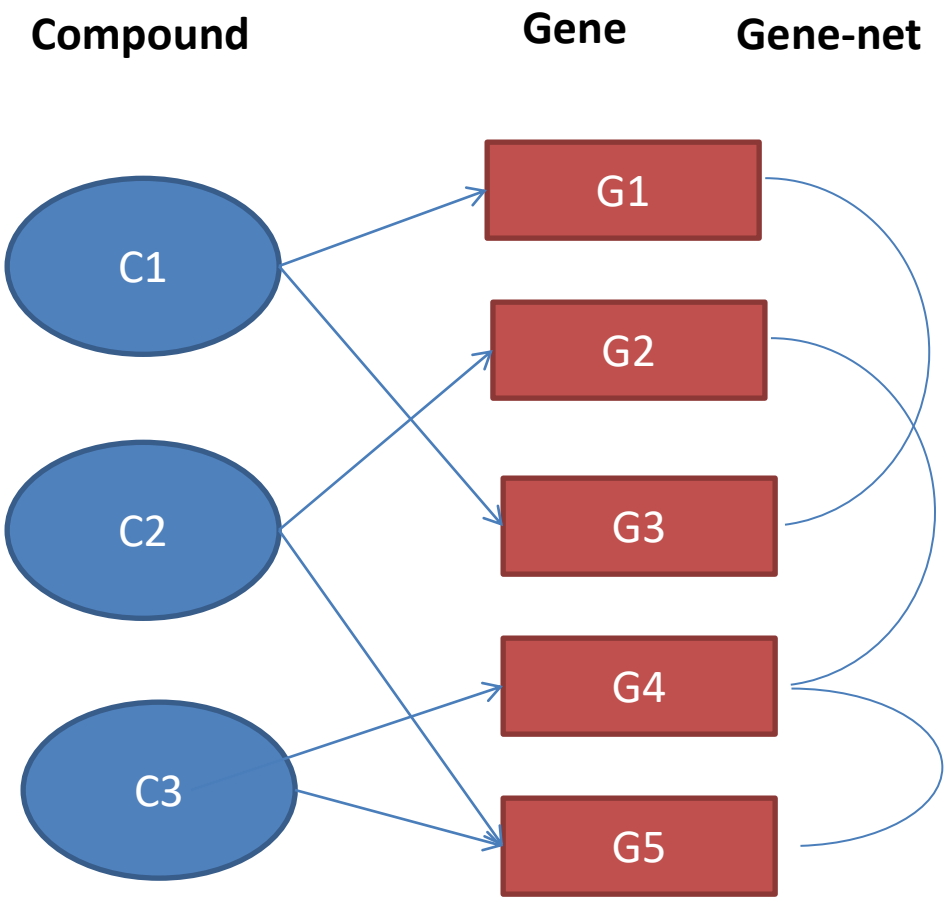
To-do slides (Dec 14, 2020)

Chen

- Static net (PPI + PDI) + DE
  - node prioritization by categories (TF, kinase, ligand)
  - Train embedding based on known perturbations
    - TF-specific signature (KO DE signature; KO embedding signature)
- Compound-gene net (data net)
  - Link prediction

L1000 formulation



Q1: given a compound (e.g. C1), hide 20% of known targets; how likely to recover the from rest 80%

Q2: hide 20% of known links; how likely to recover them from rest 80%

Q3: given a set of genes {G_i}, what are most likely up-stream change (siRNA) or compounds, which could induce/reverse such changes.

embedding (G) => C1 active or not (1/0)

Notes:

- Gene as Gene_UP and Gene_DW
- Compound could have different graph to represent relationships
- Compound experiments were also based on (i) cell-line (ii) time (iii)dosage

Simple feasibility

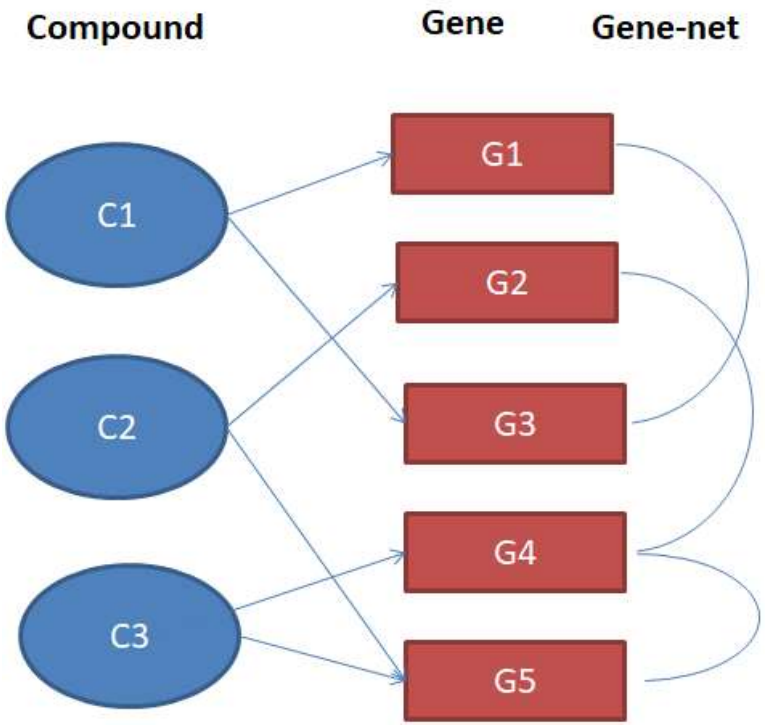
- Alternative and simple embedding
  - Spectral clustering
  - NMF
- Graph before NMF can be smoothed

(can be accounted as preparation before embedding)

  - pPageRank
    - One can add baseline p (e.g. =5%) for every nodes to ensure even unconnected nodes can be visited
  - Degree-corrected

datasets

https://maayanlab.cloud/L1000FWD/download_page



- 1. compound metadata
- 2. compound-> gene

CD_signatures_binary_42809.gmt	CD signatures (up/down gene sets) in the full space in gmt format.	115.0MB
CD_signatures_binary_42809.json	CD signatures (up/down gene sets) in the full space in json format.	169.5MB

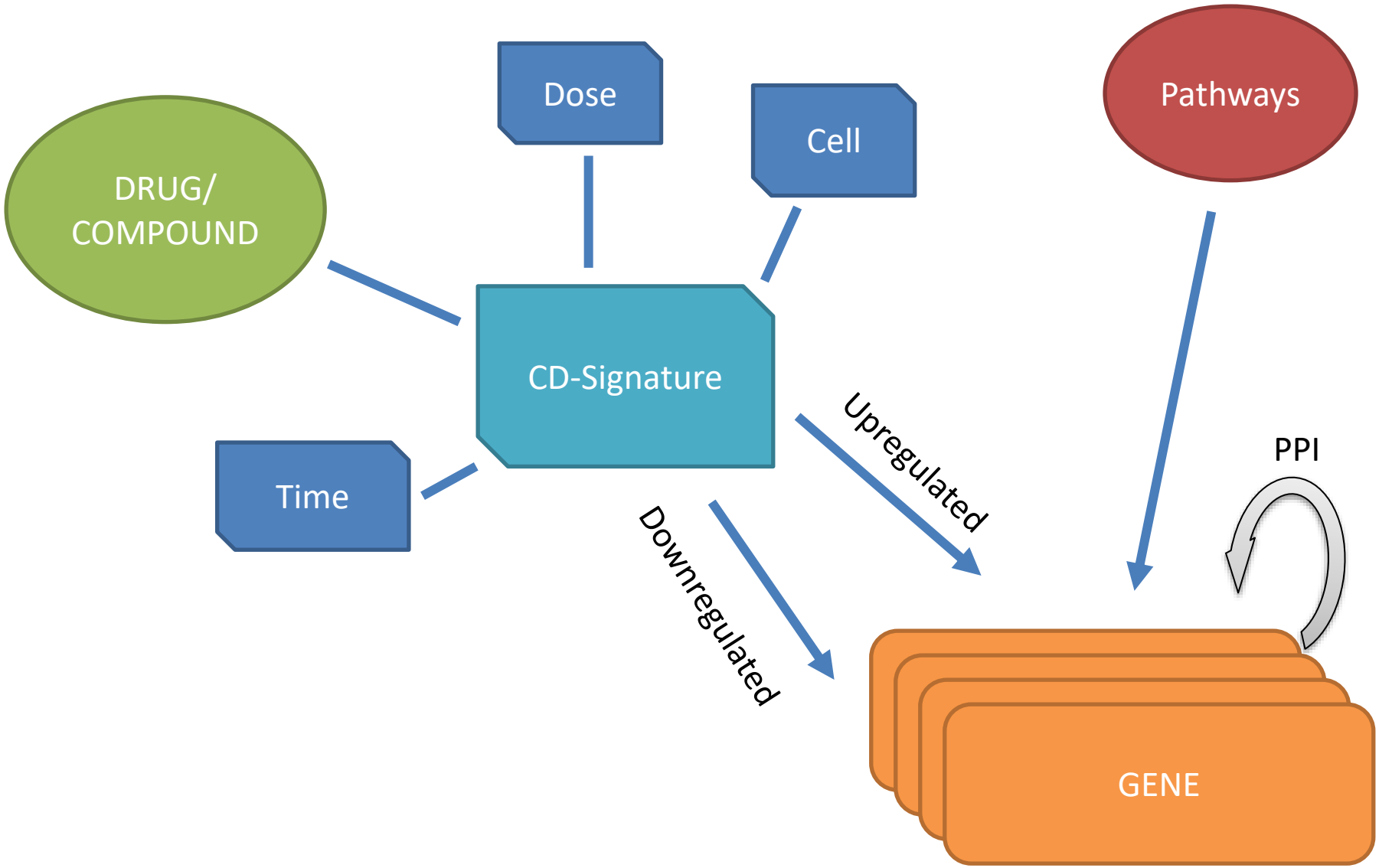
- 3. gene-net
  - Harmonizer (opt1) KEGG
  - Harmonizer (opt2) PPI & PDI

Methods & evaluations

Q1: given a compound (e.g. C1), hide 20% of known targets; how likely to recover the from rest 80%

Q2: hide 20% of known links; how likely to recover them from rest 80%

- **Performance**: CV-based AUC/AUPRC
- **Methods**:
  - Neo4j has built-in node2vec & other graph methods
  - Can they do CV within Neo4j?


DRUG/
COMPOUND
Dose
Cell
Pathways
CD-Signature
Time
Upregulated
Downregulated
PPI
GENE