Predicting Drug-Target Interaction for New Drugs Using Enhanced Similarity Measures and Super-Target Clustering

Jian-Yu Shi (NWPU)

Siu-Ming Yiu (HKU)

Yiming Li (HKU)

Henry C.M. Leung (HKU)

Francis Y.L. Chin (HKU)

神農

(*shen nong; "divine farmer"*)

Shen Nong's Organs

Interaction

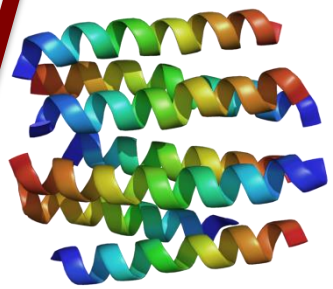
Activity related to

Treatment

Herbs

Disease

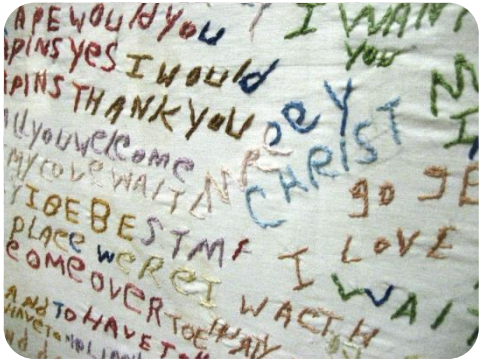
Protein



Interaction

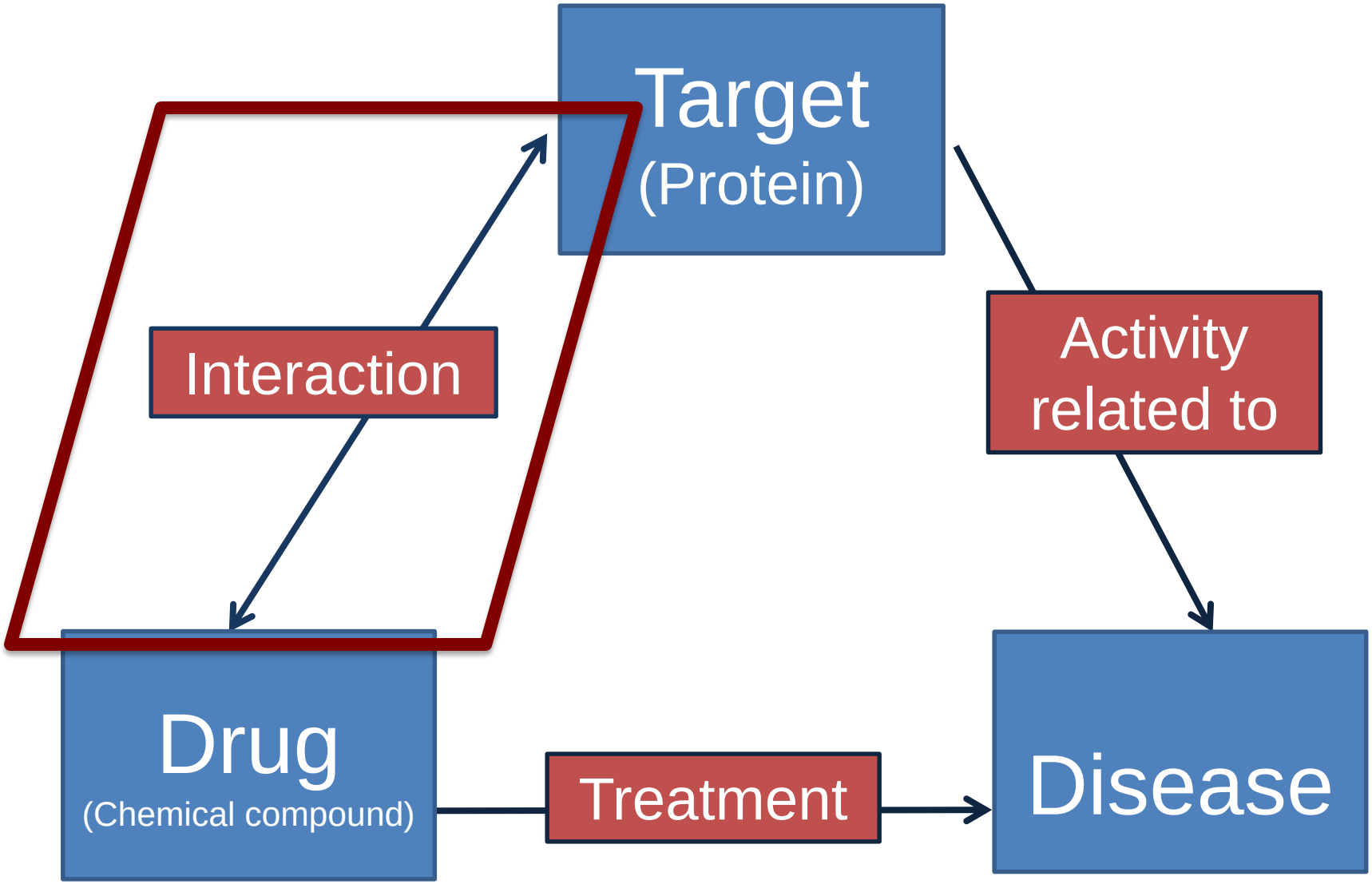
Activity related to

Treatment



Drug

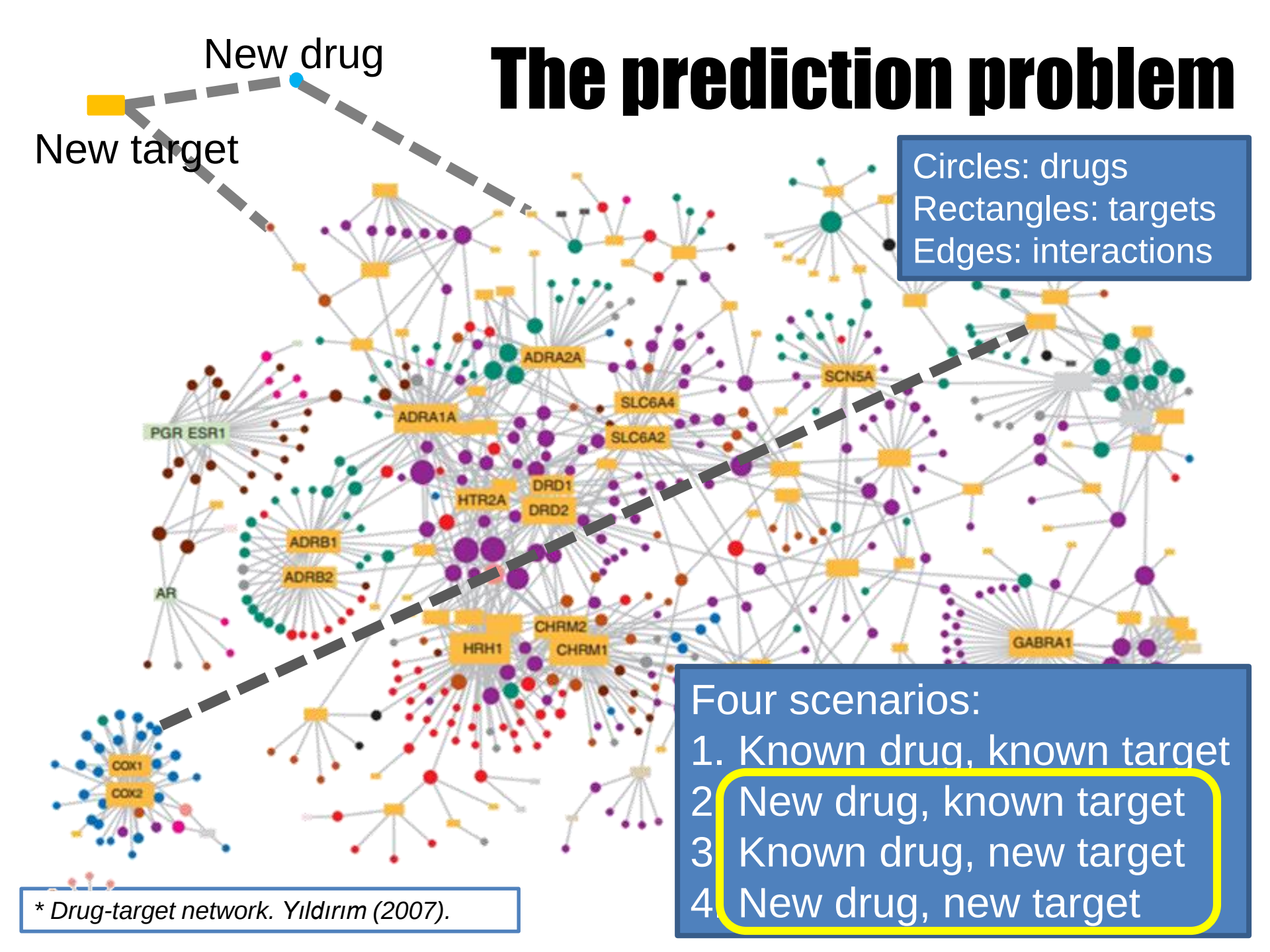
Schizophrenia (a disease)



Drug discovery:

Predicting drug-target interaction is the key!

The prediction problem



Circles: drugs
Rectangles: targets
Edges: interactions

Four scenarios:

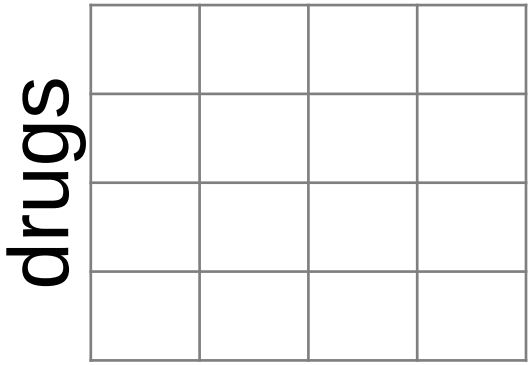
1. Known drug, known target
2. New drug, known target
3. Known drug, new target
4. New drug, new target

** Drug-target network. Yıldırım (2007).*

Input

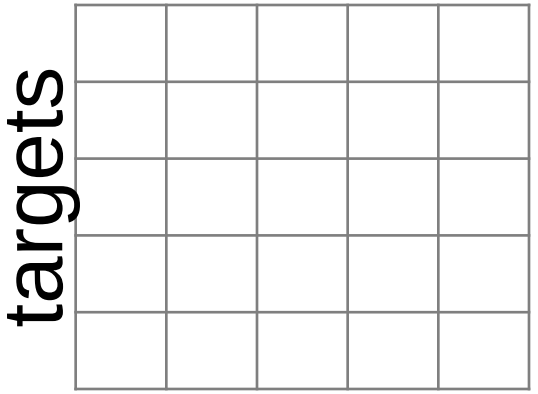
Drug **similarity**

drugs

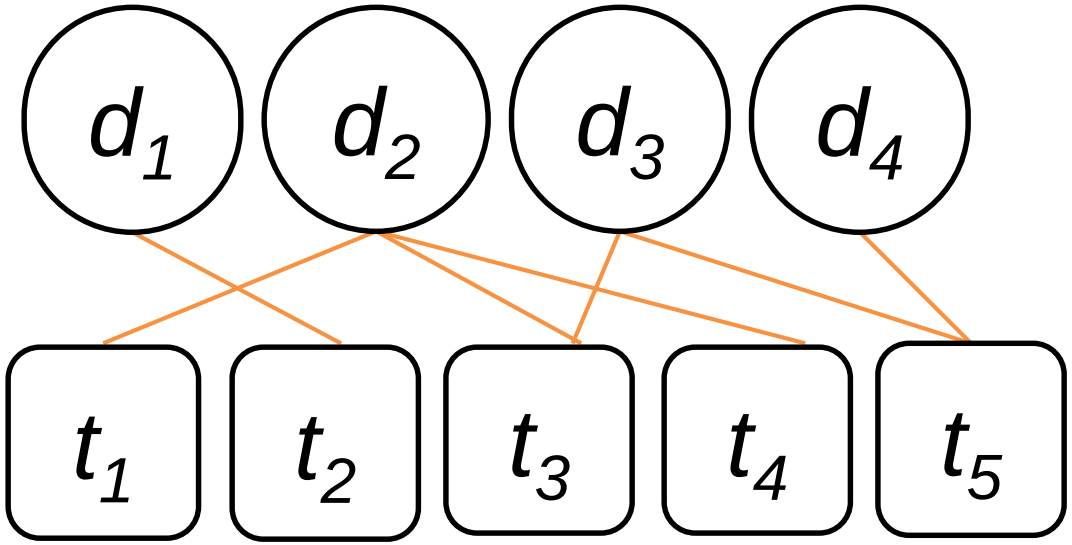


Target **similarity**

targets



Drug-target **interaction**

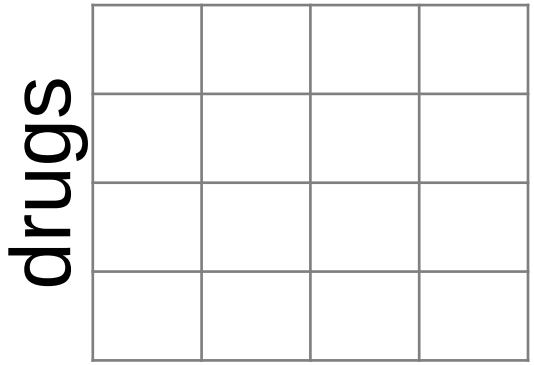


	t_1	t_2	t_3	t_4	t_5
d_1	0	1	0	0	0
d_2	1	0	1	1	0
d_3	0	0	1	0	1
d_4	0	0	0	0	1

Input

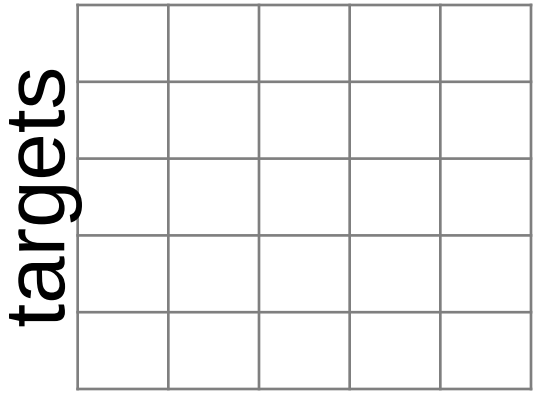
Drug **similarity**

drugs



Target **similarity**

targets



Train a model for prediction

Drug-target **interaction**

	t_1	t_2	t_3	t_4	t_5
d_1	0	1	0	0	0
d_2	1	0	1	1	0
d_3	0	0	1	0	1
d_4	0	0	0	0	1

Problem with training data: missing interactions

Existing method #1: WNN-GIP

Weighted nearest neighbor – Gaussian interaction profile

(PloS One 2013)

Drug-target **interaction**

Biased!

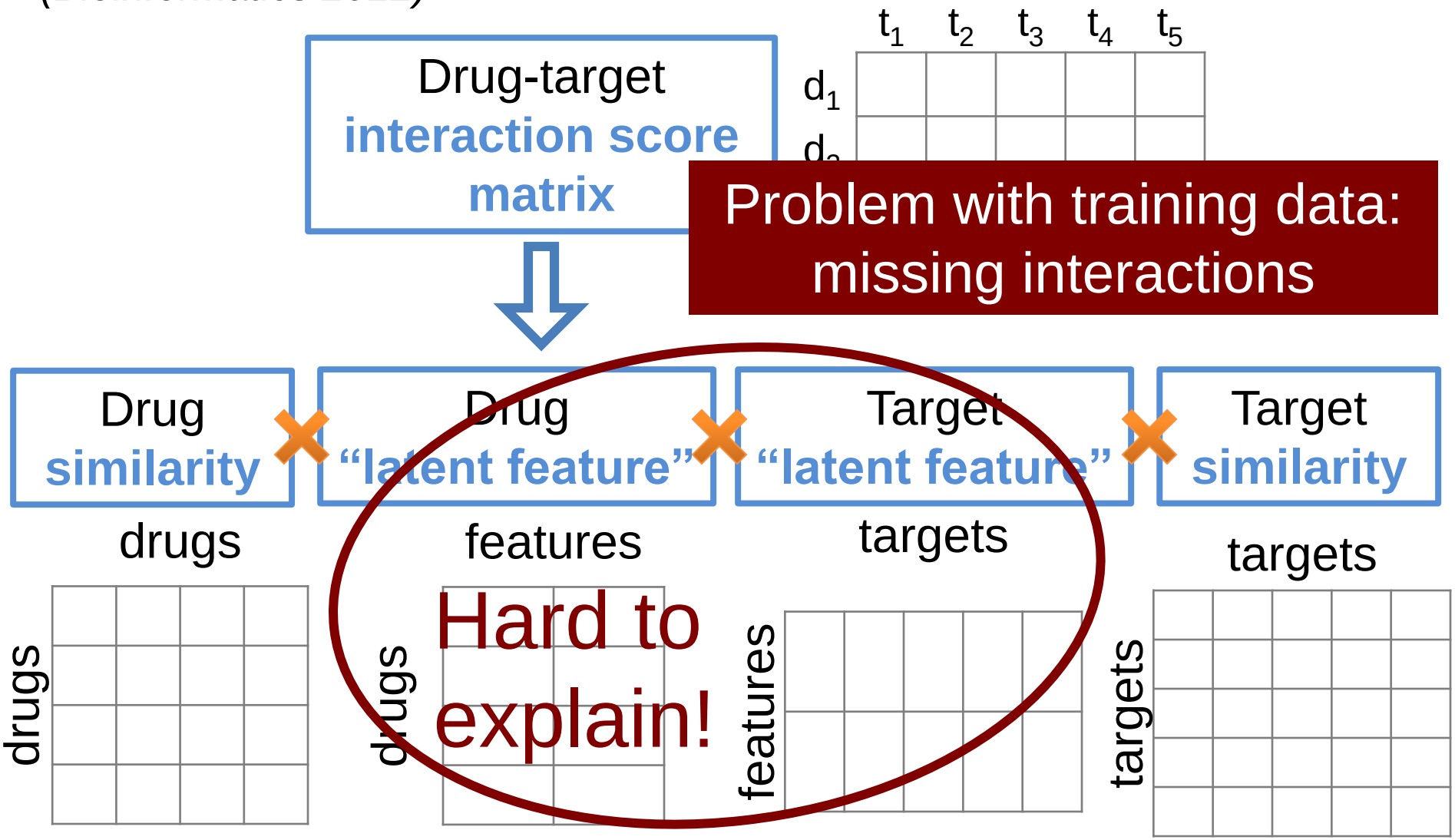
Only uses positive samples to build the model

	t_1	t_2	t_3	t_4	t_5
d_1	0	1	0	0	0
d_2	1	0	1	1	0
d_3	0	0	1	0	1
d_4	0	0	0	0	1

Existing method #2: KBMF2K

Kernelized Bayesian matrix factorization

(Bioinformatics 2012)



Limitations of the existing methods

WNN-GIP and KBMF2K

- Missing interactions
- The similarity measure
  - **Only** based on the **chemical structure** of drugs and **protein sequences** of targets

Drug-target interaction prediction as probabilistic events

The neighbor idea

- A drug's ***neighbors***: the drugs most similar to it
- Predict a new drug's behavior by its neighbors' behavior

The probability

- Event A: to be predicted

  (New) drug ***d*** interacts with target ***t***

- Event B: the observation

  # of ***d***'s neighbors interacting with target ***t***

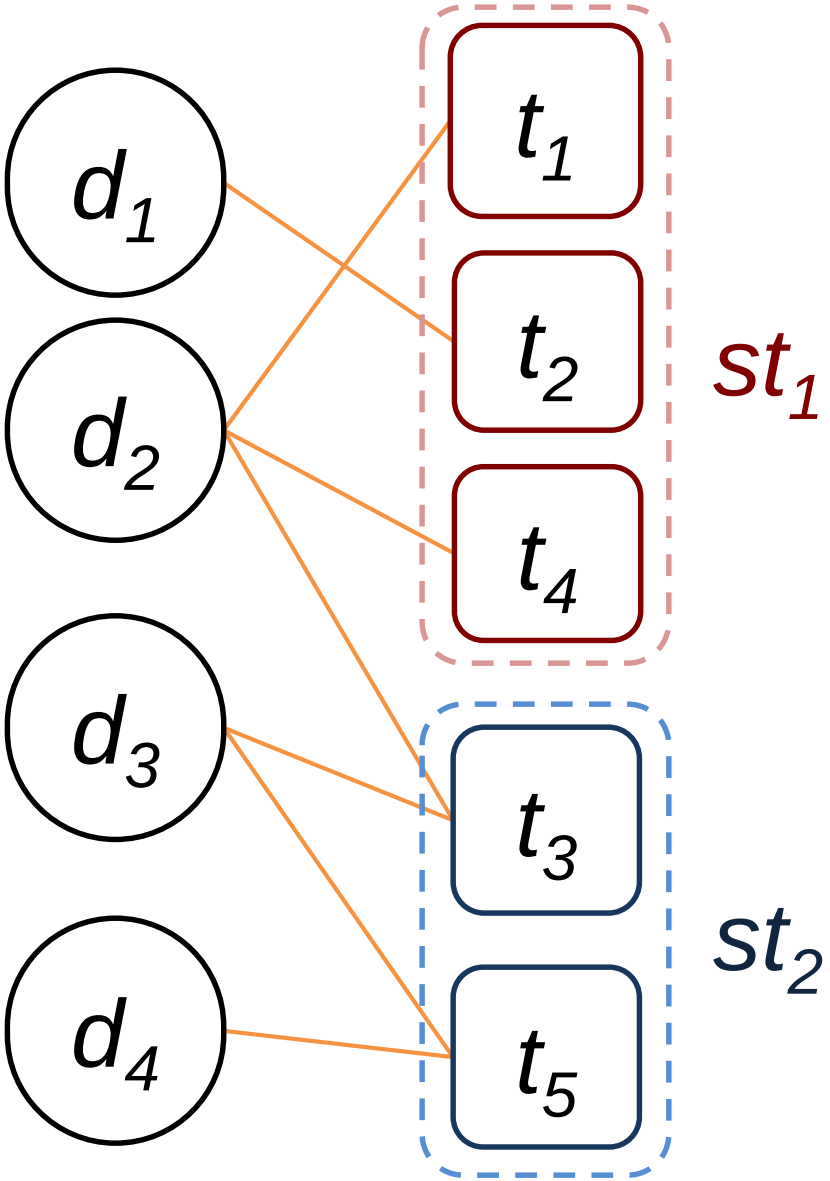
We calculate Pr(A|B) by $\frac{\Pr(AB)}{\Pr(AB) + \Pr(A^C B)}$

Probability of how likely *d* interacts with *t* given the observed number of interactions of *d*'s neighbors with *t*

Our contribution #1

“Super-targets”

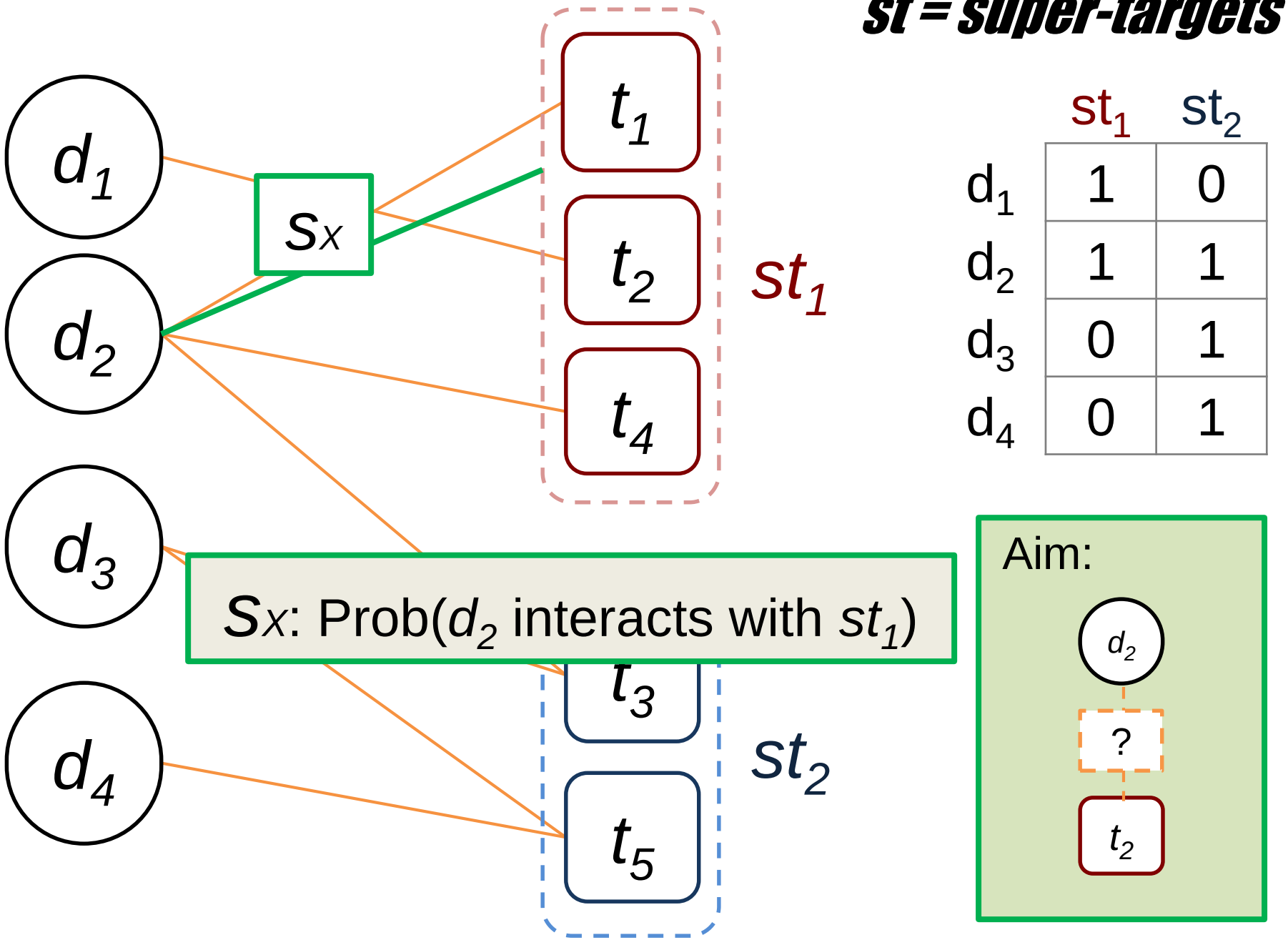
Cluster targets using similarities;
Cluster = Super-target



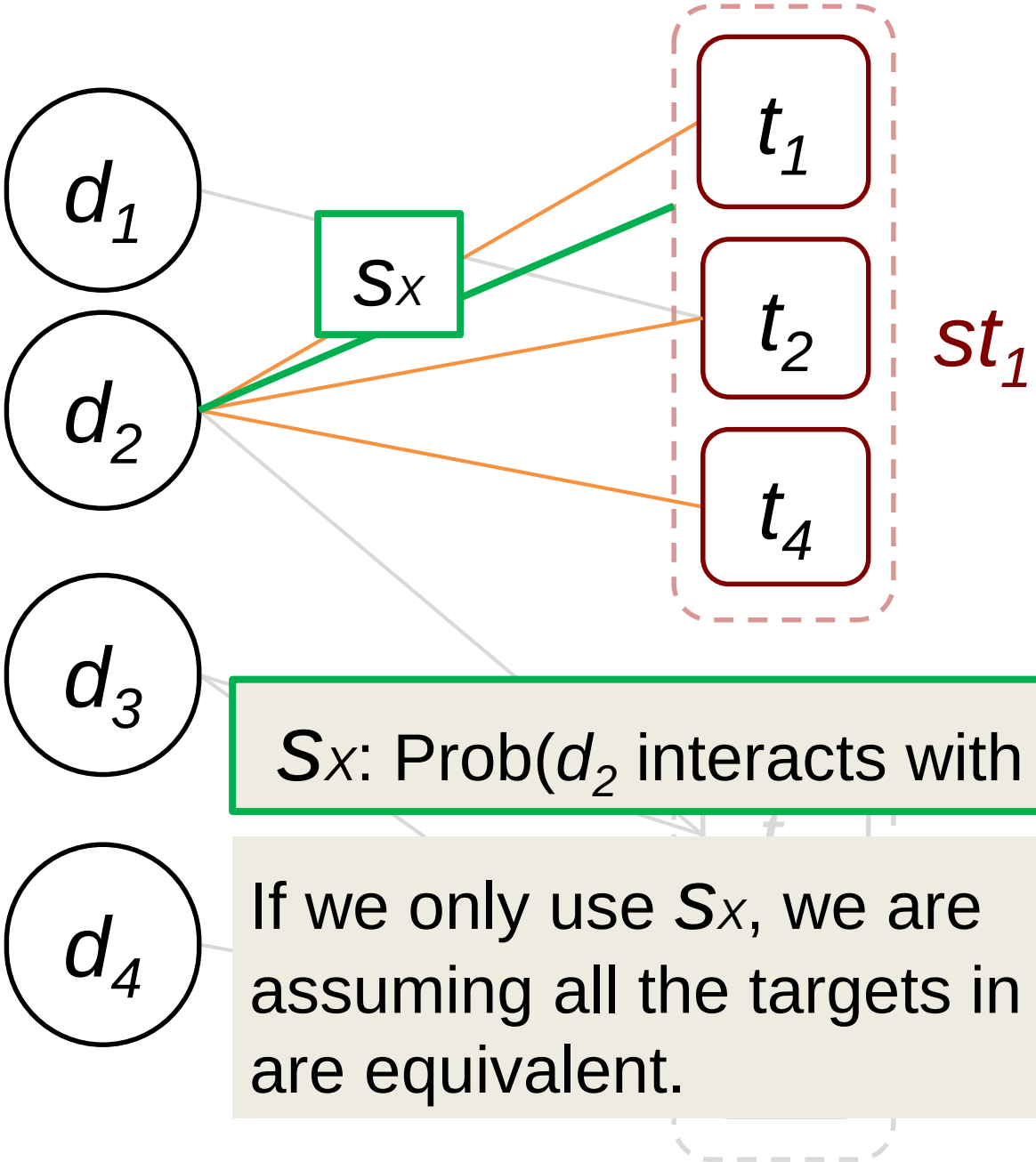
st = super-targets

	t_1	t_2	t_3	t_4	t_5
d_1	0	1	0	0	0
d_2	1	0	1	1	0
d_3	0	0	1	0	1
d_4	0	0	0	0	1

	st_1	st_2
d_1	1	0
d_2	1	1
d_3	0	1
d_4	0	1


st = super-targets
d1
d2
d3
d4
sx
t1
t2
t4
t3
t5
st1
st2
Sx: Prob(d2 interacts with st1)
st1 st2
d1 1 0
d2 1 1
d3 0 1
d4 0 1
Aim:
d2
?
t2

st = super-targets



	st_1	st_2
d_1	1	0
d_2	1	1
d_3	0	1
d_4	0	1

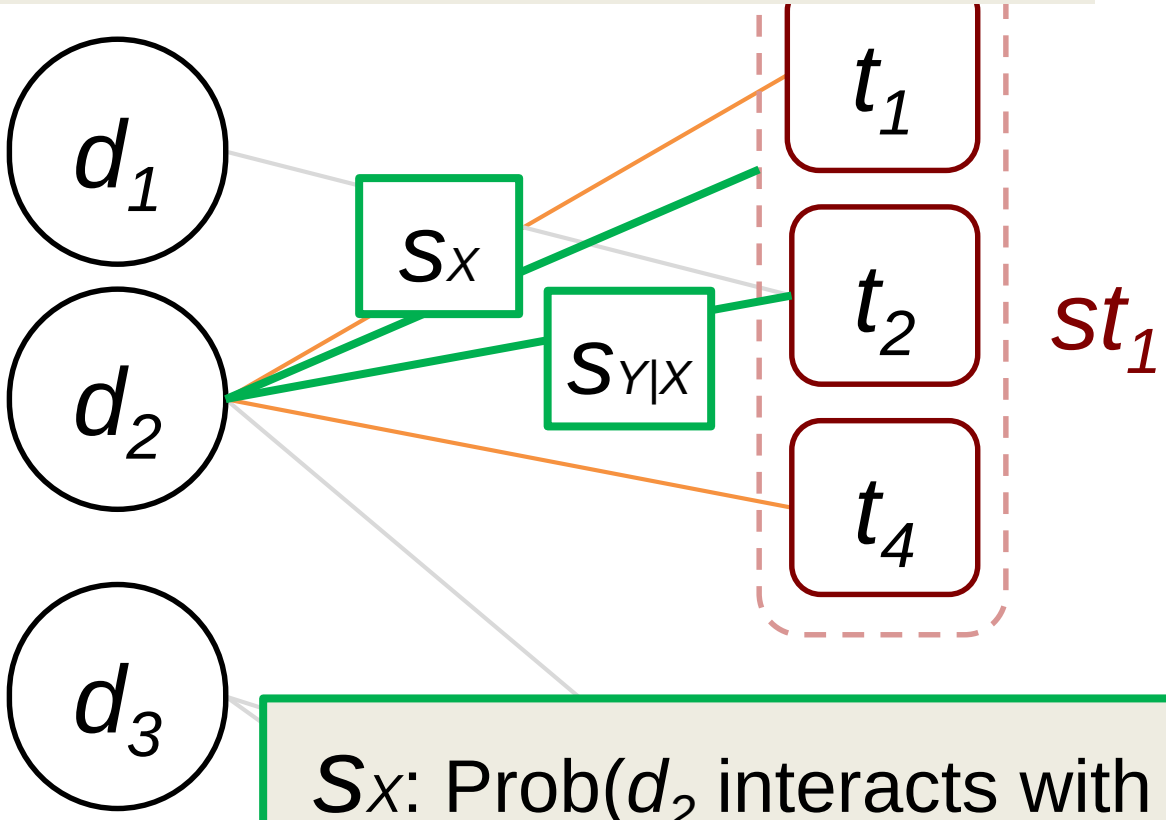
S_x: Prob(d_2 interacts with st_1)

If we only use S_x, we are assuming all the targets in st_1 are equivalent.

Aim:

For new drugs it is the same!

st = super-targets



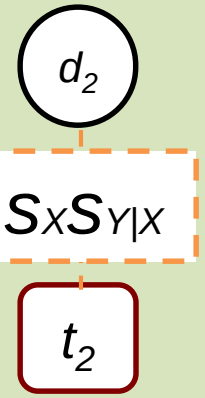
	st_1	st_2
d_1	1	0
d_2	1	1
d_3	0	1
d_4	0	1

S_X: Prob(d_2 interacts with st_1)

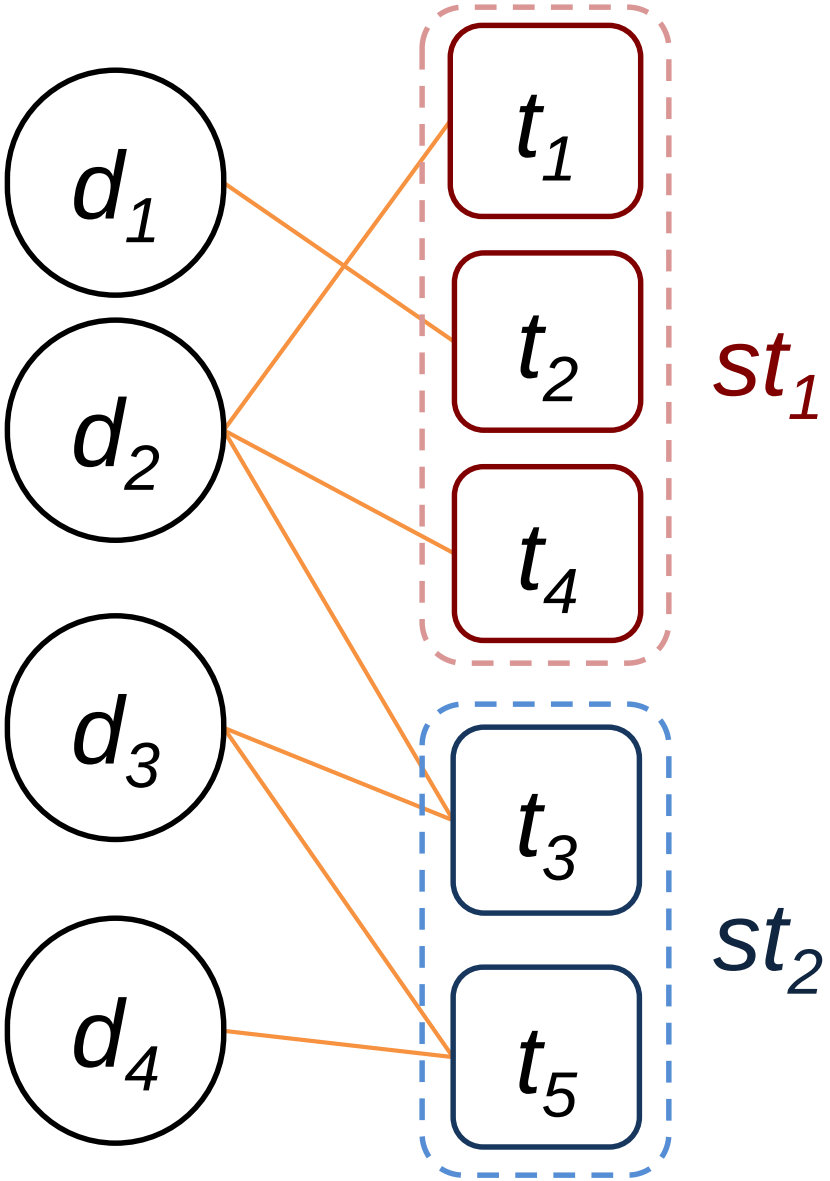
$S_{Y|X}$: Prob(d_2 interacts with t_2 | d_2 interacts with st_1)

$S_X S_{Y|X}$: Prob(d_2 interacts with t_2 in st_1)

Aim:



Cluster targets using similarities;
Cluster = Super-target



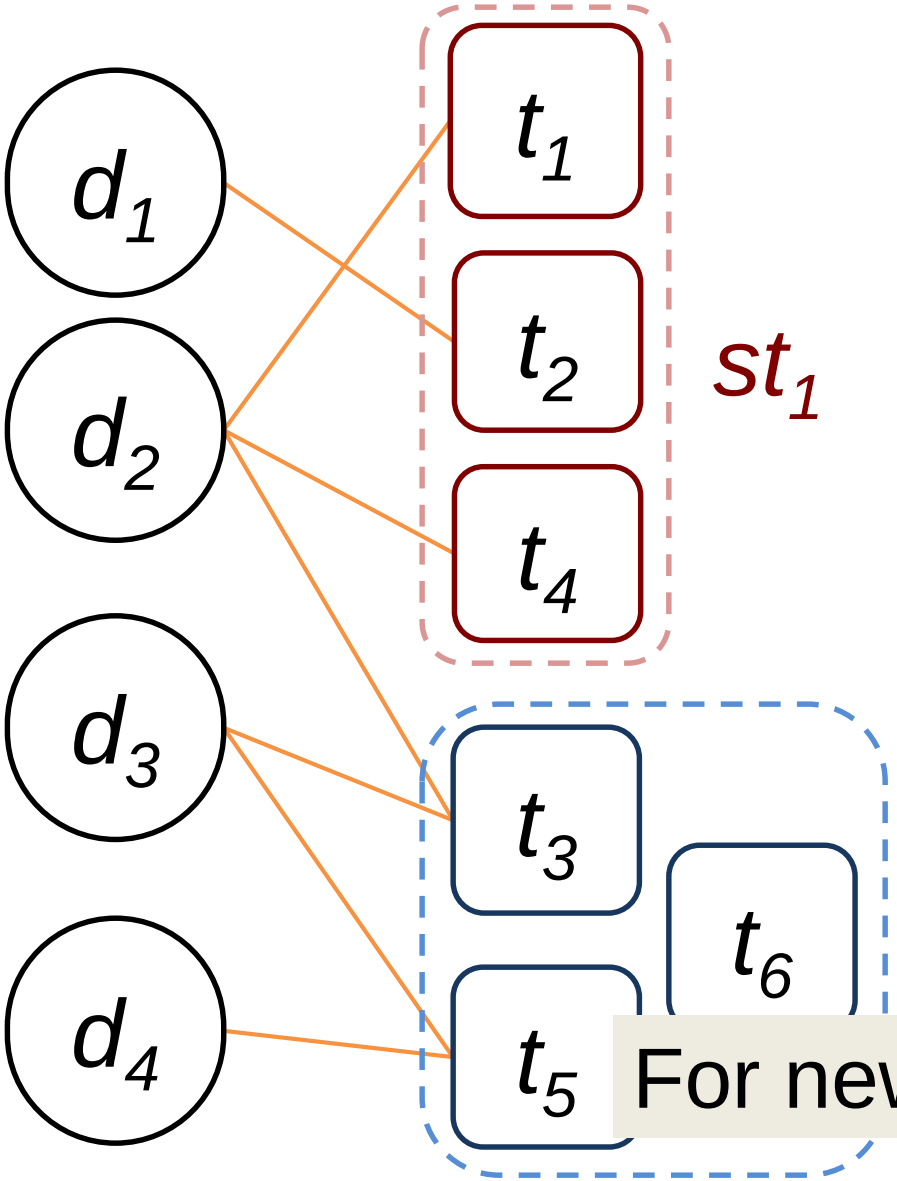
st = super-targets

	t_1	t_2	t_3	t_4	t_5
d_1	0	1	0	0	0
d_2	1	0	1	1	0
d_3	0	0	1	0	1
d_4	0	0	0	0	1

t_6

Cluster targets using similarities;
Cluster = Super-target

st = super-targets



	t_1	t_2	t_3	t_4	t_5
d_1	0	1	0	0	0
d_2	1	0	1	1	0
d_3	0	0	1	0	1
d_4	0	0	0	0	1

A new target could be clustered into one of the super-targets

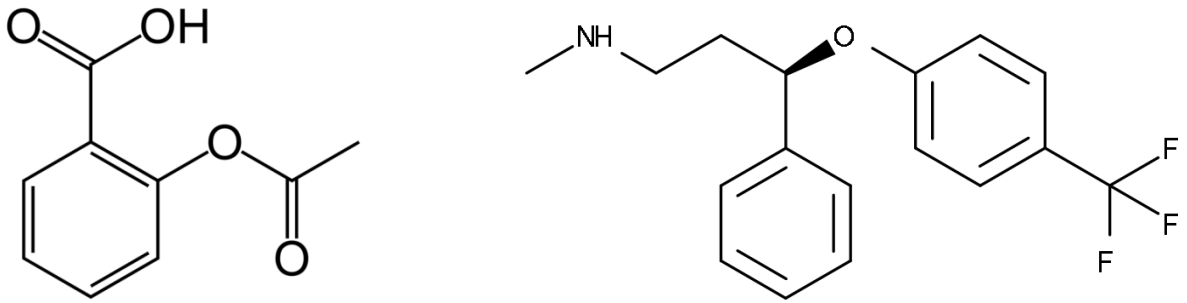
For new drugs it is the same!

Our contribution #2

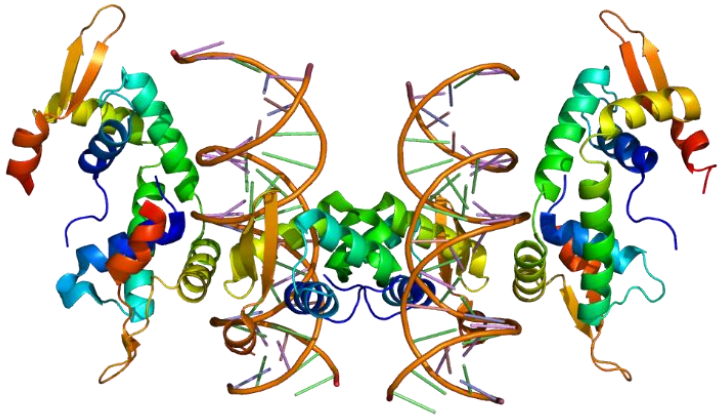
Enhanced similarity measures for drugs and targets

Existing similarity measures

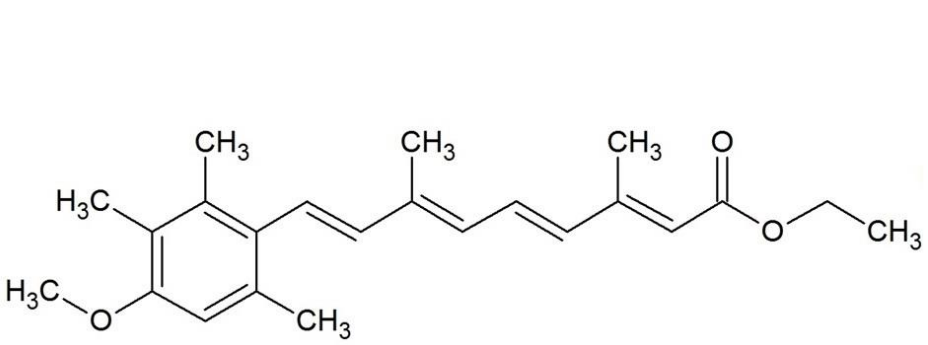
Drugs: aligning the **2D chemical structures**



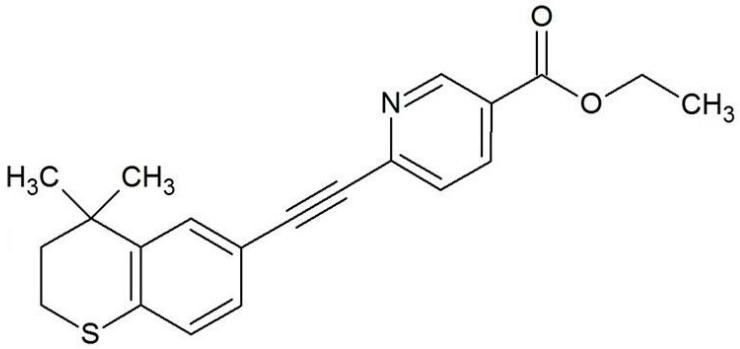
Targets: aligning the **protein sequences**



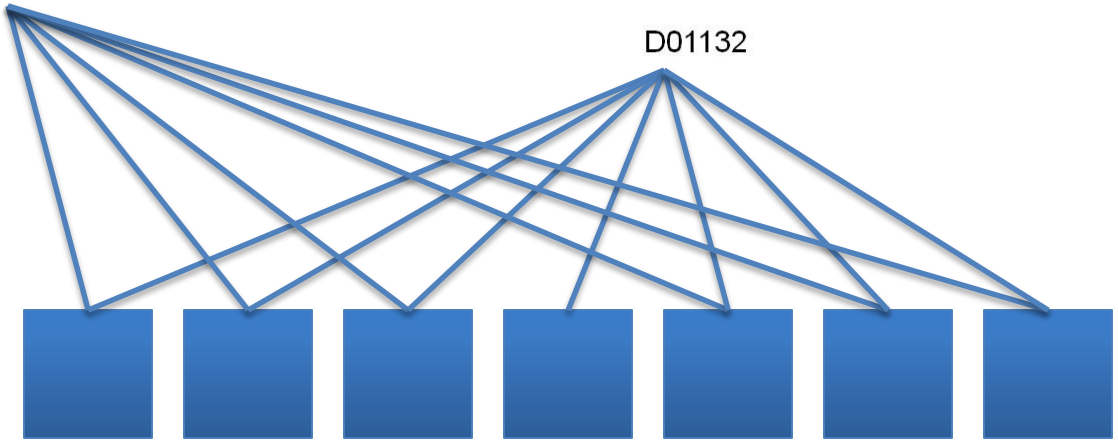
They have low structural similarity (0.275) but share many targets



D00316

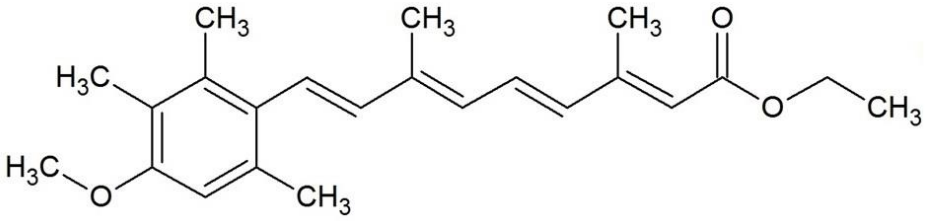


D01132

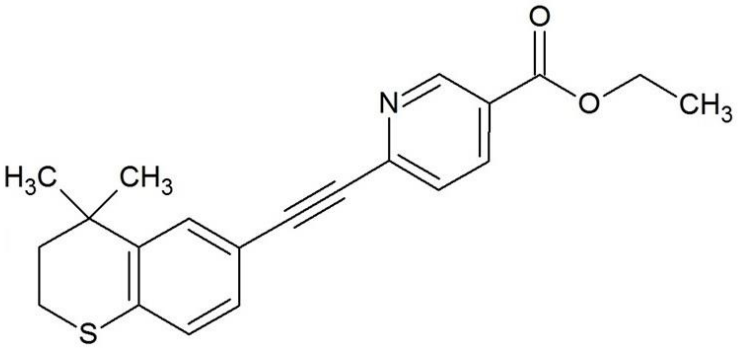


** 2D chemical structures extracted from KEGG.*

They have low structural similarity (0.275) but share many targets



D00316



D01132

Non-structural similarity measures are needed!

** 2D chemical structures extracted from KEGG.*

Anatomical Therapeutic Chemical Classification System

Furosemide

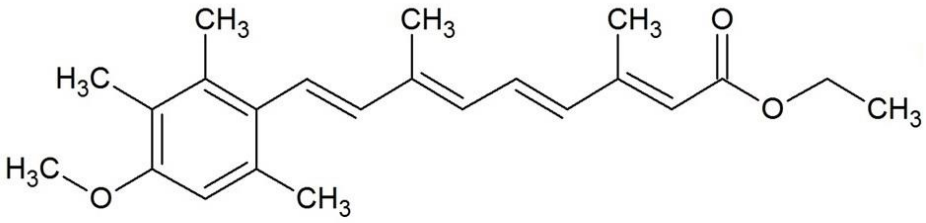
C03CA01

Hierarchical

- Level 5: chemical substance
- Level 4: therapeutic subgroup
- Level 3: therapeutic subgroup
- Level 2: therapeutic main group
- Level 1: organ or system it acts on

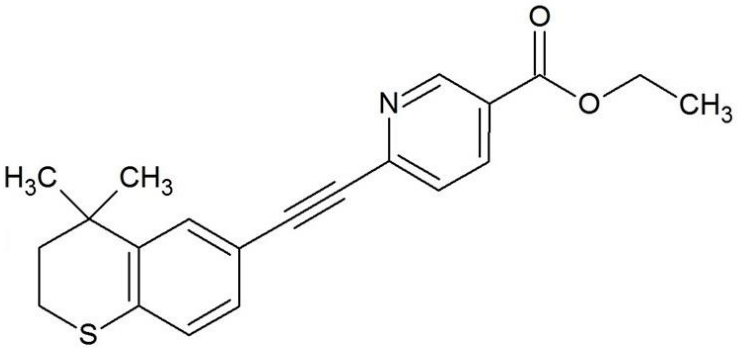
Anatomical Therapeutic Chemical Classification System

D 05 B B 01



D00316

D 05 A X 05



D01132

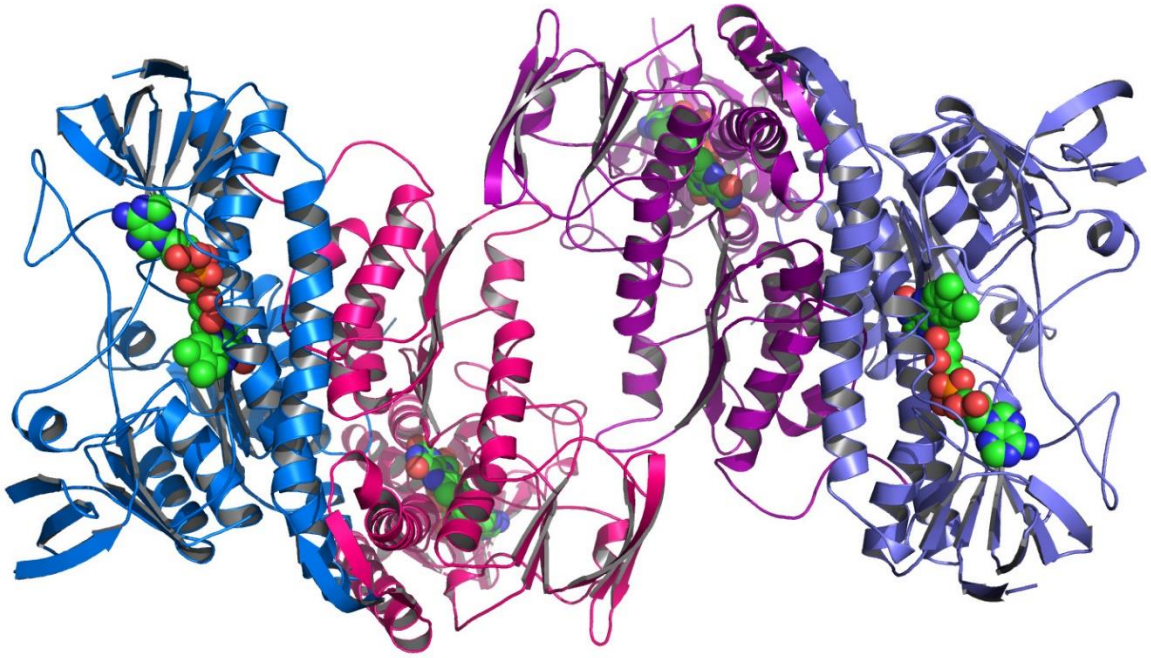
structural similarity

First two levels are the same!

ATC code similarity = 2/5 = 0.4 > 0.275

Functional categories of proteins

- Non-structural
- Describing their **functions**



Our new similarity measure

Drugs

$$\left(\text{2D chemical structure similarity} + \boxed{\text{ATC code similarity}}\right) / 2$$

Targets

$$\left(\text{protein sequence similarity} + \boxed{\text{functional category code similarity}}\right) / 2$$

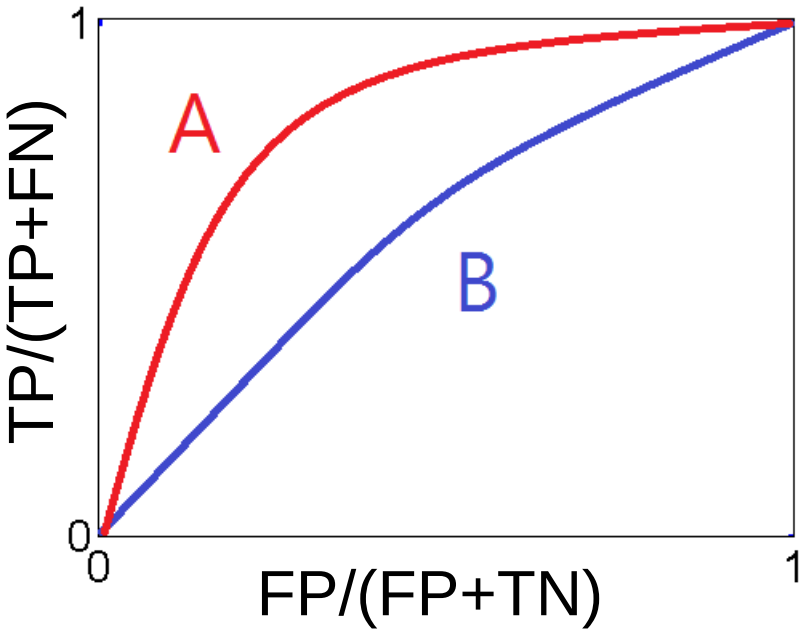
Using new similarity measures and “super-targets”

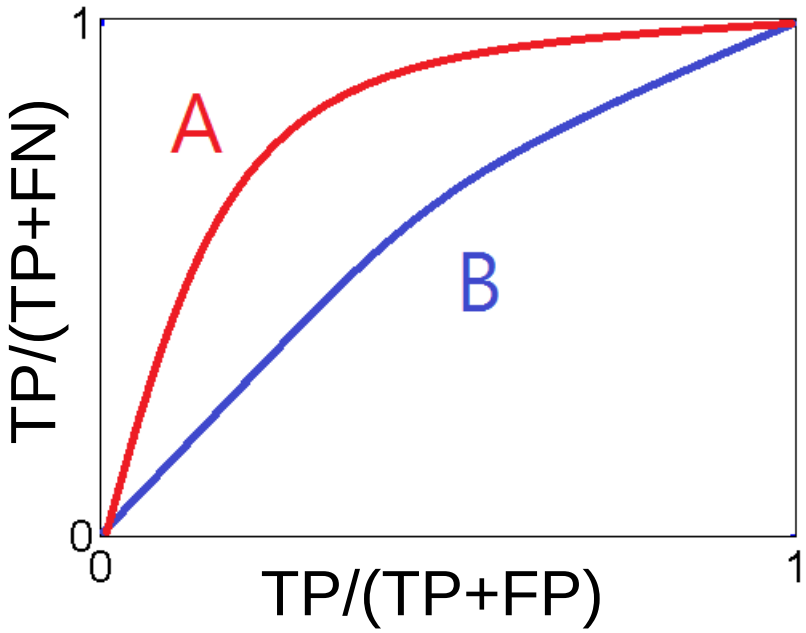
Our performance

AUC

A is better than B

AUPR





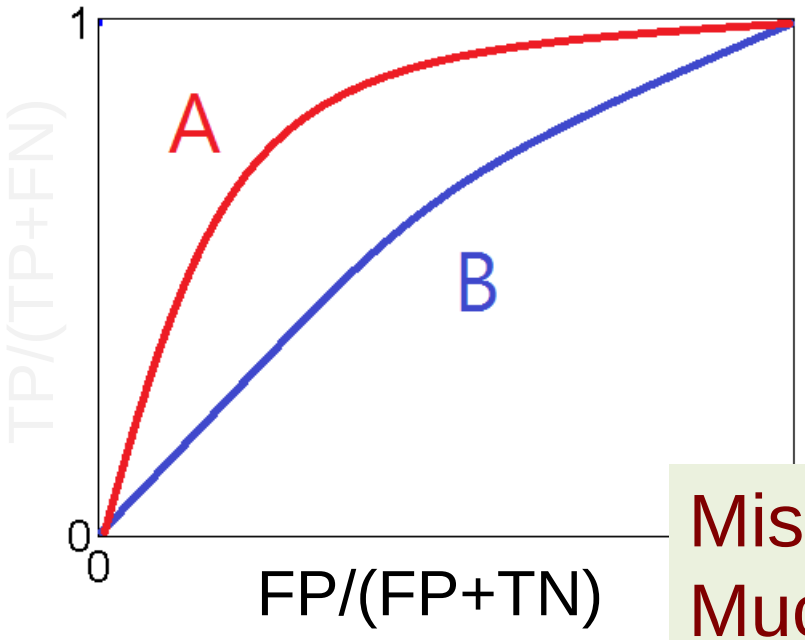
	Actual +ve	Actual -ve
Predicted +ve	TP	FP
Predicted -ve	FP	TN

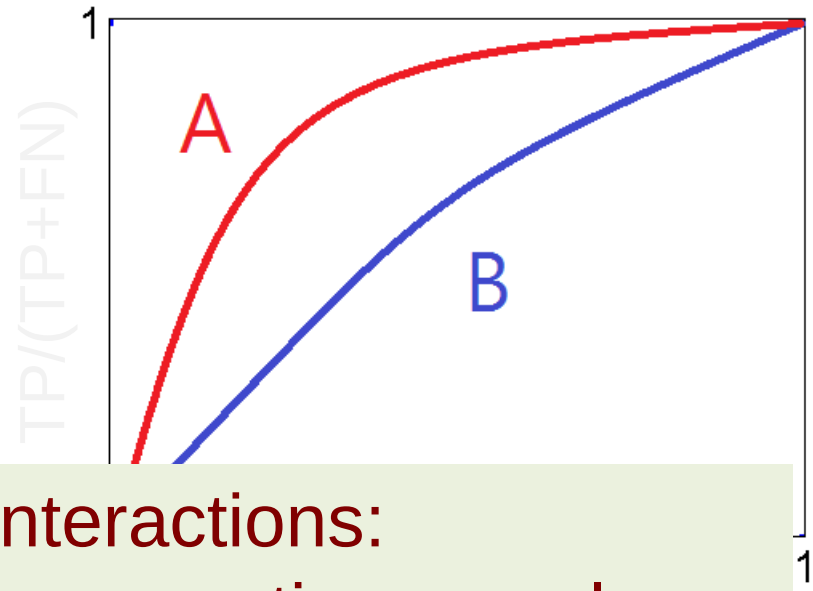
	Actual +ve	Actual -ve
Predicted +ve	TP	FP
Predicted -ve	FN	TN

AUC

A is better than B

AUPR





Missing interactions:
Much more negative samples

	Actual +ve	Actual -ve
Predicted +ve	TP	FP
Predicted -ve	FP	TN

	Actual +ve	Actual -ve
Predicted +ve	TP	FP
Predicted -ve		N

When # of FP's is big:
AUC is overly optimistic

Overall performance

	Enzyme		Ion channel		GPCR		Nuclear receptor		Total
	AUC	AUPR	AUC	AUPR	AUC	AUPR	AUC	AUPR	running time
KBMF2K	0.812	0.287	0.802	0.245	0.840	0.347	0.810	0.354	115.4 min
WNN-GIP	**0.861**	0.280	0.775	0.233	0.872	0.311	0.839	0.456	190.9 min
Ours	0.812	**0.385**	**0.811**	**0.367**	**0.875**	**0.414**	**0.871**	**0.533**	5.5 min

With and without new similarity measures

	Enzyme		Ion channel		GPCR		Nuclear receptor	
	AUC	AUPR	AUC	AUPR	AUC	AUPR	AUC	AUPR
Without new	0.805	0.332	0.776	0.296	0.854	0.304	0.860	0.476
With new	**0.812**	**0.385**	**0.811**	**0.367**	**0.875**	**0.414**	**0.871**	**0.533**

New drug, new target

- Remove known interactions from the data set to create “new” drugs and targets
- Consider if the removed interactions could be predicted
- The **mis-prediction error** measures the fraction of “new” drugs with a wrong prediction

New drug, new target

	Enzyme	Ion Channel	GPCR	Nuclear Receptor
KBMF2K	0.774	0.600	0.654	0.600
WNN-GIP	0.931	0.600	0.692	0.600
Ours	**0.657**	**0.500**	**0.500**	0.600

The numbers are mis-prediction errors.
The smaller the mis-prediction error, the better the performance.

Conclusions

- Non-structural-based similarities
- “Super-targets”

My e-mail: liym1018@hku.hk

Thank you for listening.

References

1. E. E. Bolton, Y. Wang, P. A. Thiessen et al., “PubChem: integrated platform of small molecules and biological activities,” Annual Reports in Computational Chemistry, vol. 4, pp. 217–241, 2008.
2. M. Hurle, L. Yang, Q. Xie et al., “Computational drug repositioning: from data to therapeutics,” Clin. Pharmacol. Ther., vol. 93, no. 4, pp. 335–341, 2013.
3. M. A. Yıldırım, K.-I. Goh, M. E. Cusick et al., “Drug-target network,” Nat. Biotechnol., vol. 25, no. 10, pp. 1119–1126, 2007.
4. F. Cheng, C. Liu, J. Jiang et al., “Prediction of drug-target interactions and drug repositioning via network-based inference,” PLoS Comput. Biol., vol. 8, no. 5, p. e1002503, 2012.
5. D.-S. Cao, S. Liu, Q.-S. Xu et al., “Large-scale prediction of drug–target interactions using protein sequences and drug topological structures,” Anal. Chim. Acta., vol. 752, pp. 1–10, 2012.
6. L. Jacob and J.-P. Vert, “Protein-ligand interaction prediction: an improved chemogenomics approach,” Bioinformatics, vol. 24, no. 19, pp. 2149–2156, 2008.
7. Y. Yamanishi, M. Araki, A. Gutteridge et al., “Prediction of drug–target interaction networks from the integration of chemical and genomic spaces,” Bioinformatics, vol. 24, no. 13, pp. i232–i240, 2008.
8. J.-P. Mei, C.-K. Kwoh, P. Yang et al., “Drug–target interaction prediction by learning from local information and neighbors,” Bioinformatics, vol. 29, no. 2, pp. 238–245, 2013.
9. T. van Laarhoven, S. B. Nabuurs, and E. Marchiori, “Gaussian interaction profile kernels for predicting drug–target interaction,” Bioinformatics, vol. 27, no. 21, pp. 3036–3043, 2011.
10. T. van Laarhoven and E. Marchiori, “Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile,” PloS One, vol. 8, no. 6, p. e66952, 2013.
11. M. Gonen, “Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization,” Bioinformatics, vol. 28, no. 18, pp. 2304–2310, 2012.
12. Y. Koren, R. Bell, and C. Volinsky, “Matrix factorization techniques for recommender systems,” Computer, vol. 42, no. 8, pp. 30–37, 2009.
13. M. Kanehisa and S. Goto, “KEGG: Kyoto encyclopedia of genes and genomes,” Nucleic Acids Res., vol. 28, no. 1, pp. 27–30, 2000.
14. “ATC structure and principles,” http://www.whocc.no/atc/structure_and_principles/, accessed: 2014-06-30.
15. M.-L. Zhang and Z.-H. Zhou, “ML-KNN: A lazy learning approach to multi-label learning,” Pattern Recogn., vol. 40, no. 7, pp. 2038–2048, 2007.
16. K. Bleakley and Y. Yamanishi, “Supervised prediction of drug–target interactions using bipartite local models,” Bioinformatics, vol. 25, no. 18, pp. 2397–2403, 2009.
17. M. Hattori, Y. Okuno, S. Goto et al., “Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways,” J. Am. Chem. Soc., vol. 125, no. 39, pp. 11 853–11 865, 2003.
18. T. F. Smith and M. S. Waterman, “Identification of common molecular subsequences,” J. Mol. Biol., vol. 147, no. 1, pp. 195–197, 1981.
19. M. Tsuji, “Local motifs involved in the canonical structure of the ligand-binding domain in the nuclear receptor superfamily,” J. Struct. Biol., vol. 185, no. 3, pp. 355–365, 2014.
20. I. Letunic, T. Doerks, and P. Bork, “SMART 7: recent updates to the protein domain annotation resource,” Nucleic Acids Res., vol. 40, no. D1, pp. D302–D305, 2012.
21. J. Davis and M. Goadrich, “The relationship between Precision-Recall and ROC curves,” in Proceedings of the 23rd international conference on Machine learning. ACM, 2006, pp. 233–240.
22. F. J. Azuaje, L. Zhang, Y. Devaux et al., “Drug-target network in myocardial infarction reveals multiple side effects of unrelated drugs,” Scientific Reports, vol. 1, 2011.
23. G. Tsoumakas, I. Katakis, and I. Vlahavas, “Mining multi-label data,” in Data Mining and Knowledge Discovery Handbook. Springer, 2010, pp. 667–685.

Image sources

- Shen Nong. http://upload.wikimedia.org/wikipedia/commons/c/c5/Shennong3.jpg
- Herb. https://c2.staticflickr.com/6/5046/5366857952_6c04ec9b35_b.jpg
- Zyprexa. http://da.wikipedia.org/wiki/Olanzapin#mediaviewer/File:Zyprexa.PNG
- Schizophrenia patient artwork. http://upload.wikimedia.org/wikipedia/commons/b/b2/Cloth_embroidered_by_a_schizophrenia_sufferer.jpg
- D2R. http://en.wikipedia.org/wiki/Dopamine_receptor_D2#mediaviewer/File:Protein_DRD2_PDB_1I15.png
- Yellow and blue pills. http://commons.wikimedia.org/wiki/File:Pharmaceuticals.jpg
- Aspirin 2D structure. http://commons.wikimedia.org/wiki/File:Aspirin-skeletal.png
- Prozac 2D structure. http://commons.wikimedia.org/wiki/File:Fluoxetine_structure.svg
- FOXP2 structure. http://en.wikipedia.org/wiki/FOXP2
- Protein structure #1. http://upload.wikimedia.org/wikipedia/commons/8/86/Argonne's_Midwest_Center_for_Structural_Genomics_deposits_1,000th_protein_structure.jpg
- Protein structure #2. http://upload.wikimedia.org/wikipedia/commons/8/86/Argonne's_Midwest_Center_for_Structural_Genomics_deposits_1,000th_protein_structure.jpg
- Tablets. http://pixabay.com/en/medications-cure-tablets-pharmacy-342462/
- Doxycycline. http://en.wikipedia.org/wiki/Doxycycline#mediaviewer/File:Doxycycline_100mg_capsules.jpg
- Orange tablets in yellow cup. http://pixabay.com/en/tablets-pills-medicine-disease-193666/
- Furosemide. http://upload.wikimedia.org/wikipedia/commons/thumb/5/51/Furosemide_(1).JPG/768px-Furosemide_(1).JPG
- My neighbor Totoro. http://helixaspersa.deviantart.com/art/My-Neighbor-Totoro-in-Autumm-137190257

** All the images used are labeled **for non-commercial reuse** by Google image search.*

Supplementary #1

Estimating Pr(A) and Pr(A^C)

$$\Pr\left[a(x,j)=1\right] \approx \left[1+\sum_{i=1}^{m} A(i,j)\right] /(m+2);$$
$$\Pr\left[a(x,j)=0\right] = 1-\Pr\left[a(x,j)=1\right]$$

- Event A: (New) drug ***d*** interacts with target ***t***
- Event B: ***c*** drugs in the set of ***d***'s ***K*** nearest neighbors interacts with target ***t***

Supplementary #1

Estimating Pr(B|A) and Pr(B|A^C)

$$\frac{1 + \sum_i Ind[A(i,j) = b \,\&\, n(i,j,K) = c]}{(K+1) + \sum_{c'=0}^{K} \sum_i Ind[A(i,j) = b \,\&\, n(i,j,K) = c']}$$

- Event A: (New) drug ***d*** interacts with target ***t***
- Event B: ***c*** drugs in the set of ***d***'s ***K*** nearest neighbors interacts with target ***t***

Supplementary #2

All the methods with new similarity measures

	Enzyme		Ion Channel		GPCR		Nuclear Receptor	
	AUC	AUPR	AUC	AUPR	AUC	AUPR	AUC	AUPR
KBMF2K	0.870	0.391	0.833	0.330	0878	0.414	0.860	0.403
WNN-GIP	0.846	0.323	0.813	0.263	**0.888**	0.403	0.864	0.497
Ours	**0.849**	**0.432**	**0.817**	**0.370**	**0.888**	**0.422**	**0.882**	**0.521**