Qualitatively Predicting Compound-Protein Interactions by Multi-Task Learning

Songpeng Zu
zusongpeng@gmail.com

May 13, 2016

outline

- Quantitatively predicting compound-protein interactions by multi-task learning
- Paper Review on Compound-Protein Interactions.

Part I

Quantitatively predicting compound-protein interactions by multi-task learning

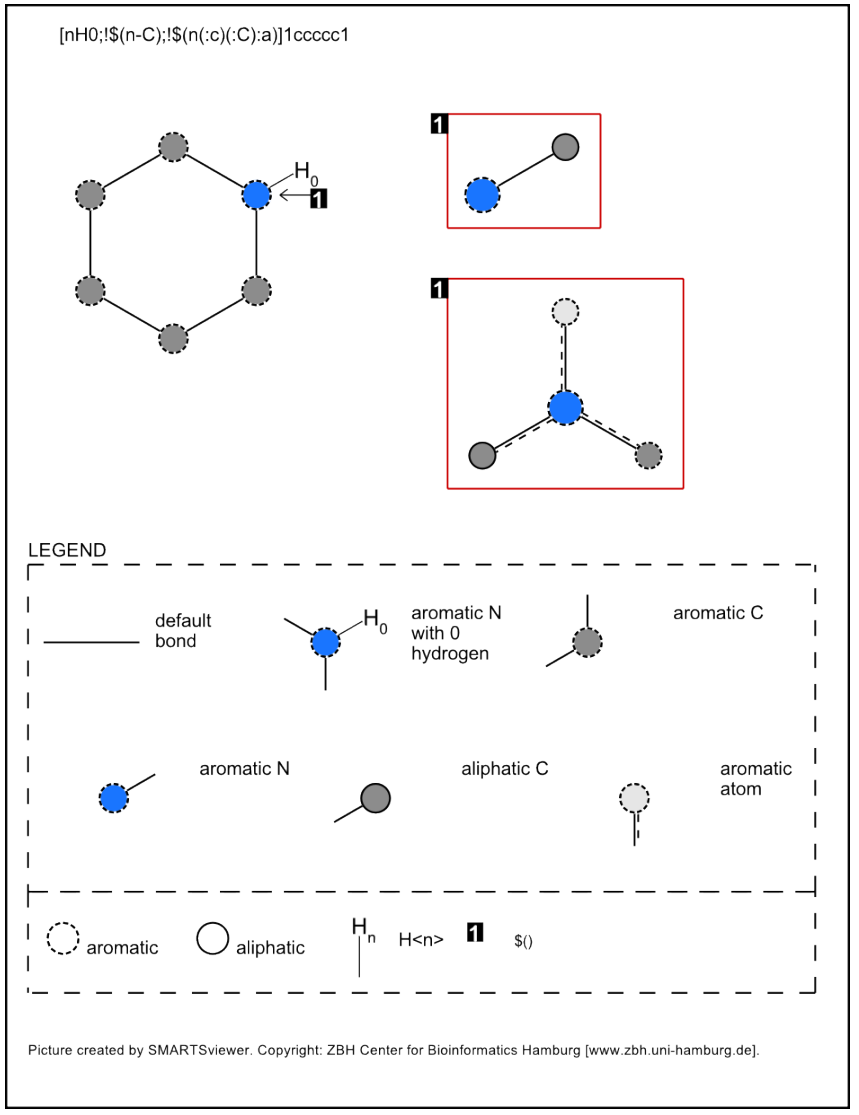
background

Methods for predicting compound-protein interactions (CPIs):

- structure-based molecular dynamics
  - depend on proteins' 3D structures.
- ligand-based method
  - can be independent of proteins' 3d structures.
  - large-scale known CPIs data
  - mainly dependent on machine learning approaches.

Machine learning on CPIs

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



Modeling on a single protein

Keiser M.J. *et al.*, *Nature* 2009, developed the SEA method to predict drugs' new molecular targets.

- Each target represented by its set of known ligands.
- Drugs computationally screen against a panel of proteins by comparing the similarity of ligands against these proteins.
- The similarities expressed as E-values, adapting the BLAST algorithm.

Modeling on a single protein

Besnard J. *et al*., *Nature* 2012, used naive Bayesian model to predict compounds' polypharmacology profiling.

- 215,000 activity data including 133,061 compounds and 784 proteins were used.
- Every compounds represented by the binary vectors of ECFP6 representations.
- For every protein, a Laplacian-modified naive models was built for classification.

Modeling on a protein family

Yabuuchi H. *et al.*, *Molecular Systems Biology* 2011 developed the CGBVS framework.

- 5207 CPIs data (including 317 GPCRs and 866 ligands)
- Compounds' structure and proteins' sequences converted into 929- and 400-dimensional vectors
- SVM then used.

Machine Learning on CPIs

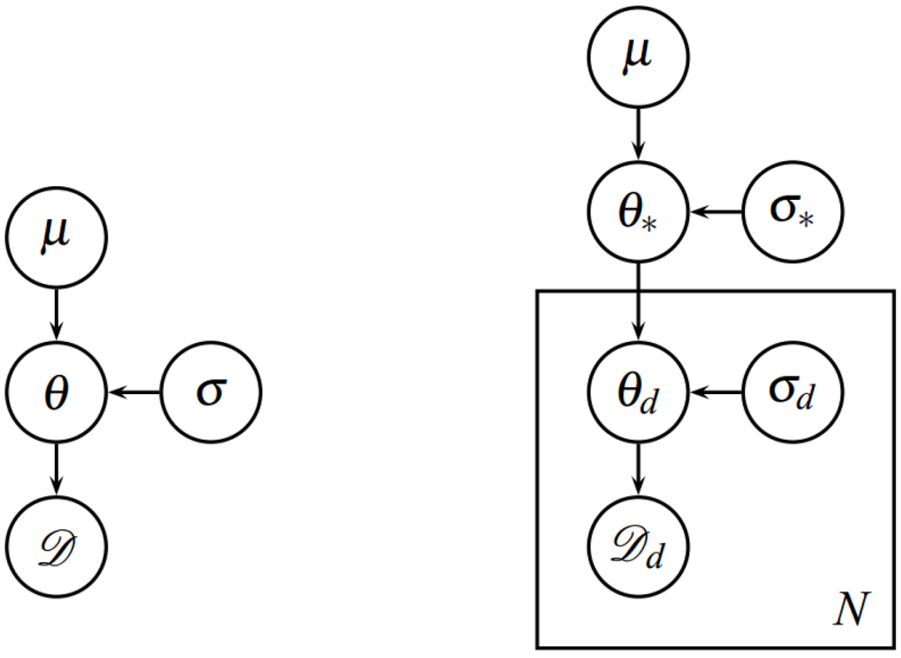
Current machine learning on predicting CPIs

- Modeling on a single protein
  More specificity Lots of data needed
- Modeling on a protein family
  Data sharing Less specificity

Multi-Task Learning

Can we combine the two approaches ?

- Learning different but similar tasks at the same time. (Finkel J.R. and Mannning C.D., 2009)
- Quantitative prediction.



(a) **(b)**

Hierarchical Bayesian Model

Suppose $\mathcal{D}_j = \{\mathbf{X}_j, \mathbf{y}_j\}, j = 1, ..., m$, and $\mathbf{X}_j \in \mathbb{R}^{d \times n_j}$. Then we have

$$\mathbf{y}_j \sim \mathcal{N}\left(\mathbf{X}_j^T \omega_j, \sigma_y^2 \mathbf{I}\right) \tag{1}$$

Since different groups data may share similar features, we assume ω_j have the same mean on the prior distribution.

$$\omega_j \sim \mathcal{N}\left(\omega_*, \sigma_j^2 \mathbf{I}\right) \tag{2}$$

In which,

$$\omega_* \sim \mathcal{N}\left(\mu, \sigma_*^2 \mathbf{I}\right) \tag{3}$$

Suppose, for simplicity, that $\mu = \mathbf{0}$, $p(\sigma_y^2) \propto 1$, and that σ_j^2 and σ_* are fixed. Let $\Theta = \left\{\omega_j, j = 1, ..., m, \omega_*, \sigma_y^2\right\}$. We have

$$
\begin{aligned}
&\mathcal{L}_{hier}(\mathcal{D};\Theta) = \mathcal{L}_{orig}(\mathcal{D}|\Theta) + logp(\Theta) \\
&= \sum_j \left(logp(\mathcal{D}_j|\omega_j) - \frac{\| \omega_j - \omega_* \|^2}{2\sigma_j^2} \right) - \frac{\| \omega_* \|^2}{2\sigma_*^2} \\
&\overbrace{- \sum_j \frac{d}{2} log(2\pi\sigma_j^2) - \frac{d}{2} log(2\pi\sigma_*^2)}^{Const} \\
&= \sum_j \left(- \frac{\| \mathbf{y}_j - \mathbf{X}_j^T \omega_j \|^2}{2\sigma_y^2} - \frac{\| \omega_j - \omega_* \|^2}{2\sigma_j^2} \right) - \frac{\| \omega_* \|^2}{2\sigma_*^2} - \sum_j \frac{n_j}{2} log(2\pi\sigma_y^2) \\
&\overbrace{- \sum_j \frac{d}{2} log(2\pi\sigma_j^2) - \frac{d}{2} log(2\pi\sigma_*^2)}^{Const}
\end{aligned}
\tag{4}
$$

L-BFGS-B optimization method is then used following the gradient below.

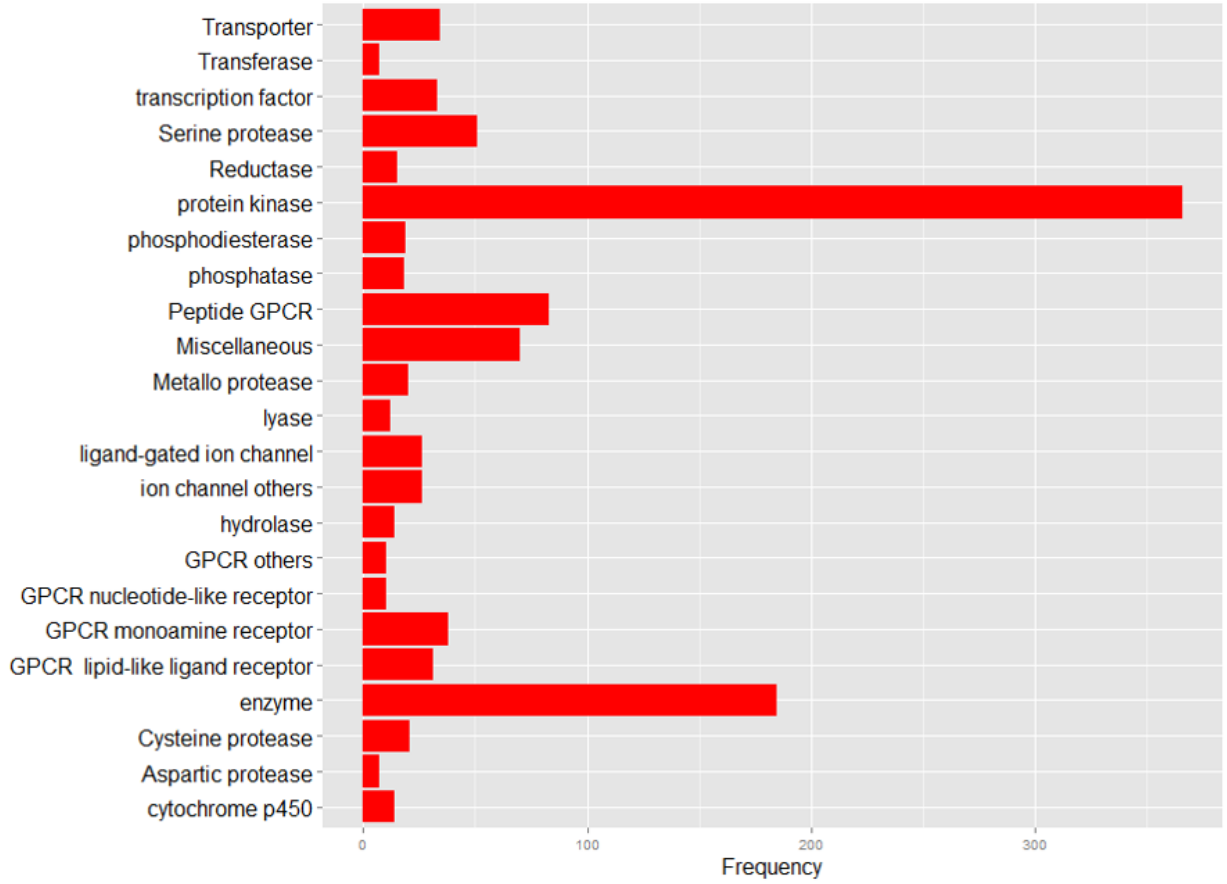
$$\begin{aligned}\frac{\partial \mathcal{L}_{hier}(\mathcal{D};\Theta)}{\partial \omega_j} &= -\frac{1}{2\sigma_y^2}\frac{\parallel \mathbf{y}_j - \mathbf{X}_j^T\omega_j \parallel^2}{\partial \omega_j} - \frac{1}{2\sigma_j^2}\frac{\parallel \omega_j - \omega_* \parallel^2}{\partial \omega_j} \\ &= \frac{\mathbf{X}_j\mathbf{y}_j}{\sigma_y^2} + \frac{\omega_*}{\sigma_j^2} - \left(\frac{\mathbf{X}_j\mathbf{X}_j^T}{\sigma_y^2} + \frac{1}{\sigma_j^2}\mathbf{I}\right)\omega_j\end{aligned} \tag{5}$$

$$\frac{\partial \mathcal{L}_{hier}(\mathcal{D};\Theta)}{\partial \omega_*} = -\sum_j \frac{\omega_* - \omega_j}{\sigma_j^2} - \frac{\omega_*}{\sigma_*^2} \tag{6}$$

$$\frac{\partial \mathcal{L}_{hier}(\mathcal{D};\Theta)}{\partial \sigma_y^2} = \frac{\sum_j \parallel \mathbf{y}_j - \mathbf{X}_j^T\omega_j \parallel^2}{2(\sigma_y^2)^2} - \frac{n}{2\sigma_y^2} \tag{7}$$

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.

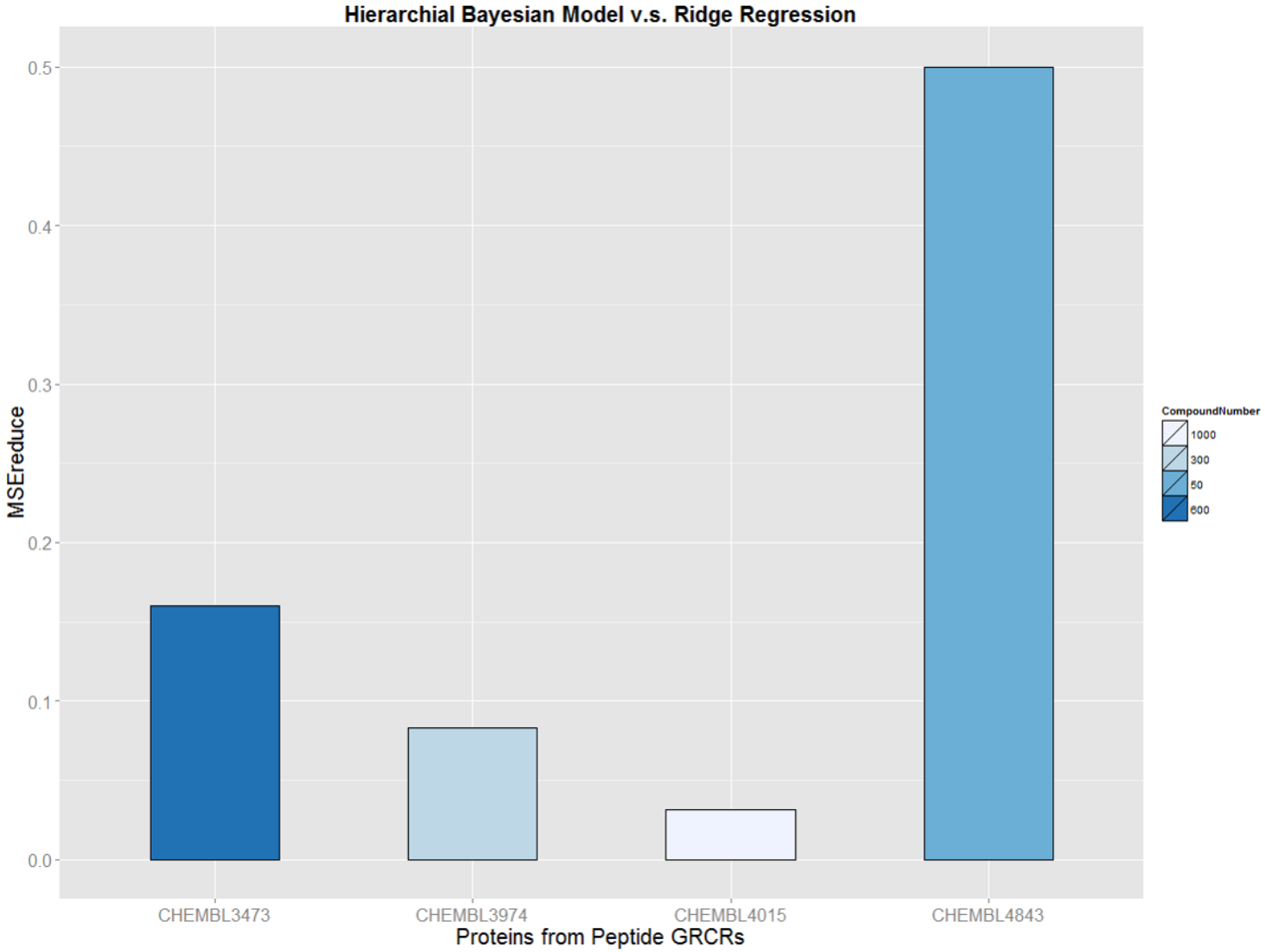


Feature Selection

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.

Comparison with Ridge Regression



Discussion

- More compounds' fingerprints are being collected by the open-source chemoinformatics and machine learning package termed RDKit.
- Computational issue:
  - High dimension v.s. Sparsity
    Colinearity
  - Linear v.s. Nonlinear

Part II

Paper Review

A systems approach to TOM

Hyun Uk Kim *et al.*, Nature Biotechnology, March 2015.

- Analyzed the structural similarities between the compounds derived from TOM and human metabolites.
- Explained 38 TOM-derived synergistic combination, i.e., connected the Major, Complementary, Neutralizing, and Delivery/retaining with the molecular biological views.
  - Complementary action: Major and Complementary
  - Neutralizing action: Major and Neutralizing
  - Facilitating action/Pharmacokinetic potentiation: Major and Delevery/retaining.

Data Summary

- 4,679 active compounds in TOMs were from TCM Database at Taiwan. 38 synergistic combinations of TOM compounds were identified by the literature since 2000.
- The human metabolites were downloaded from KEGG.
- As a control, 316 approved drugs were downloaded from the DrugBank 3.0.

Compared with Metabolites

simMetobolic.png

Complementary action

complementary.png

Neutralizing action

neutra.png

Facilitating action and Pharamacokinetic potentiation

pharmaco.png

Elucidating the mechanism of action

distrisubmeta.png

Part III

Integrating multi-level similarities to improve the network-based prediction on CPIs

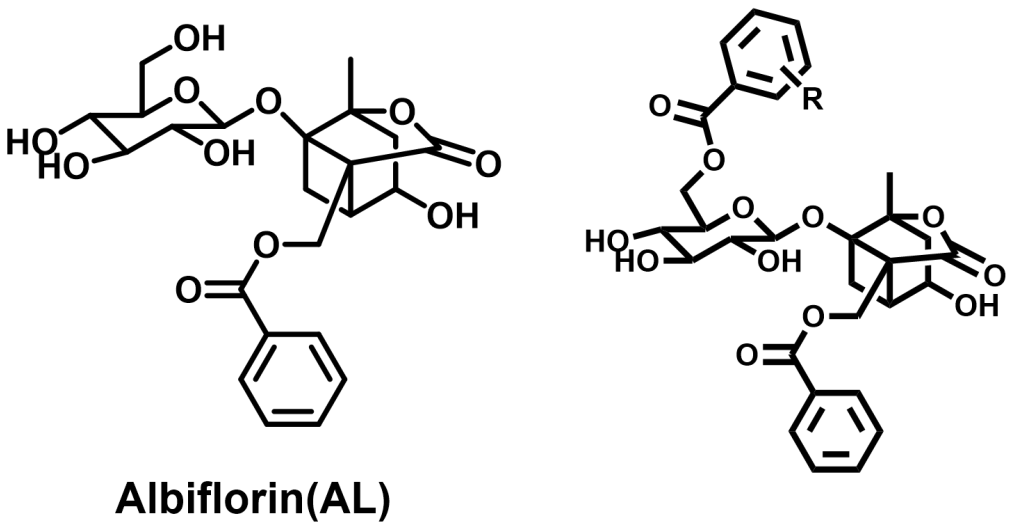
Part IV

Application on the Modification of Natural Products

Lead Discovery From TCM Herbs

- Natural products important sources for drug discovery.
- By DrugCIPHER, several compounds from traditional Chinese Medicine (TCM) Herbs are predicted to have the antitumor activities.
- Many of them have been reported, but one compound called *Albiflorin* few researches.

- Our experiments: *Albiflorin* has the antitumor activities with low potency.
- Its biological mechanism is unknown.



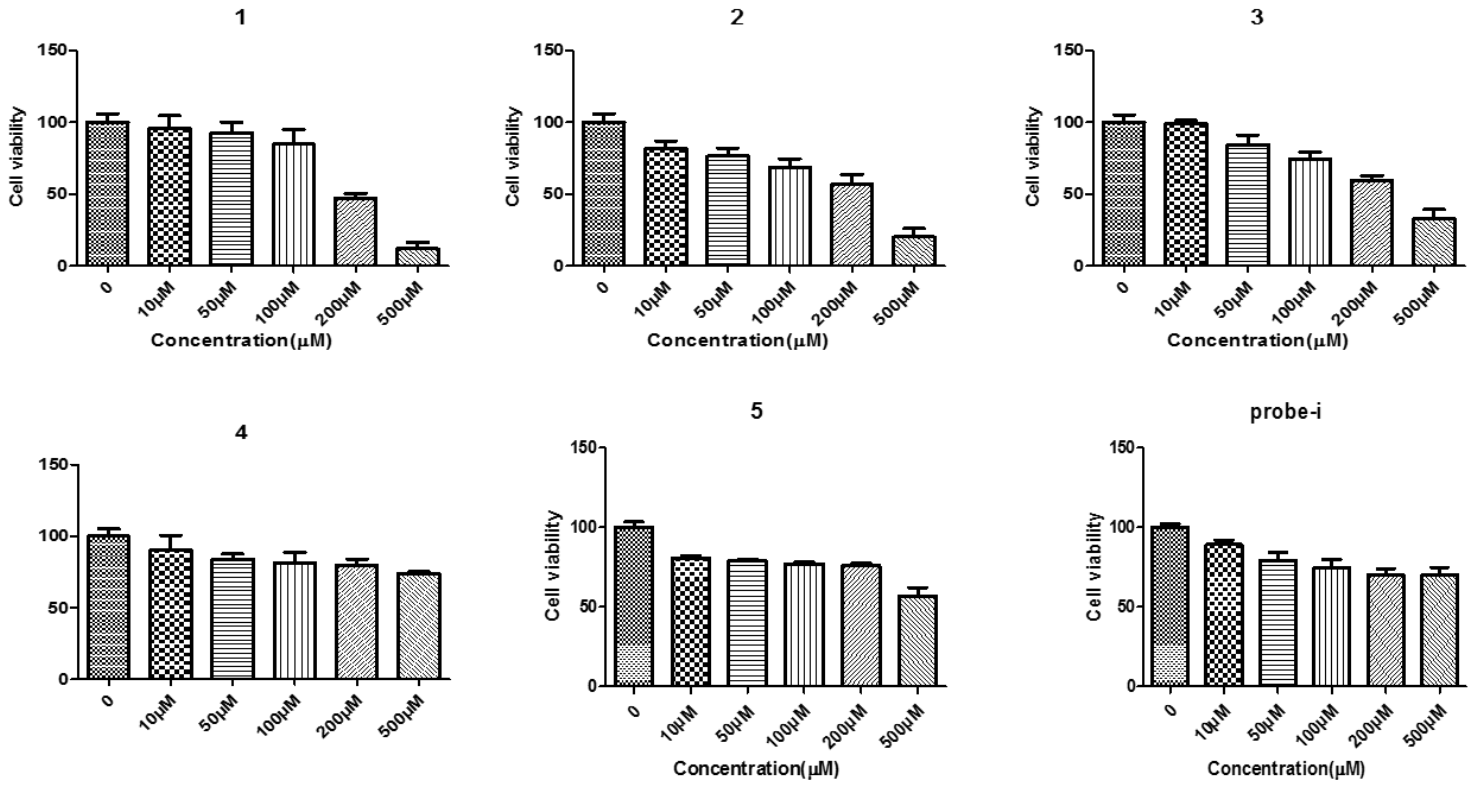
Albiflorin(AL)

- *Albiflorin*, a typical example from natural products.
  - Low potencies or activities
  - Unknown targets
- Direct experiments difficult to discover the mechanisms.
  - Low potencies $\longrightarrow$ false negative
  - Unknown targets $\longrightarrow$ hard to design analogs
  - Complex structures

Method

- Firstly, several analogs are designed based on the chemical experience as a starting point.
- Then MTT assays are used to test their biological activities on tumor growth.
- Next structure-activity relationship (SAR) analysis is performed to predicted its possible functional mechanism.
- Simulation and Filtering
  - Computational simulation of all the possible analogies.
  - Quantitatively predicting their targets.
- Experimental design and validation.

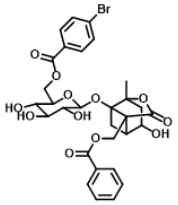
MTT Assay

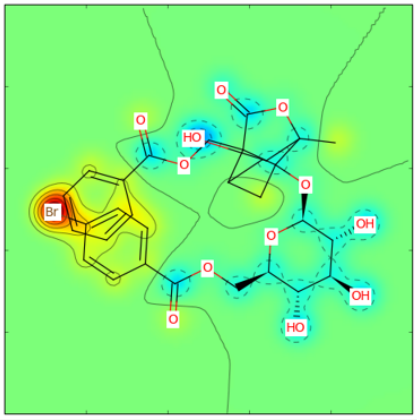


Drug	1	2	3	4	5	probe-i
HCT116(ic50μM)	205.6	184.9	269.5	16887	2811	3256

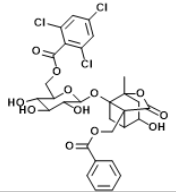
LogP Analysis

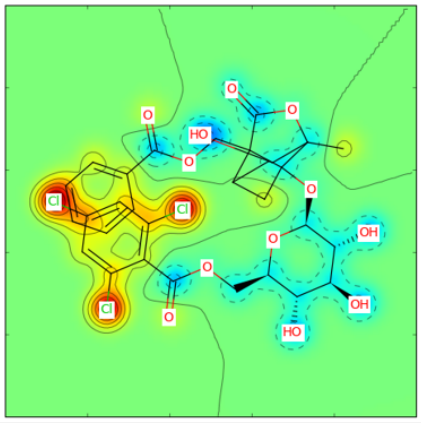
AL-1
logP: 1.11



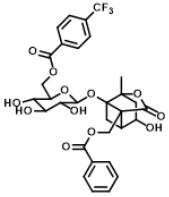


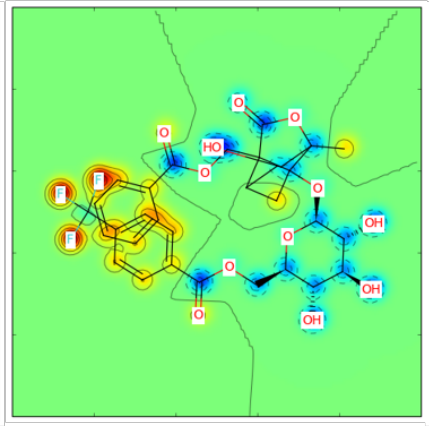
AL-2
logP: 2.31



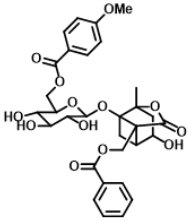


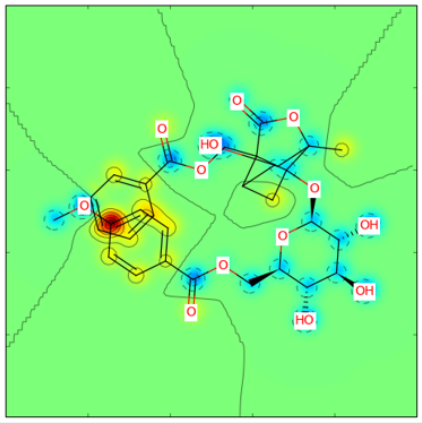
AL-3
logP: 1.37



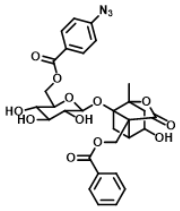


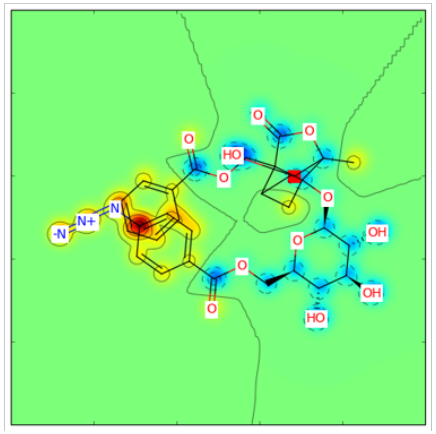
AL-4
logP: 0.36



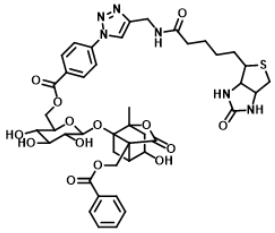


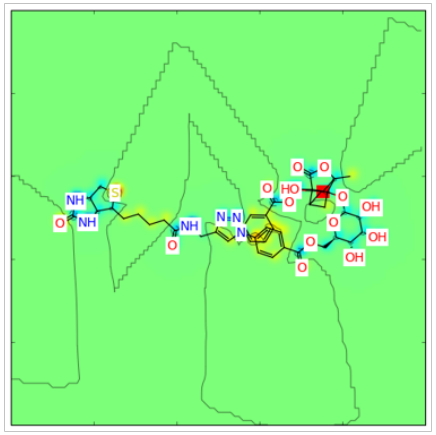
AL-5
logP: 1.29





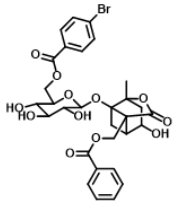
AL-Probe
logP: 0.27

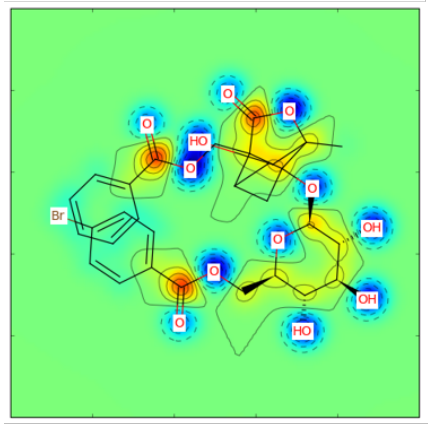




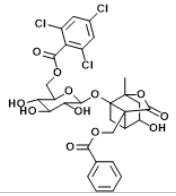
Partial Charge

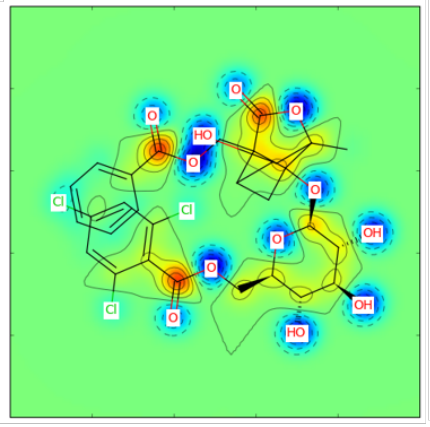
AL-1



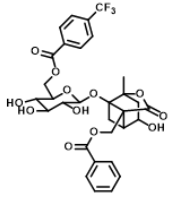


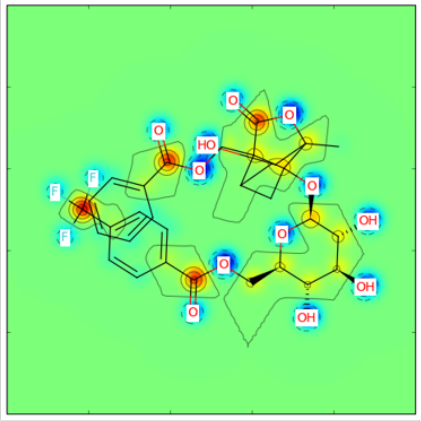
AL-2



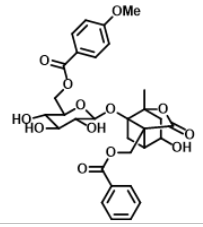


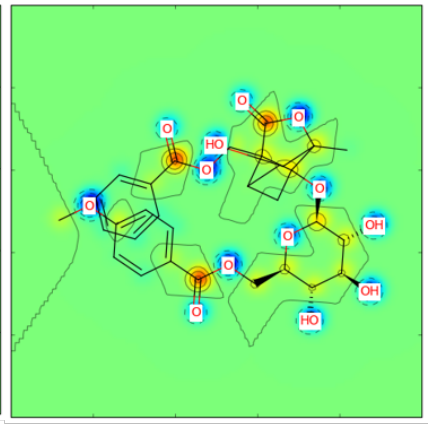
AL-3



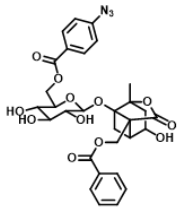


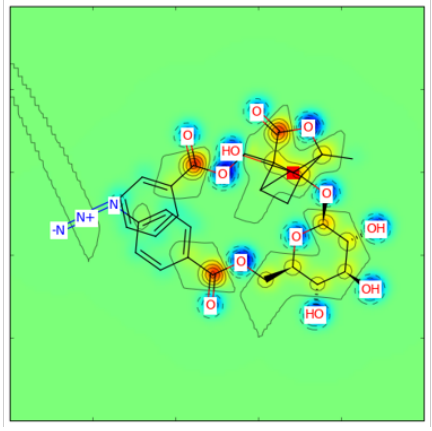
AL-4



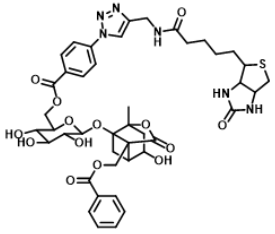


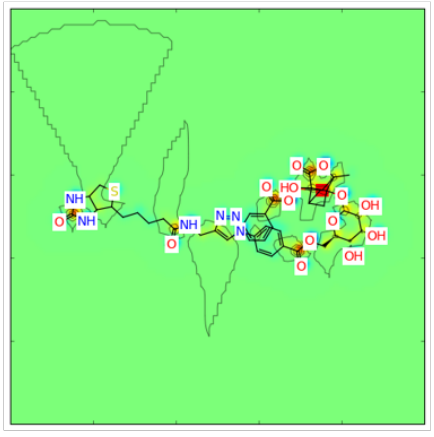
AL-5





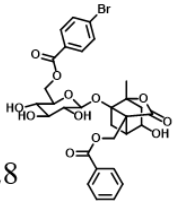
AL-Probe

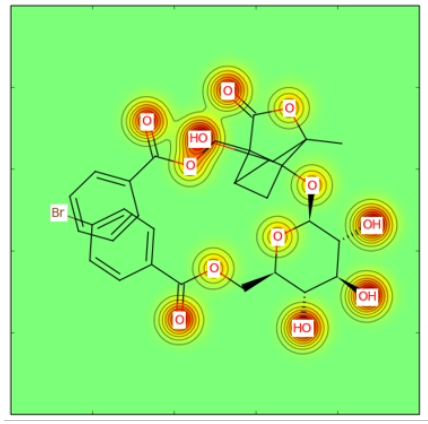




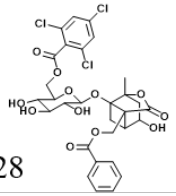
TPSA

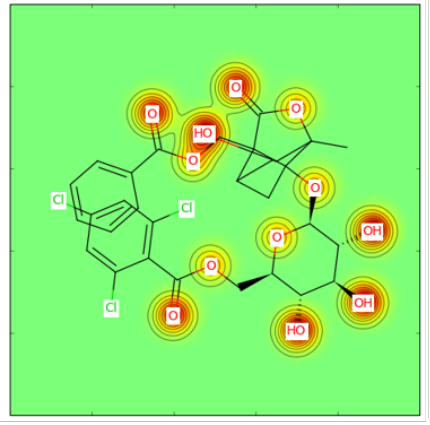
AL-1
TPSA = 178.28



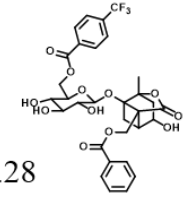


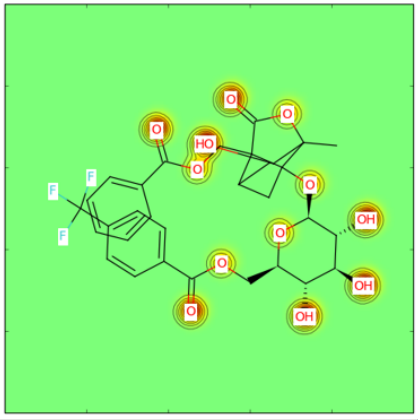
AL-2
TPSA=178.28



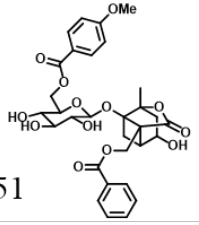


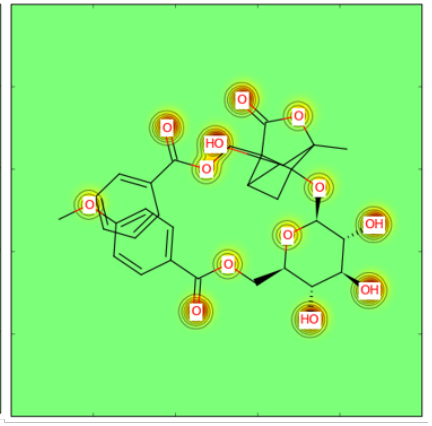
AL-3
TPSA=178.28



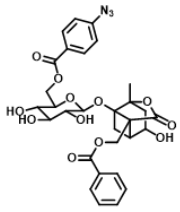


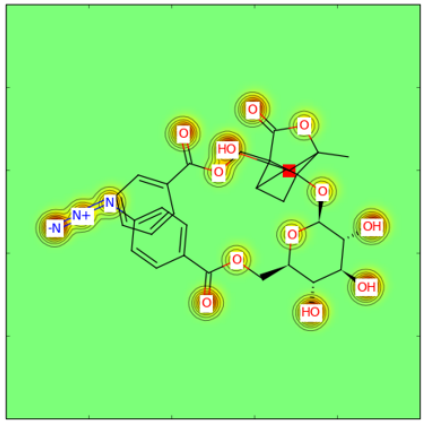
AL-4
TPSA=187.51



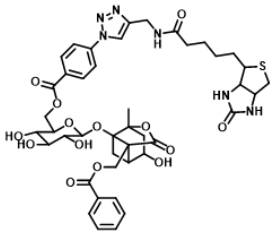


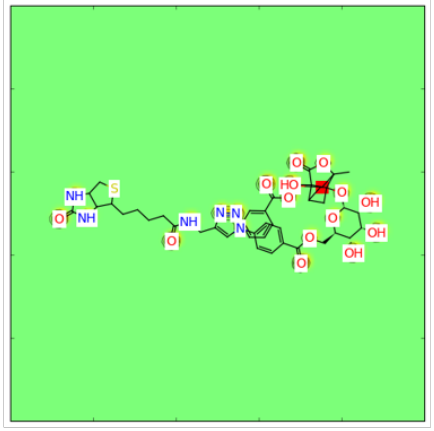
AL-5
TPSA=227.04





AL-Probe
TPSA=279.22





Discussion

Explore a new strategy to study natural products.

- Discovery by computational methods.
- Biological experiments validation.
- Computational Simulation and analysis all the possible analogs.
- Medicinal chemistry-based experiments validation.