

# Qualitatively Predicting Compound-Protein Interactions by Multi-Task Learning

PhD Candidate: Songpeng Zu

Advisor: Prof. Shao Li

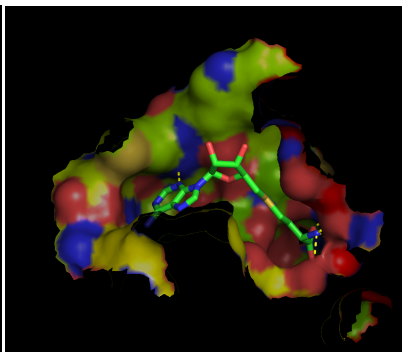
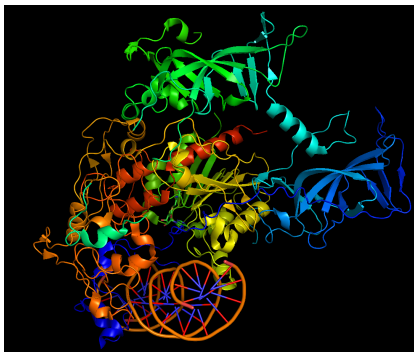
Bioinformatics Lab, Department of Automation  
Tsinghua university

May 15, 2016

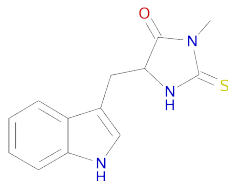
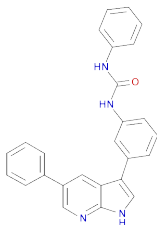
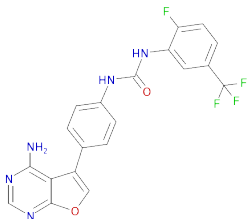
## outline

- ▶ Background:
  - Quantitative Structure-Activity Relationship (QSAR)
  - Multi-task Learning
- ▶ Method: hierarchical Bayesian model called MulTQSAR
- ▶ Result: reduce MSE on PeptideGPCR
- ▶ Discussion:
  - Sparsity and feature selection
  - Multi-task deep learning on QSAR

# Compound-Protein Interactions (CPIs)

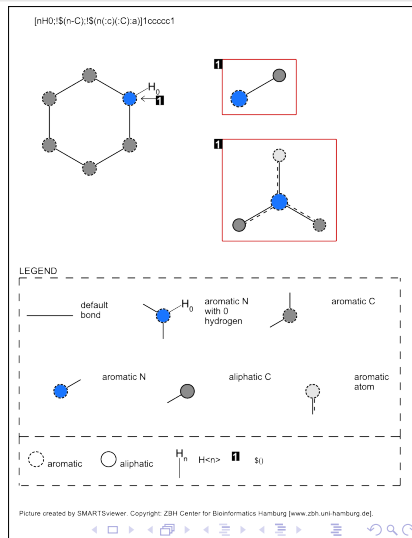


# Protein Versus Multiple Compounds



# Chemical Space: Representation of Compounds

- ▶ Compounds represented by topological fingerprints. The similar as proteins.
- ▶ CPIs recorded as binary variable or continuous variables.
- ▶ Classification or regression models then are used.



# Single Protein QSAR Model

For the protein  $l$ , we have  $n_l$  compounds. Let  $\mathbf{x}_i^l$  represents the compound  $i$ 's features in the chemical space, and  $\mathbf{y}_i^l$  represents the interaction affinity between the compound  $i$  and protein  $l$ .

QSAR is then to solve the problem:

$$f = \arg \min_f \mathcal{L}(\mathbf{y}^l, f(\mathbf{X}^l)) \quad (1)$$

in which,  $\mathbf{X}^l = (\mathbf{x}_1^l, \dots, \mathbf{x}_{n_l}^l)^t$ ,  $\mathcal{L}(\cdot, \cdot)$  is the loss function. Usually we treat it as a linear regression model, *i.e.*,

$$\omega^l = \arg \min_{\omega^l} \|\mathbf{y}^l - \mathbf{X}^l \omega^l\|^2 \quad (2)$$

# Protein Family

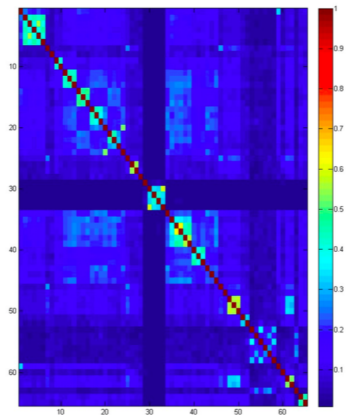
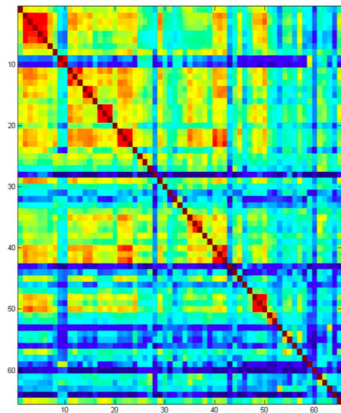
```

Q5E940_BOVIN -----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE
RLA0_HUMAN -----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE
RLA0_MOUSE -----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE
RLA0_RAT -----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE
RLA0_CHICK -----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE
RLA0_RANSY -----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE
Q7ZUG3_BRARE -----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE
RLA0_ICTPU -----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE
RLA0_DROME -----MVRENKAAWKAQYFIKVVLEDFPKCFIVGADNVGSKQMQQIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE
RLA0_DICDI -----MSGAG-SKRKKLFIEKATKLTFTYDKMIVAEADFYGSQLOKIRKSIRGI-GAVLMGKNTMIRKVIIRDADSK--PELD
Q54LP0_DICDI -----MSGAG-SKRKNVFIKATKLTFTYDKMIVAEADFYGSQLOKIRKSIRGI-GAVLMGKNTMIRKVIIRDADSK--PELD
RLA0_PLAF8 -----MAKLSKQKKQMYIEKLSLIQQYSKILIVHDVNGVNSMASYRKSLSRGK-ATILMGKNTMIRKVIIRDADSK--PELD
RLA0_SULAC -----HIGLAVTTTKKIAKWKVDEVAELTEKLTHTITIANIEGFPADKLHEIRKKLRGK-ADIKVTKNLNFIAKKNAG--YDEK
RLA0_SULTO -----MRIMAVITQERKIAKWKIEVKELEKREYHTITIANIEGFPADKLHEIRKKLRGK-ADIKVTKNLNFIAKKNAG--YDEK
RLA0_SULSO -----MKRLALALKQKRVASWKEEYKELTELKNSHTILIGNLGFPADKLHEIRKKLRGK-ADIKVTKNLNFIAKKNAG--YDEK
RLA0_AERPE HSYVSLVGYQMYKREKPIPEWKTMLRELEELFSKRVVLFADLTGSPFVYQVRVKKLWKK-YPMVYAKKRIILAMKAAGLE--LDDN
RLA0_PYRAE HMLATGKRRYVRTQYPAKVKYISEATELLQKYVYVFLDPLGLGSRILHEVRYRLERY-GVIKTIKPLFKIAFTKVVYGG--IPAL
RLA0_METAC -----MAEERHNTENIPQWKKDEIENIKELIQSHKVFQMVGIEGILATKMKIRRDLDKV-AVYLMGKNTMIRKVIIRDADSK--PELD
RLA0_METMA -----MAEERHNTENIPQWKKDEIENIKELIQSHKVFQMVGIEGILATKMKIRRDLDKV-AVYLMGKNTMIRKVIIRDADSK--PELD
RLA0_ARCFU -----MAAVRGS--PPEYKVRVAVIEIKRMISSKVVVAIVSFRNVPAGQMKIRREFRGK-AEIKVYKNTLLERLADALG--GDYL
RLA0_METKA HAVKAKGQPPSCYEKKVAENKRRVKELELMDYENYVGLVDLEGIPAPQLOEIRAKLRERETITMRRNTLMRIALEEKLDER--PELE
RLA0_METTH -----MAVVAENKKKEVQELHDLIKGYEVVGIANLADIPARQLOKMQTLRDS-ALIRMSKNTLLISALEKAGREL--ENVD
RLA0_METTL -----MITAESENKIAPIWKIEEYVKLELLKNGQIVALVDMMEVPARQLOEIRDKIR-DMTLMKMSRNTLLIRAKKEVAEETONPEFA
RLA0_METVA -----MIDAKSENKIAPIWKIEEYVKLELLKNSANVIALIDHMEVPAPQLOEIRDKIR-DMTLMKMSRNTLLIRAKKEVAEETONPEFA
RLA0_METJA -----METVKYHVAAPWKIEEYKTLKGLIKSPVVAIVDMMDVPAPQLOEIRDKIR-DMTLMKMSRNTLLIRAKKEVAEELNPKLA

```



# Can We Learn QSAR Models In A Protein Family?





# Learning Different But Similar Tasks

- ▶ Learning multiple tasks together, one type of transfer learning (Pan S. and Yang Q., 2010).
- ▶ Examples:
  - 1 Multi-task feature selection (Obozinski G., Taskar B., Jordan, M., 2006)

$$\min_{\omega} \sum_{l=1}^L \frac{1}{N_l} \sum_{i=1}^{N_l} \mathcal{L}(\omega^l, x_i^l, y_i^l) + \lambda \sum_{j=1}^p \|\omega_j\|^2 \quad (3)$$

in which,  $p$  is the feature number,  $L$  is the task number.

- 2 Adaptive multi-task LASSO (Lee S., Zhu J., and Xing E., 2010)

$$\min_{\omega} \frac{1}{2} \sum_{l=1}^L \|Y^l - X\omega^l\|_2^2 + \lambda_1 \sum_{j=1}^p \theta_j \sum_{l=1}^L |\omega_j^l| + \lambda_2 \sum_{j=1}^p \rho_j \|\omega_j\|_2 \quad (4)$$

# Statistics Behind Linear Regression

Suppose  $\mathcal{D}^l = \{\mathbf{X}^l, \mathbf{y}^l\}$ ,  $l = 1, \dots, L$ , and  $\mathbf{X}^l \in \mathbb{R}^{n^l \times p}$ . Then we have

$$\mathbf{y}^l \sim \mathcal{N}(\mathbf{X}^l \omega^l, \sigma_y^2 \mathbf{I}) \quad (5)$$

$$\omega^l \sim \mathcal{N}(\omega_*, \sigma_l^2 \mathbf{I}) \quad (6)$$

$$\omega^* \sim \mathcal{N}(\mathbf{0}, \sigma_*^2 \mathbf{I}) \quad (7)$$

Here we assume that  $p(\sigma_y^2) \propto 1$ . Let  $\Theta = \{\omega^l, l = 1, \dots, L, \omega^*, \sigma_y^2\}$ .  
We have

$$\begin{aligned}
 \mathcal{L}_{hier}(\mathcal{D}; \Theta) &= \mathcal{L}_{orig}(\mathcal{D}|\Theta) + \log p(\Theta) \\
 &= \sum_{l=1}^L \left( \log p(\mathcal{P}^l | \omega^l) - \frac{\|\omega^l - \omega^*\|^2}{2\sigma_l^2} \right) - \frac{\|\omega_*\|^2}{2\sigma_*^2} \\
 &\quad - \sum_{l=1}^L \frac{p}{2} \log(2\pi\sigma_l^2) - \frac{p}{2} \log(2\pi\sigma_*^2) \\
 &= \sum_{l=1}^L \left( -\frac{\|\mathbf{y}^l - \mathbf{X}^l \omega^l\|^2}{2\sigma_y^2} - \frac{\|\omega^l - \omega^*\|^2}{2\sigma_l^2} \right) - \frac{\|\omega^*\|^2}{2\sigma_*^2} - \sum_{l=1}^L \frac{n^l}{2} \log(2\pi\sigma_y^2) \\
 &\quad - \sum_{l=1}^L \frac{p}{2} \log(2\pi\sigma_l^2) - \frac{p}{2} \log(2\pi\sigma_*^2)
 \end{aligned} \tag{8}$$

We can use MCMC algorithm to simulate the posterior distribution of  $\Theta$ . Here  $\sigma_l^2$ ,  $\sigma_*$  are fixed, and L-BFGS-B is then used following the gradient below.

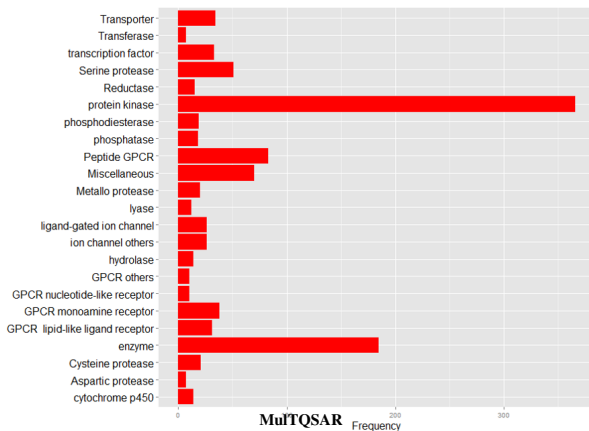
$$\begin{aligned}\frac{\partial \mathcal{L}_{hier}(\mathcal{D}; \Theta)}{\partial \omega^l} &= -\frac{1}{2\sigma_y^2} \frac{\|\mathbf{y}^l - \mathbf{X}^l \omega^l\|^2}{\partial \omega^l} - \frac{1}{2\sigma_l^2} \frac{\|\omega^l - \omega^*\|^2}{\partial \omega^l} \\ &= \frac{\mathbf{y}_l^t \mathbf{X}^l}{\sigma_y^2} + \frac{\omega^*}{\sigma_l^2} - \left( \frac{\mathbf{X}_l^t \mathbf{X}^l}{\sigma_y^2} + \frac{1}{\sigma_l^2} \mathbf{I} \right) \omega^l\end{aligned}\quad (9)$$

$$\frac{\partial \mathcal{L}_{hier}(\mathcal{D}; \Theta)}{\partial \omega^*} = -\sum_l \frac{\omega^* - \omega^l}{\sigma_l^2} - \frac{\omega^*}{\sigma_*^2} \quad (10)$$

$$\frac{\partial \mathcal{L}_{hier}(\mathcal{D}; \Theta)}{\partial \sigma_y^2} = \frac{\sum_l \|\mathbf{y}^l - \mathbf{X}^l \omega^l\|^2}{2\sigma_y^2} - \frac{n}{2\sigma_y^2} \quad (11)$$

where  $n$  is the total number of samples in all the groups.

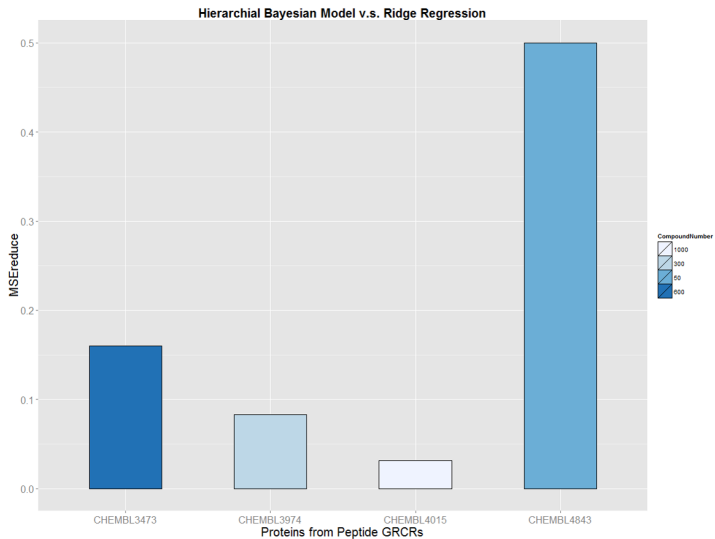
- ▶ 210,000 CPIs including more than 1,000 proteins from 20 protein families, and 150,000 compounds.
- ▶ 22 physicochemical properties and 881 chemical substructures as the compounds' features.



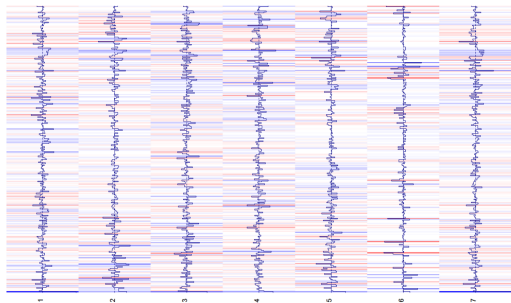
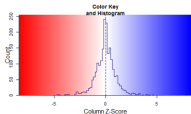
# Feature Selection

The protein family of Peptide GPCR including 85 proteins as examples.

- ▶ Based on the definitions of chemical fingerprints, SUB1-SUB115, SUB264-SUB327 are removed.
- ▶ Chemical fingerprints with too low or high frequencies are removed.
- ▶ Non-parametric dynamic slicing method for marginal feature selection.
- ▶ 284 features are finally kept.



We can involve L1 regularization for sparsity and feature selection.





## How can we involve features' combinations?

