医薬品の標的分子や副作用の予測における機械学習

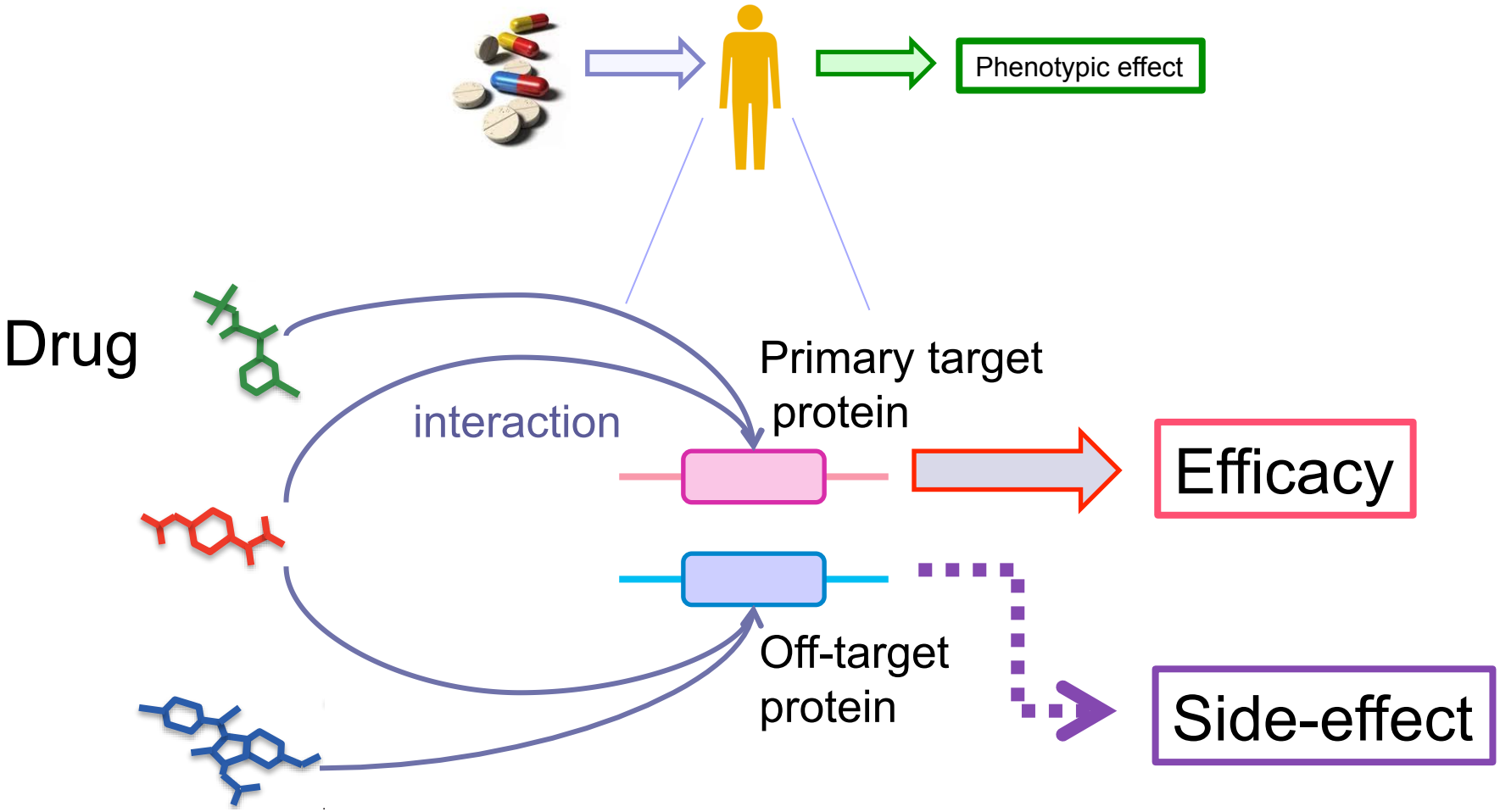
山西芳裕

九州大学生体防御医学研究所

Outline

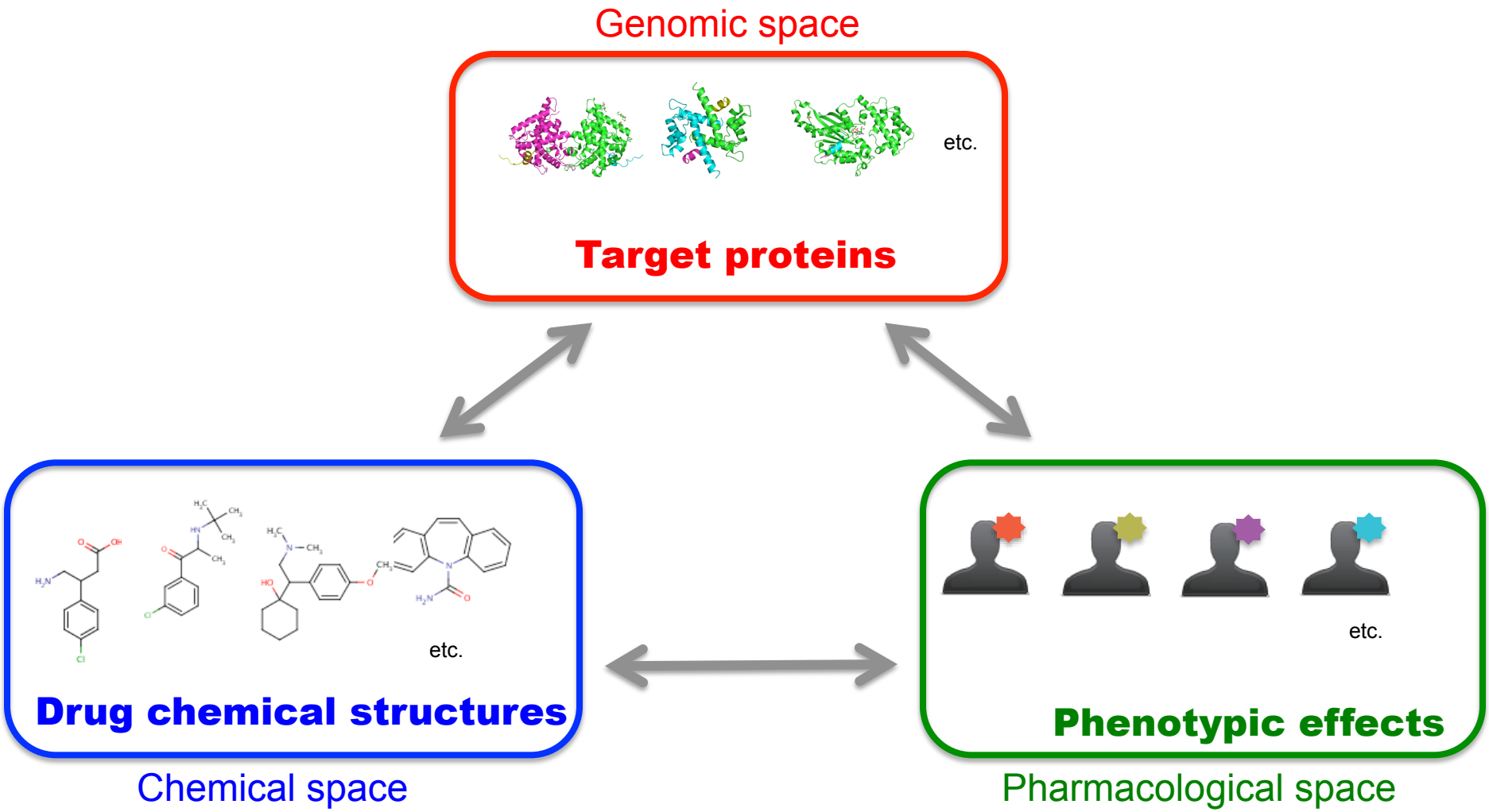
- Background
  - Chemical, genomic and pharmacological spaces
- Methods for pharmaceutical applications
  - Drug target prediction from chemical and genomic data
  - Side-effect prediction from biological data
- Results
- Concluding remarks

Drug-target interaction



Identification of interactions between drugs and target proteins is crucial in the drug development

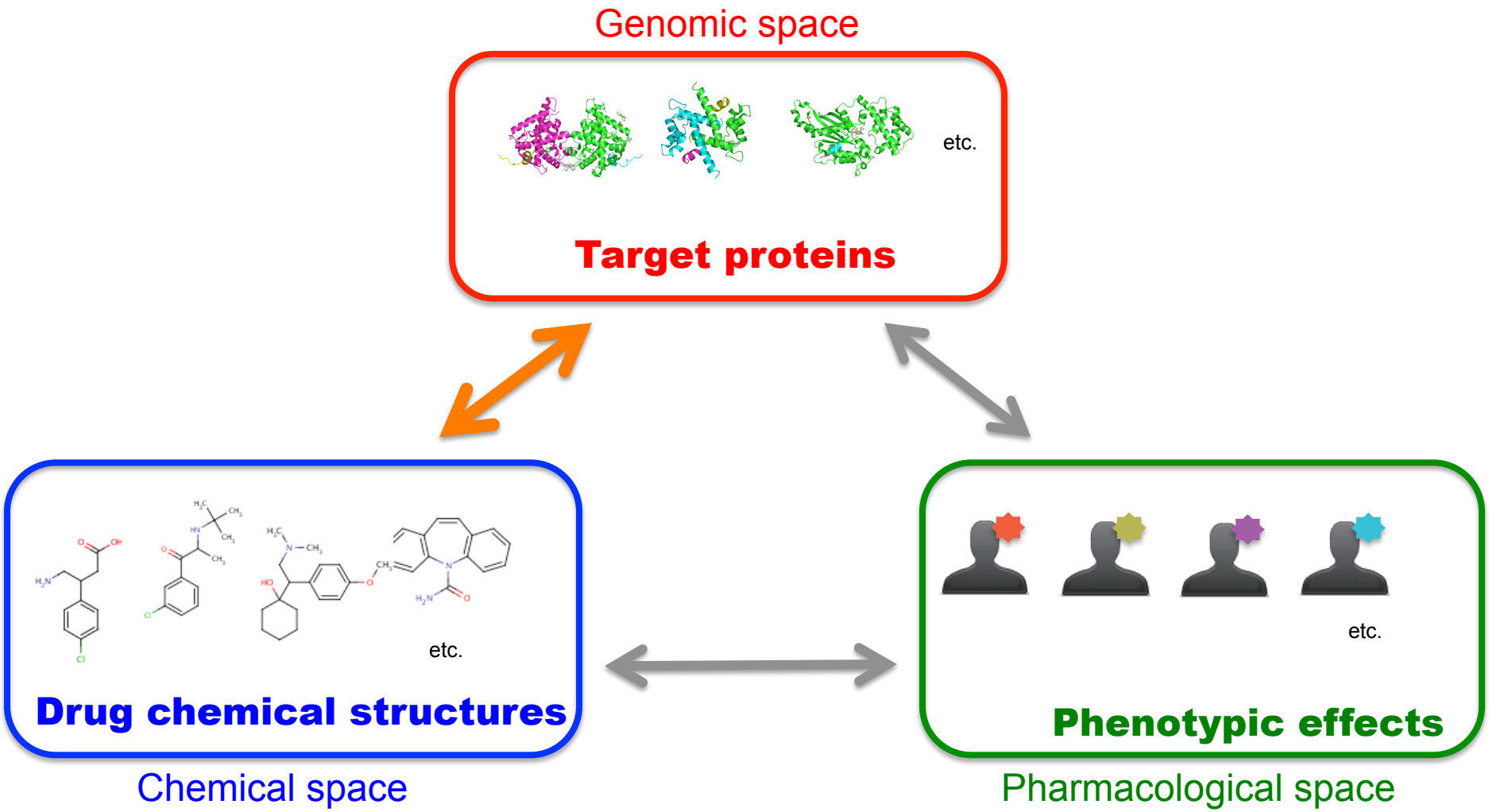
Heterogeneous omics data



Outline

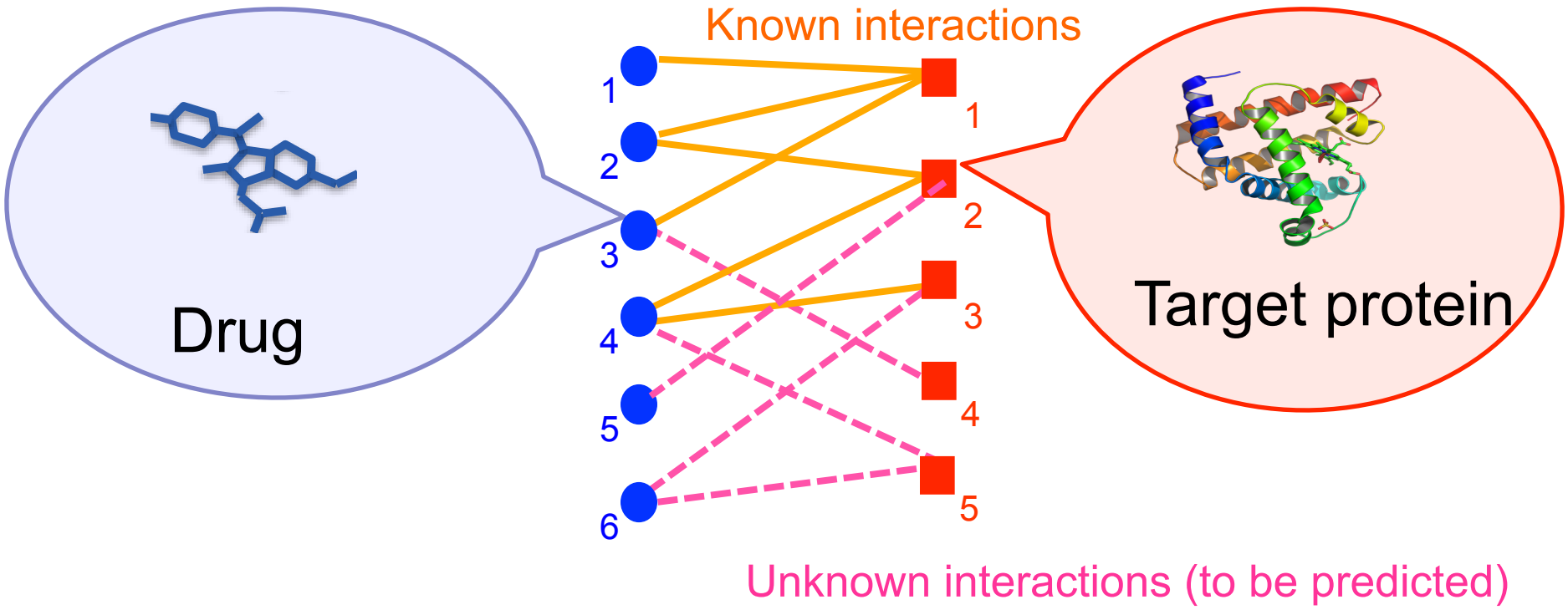
- Background
  - Chemical, genomic and pharmacological spaces
- Methods for pharmaceutical applications
  - Drug target prediction from chemical and genomic data
  - Side-effect prediction from biological data
- Results
- Concluding remarks

Heterogeneous omics data



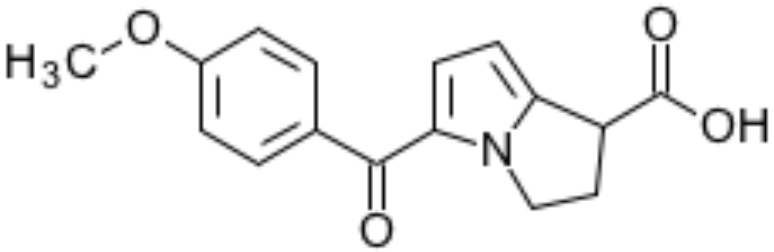
Objective

- Prediction of unknown drug-target interactions on a large scale from chemical and genomic data



Examples of the data structures

Drug



D02944

Chemical graph structure

Target protein

MAHAAQVGLQDATSPIMEELITFHDHALMIIFLICFLVLYA
LFLTLTTKLTNTNISDAQEMETVWTILPAIILVLIALPSLRIL
YMTDEVNDPSLTIKSIGHQWYWTYEYTDYGGLIFNSYML
PPLFLEPGDLRLLDVDNRVVLPIEAPIRMMITSQDVLHSW
AVPTLGLKTDAIPGRLNQTTFTATRPGVYYGQCSEICGAN
HSFMPIVLELIPLKIFEMGPVFTL

Amino acid sequence

Chemogenomic approach

Strategy: Chemically similar drugs are predicted to interact with similar target proteins

(Yamanishi et al, *Bioinformatics*, 2008; Faulon et al., *Bioinformatics*, 2008; Jacob et al, *Bioinformatics*, 2008, Yabuuchi et al, Mol Sys Bio, 2011)

Chemical structure similarity for drugs is evaluated by a graph kernel: (Mahe et al, *J Chem Inf Model,* 2005)

$$(K_x)_{ij} = k_x(\mathbf{x}_i, \mathbf{x}_j) \text{ for } i, j = 1, 2, ..., n_x$$

Genomic sequence similarity for target proteins is evaluated by a string kernel: (Saigo et al, *Bioinformatics*, 2004)

$$(K_z)_{ij} = k_z(\mathbf{z}_i, \mathbf{z}_j) \text{ for } i, j = 1, 2, ..., n_z$$

Binary classification approach

- Classification of drug-target pairs into the interaction class or non-interaction class

Support Vector Machine (SVM) with pairwise kernels

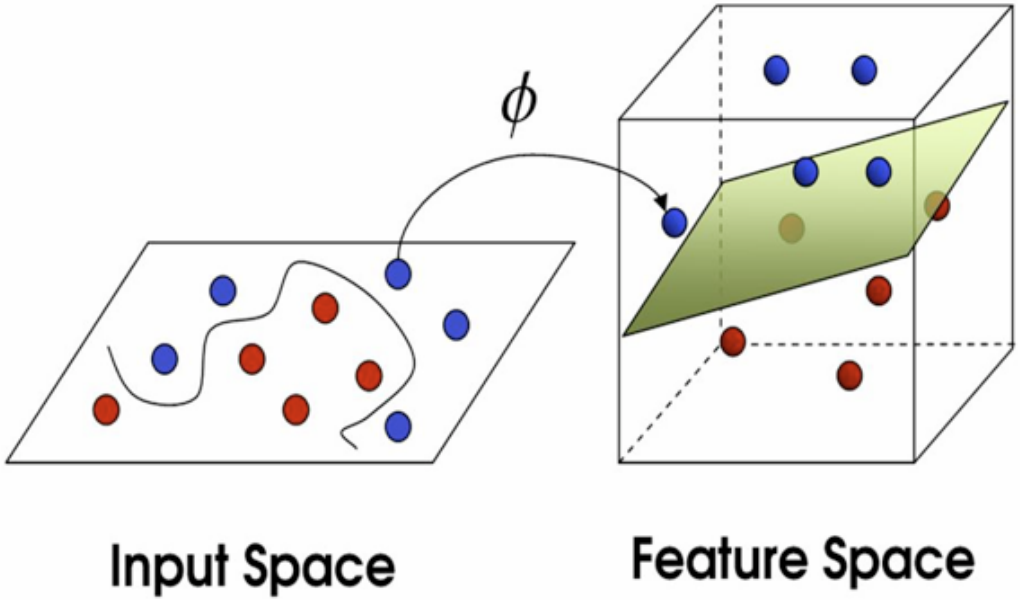
- Faulon et al., *Bioinformatics*, 24:225-233, 2008
- Jacob and Vert, *Bioinformatics*, 24:2149-2156, 2008
- Yabuuchi et al, Mol Sys Bio, 2011

SVM with pairwise kernels (pairwise SVM)

ordinary SVM :

$$f(x) = \sum_{i=1}^{n} a_i k(x_i, x') + b$$

where x : an object



pairwise SVM :

$$f(x,z) = \sum_{i=1}^{n_x \times n_z} a_i k((x,z)_i, (x',z')) + b$$

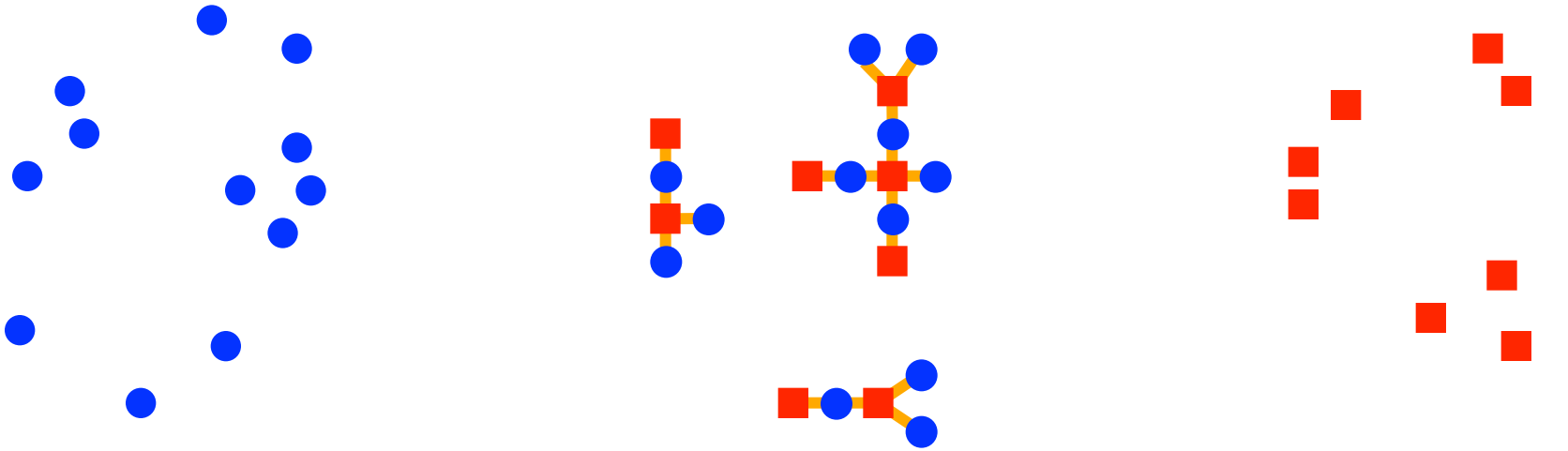
where (x,z) : a compound – protein pair

Supervised bipartie graph inference

Chemical space

Interaction space

Genomic space



known drug

known interaction

known target

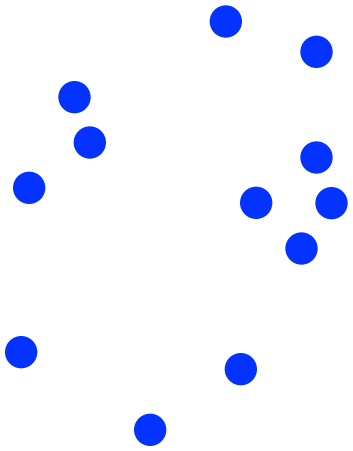
Compounds with similar structures are close to each other

Interacting drugs and targets are connected on the graph

Proteins with similar sequences are close to each other

Step 1: embedding drugs and targets on the known graph into a unified feature space $\mathbf{R}^d$

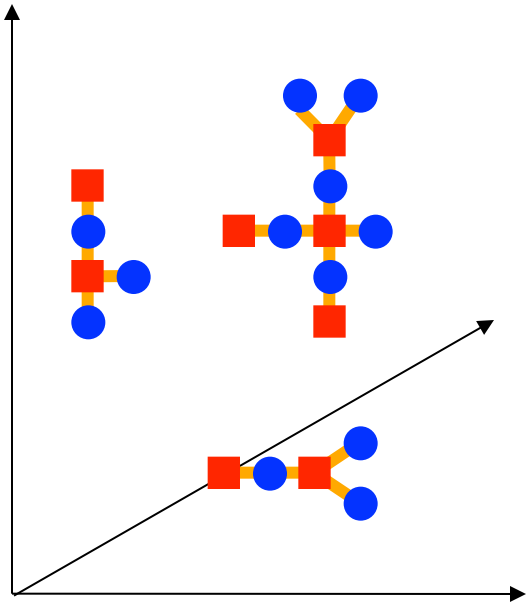
Chemical space



known drug

Compounds with similar structures are close to each other

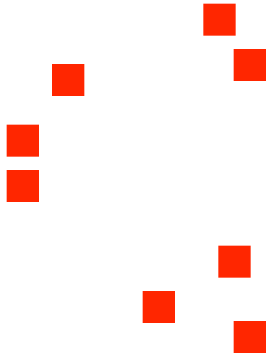
Interaction space



known interaction

Interacting drugs and targets are close to each other

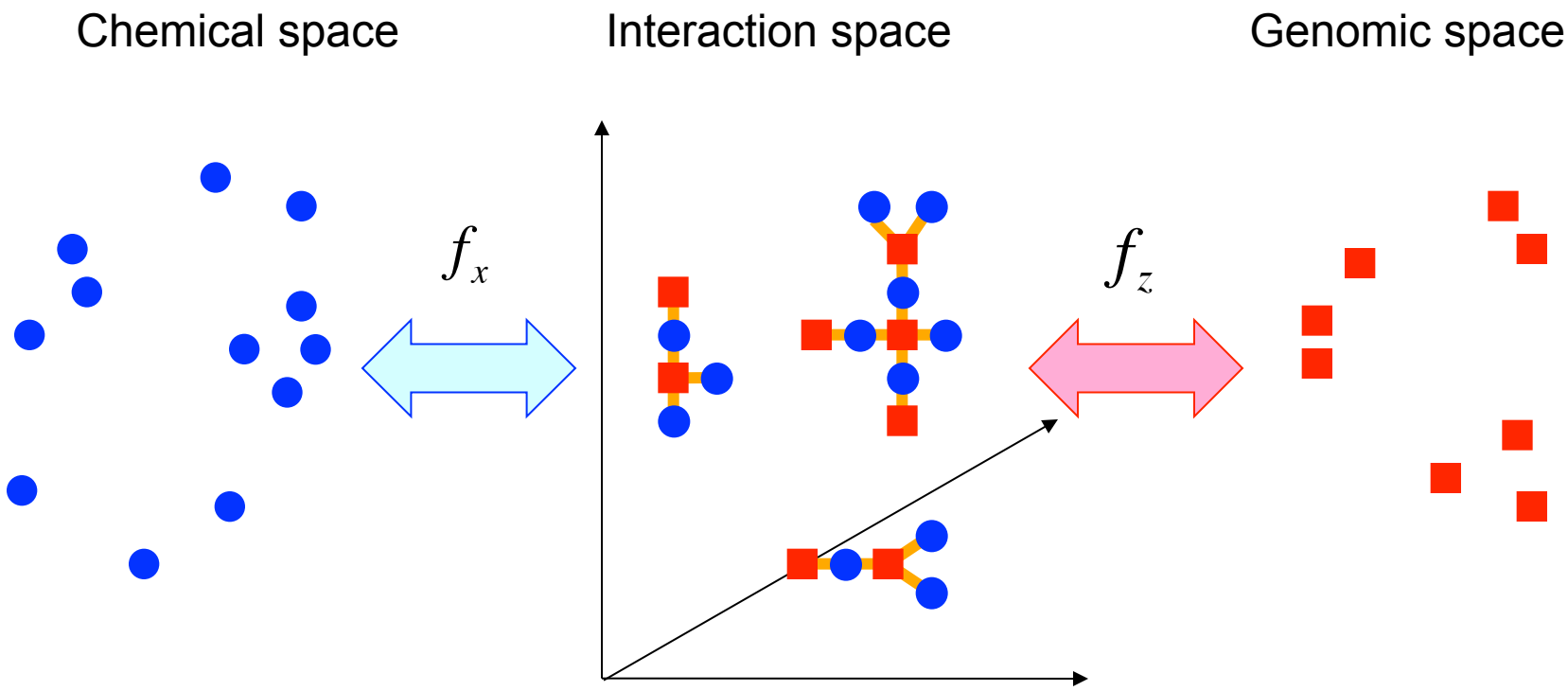
Genomic space



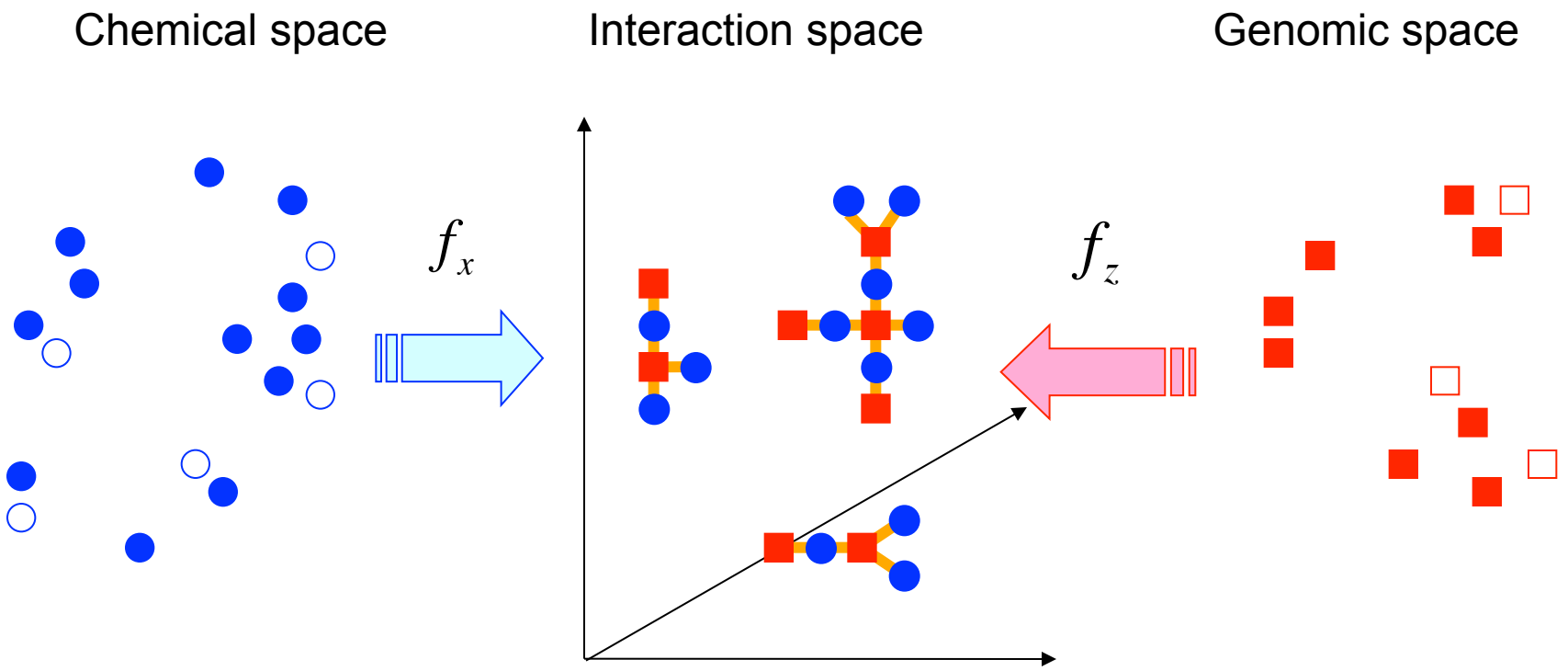
known target

Proteins with similar sequences are close to each other

Step 2. Learning a model between the chemical/ genomic space and the interaction space



Step 3. Predicting unknown interactions involving new compounds/proteins after the projection



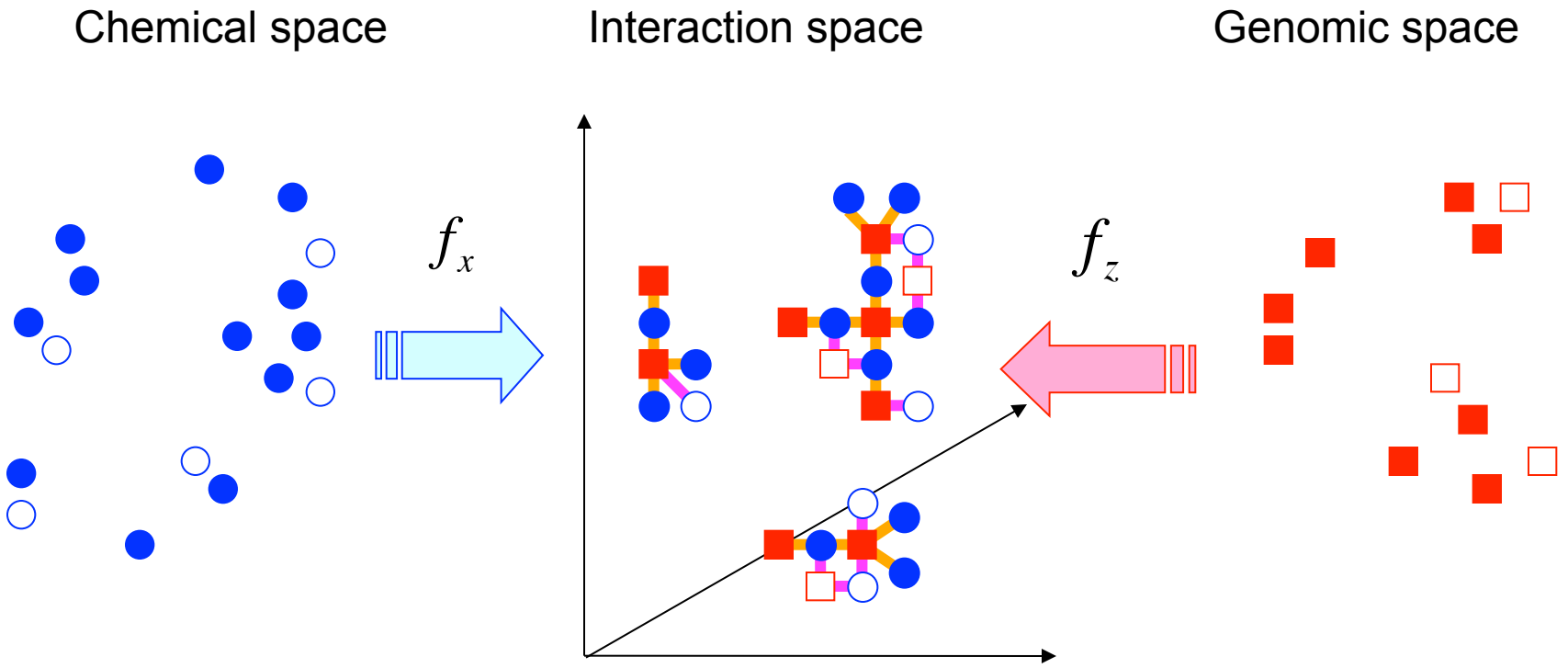
○ new compound

New compounds are mapped with f_x

□ new protein

New proteins are mapped with f_z

Step 3. Predicting unknown interactions involving new compounds/proteins after the projection



Mappting to a unified space

Let us consider two functions to map each compound **x** and each protein **z** onto a unified Euclidian space

$$\begin{cases} f_x(\mathbf{x}) = (f_x^{(1)}(\mathbf{x}), \cdots, f_x^{(d)}(\mathbf{x}))^T \in \mathbf{R}^d \\ f_z(\mathbf{z}) = (f_z^{(1)}(\mathbf{z}), \cdots, f_z^{(d)}(\mathbf{z}))^T \in \mathbf{R}^d \end{cases}$$

We find f_x and f_z which minimize

$$R(f_x, f_z) = \frac{\sum_{(\mathbf{x}_i, \mathbf{z}_j) \in E} (f_x(\mathbf{x}_i) - f_z(\mathbf{z}_j))^2 + \lambda_1 \| f_x \|^2 + \lambda_2 \| f_z \|^2}{\sum_{(\mathbf{x}_i, \mathbf{z}_j) \in V_x \times V_z} (f_x(\mathbf{x}_i) - f_z(\mathbf{z}_j))^2}$$

where V_x (resp. V_z) is a set of drugs (resp. target proteins), E is a set of interactions, and λ_1 and λ_2 are regularization parameters

(Yamanishi, Adv Neural Inf Process Syst *21,* 2009)

Extraction of multiple features

Succesive features $f_x^{(q)}$ and $f_z^{(q)}$ $(q = 1,2,\cdots,d)$ are obtained by

$$\begin{pmatrix} f_x^{(q)} \\ f_z^{(q)} \end{pmatrix} = \arg\min \frac{\sum_{(\mathbf{x}_i,\mathbf{z}_j)\in E} (f_x(\mathbf{x}_i) - f_z(\mathbf{z}_j))^2 + \lambda_1 \| f_x \|^2 + \lambda_2 \| f_z \|^2}{\sum_{(\mathbf{x}_i,\mathbf{z}_j)\in V_x \times V_z} (f_x(\mathbf{x}_i) - f_z(\mathbf{z}_j))^2}$$

under the following orthogonality constraints:

$$f_x \perp f_x^{(1)},\cdots,f_x^{(q-1)}, \quad f_z \perp f_z^{(1)},\cdots,f_z^{(q-1)}$$

Prediction score for a given pair of compound $\mathbf{x}'$ and protein $\mathbf{z}'$:

$$g(\mathbf{x}',\mathbf{z}') = \sum_{q=1}^{d} f_x^{(q)}(\mathbf{x}') \cdot f_z^{(q)}(\mathbf{z}')$$

Algorithm

By the representer theorem, features can be expanded as

$$f_x(\mathbf{x}) = \sum_{j=1}^{n_x} \alpha_j k_x(\mathbf{x}_j, \mathbf{x}), \quad f_z(\mathbf{z}) = \sum_{j=1}^{n_z} \beta_j k_z(\mathbf{z}_j, \mathbf{z})$$

Kernel Gram matrices:

$$(K_x)_{ij} = k_x(\mathbf{x}_i, \mathbf{x}_j), \quad i, j = 1, 2, \cdots, n_x$$

$$(K_z)_{ij} = k_z(\mathbf{z}_i, \mathbf{z}_j), \quad i, j = 1, 2, \cdots, n_z$$

The solution can be obtained by finding α_q and β_q which minimizes

$$R(\alpha,\beta)=\frac{\begin{pmatrix}\alpha\\ \beta\end{pmatrix}^T\begin{pmatrix}K_xD_xK_x & -K_xAK_z\\ -K_zA^TK_x & K_zD_zK_z\end{pmatrix}\begin{pmatrix}\alpha\\ \beta\end{pmatrix}+\lambda_1\alpha^TK_x\alpha+\lambda_2\beta^TK_z\beta}{\begin{pmatrix}\alpha\\ \beta\end{pmatrix}^T\begin{pmatrix}K_x^2 & 0\\ 0 & K_z^2\end{pmatrix}\begin{pmatrix}\alpha\\ \beta\end{pmatrix}}$$

under the following constraints:

$$\alpha^TK_x\alpha_1=\cdots=\alpha^TK_x\alpha_{q-1}=0,\quad \beta^TK_z\beta_1=\cdots=\beta^TK_z\beta_{q-1}=0$$

where

D_x (resp. D_z) : degree matrix of drugs (resp. target proteins),

A : adjacency matrix of drug-target interactions

It is reduced to the generalized eigenvalue problem:

$$\begin{pmatrix}K_xD_xK_x+\lambda_1K_x & -K_xAK_z\\ -K_xA^TK_x & K_zD_zK_z+\lambda_2K_z\end{pmatrix}\begin{pmatrix}\alpha\\ \beta\end{pmatrix}=\rho\begin{pmatrix}K_x^2 & 0\\ 0 & K_z^2\end{pmatrix}\begin{pmatrix}\alpha\\ \beta\end{pmatrix}$$

Drug-target interaction data for human: Gold standard data

	Statistics
Number of drugs	1874
Number of target proteins (Total in human genome)	436 (23196)
Number of drug-target interactions	6769

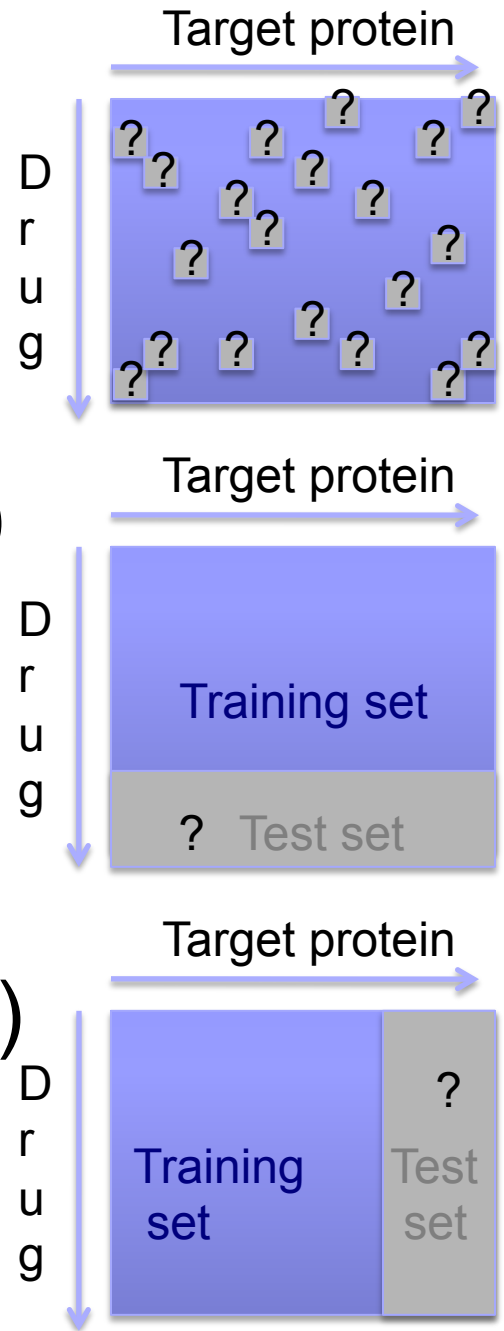
KEGG DRUG (December, 2011)

Cross-validation (CV)

i) Pairwise CV
(Missing interaction detection)

ii) Blockwise CV I
(new drug identification)

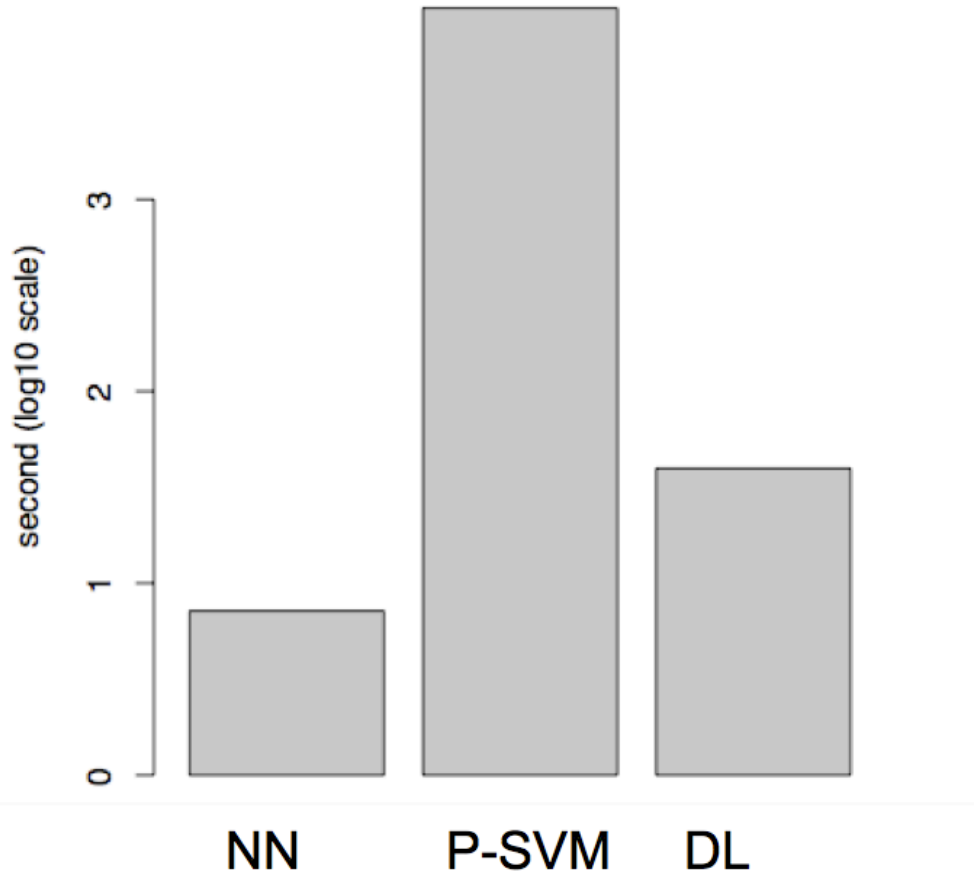
iii) Blockwise CV II
(new target identification)



Performance evaluation by 3-fold cross-validation: AUC (Area under the ROC curve)

CV	Method	AUC (S.D)
Pairwise CV (Missing interaction detection)	Nearest neighbor Pairwise SVM Proposed method	0.894 (0.009) 0.968 (0.006) 0.972 (0.007)
Blockwise CV I (New drug identification)	Nearest neighbor Pairwise SVM Proposed method	0.808 (0.010) 0.855 (0.013) 0.869 (0.006)
Blockwise CV II (New target identification)	Nearest neighbor Pairwise SVM Proposed method	0.711 (0.009) 0.805 (0.003) 0.811 (0.009)

Computational cost



NN: nearest neighbor

P-SVM: pairwise SVM

DL: distance learning
(proposed method)

Comprehensive prediction of unknown drug-target interactions

- Test drugs: all compounds in KEGG LIGAND and all drugs in KEGG DRUG
- Test target proteins: all human proteins in KEGG GENES
- All gold standard interaction data are used in the training

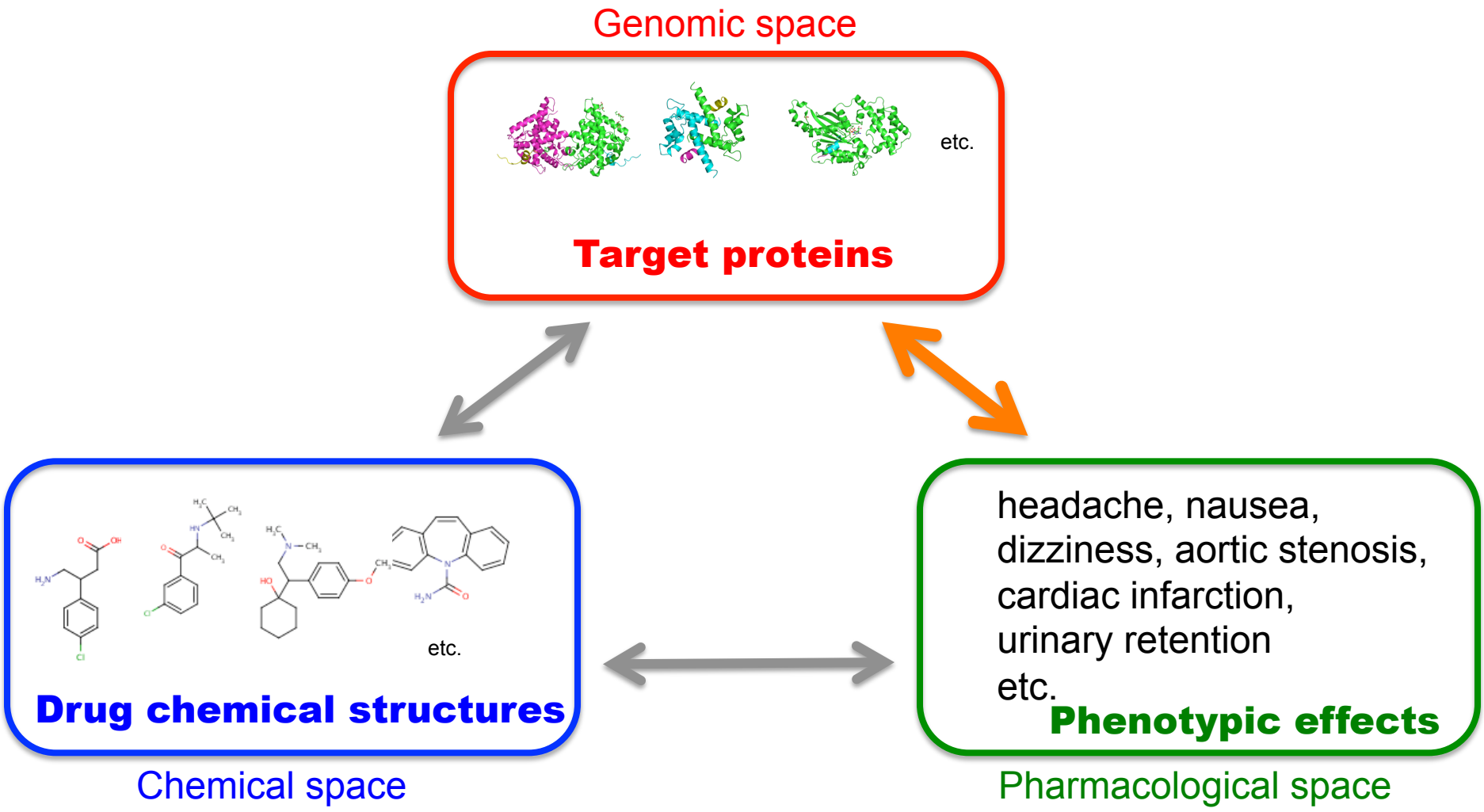
Summary

- The proposed method can predict unknown drug-target interactions on a large scale.
- The prediction is performed based on the integration of chemical and genomic data.
- It does not need 3D structure information of the target proteins.

Outline

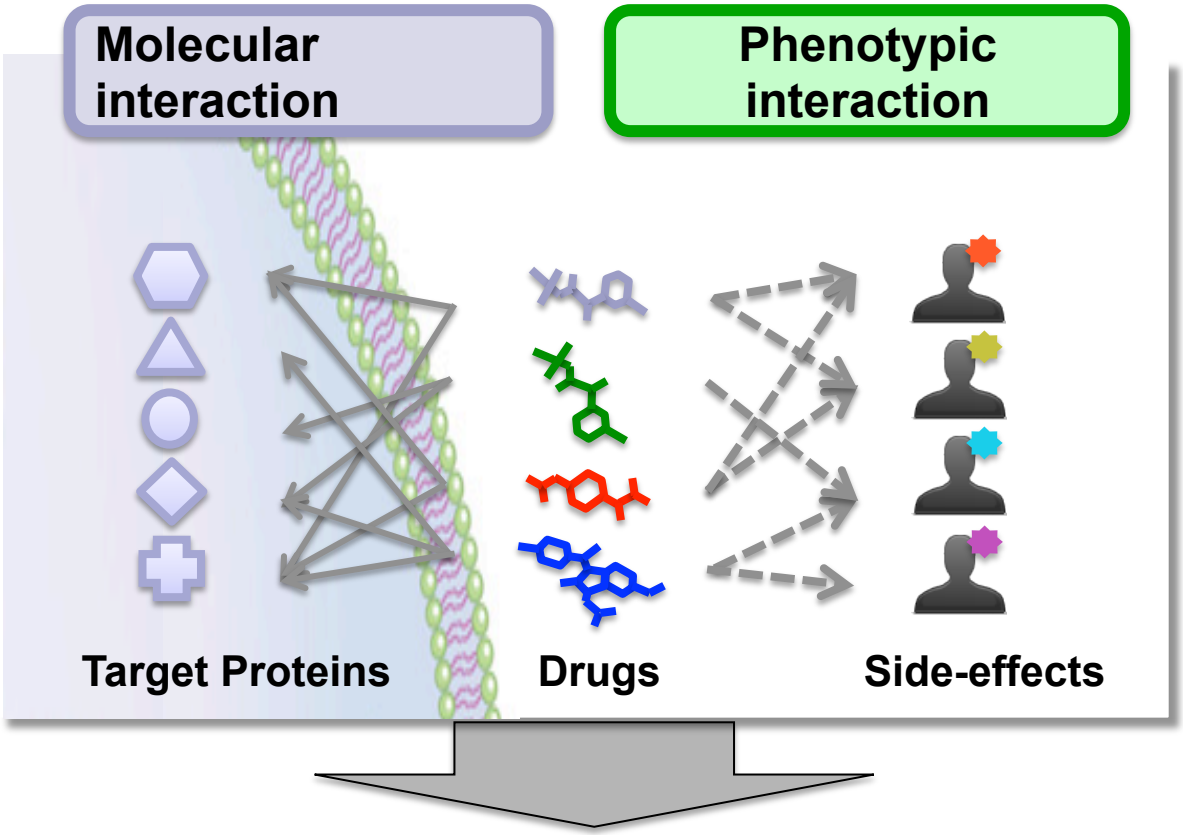
- Background
  - Chemical, genomic and pharmacological spaces
- Methods for pharmaceutical applications
  - Drug target prediction from chemical and genomic data
  - Side-effect prediction from biological data
- Results
- Concluding remarks

Heterogeneous omics data

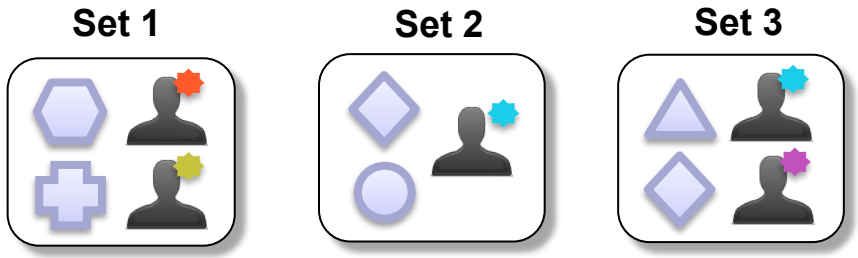


Objective

Two interaction data



Extraction of correlated sets of target proteins and side-effects



Mizutani et al, Bioinformatics, 28, i522-i528, 2012. (ECCB2012)

Correlation analysis of target proteins and side-effects

Target protein profile for each drug: $\mathbf{x}=(x_1,x_2,...,x_p)^T$

indicating presence/absence of 1368 target proteins (DrugBank, Matador)

Side-effect profile for each drug: $\mathbf{y}=(y_1,y_2,...,y_q)^T$

indicating presence/absence of 1339 side-effects (SIDER)

Goal: extraction of target proteins and side-effects which share similar sets of drugs in terms of occurence.

Ordinary canonical correlation analysis (OCCA)

1. For each drug, set two binary vectors indicating presence/absence of interactions;

$$\text{Target protein profile: } \mathbf{x} = [x_1, \ldots, x_p]^T, \quad \text{Side-effect profile: } \mathbf{y} = [y_1, \ldots, y_q]^T$$

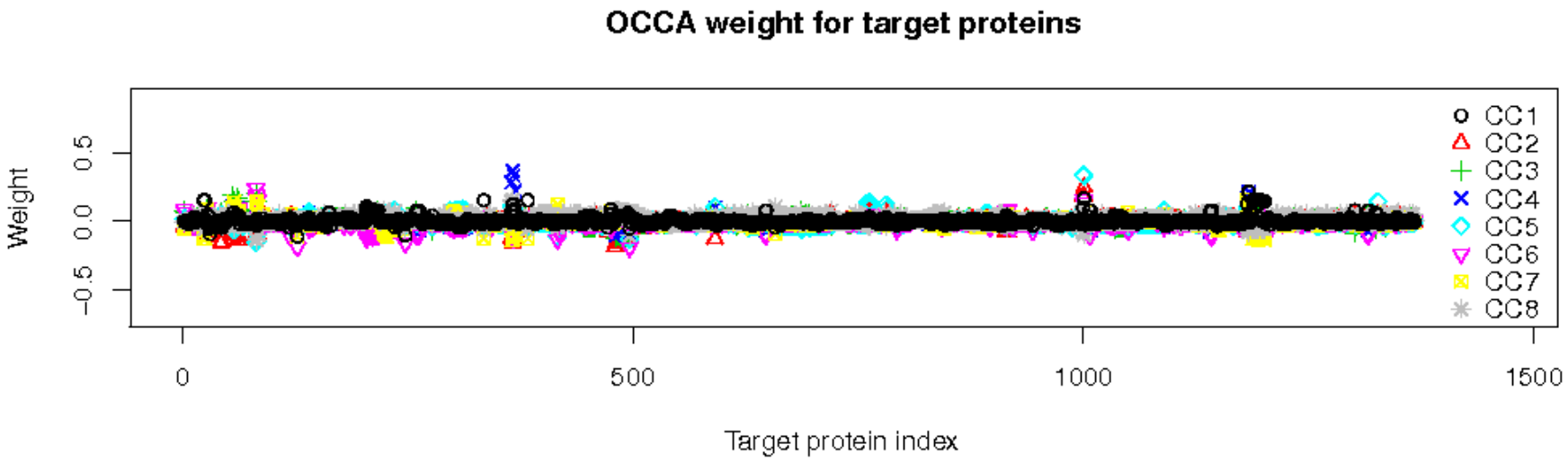
2. Consider two linear combinations for drug i with weight vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$;

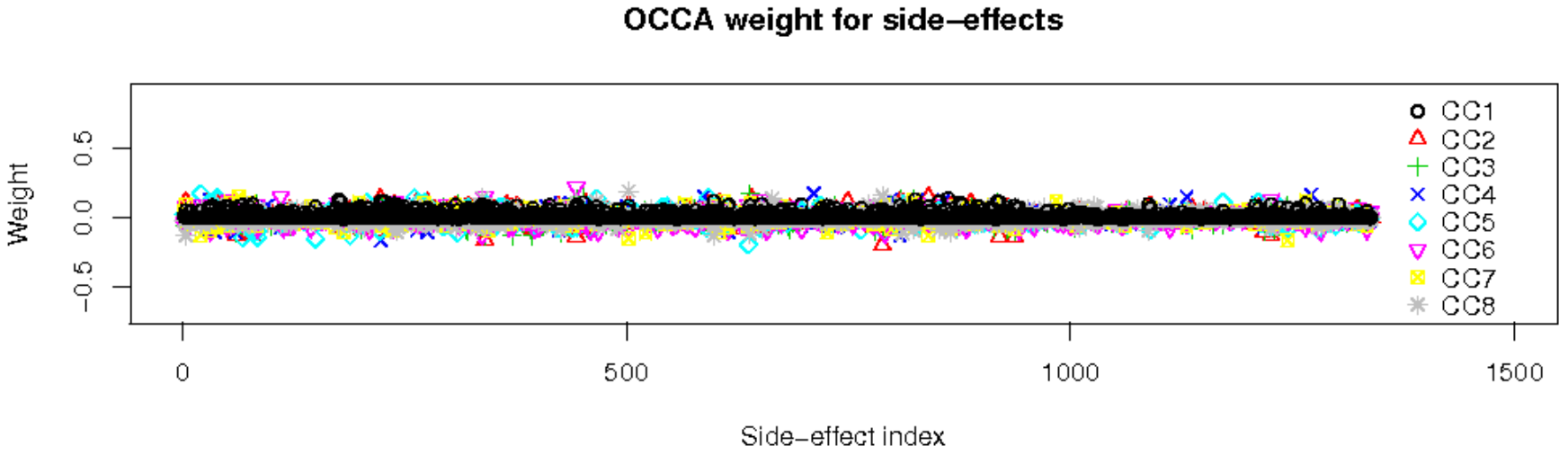
$$u_i = \boldsymbol{\alpha}^T \mathbf{x}_i \quad \text{and} \quad v_i = \boldsymbol{\beta}^T \mathbf{y}_i \quad (\ u \text{ and } v \text{ are centered })$$

Find $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that maximize the canonical correlation coefficient;

$$\rho = corr(u, v) = \frac{\sum_{i=1}^n \boldsymbol{\alpha}^T \mathbf{x}_i \cdot \boldsymbol{\beta}^T \mathbf{y}_i}{\sqrt{\sum_{i=1}^n (\boldsymbol{\alpha}^T \mathbf{x}_i)^2} \quad \sqrt{\sum_{i=1}^n (\boldsymbol{\beta}^T \mathbf{y}_i)^2}}$$

Weight vectors in OCCA





Sparse canonical correlation analysis (SCCA)

2 Consider two linear combinations for drug i with weight vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$;

$$u_i = \boldsymbol{\alpha}^T \mathbf{x}_i \quad \text{and} \quad v_i = \boldsymbol{\beta}^T \mathbf{y}_i \quad (\ u \text{ and } v \text{ are centered })$$

Find $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that maximize the canonical correlation coefficient;

$$\rho = corr(u, v) = \frac{\sum_{i=1}^{n} \boldsymbol{\alpha}^T \mathbf{x}_i \cdot \boldsymbol{\beta}^T \mathbf{y}_i}{\sqrt{\sum_{i=1}^{n} (\boldsymbol{\alpha}^T \mathbf{x}_i)^2} \quad \sqrt{\sum_{i=1}^{n} (\boldsymbol{\beta}^T \mathbf{y}_i)^2}}$$

SCCA

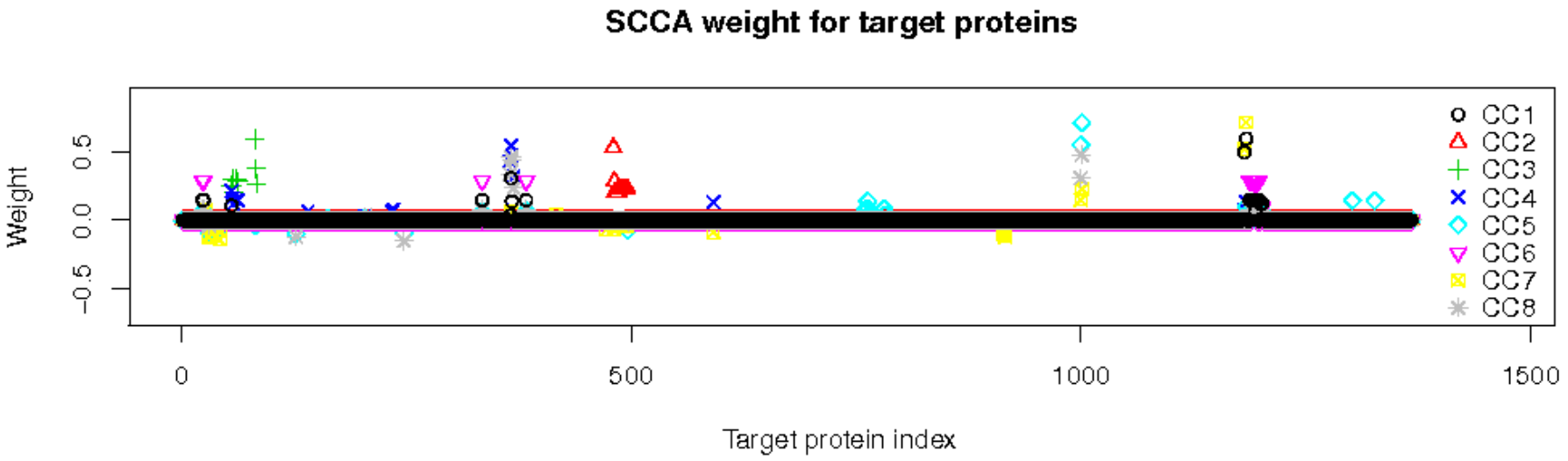
3 In order to impose sparsity on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ we give an L_1 penalty;

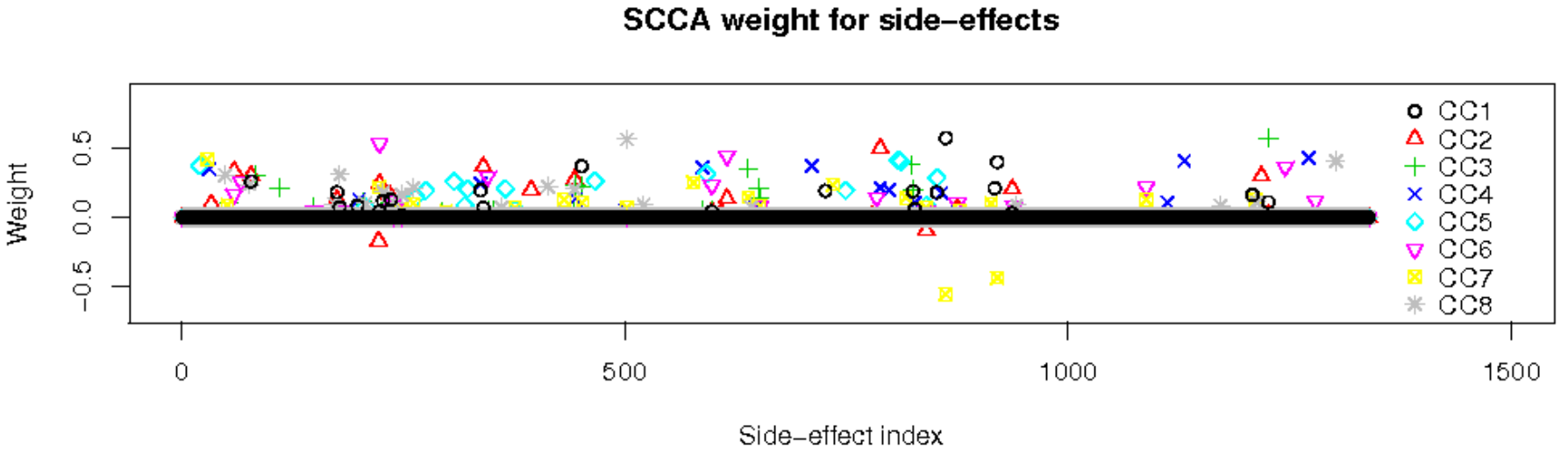
$$||\boldsymbol{\alpha}||_1 \leq c_1\sqrt{p}, \quad ||\boldsymbol{\beta}||_1 \leq c_2\sqrt{q},$$

$0 < c_1 \leq 1$ and $0 < c_2 \leq 1$ are sparsity parameters.

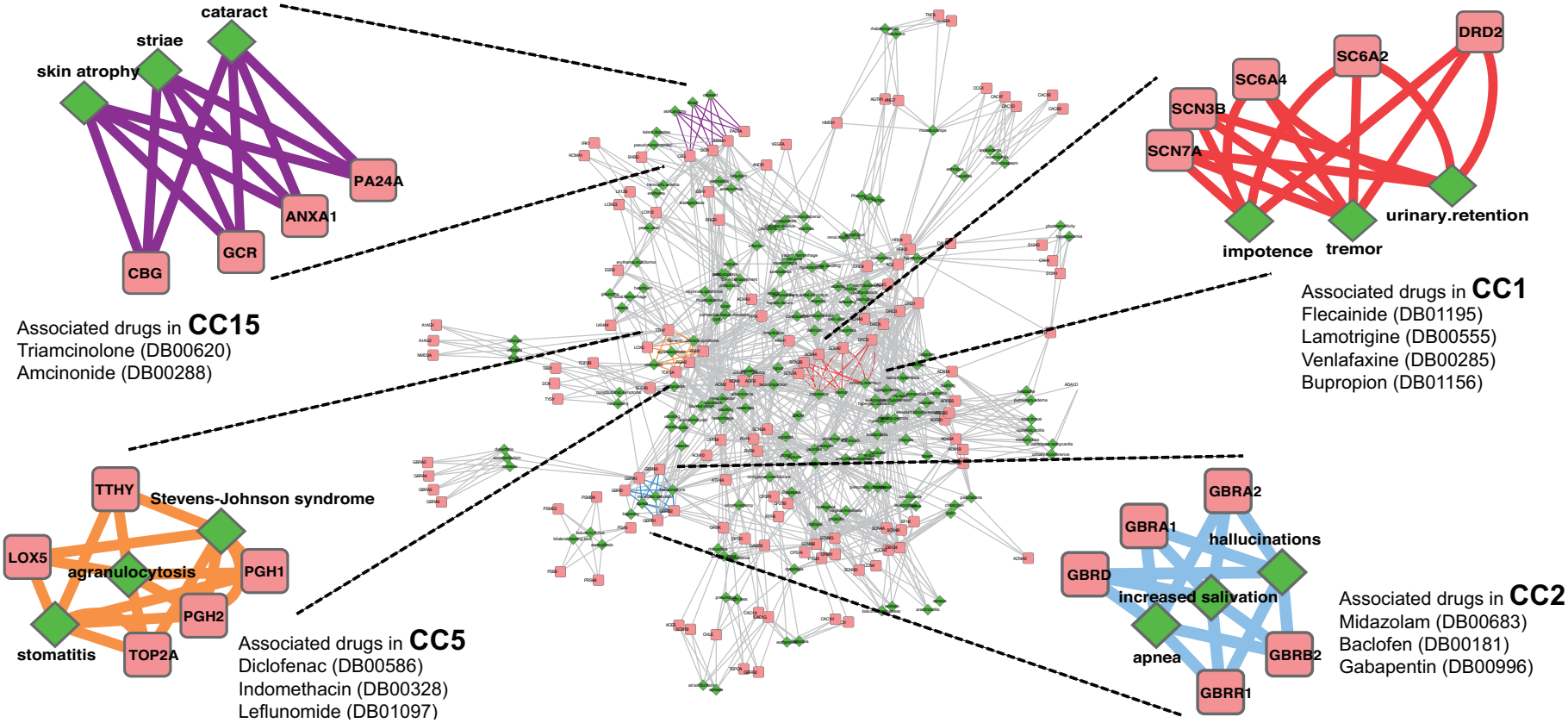
4 For n drugs, given $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$, $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T$, the weight vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be optimized by solving penalized matrix decomposition of the matrix $X^T Y$ (Witten *et al.*, *Biostatistics* 2009).

Weight vectors in SCCA





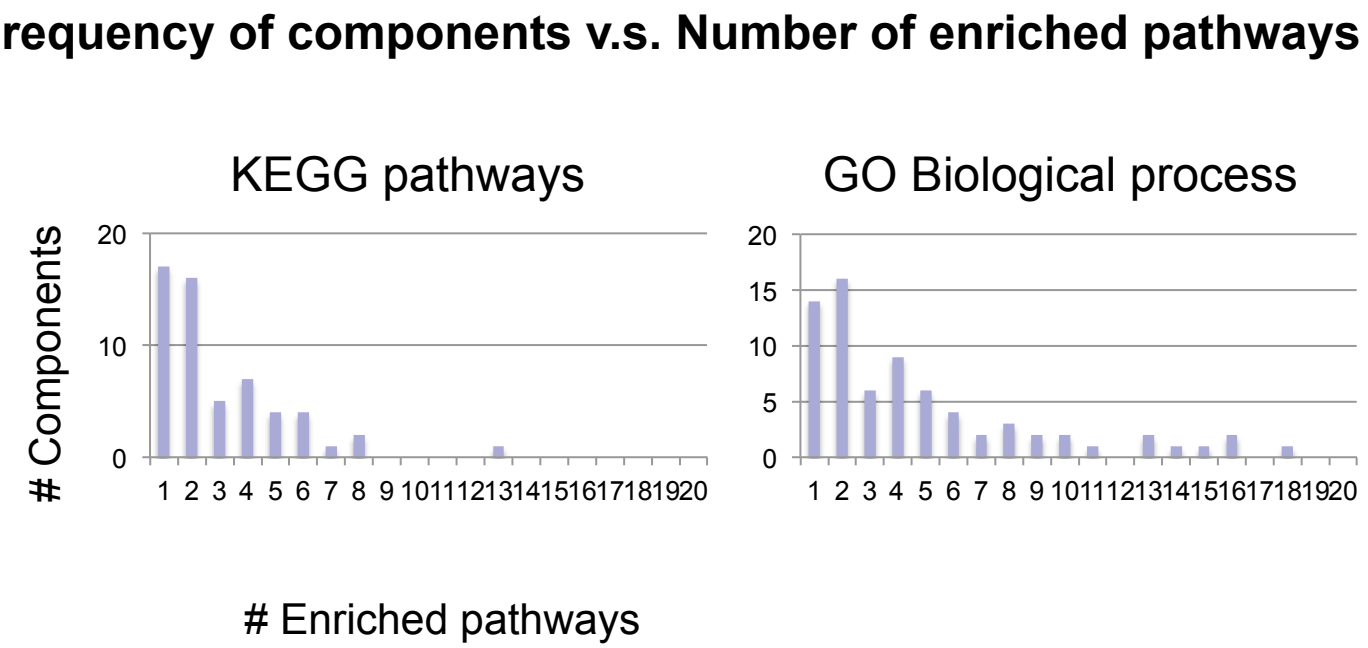
Extracted protein-effect association network across 80 components in SCCA



Biological evaluation of extracted components

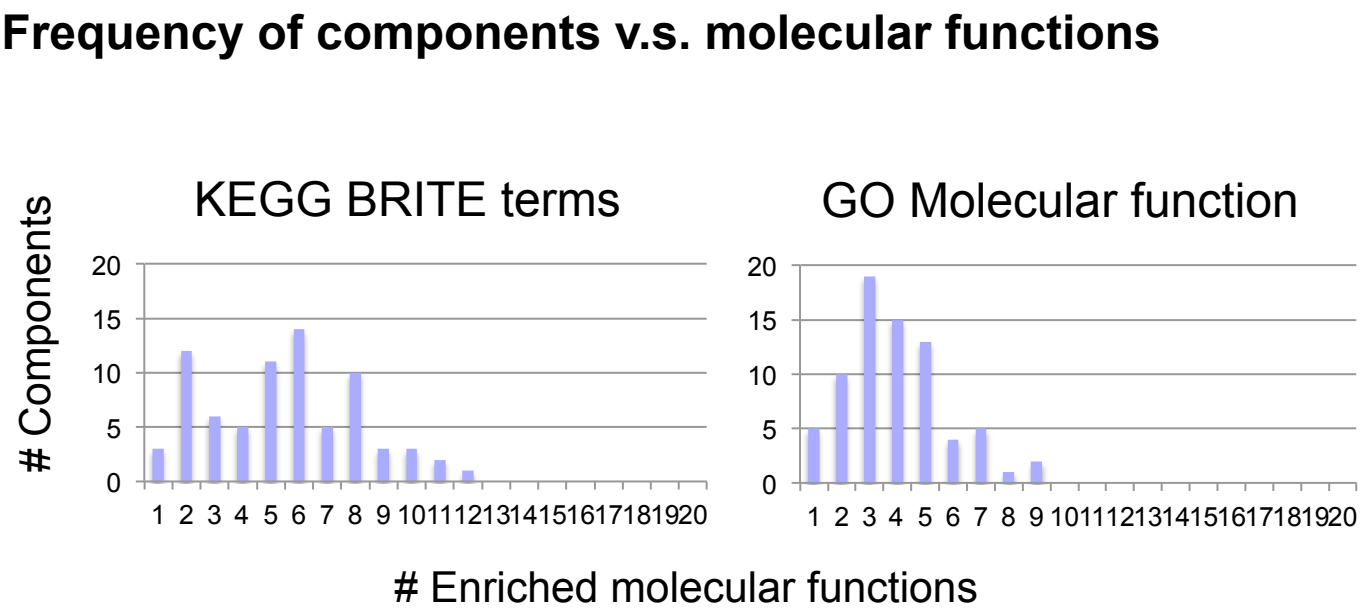
- Evaluation of the components by pathway enrichment analysis
  - How each of the correlated sets is enriched with proteins that are involved in a specific pathway?
- Pathway databases
  - KEGG : pathway maps
  - GO (Gene Ontology) : biological process terms

Pathway enrichment of proteins in components



The components were enriched with a small number of pathways.

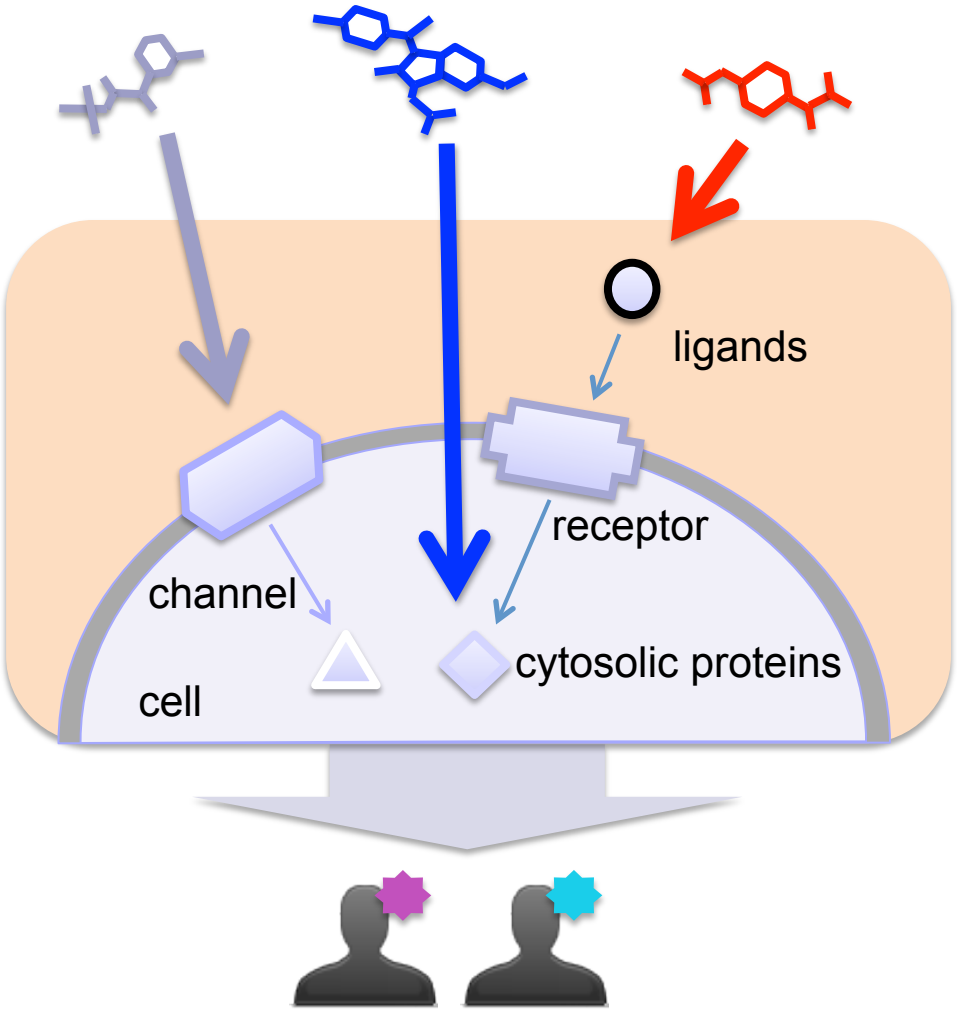
Molecular function enrichment of proteins in components



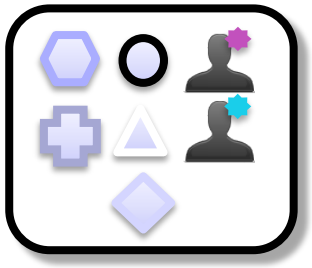
The components were enriched with a small number of pathways, but with various number of different molecular functions.

Biologically relevant interpretation of correlated sets

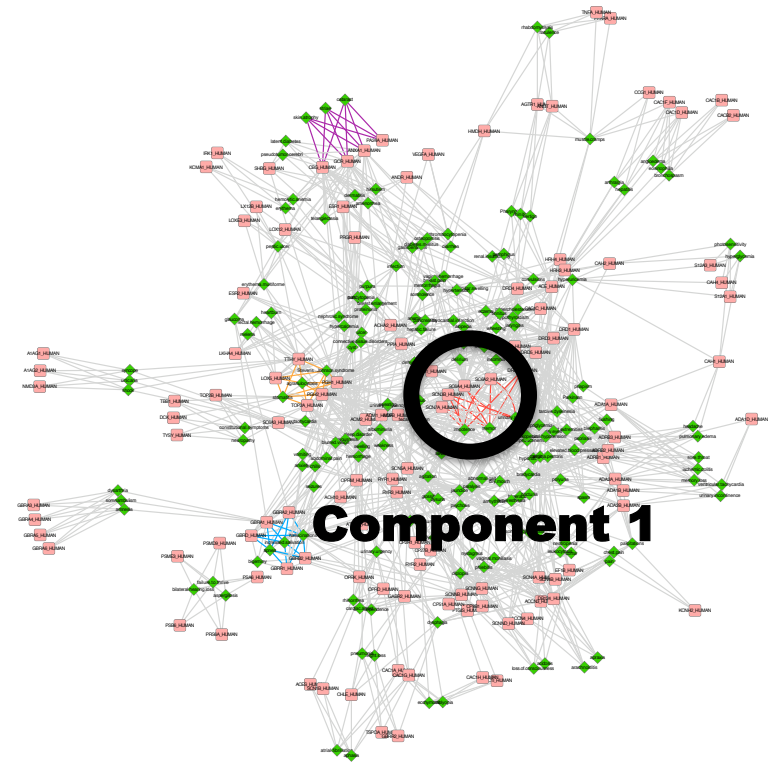
Interactions between drugs with proteins of different molecular functions tend to induce the regulation of the whole protein pathway, which lead to certain side-effects.

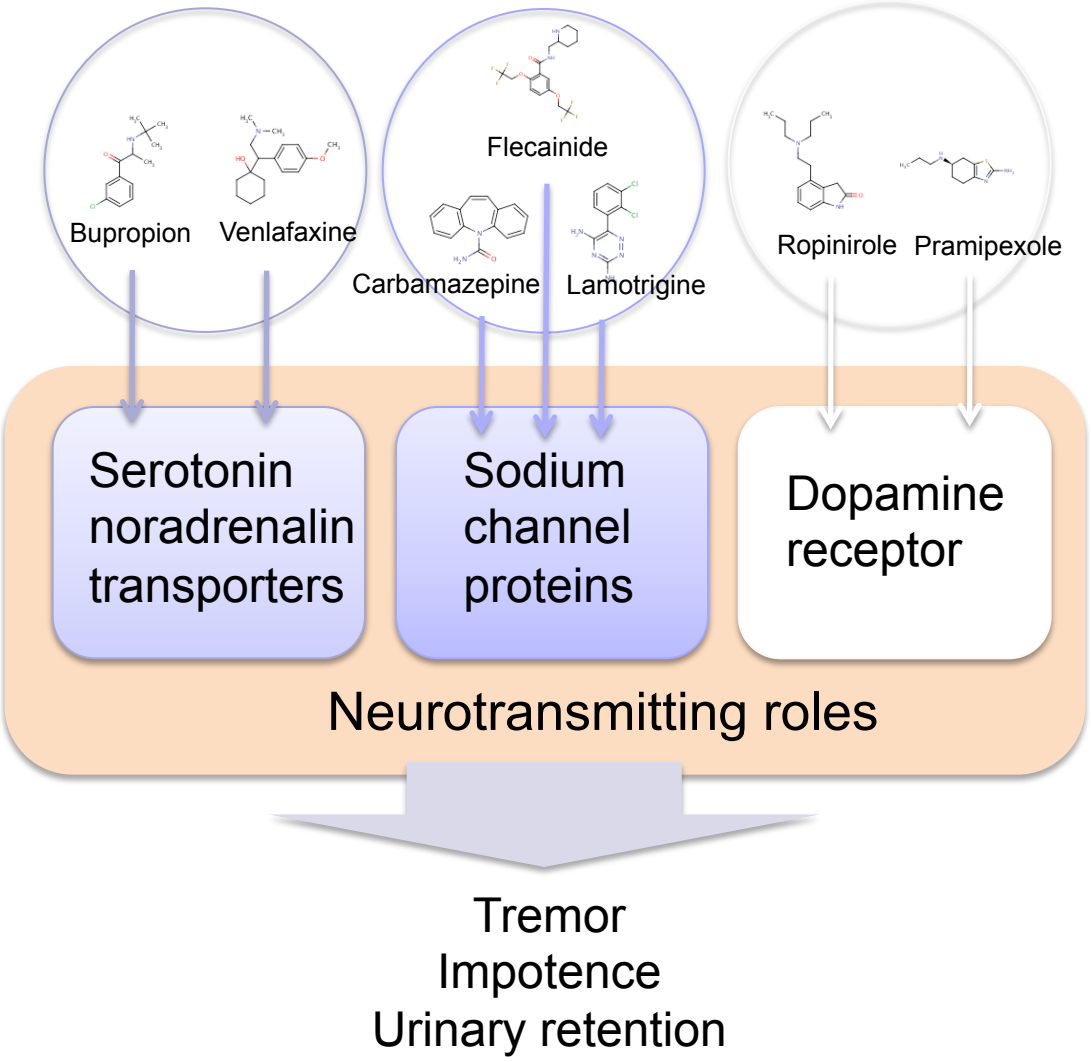


Component

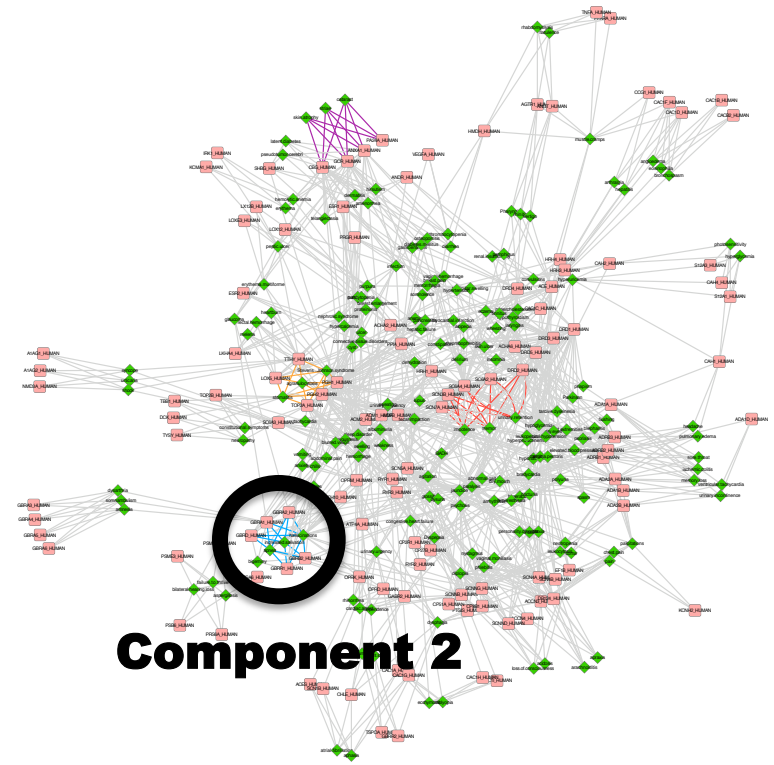


An example of extracted components

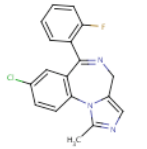


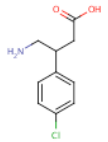


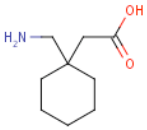
An example of extracted components



anti-anxiety, anesthesia adjuvants, …


Midazolam


Baclofen

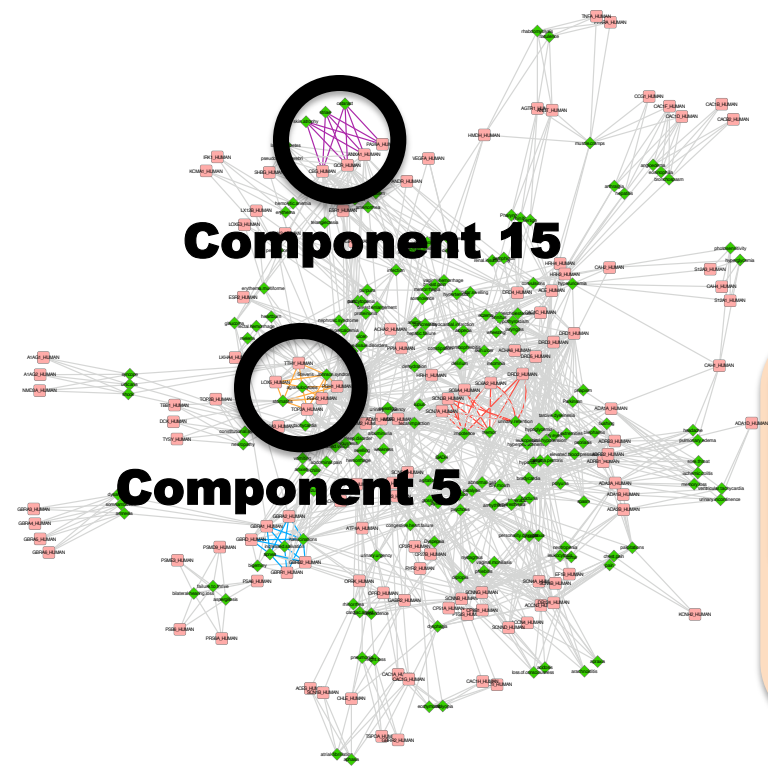

Gabapentin

GABA receptor subunits

GABA receptor

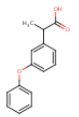
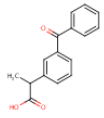
Increased salivation
Hallucinations
apnea

An example of extracted components



Component 5

Anti-inflammatory

Fenoprofen Ketoprofen

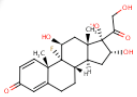
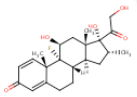
Prostaglandin G/H synthase 1 & 2
Arachidonate 5-lipoxygenase

Arachidonic acid metabolism

Stevens-Johnson syndrome
Stomatitis
Agranulocytosis

Component 15

Anti-inflammatory

Triamcinolone Dexamethazone

Annexin 1A
Glucocordicoid receptor
corticosteroid-binging globulin
phospholipase A2

Glucocorticoid effects

Striae
Skin atrophy
Cataract

Performance evaluation for side-effect prediction by 5-fold cross-validation

Prediction score :

$$\mathbf{y}_{new} = B^{-T} A^{T} \mathbf{x}_{new}$$

where $A = [\alpha_1, ..., \alpha_m]$, $B = [\beta_1, ..., \beta_m]$, m : number of components

Method	AUC ± S.D.	AUPR ± S.D.
Chemical structure based approach:		
Random	0.5000 ± 0.0000	0.0556 ± 0.0000
OCCA	0.8355 ± 0.0010	0.3753 ± 0.0016
SCCA	0.8708 ± 0.0007	0.3766 ± 0.0030
Target protein based approach:		
Random	0.5000 ± 0.0000	0.0556 ± 0.0000
OCCA	0.8850 ± 0.0007	0.4067 ± 0.0006
SCCA	**0.8895 ± 0.0002**	**0.4103 ± 0.0018**

Examples of predicted side-effects by SCCA

- For 730 uncharacterized drugs in DrugBank

Top 20 predicted side-effects

	Drug	Side-effect	Score
1	Cinnarizine	*tremor*	1.176339
2	Benzocaine	*diplopia*	1.163882
3	Cinnarizine	*constipation*	1.143190
4	Bepridil	*tremor*	1.046472
5	Cinnarizine	*somnolence*	0.996260
6	Cinnarizine	*dry.mouth*	0.961833
7	Cinnarizine	*angioedema*	0.955471
8	Cinnarizine	*insomnia*	0.950757
9	Benzocaine	*nausea*	0.947105
10	Alprenolol	*dizziness*	0.943918

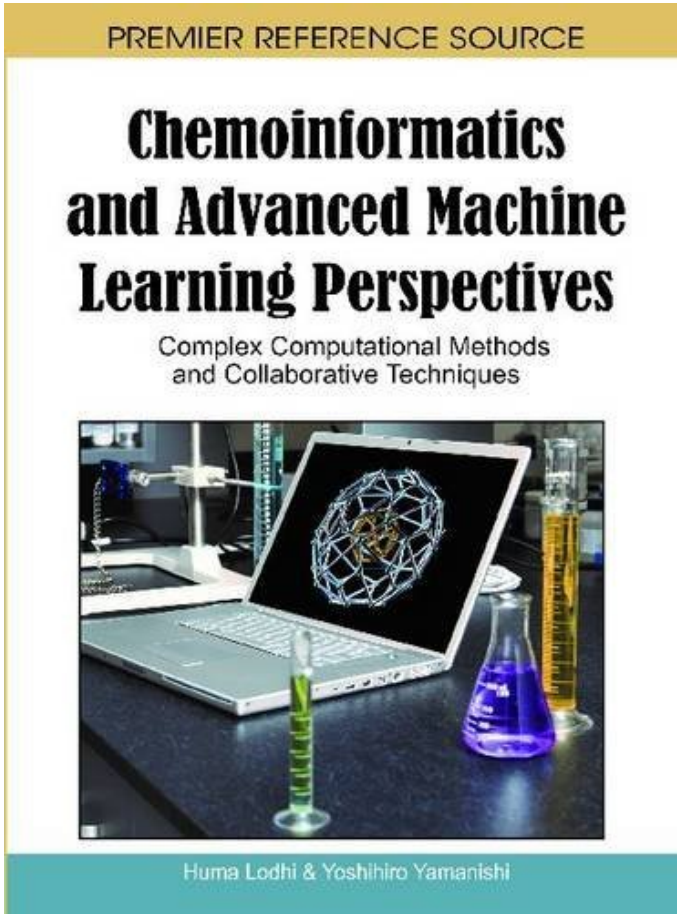
	Drug	Side-effect	Score
11	Benzocaine	*diarrhea*	0.937409
12	Nisoldipine	*tremor*	0.933258
13	Nitrendipine	*tremor*	0.933258
14	Lercanidipine	*tremor*	0.933258
15	Benzocaine	*syncope*	0.928905
16	Promazine	*tachycardia*	0.926875
17	Promazine	*somnolence*	0.924103
18	Benzocaine	*vomiting*	0.922091
19	Bepridil	*nervousness*	0.918828
20	Cinnarizine	*nervousness*	0.918615

- Side-effects reported in the literature are colored in red.

Summary

- The proposed method enables us to relate target proteins with side-effects of drugs.
- It may be useful to predict unknown side-effects in the drug development.

Machine learning in chemoinformatics



Lodhi, H. and Yamanishi, Y.,
Chemoinformatics and Advanced Machine Learning Perspectives,
IGI Global, 2010.

Acknowledgements

- Kyushu University
  - Mikita Suyama, Tetsuya Sato
- Kyushu Institute of Technology
  - Hiroto Saigo
- Kyoto University
  - Minoru Kanehisa, Susumu Goto, Masaaki Kotera, Masataka Takarabe, Sayaka Mizutani, Yosuke Nishimura
- JST
  - Yasuo Tabei
- Curie Institute - Inserm U900 - Mines ParisTech
  - Jean-Philippe Vert, Véronique Stoven, Kevin Bleakley, Edouard Pauwels