Refining Literature Curated Protein Interactions using Expert Opinions

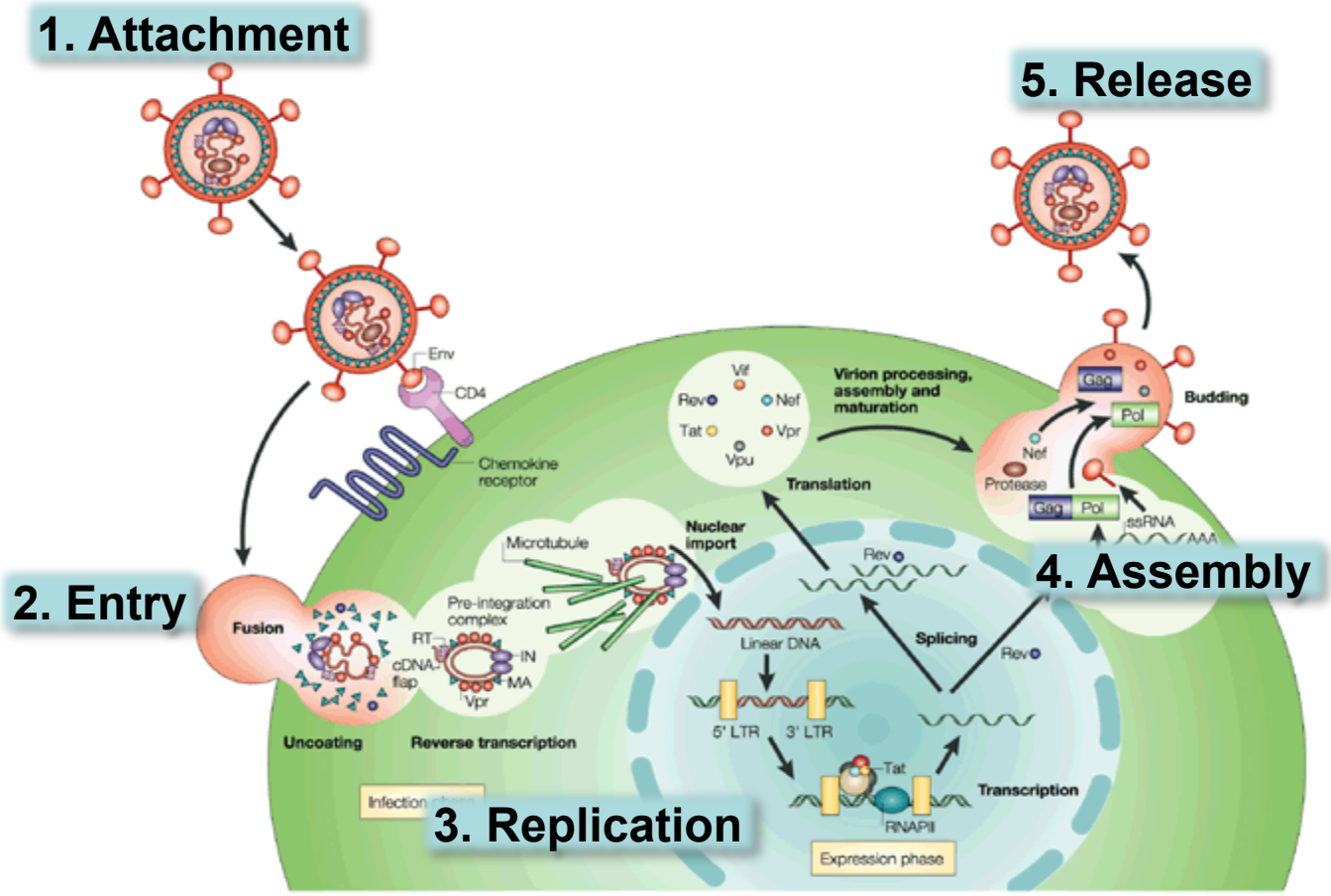
Oznur Tastan
Department of Computer Engineering, Bilkent University,
Ankara, Turkey

Yanjun Qi
Department of Computer Science, University of Virginia,
Charlottesville, VA, USA

Jaime G. Carbonell
LTI, School of Computer Science, Carnegie Mellon University,
Pittsburgh, PA, USA

Judith Klein-Seetharaman
Division of Metabolic and Vascular Health, University of Warwick,
Warwick, Coventry, UK

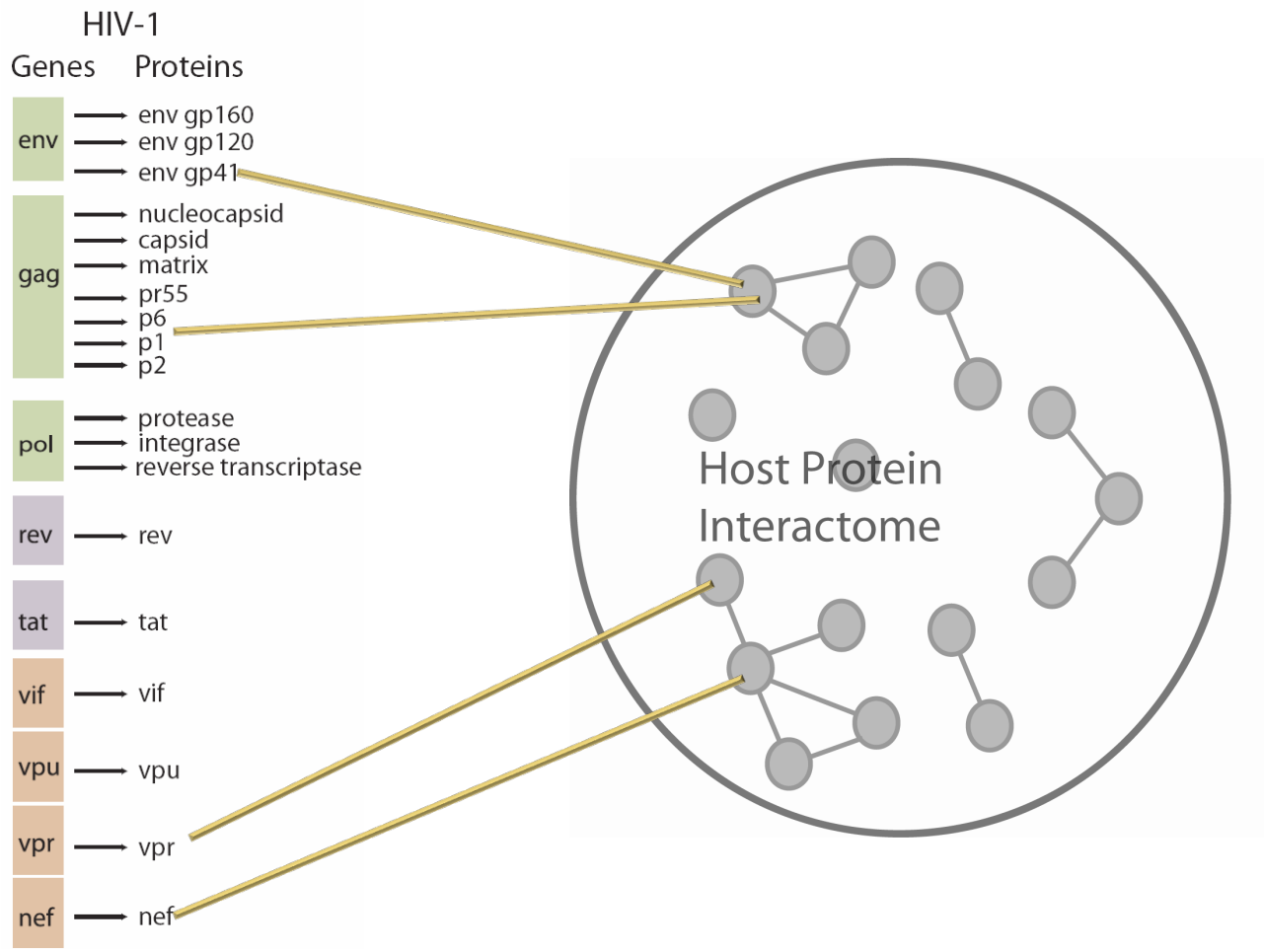
What motivated this study?



Peterlin and Trono *Nature Rev. Immu.* 3. (2003)

Host machinery is essential in the viral life cycle.
Established through host-virus protein interactions.

Predicting HIV-1,human protein-protein interactions



HIV-1, host protein-protein interaction data

NCBI

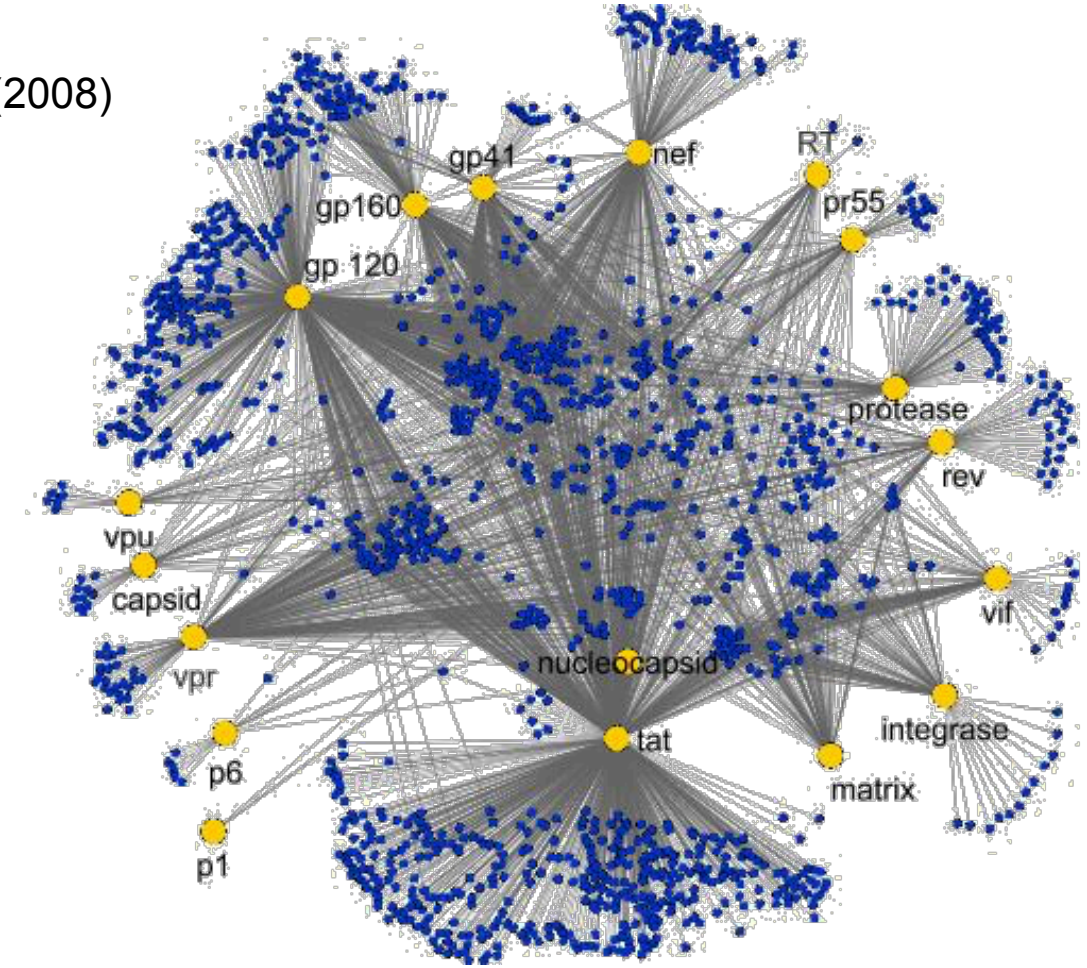
HIV-1, Human Protein Interaction Database
National Institute of Allergy & Infectious Diseases

Fu W *et al.* NAR 37:D417-22 (2009)
Ptak RG *et al.* AIDS. 24(12):1497-502 (2008)

2589 interactions
1448 human proteins

Which of these interactions are direct physical interactions?

How confident are we in each interaction being a direct physical interaction?

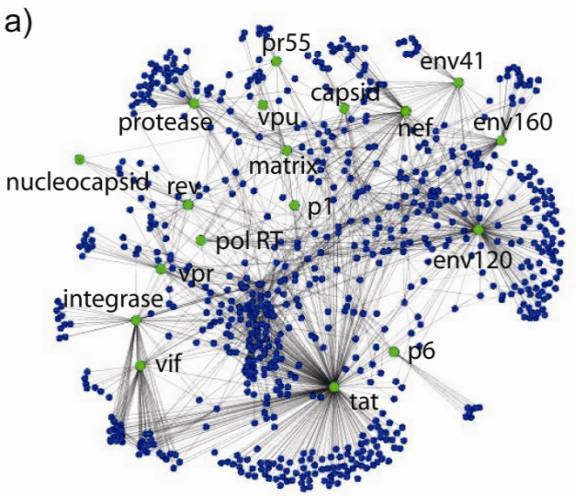


HIV-1 human protein interactions

Keywords: “Nef binds p61HCK”

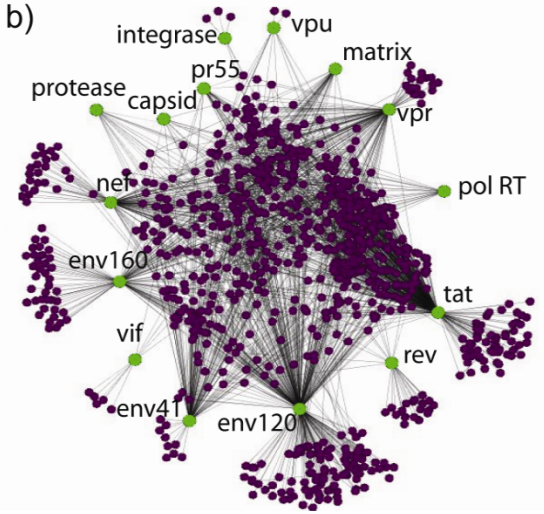
- Keywords of more likely direct interactions

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



- Keywords of indirect interactions

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



HIV-1 protein

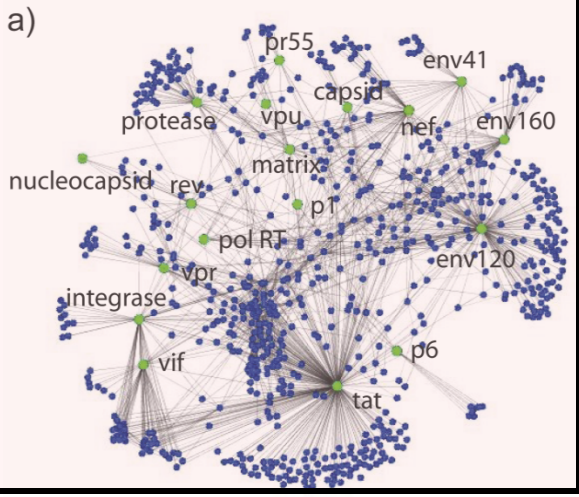
Human protein

HIV-1 human protein interactions

Keywords: “Nef binds p61HCK”

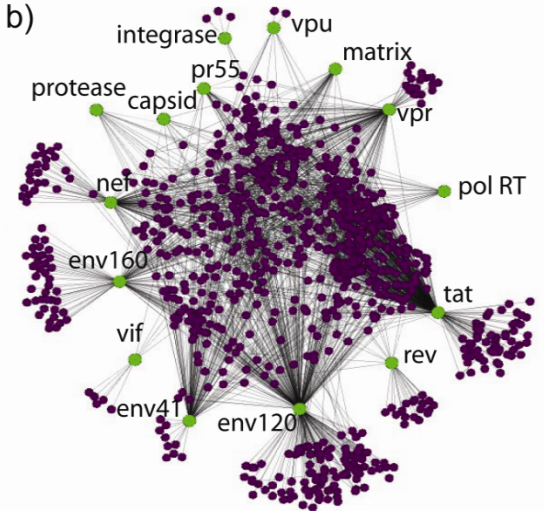
- Keywords of more likely direct interactions

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



- Keywords of indirect interactions

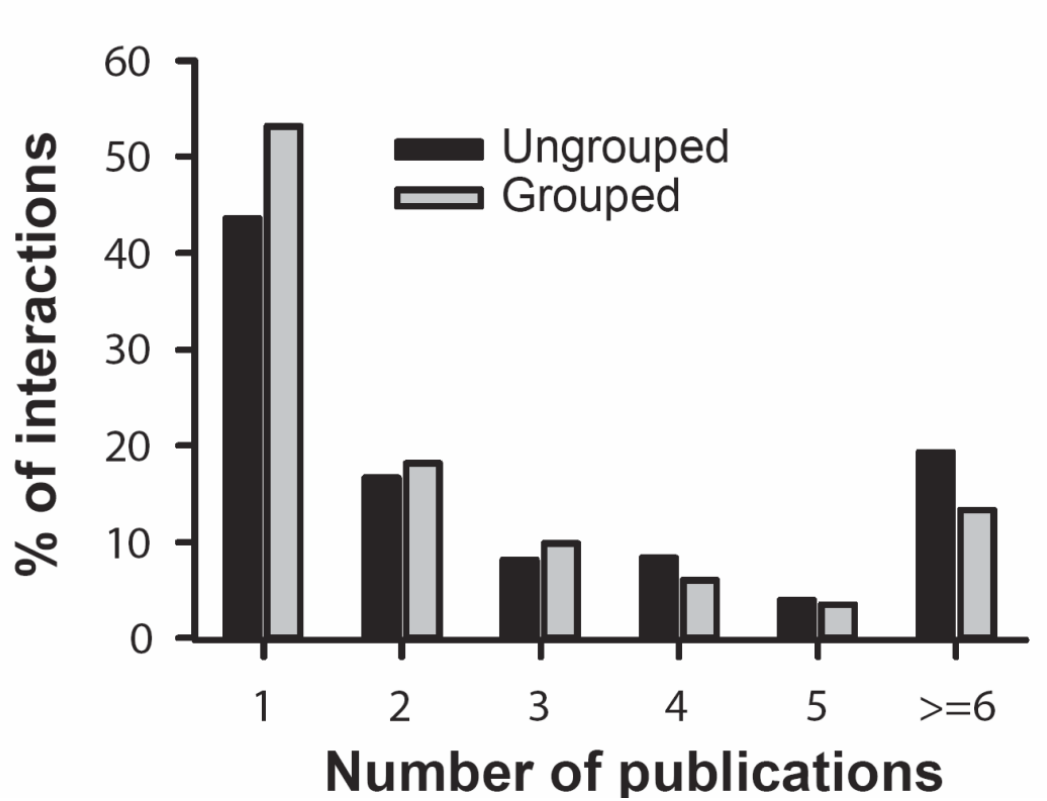
activated by, activates, antagonized by, antagonizes, associates with, causes accumulation of, co-localizes with, competes with, cooperates with, etc



HIV-1 protein

Human protein

Support of interactions



Majority of the interactions are supported by single publication!

Subsetting high-quality interaction data is challenging

- Many literature curated databases offer details on
  - the experimental techniques that found the interaction
  - the publications reporting it
  - occasionally a score based on several predefined parameters

- Yet, subsetting for high quality set of interactions is a challenge

Many techniques to detect PPIs experimentally

- There is a long list of techniques used to detect PPIs,
  - Affinity Capture-Luminescence
  - Affinity Capture-MS
  - Biochemical Activity
  - Co-crystal Structure
  - Co-fractionation
  - Co-localization
  - Co-purification
  - FRET
  - Two-hybrid
  - ….
- The strength of the evidence depends on how the experiment is conducted in what conditions, the properties of the proteins, etc

Ask HIV-1 experts

Do you think there is enough evidence to conclude the two proteins physically directly interact?

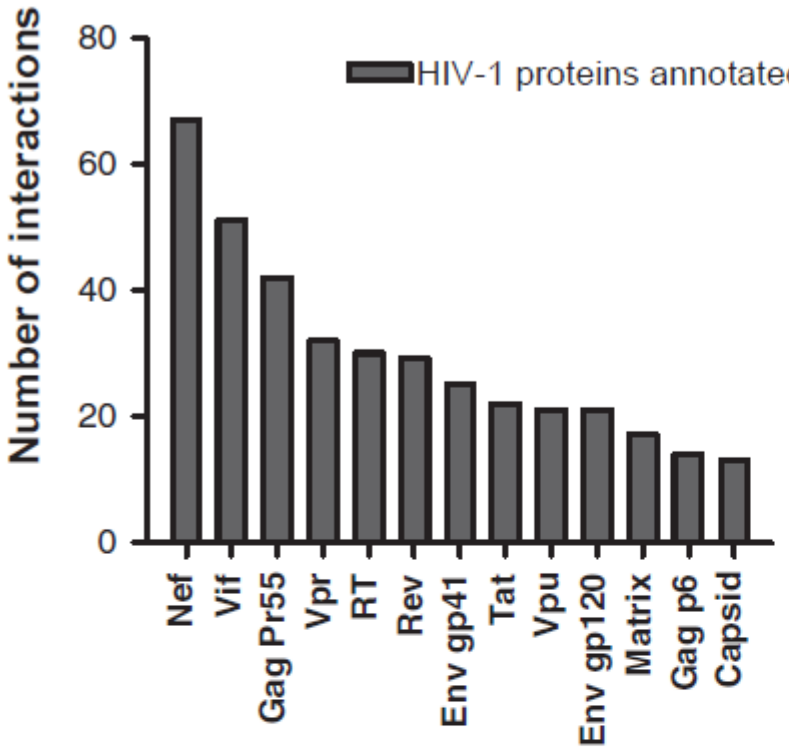
Interactions
&
Hyperlinks to publications

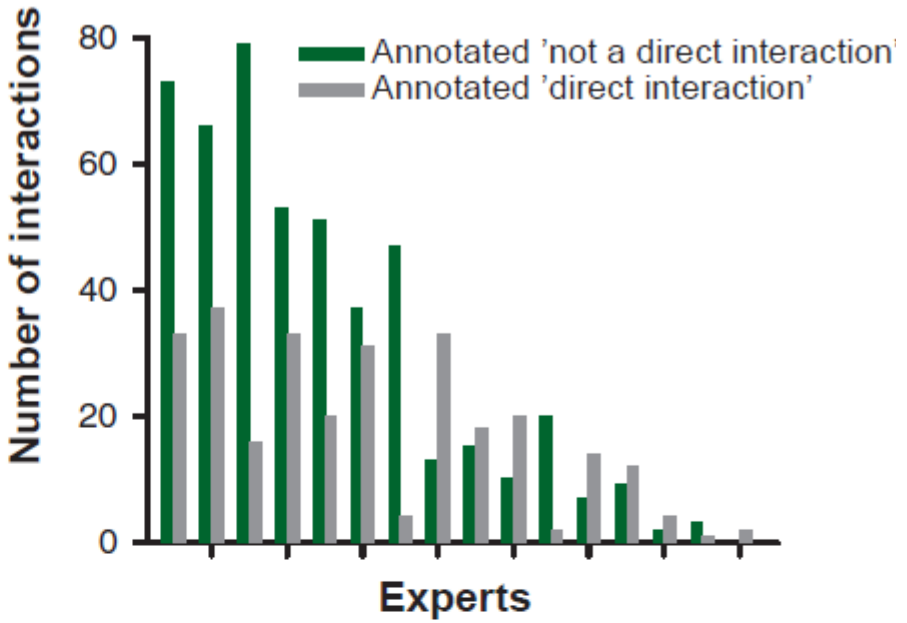
yes

no

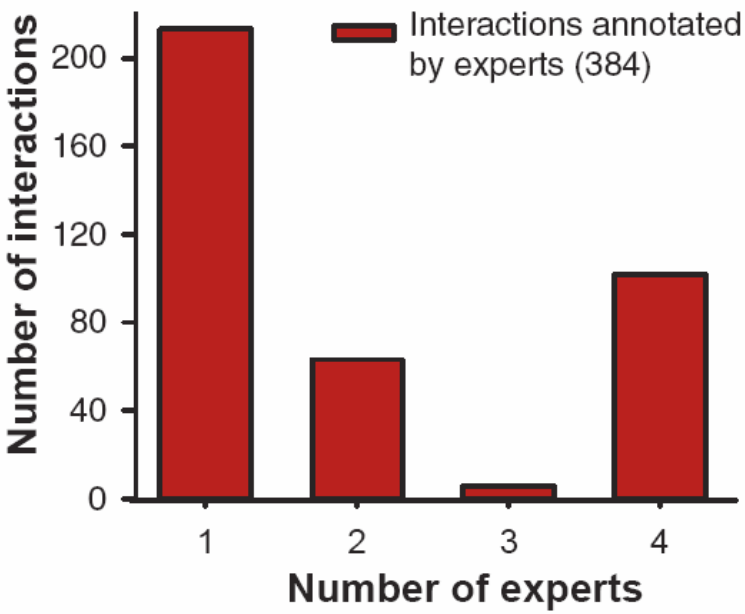
- Experts were HIV-1 biologists:
  - 15 professors well known in the field, 1 PhD student
- Experts are only asked interactions of the viral proteins that they are expert of.

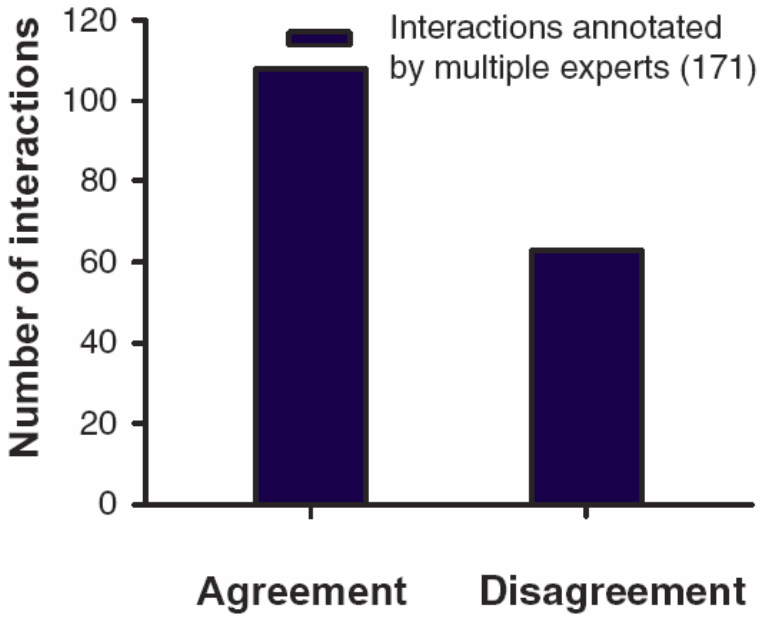
Acquired labels





Experts disagree





- Only partial evidences present
- Different stringency levels, biases

Example: 'vpu stabilizes catenin'
3 expert thinks it is a direct interaction
1 expert disagrees

Estimating the most probable label

- Given *multiple expert opinions on an interaction*, what is the most probable label and *the confidence* in the label?

- Introduce *expert labeling accuracy* to be able to account for subjectivity, bias of experts.

Expert labeling accuracy

- Let's consider N literature reported protein-protein interactions
- Let $y_i \in \mathcal{Z}$ indicate the true and hidden label for the i^{th} PPI, where $\mathcal{Z} =$ {"direct physical interaction" or "not"}
- Expert labels $y_i = \{y_i^1, y_i^2, \ldots, y_i^M\}$ provided by M different experts.
- Similiar to Raykar *et al.* biased coin model (Raykar et al. JLMR 2010) define expert j labeling accuracy for the label type z:

$$\mathbf{P}\left(y^j = z \,|\, y = z\right) = \theta_z^j$$

The probability of the label type

The probability of a label type for the interaction:

$$g_i(z \mid \Theta) \equiv \mathbf{P}\left(y_i = z \mid y_i^1, y_i^2, \ldots, y_i^M, \Theta\right) \propto \prod_{j=1}^{M} \mathbf{P}\left(y_i^j \mid y_i = z, \Theta\right) \times \mathbf{P}\left(y_i = z \mid \Theta\right)$$

$$= p_i^z \times [\theta_z^j]^{h(y_i^j = z)} \times [(1 - \theta_z^j)]^{1 - h(y_i^j = z)}$$

Most probable label for the interaction given the expert opinions:

$$\hat{y}_i = \arg\max_{z \in \mathcal{Z}} g_i(z \mid \Theta)$$

The uncertainty of this label:

$$\hat{u}_i(\hat{y}_i) = 1 - \mathbf{P}\left(y_i = \hat{y}_i \mid y_i, \Theta\right)$$

Estimating expert labeling accuracies

Estimate the parameters Θ through maximum likelihood estimation (MLE):

$$\hat{\Theta}^{\text{mle}} = \arg\max_{\theta} \mathcal{L}(\mathcal{D} \,|\, \Theta)$$

where $\mathcal{D} = \{(y_i^1, y_i^2, \ldots, y_i^M)\}_{i=1,\ldots,N}$

The log-likelihood of the observed expert opinions:

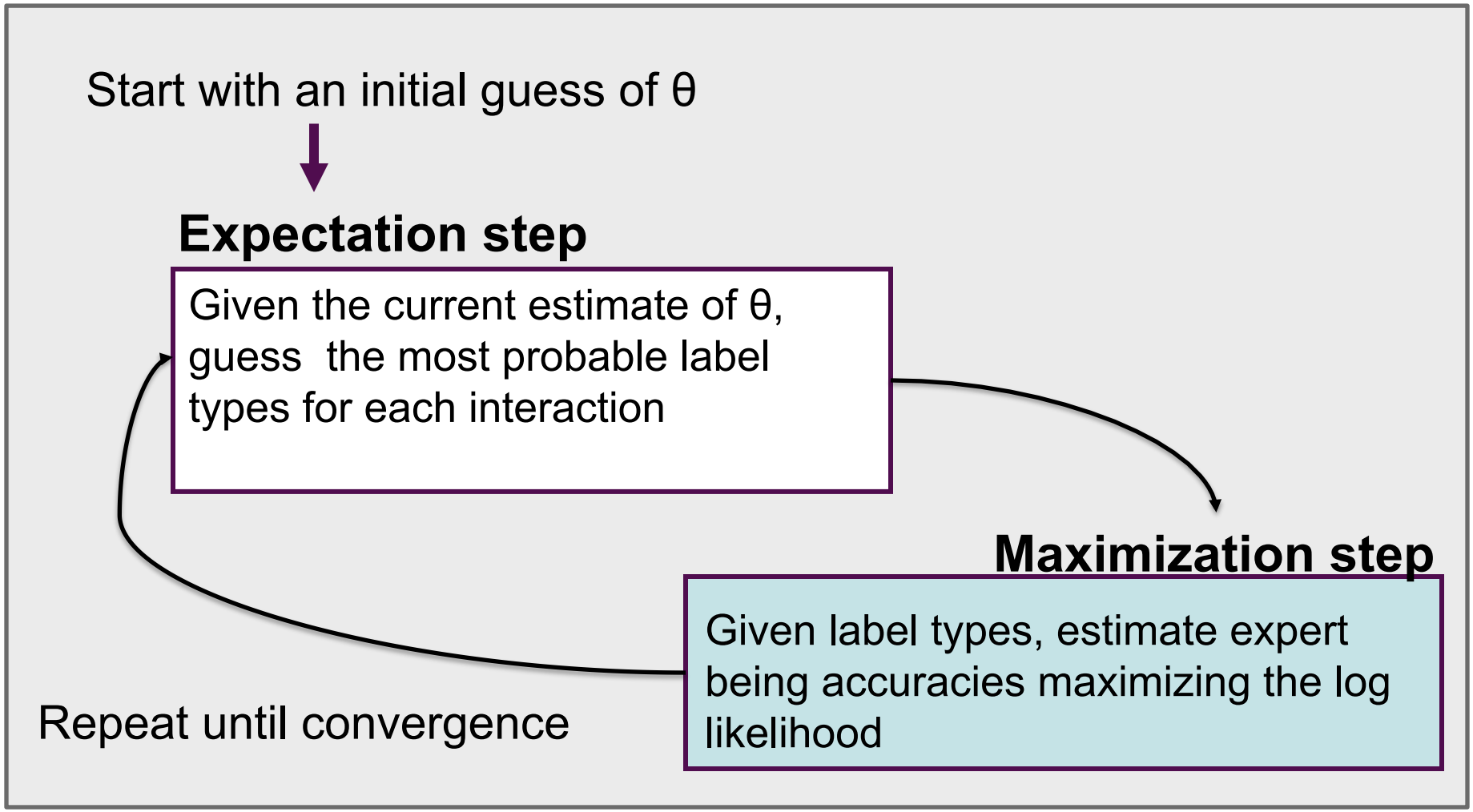
$$\mathcal{L}(\mathcal{D} \,|\, \Theta) = \sum_{i=1}^{N} \log \mathbf{P}\left(y_i \,|\, \Theta\right) = \sum_{i=1}^{N} \log \sum_{z=0}^{1} \mathbf{P}\left(y_i \,|\, y_i = z, \Theta\right) \times \mathbf{P}\left(y_i = z \,|\, \Theta\right)$$

We assume decisions by the experts are conditionally independent given the true label:

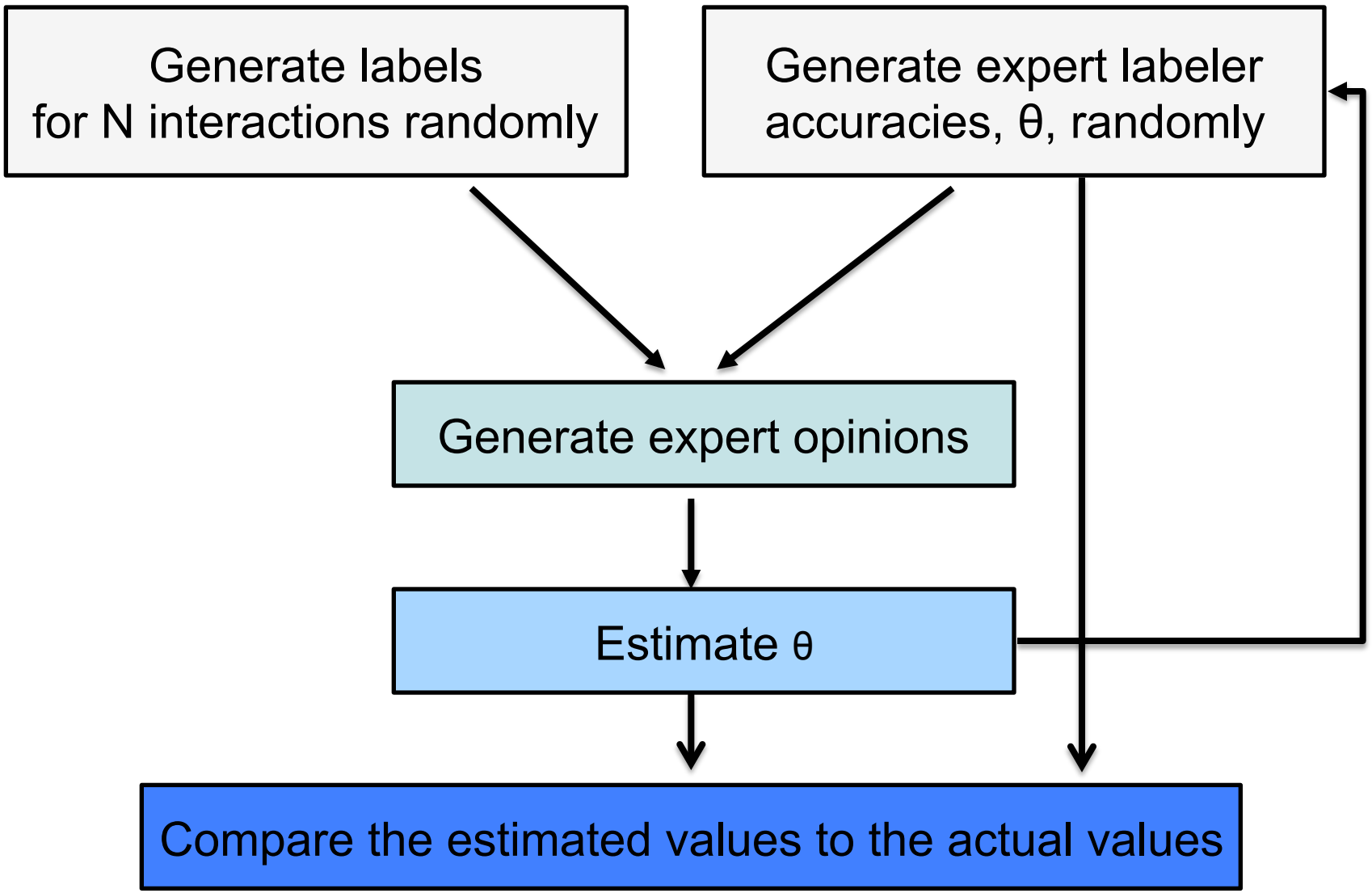
$$\mathcal{L}(\mathcal{D} \,|\, \Theta) = \sum_{i=1}^{N} \log \sum_{z=0}^{1} \left(\prod_{j=1}^{M} \mathbf{P}\left(y_i^j \,|\, y_i = z, \Theta\right) \mathbf{P}\left(y_i = z \,|\, \Theta\right) \right)$$

Finding MLE of labeler accuracies

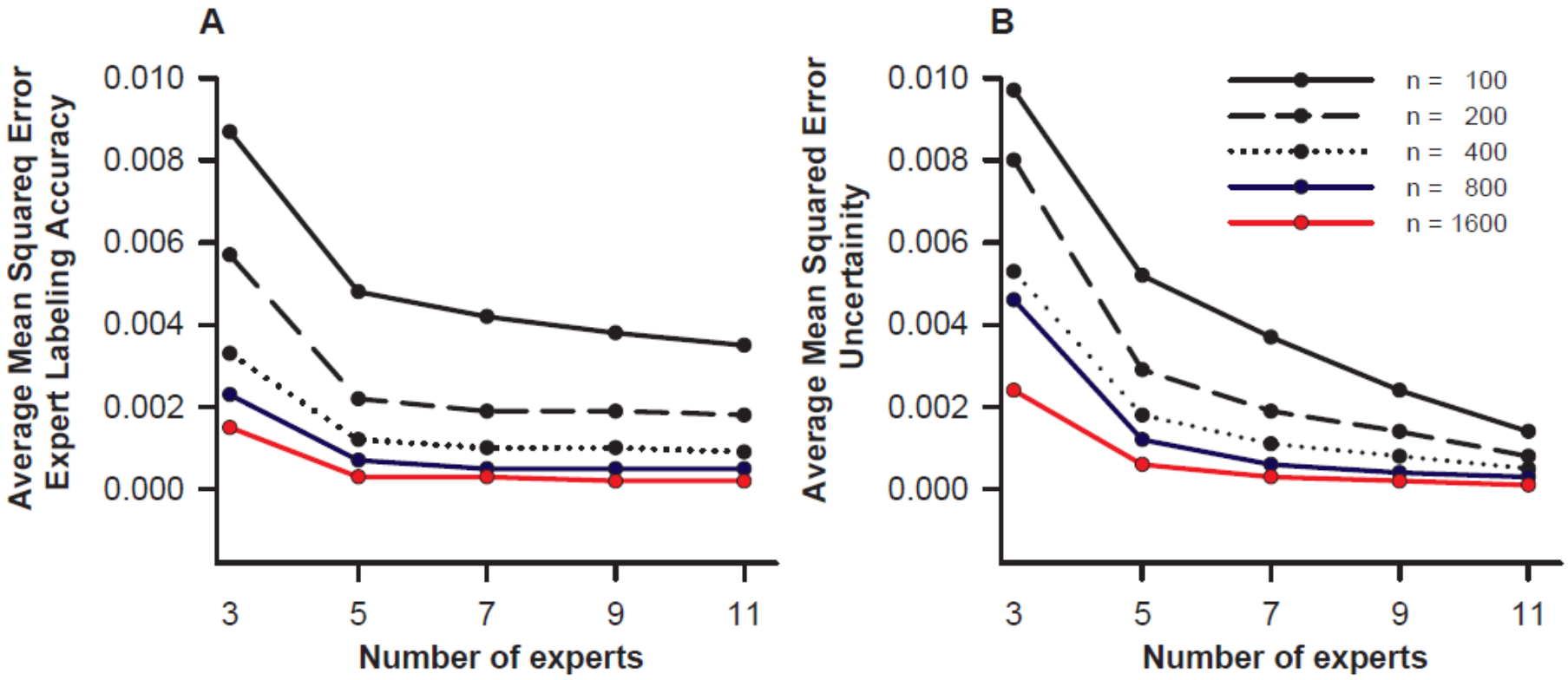
- Expectation-maximization (Dempster et al *J.R. Stat.Soc.* 1977)



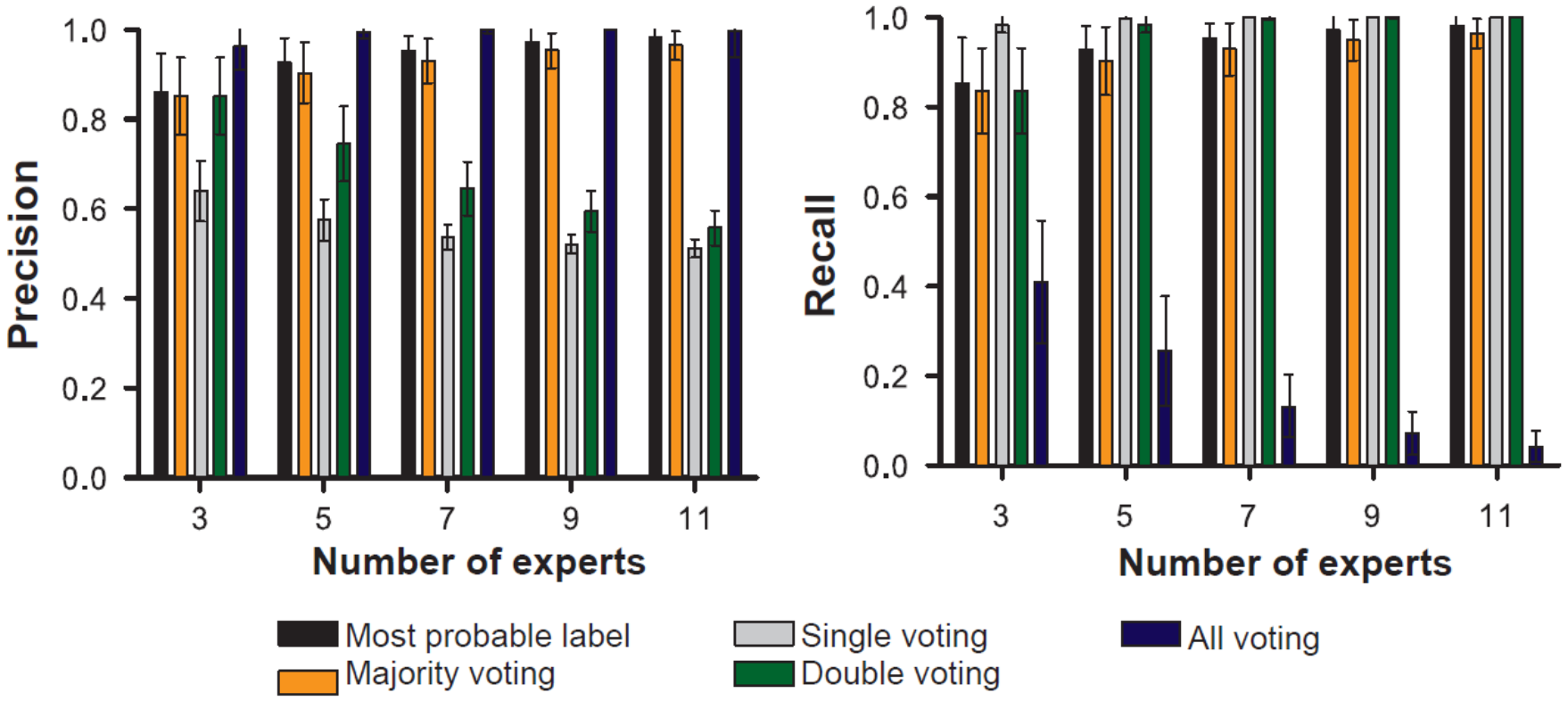
Synthetic experiments set up



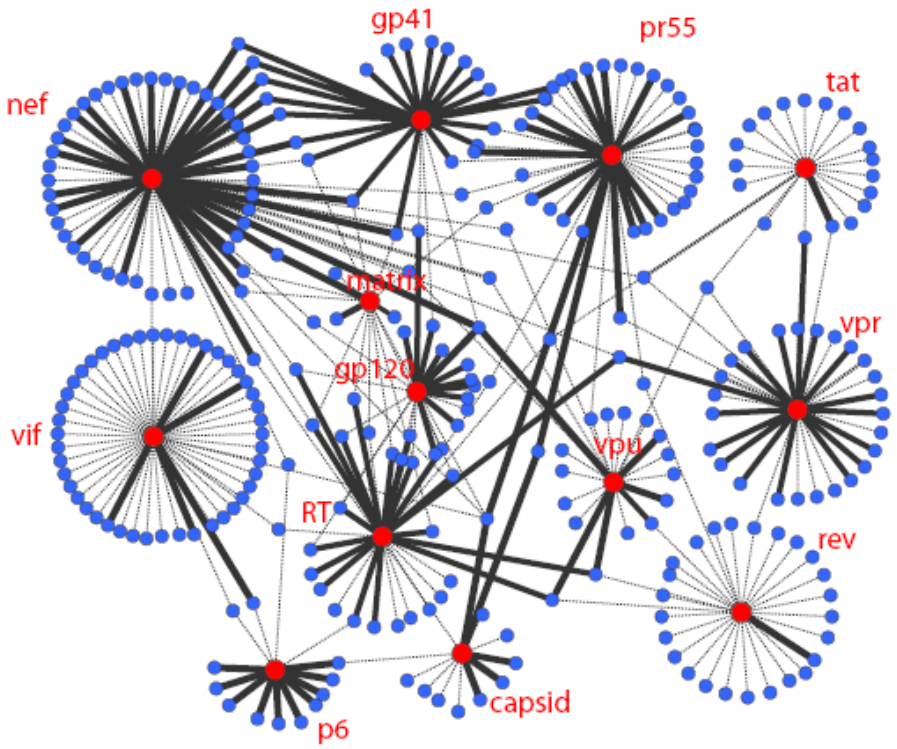
Synthetic data experiments



Comparison to baseline estimators



Refined interactome



Solid line: Estimated probability of being a direct interaction is ≥0.5
Dashed line: Estimated probability of being a direct interaction is <0.5

Edge thickness indicates confidence in the interaction

Possible directions

- Moving from experts to crowds – students?
- Providing incentives to annotate data
- Estimating over which type of interactions the labeler is better and optimizing which expert to ask which set of interactions

Acknowledgments

Jaime G. Carbonell

Allen Newell Professor,
Computer Science
Director, Language Technologies Institute, CMU

Judith Klein-Seetharaman
Associate Professor
Division of Metabolic and Vascular Health,
University of Warwick

Yanjun Qi
Assistant Professor
Department of Computer Science,
University of Virginia

Experts that helped us
Pittsburgh Center for HIV-1, Host Interaction Center

Chris Aiken, Vanderbilt University
Teresa Brosenitsch, UPITT

Questions?