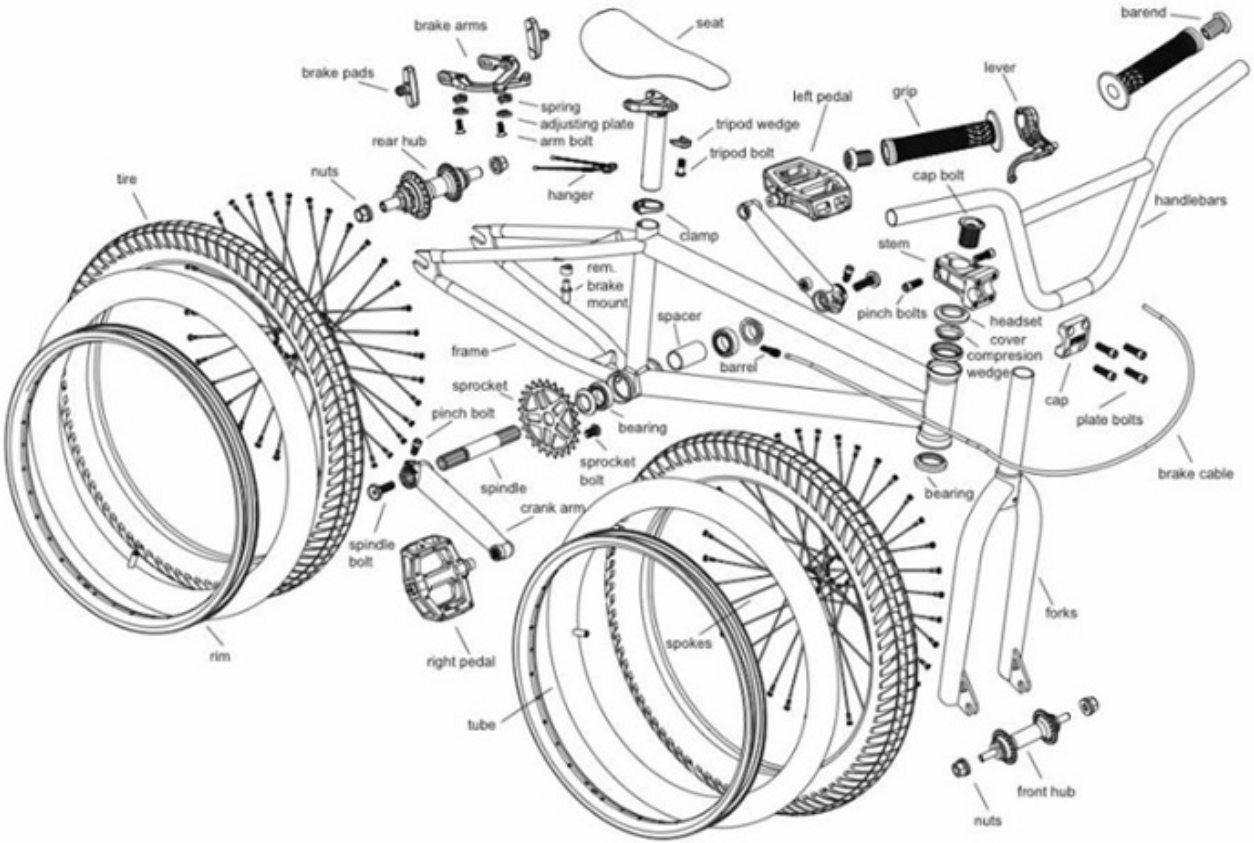
Machine Learning algorithms for making inferences on networks and answering questions in Biology and Medicine

Alberto Paccanaro

Department of Computer Science
Centre for Systems and Synthetic Biology
Royal Holloway, University of London

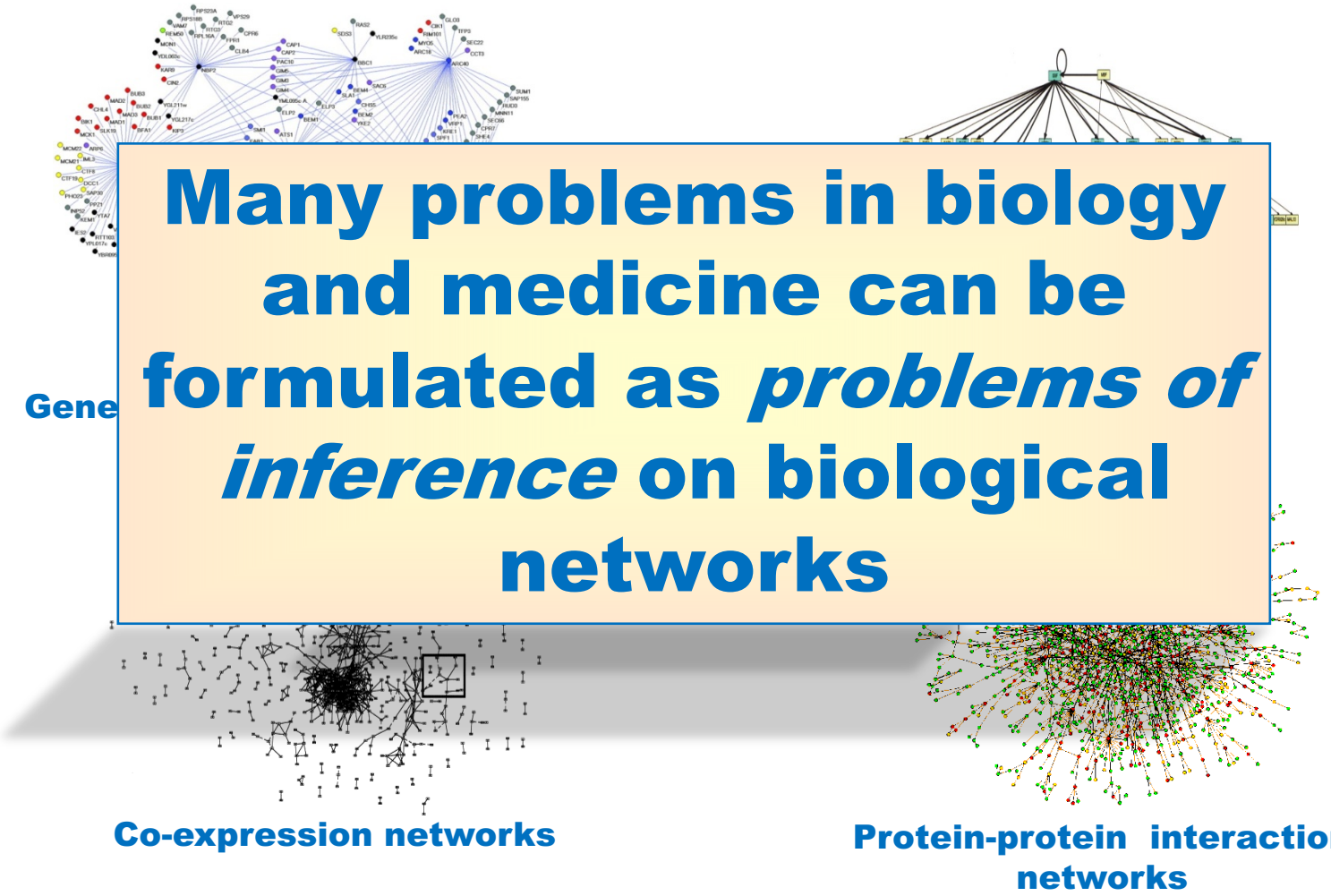
Why ML and biology

We need to analyse the cell at systems level



Biological networks

Cell as webs of interactions between biomolecules
Experimental data have a natural representation as networks



[Horak, Genes & Dev.; DeRisi, Science; Qian, J. Mol. Bio; Jeong, Nature,;Tong, Science; Goh, PNAS]

In my lab, we develop Machine Learning methods for answering questions in biology and medicine

focus on biological networks

- At the heart of our research is the biological question, not the methodology – **different areas of ML**
- **Diverse problems**
- Collaborate with **experimentalists**
- We implement **software tools** that allow biologists and clinicians to easily use the methods that we develop

Diverse problems, diverse approaches

Problem	Type of ML approach/technique	Reference
Predicting protein function (from sequence information only)	Semi-supervised learning	Jiang et al. *Gen. Biol.* 2016; Radivojac et al, *Nature Methods*, 2013
De-noising of proteomics data	Information diffusion over PPI graphs	Havugimana et al, *Cell*, 2012
Quantifying the functional similarity between genes (ontology-based)	Random Walks over ontology structures (DAGs)	Caniza et al, *Bioinf.*, 2014; Yang, Nepusz, Paccanaro *Bioinformatics*, 2012
Detection of protein complexes from protein interaction data	Overlapping clustering of large scale weighted graphs	Nepusz, Yu, Paccanaro *Nature Methods*, 2012
Selecting transcriptomics experiments for a given functional category	Supervised learning	Bhat, Yang, Paccanaro *PLoS ONE*, 2017
Selection of representative gene in a co-expression network	Function maximization (greedy, but global)	Yang, Paccanaro in preparation

Diverse problems, diverse approaches

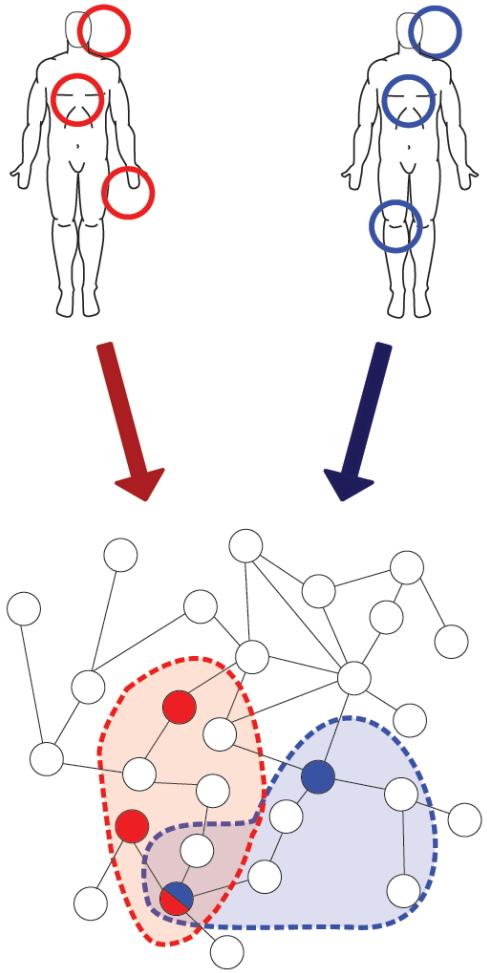
PROBLEM	TYPE OF ML APPROACH/TECHNIQUE	REFERENCE
Denoising of Hi-C data	Network modularity in random graphs	Ye et al, *Nature Methods* (under review)
Prediction of patients phenotype/outcome	Semi-supervised learning	Gliozzo et al, *PLoS Comp. Biol.* (under review)
Prediction of drug cocktails against Chagas disease	Supervised learning	Jimenez et al, *in preparation*
A measure of distance between diseases at molecular level	Analysis of sets of text labels on ontologies	Caniza, Romero, Paccanaro, *Nature Scient. Reports*, 2015
Prediction of disease genes for uncharted diseases	Semi-supervised learning	Caceres, Paccanaro, *PLoS Comp. Biol.* (to appear)
Predicting the frequency of drug side effects	Collaborative filtering (matrix factorization)	Galeano, Paccanaro bioRxiv 594465; doi: 10.1101/594465

1. Quantifying the distance between disease modules on the interactome

[Caniza, Romero, Paccanaro, Nature Scientific Reports, 2015]

Network Medicine: Disease as perturbations of molecular networks

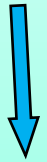
Protein-protein interaction networks



Genes associated with a specific disease tend to cluster in the same neighbourhood – the disease module

The disease modules of diseases that are phenotypically similar tend to be located in closeby regions of the interactome.

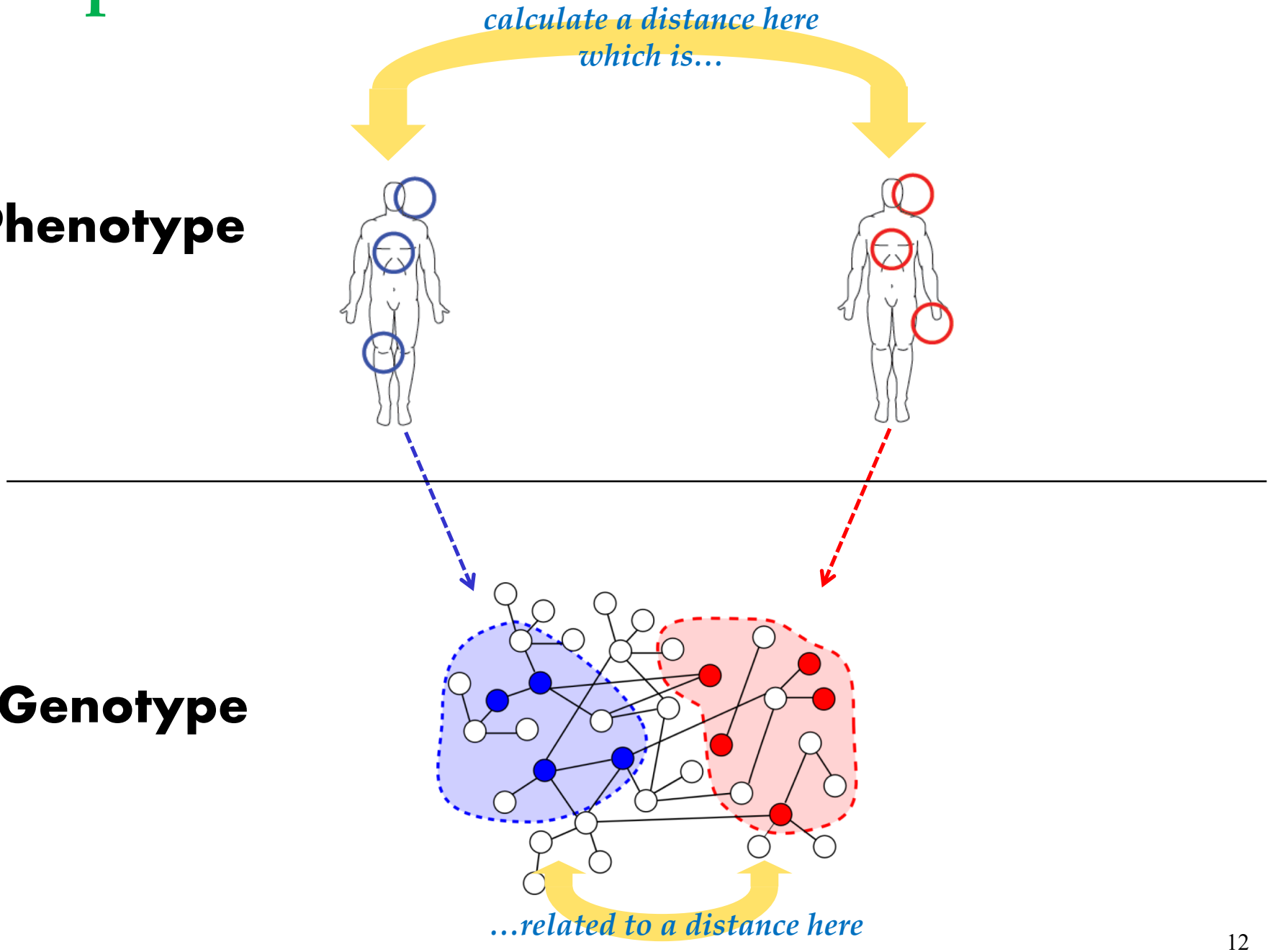
Question



Define a "distance" between diseases using the disease phenotypes such that

it is related to the distance between disease modules

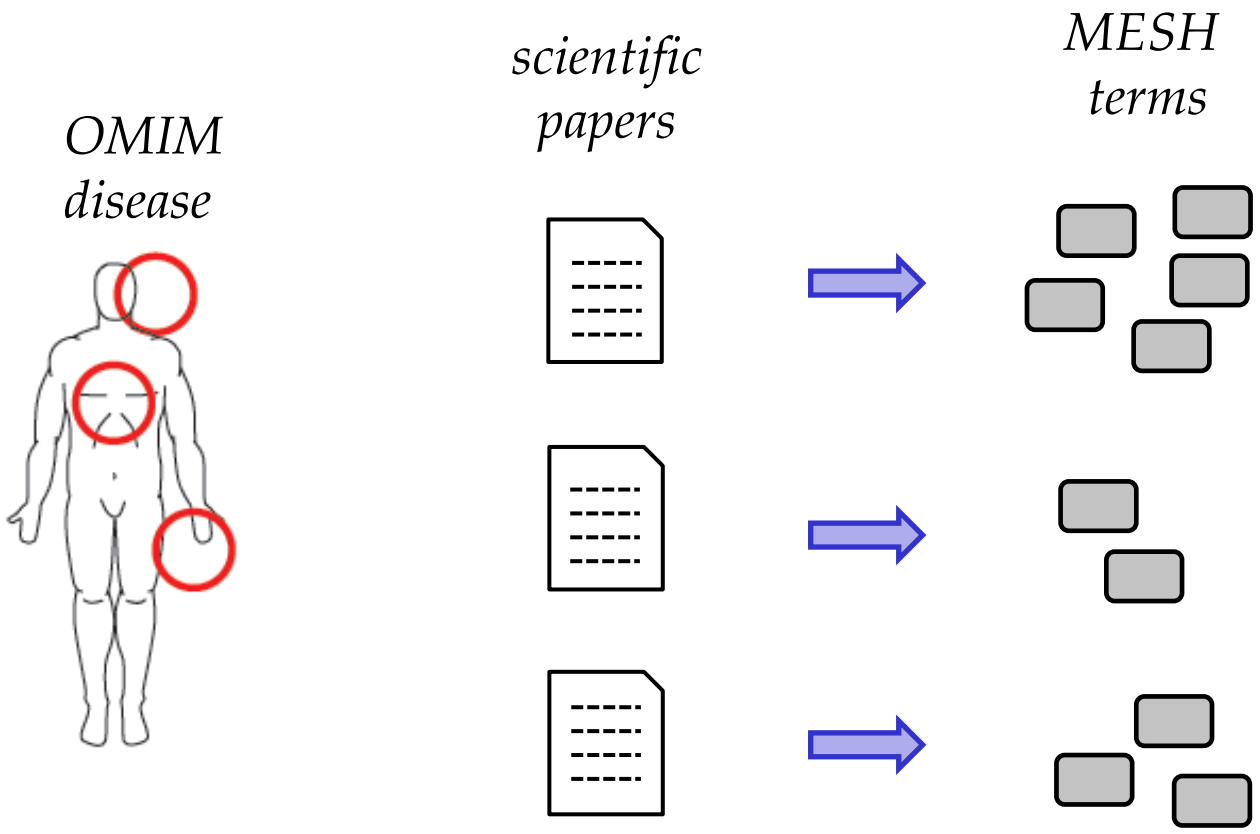
The problem


calculate a distance here
which is…
Phenotype
Genotype
…related to a distance here

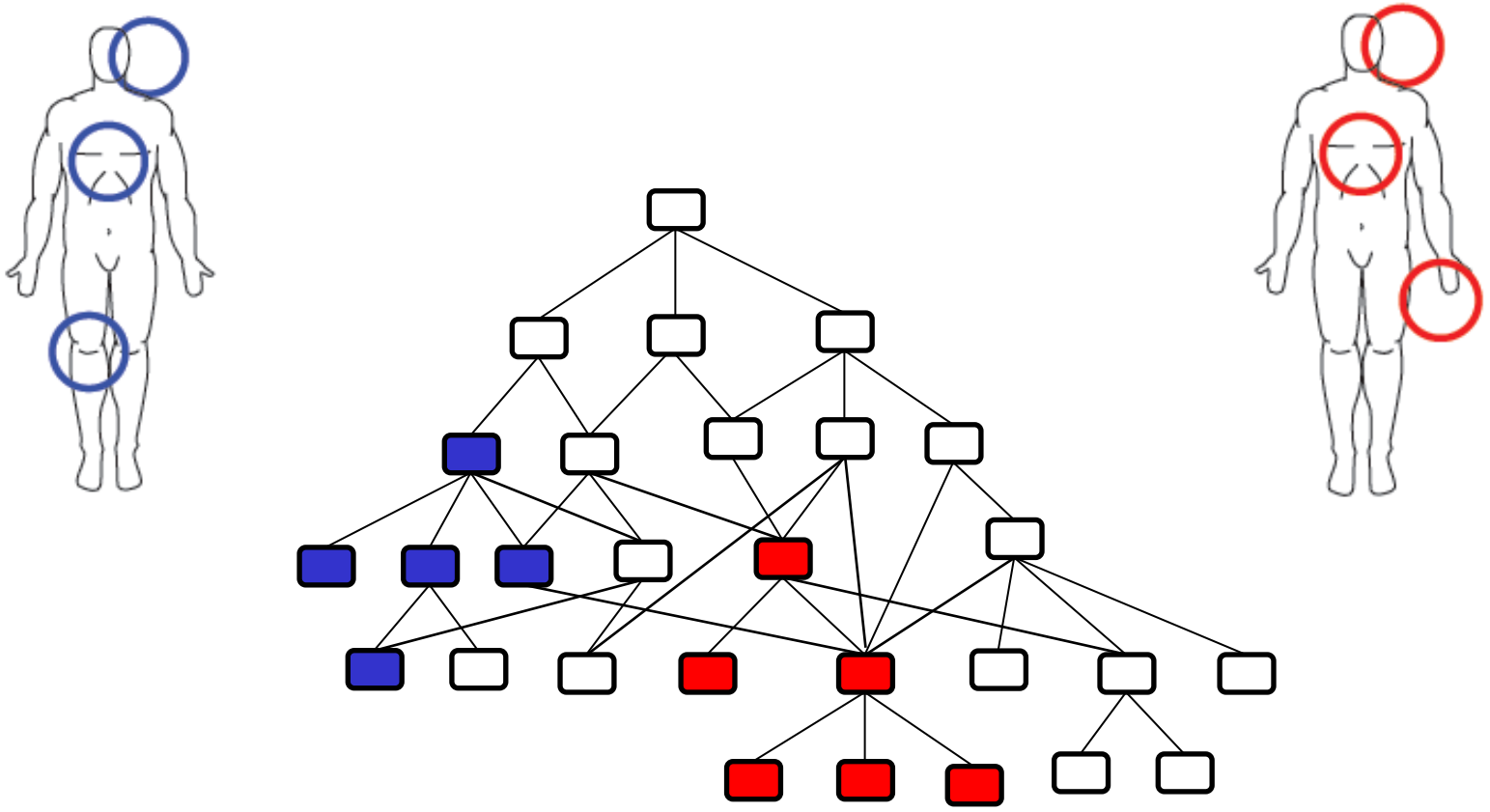
Outline of the method

[Caniza et al, *Scientific Reports*, 2015]

STEP 1: Translate a genetic disease into a set of MeSH terms



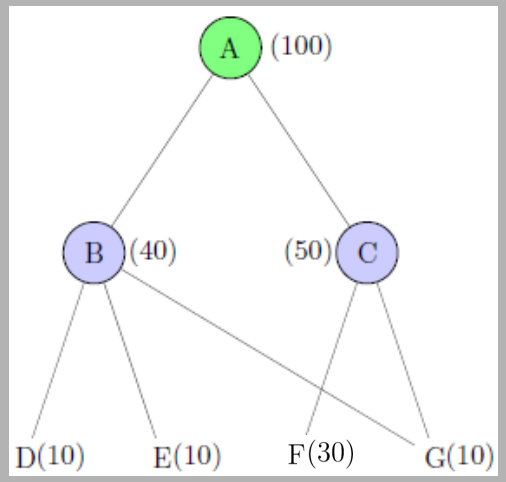
STEP 2: quantify a distance between two sets of terms on an ontology



Luckily ☺ , we had developed a measure for that !
(Yang et al, *Bioinformatics*, 2012; Caniza et al, *Bioinformatics*, 2014)

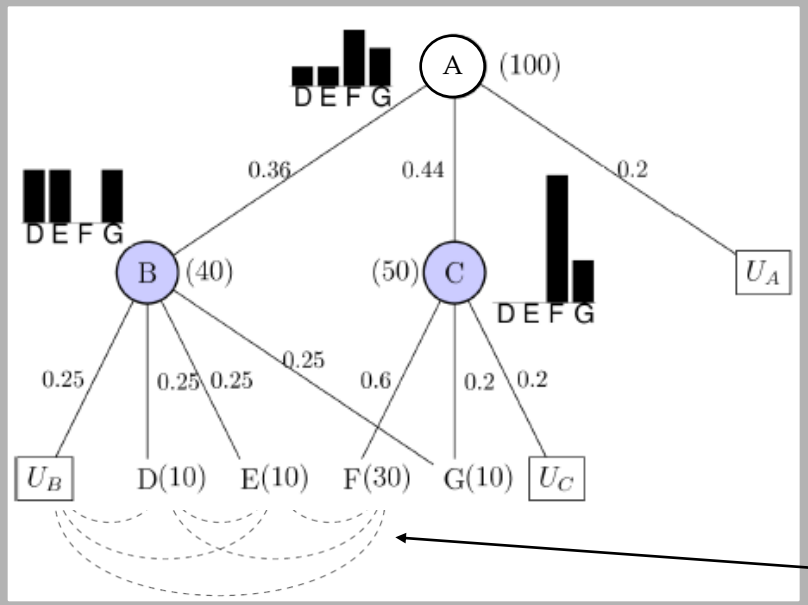
Host Similarity Measure, Random Walk Contribution

Host Similarity Measure
HSM (upward)



Yang et al, Bioinformatics, 2012
Caniza et al, Bioinformatics 2014
http://www.paccanarolab.org/gosstoweb/

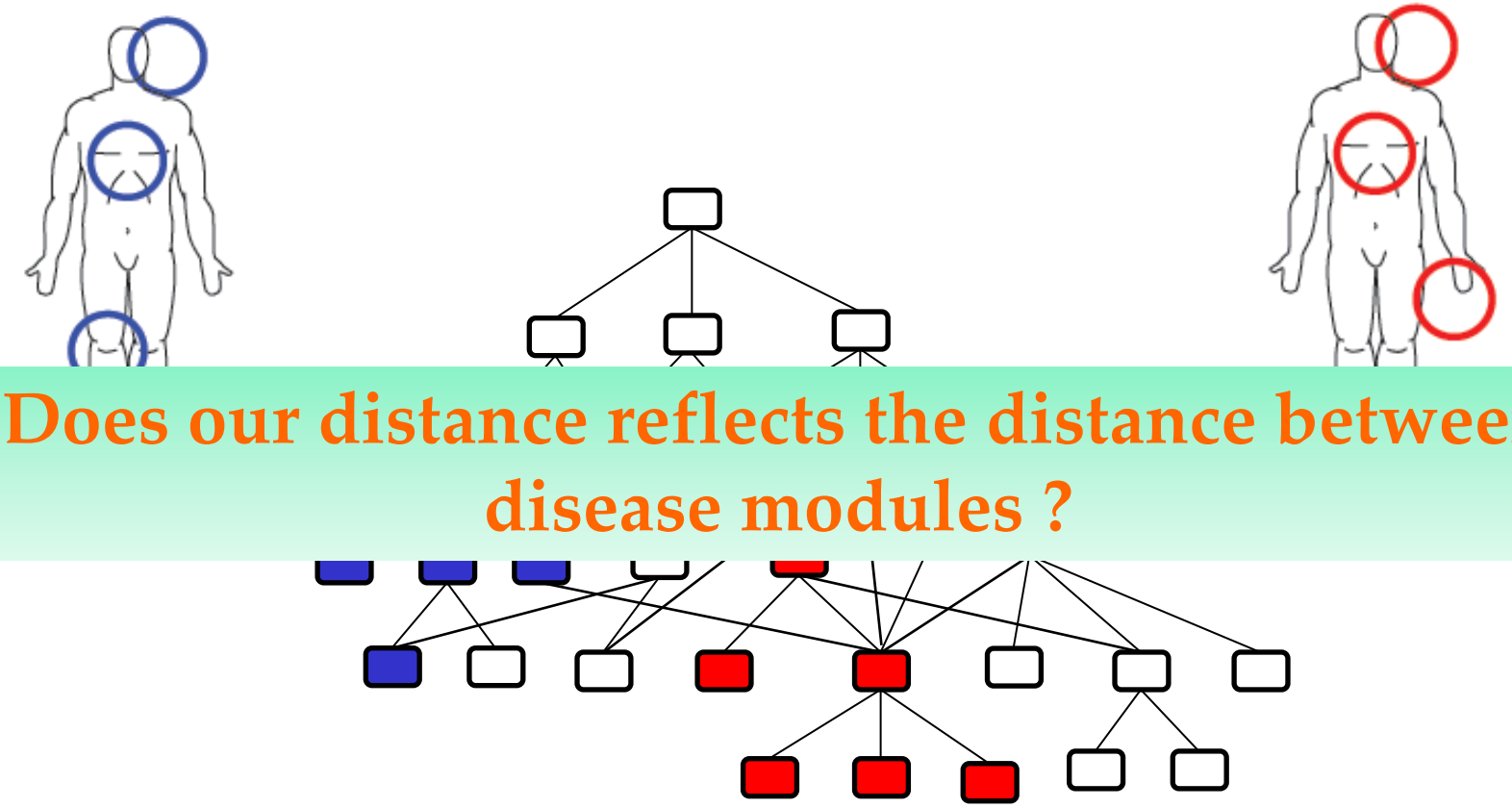
Random Walk Contribution
RWC (downward)



existence of common descendants
uncertainty
} ***affect the random walk***

HSM between every pair of leaves weighted by their probabilities

STEP 2: quantify a distance between two sets of terms on an ontology

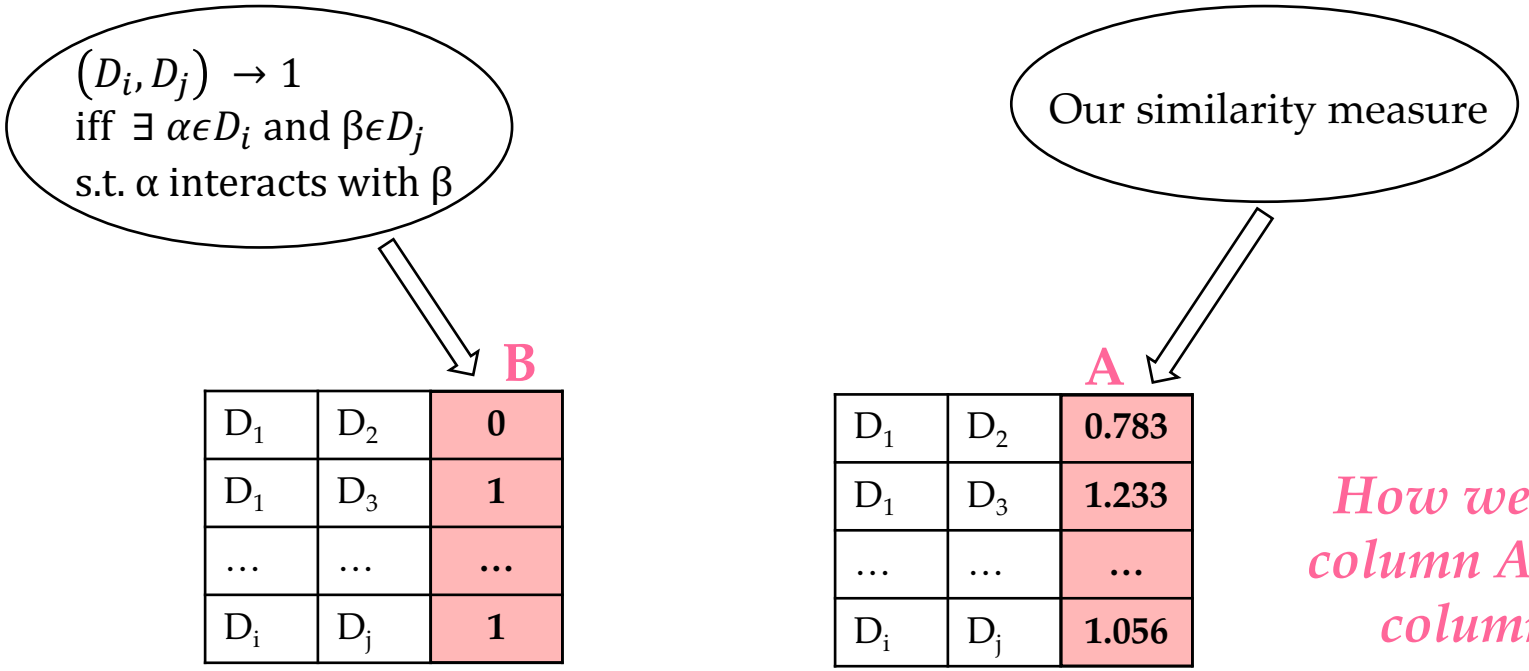


Does our distance reflects the distance between disease modules ?

Luckily ☺ , we had developed a measure for that !
(Yang et al, *Bioinformatics*, 2012; Caniza et al, *Bioinformatics*, 2014)

1. Evaluation as a prediction problem

A. Diseases related by physical interactions (PPI) of diseases proteins



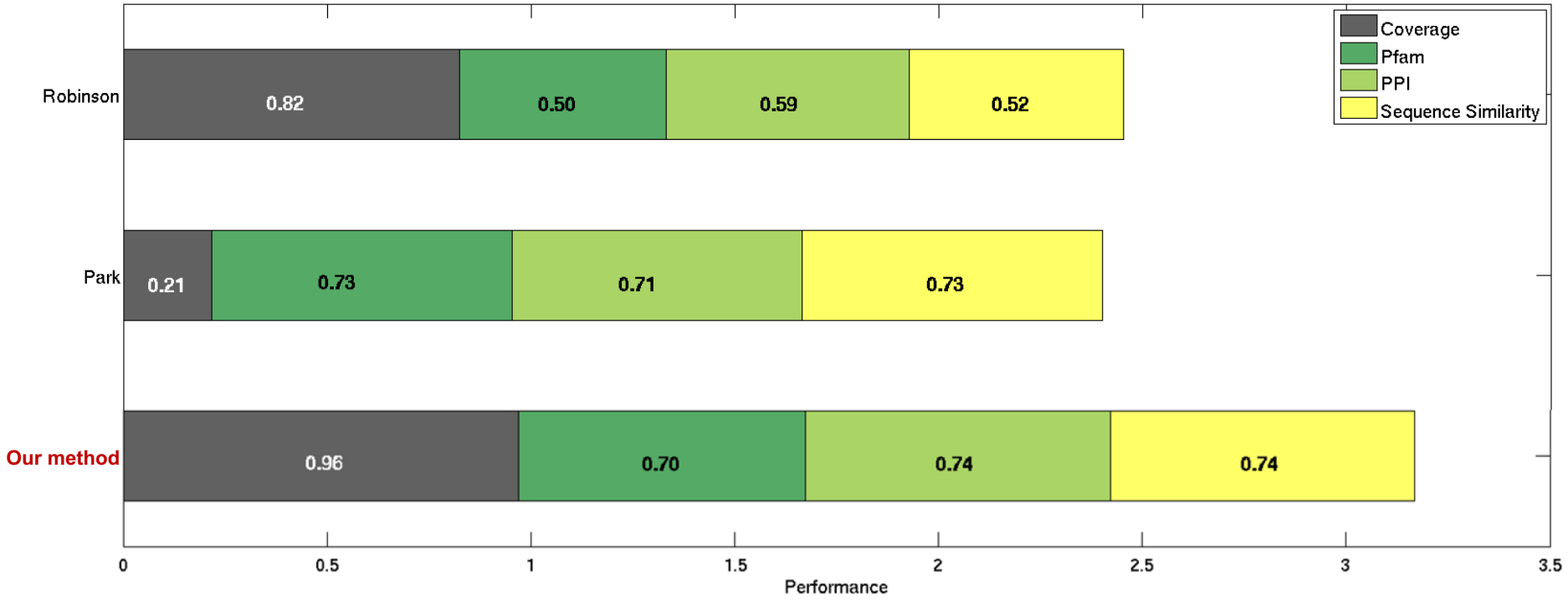
How well does column A predict column B?

B. Diseases related by sequence similarity of disease proteins

C. Diseases related by evolutionary relatedness of disease proteins (Pfam)

D. Coverage (% of OMIM diseases)

Results of AUC analysis



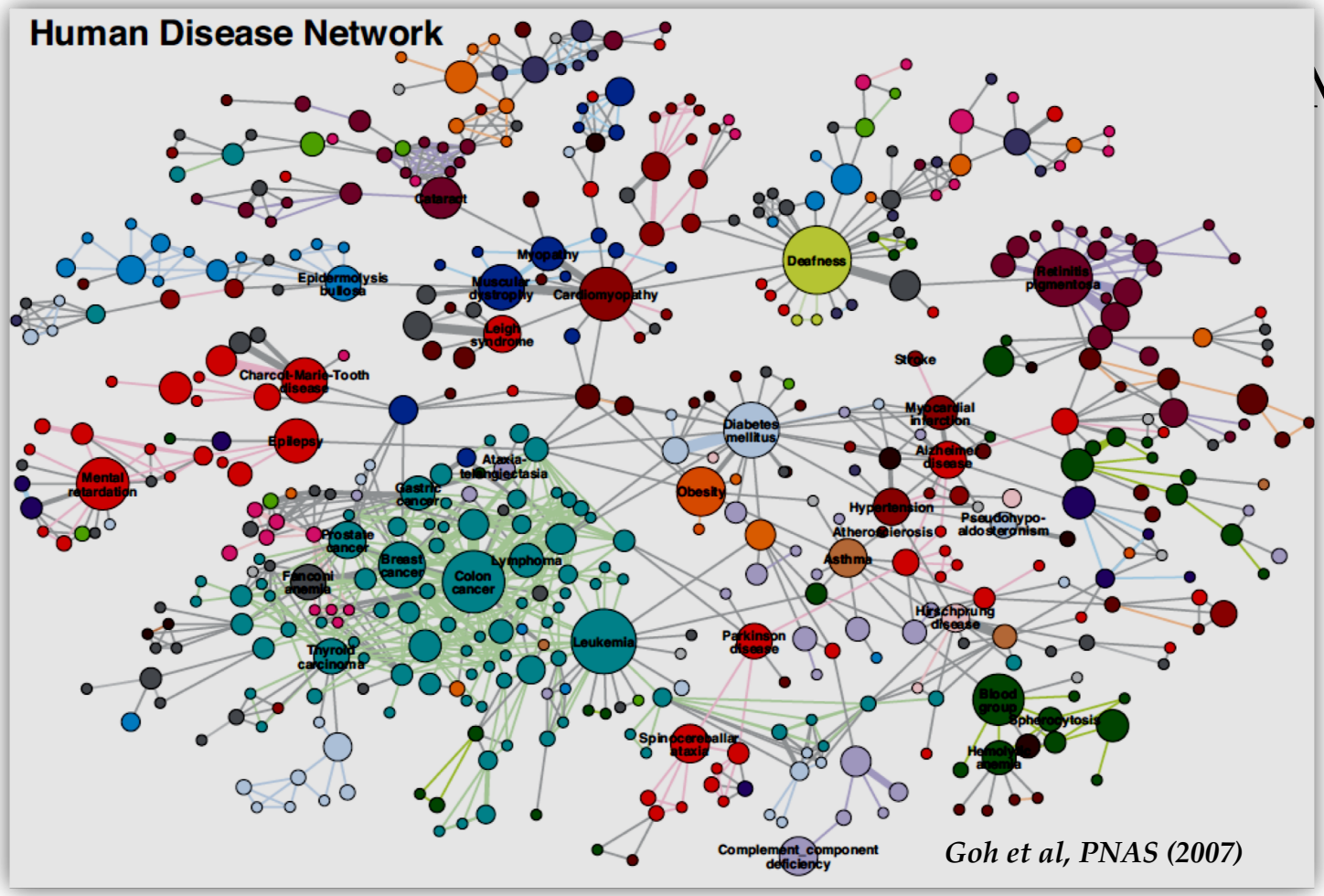
Robinson : builds and ad-hoc diseases ontology (**Human Phenotype Ontology**) and then calculates a distance on it (Köhler et al, NAR, 2013)

Park : similarity between two diseases is determined by an association score based on the **cellular co-localisation** of their disease proteins (Park et al, Mol. Sys. Bio. 2011)

2. Embedding diseases in low dimensional space

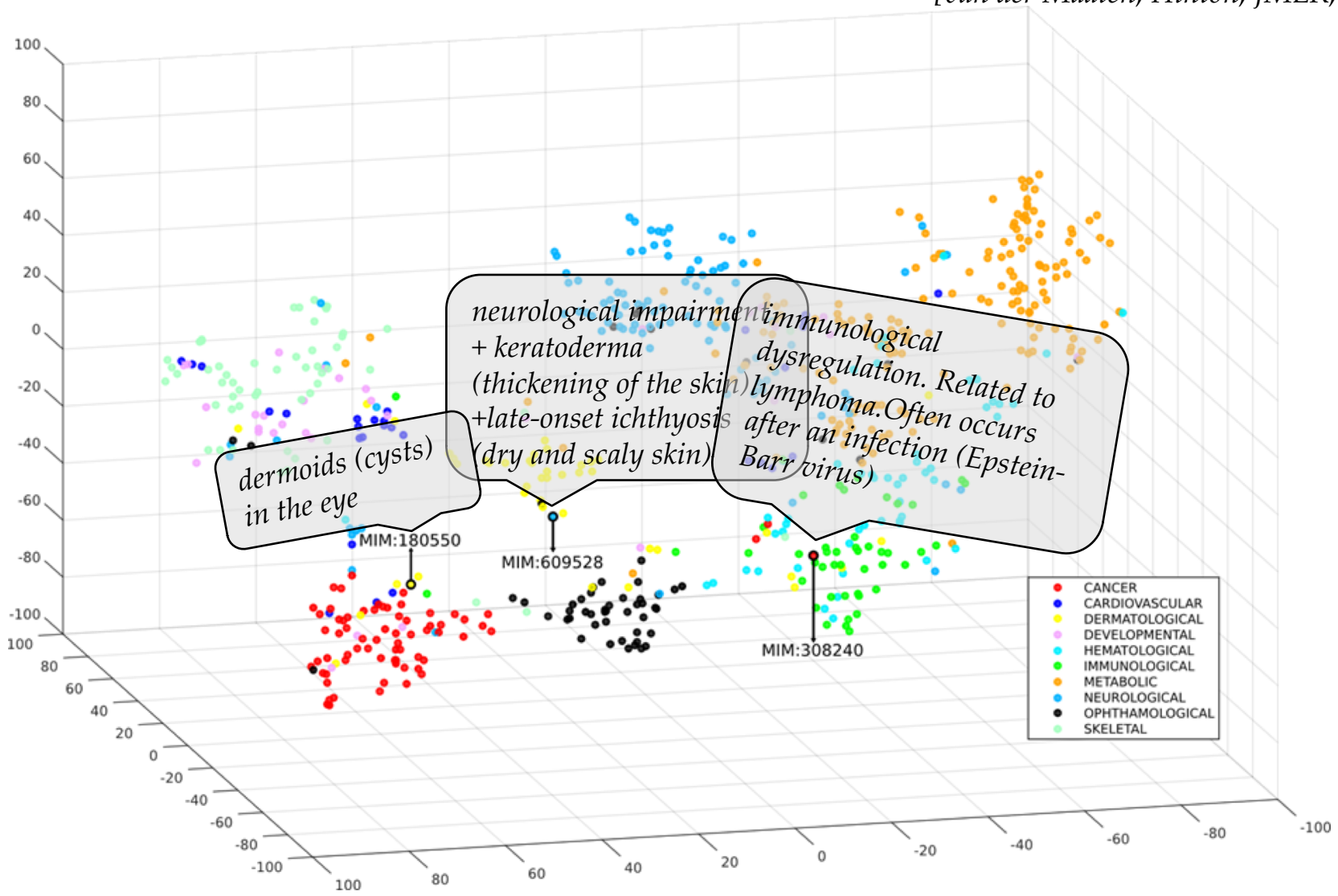
1)

2)



Embedding diseases in 3D using *t*-SNE

[van der Maaten, Hinton, JMLR, 2008]



MIM:180550 - Ring Dermoid of Cornea – cancer/dermatological/ophthalmological

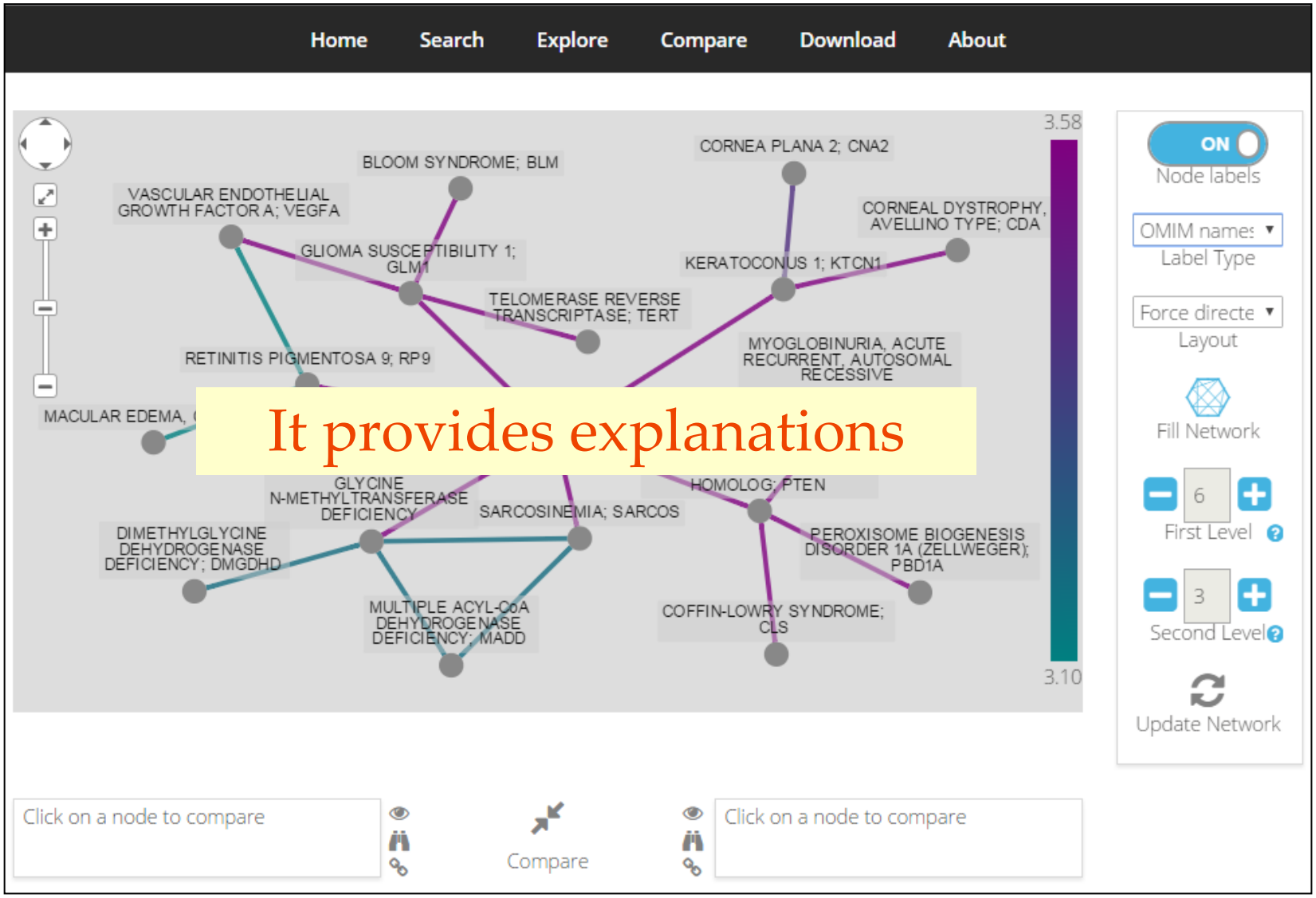
MIM:609528 - Cerebral dysgenesis, neuropathy, ichthyosis, and palmoplantar keratoderma syndrome – neurol./dermatol.

MIM:308240 - Lymphoproliferative syndrome – cancer/immunological

Landis – the Landscape of Disease Similarities

http://www.paccanarolab.org/landis

Differential diagnoses



2. Using disease distances to predict disease genes for Uncharted Diseases

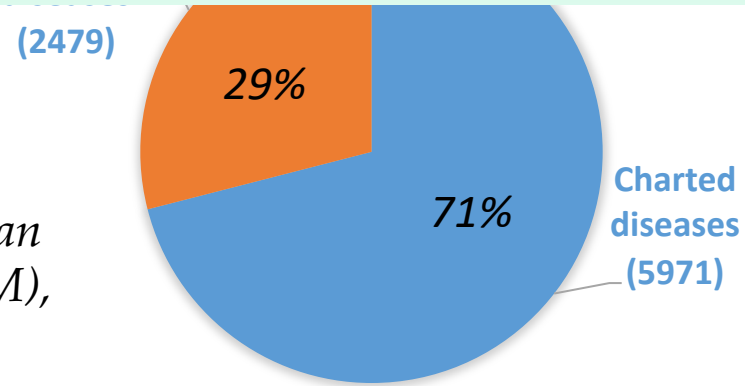
[Caceres, Paccanaro, PLoS Comp. Biology, *to appear*]

Disease gene prediction

- **Charted** diseases: some disease genes are known
- **Uncharted diseases:** no known disease genes

Disease gene prediction for charted diseases: search in a neighbourhood of known disease genes

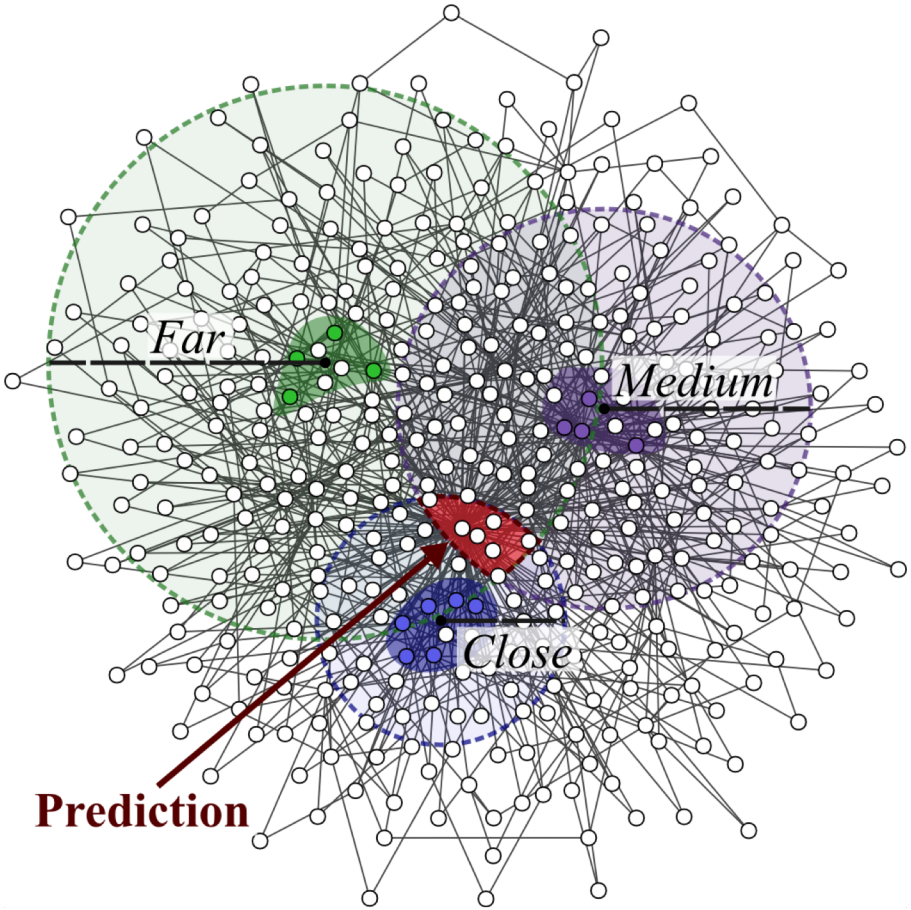
Can we use our disease similarity measure for predicting disease genes for uncharted diseases ?



Data from Online Mendelian Inheritance in Man (OMIM), Sept 2018

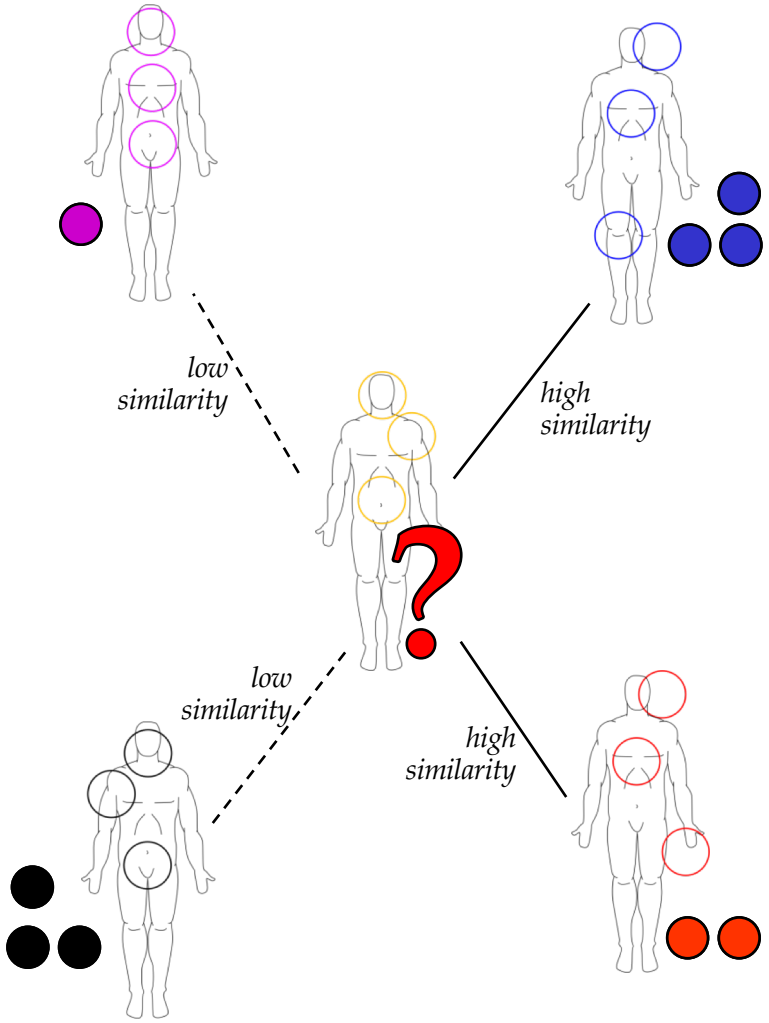
Predicting genes for *uncharted* diseases – the idea

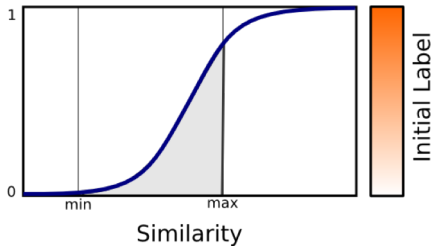
Triangulation: a mobile phone is detected within a radius from each of the towers.

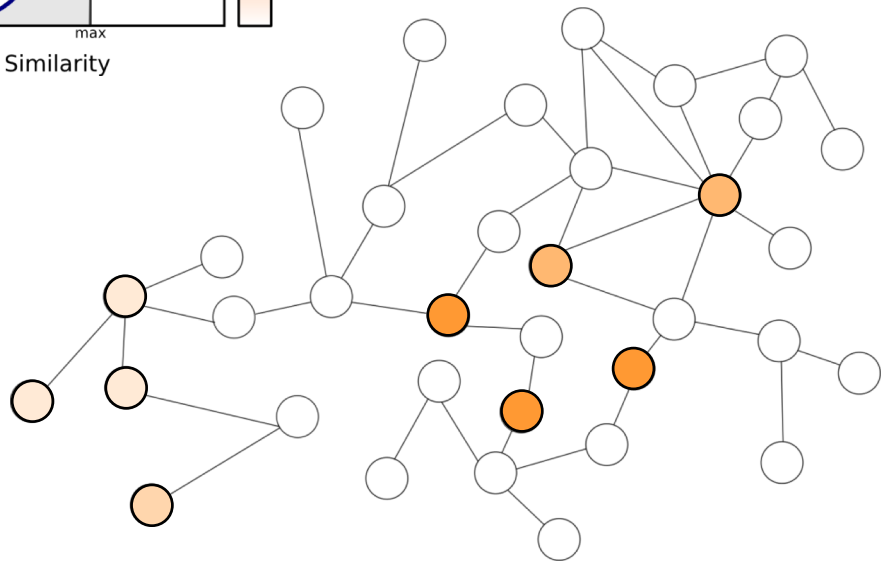


A new disease gene prediction algorithm

soft labels + diffusion







1. Calculate the similarity between our uncharted disease and each charted disease
2. Place known genes in the interactome.
3. Learn a *similarity-to-label* mapping
4. Assign a *"soft"* label to the disease genes
5. Diffuse the soft labels

Diffusing soft labels (semi-supervised learning)

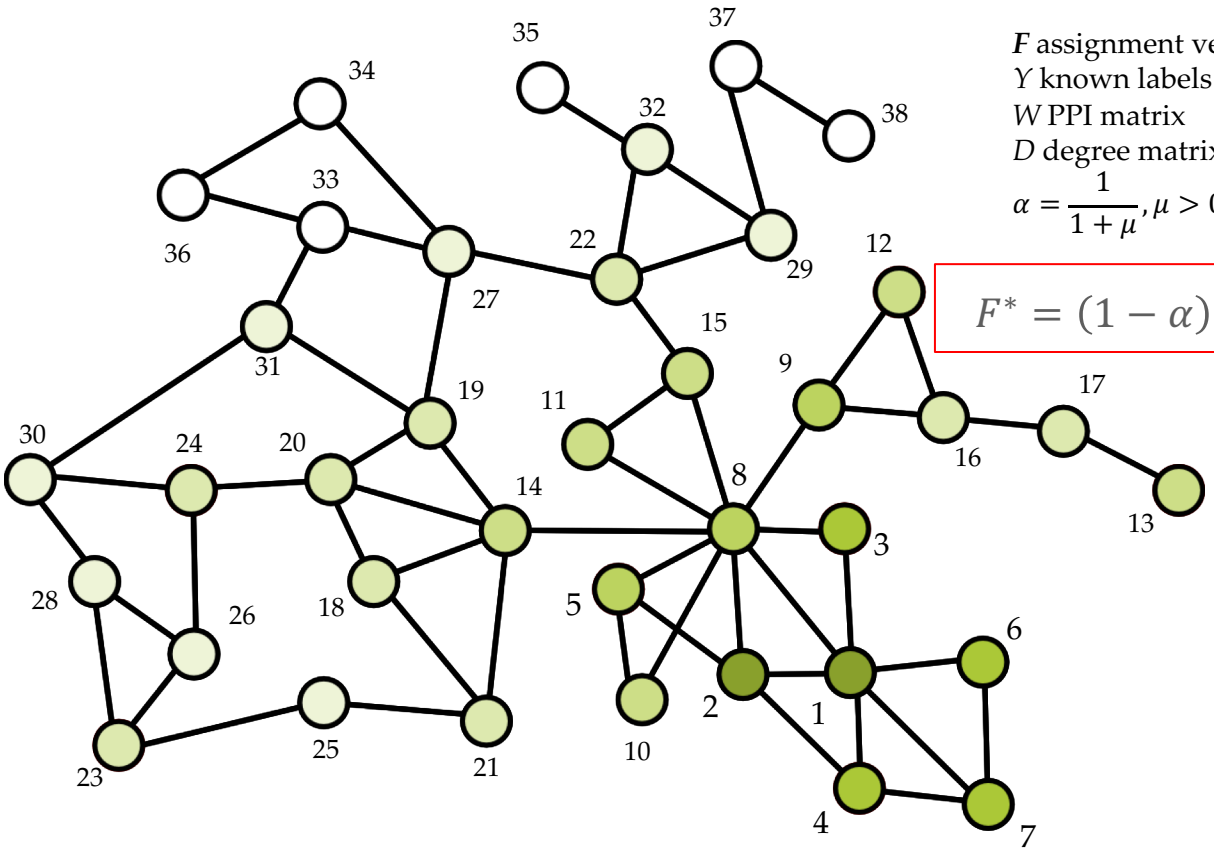
For a given disease, **the soft label is related to the probability for that gene to be a disease gene for that disease**.

Interacting nodes have similar labels

Preserve initial labelling

$$F^* = \arg\min_F Q(F)$$

$$Q(F) = \frac{1}{2}\left(\sum_{i,j=1}^{n} W_{ij} \left\| \frac{1}{\sqrt{D_{ii}}} F_i - \frac{1}{\sqrt{D_{jj}}} F_j \right\|^2 + \mu \sum_{i=1}^{n} \|F_i - Y_i\|^2 \right)$$



F assignment vector
Y known labels
W PPI matrix
D degree matrix of W
$\alpha = \frac{1}{1+\mu}, \mu > 0$

(Zhou et al, NIPS 2004, "Consistency" method)

$$F^* = (1-\alpha)\left(I - \alpha D^{-1/2} W D^{-1/2}\right)^{-1} Y$$

Testing Setup

Disease categories

Uncharted diseases

Currently there are no known disease genes

Charted diseases

Some disease genes are know

Experiment types

Prospective evaluations

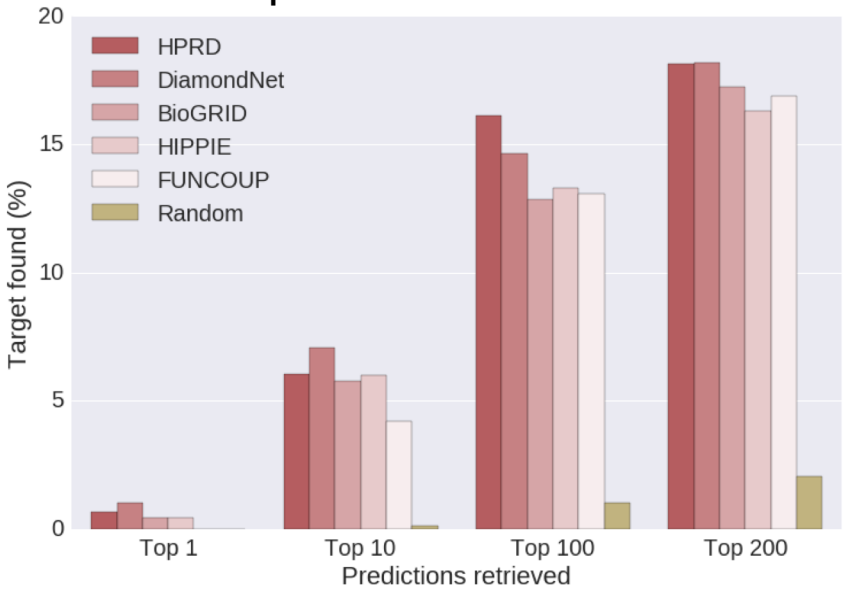
Using information from 2013, predict new disease genes known in 2018

Leave-one-out

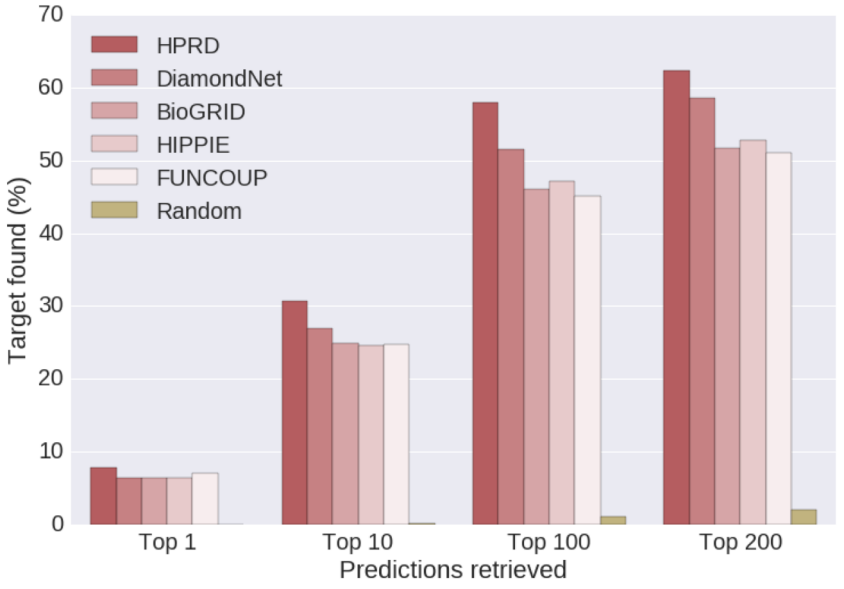
Using data from 2018, a single association is removed and is predicted back

Performance – uncharted diseases

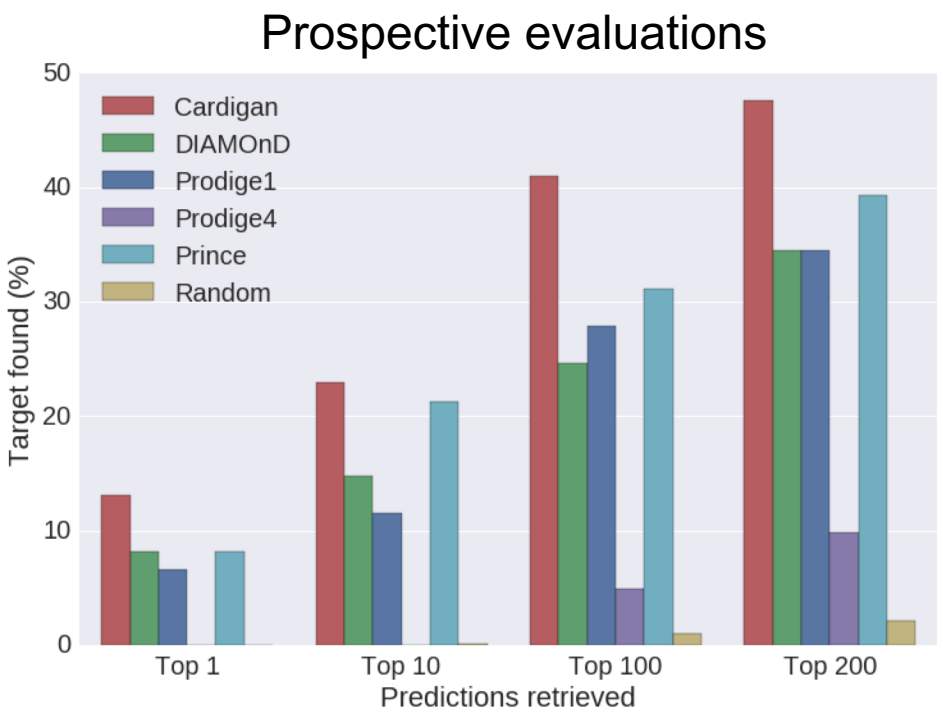
Prospective evaluations

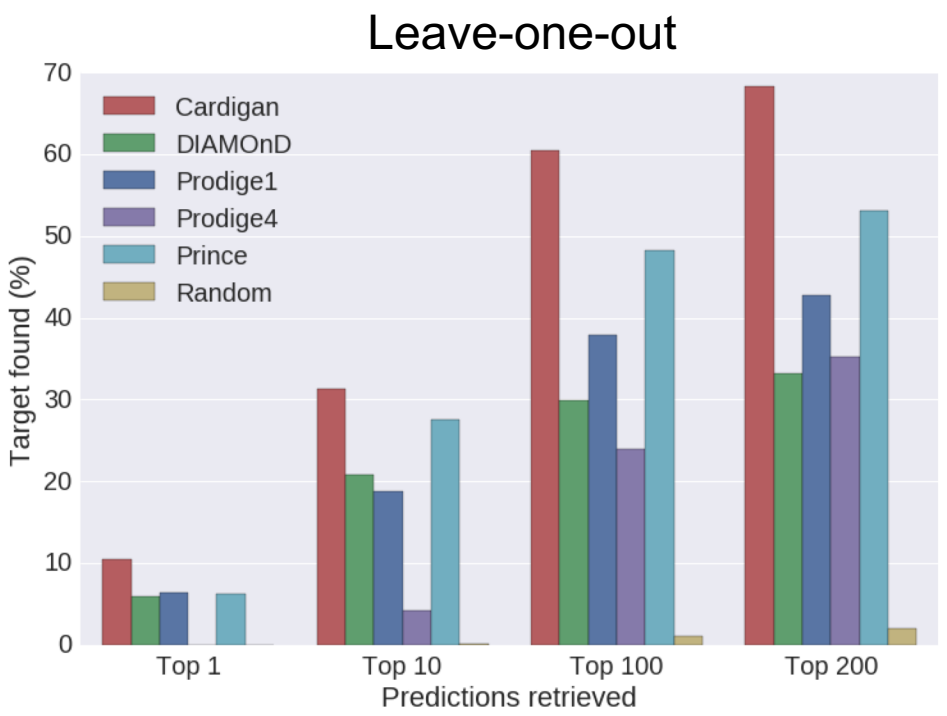


Leave-one-out



Performance – charted diseases





DIAMOnD -- *Ghiassian, Menche, Barabasi, PLoS Comp Bio 2015*
Prodige1,4 -- *Mordelet, Vert, BMC Bioinformatics, 2011*
Prince -- *Vanunu, Magger, Ruppin, Shlomi, Sharan, PLoS Comp Bio 2010*

Prospective evaluation -- Examples

Disease	2013 Status	Gene	Our Ranking	Paper
Familial Retinal Arteriolar Tortuosity (MIM:180000)	Uncharted	COL4A1	5	*Zenten J. et al. , Graefe's Arch. Clin. Exp Ophthalmology 252, 2014*
Ablepharon-macrostomia syndrome (MIM:200110)	Uncharted	TWIST2	10	*Marchegiani et al., American J. of Human Genetics 97, 2015*
Fetal Akinesia Deformation Sequence (MIM:208150)	Charted	MUSK	1	*Tan-Sindhunata et al. , Eur. J. Human Genetics 23, 2015*
Schimmelpenning-Feuerstein-Mims syndrome (MIM:163200)	Charted	NRAS	1	*Lim et al. , Human molecular genetics 23, 2014*

Conclusions of Part 1 & 2

- ✓ A distance between disease modules on the interactome which uses exclusively disease phenotype information.
- ✓ How diffusion methods + our disease similarity measure can be used to infer disease genes for uncharted diseases.
- ✓ These methods can provide **explanations**

3. A collaborative model for predicting the frequency of drug side effects

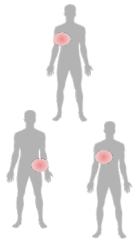
[Galeano, Paccanaro -- BioRxiv 594465, doi: 10.1101/594465]

Drugs side effects

A drug-side effect association in humans can be:

Very rare:	< 0.01%
Rare:	< 0.1%
Infrequent:	< 1%
Frequent:	< 10%
Very frequent:	> 10%

Placebo-controlled study
One disease
Limited size

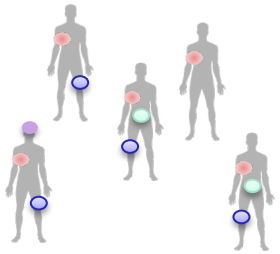


Clinical Trials
Phase I-III
(Premarketing)

Post-marketing Surveillance
Systems
(FAERS-FDA)

FDA-approved (In-market)

Observational study
Multiple diseases
Multiple medications

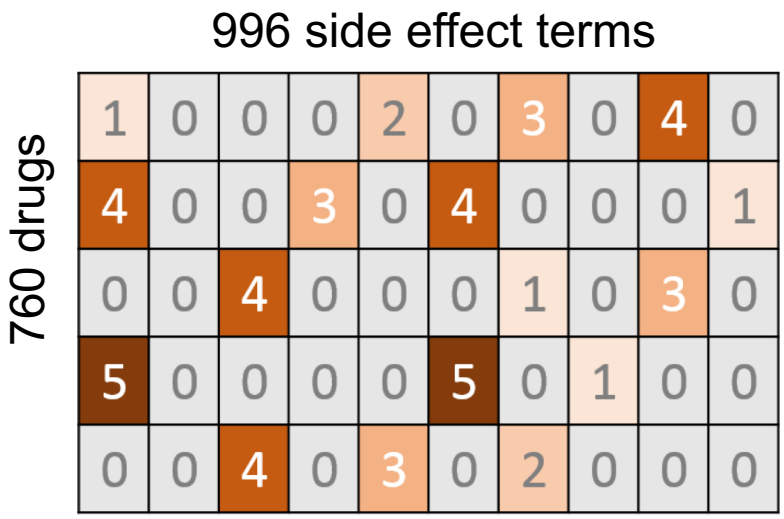


Question

Can we predict the frequency of drug side effects ?

Few methods exists which are aimed at predicting the presence/absence of side effects. These exploit molecular or cellular features.

The data

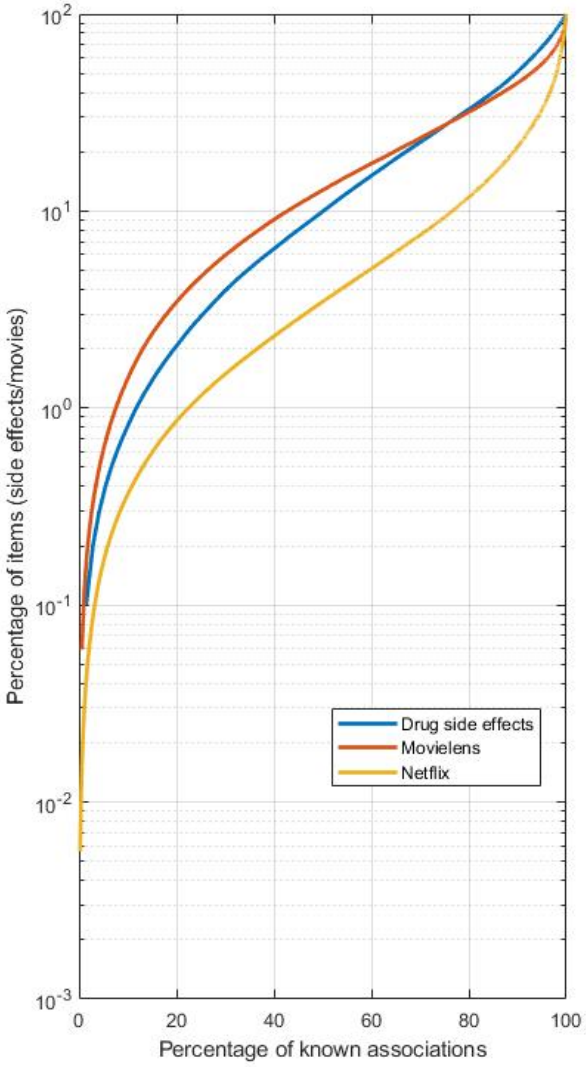


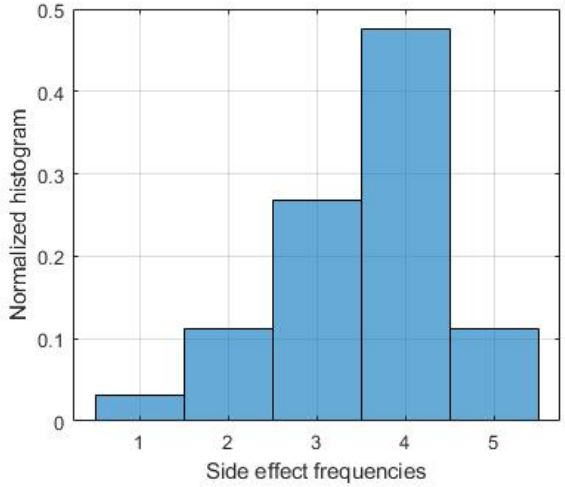
density ~ 5% (sparse)

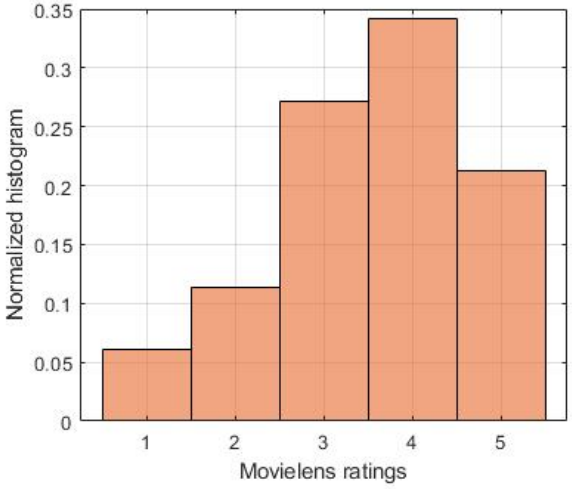
```
Very rare = 1
Rare = 2
Infrequent = 3
Frequent = 4
Very Frequent = 5
```

The Side Effect Resource (SIDER) 4.1 [Khun et al., 2015]

Let's look at the data…







How do we predict (recommend) movies?

Movies (q)

Users (p)

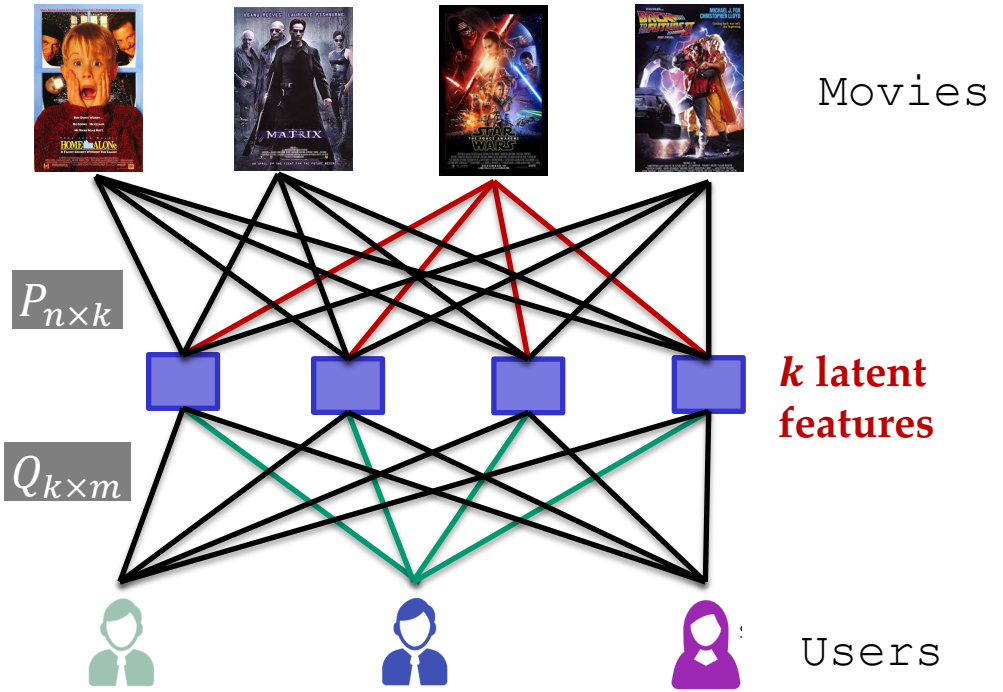
1	0	0	0	2	0	3	0	4	0
4	0	0	3	0	4	0	0	0	1
0	0	4	0	0	0	1	0	3	0
5	0	0	0	0	5	0	1	0	0
0	0	4	0	3	0	2	0	0	0

Y

Matrix decomposition models are useful for very sparse datasets with potential **latent features**

$$Y_{i,j} \approx \boldsymbol{p}_i^T \cdot \boldsymbol{q}_j$$

$$Y_{n\times m} \approx P_{n\times k} \cdot Q_{k\times m}$$



Our idea: recommending side effects to drugs

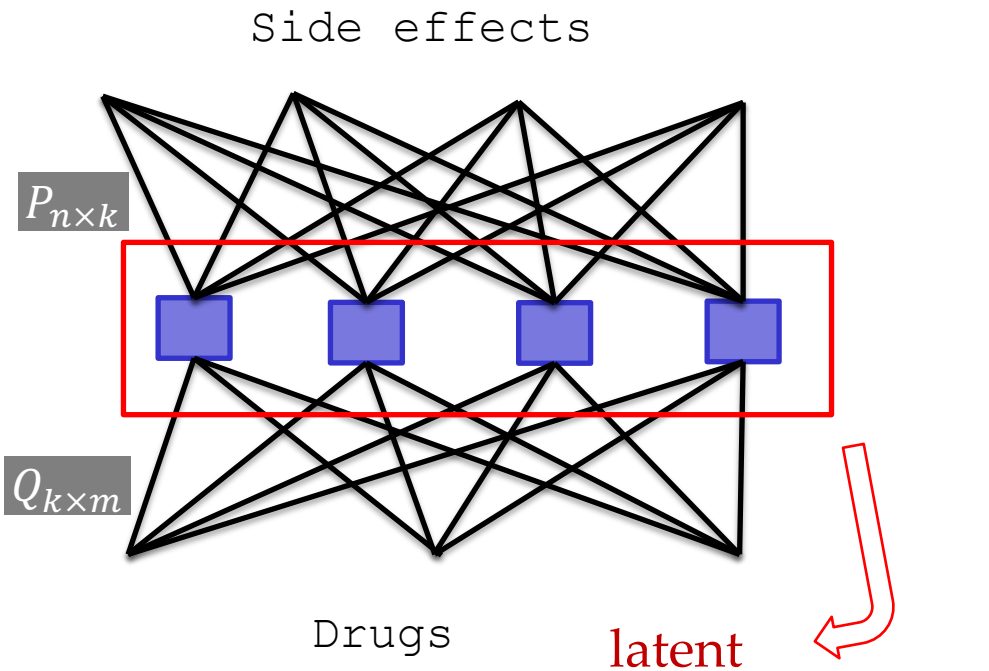
996 side effect terms

760 drugs

1	0	0	0	2	0	3	0	4	0
4	0	0	3	0	4	0	0	0	1
0	0	4	0	0	0	1	0	3	0
5	0	0	0	0	5	0	1	0	0
0	0	4	0	3	0	2	0	0	0

$Y_{n\times m}$

Very rare = 1
Rare = 2
Infrequent = 3
Frequent = 4
Very Frequent = 5



$$Y_{i,j} \approx \boldsymbol{p}_i^T \cdot \boldsymbol{q}_j$$

$$Y_{n\times m} \approx P_{n\times k} \cdot Q_{k\times m}$$

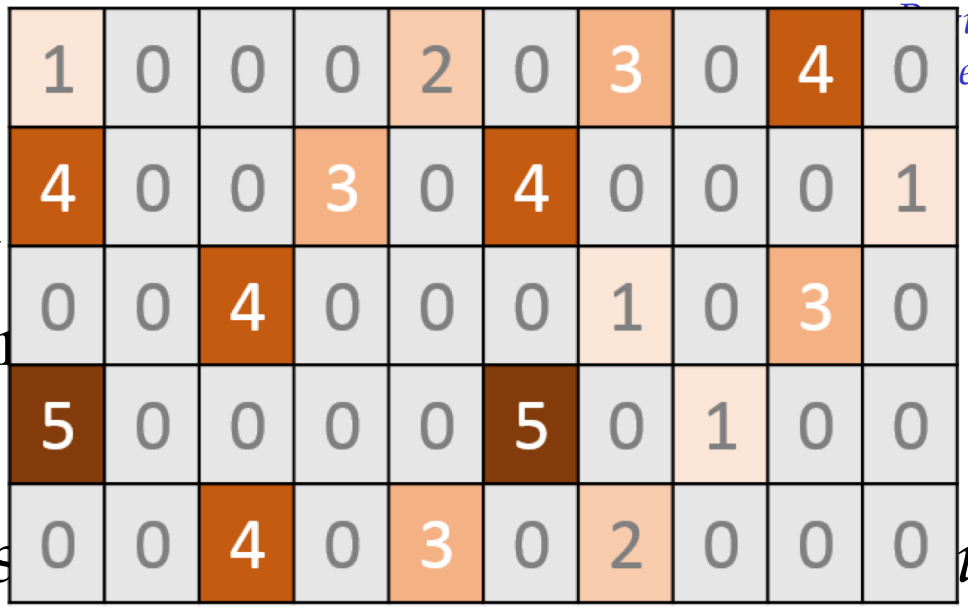
Learning the latent representations

$$\min_{P,Q} J(P,Q) = \frac{1}{2}\| Y - PQ \|_F^2 + \frac{\lambda}{2}(\| P \|_F^2 + \| Q \|_F^2)$$

ularization to ent overfitting

subject

in orde



We learn this *ilar to NMF)*
or with Conjugate Gradient Descent + projections

… it does not work ☹

Our new cost function

$$\min_{W,H \geq 0} J(W,H) = \frac{1}{2} \sum_{Y_{i,j} \in \{1,2,3,4,5\}} (Y_{i,j} - (WH)_{i,j})^2 + \frac{\alpha}{2} \sum_{Y_{i,j} = 0} ((WH)_{i,j})^2$$

Fits clinical trials frequency data

Fits unobserved associations with confidence α_{null}

$Y_{n \times m}$ of n drugs and m side effects
$W_{n \times k}$: drug signatures
$H_{k \times m}$: side effect signatures
$0 \leq \alpha \leq 1$

We are confident on clinical trials data (values 1-5) but only α-confident on the unobserved associations (0s)

Our model uses the large amount of zeros as a regularization

- Small α allows the weights in W and H to grow
- Large α keeps the weights in W and H small and induces sparsity.

Multiplicative Learning algorithm

Our cost function *converges to a local optimum* using the update rules (satisfy the Karush-Kuhn-Tucker conditions):

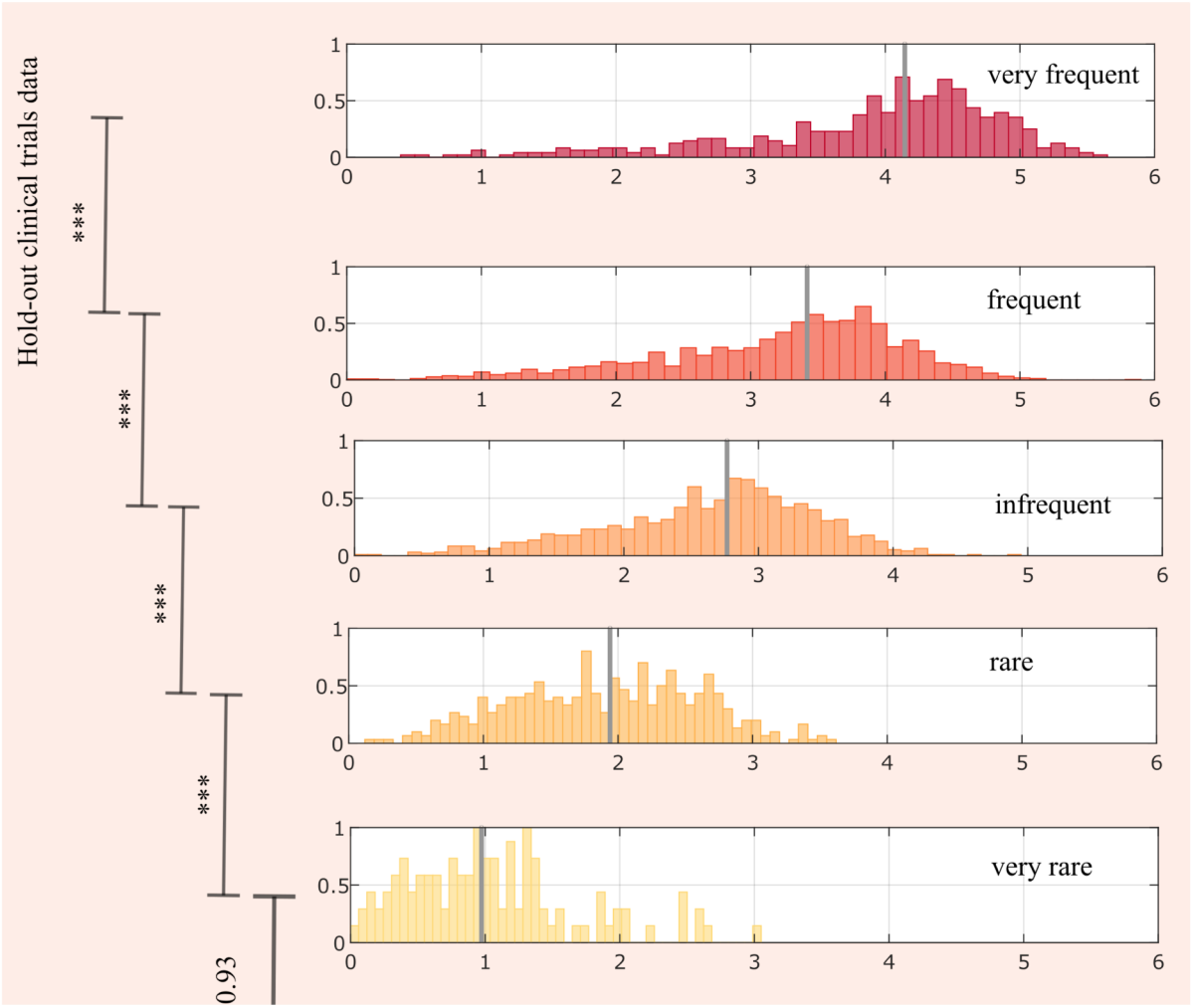
$$W \leftarrow W \circ \frac{P_\Omega(\mathrm{Y})H^T}{\left(P_\Omega(\mathrm{WH}) + \alpha \quad P_{\bar{\Omega}}(\mathrm{WH})\right)H^T}$$

$$H \leftarrow H \circ \frac{W^T P_\Omega(\mathrm{Y})}{W^T\left(P_\Omega(\mathrm{WH}) + \alpha \quad P_{\bar{\Omega}}(\mathrm{WH})\right)}$$

P_Ω: selection function for entries {1,2,3,4,5}
$P_{\bar{\Omega}}$: selection function for entries {0}
$\circ$ is the Hadamard product

Multiplicative learning rule – no learning rate, no projection function

Inspired by non-negative matrix factorization (NMF) [Lee, Seung, Nature, 1999]



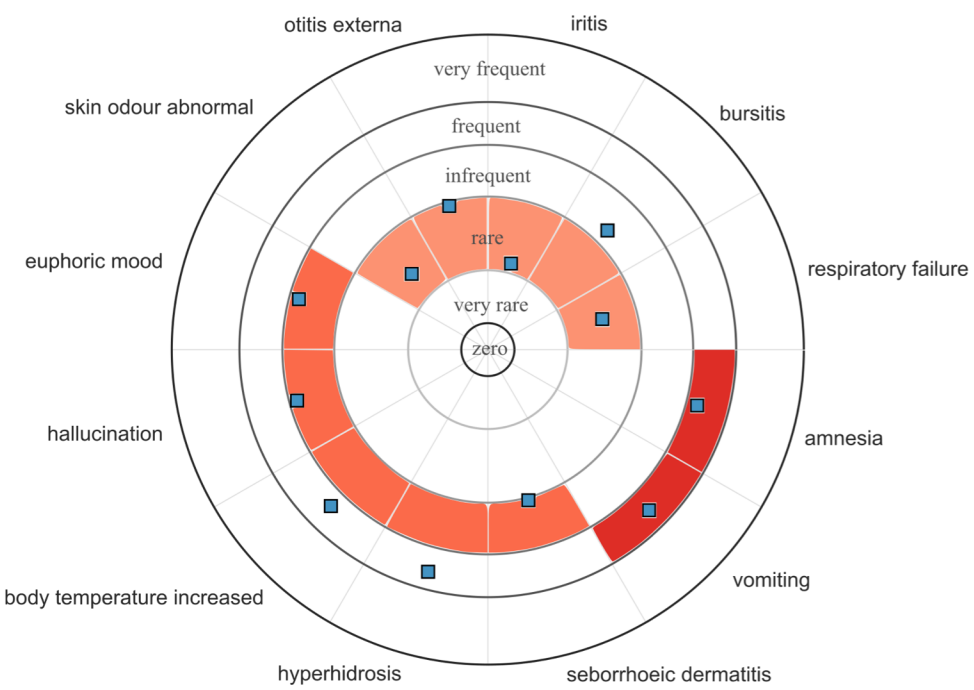
Prediction on Test Set

Higher predicted values correspond to higher side effect frequencies

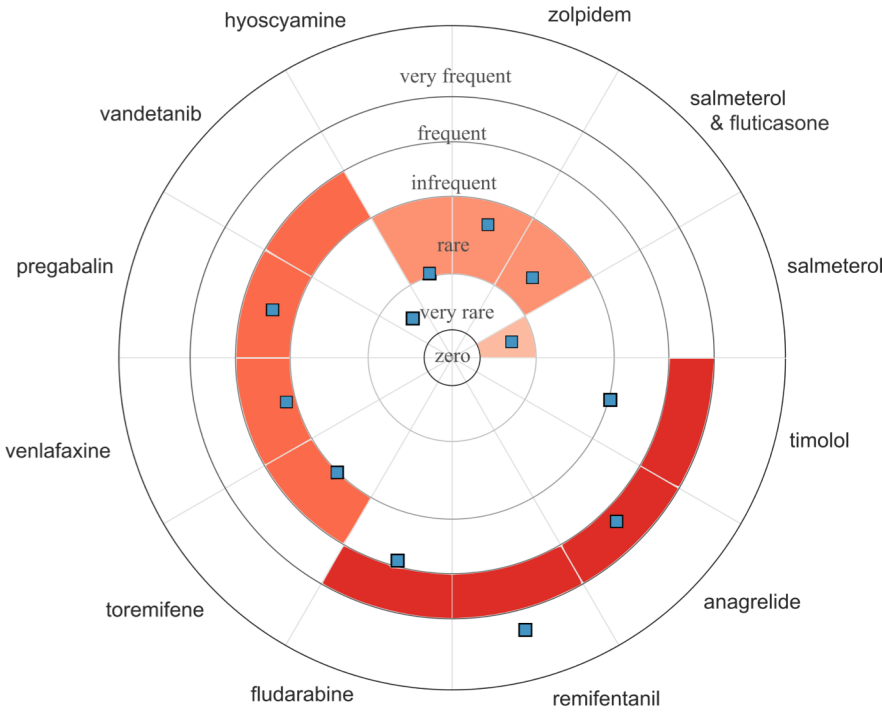
No significant differences between the predicted scores for the **very rare** side effects and the **post-marketing** side effects

Examples

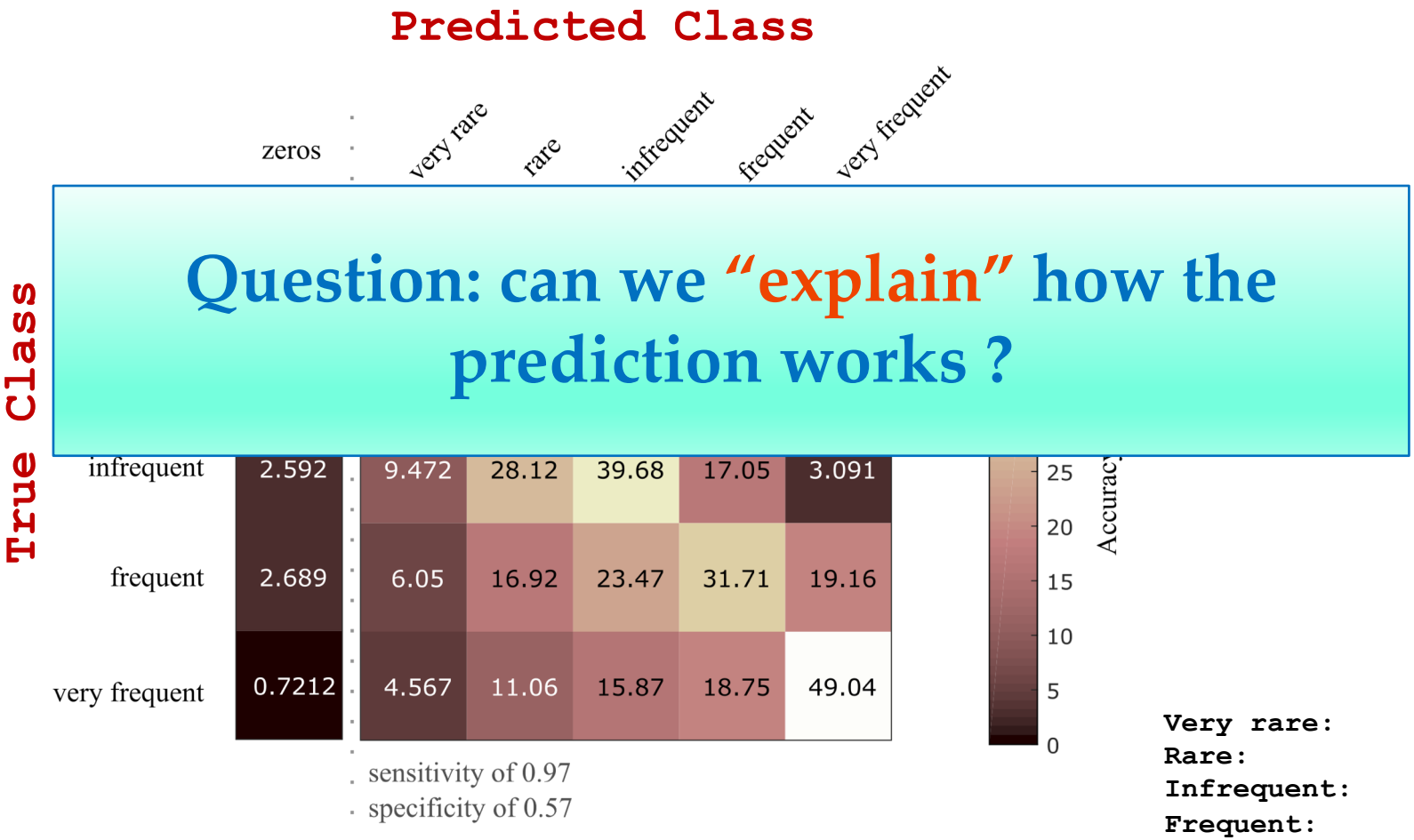
Gabapentin (anticonvulsant drug)



Arrhythmia (cardiovascular side effect)



Percentage of accuracy at predicting the frequency class of drug side effects

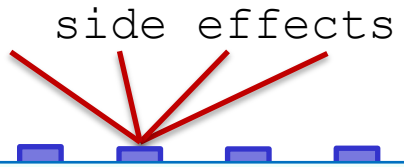


Question: can we "explain" how the prediction works ?

Very rare:	**<0.01%**
Rare:	**<0.1%**
Infrequent:	**<1%**
Frequent:	**<10%**
Very frequent:	**>10%**

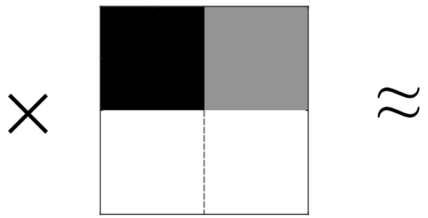
Predictions can be *explained* in terms of the latent features

***Example**: Atorvastatin is known to cause frequent respiratory and thoracic-related side effects*



Question: do the latent representations tell us something about the *biology* of the problem?

sinusitis pharyngitis bronchitis urinary tract infection rhinitis	diarrhoea dermatitis rash abdominal pain gastrointestinal pain
application site pain application site erythema erythema application site pruritus skin exfoliation application site burn eye irritation scab	personality disorder neurosis tenosynovitis muscle contractions involuntary tongue disorder hostility hyporeflexia hernia



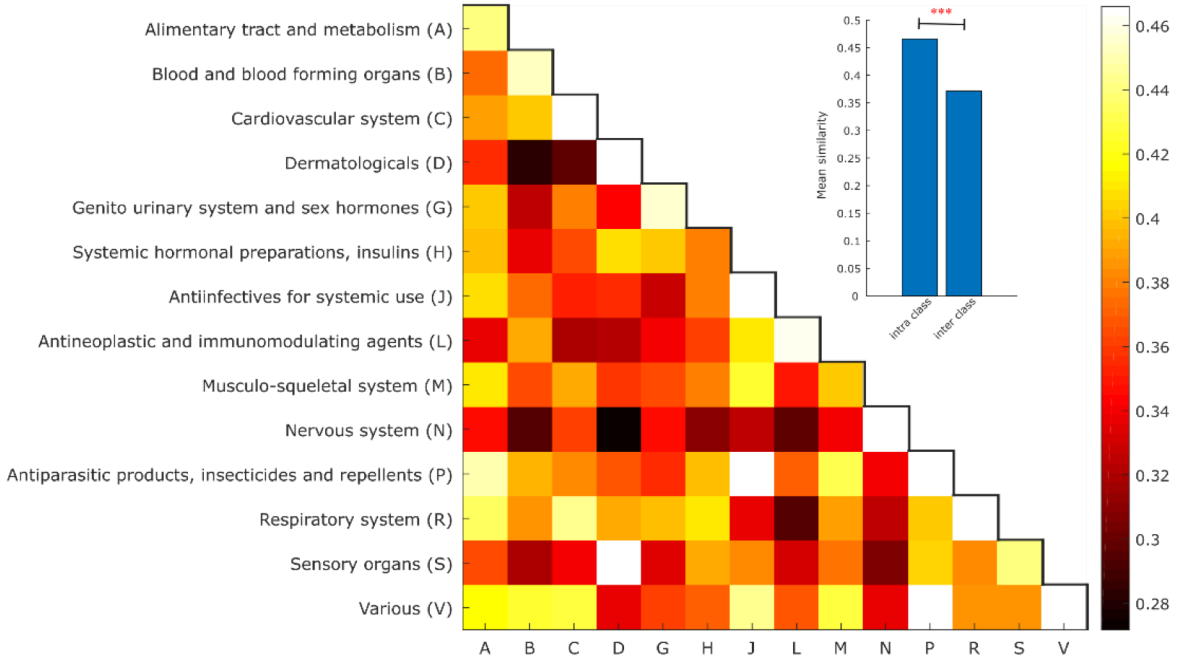
upper respiratory tract infection (4.45)
headache (4.40)
nasopharyngitis (4.20)
cough (4.04)
diarrhoea (4.01)
musculoskeletal discomfort (3.90)
abdominal pain (3.86)

Drug signature are related to clinical activity of the drug

Hierarchical categorization of drugs according to ATC (from WHO):

1. Anatomical
2. Therapeutic
3. Pharmacological
4. Chemical

Anatomical class



Drug signature similarity predicts drug clinical activity

Anatomical class | Therapeutic subclass | Pharmacological subclass | Chemical subclass

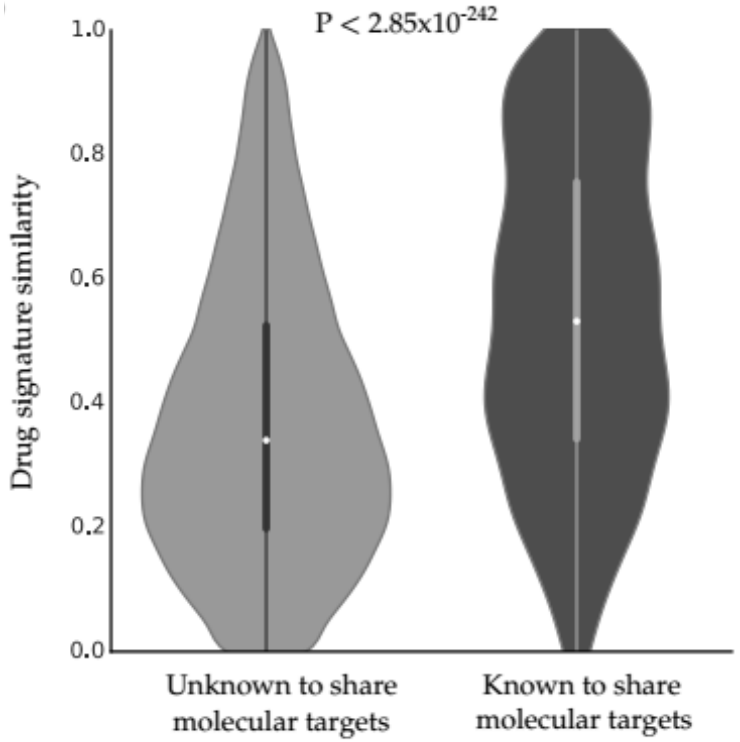
Hierarchical categorization of drugs according to

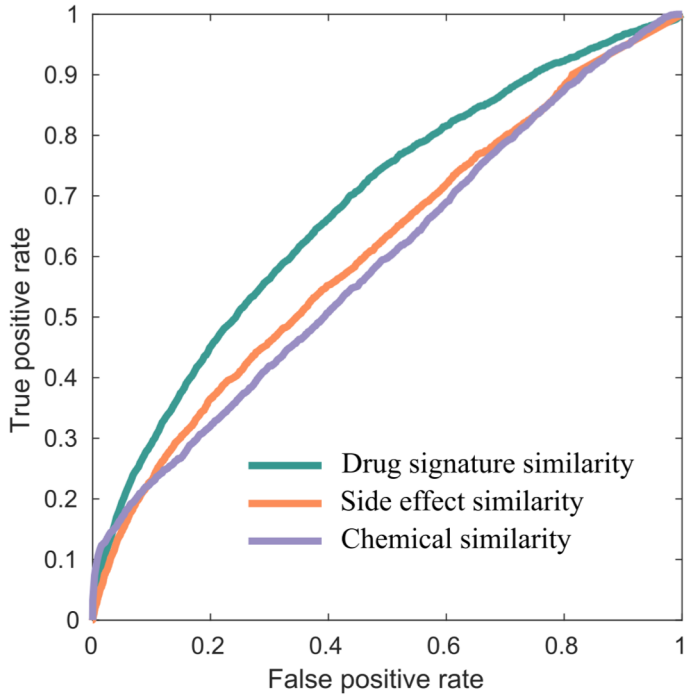
4. Chemical

Question: can we exploit the latent representations for predictions in pharmacology?

Predicting if ***2 drugs share the same category*** *using the drug signature similarity.*

Drug latent representations predict shared targets





There is a significant difference in the cosine similarity between drug ***signatures for pairs that share targets***

Prediction of whether 2 drugs share molecular targets using similarity between drug signatures

Conclusions of Part 3

- ✓ A method for predicting the frequency of side-effects in the population.
- ✓ It tells us something about the biology of the problem
- ✓ It can be used for directing clinical trials.
- ✓ It can provide **explanations**

Acknowledgements

Horacio Caniza

Alfonso E. Romero

One postdoc position available

shared position between Royal Holloway and Yale University

Juan Caceres

Haixuan Yang

http://www.paccanarolab.org

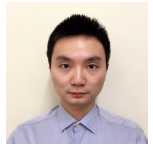
Mateo Torres

Santiago Noto

BBSRC bioscience for the future

EPSRC Engineering and Physical Sciences Research Council

Diego Galeano



Cheng Ye

MARIE CURIE ACTIONS

CONACYT

NSF

Ruben Jimenez

Jessica Gliozzo

THE ROYAL SOCIETY