人类蛋白质相关临床表型预测算法研究

Research on Prediction of Human Protein-Phenotype Associations

刘砺志

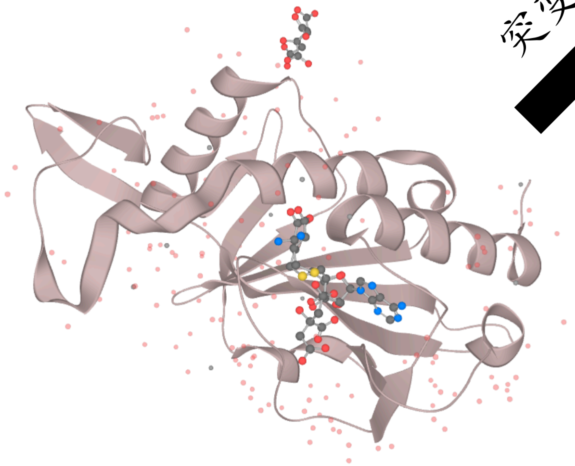
基因、遗传疾病与异常表型

基因

EMG1

表达

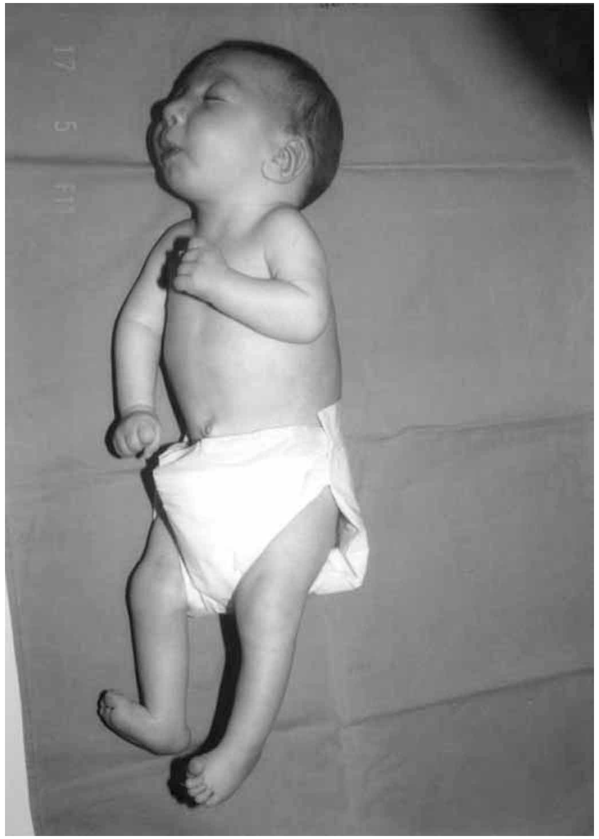
蛋白质



Ribosomal RNA small subunit methyltransferase NEP1

突变

遗传疾病



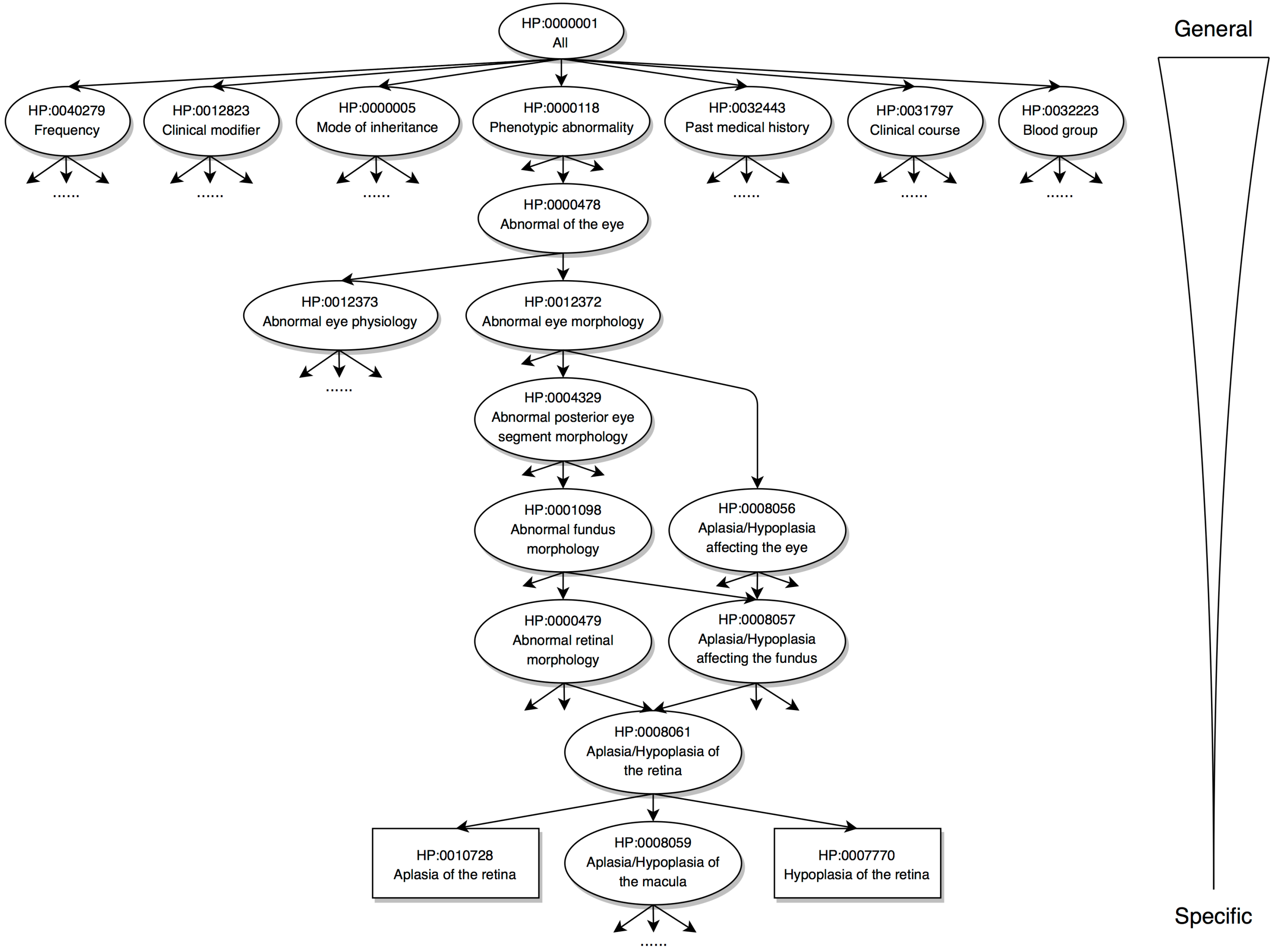
Bowen-conradi Syndrome

症状

异常表型

- □ *Small for gestational age*
- □ *Micrognathia*
- □ *Prominent nose*
- □ *Abnormal joint morphology*
- □ *Clinodactyly of the 5th finger*
- □ *Rocker bottom foot*
- □ *Microcephaly*

人类表型本体



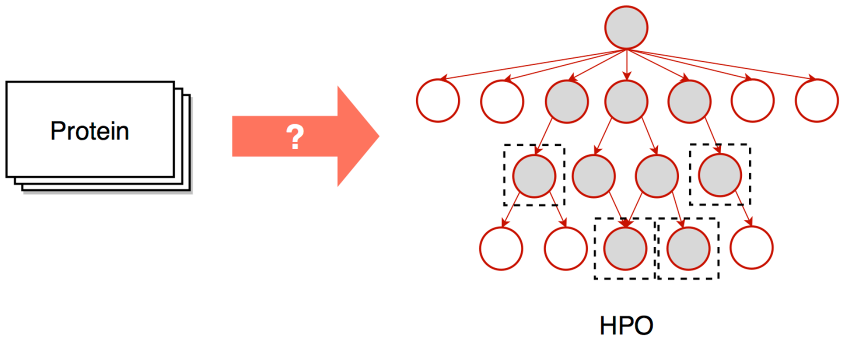
人类表型本体

Human Phenotype Ontology (HPO)

- 一个HPO术语对应一种异常表型，两个术语间的有向边为“is-a”关系
- 层次化结构：有向无环图
- 真路径规则：当基因被某个HPO术语所注释，其也被该术语的所有祖先术语所注释
- HPO包含7个子本体，其中HP:0000118 表型异常 (*Phenotypic abnormality*) 是最核心的子本体

研究内容：预测蛋白质-HPO术语关系

HPOLabeler



A

以蛋白质为中心 (protein-centric)

确定新蛋白质（或完全未被标注的蛋白质）的全部HPO注释

层次化多标签分类问题

HPOFiller

	Term 1	Term 2	Term 3	......	Term *n*
Protein 1	1	?	?	......	1
Protein 2	?	1	?	......	?
Protein 3	1	?	1	......	1
⋮	⋮	⋮	⋮		⋮
Protein *m*	?	1	?	......	1

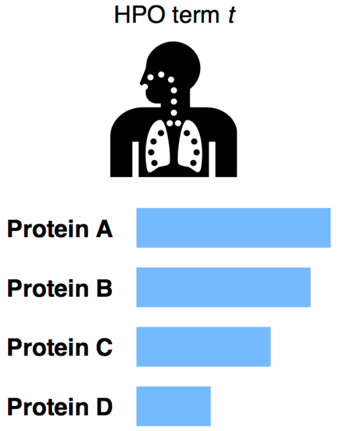
B

逐对预测 (pairwise)

识别缺失的蛋白质-HPO术语关系

矩阵填充

HPODNets



C

以HPO术语为中心 (term-centric)

对与某个HPO术语相关的候选蛋白质进行优选排序

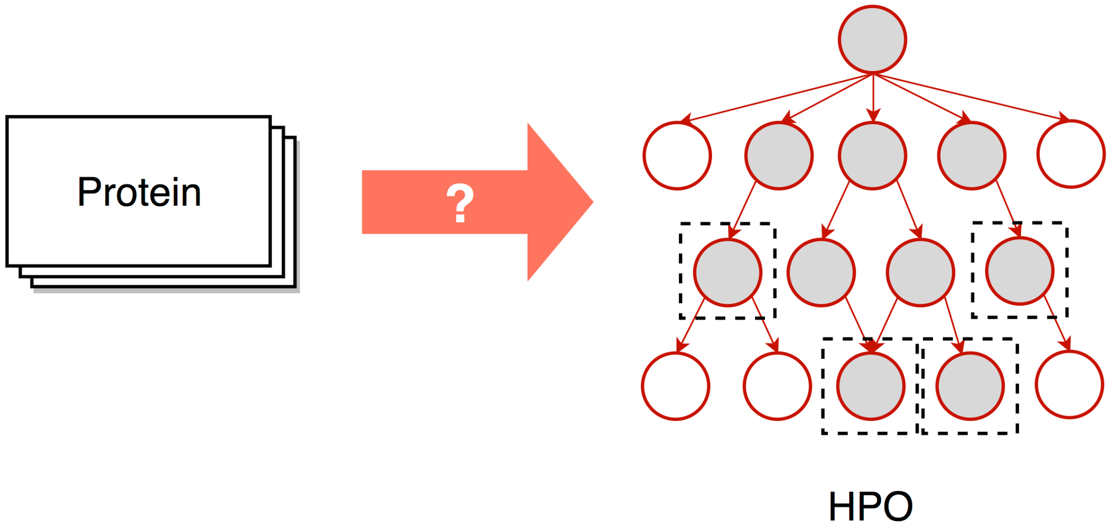
二元分类问题

1

HPOLabeler

基于排序学习的蛋白质表型标注预测算法

问题描述：预测人类蛋白质的HPO注释

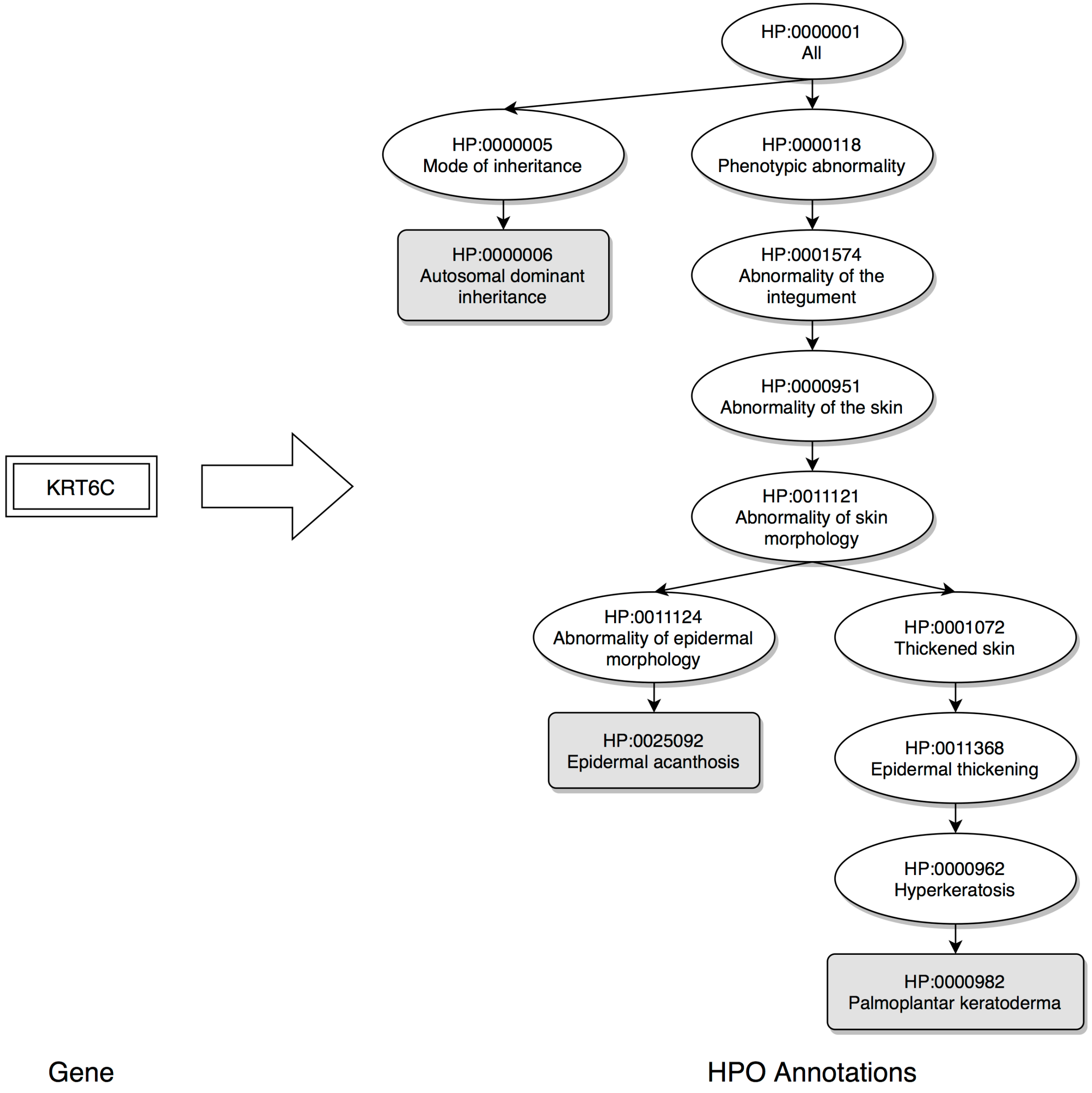

Protein

?

HPO

预测人类蛋白质的HPO标注问题

<span>研究目标：</span>利用机器学习技术，整合多种信息源，提高预测性能


KRT6C

HP:0000001
All

HP:0000005
Mode of inheritance

HP:0000006
Autosomal dominant inheritance

HP:0000118
Phenotypic abnormality

HP:0001574
Abnormality of the integument

HP:0000951
Abnormality of the skin

HP:0011121
Abnormality of skin morphology

HP:0011124
Abnormality of epidermal morphology

HP:0025092
Epidermal acanthosis

HP:0001072
Thickened skin

HP:0011368
Epidermal thickening

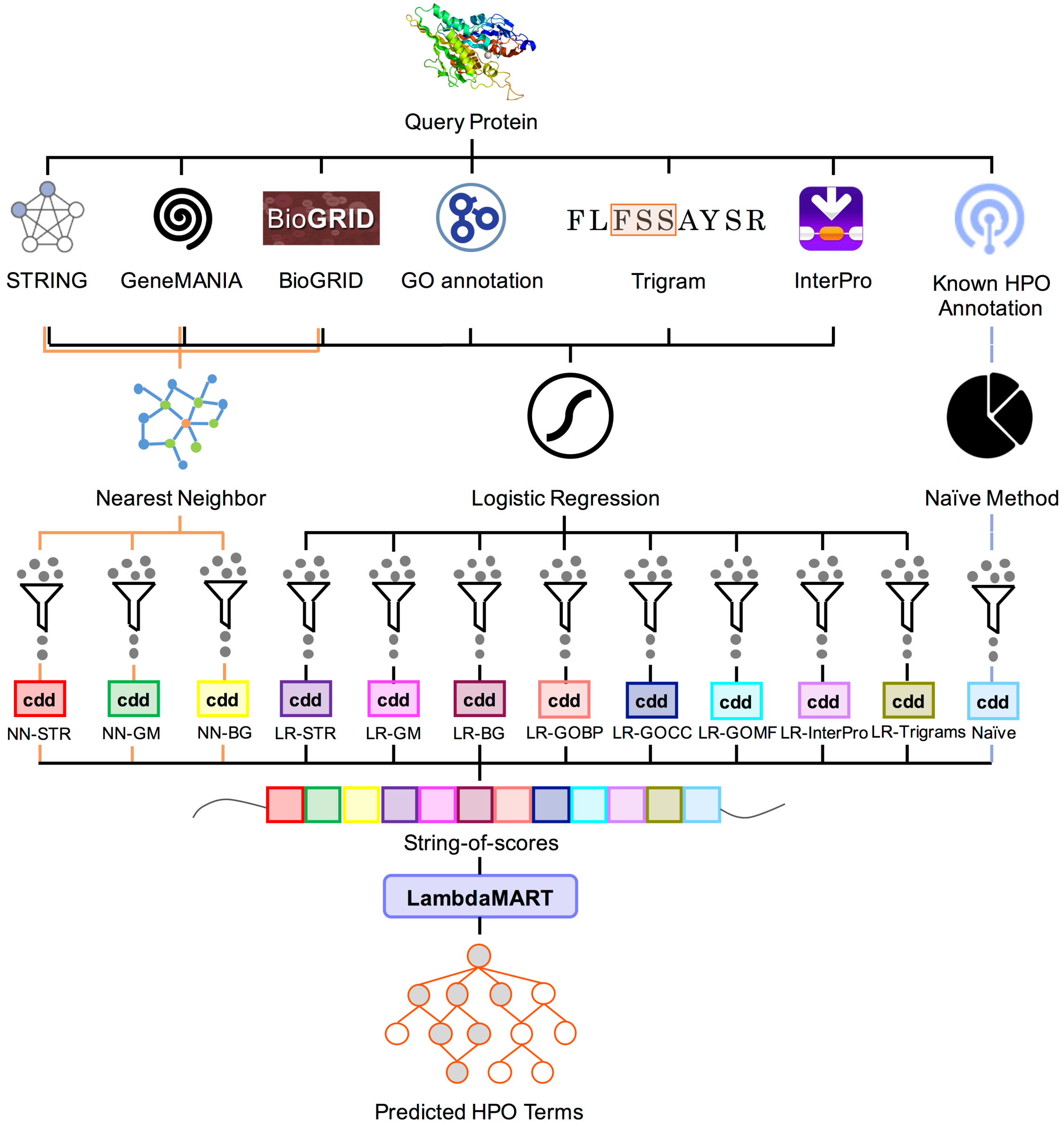
HP:0000962
Hyperkeratosis

HP:0000982
Palmoplantar keratoderma

Gene

HPO Annotations

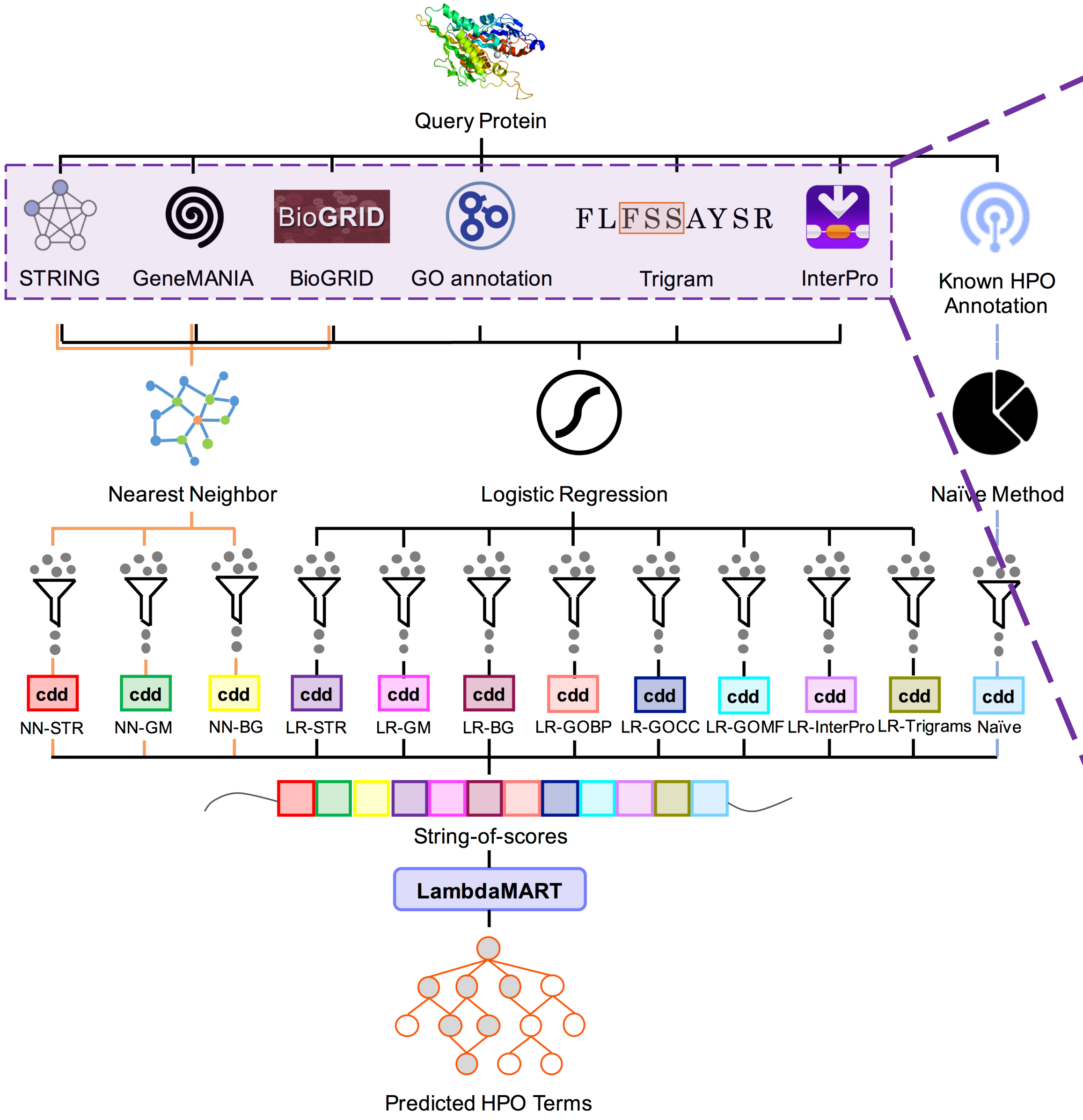
HPOLabeler — 使用排序学习提升预测性能



关键点

- 集成学习：Stacking思想
- 排序学习整合基础模型以进一步提升预测性能
- 在时序验证中唯一一个优于朴素方法的模型

特征抽取



STRING

$$\mathbf{x}_i^{(\mathrm{STR})} = \left(x_{i,1}^{(\mathrm{STR})}, x_{i,2}^{(\mathrm{STR})}, \cdots, x_{i,n^{(\mathrm{STR})}}^{(\mathrm{STR})}\right)^T \quad (1)$$

GeneMANIA

$$\mathbf{x}_i^{(\mathrm{GM})} = \left(x_{i,1}^{(\mathrm{GM})}, x_{i,2}^{(\mathrm{GM})}, \cdots, x_{i,n^{(\mathrm{GM})}}^{(\mathrm{GM})}\right)^T \quad (2)$$

BioGRID

$$\mathbf{x}_i^{(\mathrm{BGD})} = \left(x_{i,1}^{(\mathrm{BGD})}, x_{i,2}^{(\mathrm{BGD})}, \cdots, x_{i,n^{(\mathrm{BGD})}}^{(\mathrm{BGD})}\right)^T \quad (3)$$

GO BP/CC/MF

$$\mathbf{x}_i^{(\mathrm{GOXX})} = \left(x_{i,1}^{(\mathrm{GOXX})}, x_{i,2}^{(\mathrm{GOXX})}, \cdots, x_{i,n^{(\mathrm{GOXX})}}^{(\mathrm{GOXX})}\right)^T \quad (4)$$

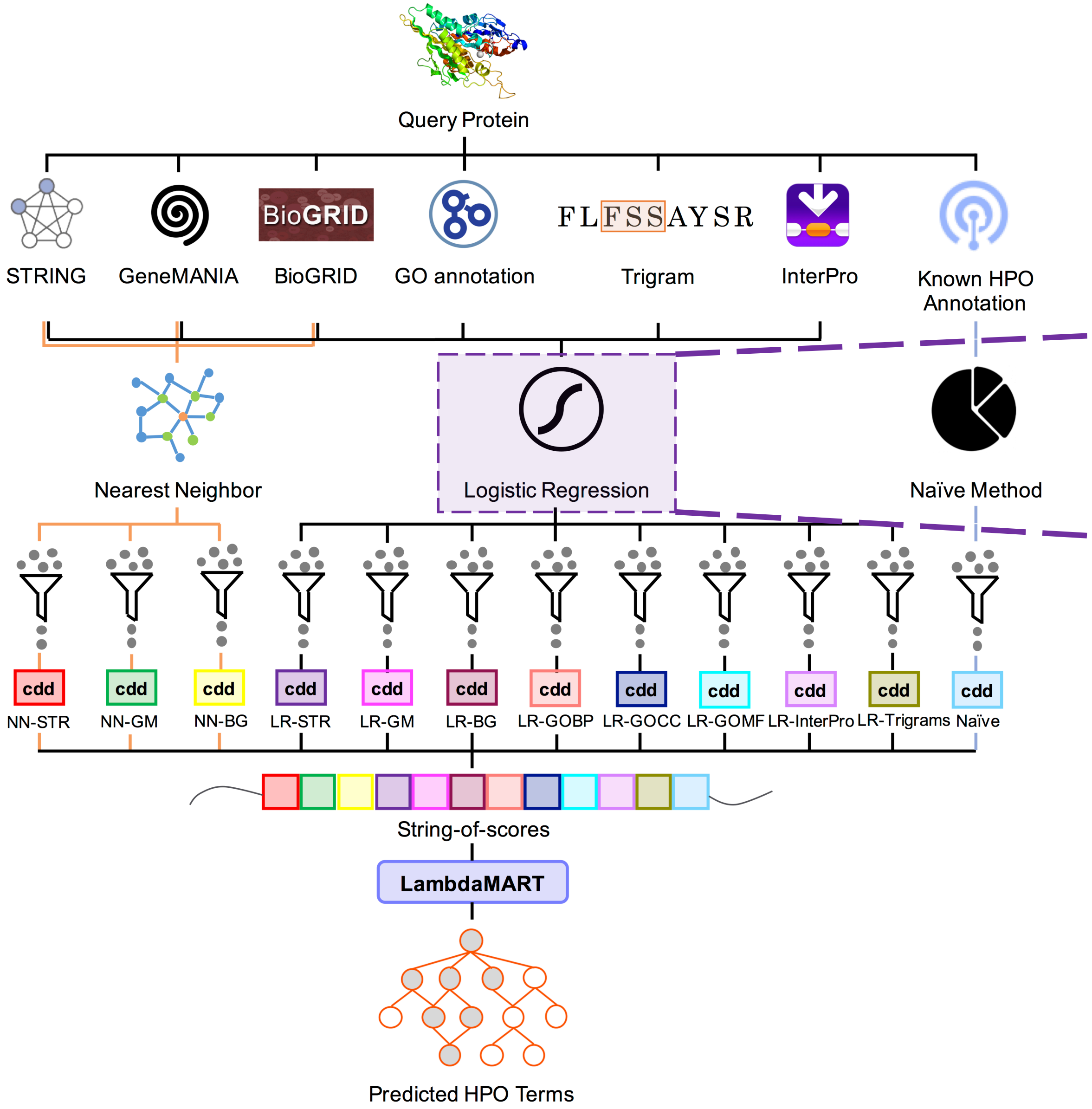
InterPro signatures

$$\mathbf{x}_i^{(\mathrm{IPR})} = \left(x_{i,1}^{(\mathrm{IPR})}, x_{i,2}^{(\mathrm{IPR})}, \cdots, x_{i,n^{(\mathrm{IPR})}}^{(\mathrm{IPR})}\right)^T \quad (5)$$

Trigrams

$$\mathbf{x}_i^{(\mathrm{TRI})} = \left(x_{i,1}^{(\mathrm{TRI})}, x_{i,2}^{(\mathrm{TRI})}, \cdots, x_{i,n^{(\mathrm{TRI})}}^{(\mathrm{TRI})}\right)^T \quad (6)$$

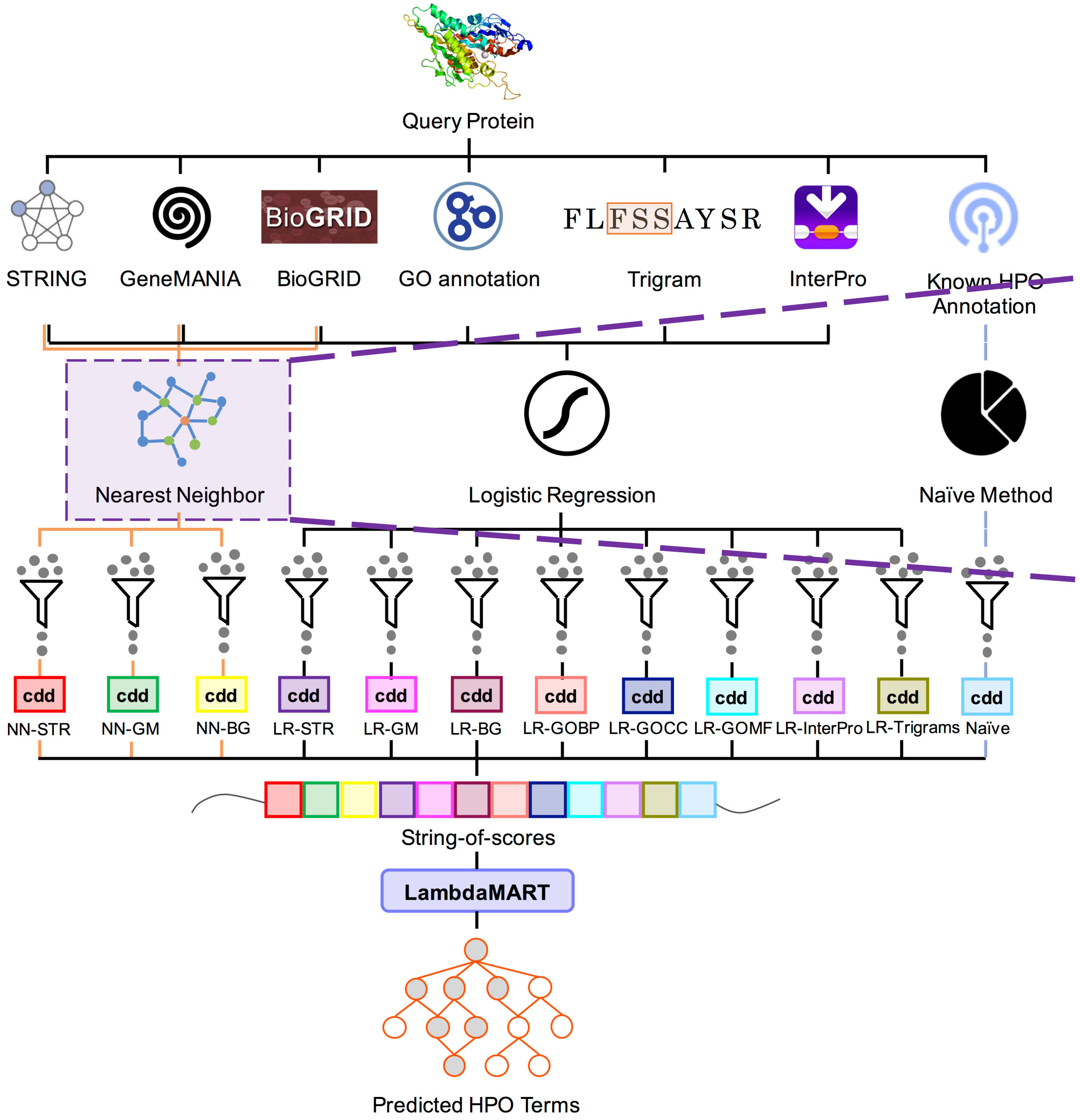
基础模型 — 逻辑斯蒂回归



LR model for each HPO term

$$S^{(f)}(p_i, t) = \mathcal{L}_t^{(f)}(\mathbf{x}_i^{(f)}) = P\left(y_{i,t} = 1|\mathbf{x}_i^{(f)}\right) \tag{7}$$

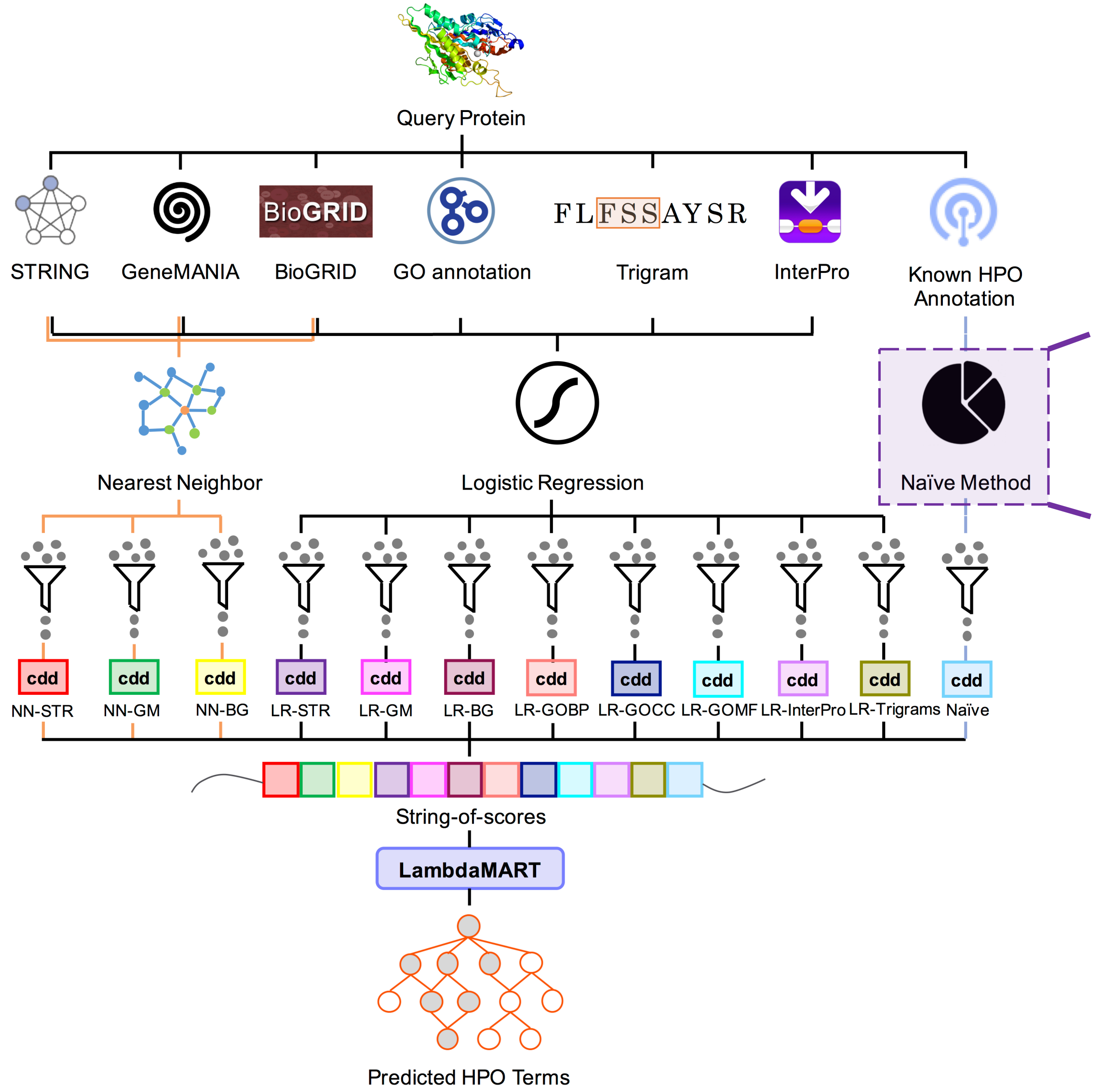
基础模型 — 近邻方法



Nearest Neighbor on STRING, GeneMANIA and BioGRID

$$S^{(\text{NBR-G})}(p_i, t) = \frac{\sum_{p_j \in N_G(p_i)} d(p_i, p_j) \cdot y_{j,t}}{\sum_{p_j \in N_G(p_i)} d(p_i, p_j)} \qquad (8)$$

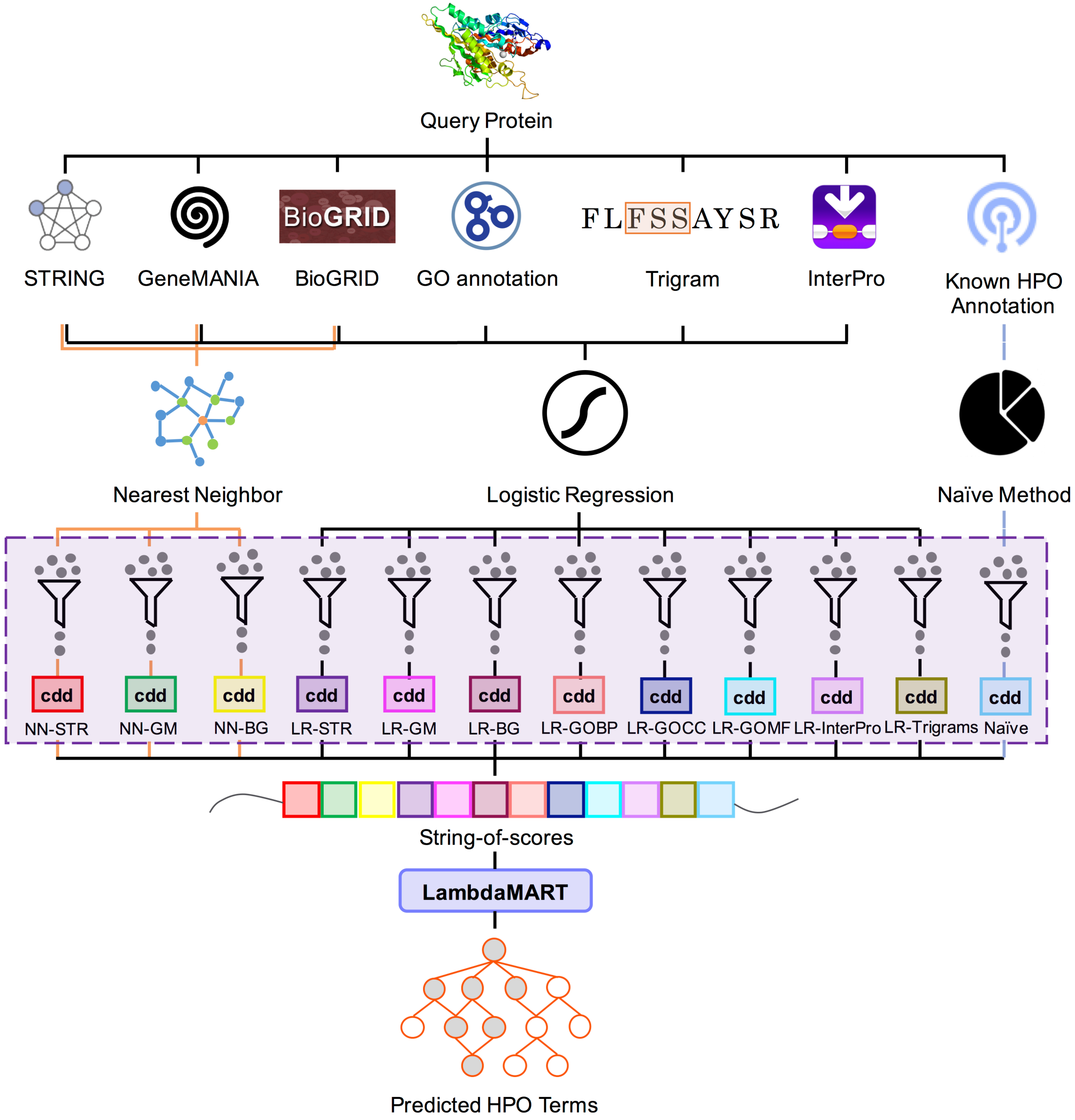
基础模型 — 朴素方法



Naïve

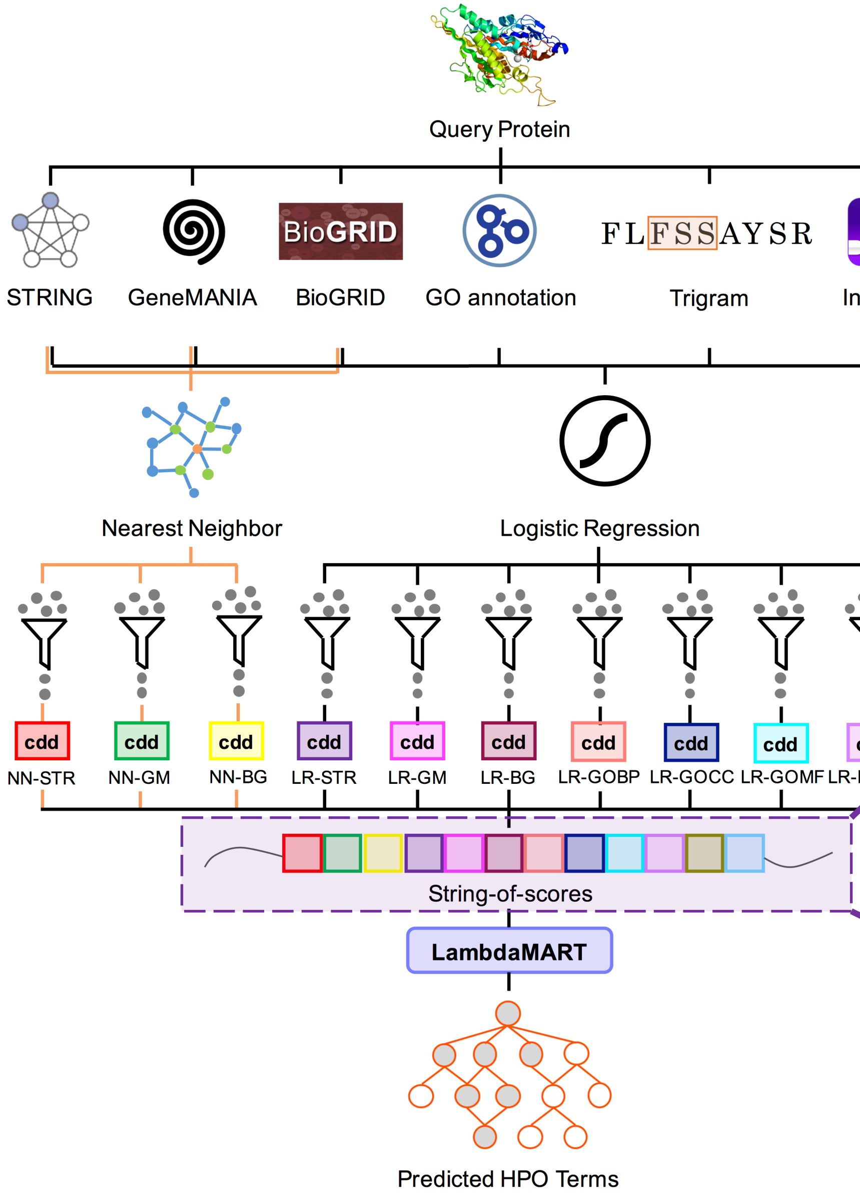
$$S^{(\text{Naïve})}(p_i, t) = \frac{|\{p_j \in \mathcal{P}_S | y_{j,t} = 1\}|}{m_S} \qquad (9)$$

HPOLabeler — 第一步：候选集产生



- 各基础模型预测结果上的前*k*个HPO术语被挑选出来
- 取这些子集的并集作为最终的候选集

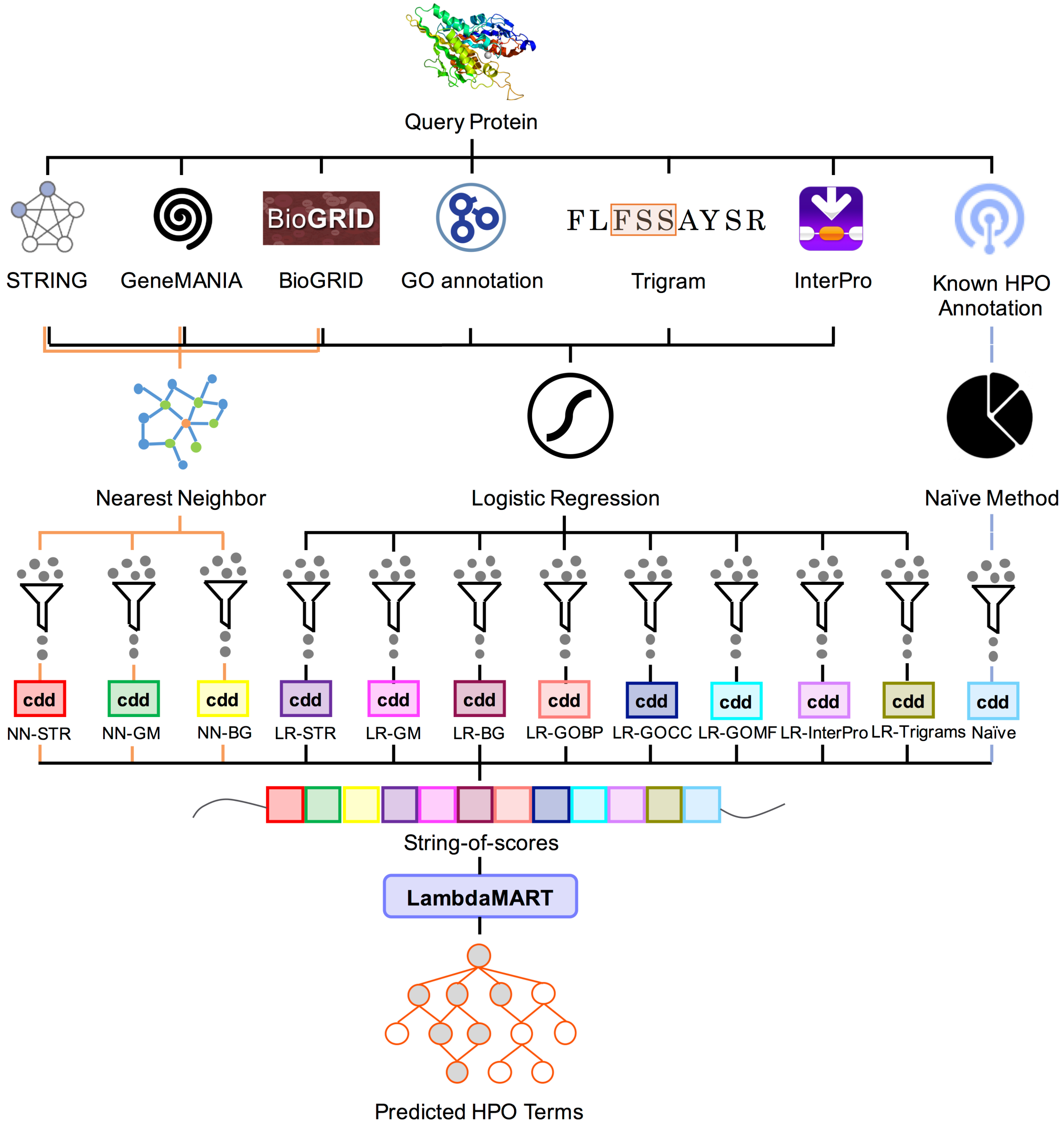
HPOLabeler — 第二步：为排序学习生成特征



String-of-scores

$$\mathbf{x}_t^{(\text{L2R})} = \begin{pmatrix} S^{(\text{STR})}(p,t) \\ S^{(\text{GM})}(p,t) \\ S^{(\text{BGD})}(p,t) \\ S^{(\text{GOBP})}(p,t) \\ S^{(\text{GOCC})}(p,t) \\ S^{(\text{GOMF})}(p,t) \\ S^{(\text{IPR})}(p,t) \\ S^{(\text{TRI})}(p,t) \\ S^{(\text{NBR-STR})}(p,t) \\ S^{(\text{NBR-GM})}(p,t) \\ S^{(\text{NBR-BGD})}(p,t) \\ S^{(\text{Naïve})}(p,t) \end{pmatrix} \tag{10}$$

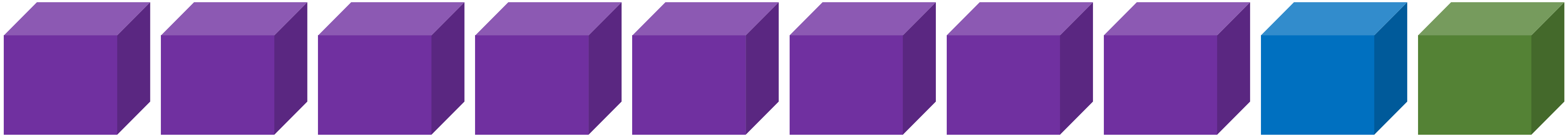
HPOLabeler — 第三步：排序学习



- 基于LambdaMART重排候选HPO术语
- 最终得到一个有序的预测打分列表

评估之一：交叉验证

2018-07-27



3,722 proteins

8,067 HPO terms

Avg. 119.4 annotations

实验结果之交叉验证 — 对比

各基础分类器的性能

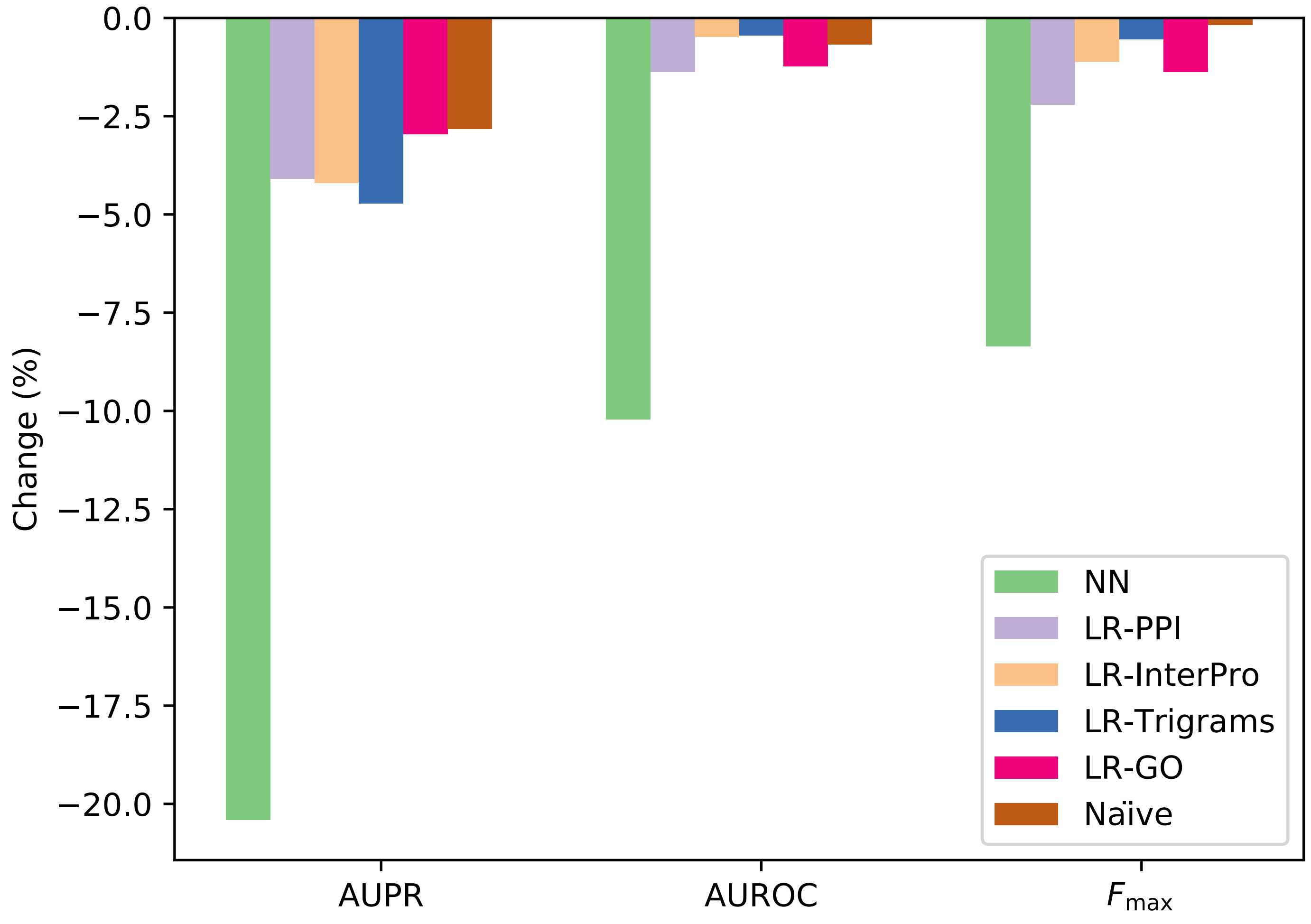
Component	F_{max}	AUC	AUPR
LR-STRING	0.4174	0.6390	0.2697
LR-GeneMANIA	0.3506	0.7282	0.2605
LR-BioGRID	0.3441	0.5941	0.2677
LR-GO BP	0.3777	0.6741	0.2926
LR-GO CC	0.3643	0.6544	0.2916
LR-GO MF	0.3343	0.6081	0.2403
LR-InterPro	0.3588	0.6041	0.2699
LR-Trigrams	0.2941	0.5136	0.1564
NN-STRING	**0.4213**	**0.7892**	**0.3635**
NN-GeneMANIA	0.4110	0.7274	0.3550
NN-BioGRID	0.3529	0.6407	0.2822
Naïve	0.3517	0.5	0.2590

整体模型同对比方法的性能

Method	F_{max}	AUC	AUPR
PHENOstruct	0.4228	0.7760	0.3596
S→D→H	0.3476	0.7606	0.2580
SVM	0.4055	0.6831	0.2900
LR	0.4242	0.6690	0.2972
HTD-DAG	0.4134	0.6832	0.2951
TPR-DAG	0.4253	0.6840	0.3170
PhenoPPIOrth	0.1430	0.5731	0.0558
HPO2GO	0.2751	0.5395	0.0936
Naïve	0.3517	0.5	0.2591
HPOLabeler (Proposed)	**0.4688***	**0.7956**	**0.4293***

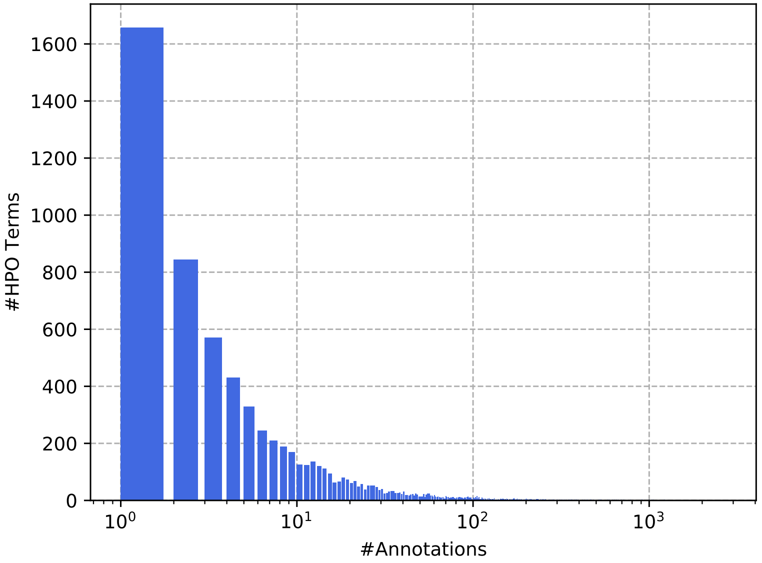
注：F_{max}是基于蛋白质计算的
AUC是基于HPO术语计算的
AUPR是就整体结果而言的

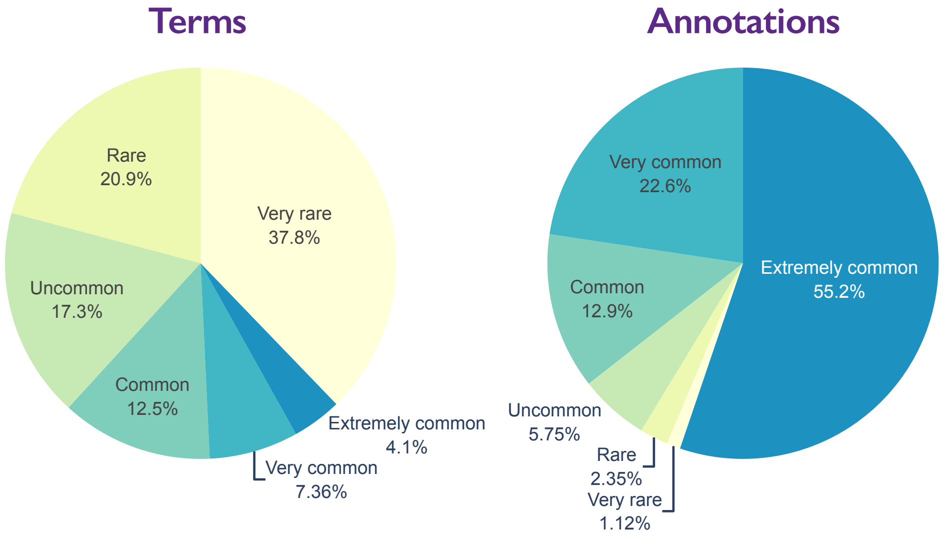
实验结果之交叉验证 — Leave-one-source-out



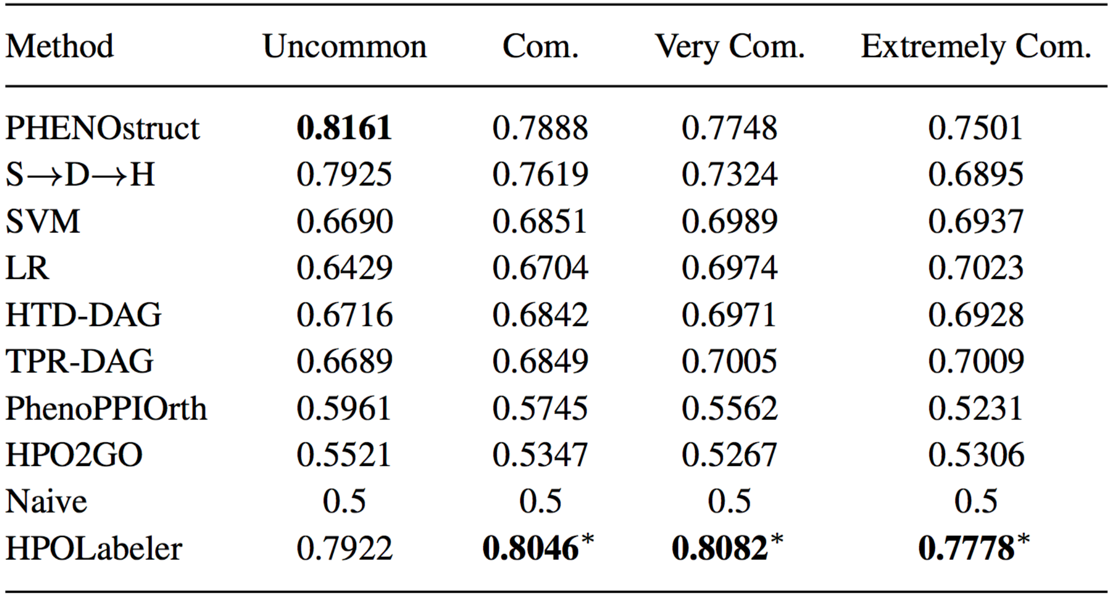
- **PPI：最有效**
- **NN：性能最好**
- **所有的变化<0：不可或缺**

实验结果之交叉验证 — 频率小组内平均AUC





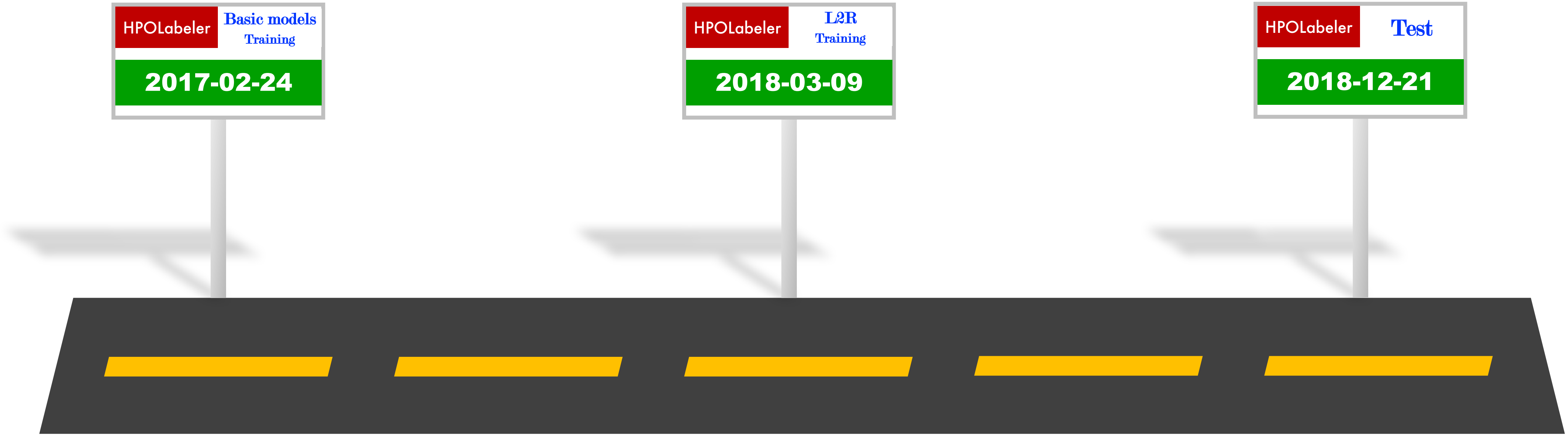
HPO及其注释是不均衡的



Method	Uncommon	Com.	Very Com.	Extremely Com.
PHENOstruct	**0.8161**	0.7888	0.7748	0.7501
S→D→H	0.7925	0.7619	0.7324	0.6895
SVM	0.6690	0.6851	0.6989	0.6937
LR	0.6429	0.6704	0.6974	0.7023
HTD-DAG	0.6716	0.6842	0.6971	0.6928
TPR-DAG	0.6689	0.6849	0.7005	0.7009
PhenoPPIOrth	0.5961	0.5745	0.5562	0.5231
HPO2GO	0.5521	0.5347	0.5267	0.5306
Naive	0.5	0.5	0.5	0.5
HPOLabeler	0.7922	**0.8046***	**0.8082***	**0.7778***

- 高频率小组 ^_^
- 低频率小组 -_-

评估之二：依时间划分验证



HPOLabeler Basic models Training

2017-02-24

HPOLabeler L2R Training

2018-03-09

HPOLabeler Test

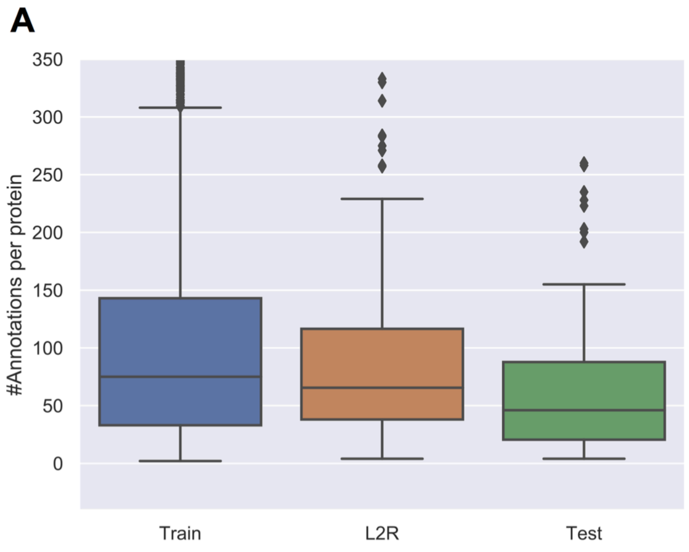
2018-12-21

	Train	L2R	Test
Proteins	3,334	304	226
Used HPO terms	7,394	2,836	2,091
Annotations	107.0936	83.9079	61.5177

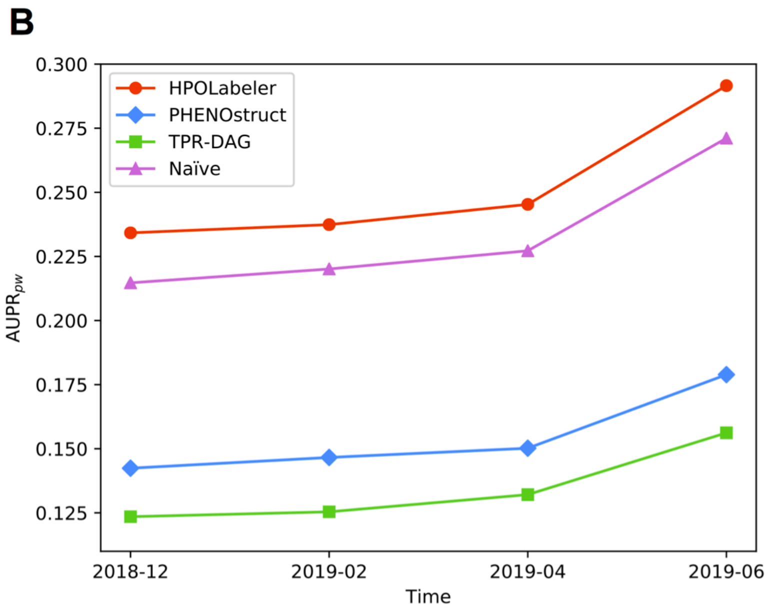
实验结果之依时间划分验证

整体模型同对比方法的性能

Method	$F_{\max}$	AUC	AUPR
PHENOstruct	0.3054	0.6362	0.1424
S→D→H	0.1461	0.5473	0.0603
SVM	0.2791	0.5929	0.1077
LR	0.2956	0.5950	0.1119
HTD-DAG	0.2933	0.5956	0.1138
TPR-DAG	0.3002	0.5962	0.1235
PhenoPPIOrth	0.0678	0.5219	0.0121
HPO2GO	0.2075	0.5083	0.0277
Naïve	0.3097	0.5	0.2147
HPOLabeler (Proposed)	**0.3415**	**0.6398**	**0.2342**

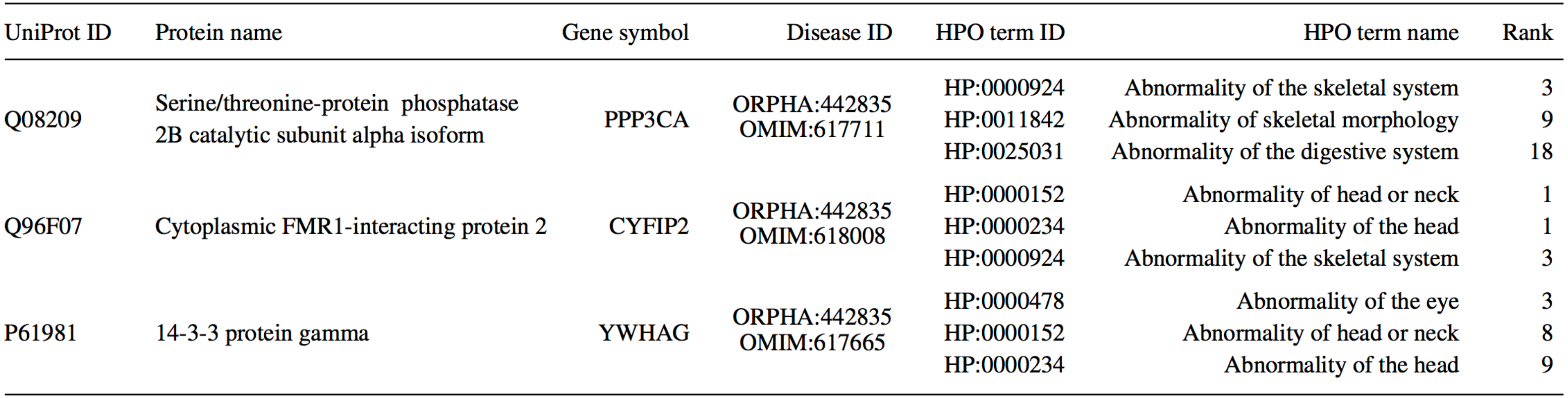


平均每个蛋白质的HPO标注条数



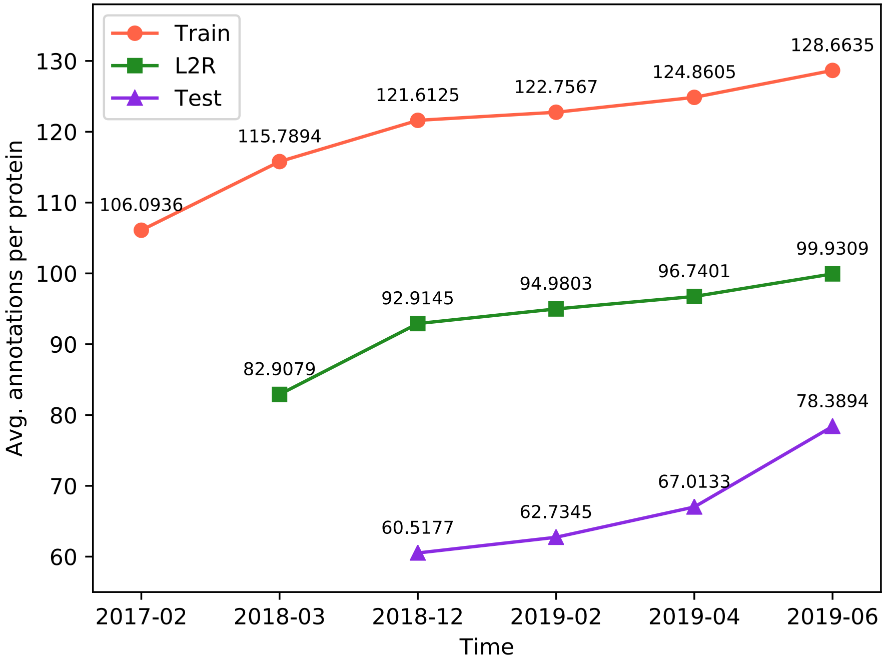
使用不同时间发布的标注文件对预测结果进行评估

HPO标注文件存在着不完善之处



UniProt ID	Protein name	Gene symbol	Disease ID	HPO term ID	HPO term name	Rank
Q08209	Serine/threonine-protein phosphatase 2B catalytic subunit alpha isoform	PPP3CA	ORPHA:442835 OMIM:617711	HP:0000924	Abnormality of the skeletal system	3
				HP:0011842	Abnormality of skeletal morphology	9
				HP:0025031	Abnormality of the digestive system	18
Q96F07	Cytoplasmic FMR1-interacting protein 2	CYFIP2	ORPHA:442835 OMIM:618008	HP:0000152	Abnormality of head or neck	1
				HP:0000234	Abnormality of the head	1
				HP:0000924	Abnormality of the skeletal system	3
P61981	14-3-3 protein gamma	YWHAG	ORPHA:442835 OMIM:617665	HP:0000478	Abnormality of the eye	3
				HP:0000152	Abnormality of head or neck	8
				HP:0000234	Abnormality of the head	9

依据旧标注文件而被判定为“错误”
但根据新发布的标注文件应当是“正确”
的预测结果（节选）



标注文件中新加入的蛋白质的平均
标注个数随着时间而不断积累增加

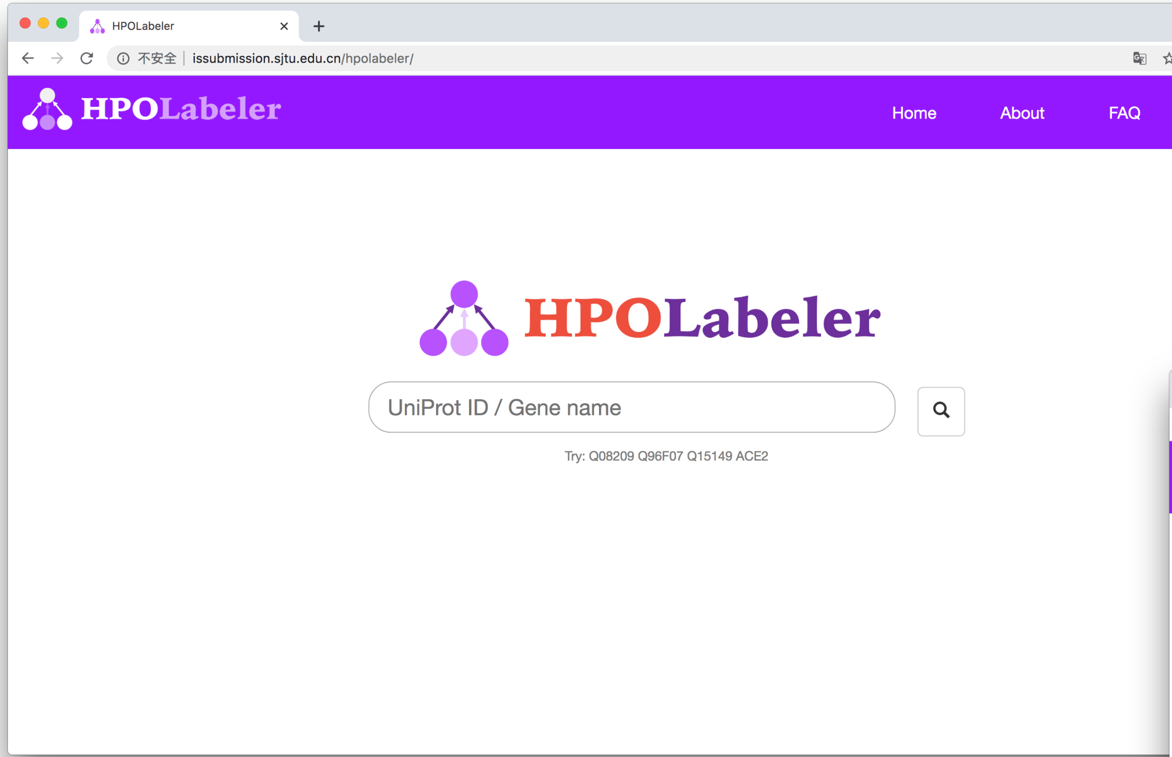
小结

- 我们提出了预测人类蛋白质的HPO标注的算法HPOLabeler，其在排序学习的框架下整合了包括PPI、GO、InterPro和序列信息等在内的多种信息源。

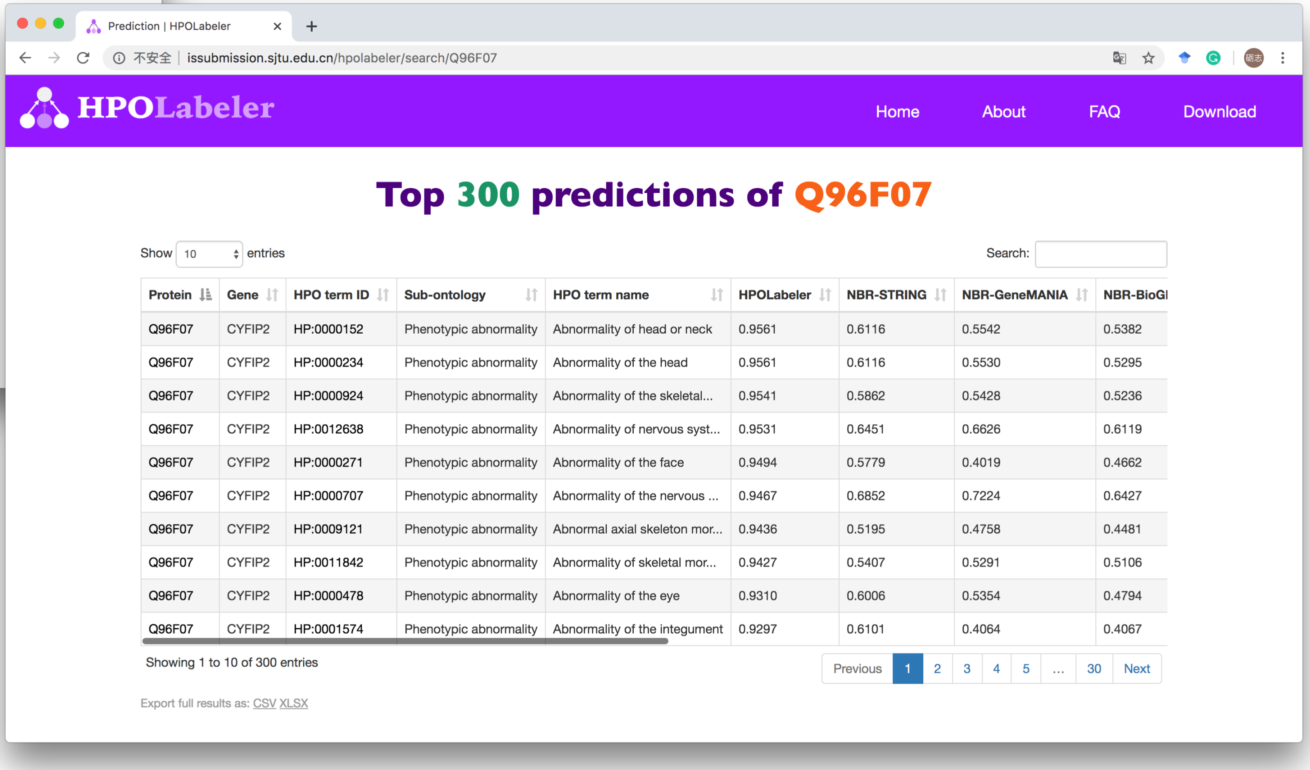
- 经过实验验证，HPOLabeler显著的优于其他对比方法。

- 进一步的实验结果表明：
  - 在所用信息源中，PPI是最有效的一个；
  - 依时间划分验证中较低的性能值可能是由新增蛋白质的HPO标注不完善所导致的。

在线查询平台



Try: Q08209 Q96F07 Q15149 ACE2



Top 300 predictions of Q96F07

Show 10 entries Search:

Protein	Gene	HPO term ID	Sub-ontology	HPO term name	HPOLabeler	NBR-STRING	NBR-GeneMANIA	NBR-BioGl
Q96F07	CYFIP2	HP:0000152	Phenotypic abnormality	Abnormality of head or neck	0.9561	0.6116	0.5542	0.5382
Q96F07	CYFIP2	HP:0000234	Phenotypic abnormality	Abnormality of the head	0.9561	0.6116	0.5530	0.5295
Q96F07	CYFIP2	HP:0000924	Phenotypic abnormality	Abnormality of the skeletal...	0.9541	0.5862	0.5428	0.5236
Q96F07	CYFIP2	HP:0012638	Phenotypic abnormality	Abnormality of nervous syst...	0.9531	0.6451	0.6626	0.6119
Q96F07	CYFIP2	HP:0000271	Phenotypic abnormality	Abnormality of the face	0.9494	0.5779	0.4019	0.4662
Q96F07	CYFIP2	HP:0000707	Phenotypic abnormality	Abnormality of the nervous ...	0.9467	0.6852	0.7224	0.6427
Q96F07	CYFIP2	HP:0009121	Phenotypic abnormality	Abnormal axial skeleton mor...	0.9436	0.5195	0.4758	0.4481
Q96F07	CYFIP2	HP:0011842	Phenotypic abnormality	Abnormality of skeletal mor...	0.9427	0.5407	0.5291	0.5106
Q96F07	CYFIP2	HP:0000478	Phenotypic abnormality	Abnormality of the eye	0.9310	0.6006	0.5354	0.4794
Q96F07	CYFIP2	HP:0001574	Phenotypic abnormality	Abnormality of the integument	0.9297	0.6101	0.4064	0.4067

Showing 1 to 10 of 300 entries

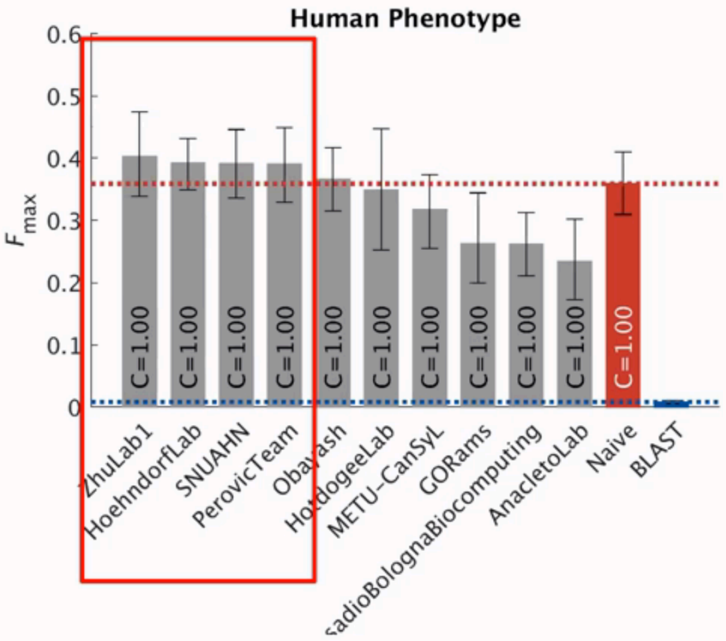
Export full results as: CSV XLSX

http://issubmission.sjtu.edu.cn/hpolabeler/

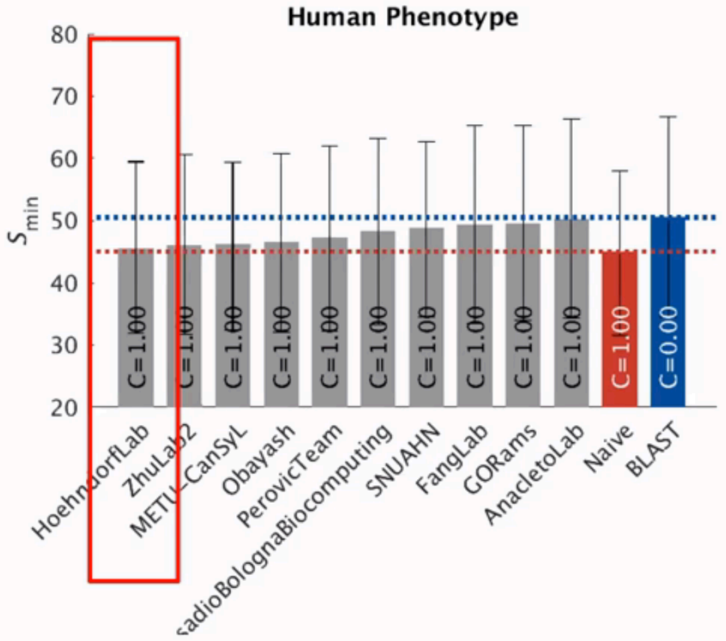
CAFA4竞赛初步评估结果

Human Phenotype Ontology

Benchmark: all species Mode: full Metrics: F_{max} and S_{min}



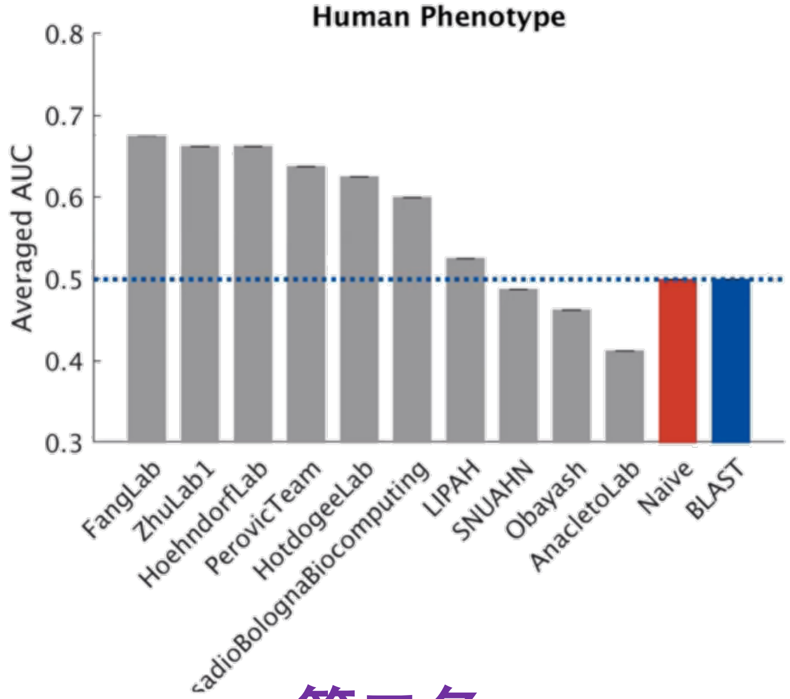
第一名



第二名

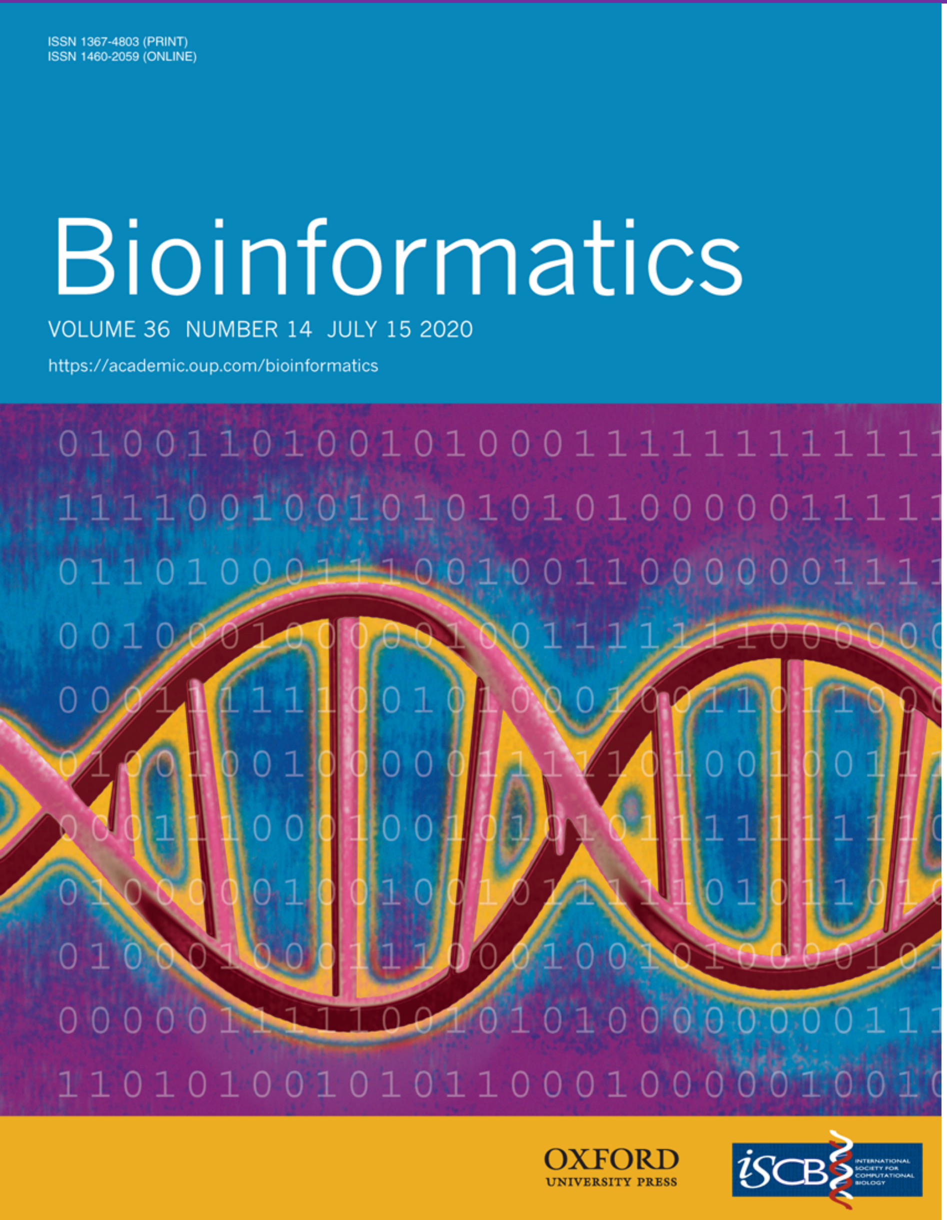
Human Phenotype, Term-Centric Evaluation

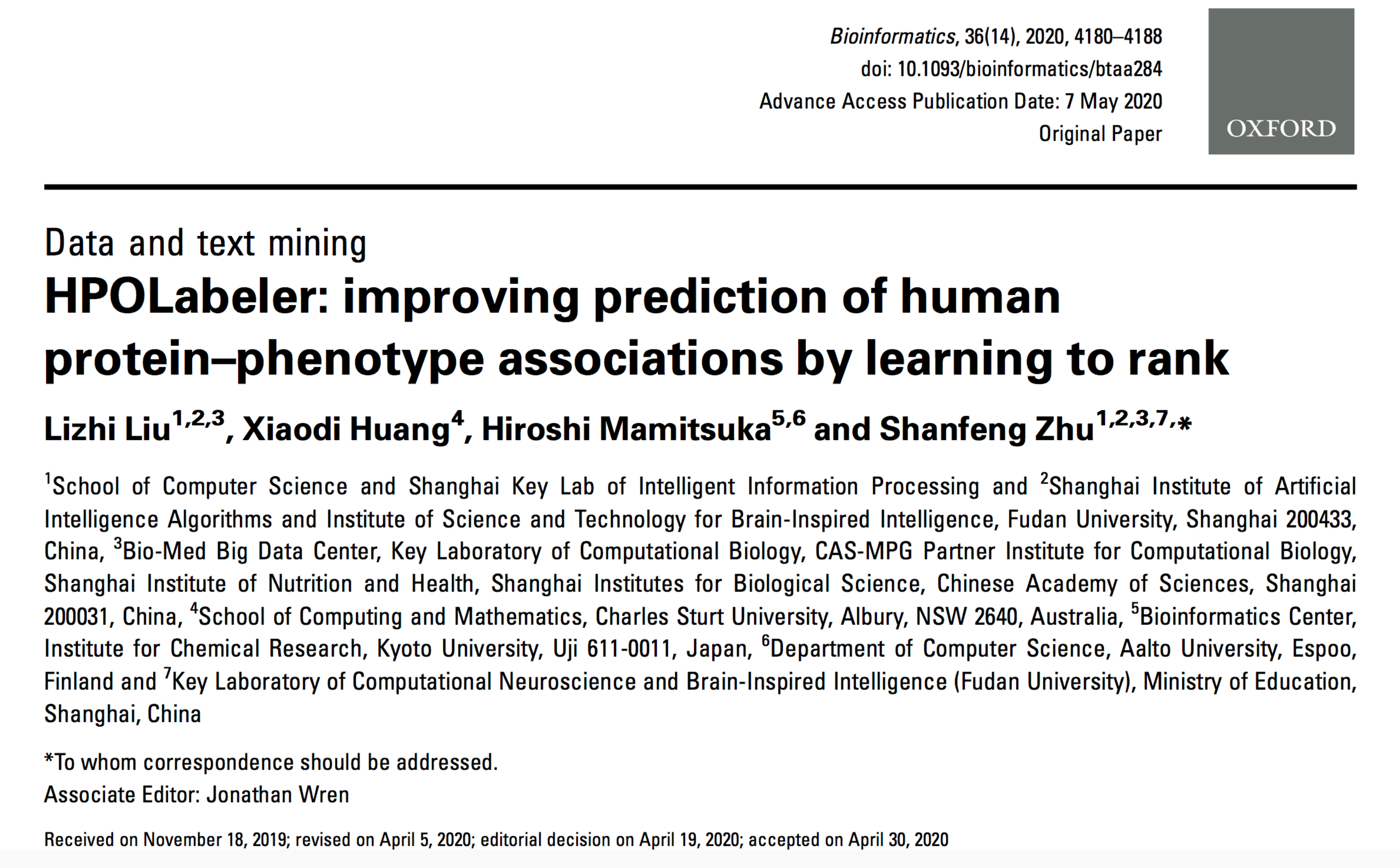
Terms: 4



第二名

论文发表


ISSN 1367-4803 (PRINT)
ISSN 1460-2059 (ONLINE)
Bioinformatics
VOLUME 36 NUMBER 14 JULY 15 2020
https://academic.oup.com/bioinformatics
OXFORD UNIVERSITY PRESS
iSCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY



Bioinformatics, 36(14), 2020, 4180–4188
doi: 10.1093/bioinformatics/btaa284
Advance Access Publication Date: 7 May 2020
Original Paper

Data and text mining

HPOLabeler: improving prediction of human protein–phenotype associations by learning to rank

Lizhi Liu[1,2,3], **Xiaodi Huang**[4], **Hiroshi Mamitsuka**[5,6] **and Shanfeng Zhu**[1,2,3,7,*]

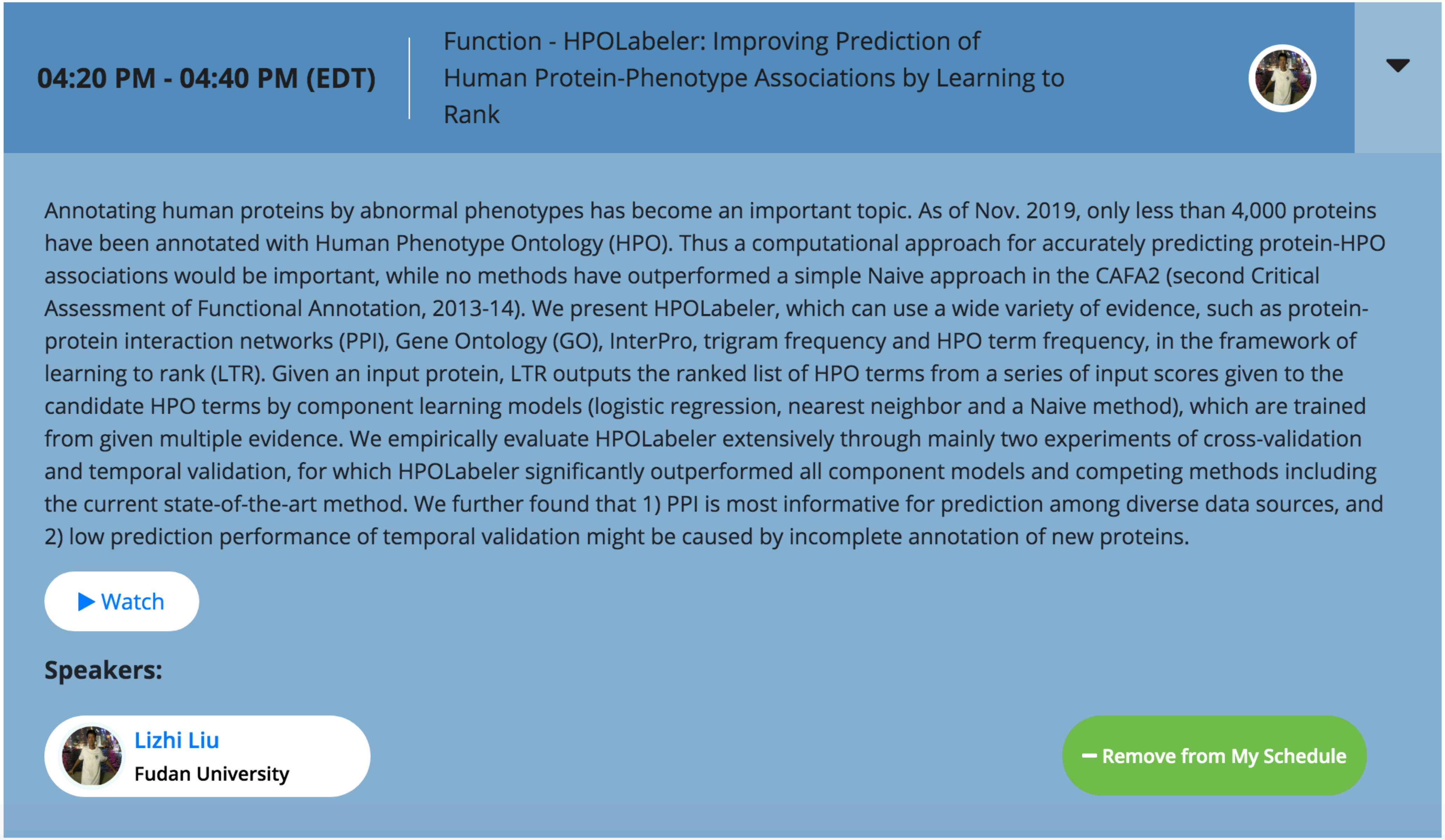
[1]School of Computer Science and Shanghai Key Lab of Intelligent Information Processing and [2]Shanghai Institute of Artificial Intelligence Algorithms and Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China, [3]Bio-Med Big Data Center, Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Science, Chinese Academy of Sciences, Shanghai 200031, China, [4]School of Computing and Mathematics, Charles Sturt University, Albury, NSW 2640, Australia, [5]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji 611-0011, Japan, [6]Department of Computer Science, Aalto University, Espoo, Finland and [7]Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (Fudan University), Ministry of Education, Shanghai, China

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on November 18, 2019; revised on April 5, 2020; editorial decision on April 19, 2020; accepted on April 30, 2020

参加学术会议



04:20 PM - 04:40 PM (EDT)

Function - HPOLabeler: Improving Prediction of Human Protein-Phenotype Associations by Learning to Rank

Annotating human proteins by abnormal phenotypes has become an important topic. As of Nov. 2019, only less than 4,000 proteins have been annotated with Human Phenotype Ontology (HPO). Thus a computational approach for accurately predicting protein-HPO associations would be important, while no methods have outperformed a simple Naive approach in the CAFA2 (second Critical Assessment of Functional Annotation, 2013-14). We present HPOLabeler, which can use a wide variety of evidence, such as protein-protein interaction networks (PPI), Gene Ontology (GO), InterPro, trigram frequency and HPO term frequency, in the framework of learning to rank (LTR). Given an input protein, LTR outputs the ranked list of HPO terms from a series of input scores given to the candidate HPO terms by component learning models (logistic regression, nearest neighbor and a Naive method), which are trained from given multiple evidence. We empirically evaluate HPOLabeler extensively through mainly two experiments of cross-validation and temporal validation, for which HPOLabeler significantly outperformed all component models and competing methods including the current state-of-the-art method. We further found that 1) PPI is most informative for prediction among diverse data sources, and 2) low prediction performance of temporal validation might be caused by incomplete annotation of new proteins.

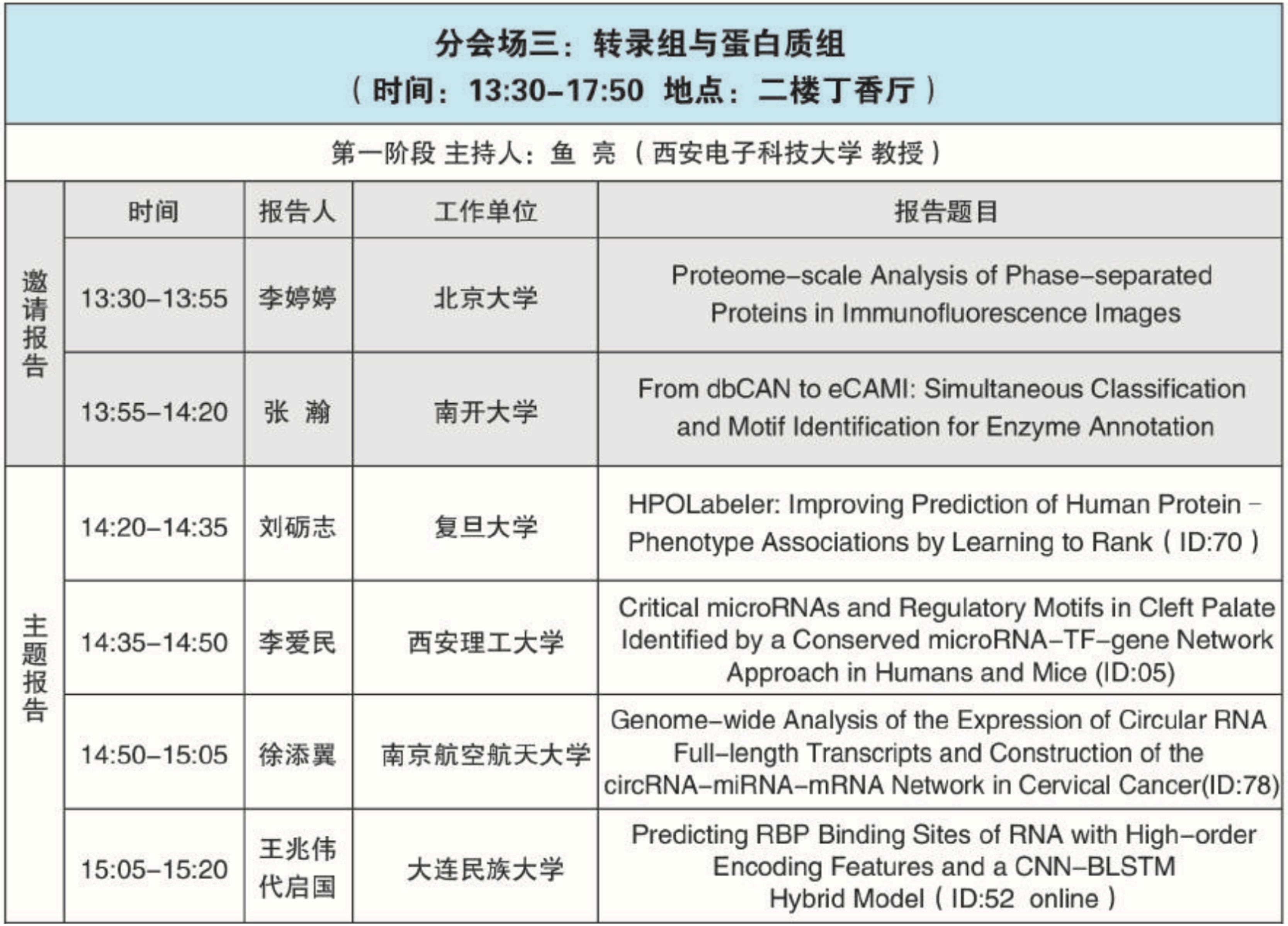
Speakers:

Lizhi Liu
Fudan University

参加学术会议

第五届中国计算机学会生物信息学会议

The Fifth CCF Bioinformatics Conference (CBC2020)



分会场三：转录组与蛋白质组

（时间：13:30–17:50　地点：二楼丁香厅）

第一阶段 主持人：鱼　亮（西安电子科技大学 教授）

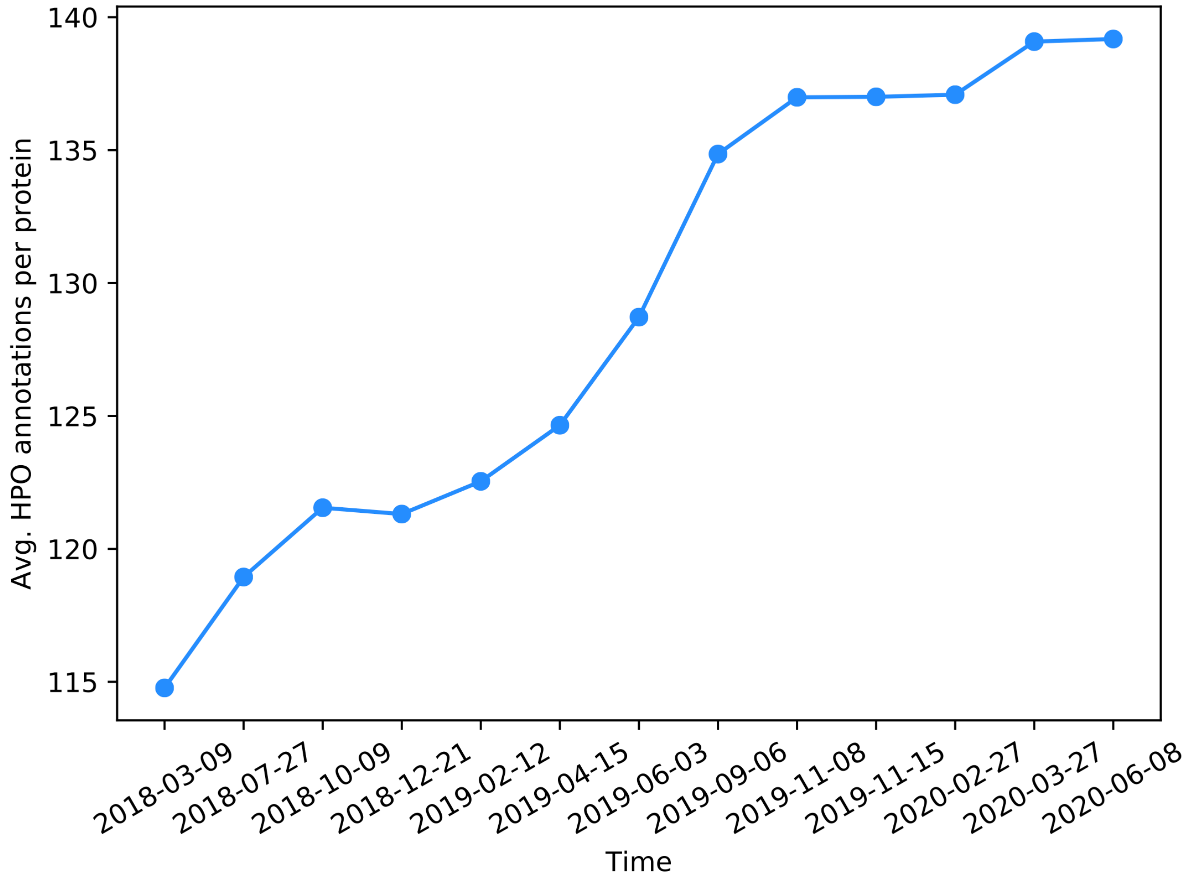
	时间	报告人	工作单位	报告题目
邀请报告	13:30–13:55	李婷婷	北京大学	Proteome–scale Analysis of Phase–separated Proteins in Immunofluorescence Images
	13:55–14:20	张　瀚	南开大学	From dbCAN to eCAMI: Simultaneous Classification and Motif Identification for Enzyme Annotation
主题报告	14:20–14:35	刘砺志	复旦大学	HPOLabeler: Improving Prediction of Human Protein – Phenotype Associations by Learning to Rank（ID:70）
	14:35–14:50	李爱民	西安理工大学	Critical microRNAs and Regulatory Motifs in Cleft Palate Identified by a Conserved microRNA–TF–gene Network Approach in Humans and Mice (ID:05)
	14:50–15:05	徐添翼	南京航空航天大学	Genome–wide Analysis of the Expression of Circular RNA Full–length Transcripts and Construction of the circRNA–miRNA–mRNA Network in Cervical Cancer(ID:78)
	15:05–15:20	王兆伟 代启国	大连民族大学	Predicting RBP Binding Sites of RNA with High–order Encoding Features and a CNN–BLSTM Hybrid Model（ID:52 online）

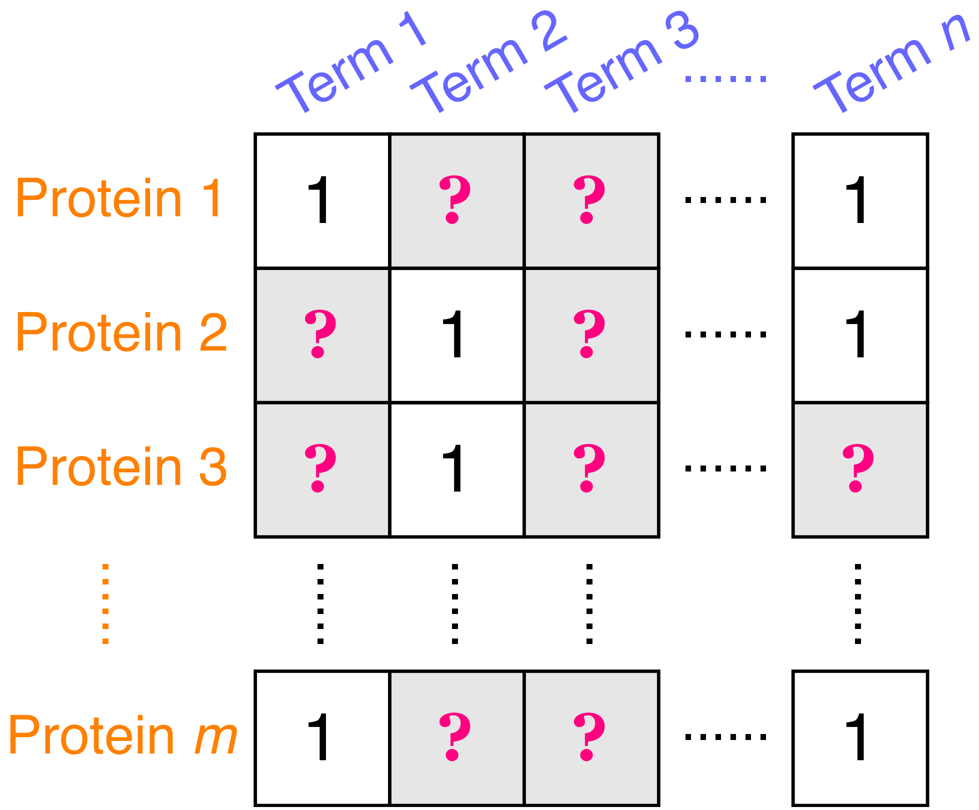
2

HPOFiller

基于图卷积网络预测缺失的蛋白质表型标注

问题描述：填补缺失的蛋白质HPO注释





	Term 1	Term 2	Term 3	……	Term *n*
Protein 1	1	?	?	……	1
Protein 2	?	1	?	……	1
Protein 3	?	1	?	……	?
⋮	⋮	⋮	⋮		⋮
Protein *m*	1	?	?	……	1

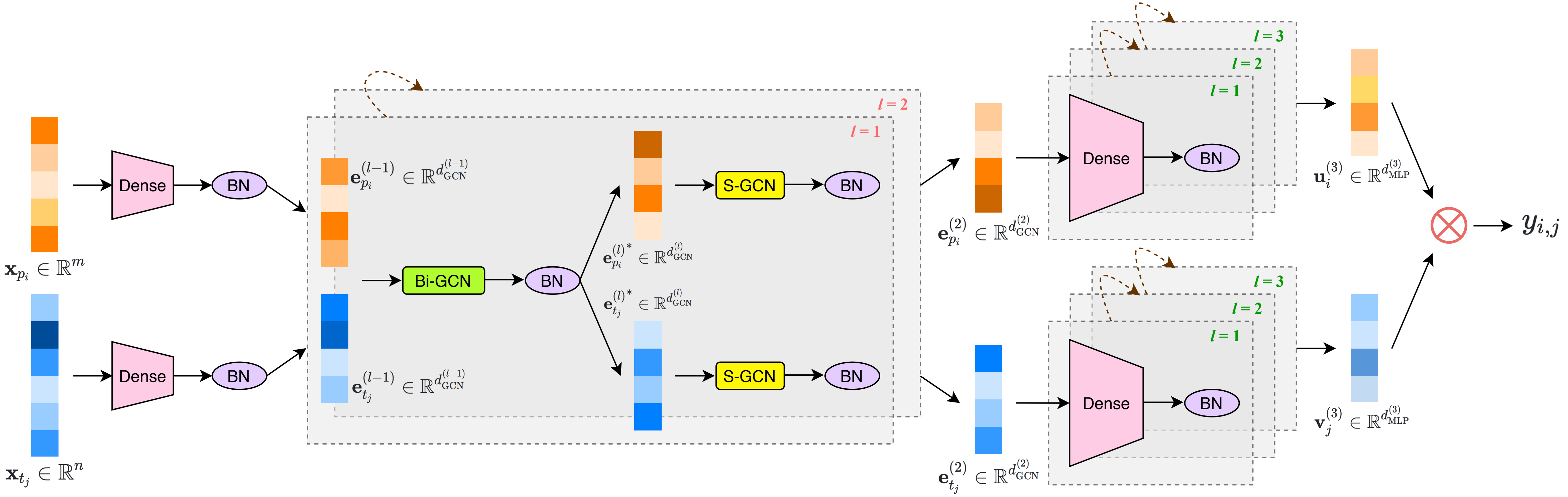
研究目标：充分利用蛋白质互作网络和HPO的层次结构，提高预测精度

现有的缺失蛋白质HPO标注预测算法

Base	Method	Data source(s)	HPO hierarchy	Optimization
Label propagation	LP [59-60]	PPI	-	Closed-form sol.
	DLP [61]	PPI	Raw HPO DAG	L-BFGS-B
	tlDLP [61]	PPI, GO annotation	Raw HPO DAG	L-BFGS-B
Matrix completion	SMC [62]	-	-	ALS
	MCNMF [63]	PPI	Jaccard coefficient	ALS
	GC^{2}NMF [64]	PPI, Pathway	Depth-adjusted DAG	ALS
	AiProAnnotator [65]	PPI	Lin method	ALS
	HPOAnnotator [66]	Multiple PPIs	Lin method	ALS

- 仅能捕捉隐藏于蛋白质-异常表型关系间的线性关联，而忽视了非线性关联
- 仅能捕捉相似度网络中的低阶拓扑结构，而忽视了高阶连通性
- 目前还没有研究人员在此类领域提出基于深度学习的预测算法

HPOFiller — 使用图神经网络预测缺失注释



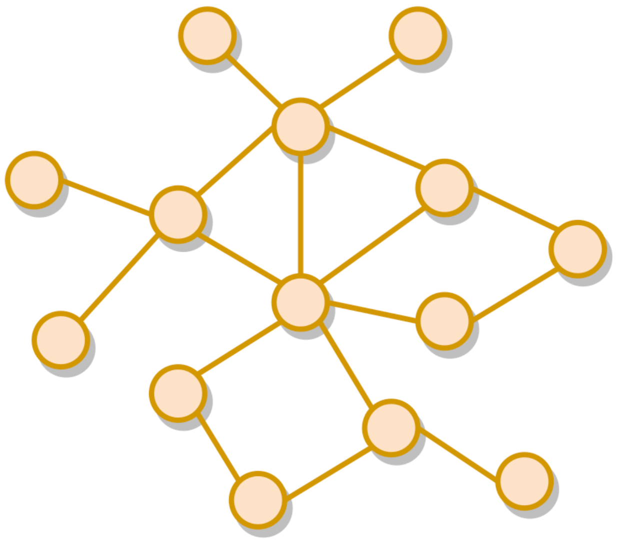
$\mathbf{x}_{p_i} \in \mathbb{R}^m$
$\mathbf{x}_{t_j} \in \mathbb{R}^n$
Dense
BN
Dense
BN
$l=2$
$l=1$
$\mathbf{e}_{p_i}^{(l-1)} \in \mathbb{R}^{d_{\mathrm{GCN}}^{(l-1)}}$
$\mathbf{e}_{t_j}^{(l-1)} \in \mathbb{R}^{d_{\mathrm{GCN}}^{(l-1)}}$
Bi-GCN
BN
$\mathbf{e}_{p_i}^{(l)*} \in \mathbb{R}^{d_{\mathrm{GCN}}^{(l)}}$
$\mathbf{e}_{t_j}^{(l)*} \in \mathbb{R}^{d_{\mathrm{GCN}}^{(l)}}$
S-GCN
BN
S-GCN
BN
$\mathbf{e}_{p_i}^{(2)} \in \mathbb{R}^{d_{\mathrm{GCN}}^{(2)}}$
$\mathbf{e}_{t_j}^{(2)} \in \mathbb{R}^{d_{\mathrm{GCN}}^{(2)}}$
$l=3$
$l=2$
$l=1$
Dense
BN
$l=3$
$l=2$
$l=1$
Dense
BN
$\mathbf{u}_i^{(3)} \in \mathbb{R}^{d_{\mathrm{MLP}}^{(3)}}$
$\mathbf{v}_j^{(3)} \in \mathbb{R}^{d_{\mathrm{MLP}}^{(3)}}$
$\otimes$
$y_{i,j}$

关键点

- 提出两种运行于相似度网络和二部网络上的**图神经网络**模块
- 使用**ε-增强损失函数**缓解标签不平衡的影响
- 设计**极为严苛的交叉验证评估**流程以避免信息泄露

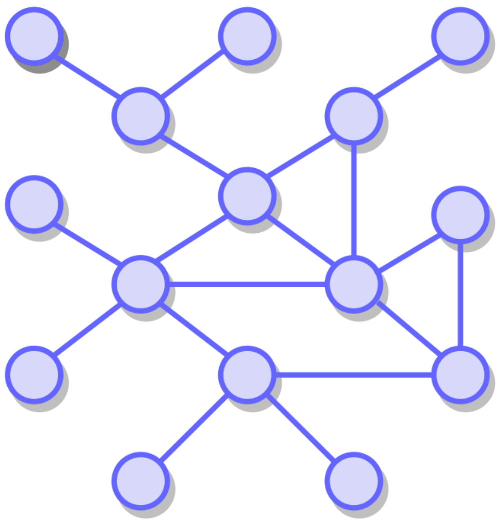
图的构造

蛋白质互作网络



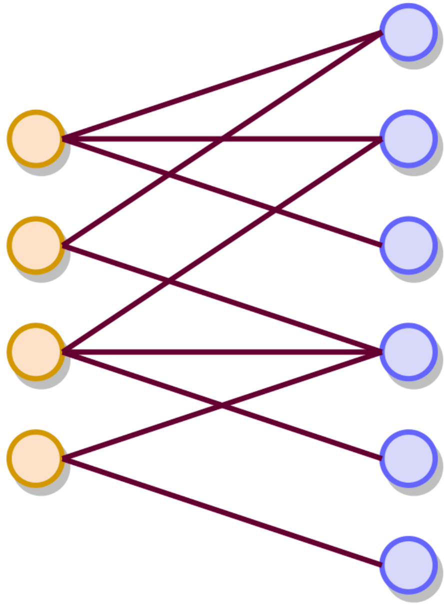
STRING

HPO术语相似度网络



$\mathrm{Sim}_{\mathrm{IC}}$

蛋白质-HPO术语二部网络



$$\mathrm{IC}(t) = -\log \frac{N_t}{N}$$

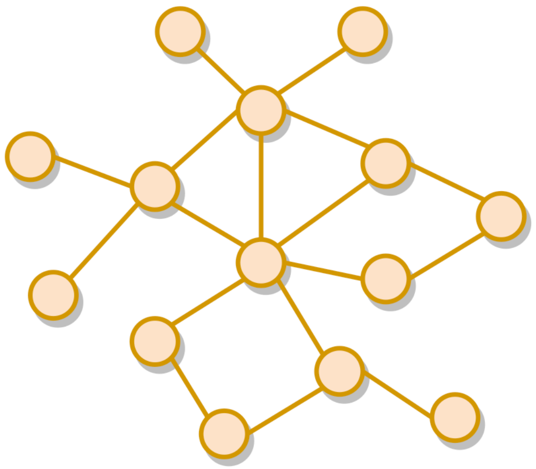
$$\mathrm{sim}_{\mathrm{IC}}(t_1, t_2) = \frac{2 \times \mathrm{IC}(t_{\mathrm{MICA}})}{\mathrm{IC}(t_1) + \mathrm{IC}(t_2)} \times \left(1 - \frac{1}{1 + \mathrm{IC}(t_{\mathrm{MICA}})}\right)$$

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} & \tilde{\mathbf{Y}} \\ \tilde{\mathbf{Y}}^{\mathsf{T}} & \mathbf{0} \end{bmatrix}$$

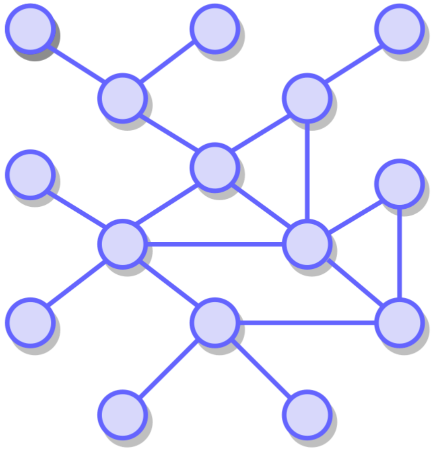
[Li et al. BIOCOMP, 2010]

特征生成

蛋白质互作网络



HPO术语相似度网络



$$\mathbf{p}_i^{t+1} = (1-\alpha)\mathbf{p}_i^t\hat{\mathbf{S}} + \alpha\mathbf{e}_i$$

带重启动的随机游走

$$\mathbf{x}_{p_i} = \mathbf{p}_i^{\infty}$$

蛋白质的特征向量

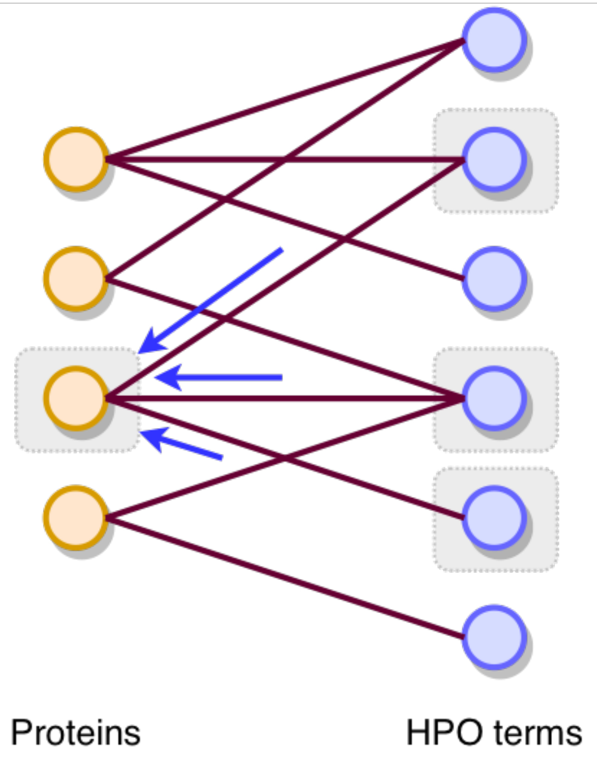
$$\mathbf{x}_{t_j} = \mathbf{p}_j^{\infty}$$

HPO术语的特征向量

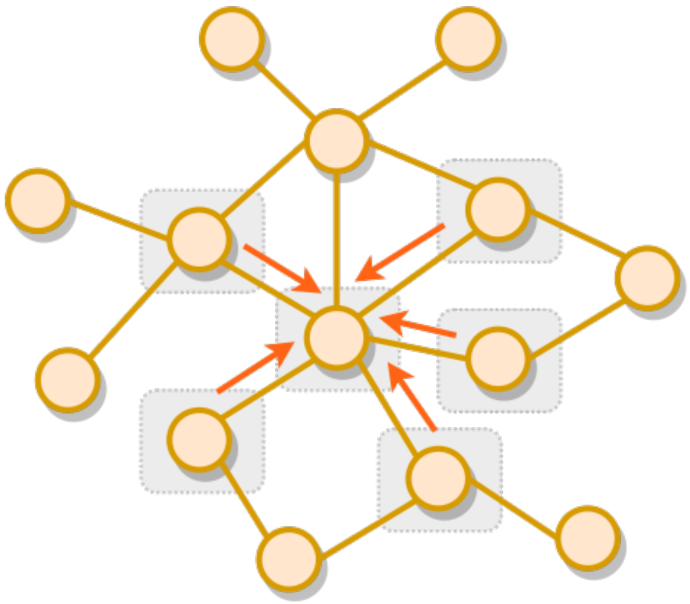
两种图神经网络模块

$$\mathbf{e}_{p_i}^{(l)} = \sigma\left(\mathbf{e}_{p_i}^{(l-1)}\boldsymbol{\Theta}_1^{(l)} + \sum_{t_j \in \mathcal{N}(p_i)} \mathbf{e}_{t_j}^{(l-1)}\boldsymbol{\Theta}_2^{(l)}\right)$$

$$\mathbf{e}_{p_i}^{(l)} = \sigma\left(\sum_{j=1}^{m}(\tilde{\mathbf{S}}_p)_{i,j}\mathbf{e}_{p_j}^{(l-1)}\boldsymbol{\Theta}_p^{(l)}\right)$$

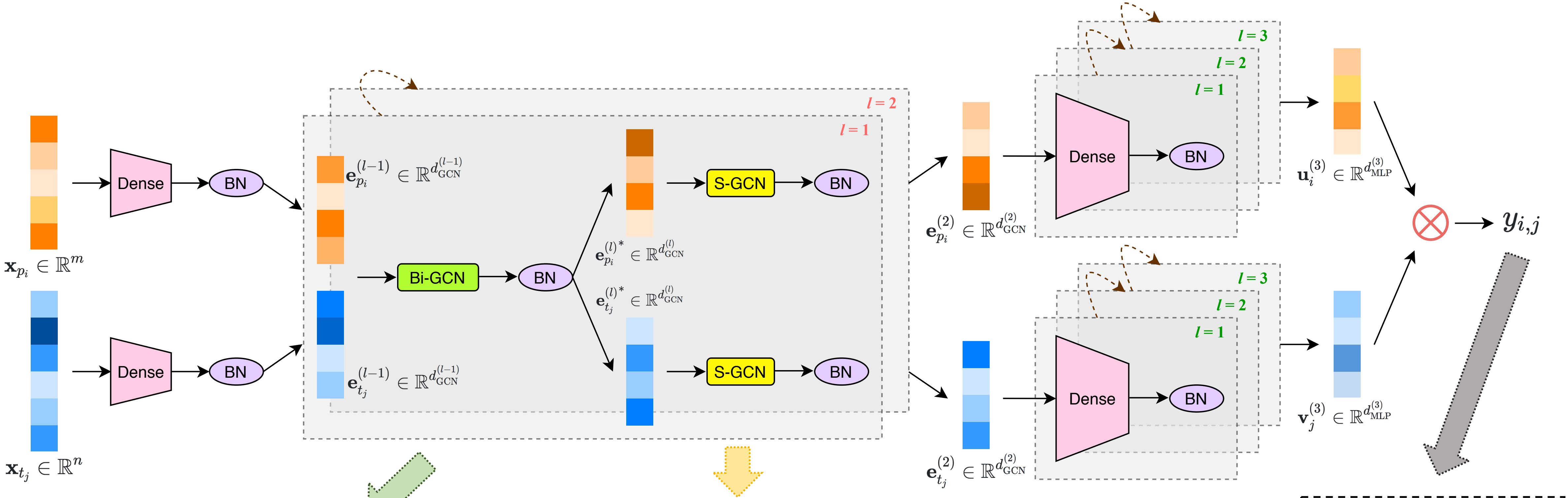


(a) Bi-GCN Block



(b) S-GCN Block

模型架构



$\mathbf{x}_{p_i} \in \mathbb{R}^m$
$\mathbf{x}_{t_j} \in \mathbb{R}^n$
Dense
BN
Dense
BN
$l=2$
$l=1$
$\mathbf{e}_{p_i}^{(l-1)} \in \mathbb{R}^{d_{\mathrm{GCN}}^{(l-1)}}$
$\mathbf{e}_{t_j}^{(l-1)} \in \mathbb{R}^{d_{\mathrm{GCN}}^{(l-1)}}$
Bi-GCN
BN
$\mathbf{e}_{p_i}^{(l)*} \in \mathbb{R}^{d_{\mathrm{GCN}}^{(l)}}$
$\mathbf{e}_{t_j}^{(l)*} \in \mathbb{R}^{d_{\mathrm{GCN}}^{(l)}}$
S-GCN
BN
S-GCN
BN
$\mathbf{e}_{p_i}^{(2)} \in \mathbb{R}^{d_{\mathrm{GCN}}^{(2)}}$
$\mathbf{e}_{t_j}^{(2)} \in \mathbb{R}^{d_{\mathrm{GCN}}^{(2)}}$
$l=3$
$l=2$
$l=1$
Dense
BN
$l=3$
$l=2$
$l=1$
Dense
BN
$\mathbf{u}_i^{(3)} \in \mathbb{R}^{d_{\mathrm{MLP}}^{(3)}}$
$\mathbf{v}_j^{(3)} \in \mathbb{R}^{d_{\mathrm{MLP}}^{(3)}}$
$y_{i,j}$

$$\left[\mathbf{E}_p^{(l)*};\mathbf{E}_t^{(l)*}\right] = \mathrm{BN}^{(l)}\left(\text{Bi-GCN}^{(l)}\left(\left[\mathbf{E}_p^{(l-1)};\mathbf{E}_t^{(l-1)}\right]\right)\right)$$

BN = Batch Normalization

$$\mathbf{E}_p^{(l)} = \mathrm{BN}_p^{(l)}\left(\text{S-GCN}_p^{(l)}\left(\mathbf{E}_p^{(l)*}\right)\right)$$

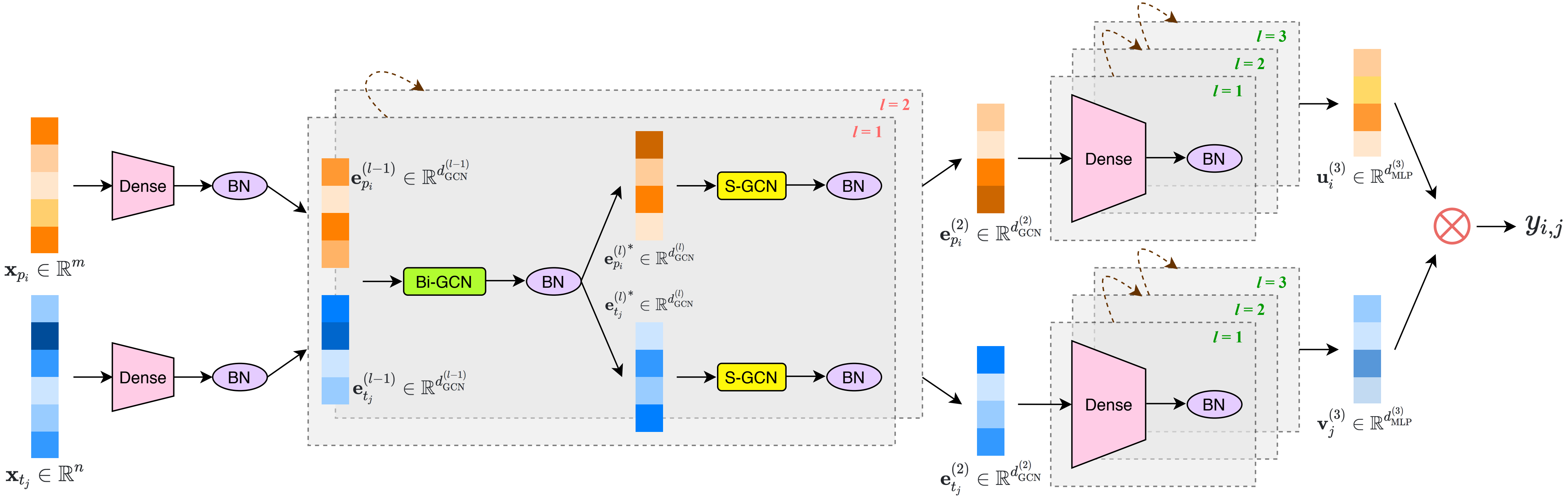
$$\mathbf{E}_t^{(l)} = \mathrm{BN}_t^{(l)}\left(\text{S-GCN}_t^{(l)}\left(\mathbf{E}_t^{(l)*}\right)\right)$$

$$y_{i,j} = \mathbf{u}_i^{(3)}\mathbf{v}_j^{(3)\top}$$

$$\mathbf{U}^{(3)} = \mathrm{Dense}_p^{(3)}\left(\mathrm{BN}_p^{(2)}\left(\cdots\mathrm{Dense}_p^{(1)}\left(\mathbf{E}_p^{(2)}\right)\cdots\right)\right)$$

$$\mathbf{V}^{(3)} = \mathrm{Dense}_t^{(3)}\left(\mathrm{BN}_t^{(2)}\left(\cdots\mathrm{Dense}_t^{(1)}\left(\mathbf{E}_t^{(2)}\right)\cdots\right)\right)$$

模型训练



$\mathbf{x}_{p_i} \in \mathbb{R}^m$
$\mathbf{x}_{t_j} \in \mathbb{R}^n$
Dense
BN
Dense
BN
$l=2$
$l=1$
$\mathbf{e}_{p_i}^{(l-1)} \in \mathbb{R}^{d_{\text{GCN}}^{(l-1)}}$
$\mathbf{e}_{t_j}^{(l-1)} \in \mathbb{R}^{d_{\text{GCN}}^{(l-1)}}$
Bi-GCN
BN
$\mathbf{e}_{p_i}^{(l)*} \in \mathbb{R}^{d_{\text{GCN}}^{(l)}}$
$\mathbf{e}_{t_j}^{(l)*} \in \mathbb{R}^{d_{\text{GCN}}^{(l)}}$
S-GCN
BN
S-GCN
BN
$\mathbf{e}_{p_i}^{(2)} \in \mathbb{R}^{d_{\text{GCN}}^{(2)}}$
$\mathbf{e}_{t_j}^{(2)} \in \mathbb{R}^{d_{\text{GCN}}^{(2)}}$
$l=3$
$l=2$
$l=1$
Dense
BN
$l=3$
$l=2$
$l=1$
Dense
BN
$\mathbf{u}_i^{(3)} \in \mathbb{R}^{d_{\text{MLP}}^{(3)}}$
$\mathbf{v}_j^{(3)} \in \mathbb{R}^{d_{\text{MLP}}^{(3)}}$
$\otimes$
$y_{i,j}$

Enhanced annotation matrix

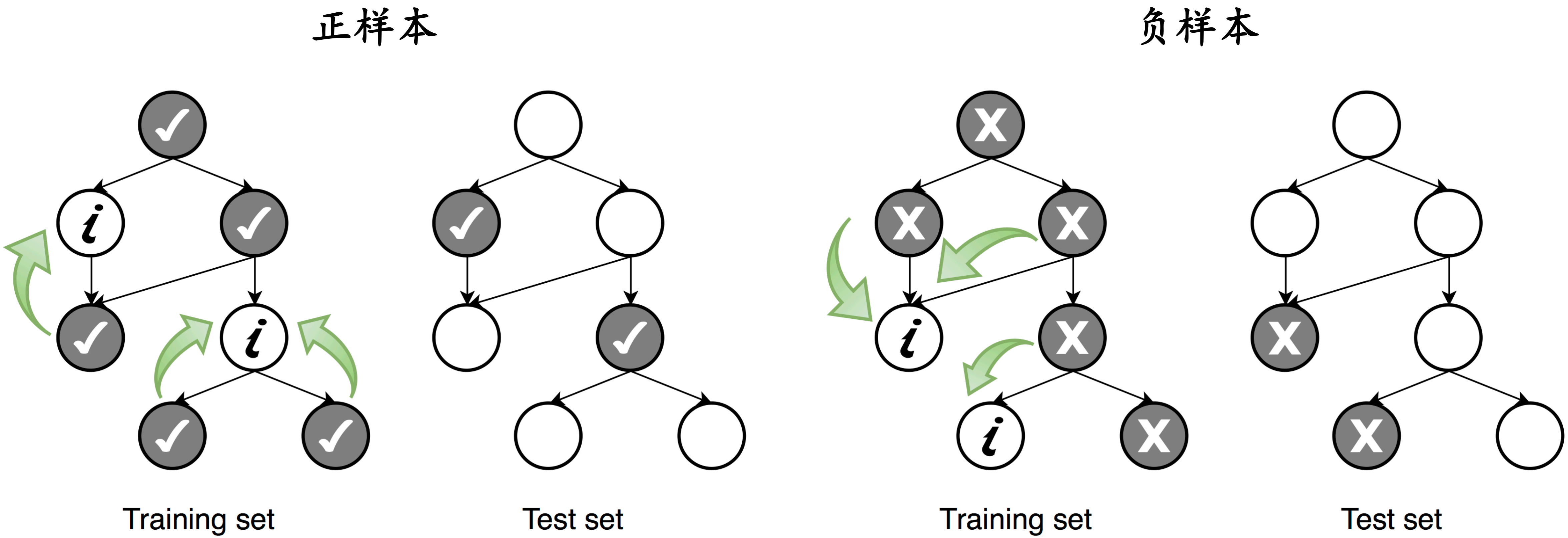
$$\tilde{\mathbf{Y}}'_{i,j} = \begin{cases} \underline{\epsilon} & \text{if } \tilde{\mathbf{Y}}_{i,j} = 1, \\ 0 & \text{otherwise.} \end{cases}$$

ε-enhanced loss function

$$\mathcal{L} = \|\mathbf{\Omega} \circ (\mathbf{Y} - \tilde{\mathbf{Y}}')\|_F^2 + \lambda \|\mathbf{\Theta}\|_2^2$$

$$\Omega_{ij} = \begin{cases} 1 & \tilde{\mathbf{Y}}_{ij} \text{ in the training set,} \\ 0 & \text{otherwise} \end{cases}$$

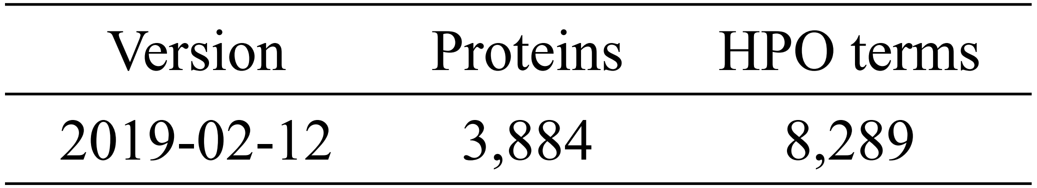
评估之一：交叉验证



信息泄露

[Petegrosso et al. Bioinformatics, 2017]
[Gao et al. BIBM, 2018]
[Gao et al. BMC Med Genomics, 2019]

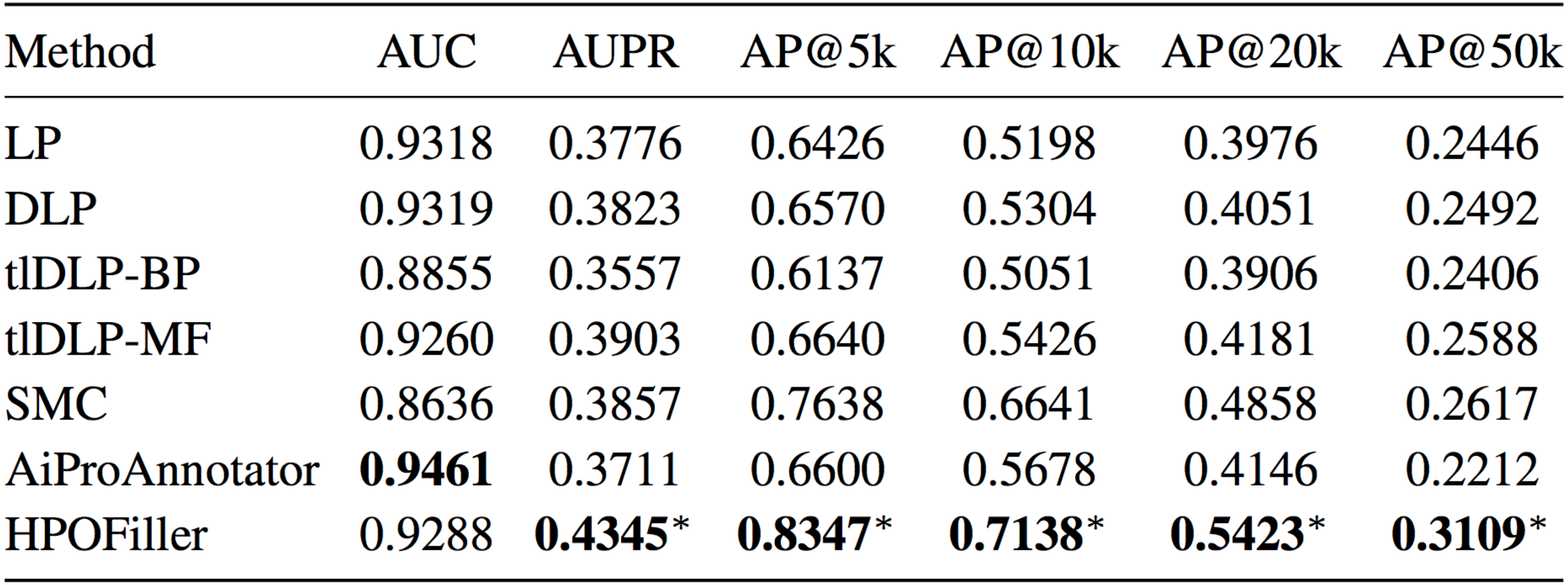
实验结果之交叉验证 — 对比



Version	Proteins	HPO terms
2019-02-12	3,884	8,289

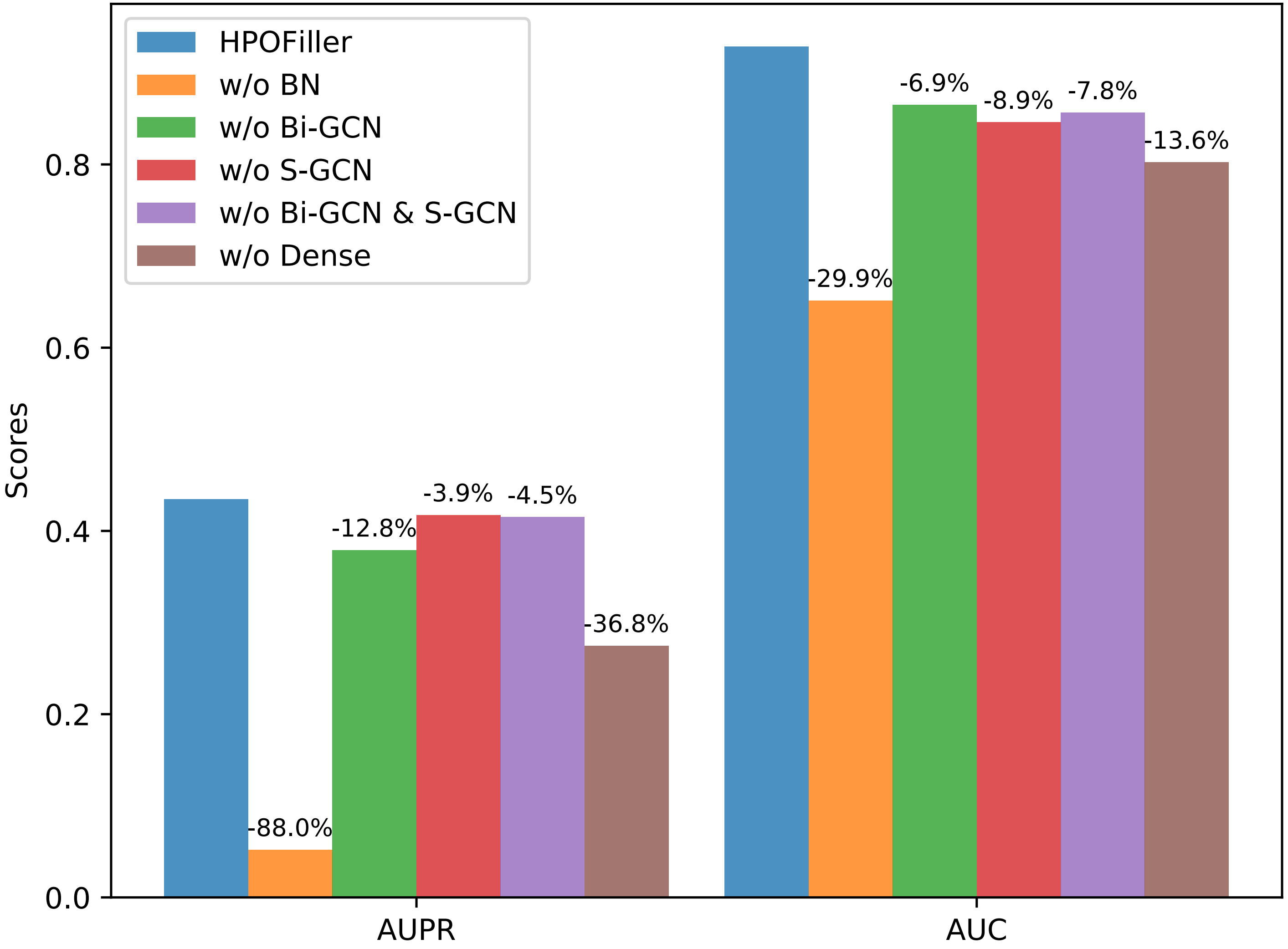
▲所用数据集统计

▼性能评估结果

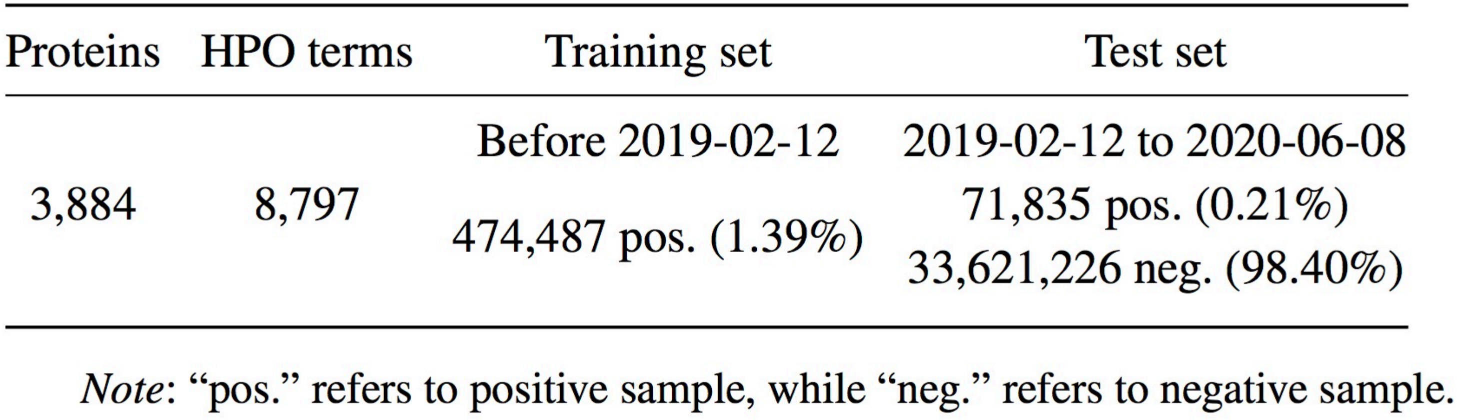


Method	AUC	AUPR	AP@5k	AP@10k	AP@20k	AP@50k
LP	0.9318	0.3776	0.6426	0.5198	0.3976	0.2446
DLP	0.9319	0.3823	0.6570	0.5304	0.4051	0.2492
tlDLP-BP	0.8855	0.3557	0.6137	0.5051	0.3906	0.2406
tlDLP-MF	0.9260	0.3903	0.6640	0.5426	0.4181	0.2588
SMC	0.8636	0.3857	0.7638	0.6641	0.4858	0.2617
AiProAnnotator	**0.9461**	0.3711	0.6600	0.5678	0.4146	0.2212
HPOFiller	0.9288	**0.4345***	**0.8347***	**0.7138***	**0.5423***	**0.3109***

实验结果之交叉验证 — 消融实验



评估之二：依时间划分验证

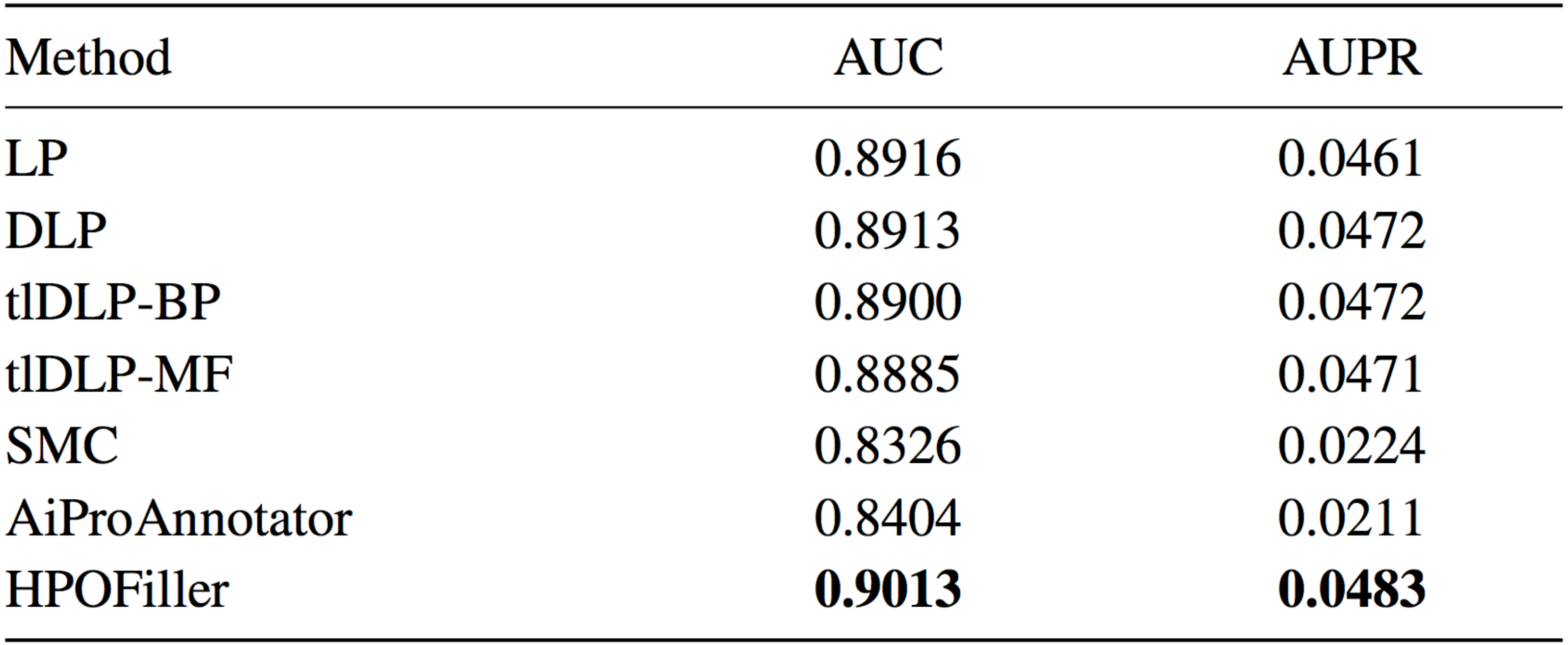


Proteins	HPO terms	Training set	Test set
3,884	8,797	Before 2019-02-12 474,487 pos. (1.39%)	2019-02-12 to 2020-06-08 71,835 pos. (0.21%) 33,621,226 neg. (98.40%)

Note: “pos.” refers to positive sample, while “neg.” refers to negative sample.

▲所用数据集统计

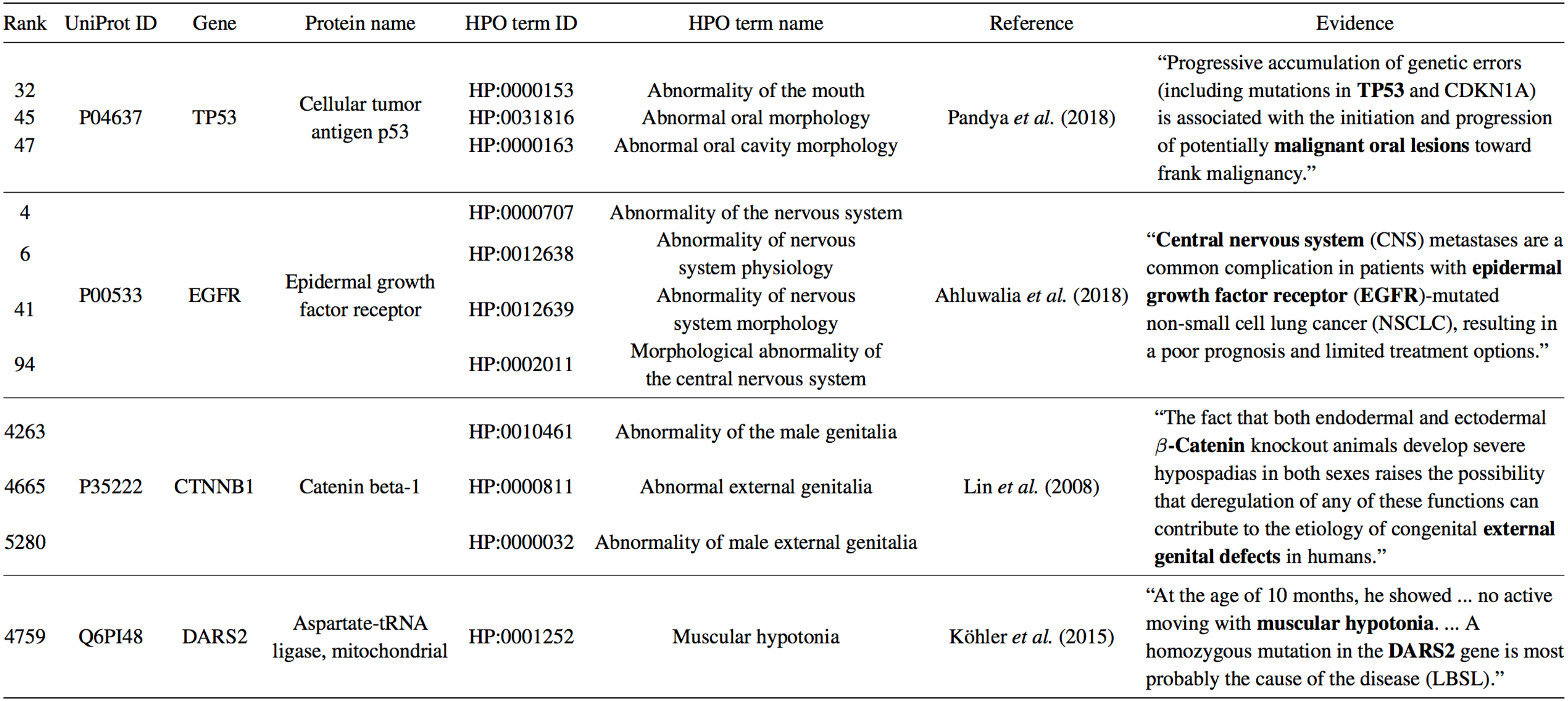
▼性能评估结果



Method	AUC	AUPR
LP	0.8916	0.0461
DLP	0.8913	0.0472
tlDLP-BP	0.8900	0.0472
tlDLP-MF	0.8885	0.0471
SMC	0.8326	0.0224
AiProAnnotator	0.8404	0.0211
HPOFiller	**0.9013**	**0.0483**

案例分析之一：有文献佐证的前部预测结果

发现尚未录入的HPO标注



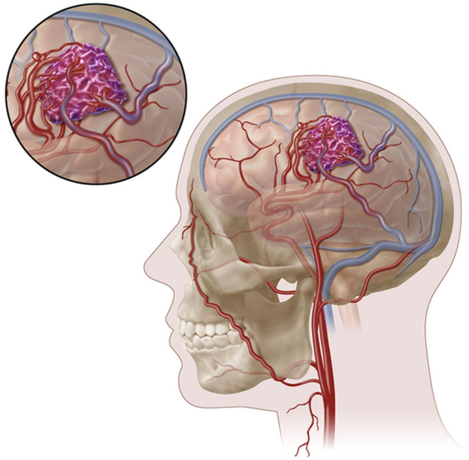
Rank	UniProt ID	Gene	Protein name	HPO term ID	HPO term name	Reference	Evidence
32 45 47	P04637	TP53	Cellular tumor antigen p53	HP:0000153 HP:0031816 HP:0000163	Abnormality of the mouth Abnormal oral morphology Abnormal oral cavity morphology	Pandya *et al.* (2018)	"Progressive accumulation of genetic errors (including mutations in **TP53** and CDKN1A) is associated with the initiation and progression of potentially **malignant oral lesions** toward frank malignancy."
4 6 41 94	P00533	EGFR	Epidermal growth factor receptor	HP:0000707 HP:0012638 HP:0012639 HP:0002011	Abnormality of the nervous system Abnormality of nervous system physiology Abnormality of nervous system morphology Morphological abnormality of the central nervous system	Ahluwalia *et al.* (2018)	"**Central nervous system** (CNS) metastases are a common complication in patients with **epidermal growth factor receptor** (**EGFR**)-mutated non-small cell lung cancer (NSCLC), resulting in a poor prognosis and limited treatment options."
4263 4665 5280	P35222	CTNNB1	Catenin beta-1	HP:0010461 HP:0000811 HP:0000032	Abnormality of the male genitalia Abnormal external genitalia Abnormality of male external genitalia	Lin *et al.* (2008)	"The fact that both endodermal and ectodermal **β-Catenin** knockout animals develop severe hypospadias in both sexes raises the possibility that deregulation of any of these functions can contribute to the etiology of congenital **external genital defects** in humans."
4759	Q6PI48	DARS2	Aspartate-tRNA ligase, mitochondrial	HP:0001252	Muscular hypotonia	Köhler *et al.* (2015)	"At the age of 10 months, he showed ... no active moving with **muscular hypotonia**. ... A homozygous mutation in the **DARS2** gene is most probably the cause of the disease (LBSL)."

案例分析之二：寻找疾病-基因关联的应用

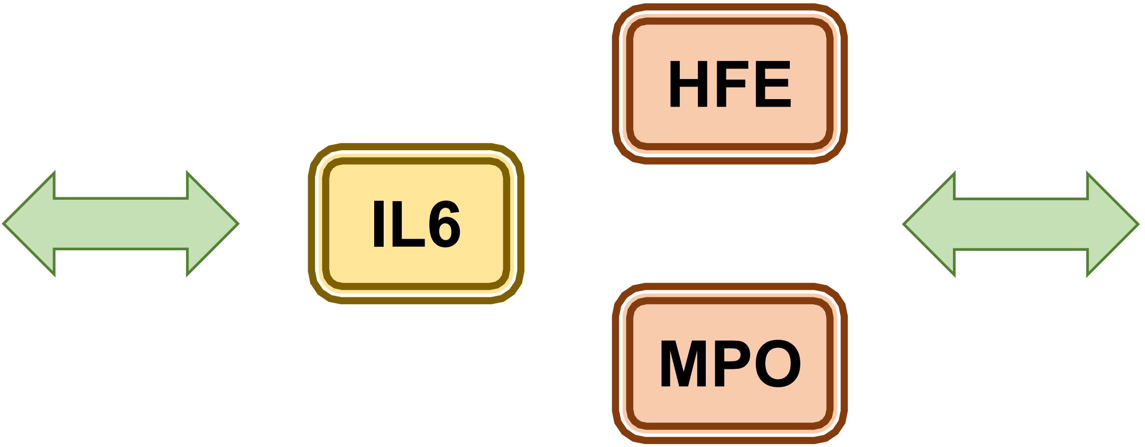
发现新增的疾病—基因关联

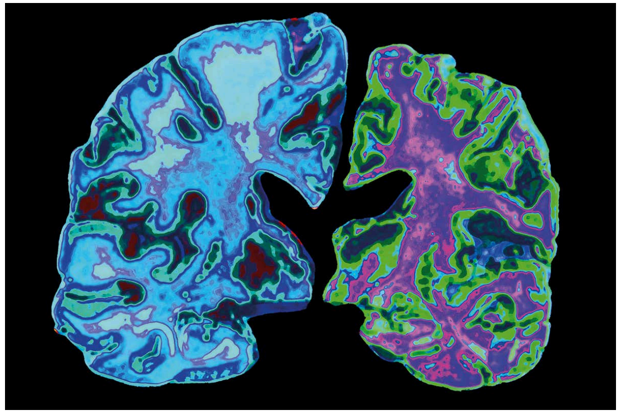
Rank	Protein ID	Gene	Protein name	HPO term ID	HPO term name	Disease ID	Disease name
114	P05231	IL6	Interleukin-6	HP:0002408	Cerebral arteriovenous malformation	OMIM:108010	Arteriovenous malformations of the brain (BAVM)
1323	Q30201	HFE	Hereditary hemochromatosis protein	HP:0000726	Dementia	OMIM:104300	Alzheimer disease (AD)
4032	P05164	MPO	Myeloperoxidase	HP:0002423	Long-tract signs		

Note: 'HPO term' refers to the predicted missing HPO annotation of corresponding protein by HPOFiller.



Arteriovenous malformations of the brain (BAVM)





Alzheimer disease (AD)

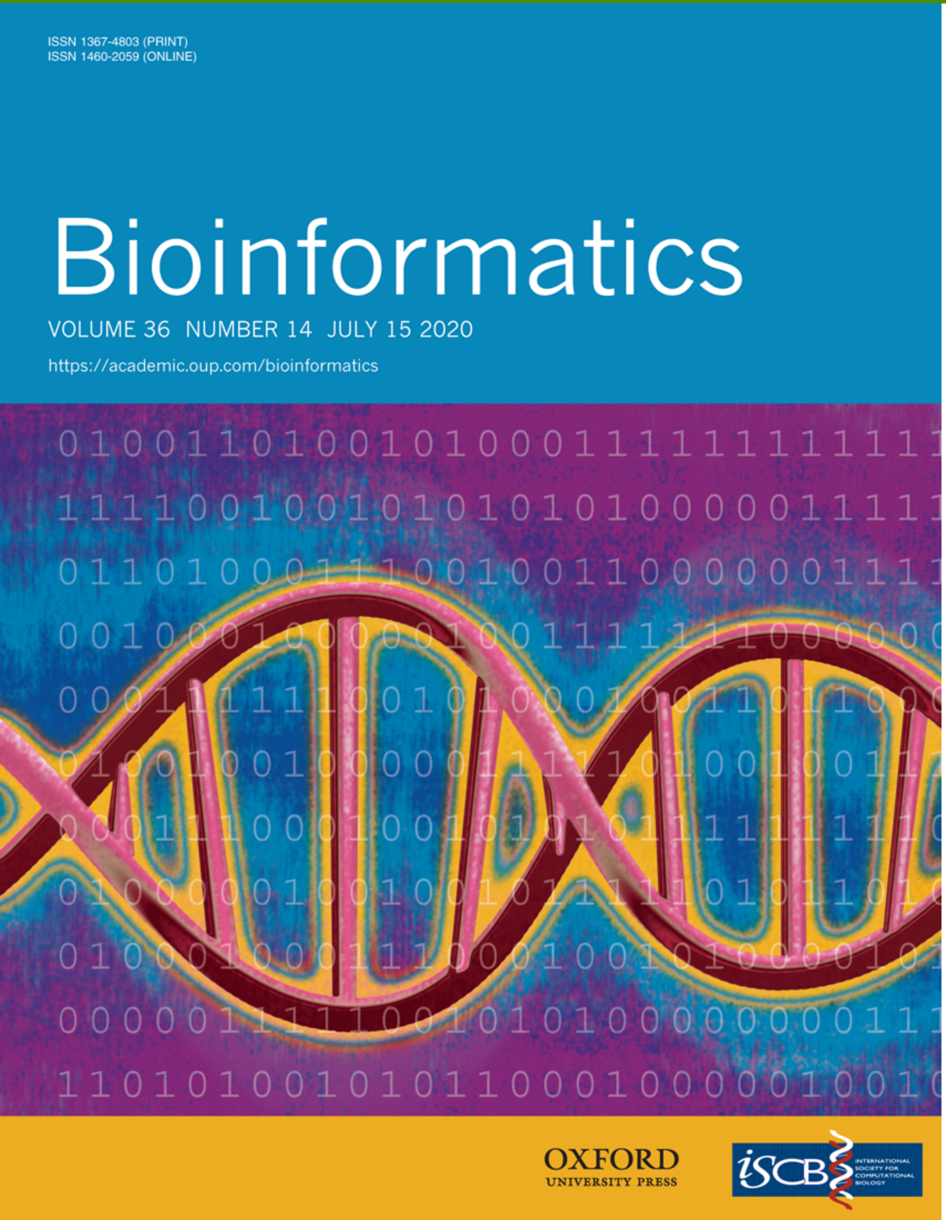
小结

- 我们提出了第一个基于图卷积网络的缺失HPO注释预测模型HPOFiller算法，使用S-GCN和Bi-GCN两种图卷积网络模块，从蛋白质互作网络、HPO语义相似度网络和蛋白质–HPO术语二分网络中充分捕捉非线性关系和高阶拓扑结构。

- 我们使用ε-增强损失函数缓解标签不平衡对训练带来的影响。

- 我们设计了极为严格的评估流程以避免潜在的信息泄露。实验结果显示，HPOFiller显著优于基于标签传播和矩阵分解的对比方法。

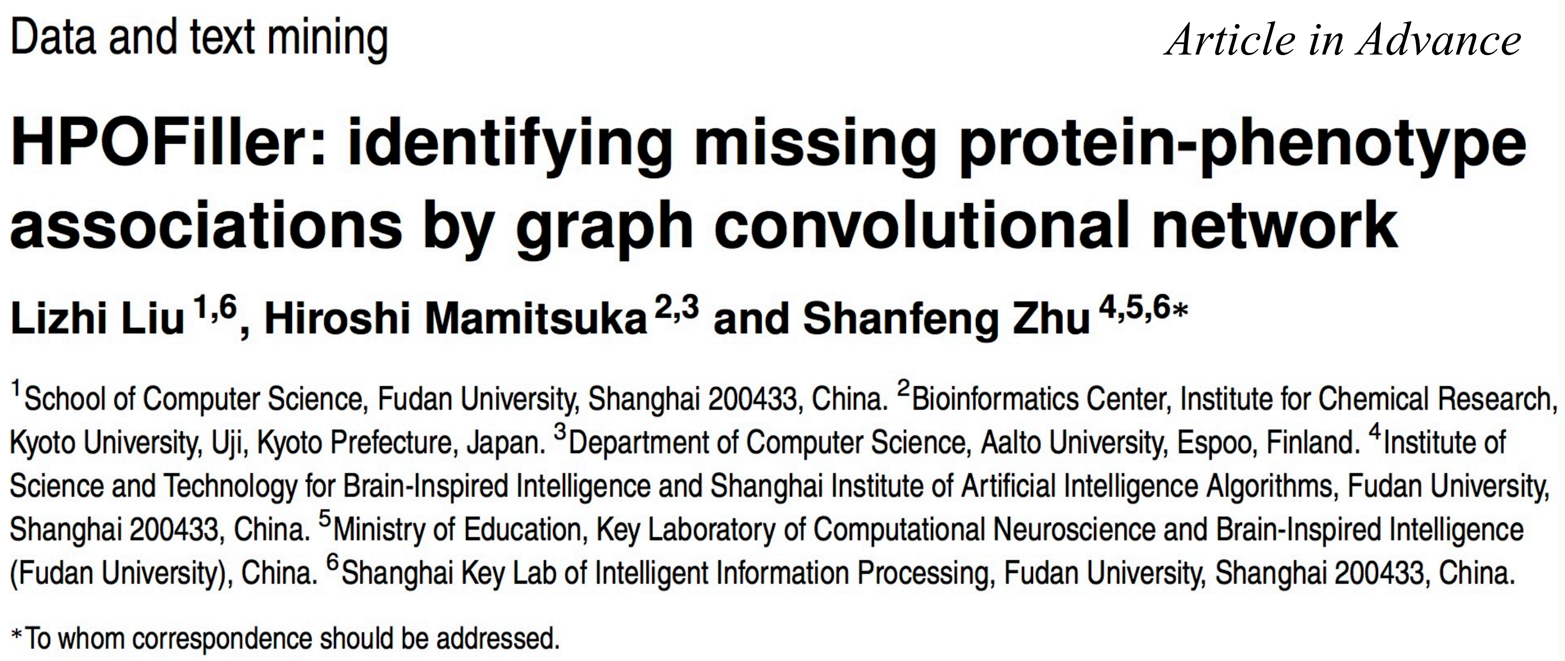
论文发表



ISSN 1367-4803 (PRINT)
ISSN 1460-2059 (ONLINE)

Bioinformatics

VOLUME 36 NUMBER 14 JULY 15 2020
https://academic.oup.com/bioinformatics



Data and text mining *Article in Advance*

HPOFiller: identifying missing protein-phenotype associations by graph convolutional network

Lizhi Liu [1,6], **Hiroshi Mamitsuka** [2,3] **and Shanfeng Zhu** [4,5,6*]

[1] School of Computer Science, Fudan University, Shanghai 200433, China. [2] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto Prefecture, Japan. [3] Department of Computer Science, Aalto University, Espoo, Finland. [4] Institute of Science and Technology for Brain-Inspired Intelligence and Shanghai Institute of Artificial Intelligence Algorithms, Fudan University, Shanghai 200433, China. [5] Ministry of Education, Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (Fudan University), China. [6] Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China.

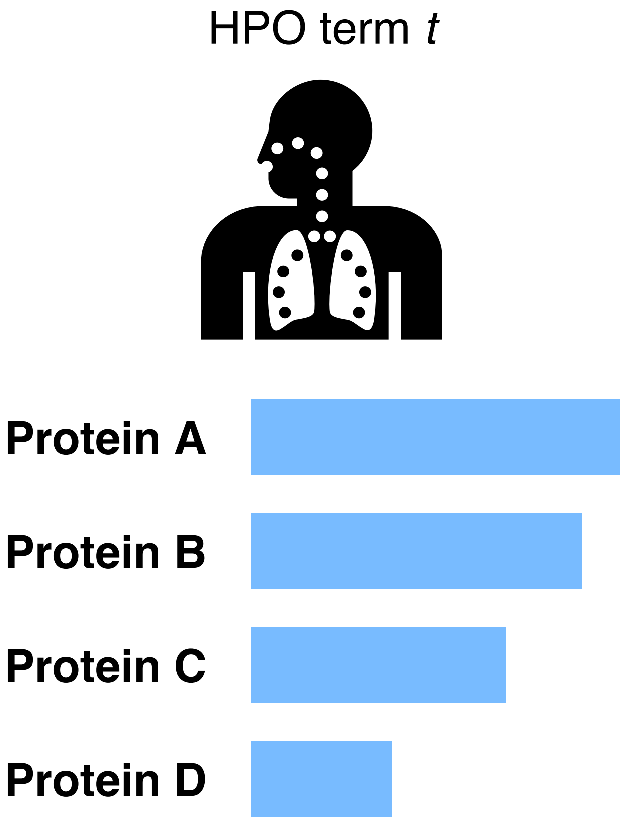
*To whom correspondence should be addressed.

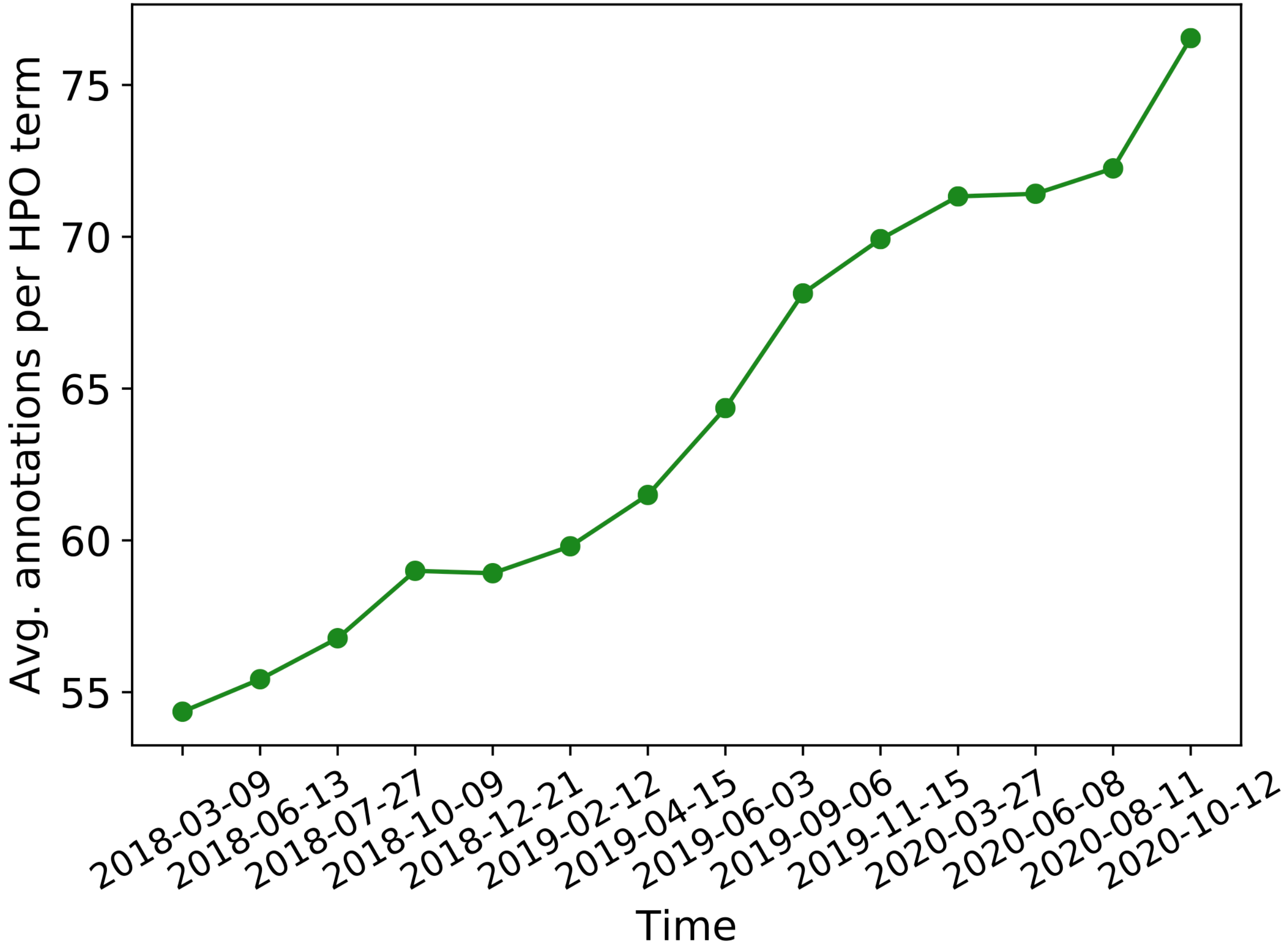
3

HPODNets

基于深度图卷积网络
整合多种互作网络
预测异常表型的相关蛋白质

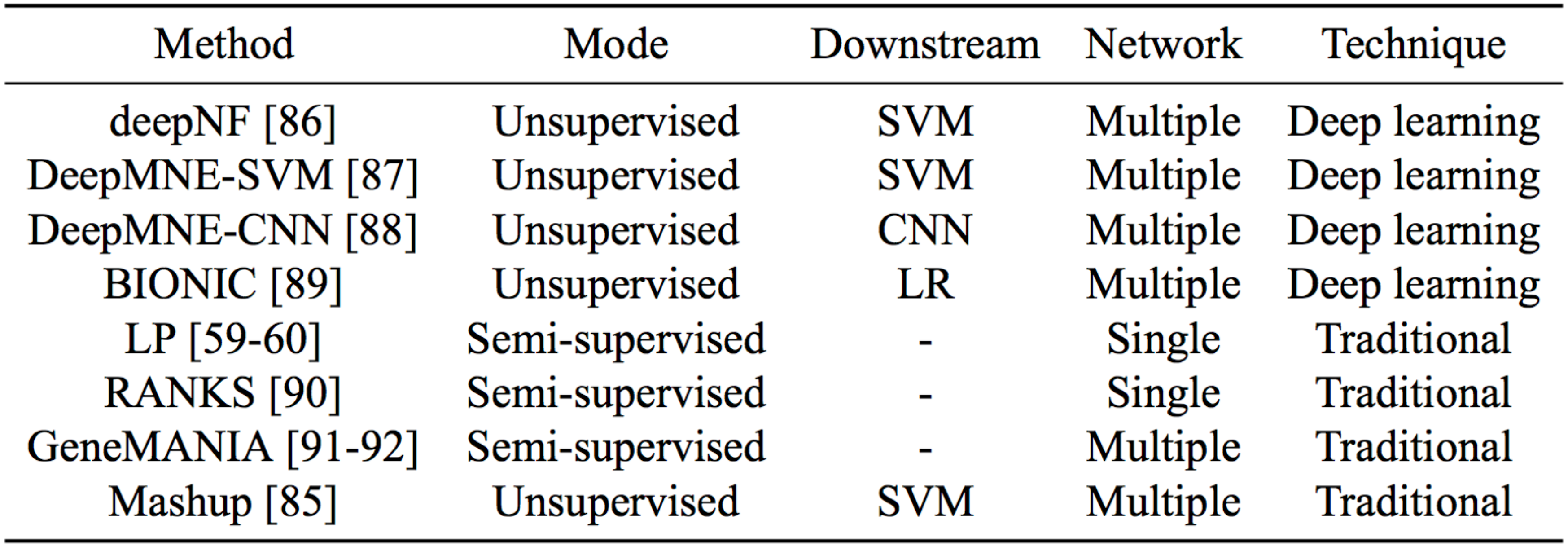
问题描述：对与异常表型相关候选蛋白质排序





研究目标：充分整合利用多种蛋白质互作网络信息，提高预测精度

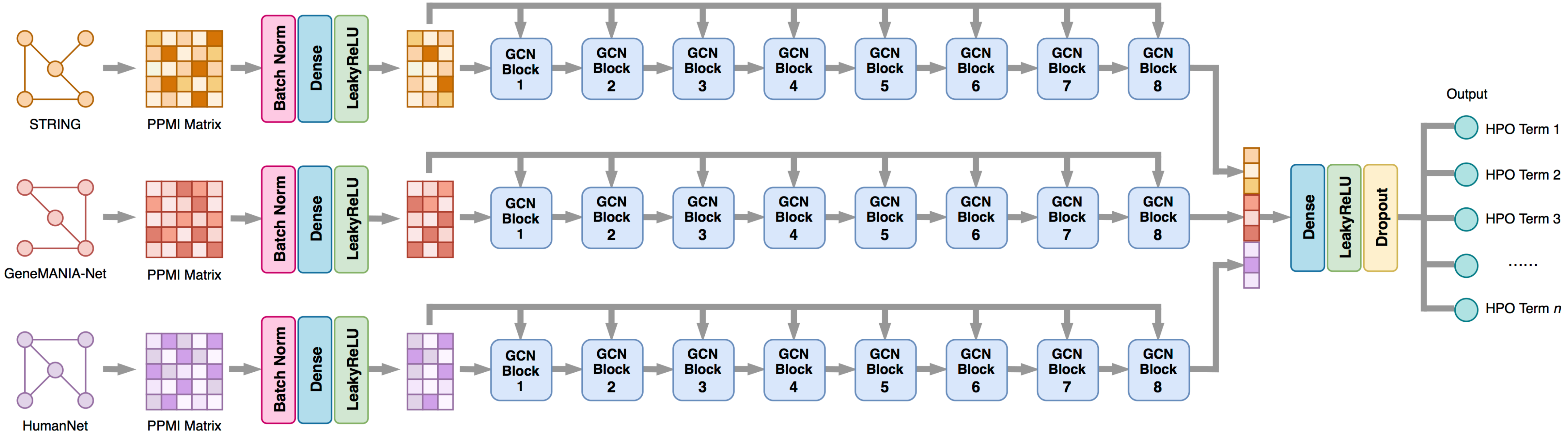
现有的以术语为中心的基因功能预测算法



Method	Mode	Downstream	Network	Technique
deepNF [86]	Unsupervised	SVM	Multiple	Deep learning
DeepMNE-SVM [87]	Unsupervised	SVM	Multiple	Deep learning
DeepMNE-CNN [88]	Unsupervised	CNN	Multiple	Deep learning
BIONIC [89]	Unsupervised	LR	Multiple	Deep learning
LP [59-60]	Semi-supervised	-	Single	Traditional
RANKS [90]	Semi-supervised	-	Single	Traditional
GeneMANIA [91-92]	Semi-supervised	-	Multiple	Traditional
Mashup [85]	Unsupervised	SVM	Multiple	Traditional

- 半监督学习算法需要将多种互作网络事先整合成一个复合网络，可能会丢失原来单个网络中的部分信息
- 无监督学习方法采用自编码器学习蛋白质嵌入表示，未融入已知的蛋白质功能注释这一监督信息，过于通用，缺乏针对蛋白质功能注释预测任务相应的判别能力
- 尚未提出以术语为中心的蛋白质异常表型预测模型，更无基于深度学习技术的监督学习算法

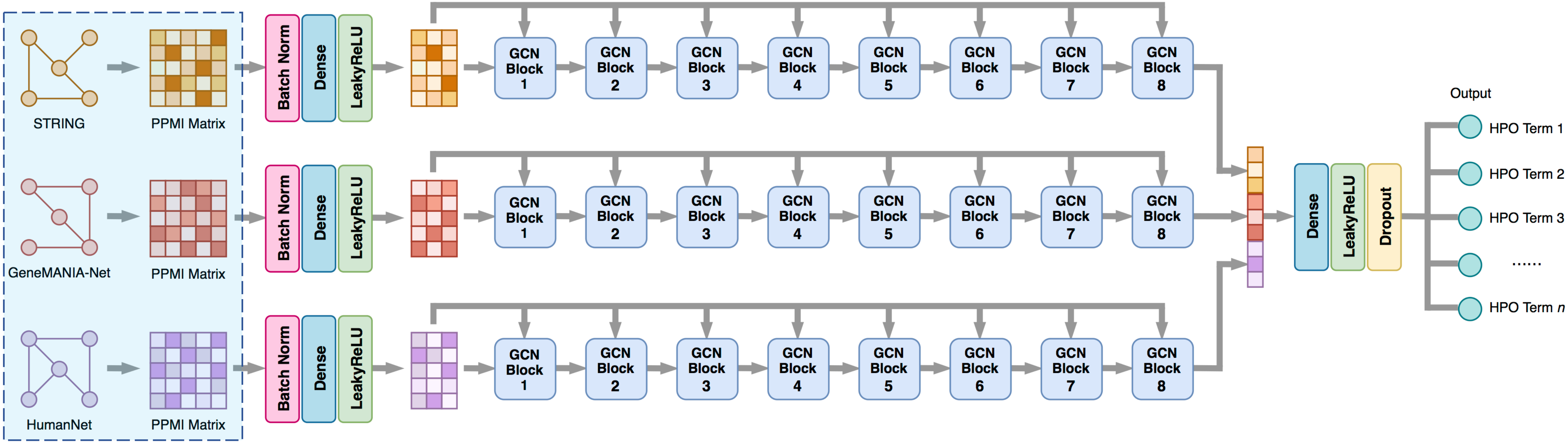
HPODNets — 基于深度图神经网络进行预测



关键点

- 提出第一个基于深度图卷积网络的半监督学习算法，整合多种蛋白质互作网络，以实现以术语为中心的蛋白质表型标注预测
- 在传统图卷积操作中引入初始表示和恒等映射，并合理排放图卷积、批标准化、随机失活和激活函数等组件，以缓解“过平滑”对性能的影响，并能充分捕捉网络中的低阶和高阶拓扑结构

HPODNets — 第一步：预处理与特征生成



- 甲、对蛋白质互作网络邻接矩阵进行**对称规范化**

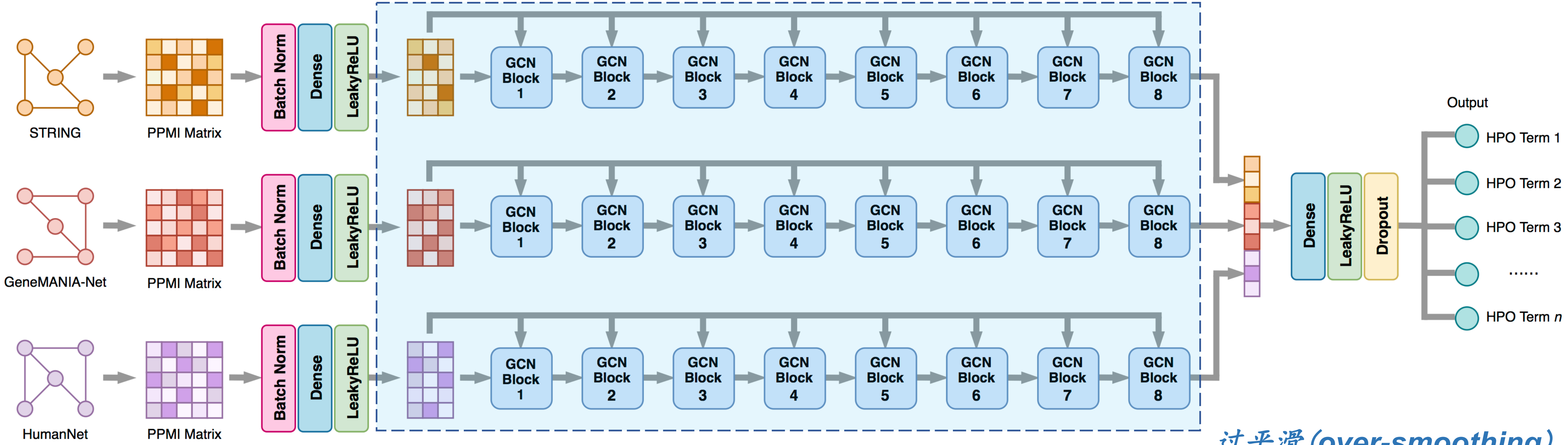
$$\bar{\mathbf{A}}_G = \mathbf{D}_G^{-\frac{1}{2}} \mathbf{A}_G \mathbf{D}_G^{-\frac{1}{2}}$$

- 乙、构建正值逐点互信息(**Positive Pointwise Mutual Information, PPMI**)矩阵，并作为特征向量

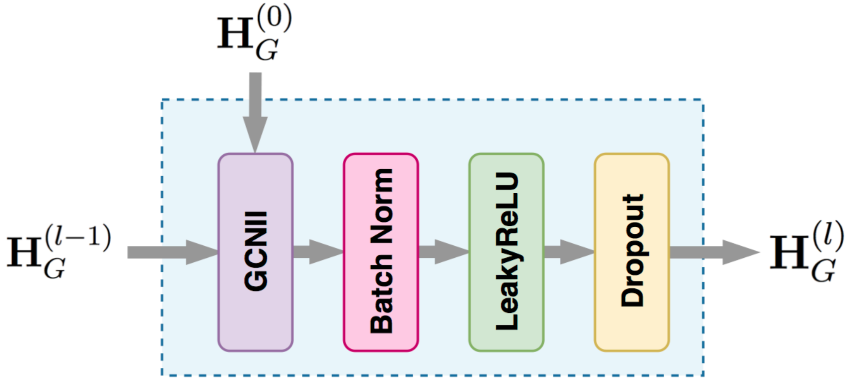
$$\mathbf{X}_{G,ij} = \max\left(0, \log_2\left(\frac{\bar{\mathbf{A}}_{G,ij}\sum_s\sum_t \bar{\mathbf{A}}_{G,st}}{\sum_s \bar{\mathbf{A}}_{G,sj}\sum_t \bar{\mathbf{A}}_{G,it}}\right)\right)$$

[Cao et al. AAAI, 2016]

HPODNets — 第二步：深度图卷积编码器



过平滑（*over-smoothing*）



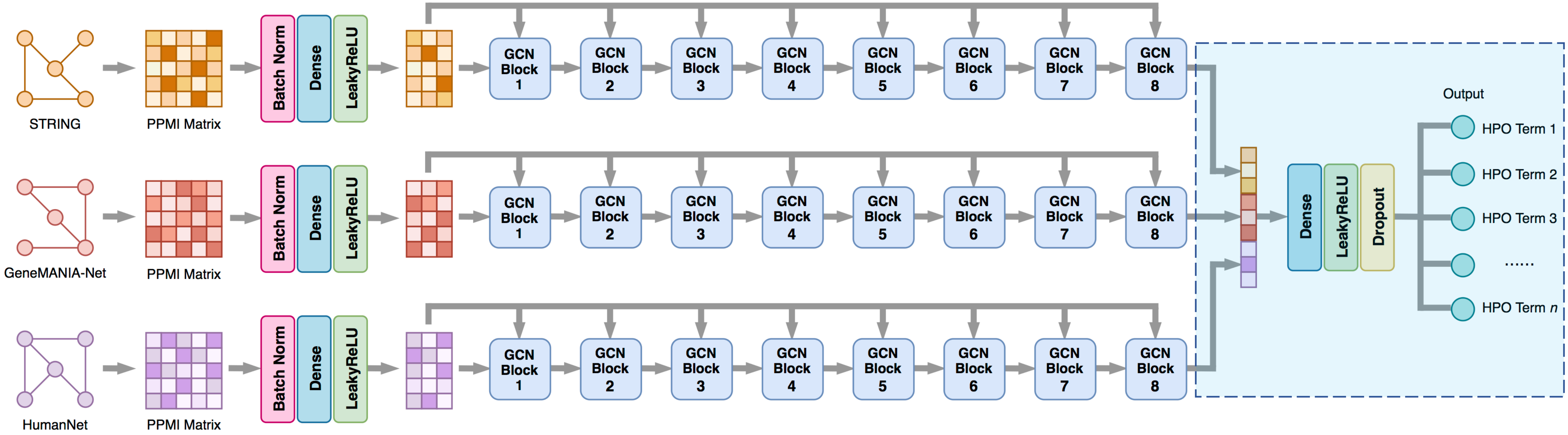
- 甲、在传统的图卷积操作中引入初始表示（Initial representation）和恒等映射（Identity mapping），构成GCNII层：[Chen et al. ICML, 2020]

$$\mathbf{H}_G^{(l)} = \sigma\left(\left((1-\alpha)\tilde{\mathbf{P}}_G\mathbf{H}_G^{(l-1)} + \alpha\mathbf{H}_G^{(0)}\right)\left((1-\beta_l)\mathbf{I} + \beta_l\mathbf{W}^{(l)}\right)\right)$$

- 乙、图卷积模块内各组件的排列顺序很重要！ [Li et al. 2020]

GCNII → Batch Normalization → LeakyReLU → Dropout

HPODNets — 第三步：融合与输出预测结果



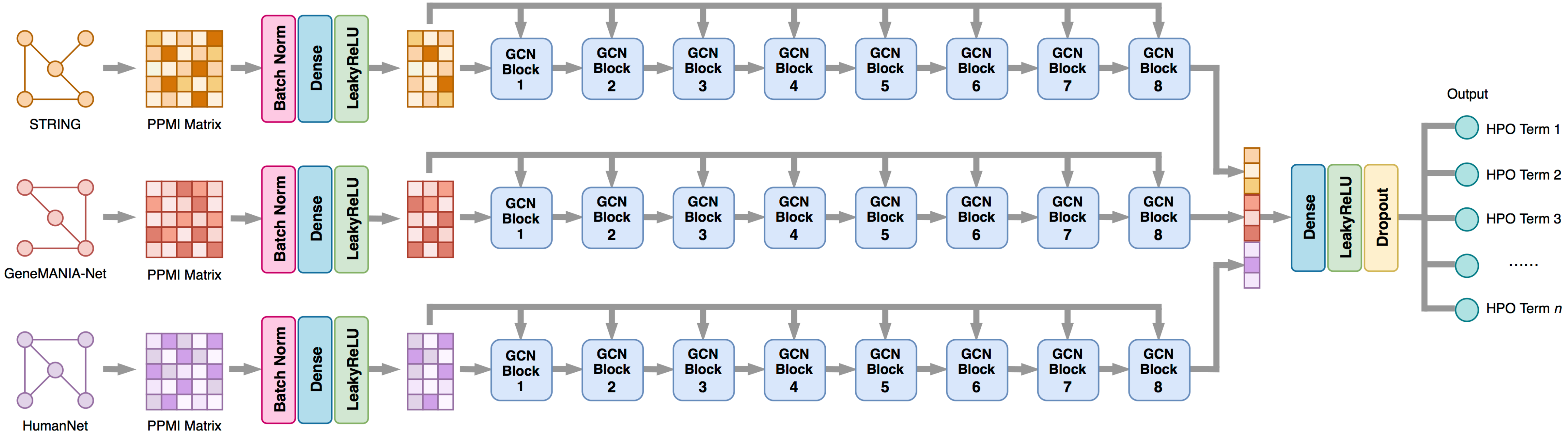
- 甲、将各分支得到的隐含表示拼接起来组成融合表示

$$\mathbf{H} = \mathbf{H}_{STR}^{(L)} \| \mathbf{H}_{GM}^{(L)} \| \mathbf{H}_{HN}^{(L)}$$

- 乙、将最终的蛋白质嵌入表示分配给各输出神经元，并为相应的**HPO**术语输出预测打分

$$\hat{y}_{i,j} = \operatorname{sigmoid}\left(\mathbf{e}_i \cdot \boldsymbol{\theta}_j\right) = \frac{1}{1 + \exp(-\mathbf{e}_i \cdot \boldsymbol{\theta}_j)}$$

模型训练

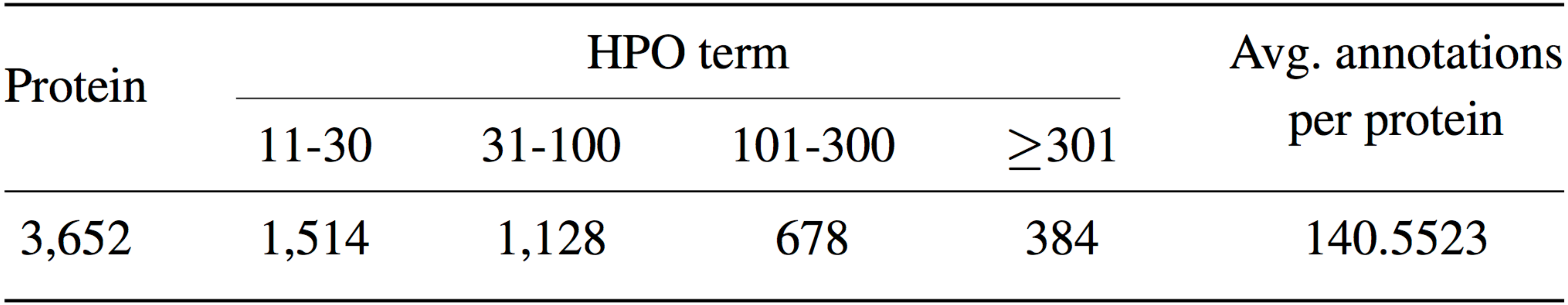


Binary cross-entropy loss

$$\mathcal{L} = \sum_{t=1}^{n} \mathcal{L}_t = -\sum_{t=1}^{n}\sum_{i=1}^{l}[\gamma_t y_{i,t} \log(\hat{y}_{i,t}) + (1 - y_{i,t}) \log(1 - \hat{y}_{i,t})]$$

Adjustment weight: $\gamma_t = \frac{m_t^-}{m_t^+}$ *类别不平衡(class-imbalance)*

评估之一：交叉验证

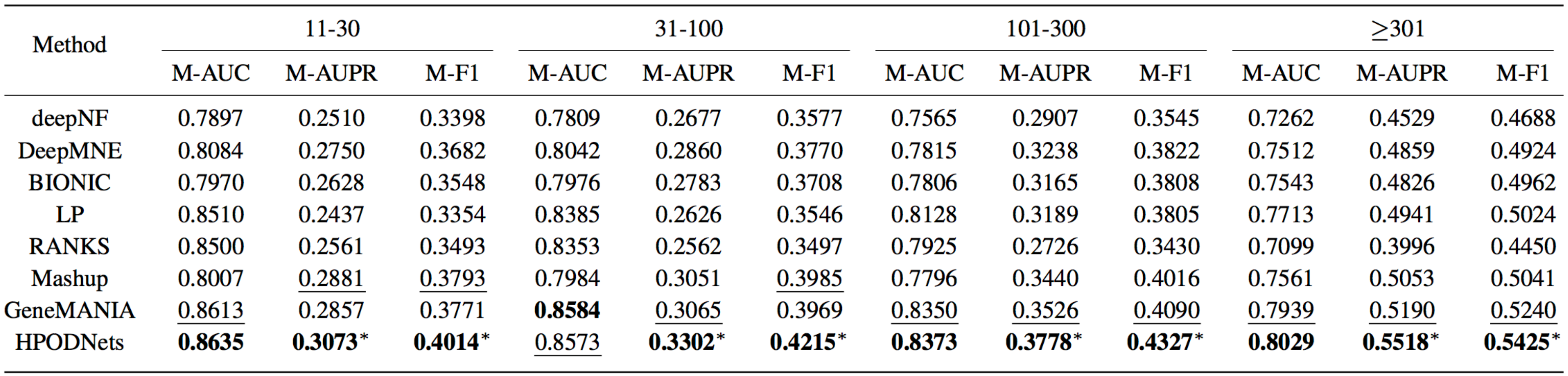


Protein	HPO term				Avg. annotations per protein
	11-30	31-100	101-300	≥301	
3,652	1,514	1,128	678	384	140.5523

Version: 2020-03-27

实验结果之交叉验证 —— 对比

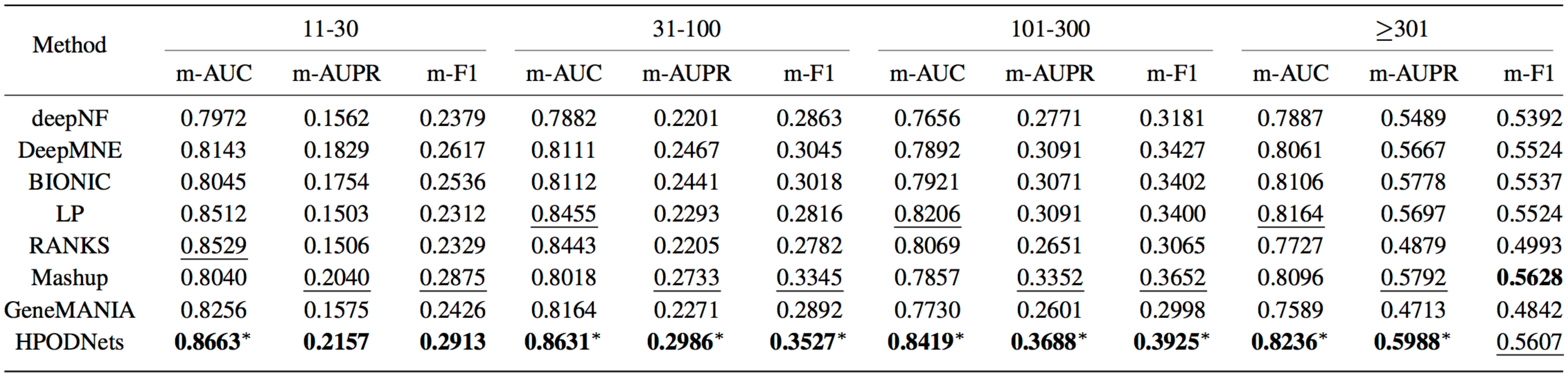
Table 3. Cross-validation performance under macro-averaged metrics



Method	11-30			31-100			101-300			≥301		
	M-AUC	M-AUPR	M-F1	M-AUC	M-AUPR	M-F1	M-AUC	M-AUPR	M-F1	M-AUC	M-AUPR	M-F1
deepNF	0.7897	0.2510	0.3398	0.7809	0.2677	0.3577	0.7565	0.2907	0.3545	0.7262	0.4529	0.4688
DeepMNE	0.8084	0.2750	0.3682	0.8042	0.2860	0.3770	0.7815	0.3238	0.3822	0.7512	0.4859	0.4924
BIONIC	0.7970	0.2628	0.3548	0.7976	0.2783	0.3708	0.7806	0.3165	0.3808	0.7543	0.4826	0.4962
LP	0.8510	0.2437	0.3354	0.8385	0.2626	0.3546	0.8128	0.3189	0.3805	0.7713	0.4941	0.5024
RANKS	0.8500	0.2561	0.3493	0.8353	0.2562	0.3497	0.7925	0.2726	0.3430	0.7099	0.3996	0.4450
Mashup	0.8007	0.2881	0.3793	0.7984	0.3051	0.3985	0.7796	0.3440	0.4016	0.7561	0.5053	0.5041
GeneMANIA	0.8613	0.2857	0.3771	**0.8584**	0.3065	0.3969	0.8350	0.3526	0.4090	0.7939	0.5190	0.5240
HPODNets	**0.8635**	**0.3073***	**0.4014***	0.8573	**0.3302***	**0.4215***	**0.8373**	**0.3778***	**0.4327***	**0.8029**	**0.5518***	**0.5425***

Notes: *Statistical significance ($P < 0.05$) by pairwise t-test. The boldface items in the table represent the best performance, and the runner-ups are underlined.

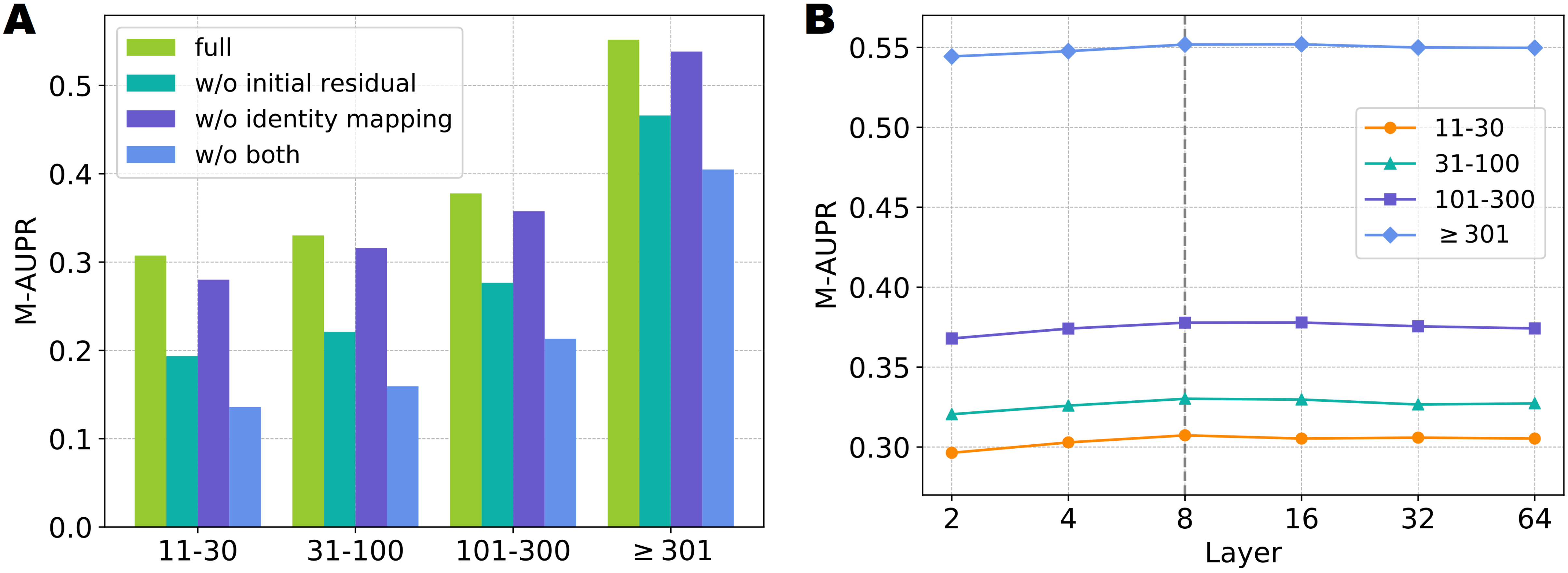
Table 4. Cross-validation performance under micro-averaged metrics



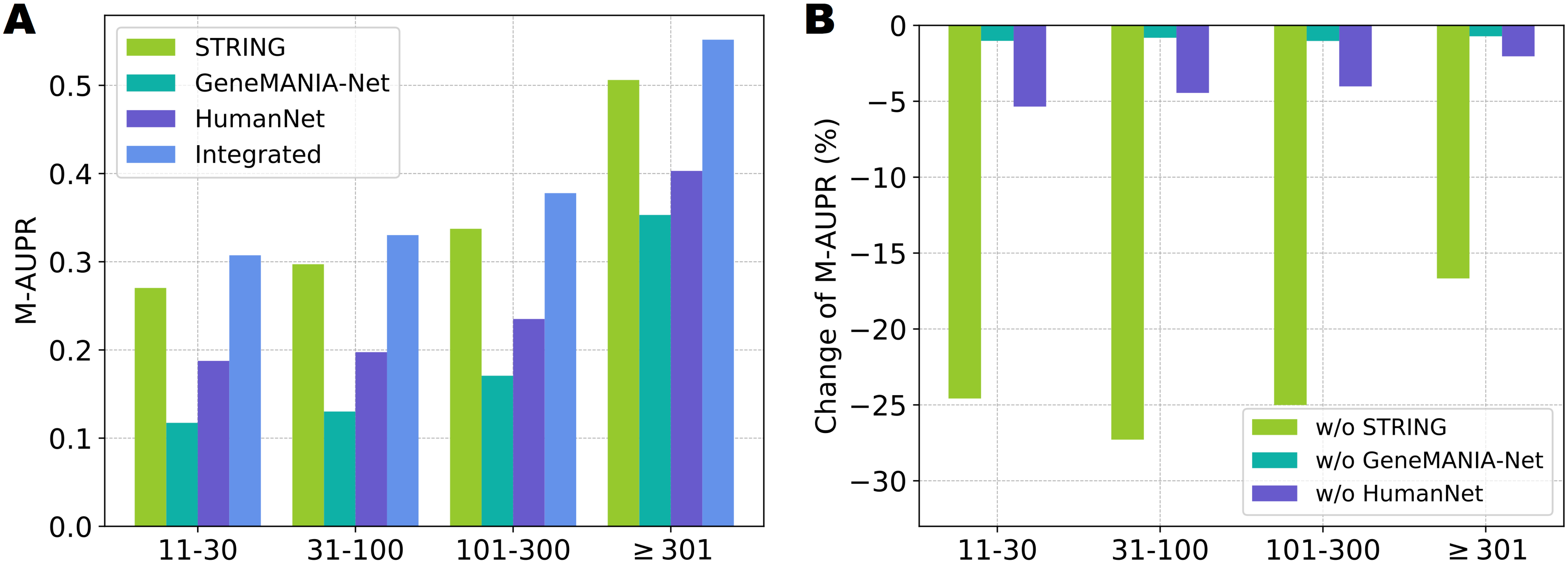
Method	11-30			31-100			101-300			≥301		
	m-AUC	m-AUPR	m-F1	m-AUC	m-AUPR	m-F1	m-AUC	m-AUPR	m-F1	m-AUC	m-AUPR	m-F1
deepNF	0.7972	0.1562	0.2379	0.7882	0.2201	0.2863	0.7656	0.2771	0.3181	0.7887	0.5489	0.5392
DeepMNE	0.8143	0.1829	0.2617	0.8111	0.2467	0.3045	0.7892	0.3091	0.3427	0.8061	0.5667	0.5524
BIONIC	0.8045	0.1754	0.2536	0.8112	0.2441	0.3018	0.7921	0.3071	0.3402	0.8106	0.5778	0.5537
LP	0.8512	0.1503	0.2312	0.8455	0.2293	0.2816	0.8206	0.3091	0.3400	0.8164	0.5697	0.5524
RANKS	0.8529	0.1506	0.2329	0.8443	0.2205	0.2782	0.8069	0.2651	0.3065	0.7727	0.4879	0.4993
Mashup	0.8040	0.2040	0.2875	0.8018	0.2733	0.3345	0.7857	0.3352	0.3652	0.8096	0.5792	**0.5628**
GeneMANIA	0.8256	0.1575	0.2426	0.8164	0.2271	0.2892	0.7730	0.2601	0.2998	0.7589	0.4713	0.4842
HPODNets	**0.8663***	**0.2157**	**0.2913**	**0.8631***	**0.2986***	**0.3527***	**0.8419***	**0.3688***	**0.3925***	**0.8236***	**0.5988***	0.5607

Notes: *Statistical significance ($P < 0.05$) by pairwise t-test. The boldface items in the table represent the best performance, and the runner-ups are underlined.

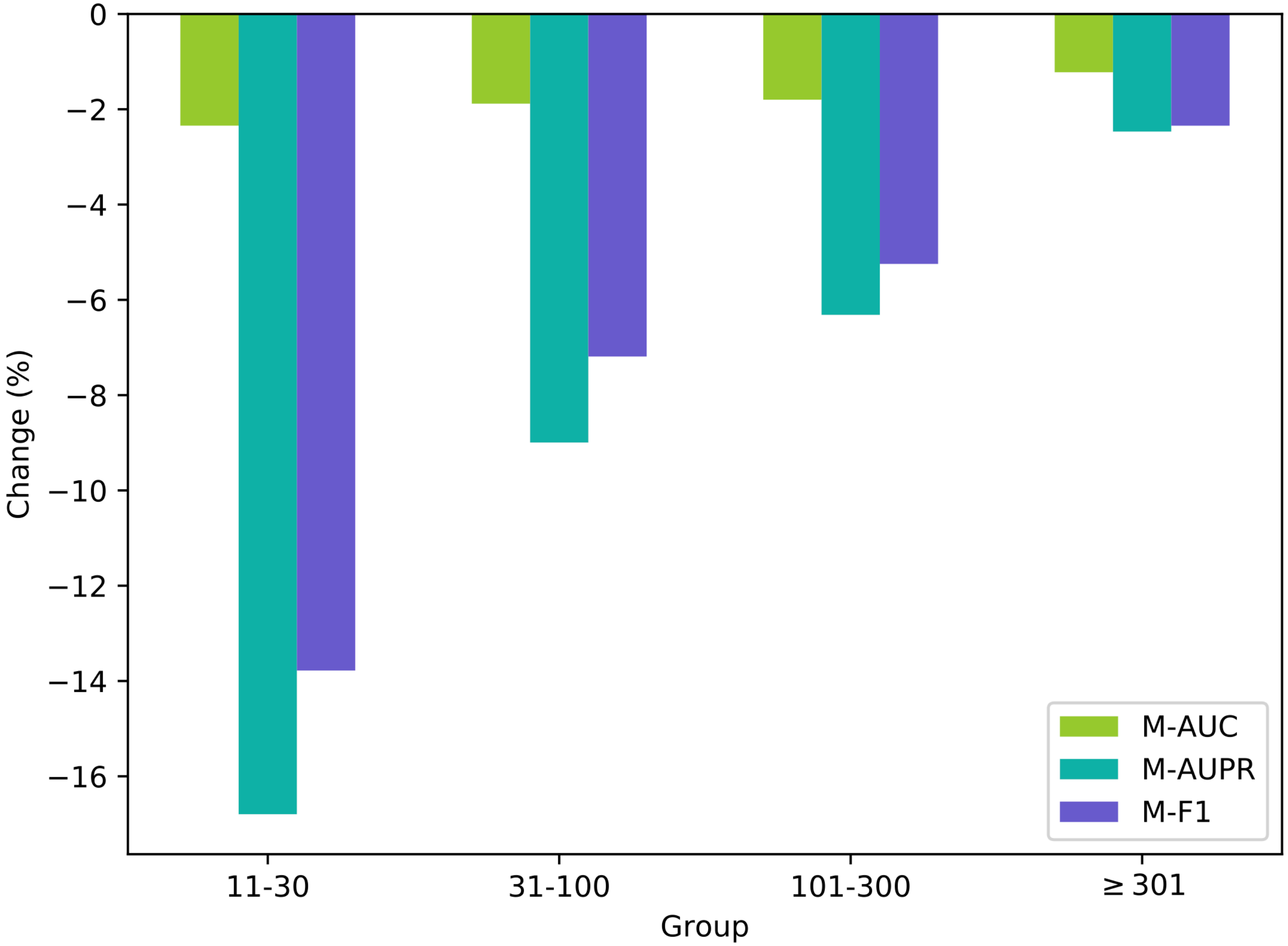
为什么要使用GCNII图卷积模块？



为什么要利用多种蛋白质互作网络？

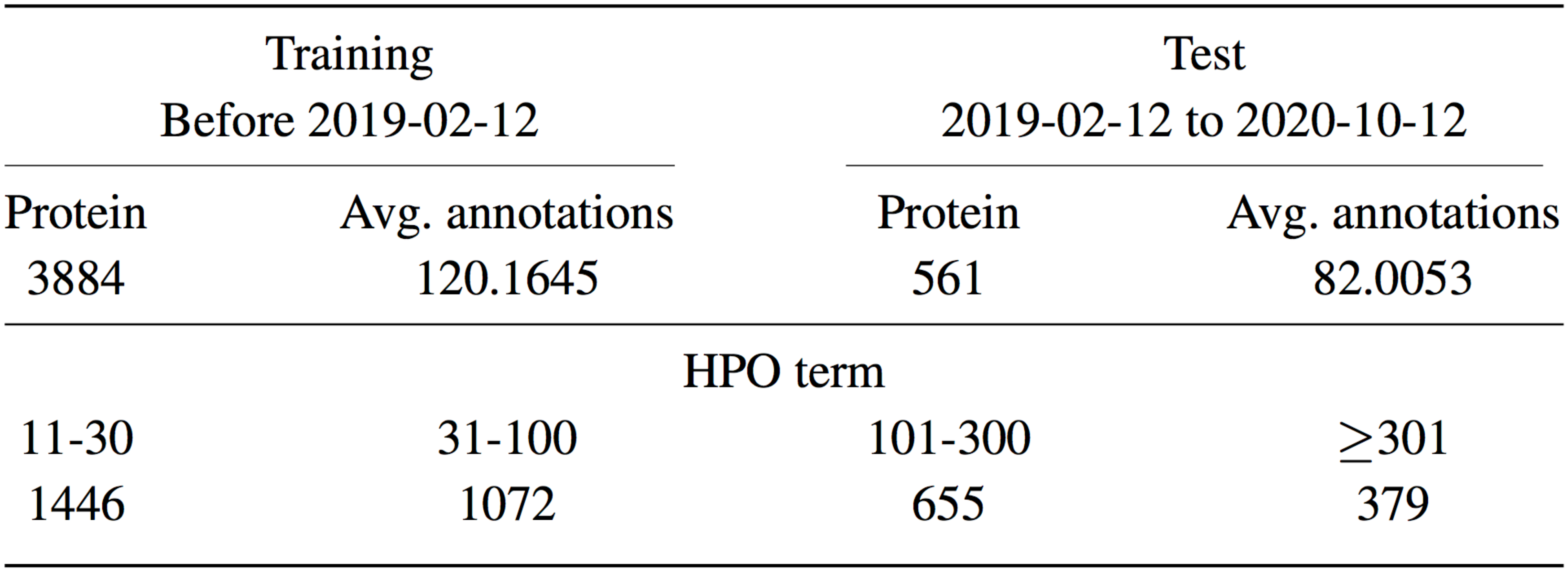


为什么要在损失函数中增加调节权值?



移除损失函数中的调节权值后，特别是低频**HPO**术语组上的性能显著下降

评估之二：依时间划分验证



Training		Test	
Before 2019-02-12		2019-02-12 to 2020-10-12	
Protein	Avg. annotations	Protein	Avg. annotations
3884	120.1645	561	82.0053
HPO term			
11-30	31-100	101-300	≥301
1446	1072	655	379

实验结果之依时间划分验证

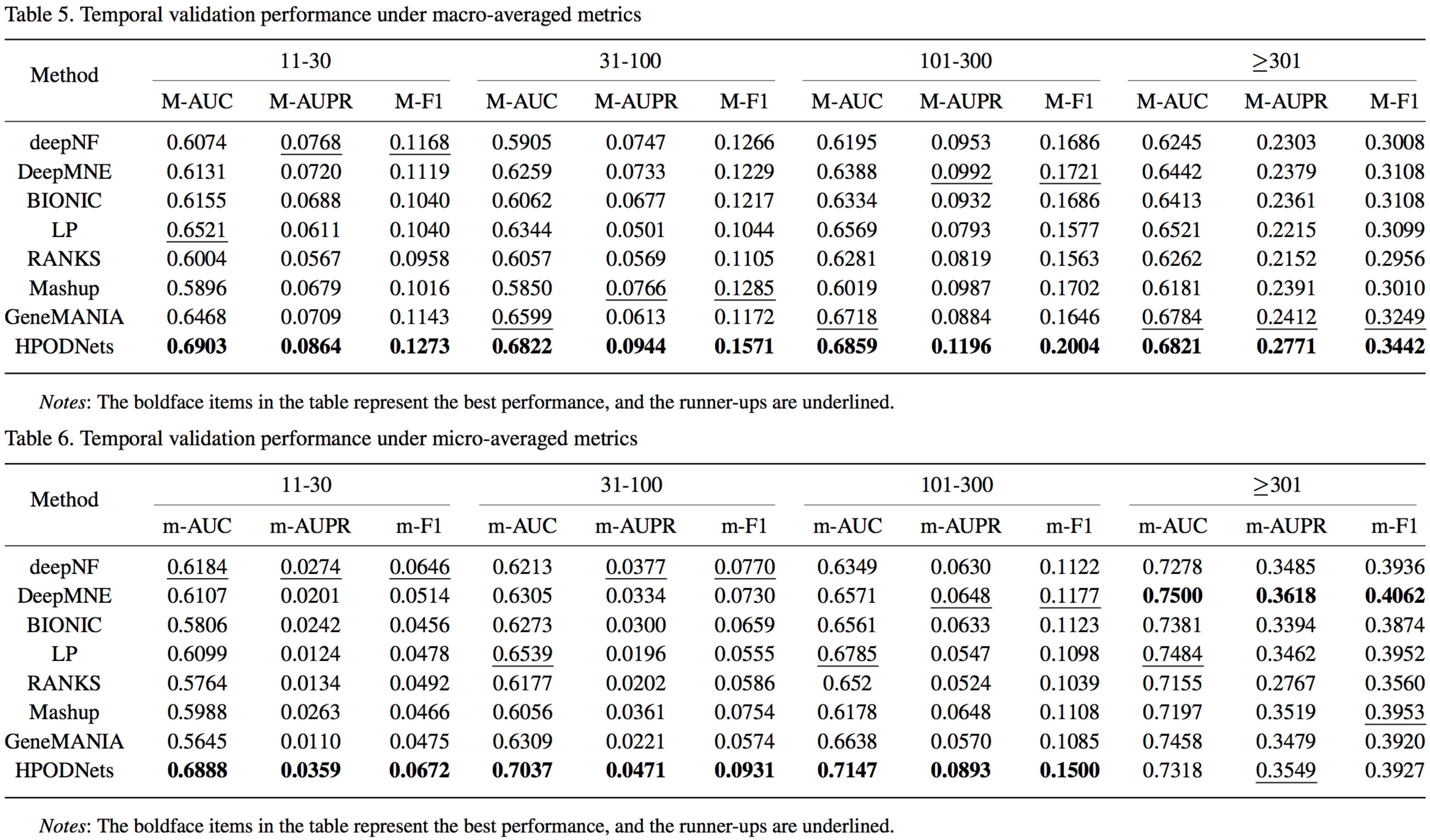


Table 5. Temporal validation performance under macro-averaged metrics

Method	11-30			31-100			101-300			≥301		
	M-AUC	M-AUPR	M-F1	M-AUC	M-AUPR	M-F1	M-AUC	M-AUPR	M-F1	M-AUC	M-AUPR	M-F1
deepNF	0.6074	0.0768	0.1168	0.5905	0.0747	0.1266	0.6195	0.0953	0.1686	0.6245	0.2303	0.3008
DeepMNE	0.6131	0.0720	0.1119	0.6259	0.0733	0.1229	0.6388	0.0992	0.1721	0.6442	0.2379	0.3108
BIONIC	0.6155	0.0688	0.1040	0.6062	0.0677	0.1217	0.6334	0.0932	0.1686	0.6413	0.2361	0.3108
LP	0.6521	0.0611	0.1040	0.6344	0.0501	0.1044	0.6569	0.0793	0.1577	0.6521	0.2215	0.3099
RANKS	0.6004	0.0567	0.0958	0.6057	0.0569	0.1105	0.6281	0.0819	0.1563	0.6262	0.2152	0.2956
Mashup	0.5896	0.0679	0.1016	0.5850	0.0766	0.1285	0.6019	0.0987	0.1702	0.6181	0.2391	0.3010
GeneMANIA	0.6468	0.0709	0.1143	0.6599	0.0613	0.1172	0.6718	0.0884	0.1646	0.6784	0.2412	0.3249
HPODNets	**0.6903**	**0.0864**	**0.1273**	**0.6822**	**0.0944**	**0.1571**	**0.6859**	**0.1196**	**0.2004**	**0.6821**	**0.2771**	**0.3442**

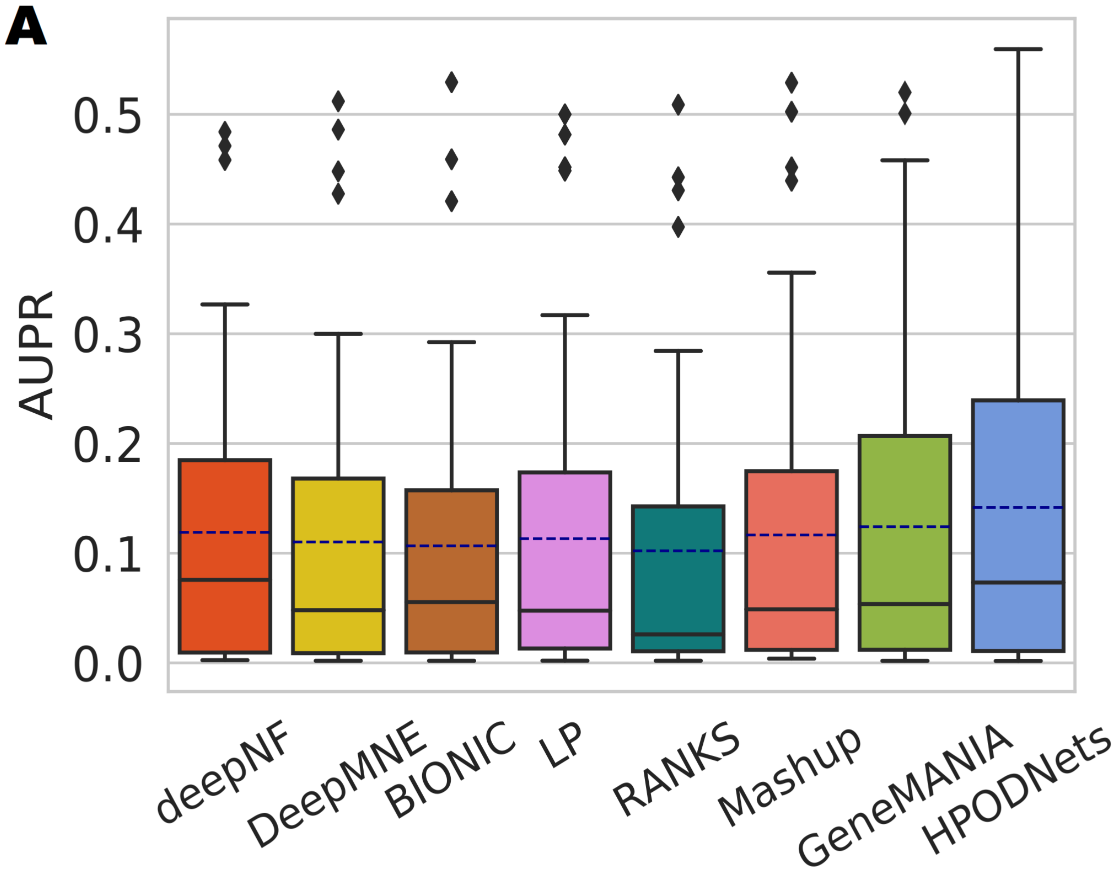
Notes: The boldface items in the table represent the best performance, and the runner-ups are underlined.

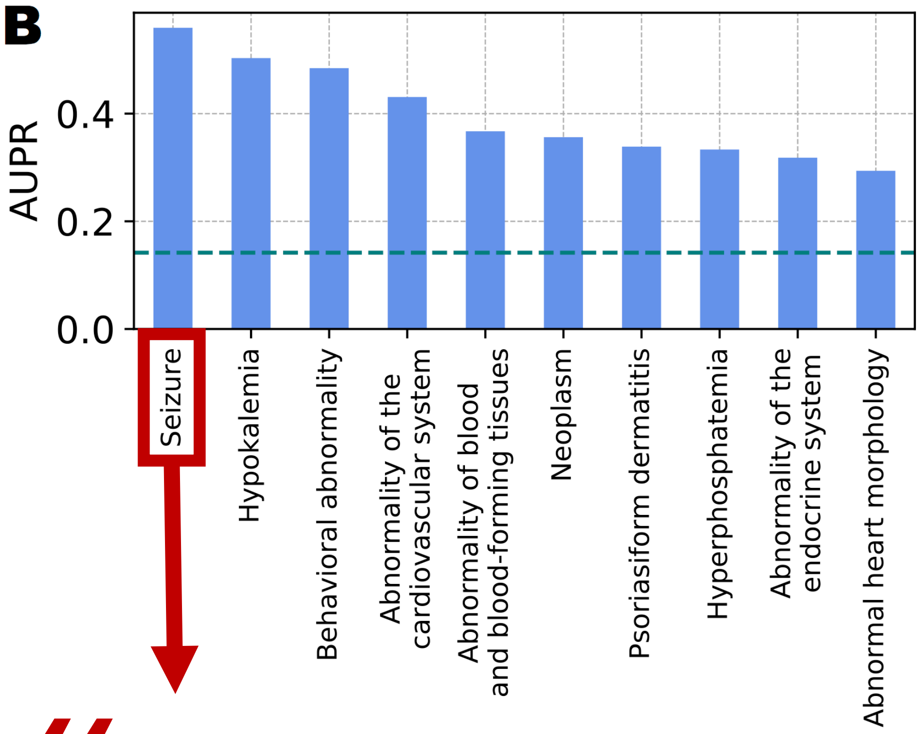
Table 6. Temporal validation performance under micro-averaged metrics

Method	11-30			31-100			101-300			≥301		
	m-AUC	m-AUPR	m-F1	m-AUC	m-AUPR	m-F1	m-AUC	m-AUPR	m-F1	m-AUC	m-AUPR	m-F1
deepNF	0.6184	0.0274	0.0646	0.6213	0.0377	0.0770	0.6349	0.0630	0.1122	0.7278	0.3485	0.3936
DeepMNE	0.6107	0.0201	0.0514	0.6305	0.0334	0.0730	0.6571	0.0648	0.1177	**0.7500**	**0.3618**	**0.4062**
BIONIC	0.5806	0.0242	0.0456	0.6273	0.0300	0.0659	0.6561	0.0633	0.1123	0.7381	0.3394	0.3874
LP	0.6099	0.0124	0.0478	0.6539	0.0196	0.0555	0.6785	0.0547	0.1098	0.7484	0.3462	0.3952
RANKS	0.5764	0.0134	0.0492	0.6177	0.0202	0.0586	0.652	0.0524	0.1039	0.7155	0.2767	0.3560
Mashup	0.5988	0.0263	0.0466	0.6056	0.0361	0.0754	0.6178	0.0648	0.1108	0.7197	0.3519	0.3953
GeneMANIA	0.5645	0.0110	0.0475	0.6309	0.0221	0.0574	0.6638	0.0570	0.1085	0.7458	0.3479	0.3920
HPODNets	**0.6888**	**0.0359**	**0.0672**	**0.7037**	**0.0471**	**0.0931**	**0.7147**	**0.0893**	**0.1500**	0.7318	0.3549	0.3927

Notes: The boldface items in the table represent the best performance, and the runner-ups are underlined.

与冠状病毒感染相关的HPO术语上的预测性能





> ***“How does the COVID-19 cause seizure and epilepsy in patients? The potential mechanisms”***

[Nikbakht et al. Mult Scler Ralat Disord, 2020]

从*http://covidresearchtrials.com*上下载了**50**个与冠状病毒感染（*Coronavirus infection*）相关的**HPO**术语

[Alag, PLoS One, 2020]

HPODNets发现了新的与肺炎相关的致病基因

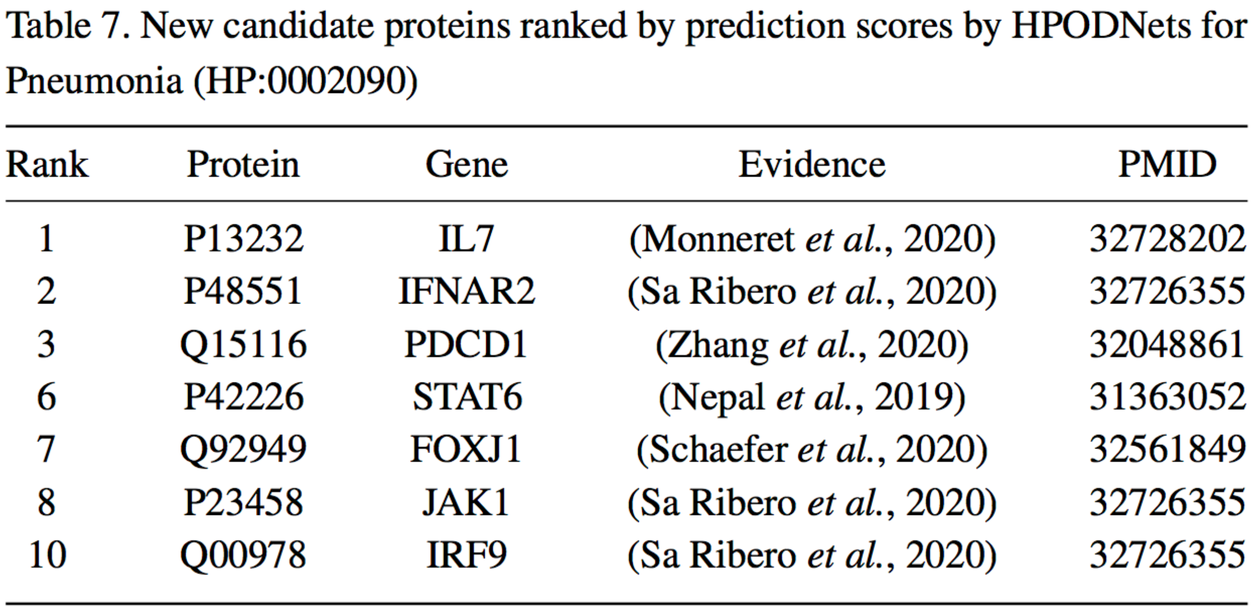
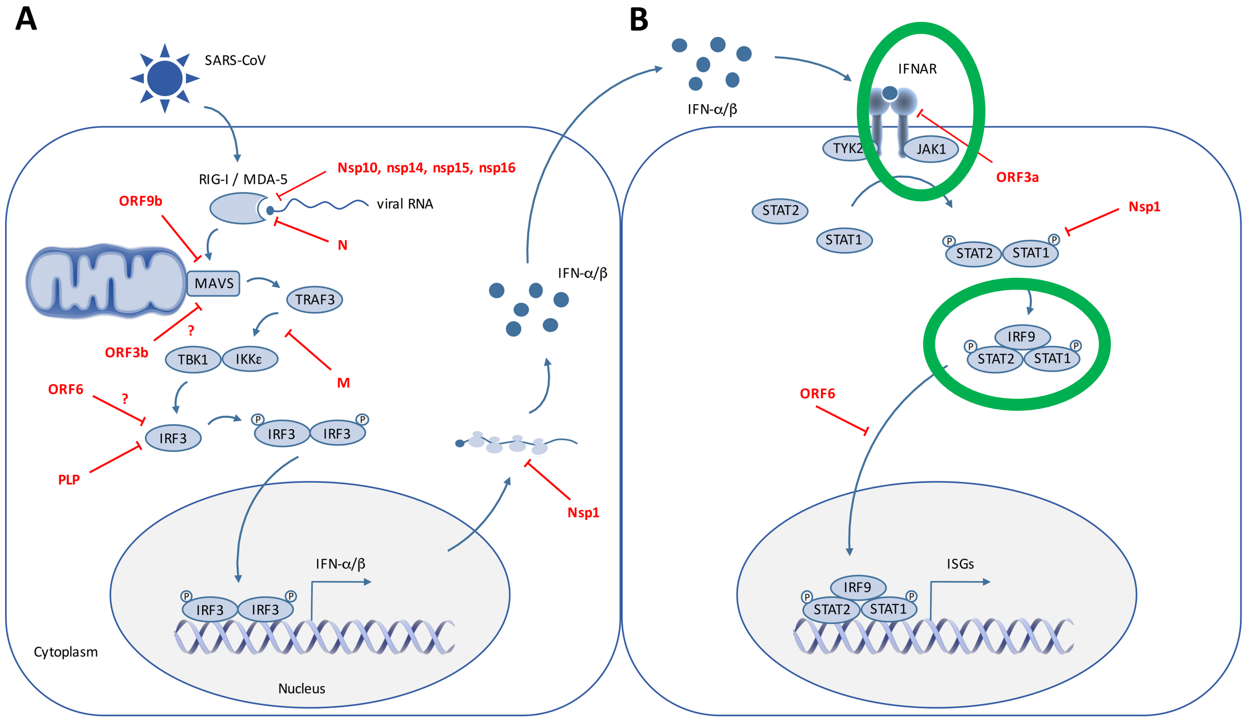


Table 7. New candidate proteins ranked by prediction scores by HPODNets for Pneumonia (HP:0002090)

Rank	Protein	Gene	Evidence	PMID
1	P13232	IL7	(Monneret *et al.*, 2020)	32728202
2	P48551	IFNAR2	(Sa Ribero *et al.*, 2020)	32726355
3	Q15116	PDCD1	(Zhang *et al.*, 2020)	32048861
6	P42226	STAT6	(Nepal *et al.*, 2019)	31363052
7	Q92949	FOXJ1	(Schaefer *et al.*, 2020)	32561849
8	P23458	JAK1	(Sa Ribero *et al.*, 2020)	32726355
10	Q00978	IRF9	(Sa Ribero *et al.*, 2020)	32726355

SARS-CoV干扰IFN的诱导和信号传导通路



[Ribero et al. PLoS Pathog., 2020]

HPODNets发现了新的蛋白质HPO标注

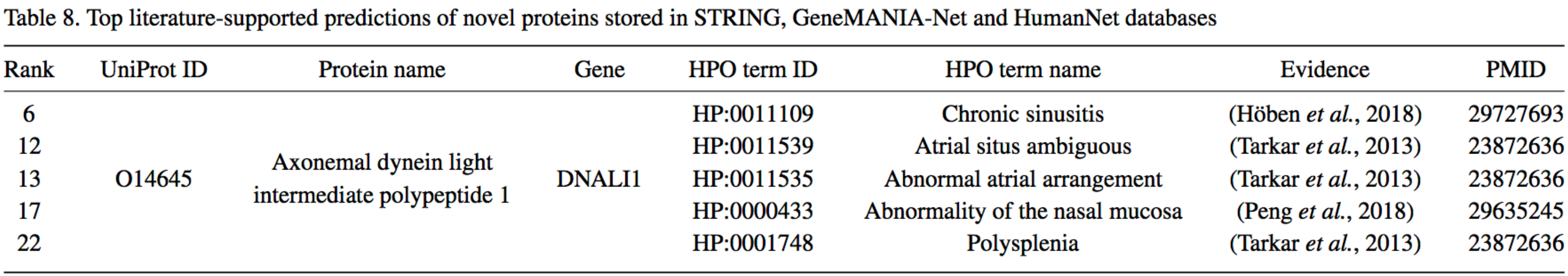


Table 8. Top literature-supported predictions of novel proteins stored in STRING, GeneMANIA-Net and HumanNet databases

Rank	UniProt ID	Protein name	Gene	HPO term ID	HPO term name	Evidence	PMID
6	O14645	Axonemal dynein light intermediate polypeptide 1	DNALI1	HP:0011109	Chronic sinusitis	(Höben *et al.*, 2018)	29727693
12				HP:0011539	Atrial situs ambiguous	(Tarkar *et al.*, 2013)	23872636
13				HP:0011535	Abnormal atrial arrangement	(Tarkar *et al.*, 2013)	23872636
17				HP:0000433	Abnormality of the nasal mucosa	(Peng *et al.*, 2018)	29635245
22				HP:0001748	Polysplenia	(Tarkar *et al.*, 2013)	23872636

HPODNets帮助发现新的致病基因关联

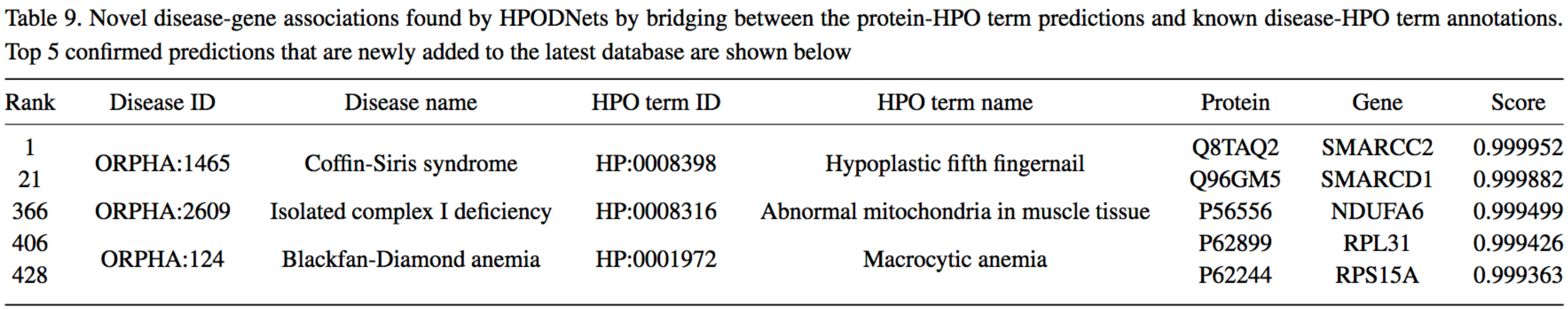
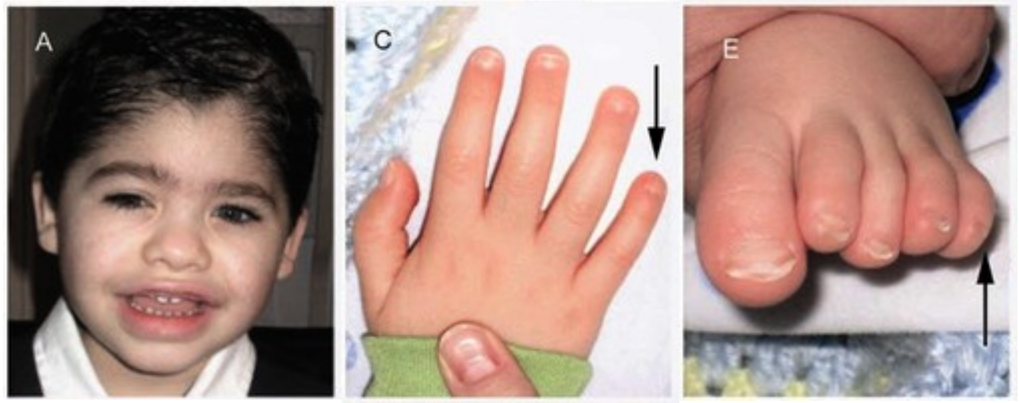


Table 9. Novel disease-gene associations found by HPODNets by bridging between the protein-HPO term predictions and known disease-HPO term annotations. Top 5 confirmed predictions that are newly added to the latest database are shown below

Rank	Disease ID	Disease name	HPO term ID	HPO term name	Protein	Gene	Score
1	ORPHA:1465	Coffin-Siris syndrome	HP:0008398	Hypoplastic fifth fingernail	Q8TAQ2	SMARCC2	0.999952
21					Q96GM5	SMARCD1	0.999882
366	ORPHA:2609	Isolated complex I deficiency	HP:0008316	Abnormal mitochondria in muscle tissue	P56556	NDUFA6	0.999499
406	ORPHA:124	Blackfan-Diamond anemia	HP:0001972	Macrocytic anemia	P62899	RPL31	0.999426
428					P62244	RPS15A	0.999363



Coffin-Siris syndrome

SMARCC2 SMARCD1

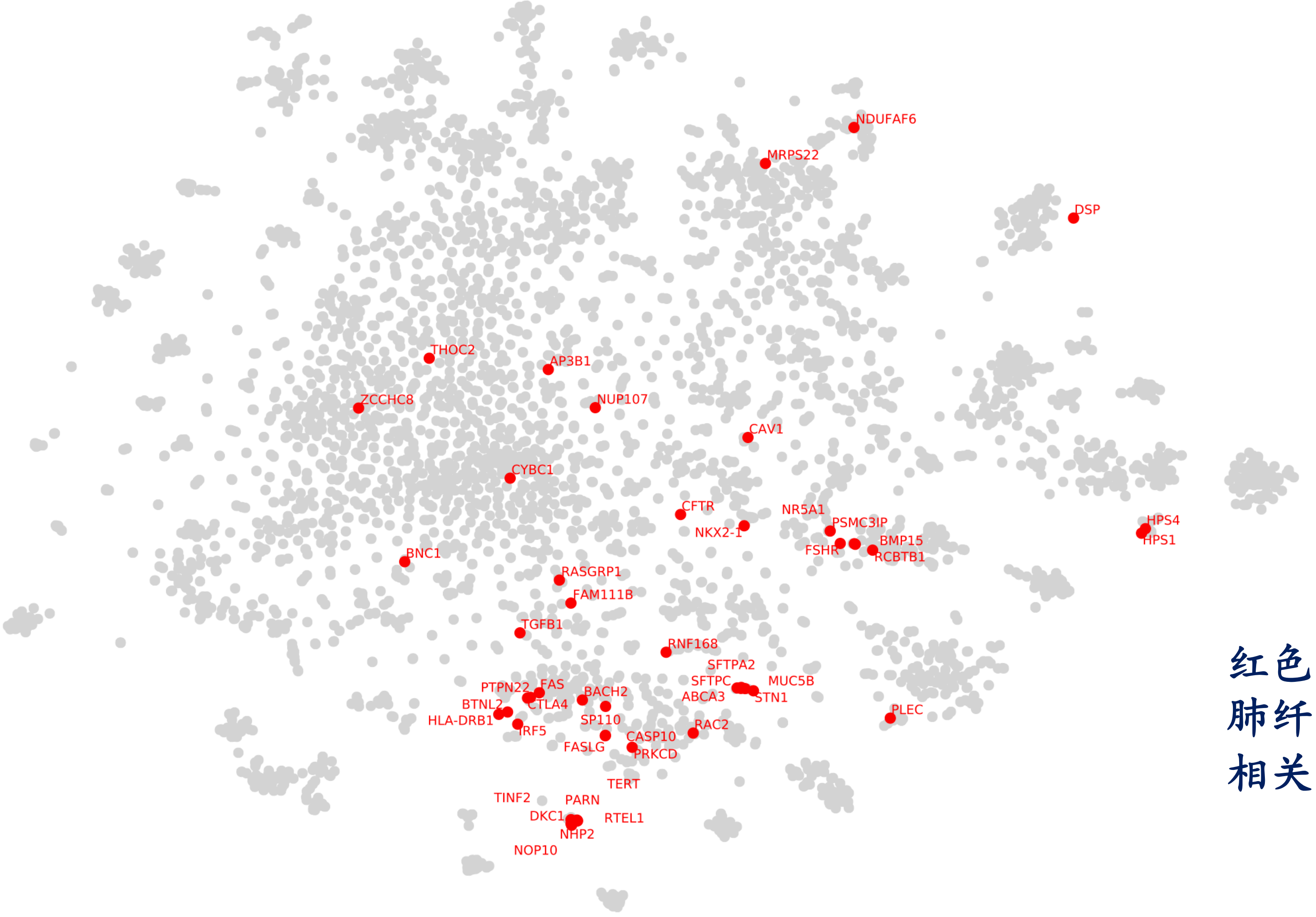
Isolated complex I deficiency

NDUFA6

Blackfan-Diamond anemia

RPL31 RPS15A

探索HPODNets生成的蛋白质嵌入表示



红色圆点表示与HP:0002206 肺纤维化（*Pulmonary fibrosis*）相关的基因

小结

- 我们提出了第一个基于深度图卷积神经网络的、整合了多种蛋白质互作网络的、以HPO术语为中心的人类蛋白质异常表型标注预测算法HPODNets。

- GCNII模块将初始表示和恒等映射引入传统的图卷积操作中，并巧妙摆放各组件顺序，有效缓解了深度图神经网络的过平滑现象，不仅捕捉了低阶也探索了高阶拓扑结构。

- HPODNets作为半监督学习算法以端到端方式呈现，引入了已知的蛋白质表型标注监督信息，克服了其它半监督学习算法需事先根据输入的多种互作网络构造复合网络导致潜在信息损失的缺点。

谢谢大家