Classifiers
- Results

Features



Features

- Expression, binding, protein data not features
  - global values not dependant upon a given edge
- Must exploit topology of data networks



Outline

- Data Sources
- Feature Selection
  - sparse and precise
  - dense and aggregate
- Classifiers
- Results

Sparse and Precise

- Several attributes for *every* gene
  - binding pvalue for gene with factor/target
  - expression of gene if gene *can* bind factor/target
  - expression of gene if factor/target *can* bind gene
  - expression of gene if protein interaction with factor/target exists
  - expression of gene if gene is factor/target

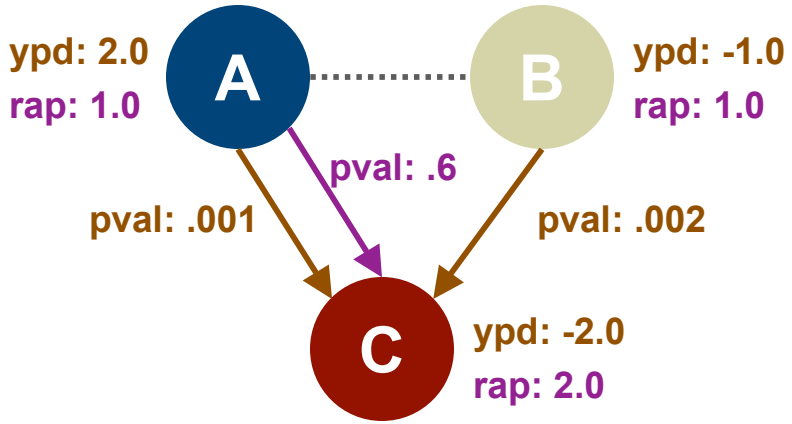
Example

Nonzero attributes (16)

factor_binds_target_rap	**.6**
factor_binds_target_ypd	.001
factor_binds_C_ypd	.001
B_binds_target_ypd	.002
factor_self_A_ypd	2.0
factor_ldown_C_ypd	-2.0
factor_lpp_B_ypd	-1.0
factor_self_A_rap	1.0
factor_ldown_C_rap	2.0
factor_lpp_B_rap	1.0
target_self_C_ypd	-2.0
target_lup_A_ypd	2.0
target_lup_B_ypd	-1.0
target_self_C_rap	2.0
target_lup_A_rap	1.0
target_lup_B_rap	1.0

Zero attributes (46)

factor_binds_A_ypd
factor_binds_B_ypd
target_binds_A_ypd
target_binds_B_ypd
<etc.>



Pros and Cons

- Pros
  - Precisely captures all the data
  - Sparse dataset results in compact representation
    - Solvers can take advantage of sparseness
- Cons
  - Susceptible to over-fitting
  - Huge number of attributes
  - Solvers require binary attributes

Outline

- Data Sources
- Feature Selection
  - sparse and precise
  - dense and aggregate
- Classifiers
- Results

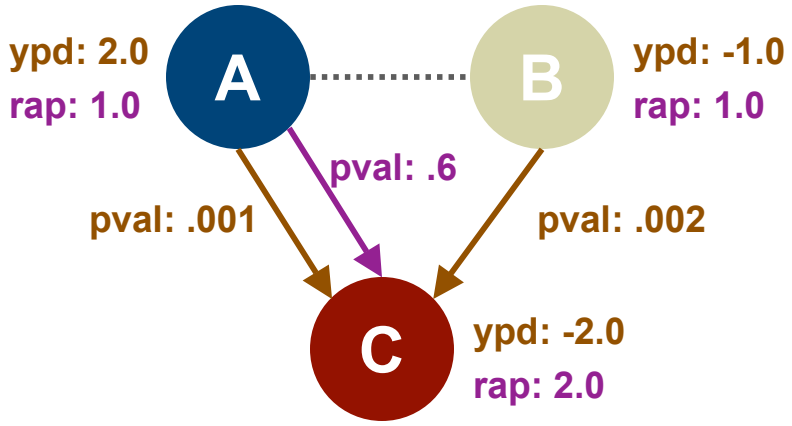
Dense and Aggregate

- Use averages of data based on topological relationship in network
  - genes that can bind factor/target
  - genes that factor/target can bind
  - genes with protein interactions with factor/target
- YPD binding data

Example

Nonzero attributes (12)	
rap_bind	**.6**
ypd_bind	0.001
factor_expr_ypd	2.0
factor_expr_rap	1.0
target_expr_ypd	-2.0
target_expr_rap	2.0
target_ave_expr_up_YPD	**0.5**
target_ave_expr_up_RAP	**1.0**
factor_ave_expr_down_YPD	-2.0
factor_ave_expr_down_RAP	2.0
factor_ave_expr_pp_YPD	-1.0
factor_ave_expr_pp_RAP	1.0

Zero attributes (6)	
target_ave_expr_down_YPD	0
target_ave_expr_down_RAP	0
target_ave_expr_pp_YPD	0
target_ave_expr_pp_RAP	0
factor_ave_expr_up_YPD	0
factor_ave_expr_up_RAP	0



Pros and Cons

- Pros
  - Small, constant, number of attributes
  - Low penalty for adding additional attributes
- Cons
  - Information lost

Outline

- Data Sources
- Feature Selection
- Classifiers
  - Logistic Regression
  - K Nearest Neighbor
  - Naïve Bayes
  - Learned Bayes Net
- Results

Logistic Regression

- Find β such that μ best approximate the training data outputs **y** where

$$\mu_i = \frac{e^{(\beta \bullet \mathbf{x}_i)}}{1 + e^{(\beta \bullet \mathbf{x}_i)}}$$

- Solved with iterative re-weighted least squares
  - Newton-Raphson

K Nearest Neighbors

- Classify a point based on value of training points close by in attribute space

Naïve Bayes

- Makes simplifying assumption that attributes are conditional independent given class
- Uses training data to estimate conditional probabilities
- Classifies based on what class assignment maximizes joint probability

Learned Bayes Net

- Use training data to find a “good” network of conditional dependencies

Outline

- Data Sources
- Feature Selection
- Classifiers
- Results

Tools

- Auton Fast Classifiers
  - http://www.autonlab.org/
- Bayes Net Inference
  - BNT/Matlab
    - http://www.ai.mit.edu/~murphyk/Software/BNT/bnt.html

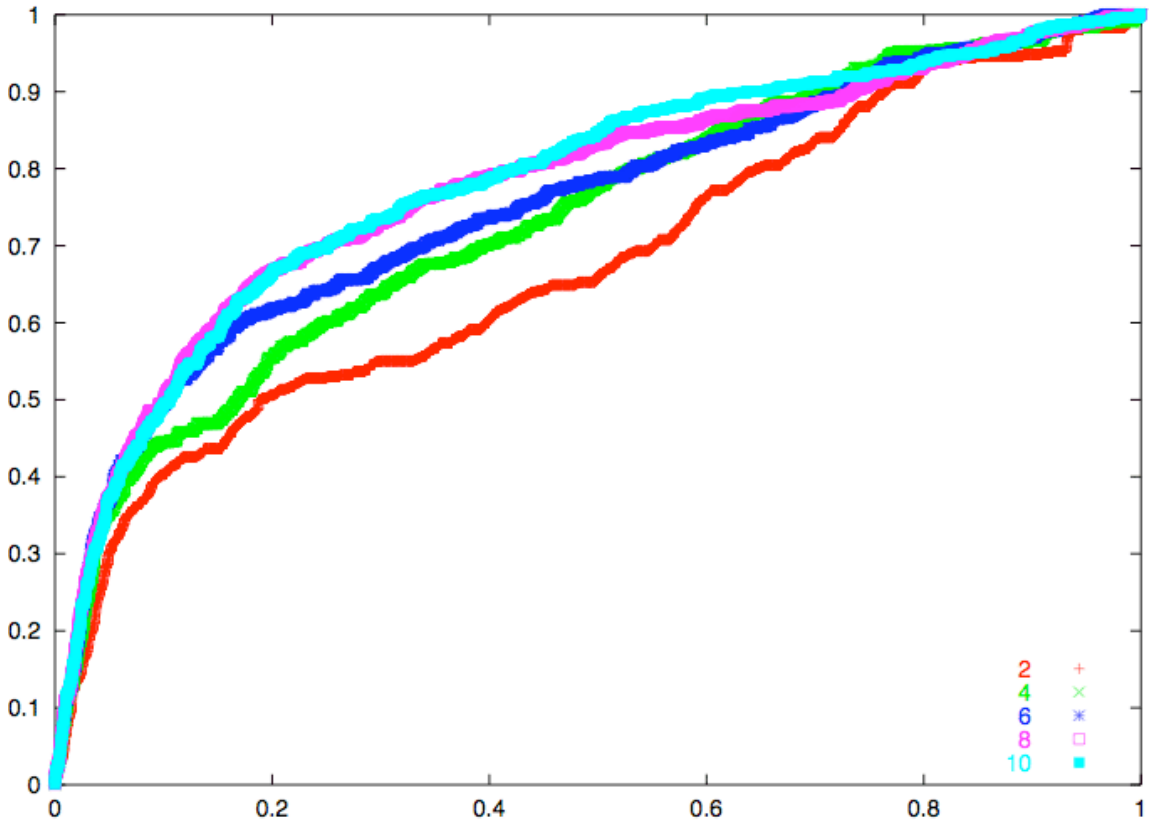
Goals

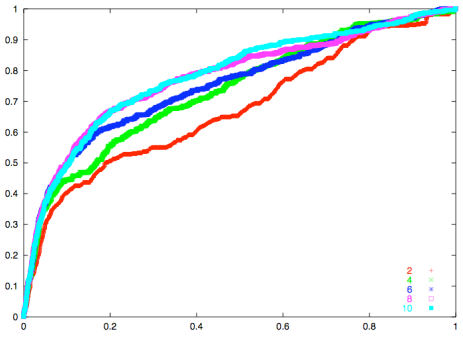
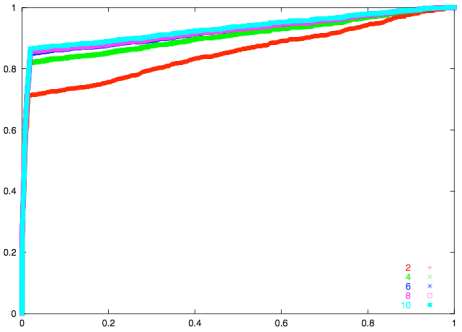
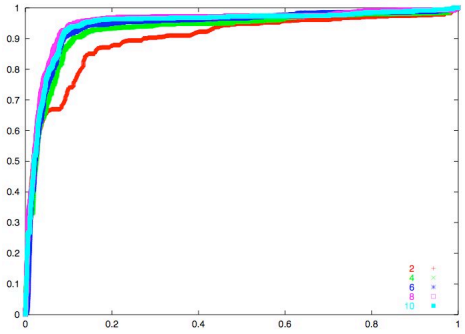
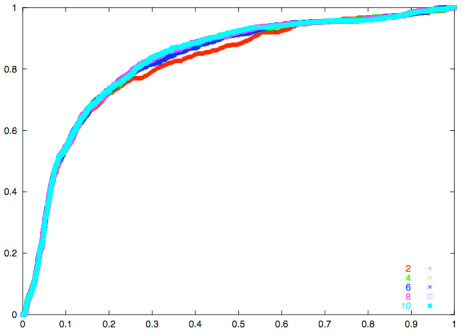
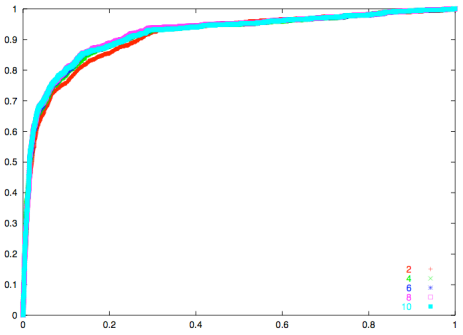
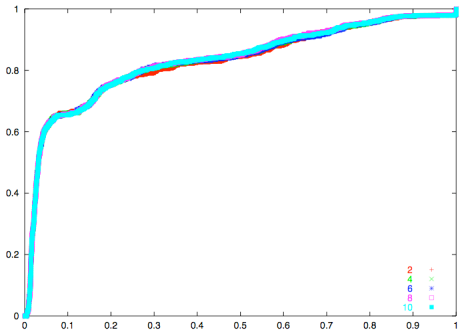
- Can we fill in the missing values of a binding experiment?
- Can we predict all the values of a binding experiment?
- Can we explain the differences in binding?

Evaluation

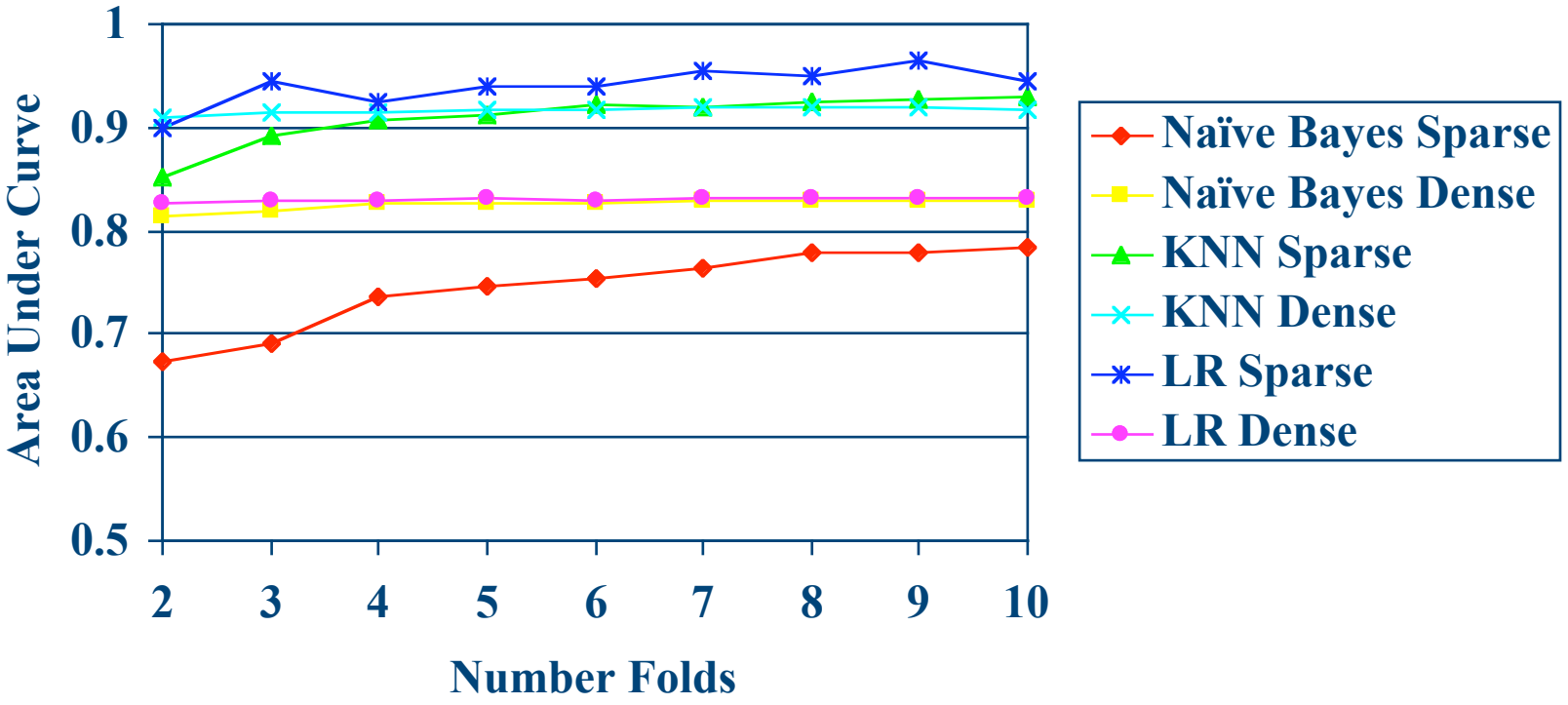
- Use data from all 12 transcription factors
- Training set
  - all edges with binding in either condition
  - randomly selected nonbinding edges
- k-fold validation
  - use 1/k'th of data as test set
  - simulates missing values

ROC Curve: Sparse Naïve Bayes

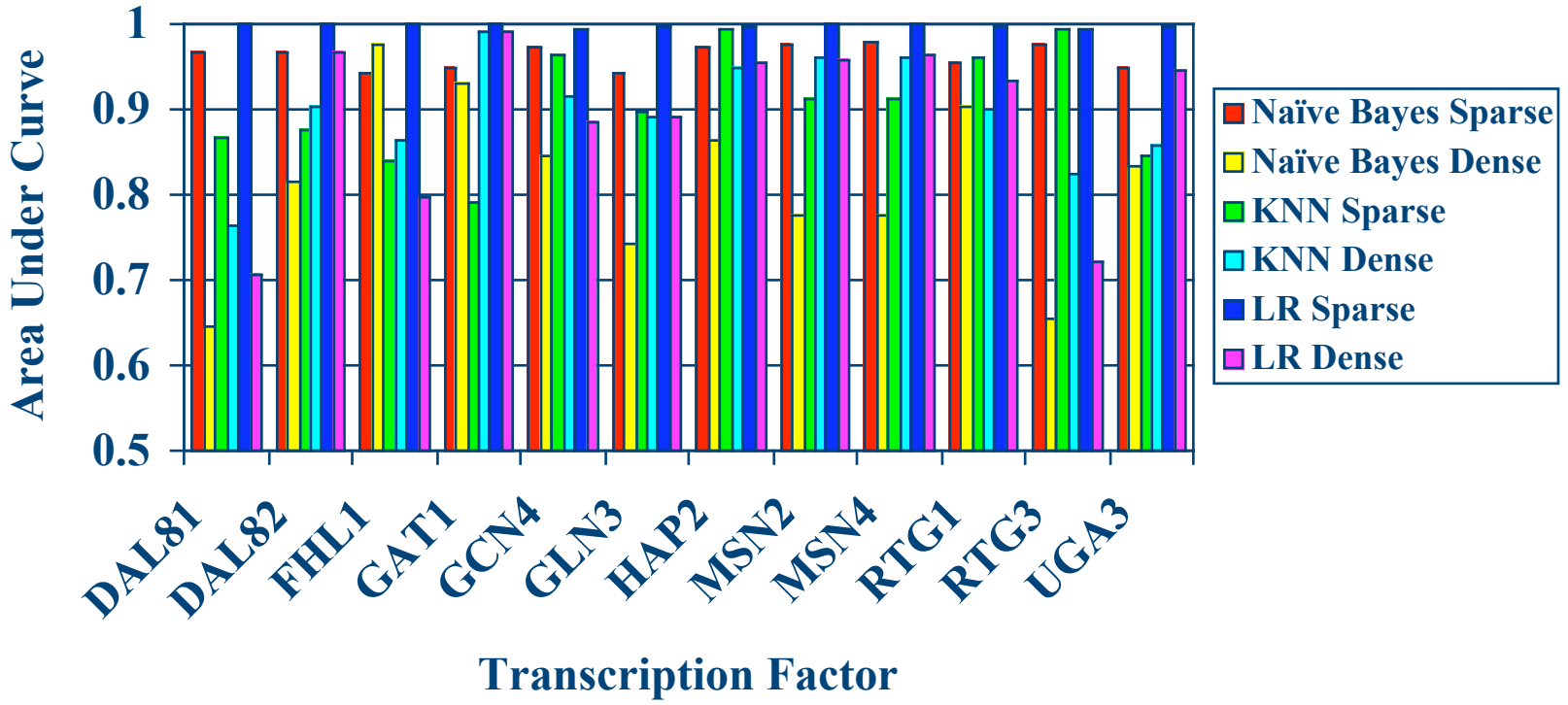


	N. Bayes	KNN	LR
Sparse			
Dense			

K-Folds AUC



8-Fold, Single Factor



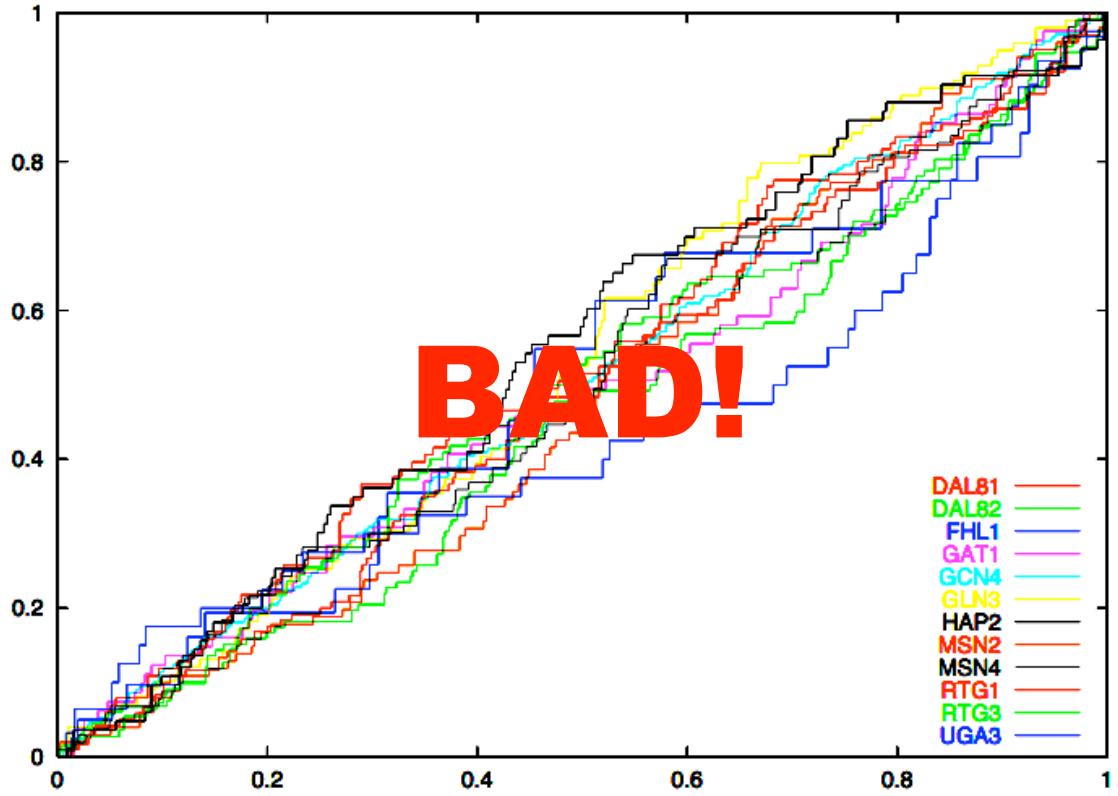
Goals

- Can we fill in the missing values of a binding experiment?
- Can we predict all the values of a binding experiment?
- Can we explain the differences in binding?

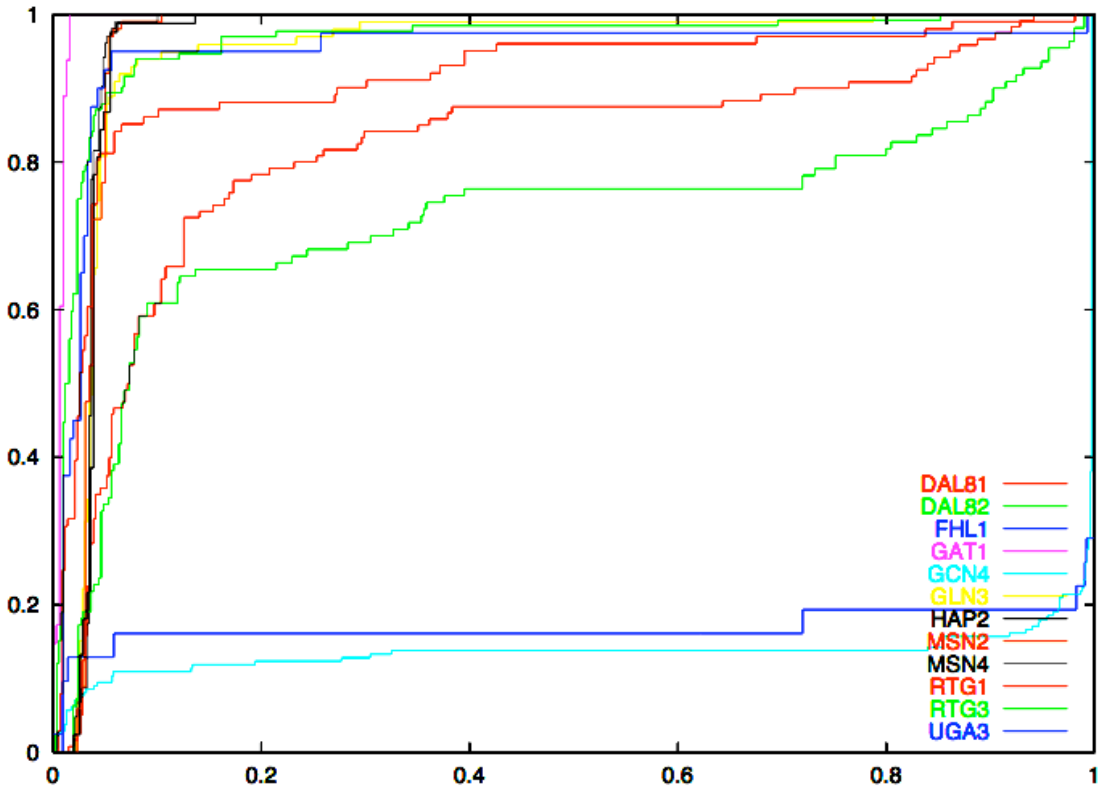
Evaluation

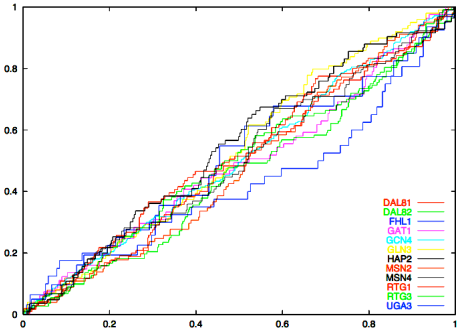
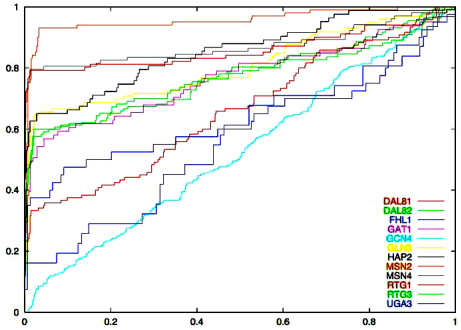
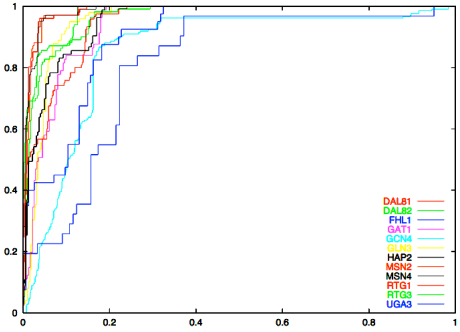
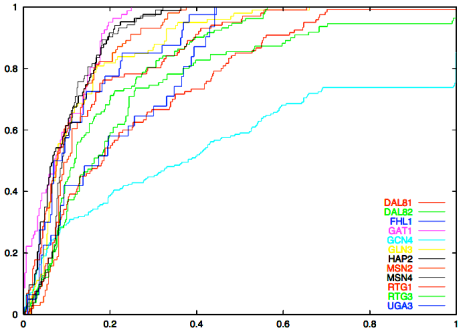
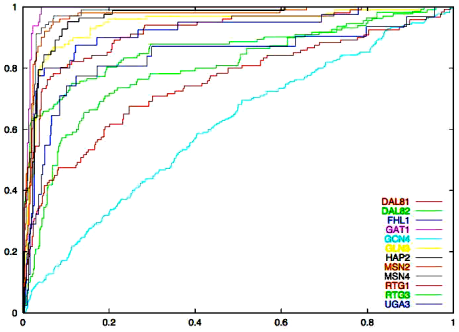
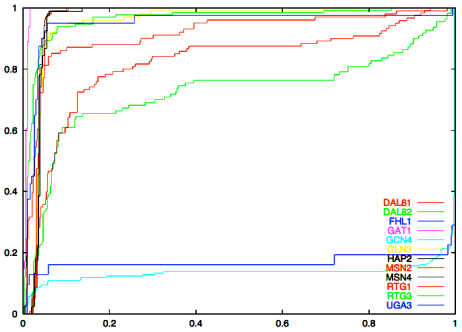
- Training set
  - full data for 11 transcription factors
- Test set
  - full data of remaining transcription factor

ROC Curves: Sparse N. Bayes

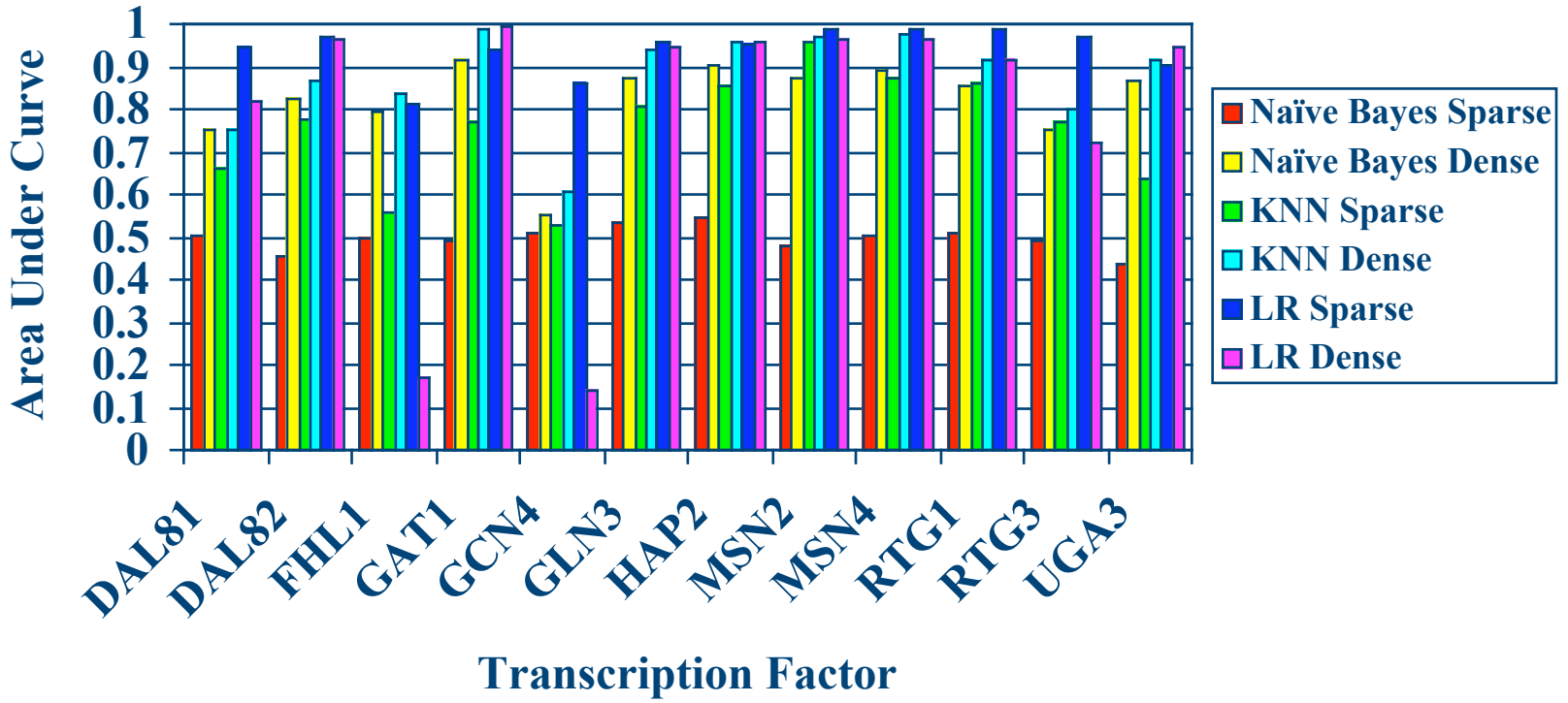


ROC Curves: Dense LR



	N. Bayes	KNN	LR
Sparse			
Dense			

AUC: Leave One Out



Unknown Transcription Factors

- Rapamycin data for only 12 factors
- YPD data for 106 factors
- What is predicted for additional factors?
  - Use sparse LR
  - Only consider already binding YPD edges

Top 20 Most Differing Factors

	FHL1	**94%**		**DAL81**	**62%**
	GAT1	**93%**		**GLN3**	**61%**
	DAL82	**91%**		**RTG1**	**60%**
	UGA3	**90%**		REB1	59%
4	RAP1	88%		MCM1	48%
	MSN4	**82%**	1	FKH1	47%
	MSN2	**82%**		RCS1	46%
1	ABF1	79%		SWI4	44%
	HAP2	**67%**		**RTG3**	**43%**
	CIN5	64%		FZF1	41%

PubMed Hits

Goals

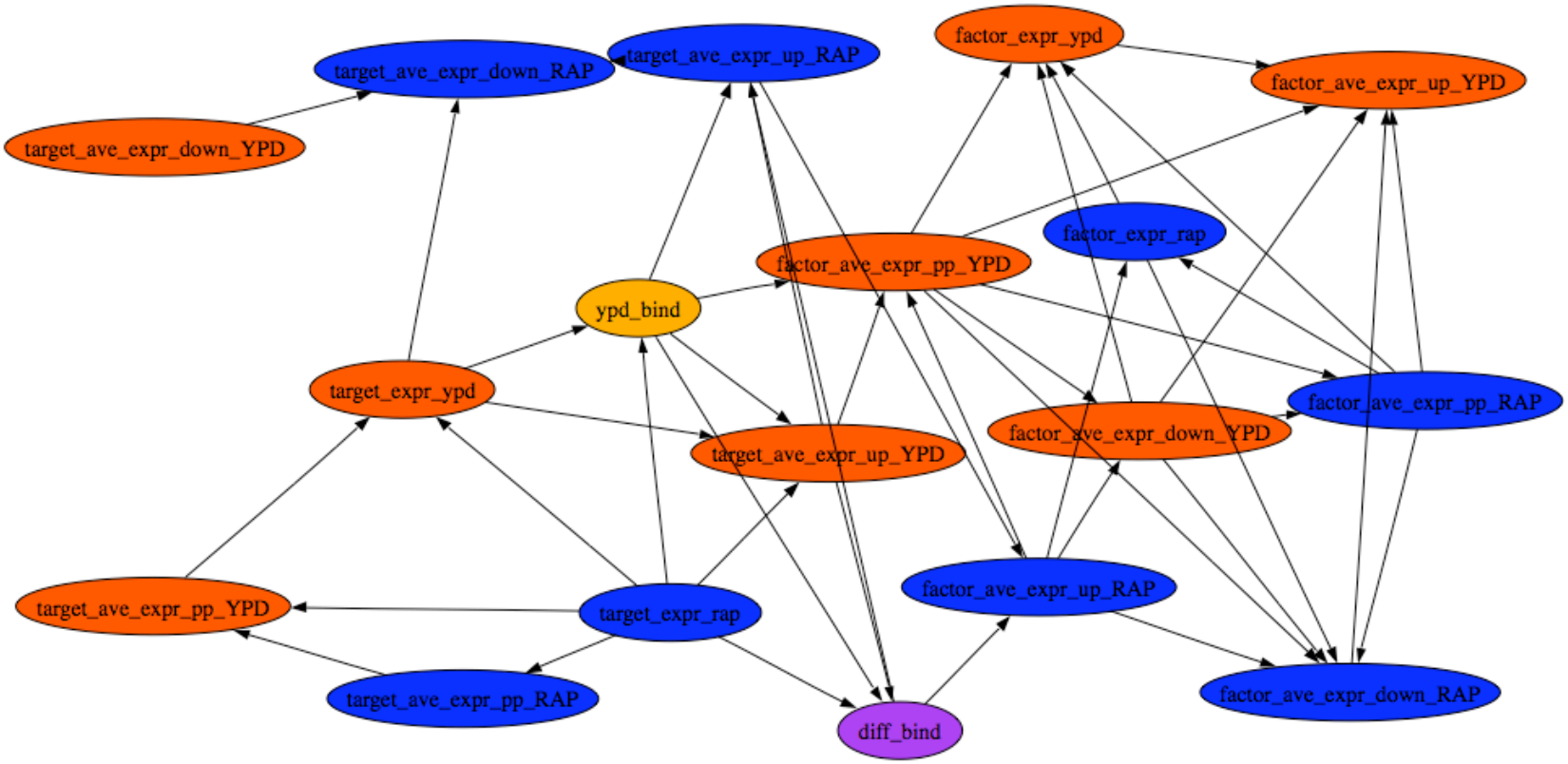
- Can we fill in the missing values of a binding experiment?
- Can we predict all the values of a binding experiment?
- Can we explain the differences in binding?

Learned Bayes Network

- Simple classifiers may be successful
  - but don't generate intuitive models
- Bayesian network might infer causality

- Find network that explains (dense) data well

Learned Baysian Network



Conclusion

- Classifiers very good at filling in missing values
- Classifiers can sometimes predict results of an experiment
  - but sometimes way off
- Results may be used as guide to experimentation
- There may be some biological meaning within the classifier's model