Qualitatively Predicting Compound-Protein Interactions by Multi-Task Learning

PhD Candidate: Songpeng Zu
Advisor: Shao Li

FIT 1-108, Tsinghua University

January 26, 2015

Outline

- Quantitatively Predicting Compound-Protein Interactions by Multi-Task Learning
- Inference on Chemogenomic Features from Drug-Target Interactions
- Application on the Modification of Natural Products.

Part I

Quantitatively Predicting Compound-Protein Interactions by Multi-Task Learning

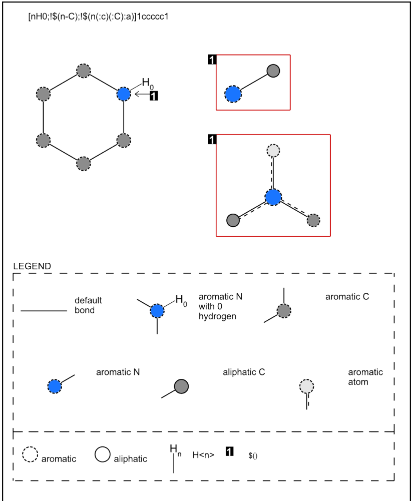
Background

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.



Picture created by SMARTSviewer. Copyright: ZBH Center for Bioinformatics Hamburg [www.zbh.uni-hamburg.de].

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

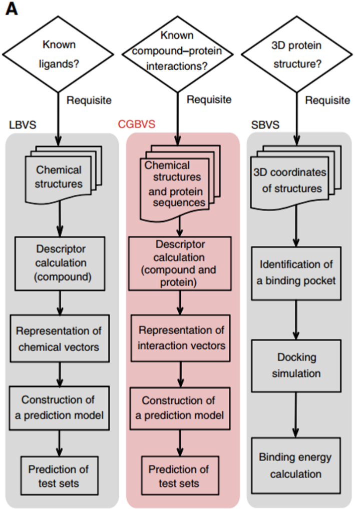
Besnard J. *et al.*, *Nature* 2012, used naive Bayesian model to predict compounds' polypharmacology profiling.

- 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.


A
Known ligands?
Requisite
LBVS
Chemical structures
Descriptor calculation (compound)
Representation of chemical vectors
Construction of a prediction model
Prediction of test sets
Known compound–protein interactions?
Requisite
CGBVS
Chemical structures and protein sequences
Descriptor calculation (compound and protein)
Representation of interaction vectors
Construction of a prediction model
Prediction of test sets
3D protein structure?
Requisite
SBVS
3D coordinates of structures
Identification of a binding pocket
Docking simulation
Binding energy calculation

Machine Learning on CPIs

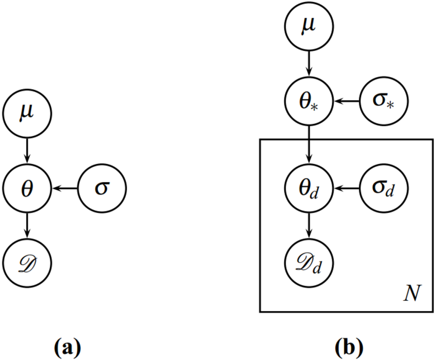
Current machine learning on predicting CPIs

- 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.



Hierarchical Bayesian Model

Suppose $\mathcal{D}_j = \left\{\mathbf{X}_j, \mathbf{y}_j\right\}, j = 1, ..., m$, 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 $\Theta = \{\omega_j, j = 1, ..., m, \omega_*, \sigma_y^2\}$. We have

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

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

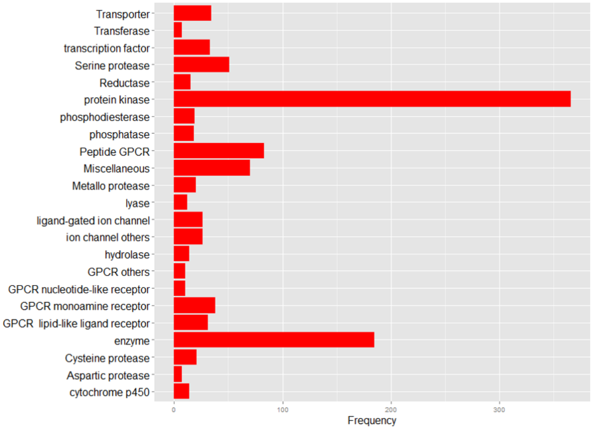
$$\begin{aligned}\frac{\partial \mathcal{L}_{hier}(\mathcal{D};\Theta)}{\partial \omega_j} &= -\frac{1}{2\sigma_y^2}\frac{\parallel \mathbf{y}_j - \mathbf{X}_j^T\omega_j \parallel^2}{\partial \omega_j} - \frac{1}{2\sigma_j^2}\frac{\parallel \omega_j - \omega_* \parallel^2}{\partial \omega_j} \\ &= \frac{\mathbf{X}_j\mathbf{y}_j}{\sigma_y^2}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

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

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

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.

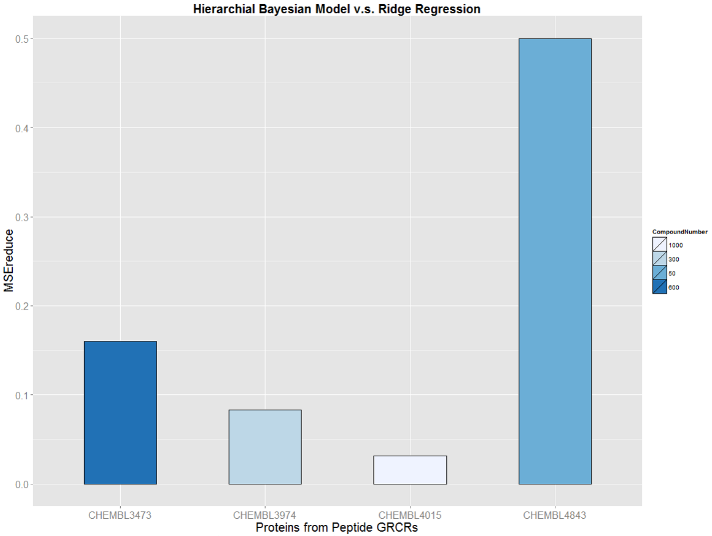


Feature Selection

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



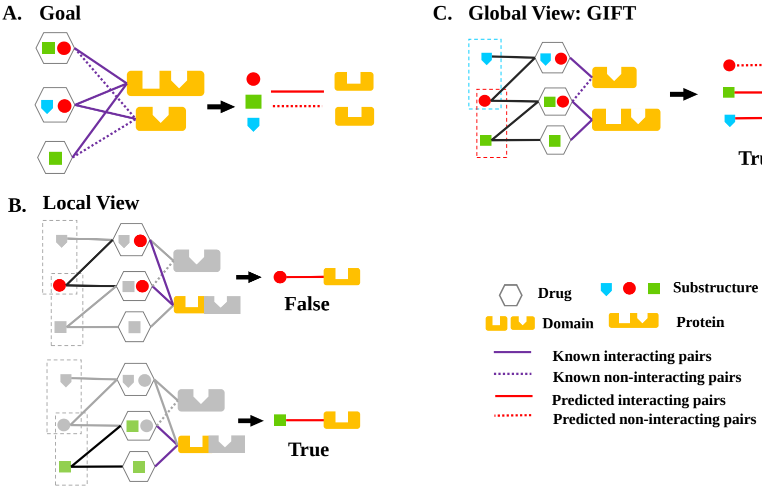
Discussion

- More computational tests
- The relationship between proteins' pharmacological and genomic information
- Deficiency:
  - High dimension v.s. Sparsity
  - Linear v.s. Nonlinear

Part II

Inference on Chemogenomic Features from Drug-Target Interactions

Background



A. Goal

C. Global View: GIFT

True

B. Local View

False

True

⬡ Drug

● ■ Substructure

Domain

Protein

Known interacting pairs

Known non-interacting pairs

Predicted interacting pairs

Predicted non-interacting pairs

Definition

O_{ij} The observations of drug and protein interactions.

YP_{ij} The binary variable of drug i and protein j interactions.

$D_{mn}^{(ij)}$ The interaction result of substructure m from drug i and domain n from protein j

θ_{mn} $\theta_{mn} = Pr(ZD_{mn} = 1)$

fn $\mathrm{fn} = Pr(O_{ij} = 0 | YP_{ij} = 1)$

fp $\mathrm{fp} = Pr(O_{ij} = 1 | YP_{ij} = 0)$

Assumption

- **Consistency**

$$\theta_{mn} = Pr(D_{mn}^{(ij)} = 1) \tag{8}$$

- **Independence**

$$Pr(YP_{ij} = 1|\theta) = 1 - \prod_{D_{mn}^{(ij)}} (1 - \theta_{mn}) \tag{9}$$

EM Algorithm

- Then the log likelihood function is followed:

$$l(\theta) = log\left(Pr(O|\theta)\right) \tag{10}$$

$$Pr(O_{ij} = 1|\theta) = (1 - fn)Pr(YP_{ij} = 1|\theta) + fp \cdot Pr(YP_{ij} = 0|\theta) \tag{11}$$

- The EM Algorithm is used to get the MLE estimation due to the missing data of D_{mn}.

EM Algorithm

- E Step:

$$E(D_{mn}^{(ij)}|O,\theta^{(t-1)}) = \frac{\theta_{mn}^{(t-1)}(1-fn)^{O_{ij}}fn^{1-O_{ij}}}{Pr(O_{ij}|\theta^{(t-1)})} \tag{12}$$

- M Step:

$$\theta_{mn}^{(t)} = \frac{1}{N_{mn}} \sum_{i,j:Zm\in Y_i,Dn\in P_j} E(D_{mn}^{(ij)}|O_{ij},\theta^{(t-1)}) \tag{13}$$

Variance Estimation

- The variance of the parameters are estimated by the observed Fisher information.

$$var(\hat{\theta}) = \frac{1}{I(\hat{\theta})}, I(\theta) = -\frac{d^2 log(Pr(O|\theta))}{d\theta^2} \tag{14}$$

- In our model, the observed Fisher information is followed:

$$I(\theta_{mn}) = \sum_{i,j:Zm \in Y_i, Dn \in P_j} \delta_{mn}^{(i,j)^2} \left(\frac{O_{mn}^{(ij)}}{\mu_{mn}^{(ij)2}} + \frac{1 - O_{mn}^{(ij)}}{(1-\mu_{mn}^{(ij)})^2} \right) \tag{15}$$

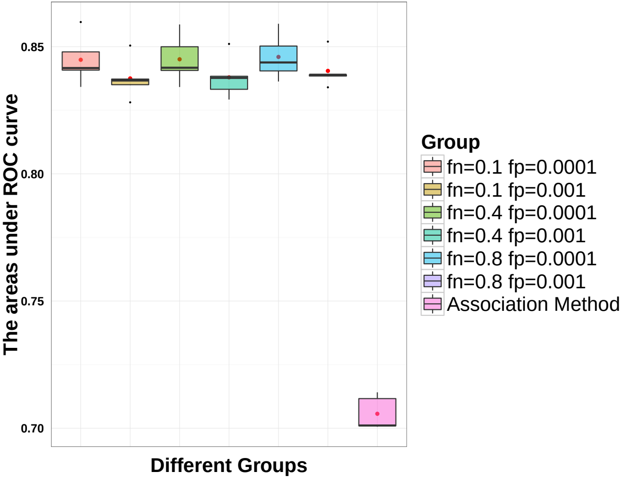
In which,

$$\delta_{mn}^{(ij)} = \frac{\mu_{mn}^{(ij)}}{\partial \theta_{mn}}, \mu_{mn}^{(ij)} = Pr(O_{mn}^{(ij)} = 1|\theta) \tag{16}$$

Data Source

- 1862 drugs are represented by 881-dimensional chemical substructure binary vectors defined by the PubChem database.
- 1554 proteins are represented by 876-dimensional protein domain binary vectors from the Pfam database.
- 4809 interactions between drugs and proteins.

- Different combinations of fn and fp.



- Comparison with other methods.

Ratio	GIFT	L1-Log	L1-SVM	SCCA
1	0.835	0.829	0.830	0.798
5	0.847	0.838	0.855	0.798

- Results of predictions on known drug-domain interactions.

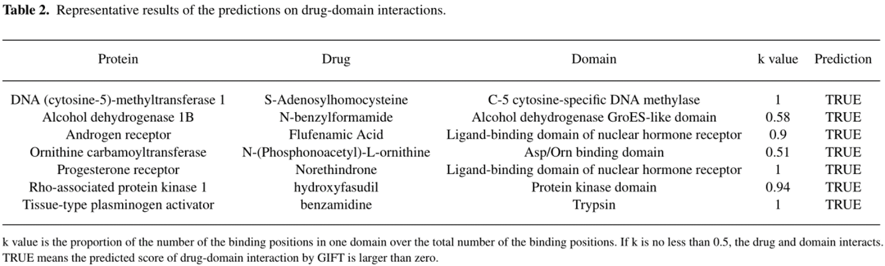
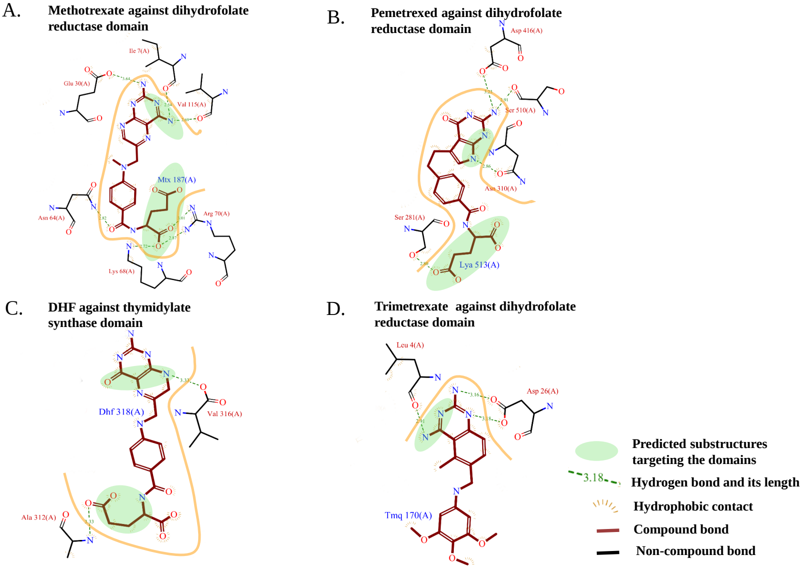


Table 2. Representative results of the predictions on drug-domain interactions.

Protein	Drug	Domain	k value	Prediction
DNA (cytosine-5)-methyltransferase 1	S-Adenosylhomocysteine	C-5 cytosine-specific DNA methylase	1	TRUE
Alcohol dehydrogenase 1B	N-benzylformamide	Alcohol dehydrogenase GroES-like domain	0.58	TRUE
Androgen receptor	Flufenamic Acid	Ligand-binding domain of nuclear hormone receptor	0.9	TRUE
Ornithine carbamoyltransferase	N-(Phosphonoacetyl)-L-ornithine	Asp/Orn binding domain	0.51	TRUE
Progesterone receptor	Norethindrone	Ligand-binding domain of nuclear hormone receptor	1	TRUE
Rho-associated protein kinase 1	hydroxyfasudil	Protein kinase domain	0.94	TRUE
Tissue-type plasminogen activator	benzamidine	Trypsin	1	TRUE

k value is the proportion of the number of the binding positions in one domain over the total number of the binding positions. If k is no less than 0.5, the drug and domain interacts. TRUE means the predicted score of drug-domain interaction by GIFT is larger than zero.



A. Methotrexate against dihydrofolate reductase domain

B. Pemetrexed against dihydrofolate reductase domain

C. DHF against thymidylate synthase domain

D. Trimetrexate against dihydrofolate reductase domain

🟩 Predicted substructures targeting the domains

3.18 Hydrogen bond and its length

Hydrophobic contact

Compound bond

Non-compound bond

Discussion

- Here we propose an efficient method to extract meaningful chemogenomic features, and it also shows the power to predict drug-protein interactions.
- The predicted chemical substructures might be a useful source to design the compounds' analogs against a given protein or its domain.
- Large-scale compound-protein interactions are accumulated in the PDB database (known 3D structures), BindDB and ChEMBL database, which can be further studied by our method.

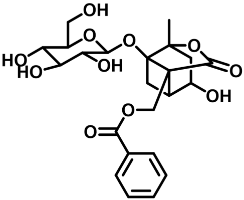
Part III

Application on the Modification of Natural Products

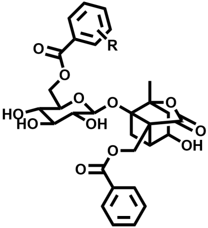
Lead Discovery From TCM Herbs

- 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)

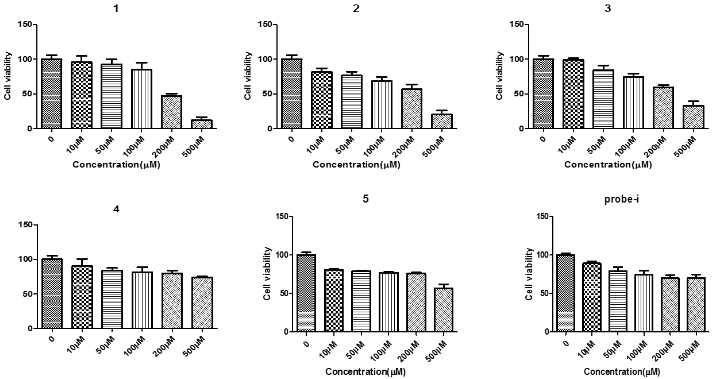


- *Albiflorin*, a typical example from natural products.
  - Low potencies or activities
  - Unknown targets
- Direct experiments difficult to discover the mechanisms.
  - Low potencies $\longrightarrow$ false negative
  - Unknown targets $\longrightarrow$ hard to design analogs
  - Complex structures

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.

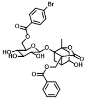
MTT Assay

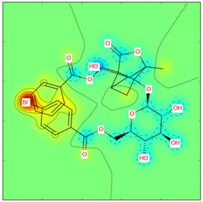


Drug	1	2	3	4	5	probe-i
HCT116(ic50μM)	205.6	184.9	269.5	16887	2811	3256

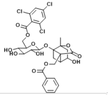
LogP Analysis

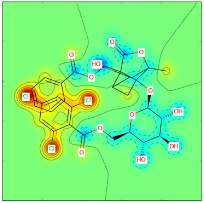
AL-1
logP: 1.11



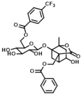


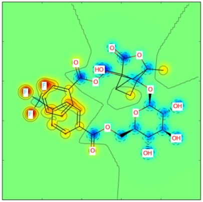
AL-2
logP: 2.31



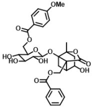


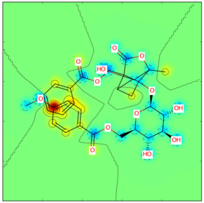
AL-3
logP: 1.37



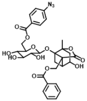


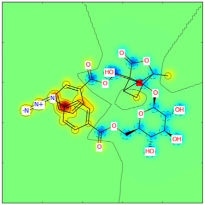
AL-4
logP: 0.36



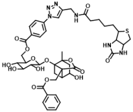


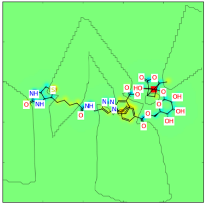
AL-5
logP: 1.29





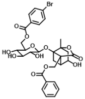
AL-Probe
logP: 0.27

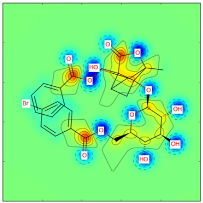




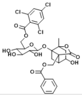
Partial Charge

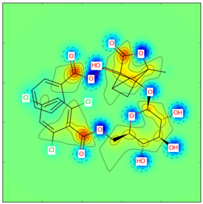
AL-1



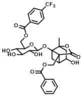


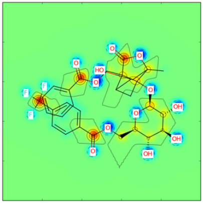
AL-2



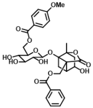


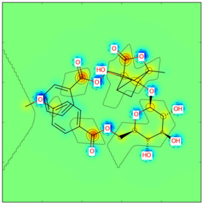
AL-3



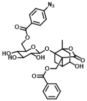


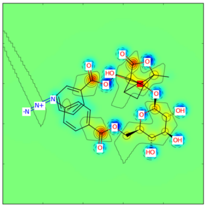
AL-4



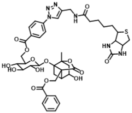


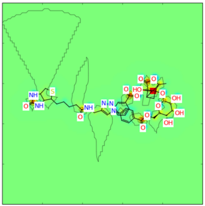
AL-5





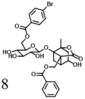
AL-Probe

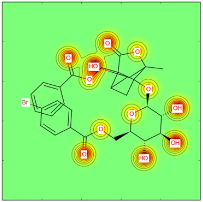




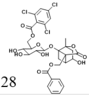
TPSA

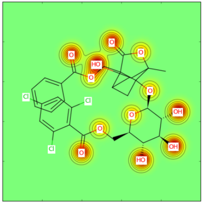
AL-1
TPSA = 178.28



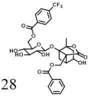


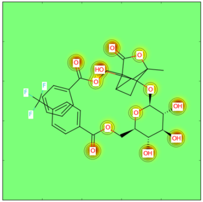
AL-2
TPSA=178.28



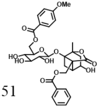


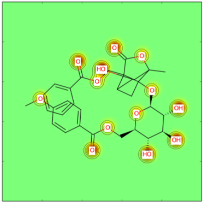
AL-3
TPSA=178.28



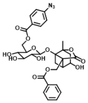


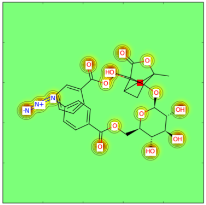
AL-4
TPSA=187.51



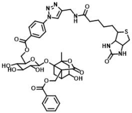


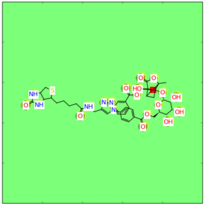
AL-5
TPSA=227.04





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.