Learning the Structure of Generative Models without Labeled Data

Stephen Bach Bryan He Alex Ratner Chris Ré
Stanford University

This Talk

- We study structure learning for generative models in which a latent variable generates weak signals

- The challenge is distinguishing between dependencies directly between the weak signals and those induced by the latent class

This Talk

- We propose an l1- regularized pseudolikelihood approach

- We develop a new analysis technique, since previous analyses of related approaches only apply to the fully supervised case

Roadmap

- Motivation: Denoising Weak Supervision with Generative Models
- Our Work: Learn their Structure without Ground Truth
- Results
  - Provable Recovery
  - Consistent Performance Improvements on Existing Systems

Motivation: Denoising Weak Supervision with Generative Models

Training Data Creation: $$$, Slow, Static

- Expensive & Slow:
  - *Especially when domain expertise needed*

Grad Student Labeler

- With deep learning replacing feature engineering, collecting training data is now often *the biggest* ML bottleneck

Snorkel

- Open-source system to build ML models with weak supervision
- Users write labeling functions, model their accuracies and correlations, and train models

snorkel.stanford.edu

snorkel

Example: Chemical-Disease Relations

TITLE:
Myasthenia gravis presenting as weakness after magnesium administration.
ABSTRACT:
We studied a patient with no prior history of neuromuscular disease who became virtually quadriplegic after parenteral magnesium administration for preeclampsia. The serum magnesium concentration was 3.0 mEq/L, which is usually well tolerated. The magnesium was stopped and she recovered over a few days. While she was weak, 2-Hz repetitive stimulation revealed a decrement without significant facilitation at rapid rates or after exercise, suggesting postsynaptic neuromuscular blockade. After her strength returned, repetitive stimulation was normal, but single fiber EMG revealed increased jitter and blocking. Her acetylcholine receptor antibody level was markedly elevated. Although paralysis after magnesium administration has been described in patients with known myasthenia gravis, it has not previously been reported to be the initial or only manifestation of the disease. Patients who are unusually sensitive to the neuromuscular effects of magnesium should be suspected of having an underlying disorder of neuromuscular transmission.

- We have entity mentions:
  - Chemicals
  - Diseases
- Goal: Populate table with relation mentions

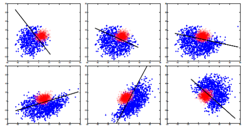
ID	Chemical	Disease	Prob.
00	magnesium	Myasthenia gravis	0.84
01	magnesium	quadriplegic	0.73
02	magnesium	paralysis	0.96

How can we train without hand-labeling examples?

Weak Supervision

Noisy, less expensive labels

Example types:

- Domain heuristics 
- Distant supervision 
- Crowdsourcing 
- Weak classifiers 

Generative Models for Weak Supervision

- Crowdsourcing
  [Dawid and Skene, 1979, Dalvi et al., WWW 2013]
- Hierarchical topic models for relation extraction
  [Alfonseca et al., ACL 2012, Roth and Klakow, EMNLP 2013]
- Generative models for denoising distant supervision
  [Takamatsu et al., ACL 2012]
- Generative models for arbitrary labeling functions
  [Ratner et al., NIPS 2016]

Labeling Functions – Domain Heuristics

"In our study, administering Chemical A caused Disease B under certain conditions…"

```
def LF_1(x):
  m = re.match('.*caused.*', x.sentence)
  return True if m else None
```

Labeling Functions – Distant Supervision

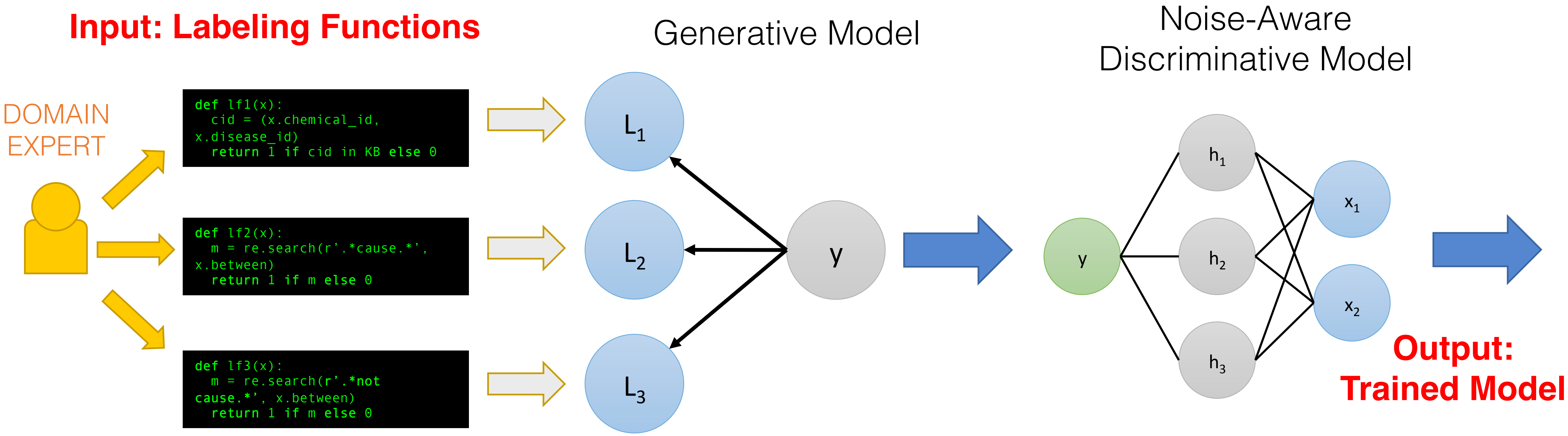
"In our study, administering Chemical A caused Disease B under certain conditions…"

```
def LF_2(x):
  in_kb = (x.chemical, x.disease) in ctd
  return True if in_kb else None
```

ctd

Comparative Toxicogenomics Database
http://ctdbase.org

Weak Supervision Pipeline in Snorkel

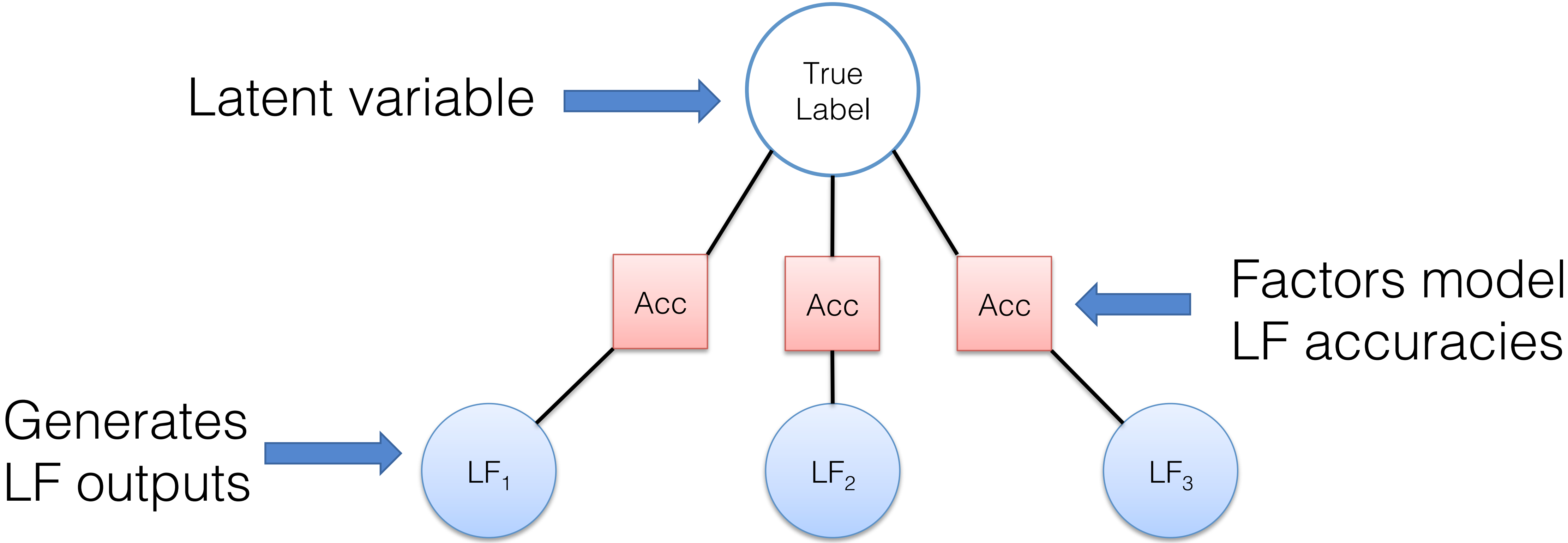


Users write functions to label training data

We model functions' behavior to denoise it

We use estimated labels to train a model

Denoising Weak Supervision



We maximize the marginal likelihood of the noisy labels

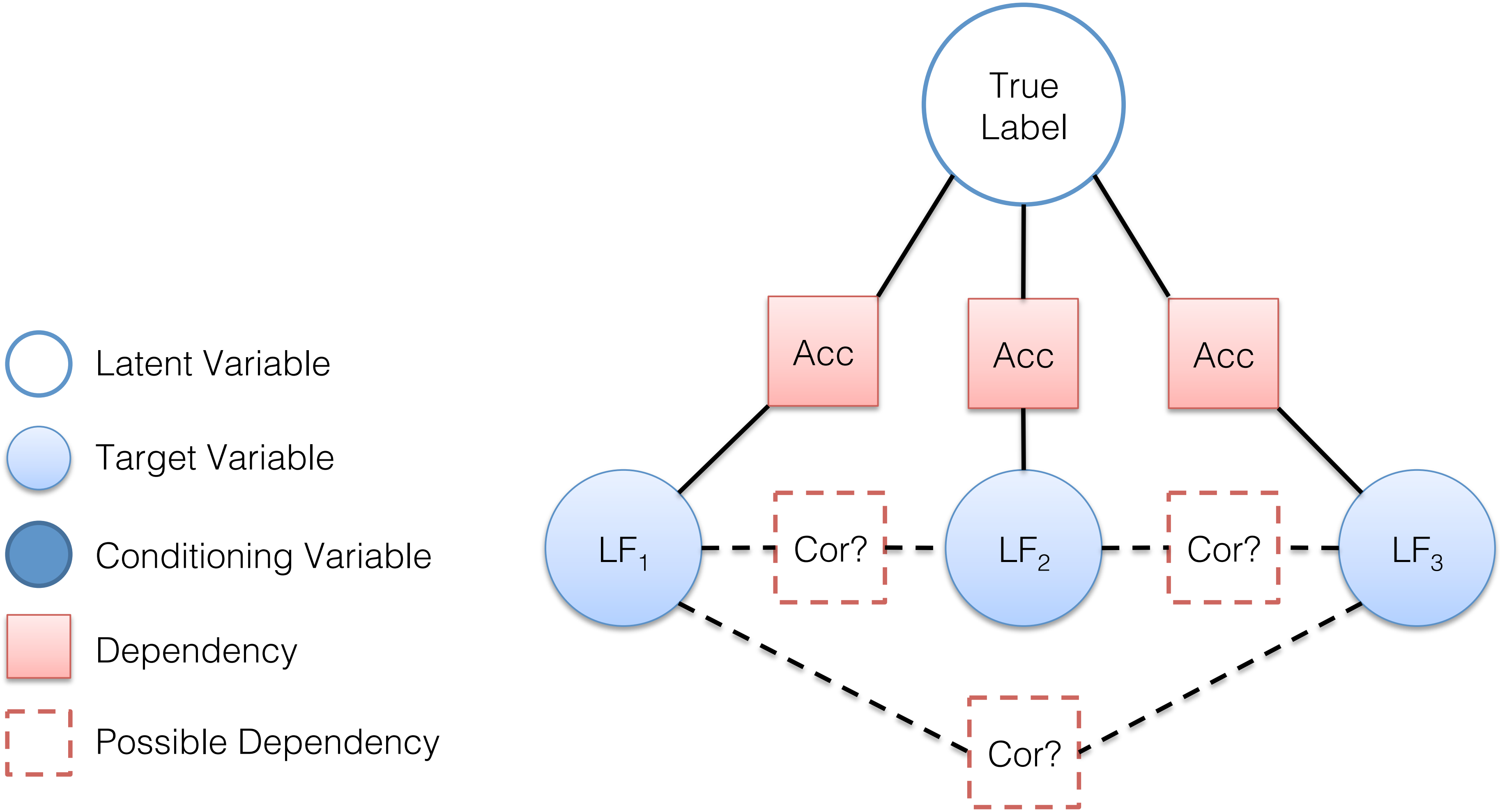
Intuitively, compares their agreements and disagreements

Dependent Labeling Functions

- Correlated heuristics
  - E.g., looking for keywords in different sized windows of text
- Correlated inputs
  - E.g., looking for keywords in raw tokens or lemmas
- Correlated Knowledge Sources
  - E.g., distant supervision from overlapping knowledge bases

Structure Learning

Structure Learning



Structure Learning for Factor Graphs

Challenges

- Gradient requires approximation
- Possible dependencies grow quadratically or worse

Prior Work

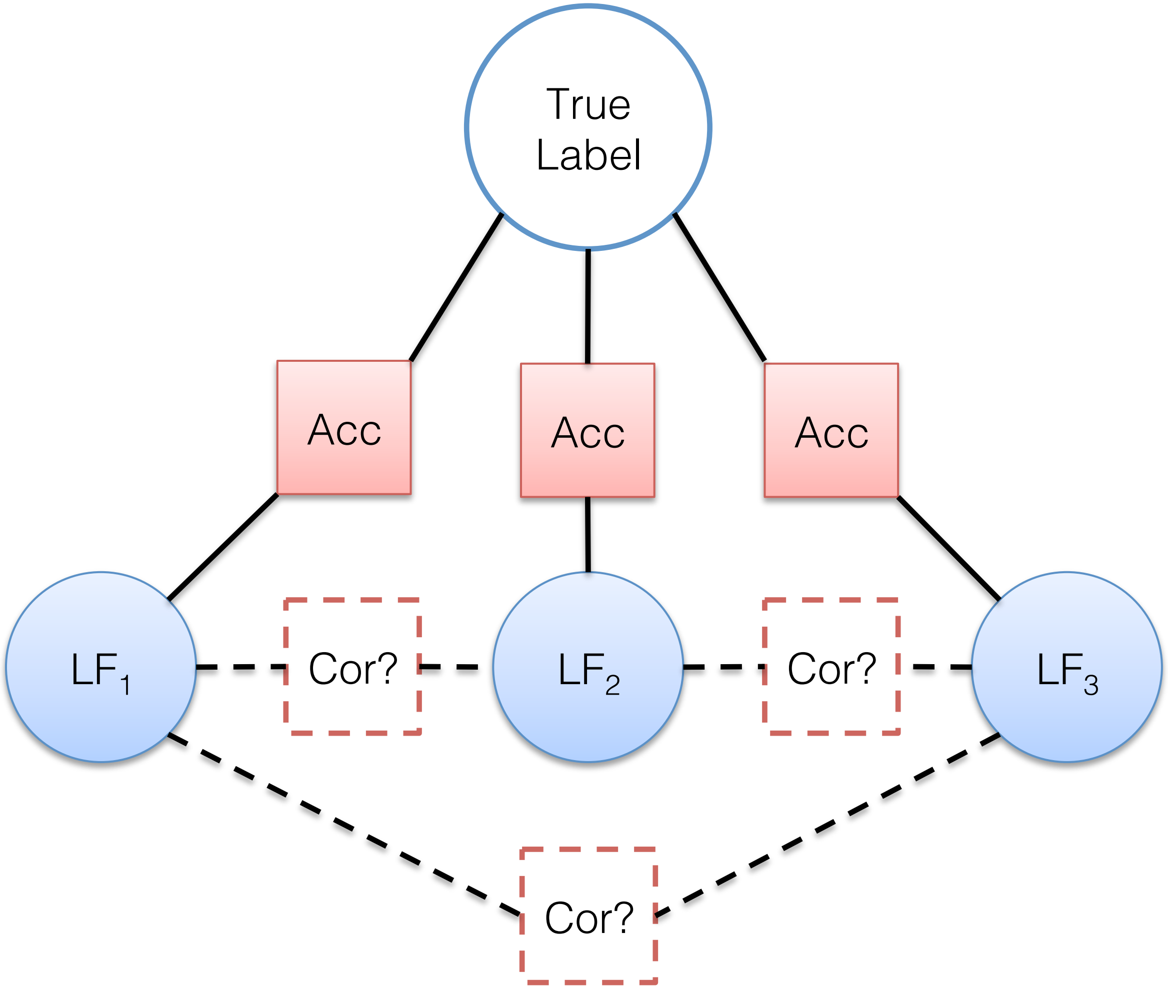
- Ravikumar et al. (Ann. of Stats., 2010) proposed using l1-regularized pseudolikelihood for supervised Ising models

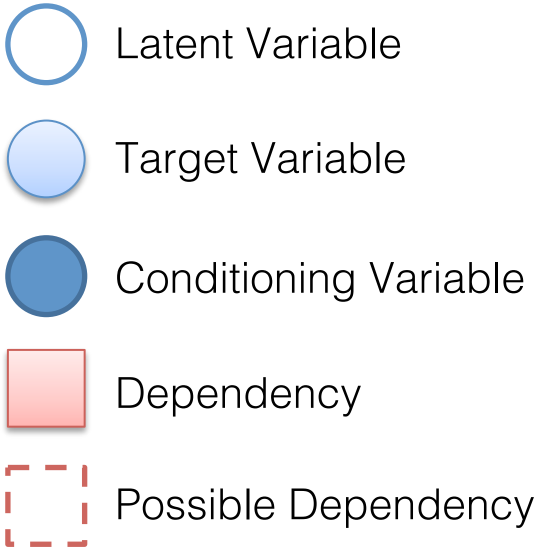
Structure Learning for Generative Models

- We maximize the l1-regularized *marginal* pseudolikelihood

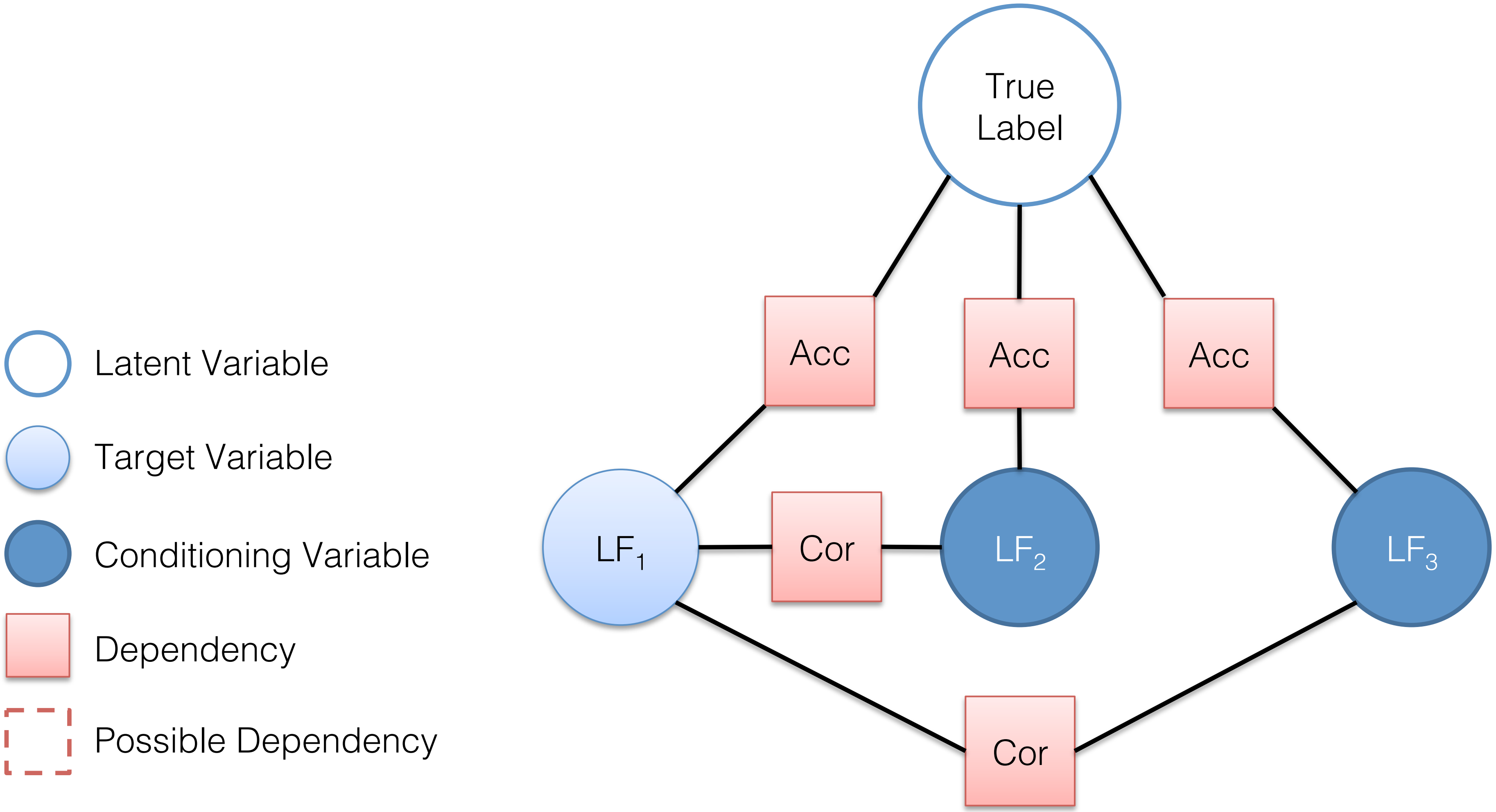
- One target variable and one latent variable means gradient can be computed exactly, efficiently

Structure Learning for Generative Models

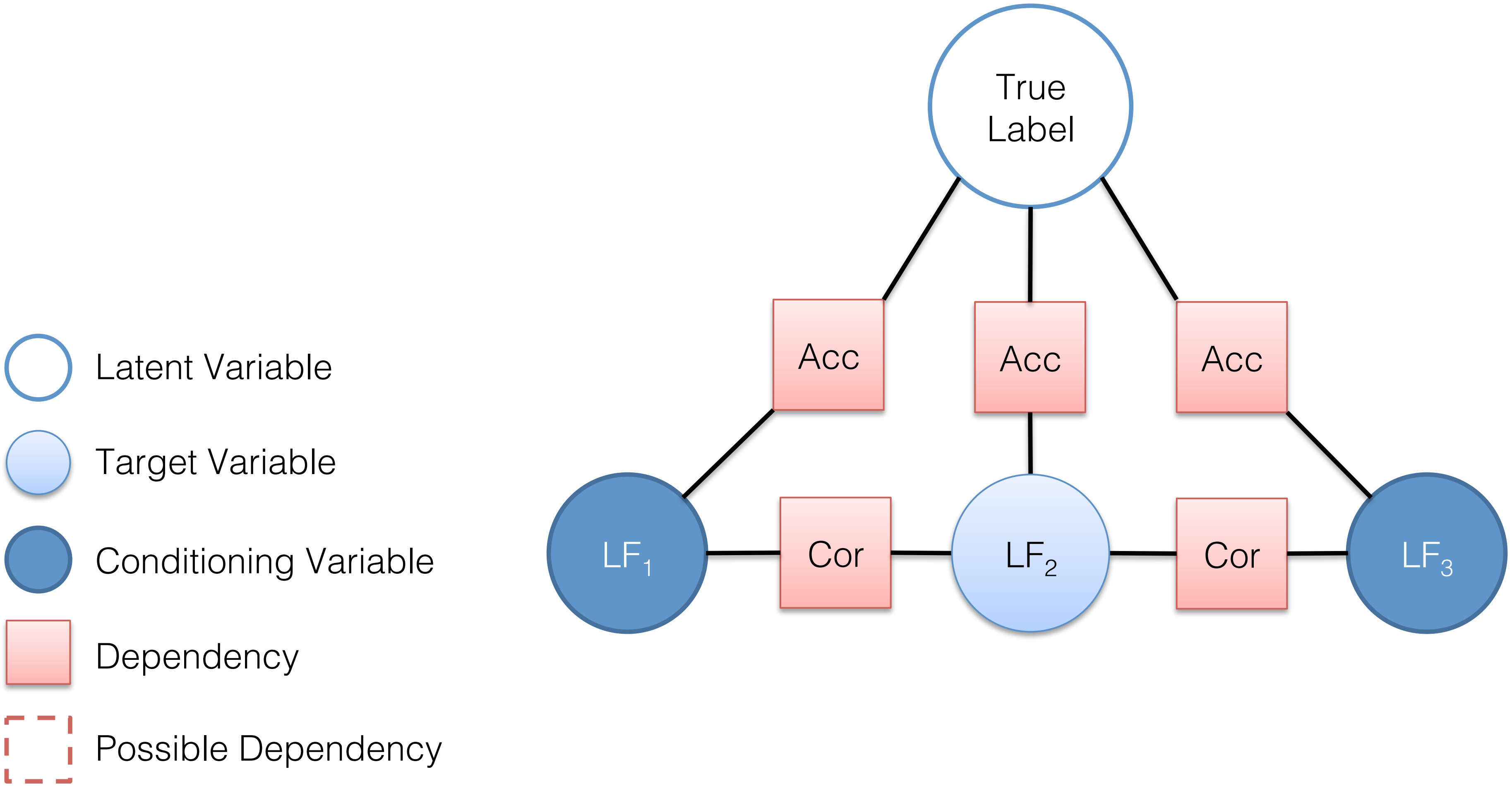




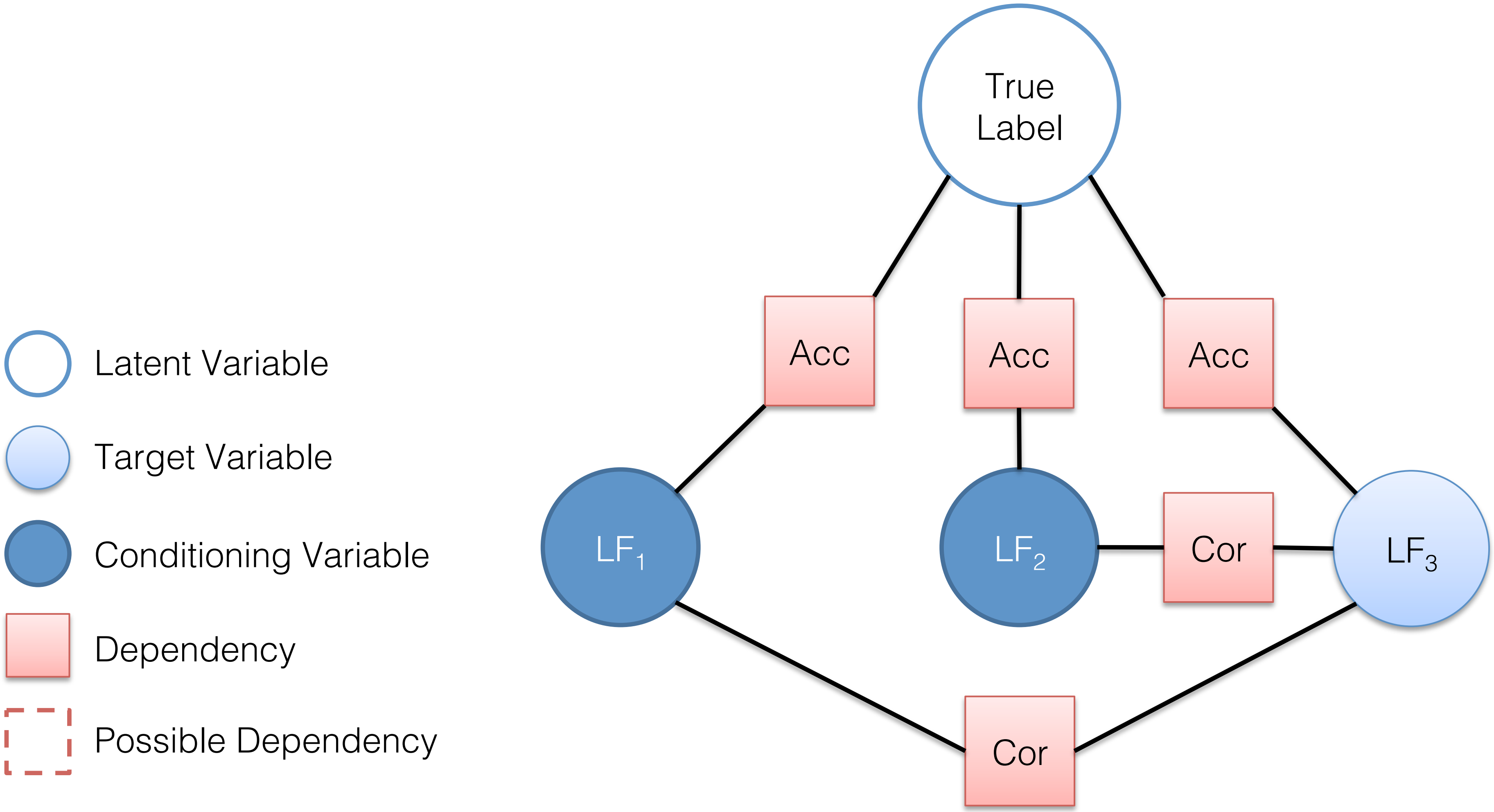
Structure Learning for Generative Models



Structure Learning for Generative Models



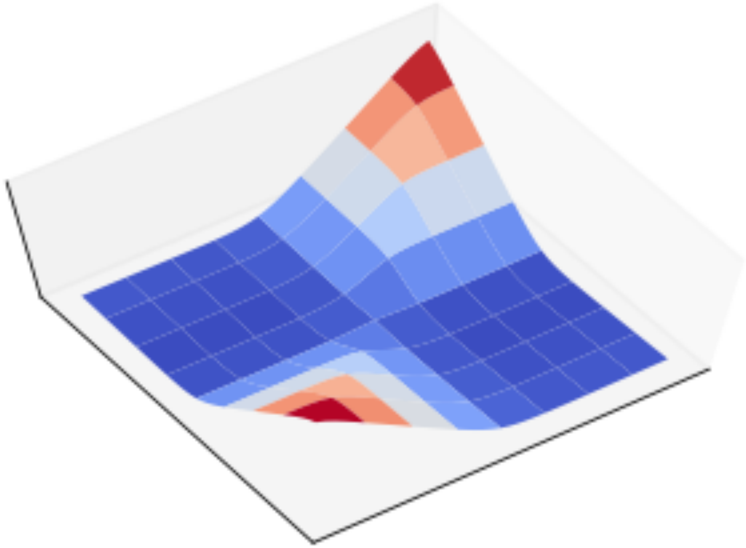
Structure Learning for Generative Models



Structure Learning for Generative Models

- Without ground truth, the problem becomes harder

- Latent variable means marginal likelihood is nonconvex



Analysis

Analysis

- Strategy
  - Focus on case in which most labeling functions are non-adversarial
  - Show that true model contained in region in which objective is locally strongly convex
- Assumptions
  - Feasible set of parameters that contains the true model
  - Over the feasible set, conditioning on a labeling function provides more information than marginalizing it out

Theorem: Guaranteed Recovery

For pairwise dependencies, such as correlations,

$$m \geq \Omega\left(n \log \frac{n}{\delta}\right)$$

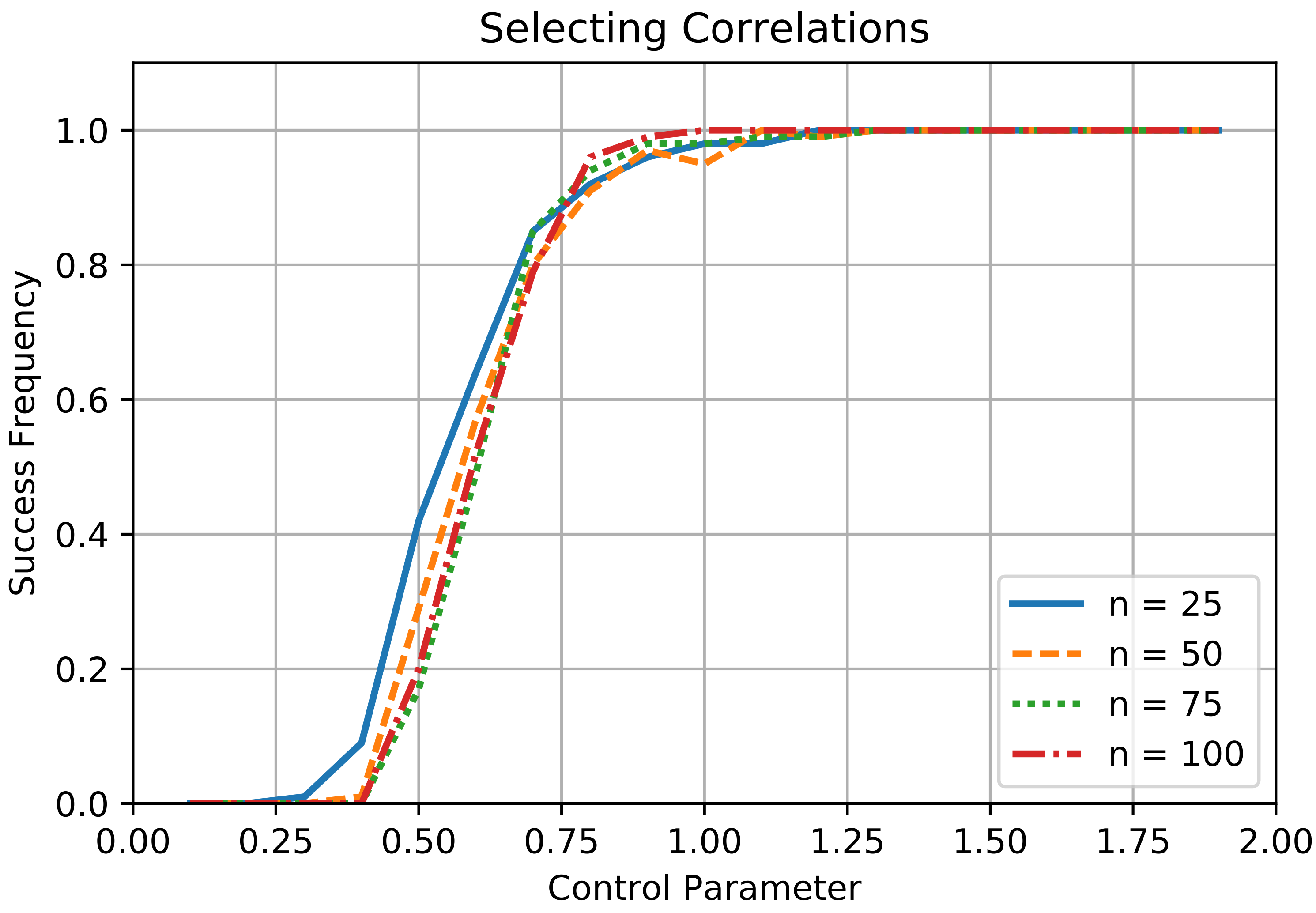
samples are sufficient to recover true dependency structure over n labeling functions with probability at least 1 - δ.

Empirical Results

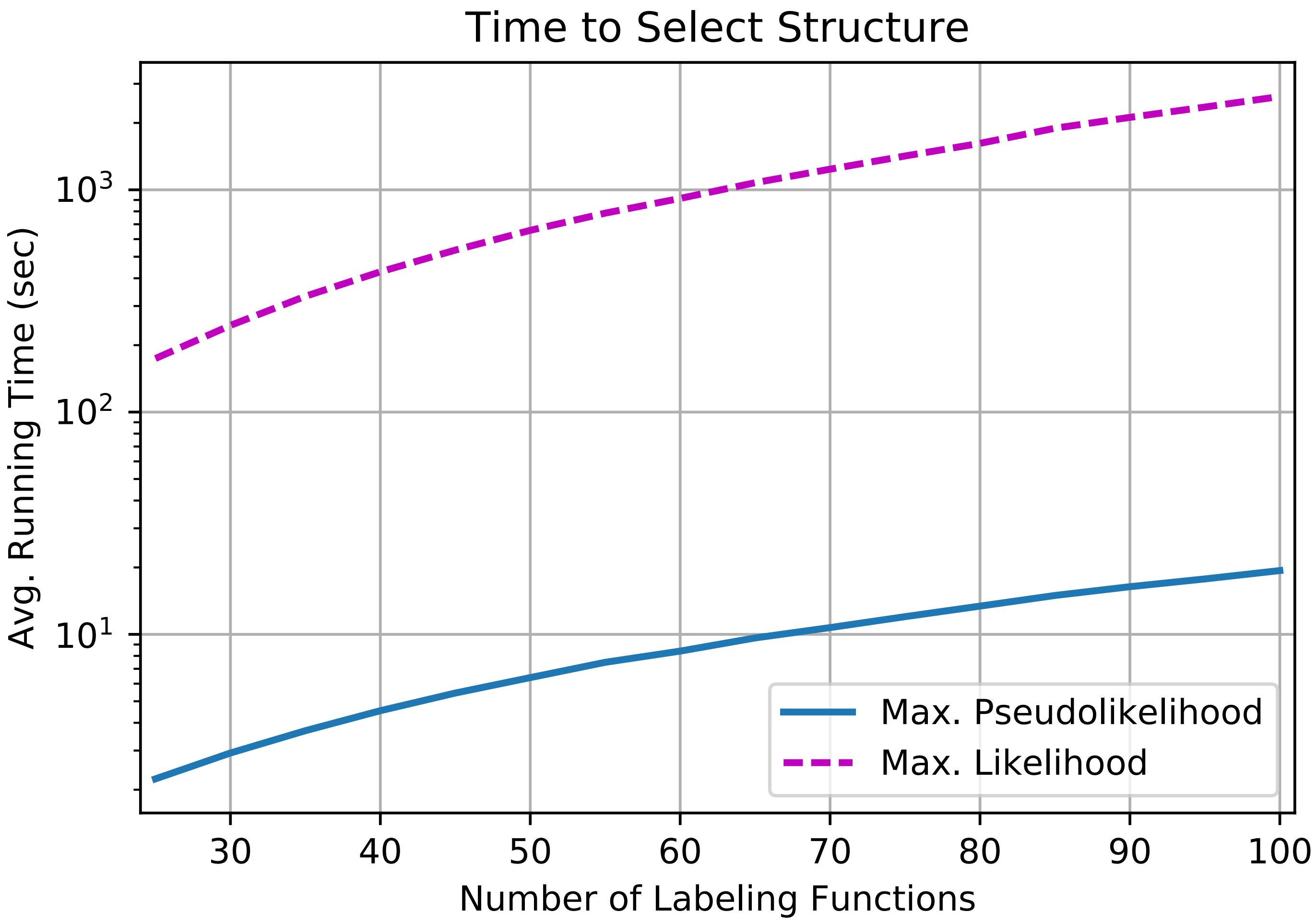
Empirical Sample Complexity

- Better in practice

- Same as observed in supervised setting



Speed Up: 100x



Improvement to End Models

Application	Ind. F1	Struct. F1	F1 Diff	# LF	# Dep.
Disease Tagging	66.3	68.9	+2.6	233	315
Chemical-Disease	54.6	55.9	+1.3	33	21
Device-Polarity	88.1	88.7	+0.6	12	32

Conclusion

- Generative models can help us get around the training data bottleneck, but we need to learn their structure
- Maximum pseudolikelihood gives
  - provable recovery
  - 100x speedup
  - end-model improvement

snorkel.stanford.edu

Thank you!