Prediction of Interactions between HIV-1 and Human Proteins by Information Integration

Oznur Tastan[1], Yanjun Qi[1,Ŧ], Jaime G. Carbonell[1] and Judith Klein-Seetharaman[1,2,*]

[1] Language Technologies Institute,
School of Computer Science, Carnegie Mellon University

[2] Department of Structural Biology,
School of Medicine, University of Pittsburgh

[Ŧ] NEC Laboratories America, Inc.

*University of London, Royal Holloway

Human Immunodeficiency Virus-1 (HIV-1)

❑ Causative agent of AIDS

- Destructs the immune system
- Leads to opportunistic infections and malignancies

❑ Current antiviral therapy prolonged the patients' survival rates

- Not accessible to everyone
- Cannot eradicate HIV from the body
- Drug resistance problems

❑ No vaccine

World Health Organization

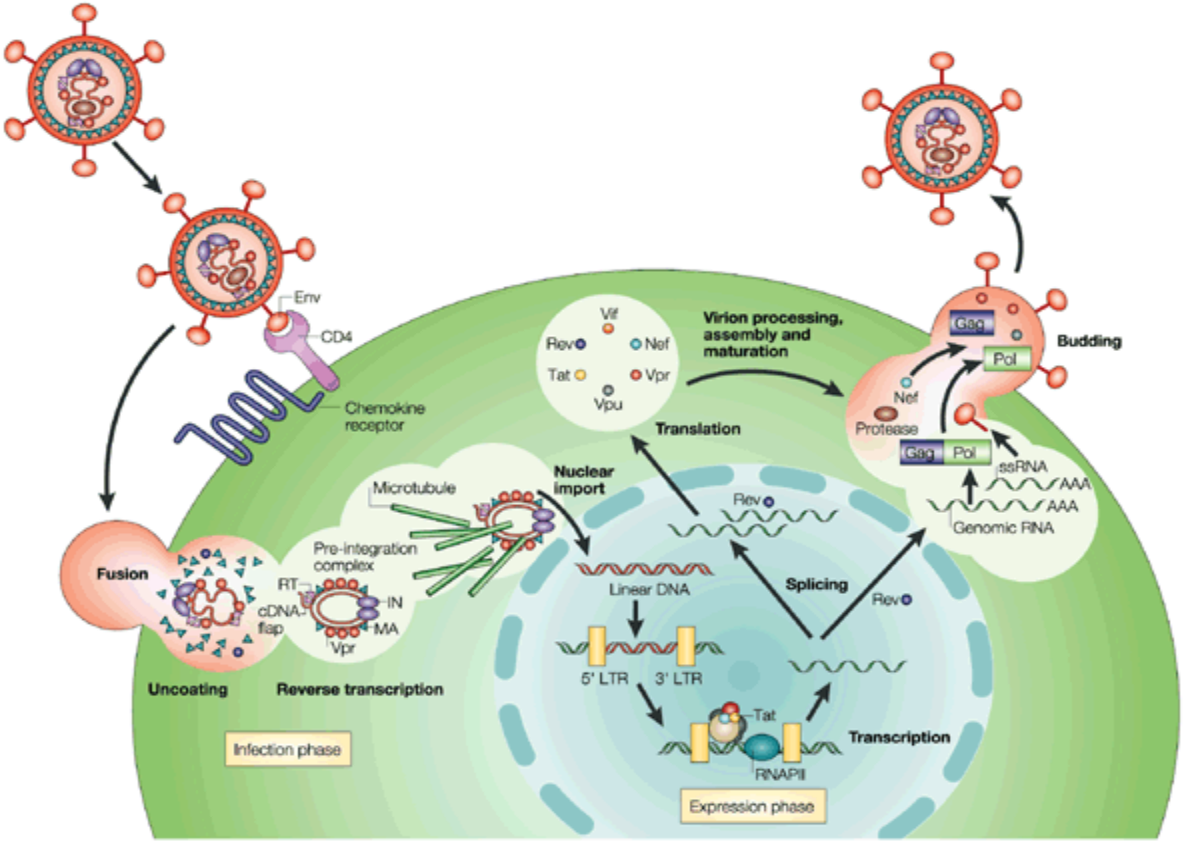
UNAIDS JOINT UNITED NATIONS PROGRAMME ON HIV/AIDS

Global Summary of AIDS epidemic, December 2007

Number of people living with HIV in 2007	**Total** **Children under 15 years**	**33 million** **2 million**
AIDS related deaths in 2007	**Total** **Children under 15 years**	**2.0 million** **270 000**

HIV-1 Genome and Life Cycle

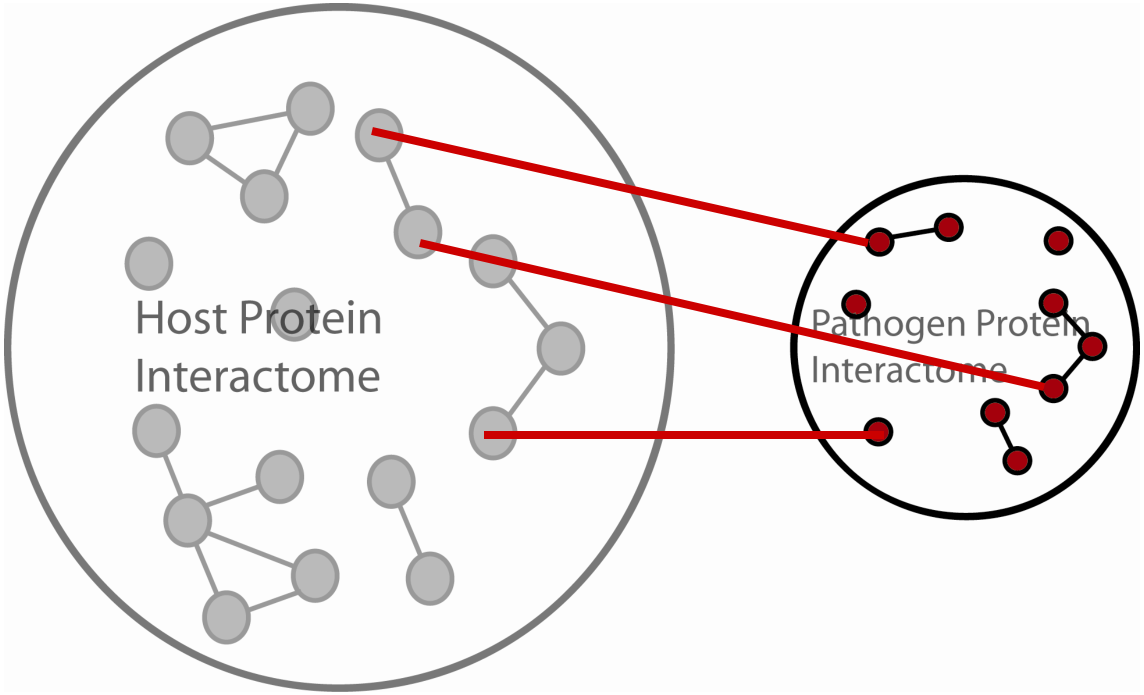
Genes	Proteins
env	env gp160 env gp120 env gp41
gag	nucleocapsid capsid matrix pr55 p6 p1
pol	protease integrase reverse transcriptase
vif	vif
vpu	vpu
vpr	vpr
tat	tat
nef	nef
rev	rev



Peterlin and Trono *Nature Rev. Immu.*(2003) 3: 97-107

Aim

Predict novel direct physical interactions between HIV-1 and human proteins



Prediction of Host Pathogen Interactions

- □ Dyer *et al. Bioinformatics* (2007) 23(13): i159-66
  - Human *Plasmodium falciparum*
  - Co-occurrence of domain sequence signatures

- □ Davis *et al.*, *Protein Sci* (2007) 16(12): 2585-96
  - Inter-PPI of human with 10 pathogens (does not include HIV)
  - Comparative modeling

- □ Konig *et al.* Cell (2008) 135(1): 49-60
  - Functional siRNA knockout screen filtered by multiple evidences

No work to date to predict global interactome of direct physical interactions between HIV-1 and human proteins

Our Approach

Supervised Learning

- □ HIV-1 human protein pair is described with a feature vector and a class label :

$$(\vec{x}_i, y) \quad y \in \{\text{'Interact','Not Interact'}\}$$

↓

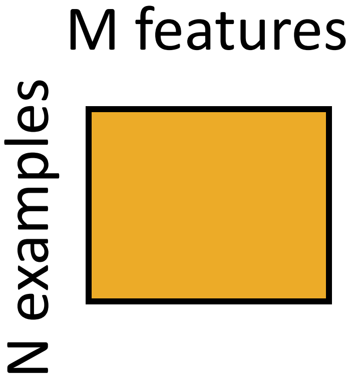
Each feature summarizes a biological information

- □ Given data learn *a function* that would *map feature space into one of the two classes:*

$$f : X \rightarrow Y$$

Random Forest Classifier

Training Data



M features

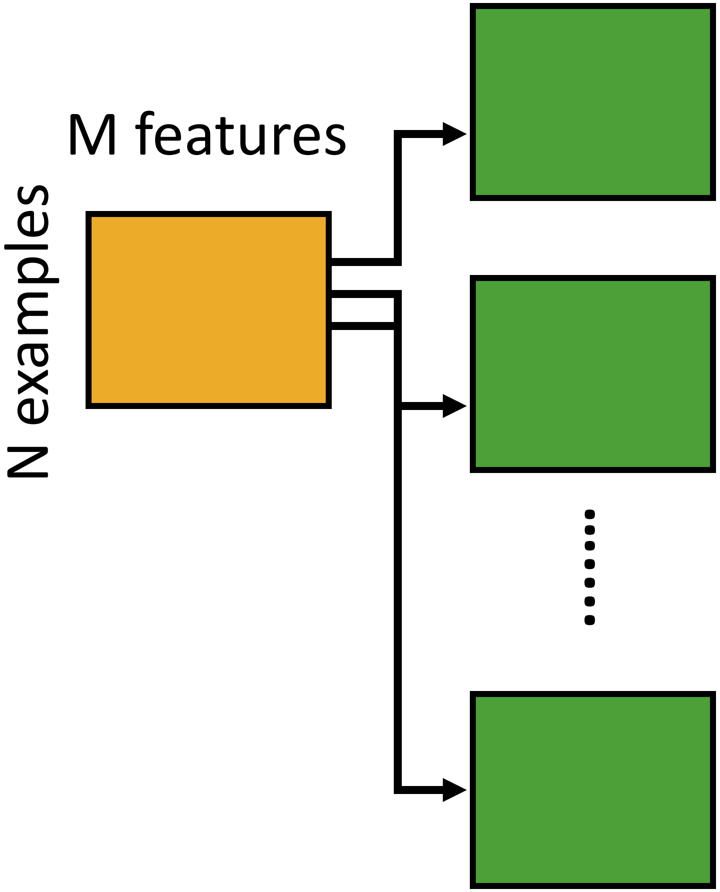
N examples

Qi *et al. Proteins*. (2006) 63: 490-500

Breiman *Machine Learning* (2001) 5-32

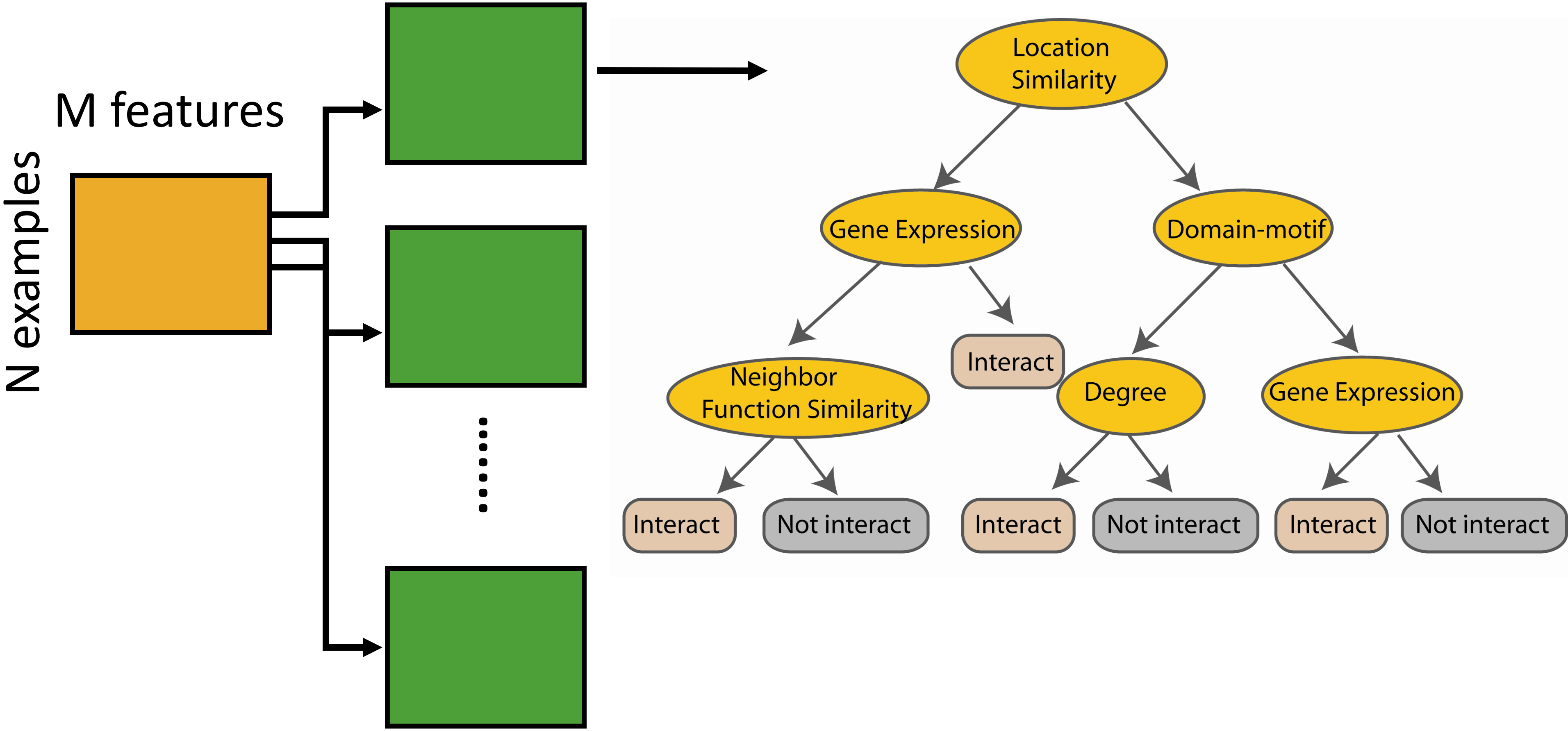
Random Forest Classifier

Create bootstrap samples from the training data



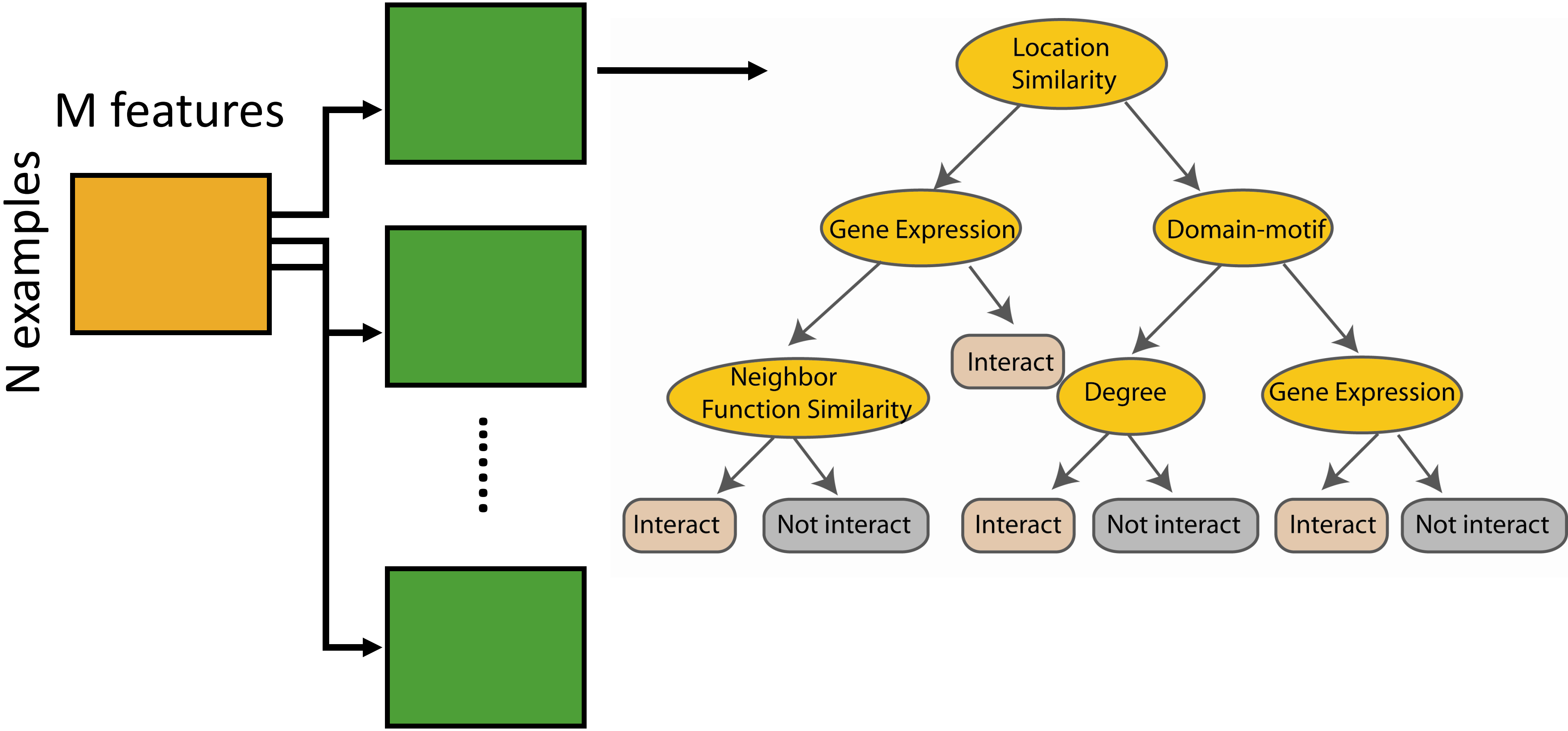
Random Forest Classifier

Construct a decision tree
Use Gini Gain for splitting the nodes



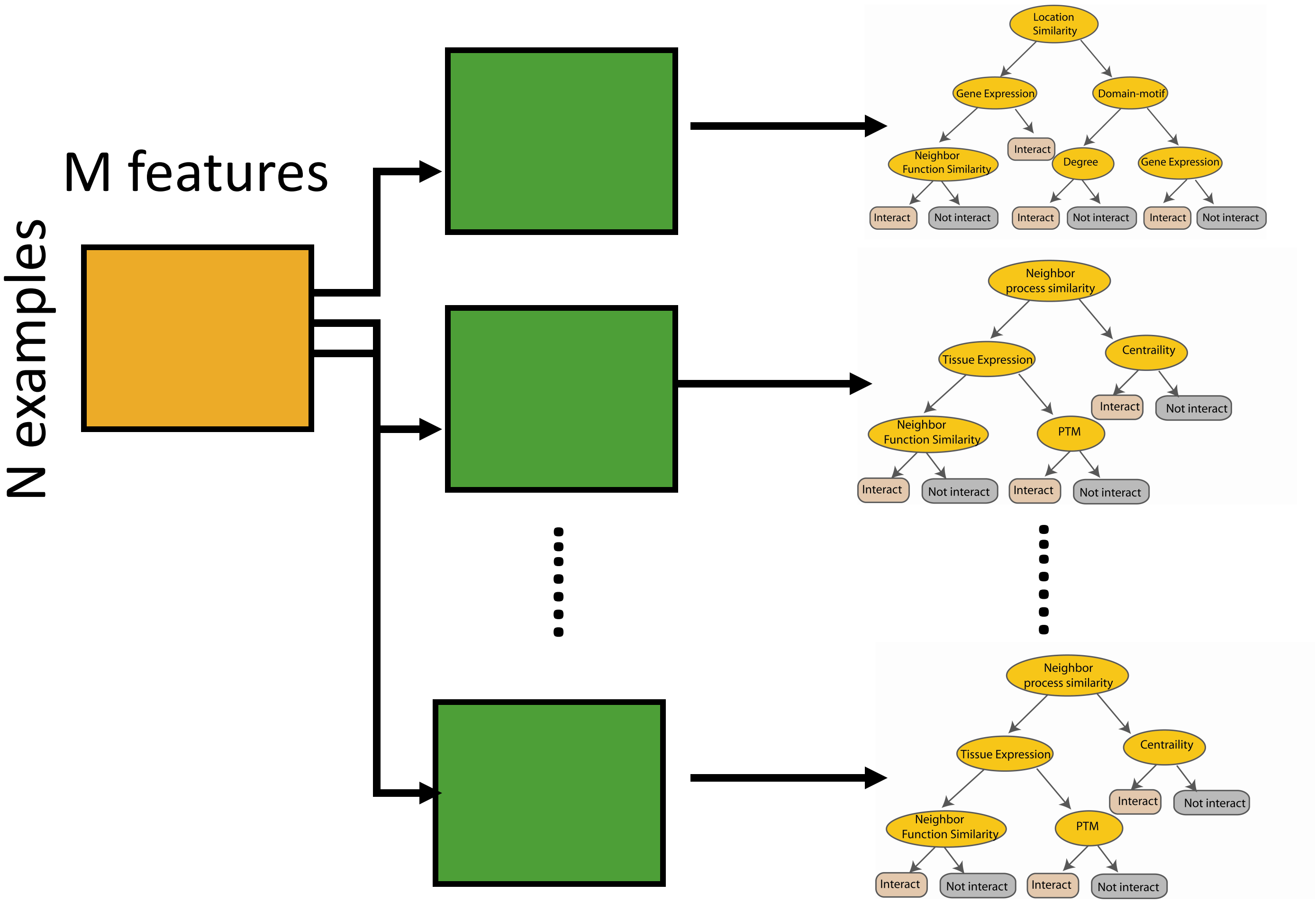
Random Forest Classifier

At each node in choosing the split feature choose only among *m*<*M* features

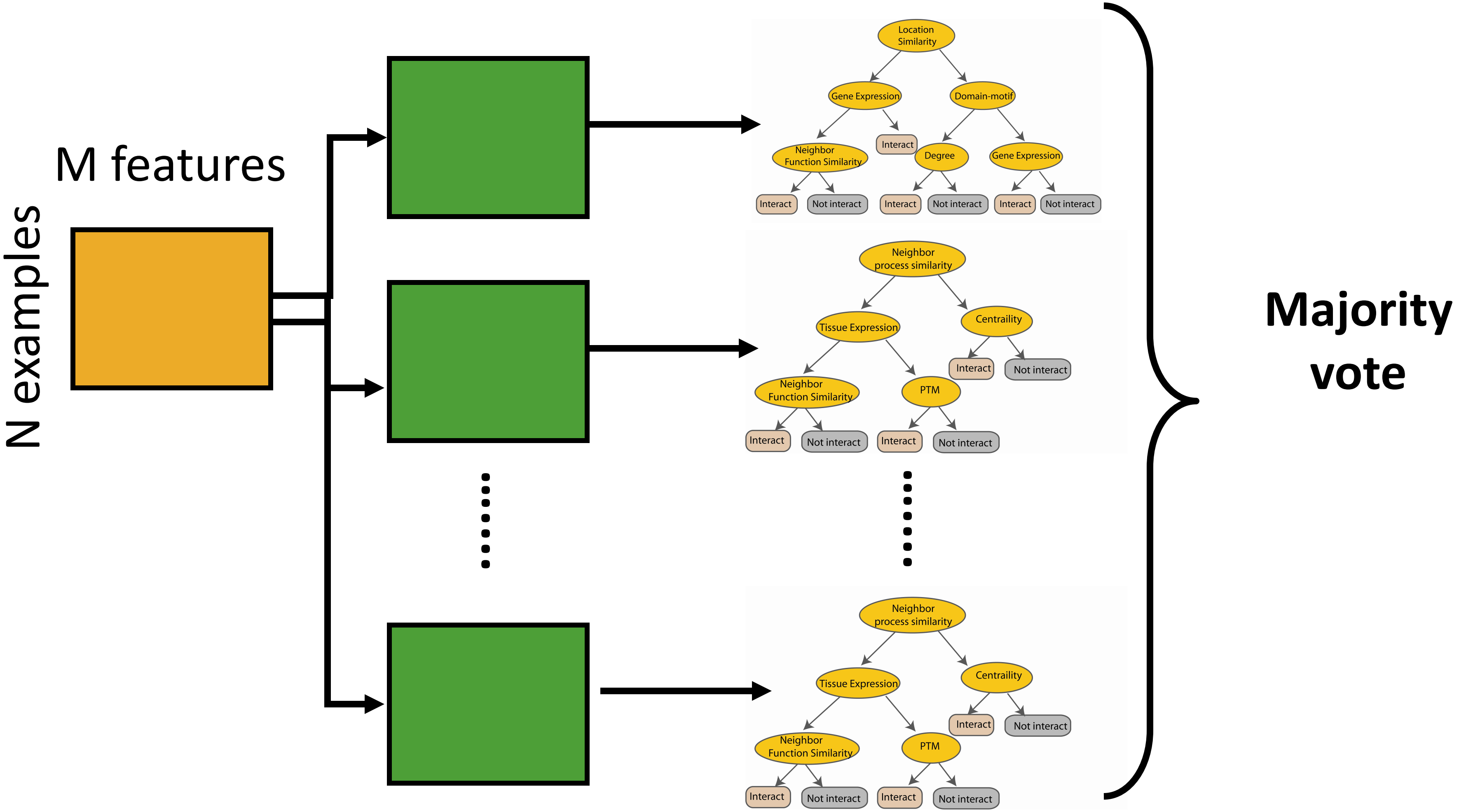


Random Forest Classifier

Create decision tree from each bootstrap sample



Random Forest Classifier



M features

N examples

Location Similarity

Gene Expression

Domain-motif

Neighbor Function Similarity

Interact

Degree

Gene Expression

Interact

Not interact

Interact

Not interact

Interact

Not interact

Neighbor process similarity

Tissue Expression

Centraility

Interact

Not interact

Neighbor Function Similarity

PTM

Interact

Not interact

Interact

Not interact

Neighbor process similarity

Tissue Expression

Centraility

Interact

Not interact

Neighbor Function Similarity

PTM

Interact

Not interact

Interact

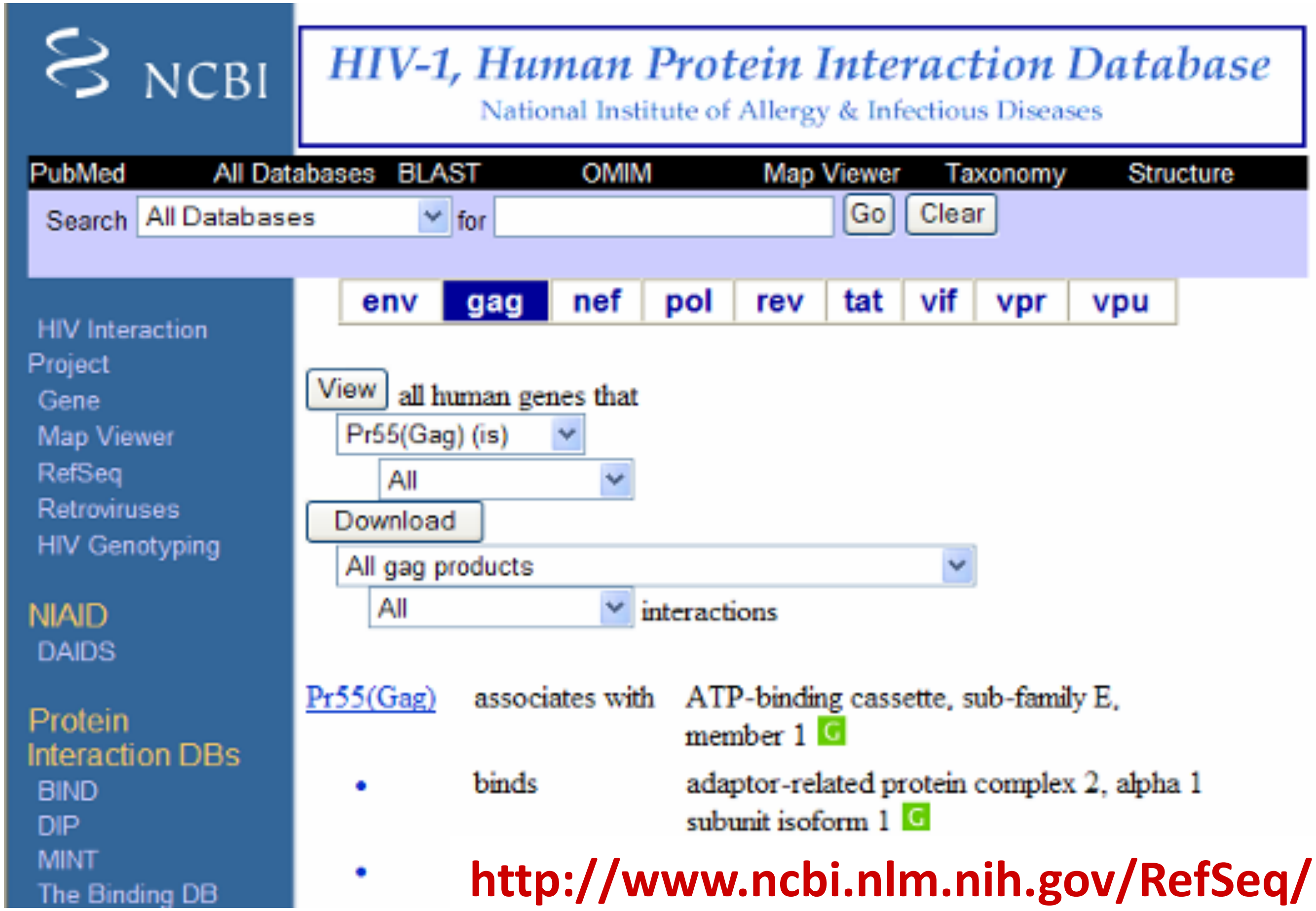
Not interact

Majority vote

Interaction Data

HIV-1 Human Protein Interactions

- ❑ NIAID database of human HIV-1 protein interactions curated from literature



NCBI

HIV-1, Human Protein Interaction Database

National Institute of Allergy & Infectious Diseases

PubMed All Databases BLAST OMIM Map Viewer Taxonomy Structure

Search All Databases for Go Clear

env **gag** nef pol rev tat vif vpr vpu

HIV Interaction Project
Gene
Map Viewer
RefSeq
Retroviruses
HIV Genotyping

NIAID
DAIDS

Protein Interaction DBs
BIND
DIP
MINT
The Binding DB

View all human genes that
Pr55(Gag) (is)
All
Download
All gag products
All interactions

Pr55(Gag) associates with ATP-binding cassette, sub-family E, member 1 G

- binds adaptor-related protein complex 2, alpha 1 subunit isoform 1 G

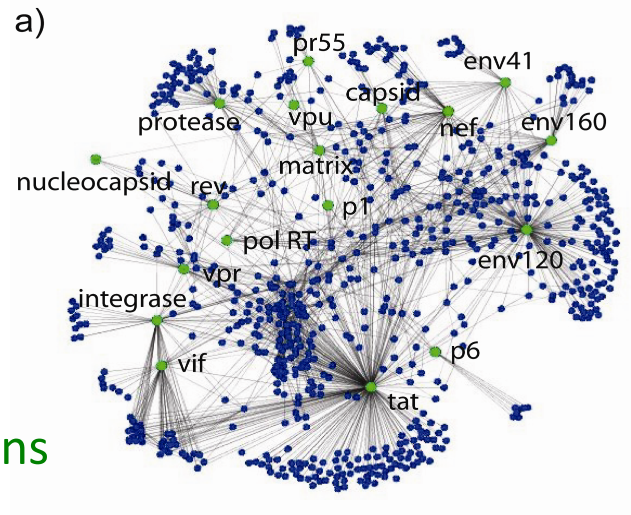
http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions

Sanders-Beer *et al. NAR* (2008) doi: 10.1093/nar/gkn708

HIV-1 Human Protein Interactions

Keywords: “Nef binds hemopoietic cell kinase isoform p61HCK”

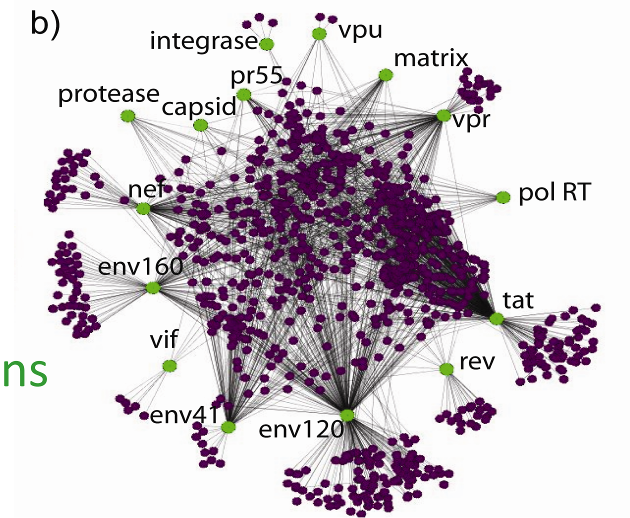
□ Group 1: more likely direct interactions



acetylated by, acetylates, binds, cleaved by, cleaves, degraded by, dephosphorylates, interacts with, methylated by, myristoylated by, phosphorylated by, phosphorylates, ubiquitinated by

1063 interactions, 721 human proteins, 17 HIV-1 proteins

□ Group 2: could be indirect interactions



activated by, activates, antagonized by, antagonizes, associates with, causes accumulation of, co-localizes with, competes with, cooperates with ...etc

1454 interactions, 914 human proteins, 16 HIV-1 proteins

● HIV-1 protein ● ● Human protein

Training and Testing Data

The 'interaction' class:

Group 1, the more likely direct interactions

1063 interactions, 721 human proteins, 17 HIV-1 proteins

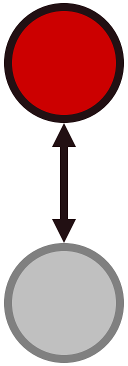
The 'non-interaction' class:

Select randomly from the pairs that are not reported in NIAID database

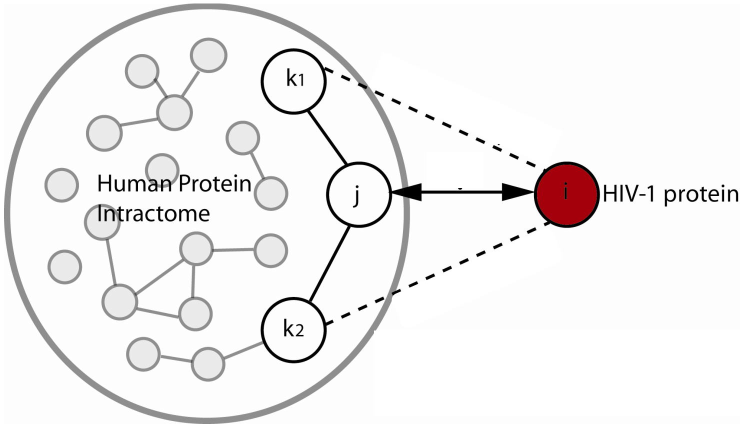
Features

35 Features

- □ Differential gene expression in HIV infected vs uninfected cells (4)
- □ Human protein expression in HIV-1 susceptible tissues (1)
- □ Similarity of the two proteins in terms of (4)
  - Cellular location
  - Molecular process
  - Molecular function
  - Sequence



- □ HIV-1 protein type (17)
- □ ELM-ligand feature (1)
- □ Human PPI interactome features (8)



ELM-Ligand Feature

- Functional interaction motifs obtained Eukaryotic Linear Motif (ELM) database

[RKY]XXPXXP	*motif involved in protein-protein interaction mediated by SH3 domains*

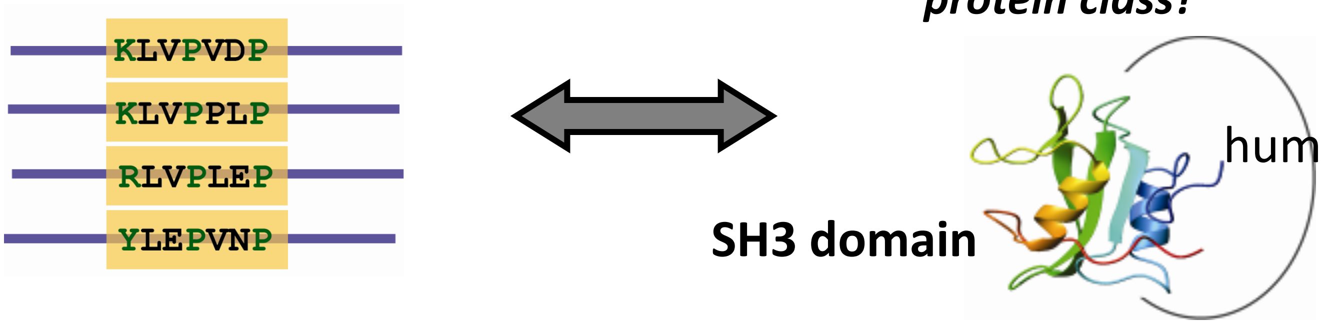
Motif-Ligand Feature

❑ Functional interaction motifs obtained Eukaryotic Linear Motif database

[RKY]XXPXXP *motif involved in protein-protein interaction mediated by SH3 domains*

Is the motif conserved in HIV-1 sequences?

Does the human protein contain the ligand domain or belongs to the ligand protein class?

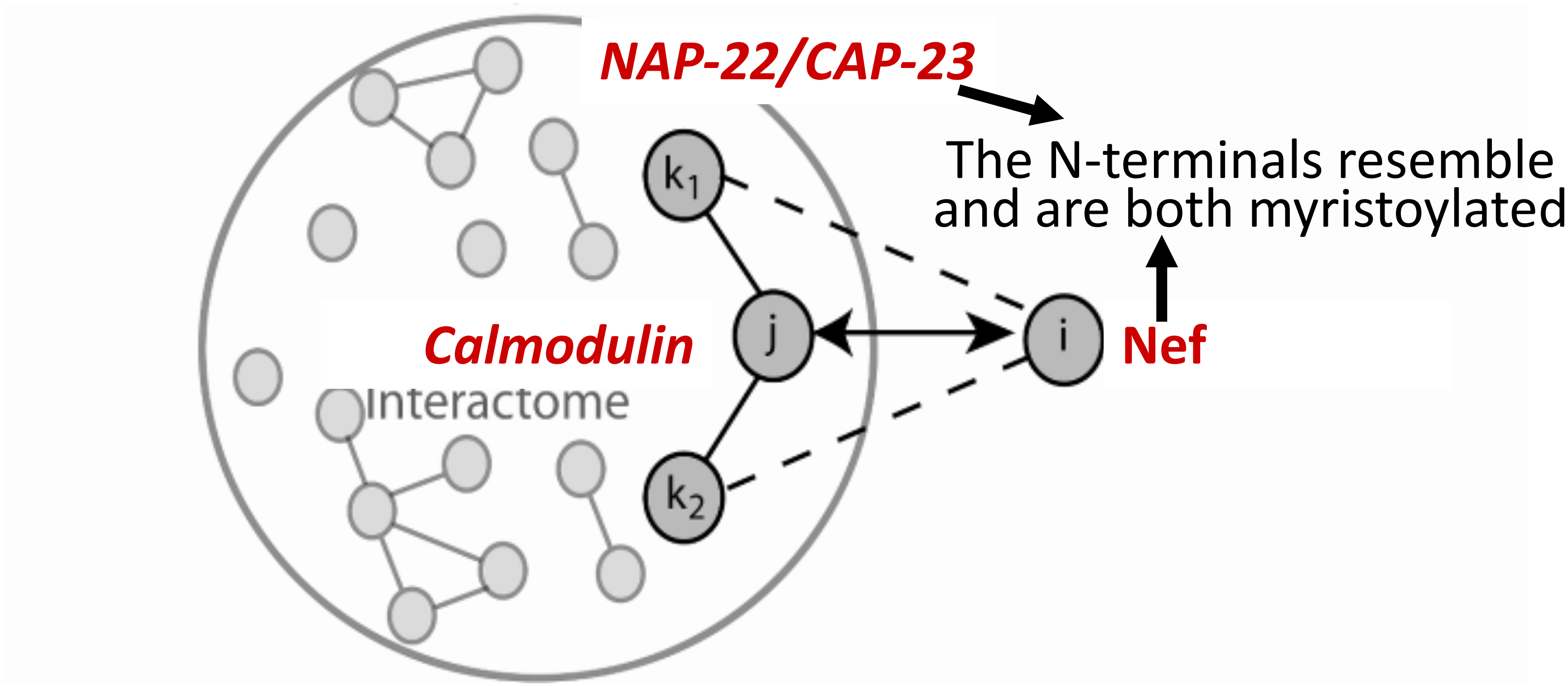


human protein

SH3 domain

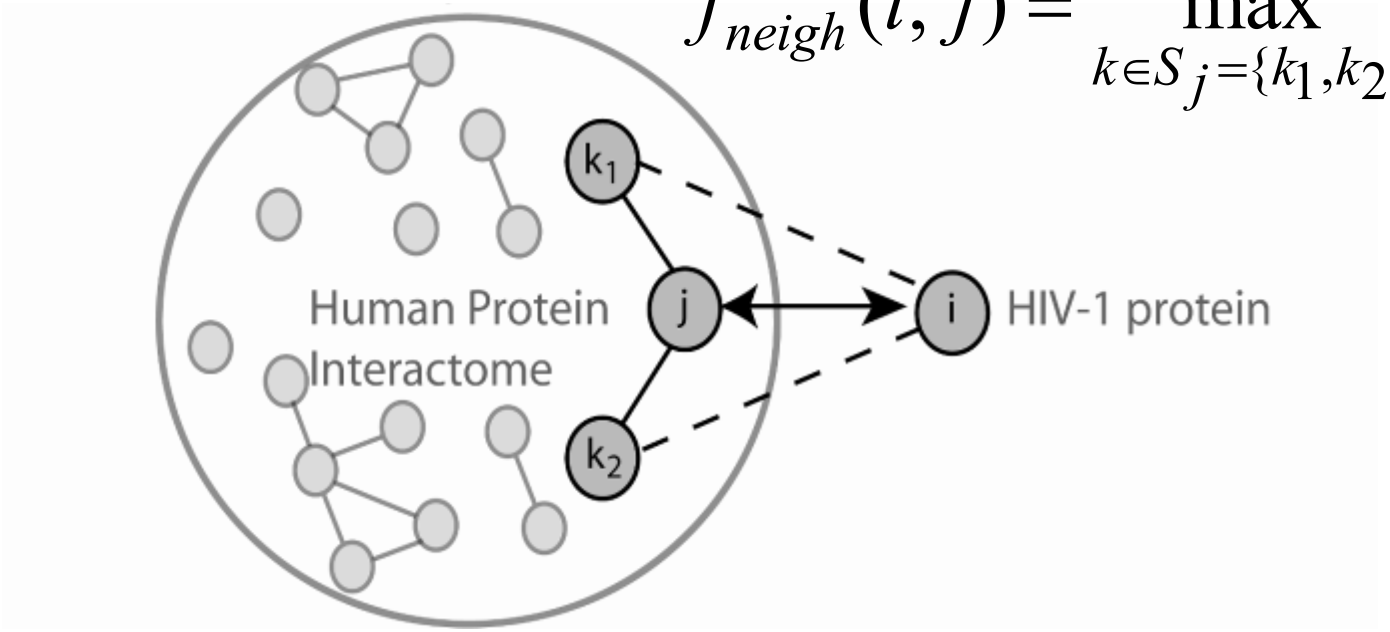
$$f_{motif} = q, \text{ where } 0 \le q \le 1$$

Making Use of the Human PPI Interactome: Mimicry of Human Protein Interaction Partners’



Making Use of the Human PPI Interactome: Mimicry of Human Protein's Interaction Partners

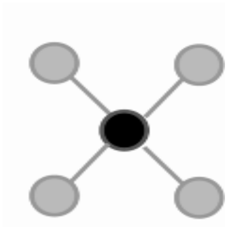
$$f_{neigh}(i, j) = \max_{k \in S_j = \{k_1, k_2\}} f_{pairwise}(i, k)$$



❑ Similarity of HIV-1 protein to human protein's interaction partner

- Sequence
- Post translational modification
- Cellular location
- Molecular process
- Molecular function

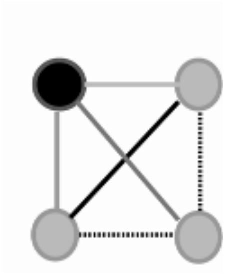
Making Use of the Human PPI Interactome: Human Protein’s Topological Properties the Human PPI network



Degree

Number of neighbors

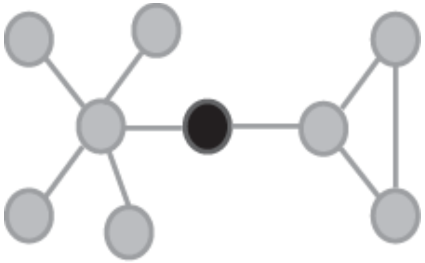
$$k_v$$



Clustering coefficient

The extent the neighbors are connected with each other

$$\frac{2n_v}{k_v(k_v-1)}$$



Betweenness Centrality

The fraction of shortest paths pass through the node

$$\sum_{\substack{u,w\in V\\ u,w\neq v}} \frac{\sigma_{uw}(v)}{\sigma_{uw}}$$

Evaluation

Performance Measures

❑ Precision Recall Curve

- Precision : TP/(TP+ FP)
- Recall (Sensitivity) : TP/(TP+ FN)

Performance Measures

❑ Precision Recall Curve

- Precision : TP/(TP+ FP)
- Recall (Sensitivity) : TP/(TP+ FN)

❑ The Mean Average Precision (MAP):

- Mean of the average precisions where each average precision is calculated when recall increases.

Performance Measures

❑ Precision Recall Curve

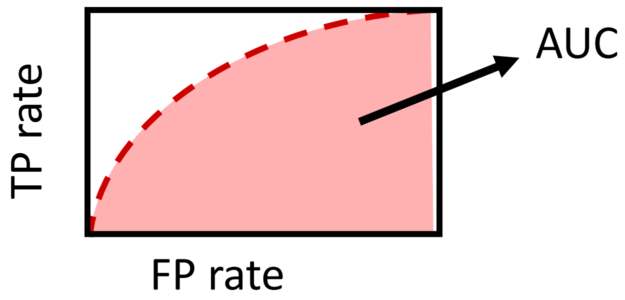
- Precision : TP/(TP+ FP)
- Recall (Sensitivity) : TP/(TP+ FN)

❑ The Mean Average Precision (MAP):

- Mean of the average precisions where each average precision is calculated when recall increases.

❑ Area Under the Receiver Operating Curve (AUC):

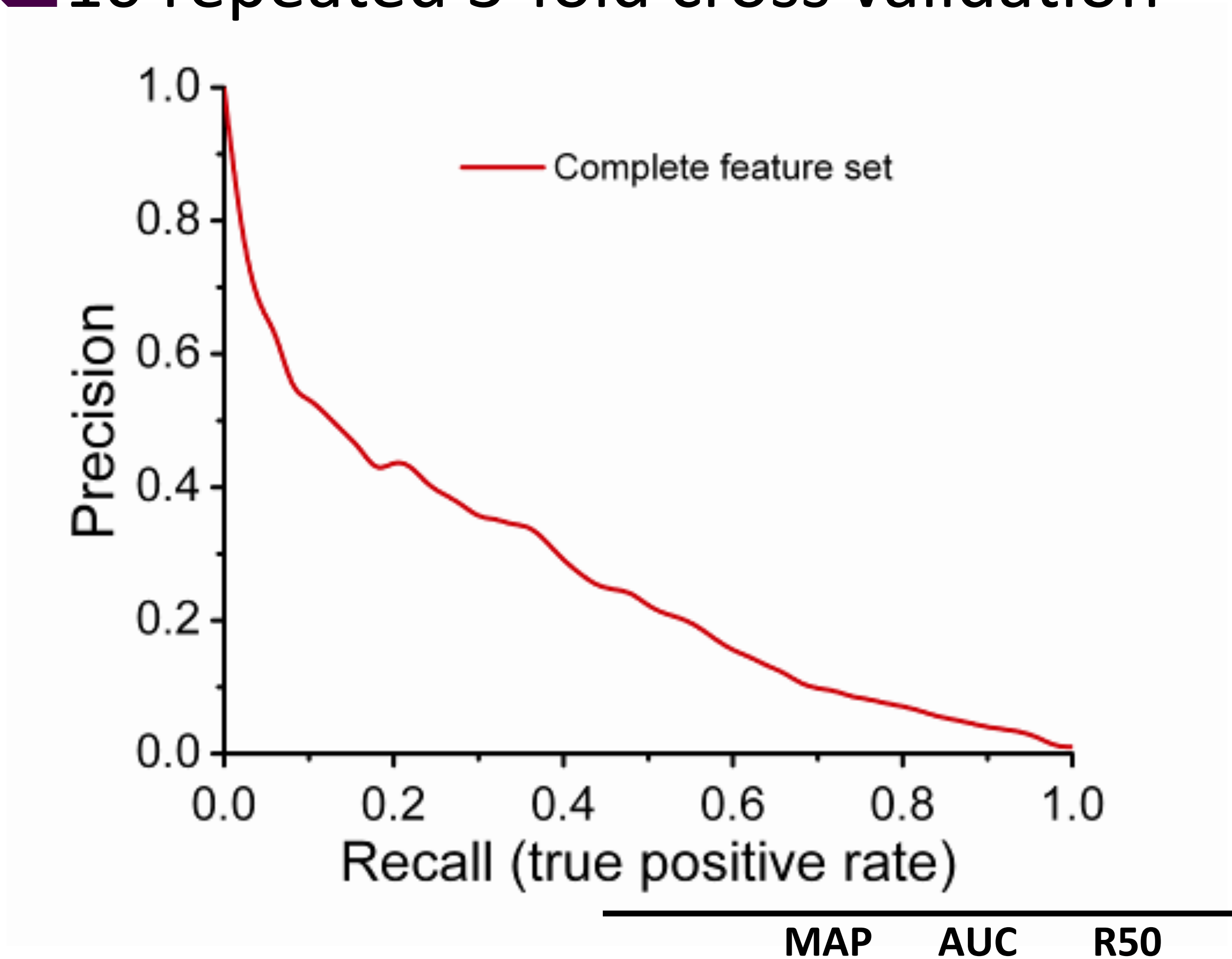
ROC curve



- Partial AUC scores :
  Area under the curve
  until reaching N false positives

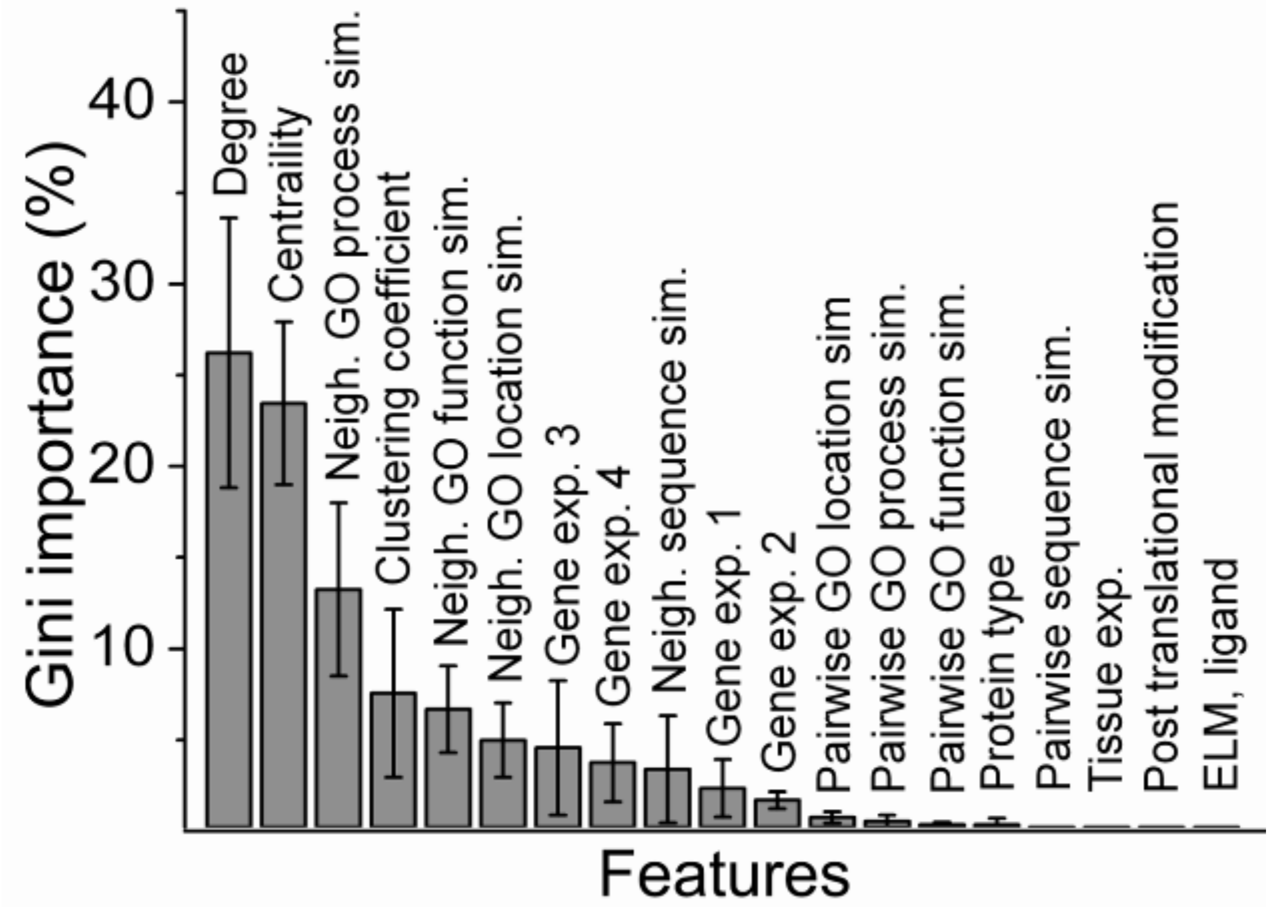
Performance Evaluation

- 10 repeated 3-fold cross validation



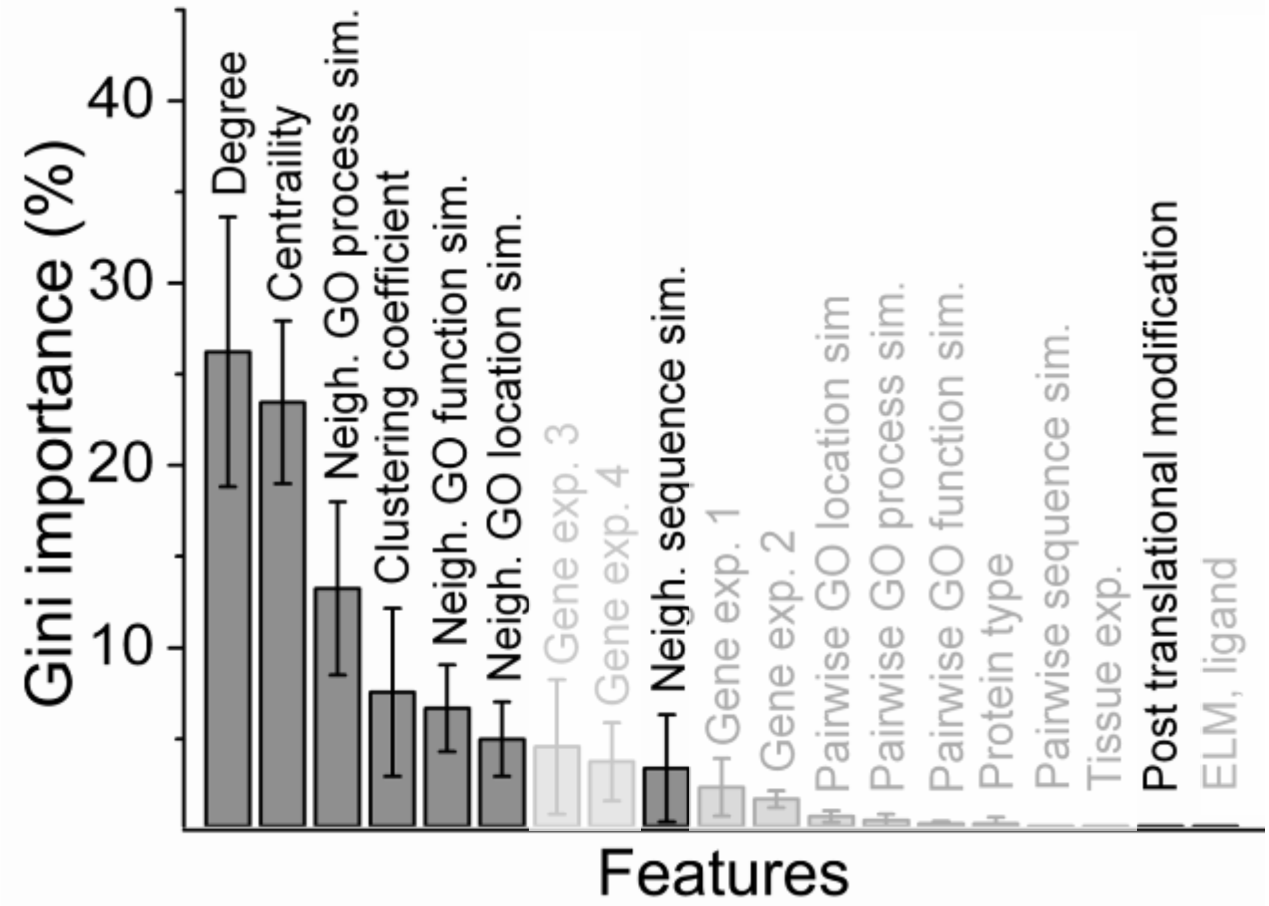
	MAP	AUC	R50	R100	R200	R300
Avg	0.23	0.92	0.07	0.11	0.17	0.22
Std	0.02	0.01	0.01	0.02	0.02	0.02

Feature Importance



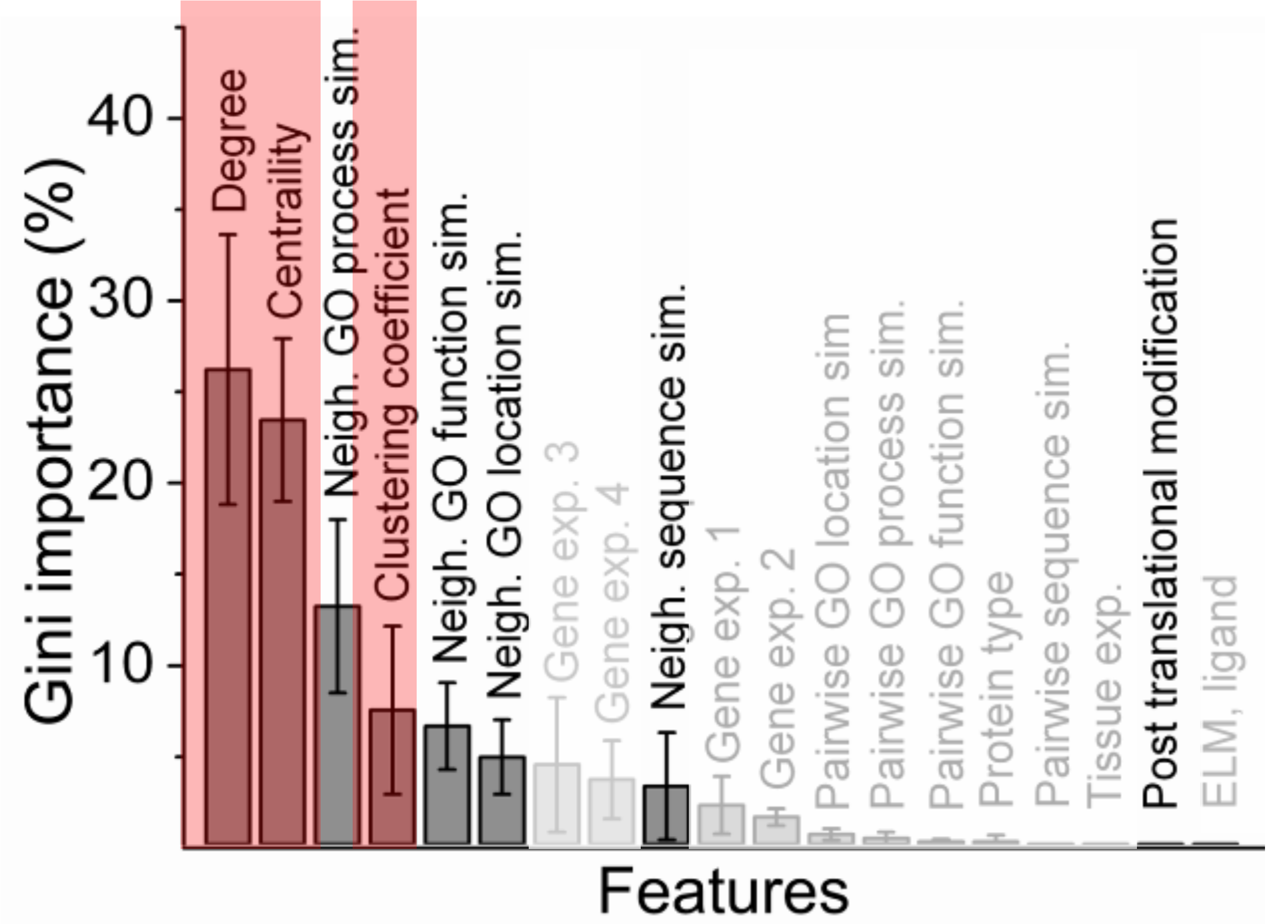
Gini importance: Normalized sum of improvement in the "Gini gain" due a given feature in the forest

Feature Importance



Majority of the human interactome features are highly informative

Feature Importance



The network topology features are highly ranked

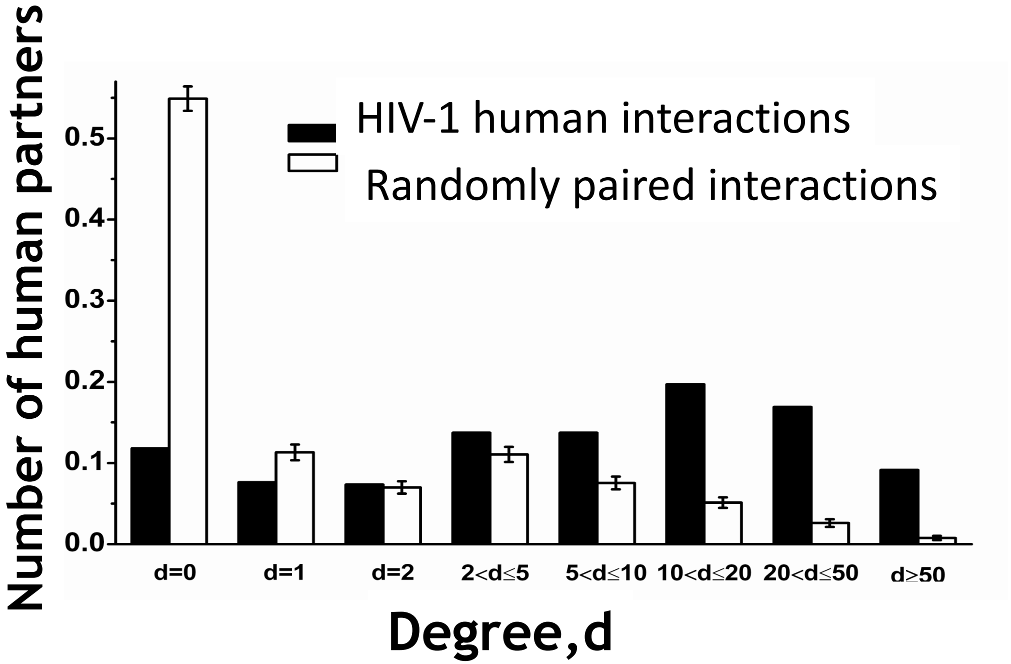
Network Features

- ❑Epstein–Barr virus targets high degree human proteins

  Calderwood *et al.*, *PNAS* (2007) 104: 7606-11

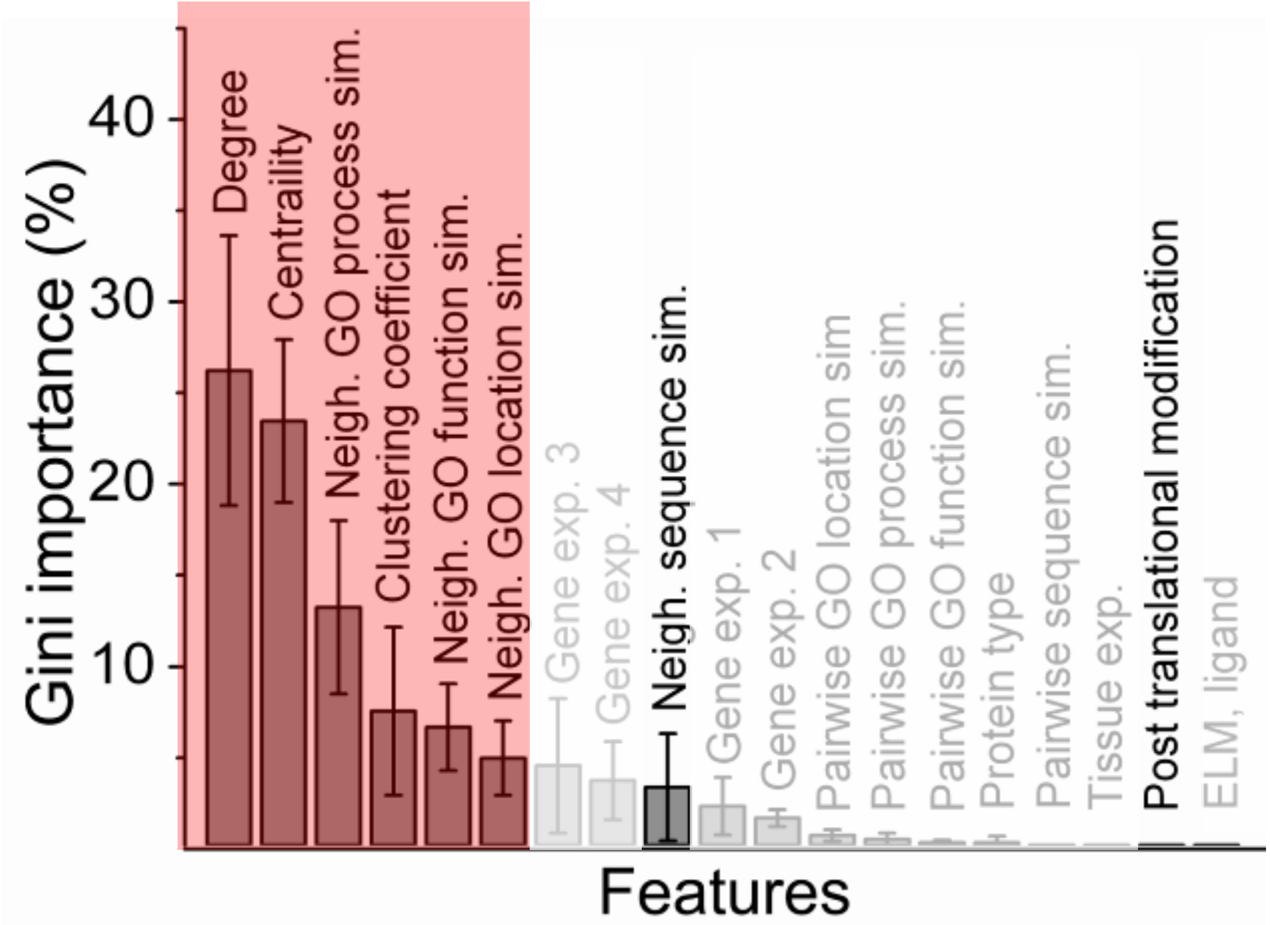
- ❑Pathogens tend to interact with host proteins with high degrees and betweenness centrality

  Dyer et. al. *PLoS Pathog* (2008) 4, e32



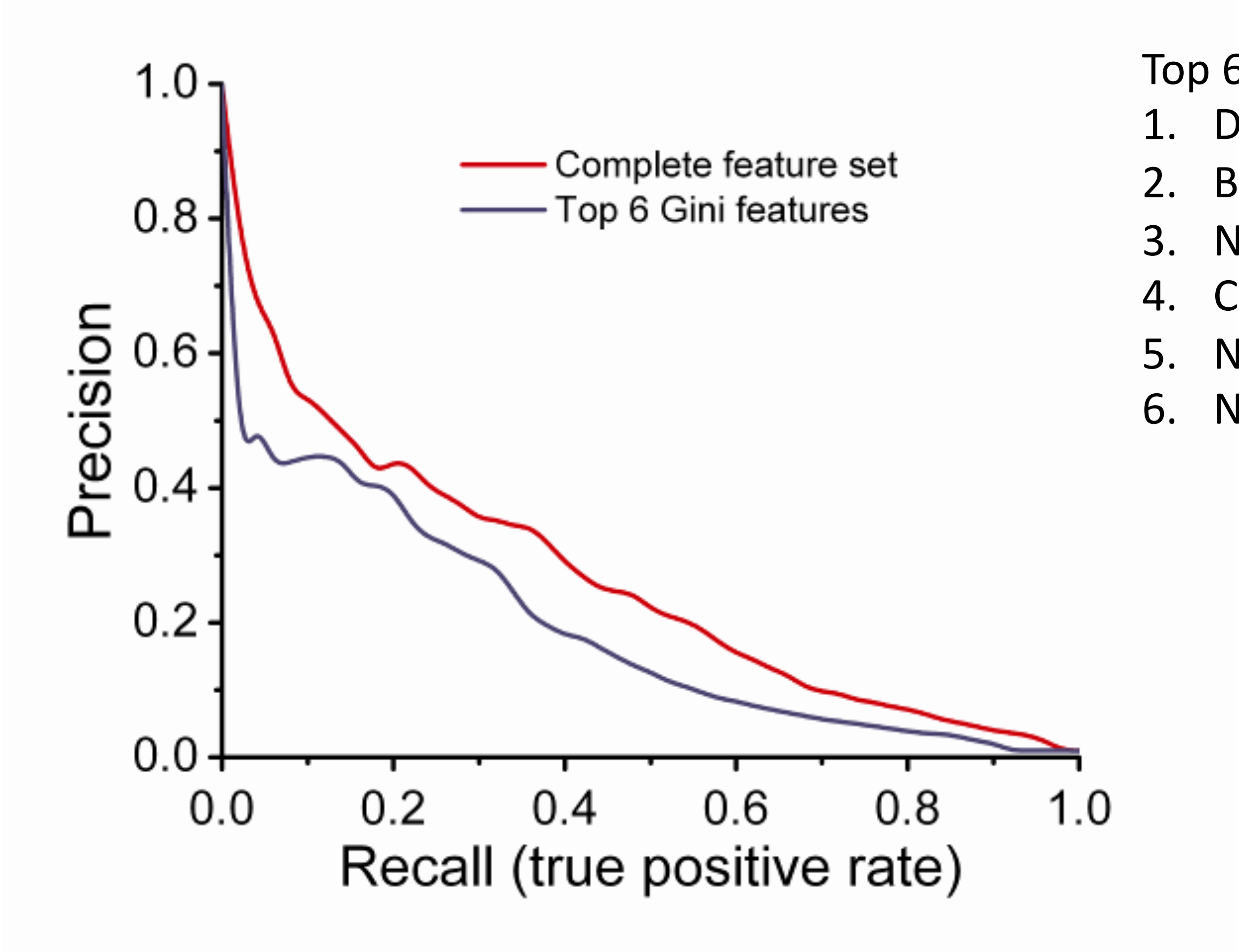
Feature Importance

Top 6 features



How can we perform using only the top 6 features?

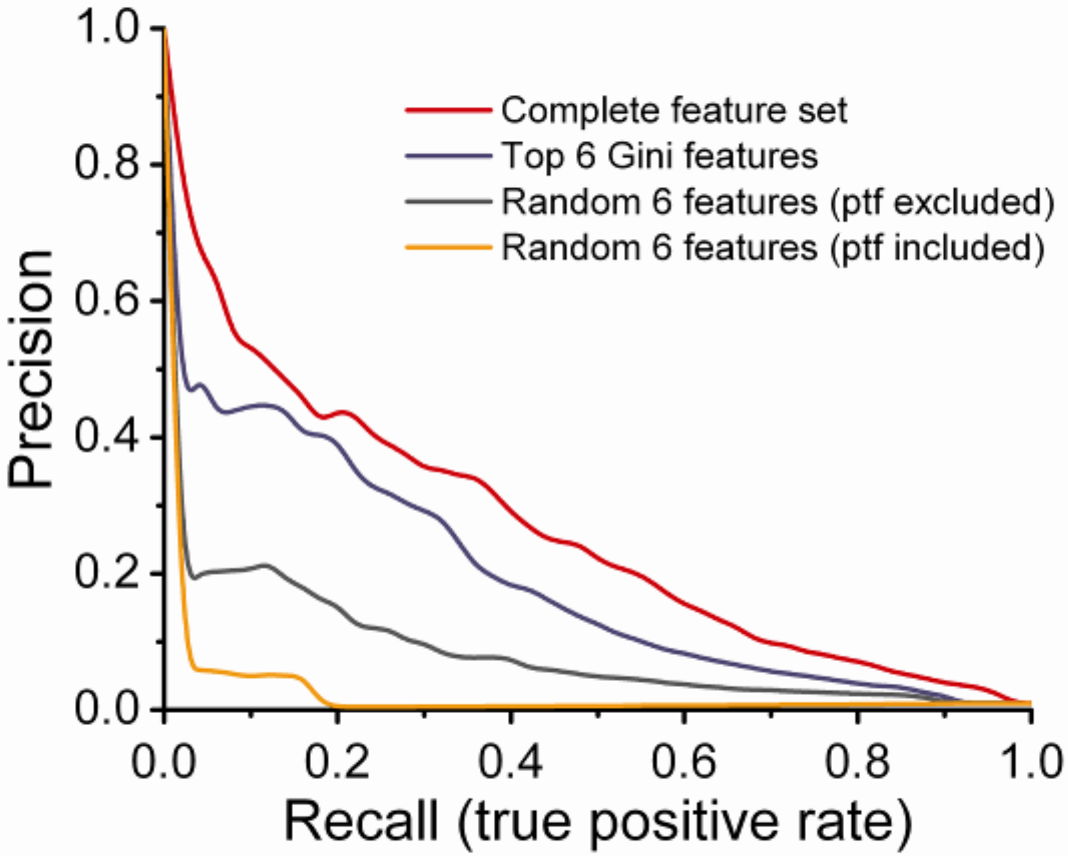
When only the Top Ranked Features Used



Top 6 Gini Features:

1. Degree
2. Betweenness centrality
3. Neighbor process similarity
4. Clustering coefficient
5. Neighbor function similarity
6. Neighbor location similarity

When only the Top Ranked Features Used

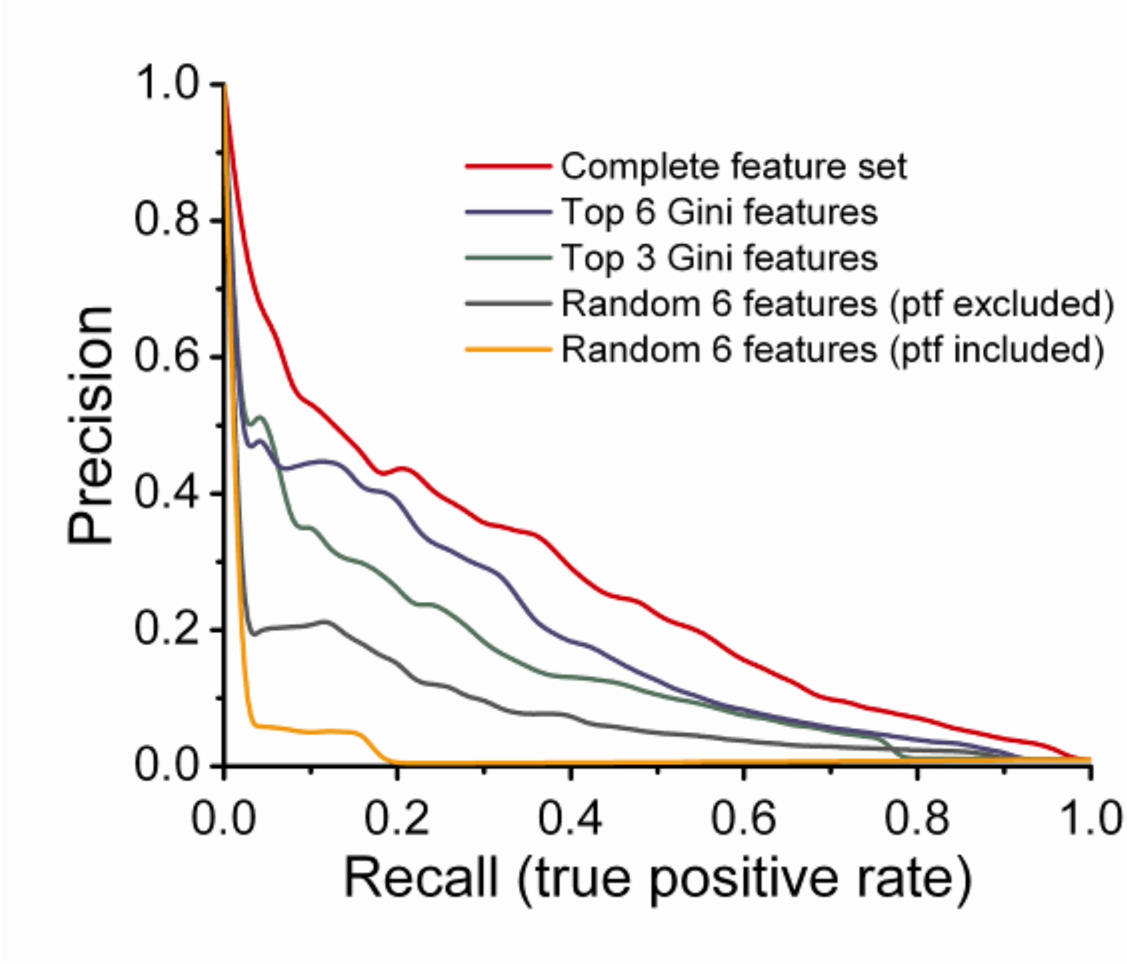


Top 6 Gini Features:

1. Degree
2. Betweenness centrality
3. Neighbor process similarity
4. Clustering coefficient
5. Neighbor function similarity
6. Neighbor location similarity

PTF: Protein type features

When only the Top Ranked Features Used



Top 6 Gini Features:

1. Degree
2. Betweenness centrality
3. Neighbor process similarity
4. Clustering coefficient
5. Neighbor function similarity
6. Neighbor location similarity

PTF: Protein type features

Predicted Interactions

Predictions

- ❑ Apply the model to all possible HIV-1, human protein pairs

Increasing Novelty →

High Recall

Score Cutoff	Total Pairs	Group 1	Group 2	Novel	Recall	Precision
≥ 0.00	3372	1040	232	2100	0.51	0.20
≥ 0.50	1942	1034	141	767	0.37	0.29
≥ 1.00	1440	1023	68	349	0.26	0.36
≥ 1.50	1085	894	34	157	0.18	0.41
≥ 2.00	622	538	15	69	0.13	0.47
≥ 2.50	279	243	8	28	0.09	0.47

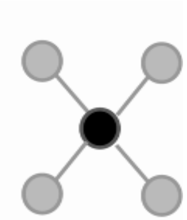
Used in model construction

Predictions

High Precision

Functionally Interesting Interactions

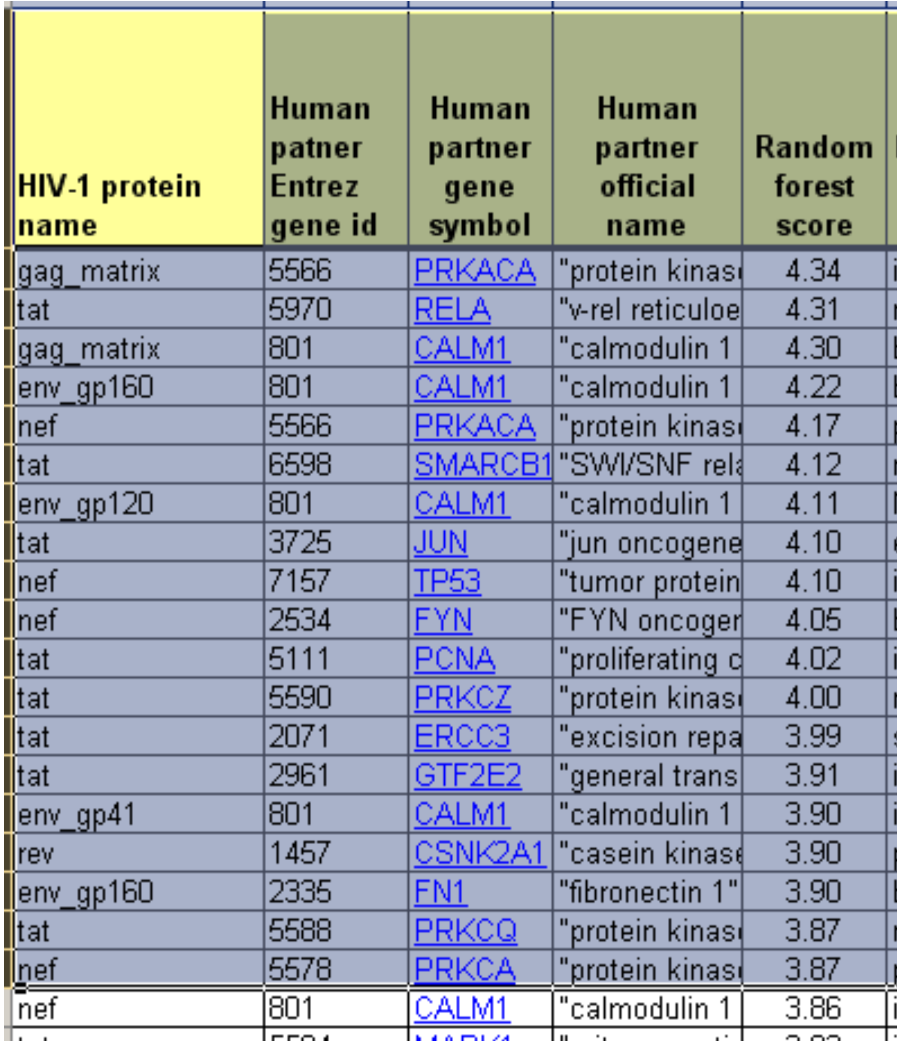
- 304 cellular proteins detected in virion
  - Ott *Rev Med Bio (*2008 17: 159-75)
- 273 genes that had an effect in the Brass siRNA screen
  - Brass *et al*, *Science* (2008) 319: 921-6
- 295 genes that had an effect in the Konig siRNA screen
  - Konig *et al.* Cell (2008) 1: 49-60
- The interactors of the siRNA genes



Recall	**Precision**	**in Virion**	**Brass *et al.* siRNA screen**		**Konig *et al.* screen**	
			Genes	**Interactors**	**Genes**	**Interactors**
0.51	0.20	246	46	1064	77	422
0.37	0.29	101	13	441	21	181
0.26	0.36	48	5	212	11	99
0.18	0.41	17	2	99	7	53
0.13	0.47	8	1	49	4	28
0.09	0.47	4	0	25	2	14

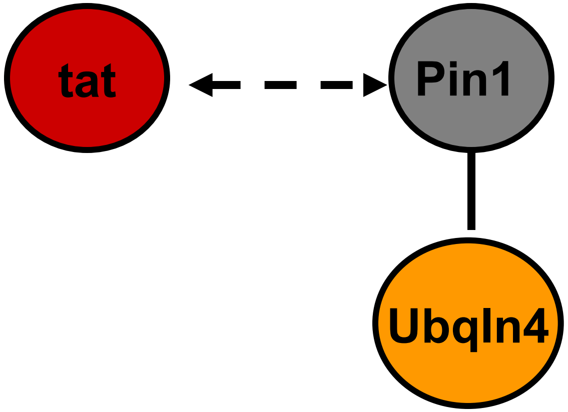
Tat interacts with Pin1

www.cs.cmu.edu/~HIV/hivPPI.html



HIV-1 protein name	Human patner Entrez gene id	Human partner gene symbol	Human partner official name	Random forest score
gag_matrix	5566	PRKACA	"protein kinas	4.34
tat	5970	RELA	"v-rel reticuloe	4.31
gag_matrix	801	CALM1	"calmodulin 1	4.30
env_gp160	801	CALM1	"calmodulin 1	4.22
nef	5566	PRKACA	"protein kinas	4.17
tat	6598	SMARCB1	"SWI/SNF rel	4.12
env_gp120	801	CALM1	"calmodulin 1	4.11
tat	3725	JUN	"jun oncogene	4.10
nef	7157	TP53	"tumor protein	4.10
nef	2534	FYN	"FYN oncoger	4.05
tat	5111	PCNA	"proliferating c	4.02
tat	5590	PRKCZ	"protein kinas	4.00
tat	2071	ERCC3	"excision repa	3.99
tat	2961	GTF2E2	"general trans	3.91
env_gp41	801	CALM1	"calmodulin 1	3.90
rev	1457	CSNK2A1	"casein kinase	3.90
env_gp160	2335	FN1	"fibronectin 1"	3.90
tat	5588	PRKCQ	"protein kinas	3.87
nef	5578	PRKCA	"protein kinas	3.87
nef	801	CALM1	"calmodulin 1	3.86

detected in virion



detected in siRNA screen

Pin1 interacts with and reduces expression of APOBEC3G.

Watashi *JV* (2008) *82: 9928-36*

Conclusions

- ❑ Collected data from multiple biological information sources and encoded as features
- ❑ Developed a model to predict HIV-1,human protein interaction network
- ❑ Features containing human proteome knowledge is highly informative
- ❑ Specific protein interactions are being tested
- ❑ Predictions available

  www.cs.cmu.edu/~HIV/hivPPI.html

  www.hivppi.pitt.edu

Acknowledgements

Yanjun Qi
NEC Laboratories America, Inc

Jaime G. Carbonell
Carnegie Mellon University

Judith Klein-Seetharaman

Carnegie Mellon University
University of Pittsburgh

Thanks to :

-PSB organizers

-National Institutes of Health

-Pittsburgh Center for HIV Protein Interactions

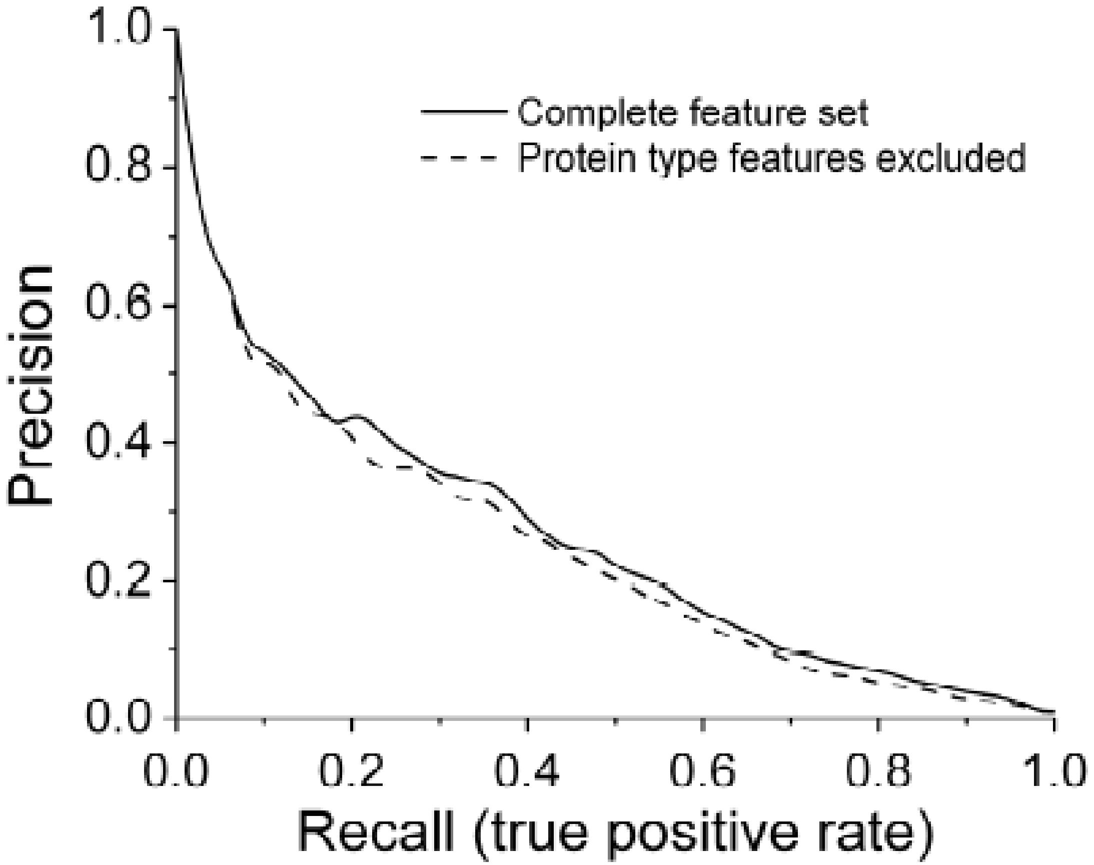
Click Here &rarr Funding Opportunity: PCHPI Collaboration Development Program

Welcome to the website of the Pittsburgh Center for HIV Protein Interactions (PCHPI).

We invite you to explore our web pages and learn more about our center and the biology of HIV. After reading through these pages, if you have any questions or comments or would like to begin a scientific collaboration, please contact the PCHPI coordinator Teresa Brosenitsch. We hope you find these pages helpful and look forward to hearing from you.

❑ Extra slides

Precision/Recall Curve When Protein Type Features Excluded



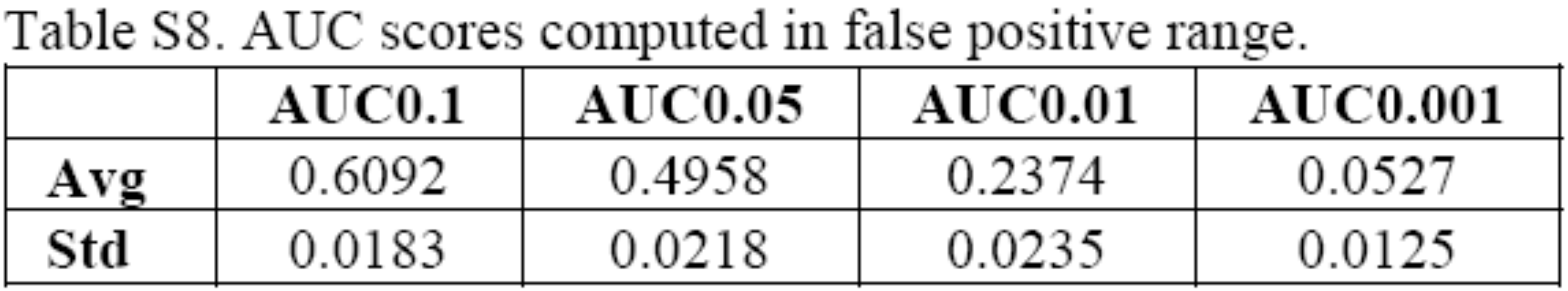
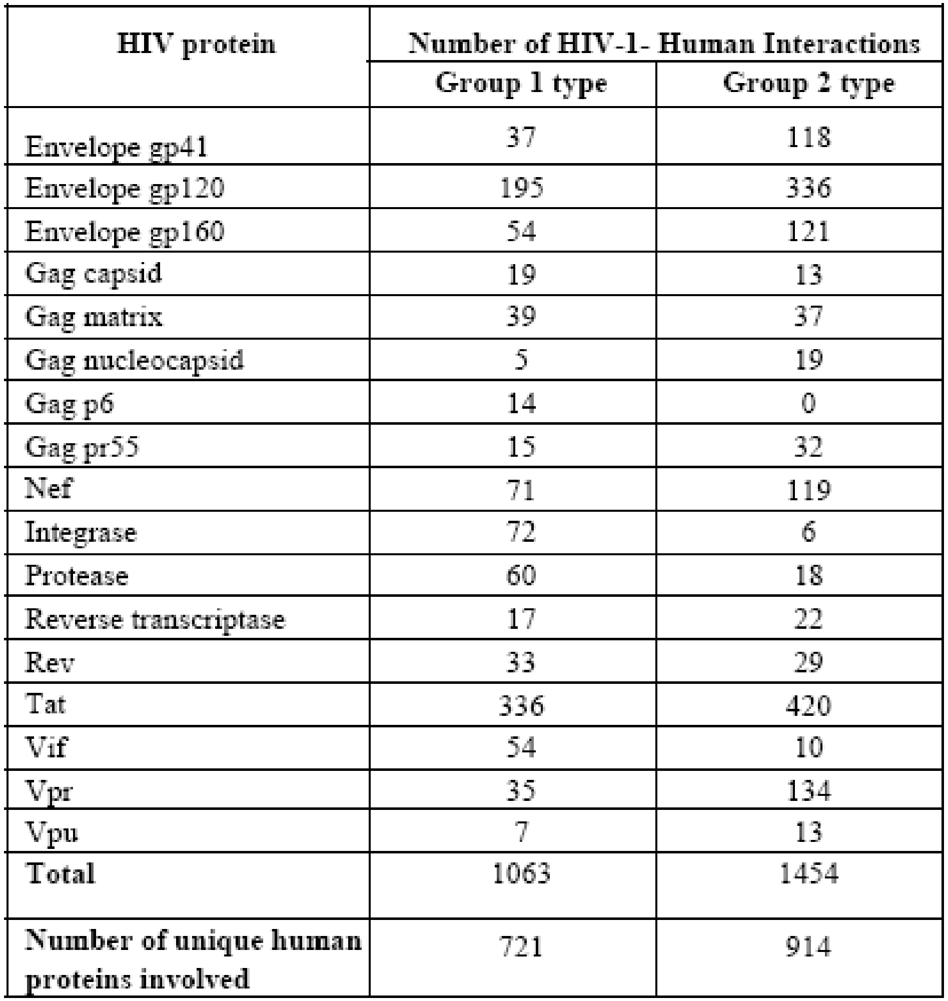


Table S8. AUC scores computed in false positive range.

	AUC0.1	AUC0.05	AUC0.01	AUC0.001
Avg	0.6092	0.4958	0.2374	0.0527
Std	0.0183	0.0218	0.0235	0.0125

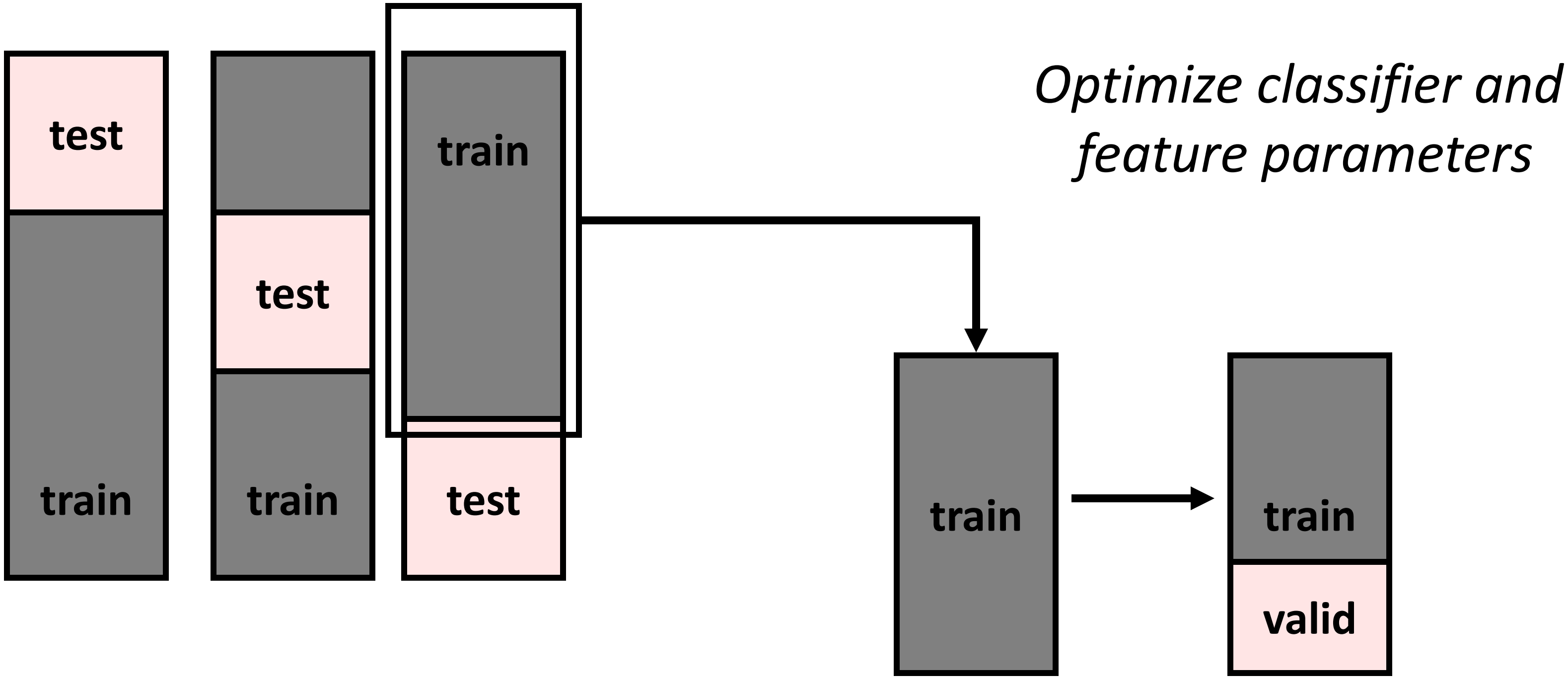
The Interaction Data Counts



HIV protein	Number of HIV-1- Human Interactions	
	Group 1 type	Group 2 type
Envelope gp41	37	118
Envelope gp120	195	336
Envelope gp160	54	121
Gag capsid	19	13
Gag matrix	39	37
Gag nucleocapsid	5	19
Gag p6	14	0
Gag pr55	15	32
Nef	71	119
Integrase	72	6
Protease	60	18
Reverse transcriptase	17	22
Rev	33	29
Tat	336	420
Vif	54	10
Vpr	35	134
Vpu	7	13
Total	1063	1454
Number of unique human proteins involved	721	914

Evaluation Design

1. Randomly select the negative examples from non-interacting pairs
2. Repeated 3-fold cross validation

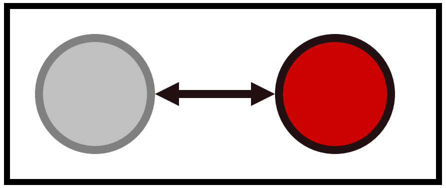


Repeated 10 times.

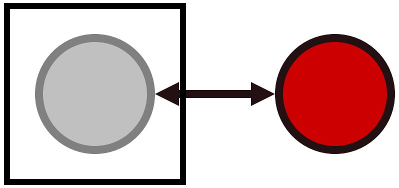
The performance is average of 30 runs.

Features

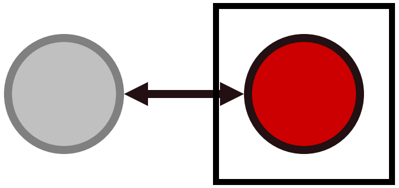
- 35 features calculated for e HIV-1 , human protein pair



8 features specific to HIV-1, human protein pair



10 features specific to human protein



17 features specific to human protein