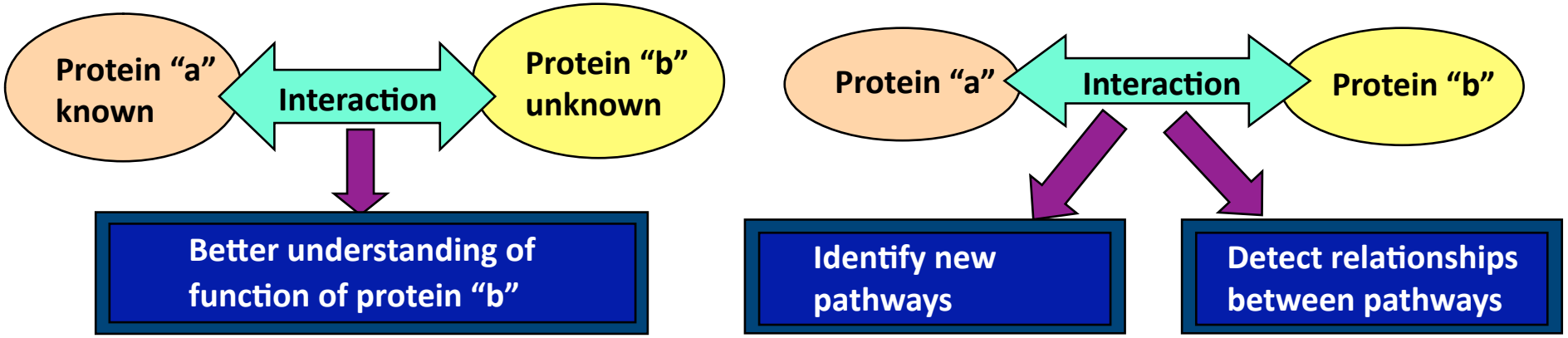
Semi-Supervised Multi-task Learning for Predicting Interactions between HIV-1 and Human Proteins

Yanjun Qi[1], Oznur Tastan[2], Jaime G. Carbonell[2], Judith Klein-Seetharaman[2], Jason Weston[3]

[1] Machine Learning Department, NEC Labs America.
[2] School of Computer Science, Carnegie Mellon University
[3] Google Research, NY

Importance of Protein Interactions



- Need comprehensive identification of Protein-Protein Interactions (PPI)
  - To systematically define proteins' functions
  - To decipher molecular mechanisms underlying given biological functions
  - Essential for diseases studies & drug discoveries

Previous Approaches

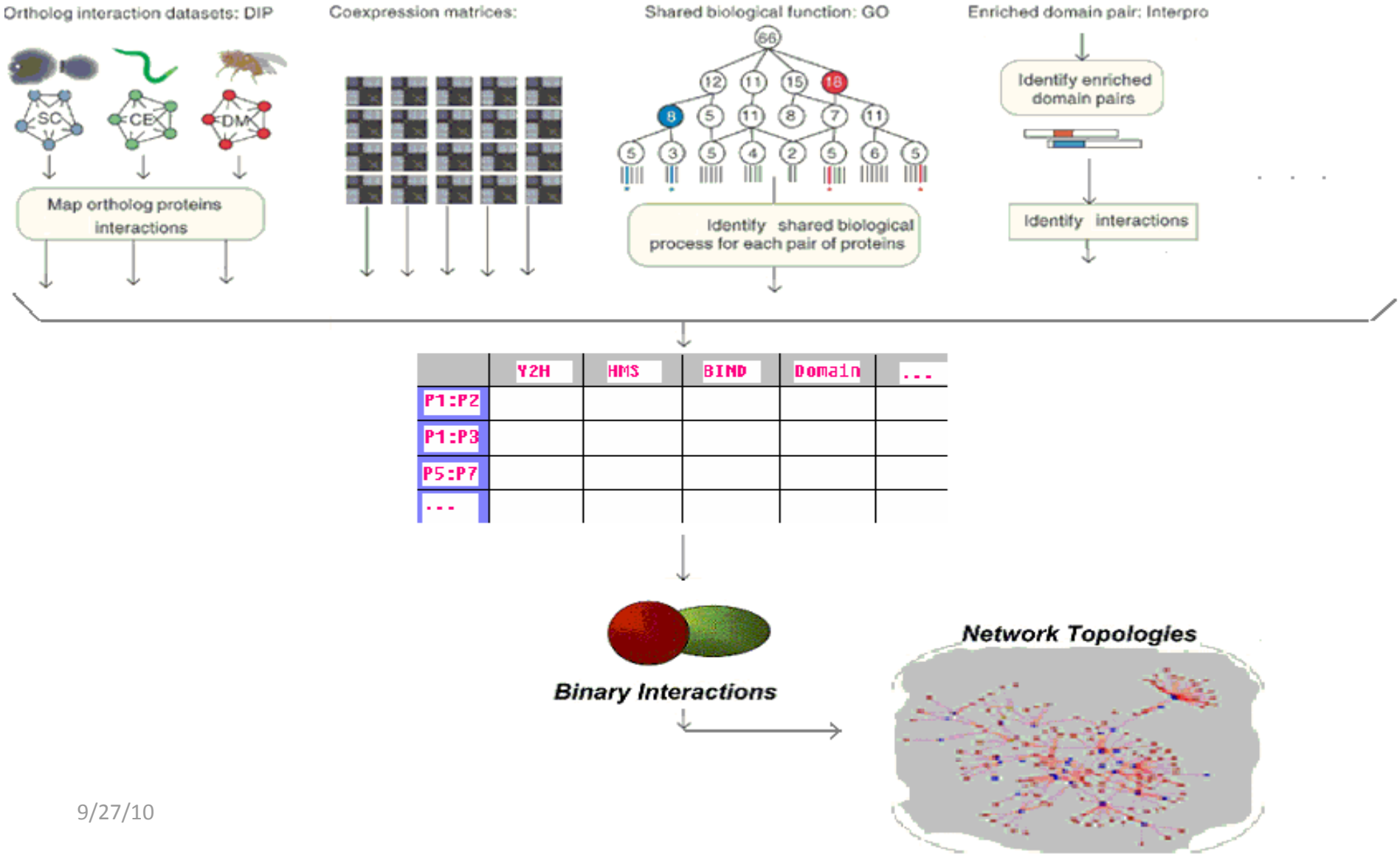
- Experimental:
  - Direct large scale experimental data
    - High false-positive and false-negative rate,
    - Incomplete, with majority remains to be discovered, especially for human
    - Surprisingly small overlap among different sets

- Computational:
  - Combine direct evidence and other implicitly related biological information as features
    - Example: If two proteins are co-expressed, they may interact.

➔ Large portion of the PPIs still missing or noisy !

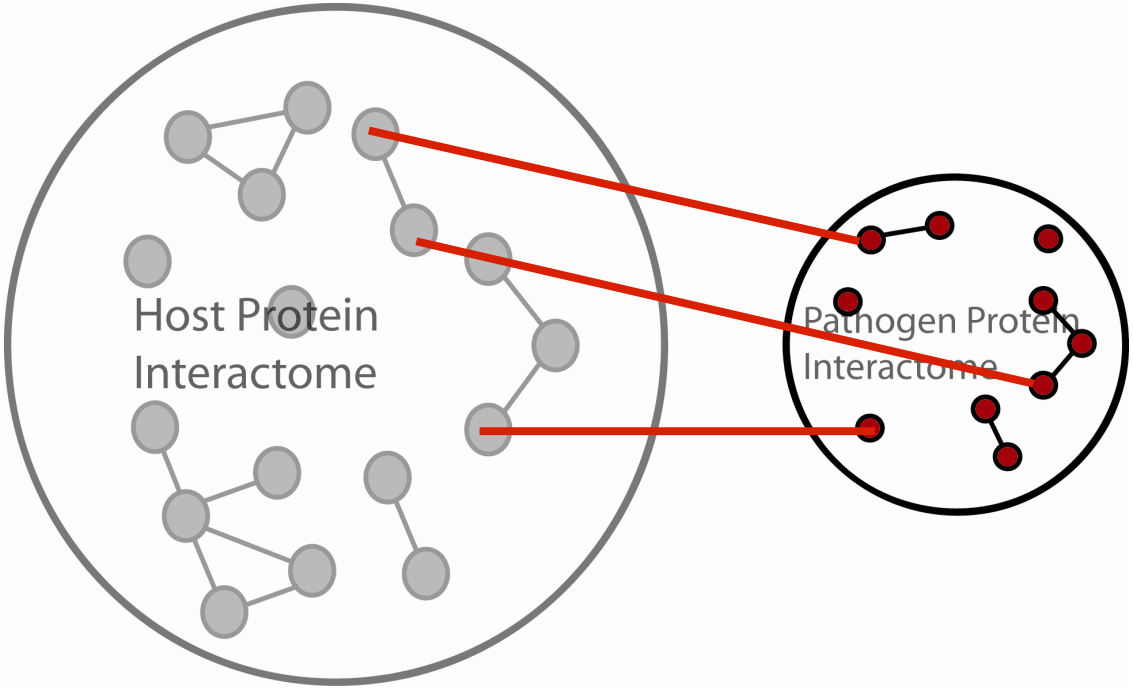
Background

Computational PPI Prediction through Data Fusion



Our Aim

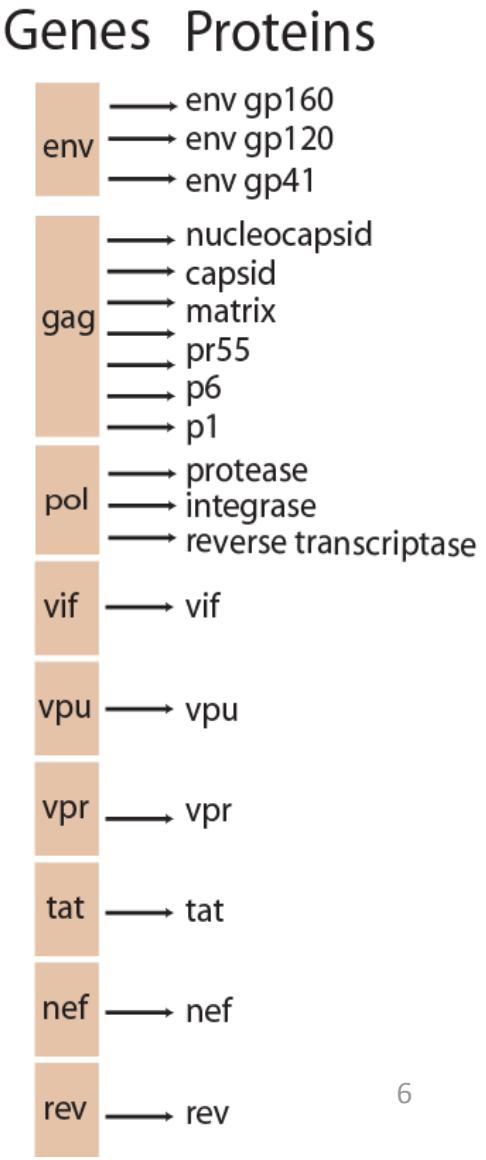
Predict novel direct physical interactions between HIV-1 and human proteins



Critical for designing strategies to get HIV-1 under control !

HIV-1: Human Immunodeficiency Virus-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



Previous Work: Supervised Classification

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

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

Each feature summarizes a biological information

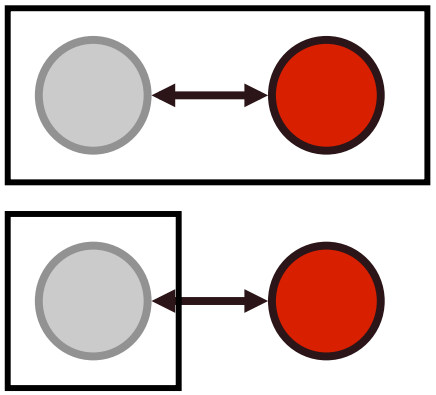
- Given data learn *a function* that would *map feature space into one of the two classes:* $f : X \rightarrow Y$

- State-of-the-art performance: Random forest (Tastan et al. (PSB 2009))

Target Problem: Features

- Features and reference sets are from paper:
  - Tastan et al. (PSB 2009)

- 18 features calculated for each HIV-1 , human protein pair

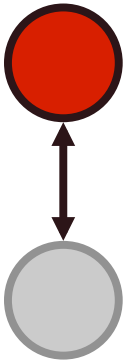


10 features specific to HIV-1, human protein pair

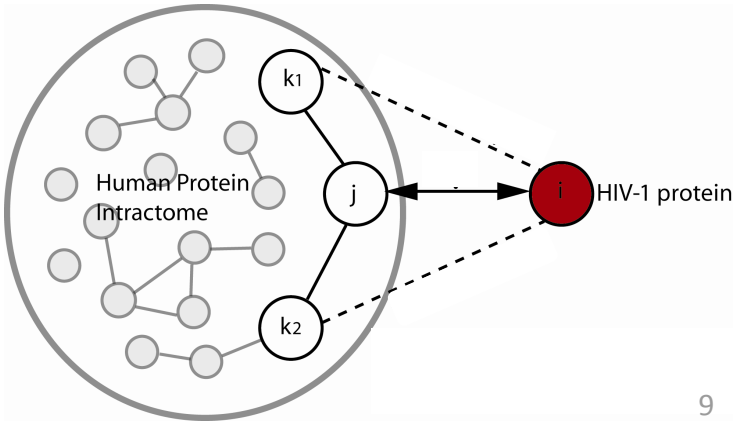
8 features of human protein

Target Problem: 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



- ELM-ligand feature (1)
- Human PPI interactome features (8)
  - Similarity of HIV-1 protein to human protein's interaction partner (5)
  - Topological properties of human interaction graph (3)



Target Problem: Data Situation

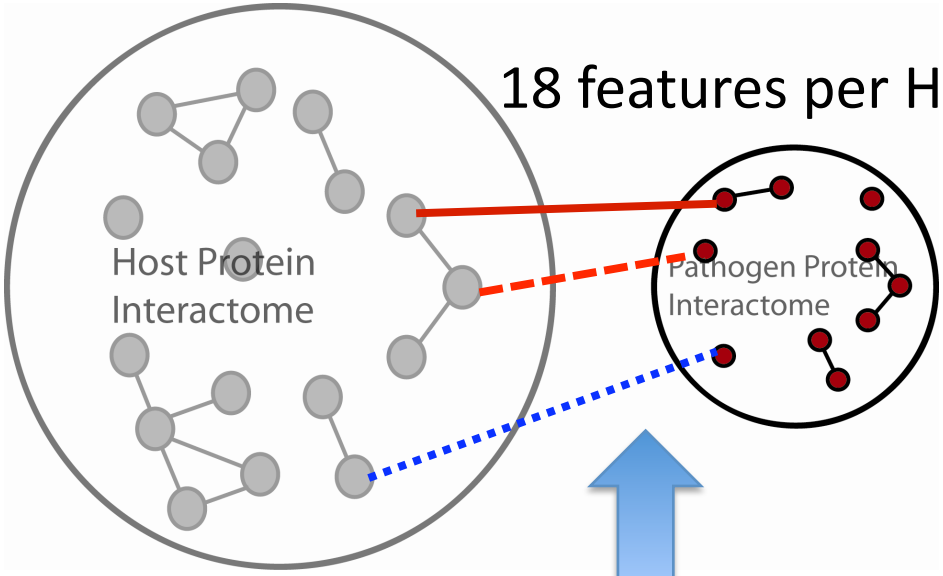
- Partial positive labels from NIAID database
  - ~2100 protein interaction pairs (extracted from literature)
  - Not enough evidence supporting reliabilities (partial positive)
  - Each associated with keywords
    - (e.g. “interacts”, “binds”, “up-regulates”, ...)
    - Some strong indication, some weak

- Positive labels annotated by HIV experts
  - 361 possible pairs given to experts
  - 158 out of above annotated as interaction (positive)

Target Problem: Data Situation

- No negative (not interacting) set available
- Highly skewed class distribution
  - Much more non-interacting pairs than interacting pairs

20873 Human Proteins



17 HIV Proteins

Positive	Partial Positive	Remaining
158	2119	~352000

Method: Multi-Tasking with Semi-Supervised Auxiliary Task

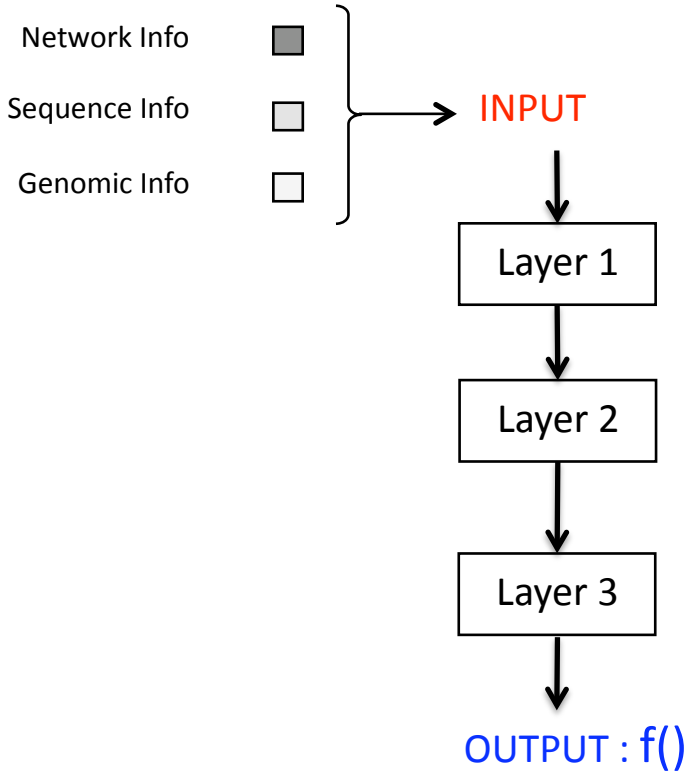
- Multi-tasking two tasks
  - *Supervised main* PPI *classification task*
  - *Semi-supervised auxiliary task* with partial labels
    - (1) Classification
    - (2) Ranking
    - (3) Embedding

- Add auxiliary task as a regularizer on the supervised MLP

$$\sum_{i=1}^{L} \ell(f(x_i), y_i) + \lambda \quad \text{Loss (Auxiliary Task)}$$

Main: (0) Supervised PPI Classification (MLP)

- Main task: *Supervised PPI classification*
  - Multiple Layer Perceptron (MLP)
  - Binary classification (interact "1", not interact "-1")
  - Train with stochastic gradient descent
  - Toward hinge loss

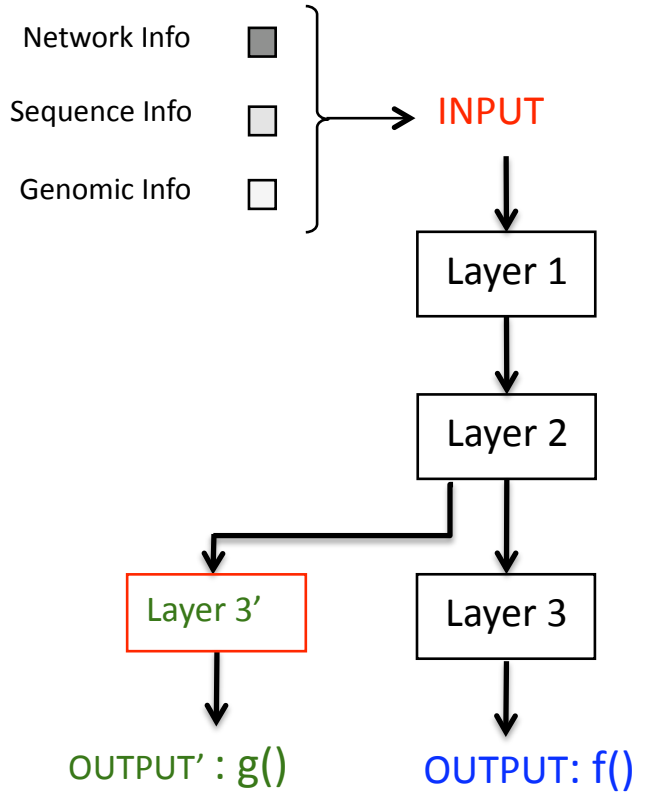


$$\sum_{i=1}^{L} \ell(f(x_i), y_i) = \sum_{i=1}^{L} \max(0, 1 - y_i f(x_i))$$

•Assuming labeled data (x_i, y_i), $i = 1, \ldots, L$

Auxiliary: (1) Classification with Partial Labels (SMLC)

- Auxiliary task: *Pseudo-Supervised classification*
  - MLP shares layers with main task
  - Binary classification
    - partial positive "1"
    - not interact "-1"
  - Toward hinge loss with pseudo-labels

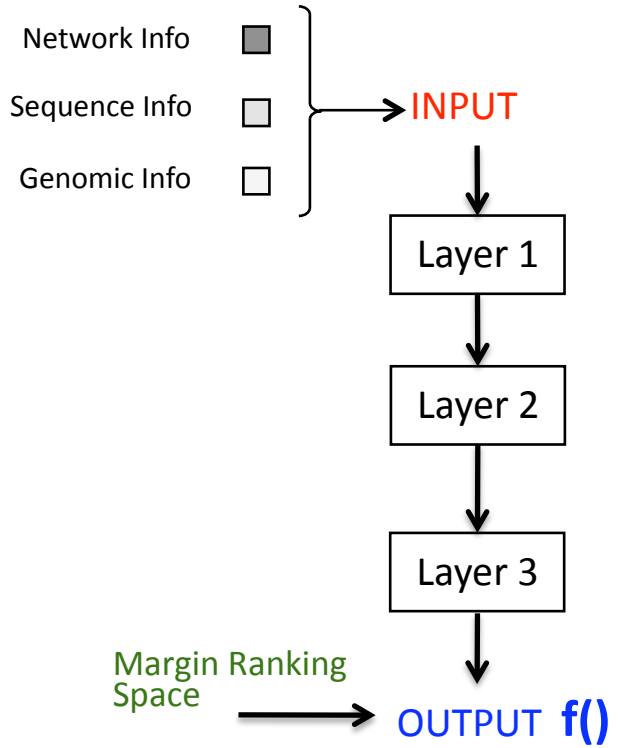


$$\text{Loss (Auxiliary Task)} = \sum_{j=L+1}^{L+U} \max(0, 1 - y'_j g(x_j))$$

•Assuming partial Labeled data (x_i , y'$_i$), $i = L+1 , \ldots, L + U$

Auxiliary: (2) Ranking with Partial Labels (SMLR)

- Auxiliary task: *Pseudo-Supervised ranking*
  - MLP shares the same network as main task
  - Preference ranking
    - Rank "partial positive" more likely than "negative"
  - Toward margin rank loss



$$\text{Loss(Aux.)} = \sum_{p \in P} \sum_{n \in N} \max\left(0, 1 - f(x_p) + f(x_n)\right)$$

P the set of partial positives and N the set of negative examples

Auxiliary: (3) Embedding with Partial Labels (SMLE)

- Embedding: Given data $x_1, \ldots, x_P$, find an embedding function $f(x_i)$ by minimizing pairwise distance margin loss

$$L(f_i, f_j, W_{ij}) = \begin{cases} ||f_i - f_j||_1 & \text{if } W_{ij} = 1, \\ \max(0, m - ||f_i - f_j||_1) & \text{if } W_{ij} = 0 \end{cases}$$

  - W matrix should be supplied in advance and specify the similarity between examples x_i and x_j

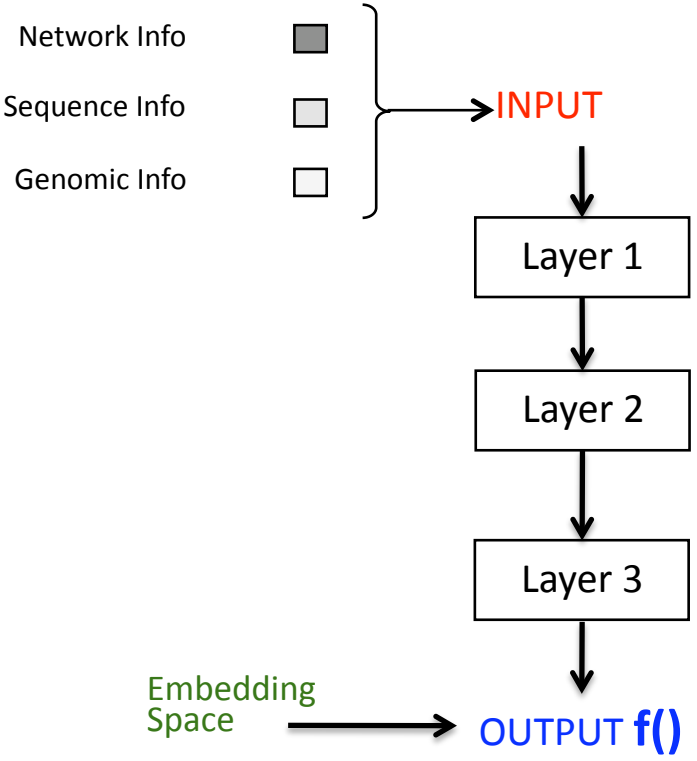
- Motivation: embedding could uncover hidden cluster structure within the data based on partial examples' similarities to the remaining examples

- We use *partial labels* to *build matrix W* for embedding
  - *Wij = 1 if both examples from partial positive set*
  - *Wij = 0 if one partial positive example and the other a negative example*

Auxiliary: (3) Embedding with Partial Labels (SMLE)

✓Auxiliary task: *Pseudo-Embedding*

- MLP shares the same network as main task
- Embedding adds an extra distance layer on output
- Motivation: embedding could improve accuracy by helping data clusters get similar labels



- The whole network optimize toward the loss

$$\sum_{i=1}^{L} \ell(f(x_i), y_i) + \lambda \sum_{i,j=1}^{L+U} L(f(x_i), f(x_j), W_{ij})$$

Method: *Semi-Multi-Embed* **Algorithm (SMLE case)**

Input:

- Labeled data (x_i, y_i), $i = 1, \ldots, L$
- Partial Labeled data x_i, $i = L+1, \ldots, L + P$

Repeat:

- Pick a random labeled example (x_i, y_i)
- Make a gradient step to optimize $l(f(x_i), y_i)$
- Pick a random partially labeled example x_p
- Pick another random example x_q where $W_{pq} = 1$
- Make a gradient step to optimize $\lambda L(f(x_p), g(x_q), 1)$
- Pick a random partially labeled example x_m
- Pick another random example x_n where $W_{mn} = 0$
- Make a gradient step to optimize $\lambda L(f(x_m), g(x_n), 0)$

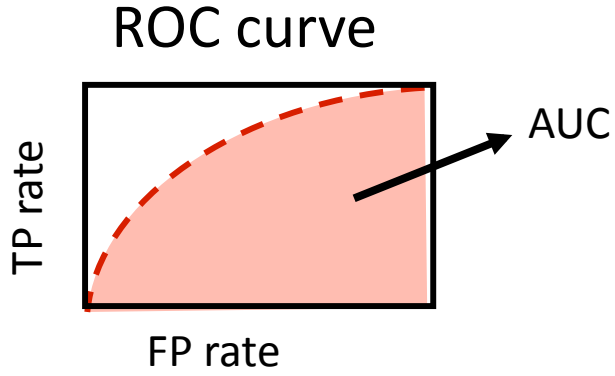
Until stopping criteria is met

Evaluation: Performance Measures

- The Mean Average Precision (MAP)
  - Mean of the average precisions where each average precision is calculated when recall increases

- Precision-Recall breakpoint (PRB)
  - Value where precision is equal to recall

- Area Under the Receiver Operating Curve (AUC):



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

Evaluation: Performance Comparison

- 20 times of randomly repeated 5 folds cross validation
- Compare: SMLC, SMLR, SMLE, MLP and RF
  - Torch for SMLC, SMLR, SMLE, and MLP
  - RF Berkely package for RF

METHOD	AUC 50	MAP	PRB	AUC
SMLC	0.277	0.263	0.312	0.905
SMLR	**0.31**	0.268	0.311	**0.919**
SMLE	0.309	**0.277**	**0.326**	0.908
RF	0.199	0.135	0.18	0.893
RF-P	0.23	0.213	0.281	0.896
MLP	0.204	0.197	0.257	0.859
MLP-P	0.229	0.21	0.282	0.893

Evaluation: Validation

- Statistics of overlaps between top predicted human partners to those found in
  - *(i)(Ott, 2008) virion screen list,*
  - *(ii) Combined 4 siRNA screens (Brass et al., 2008; König et al., 2008; Yeung et al., 2009; Zhou et al., 2008)*

Predicted Interactions	2434
Interaction Confirmed by Partial Positive	223
Novel Interactions	2172
Human Gene in Predicted Interactions	721
Confirmed with Virion (316 genes)	61
Combined Four siRNA (1049 genes)	72

All experts labels / partial labels / top predicted interactions are shared online !
http://www.cs.cmu.edu/~qyj/HIVsemi

Conclusion

- Semi-supervised multitasking is promising for HIV-Human PPI prediction task
- Easily extendable for incorporating other auxiliary information, such as large-scale noisy experimental PPI evidence
- Easily extendable for other PPI tasks, such as PPI predictions in yeast or human

Thanks !

All experts labels / partial labels / top predicted interactions are shared online !
http://www.cs.cmu.edu/~qyj/HIVsemi