HPOLabeler

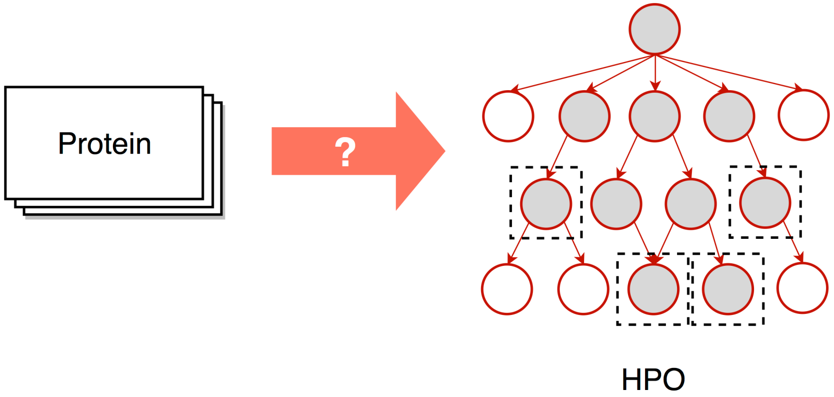
HPOLabeler: Improving Prediction of Human Protein-Phenotype Associations by Learning to Rank

刘砺志

复旦大学计算机科学技术学院

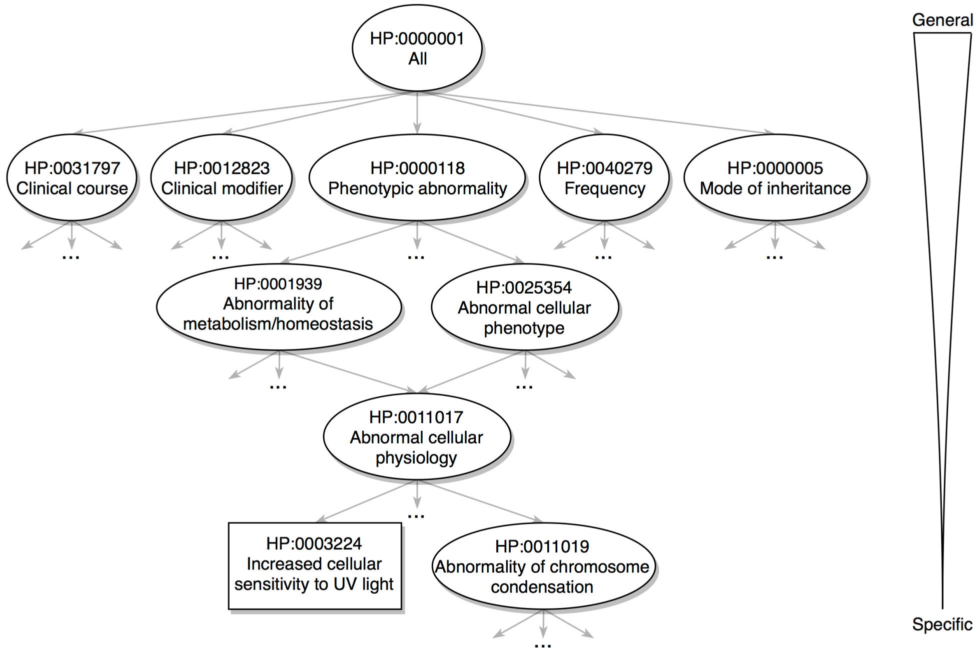
CBC 2020

问题描述：预测人类蛋白质的表型本体注释



预测人类蛋白质的HPO标注问题

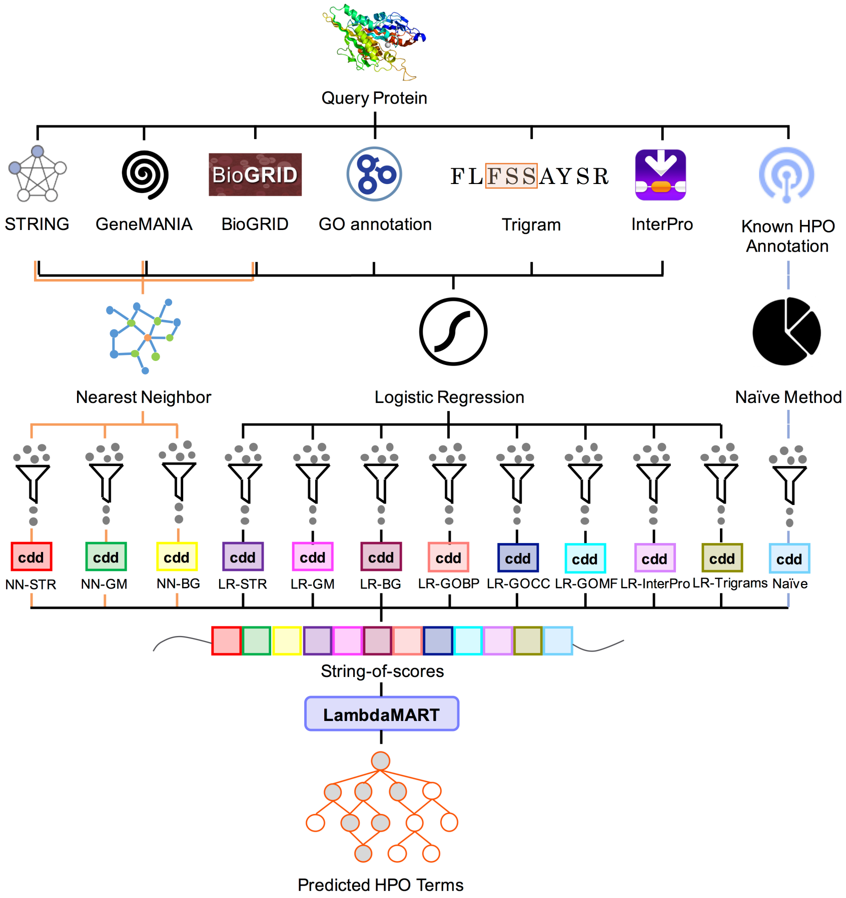
研究目标： 利用机器学习技术，整合多种信息源，提高预测性能



人类表型本体

HPO (Human Phenotype Ontology)

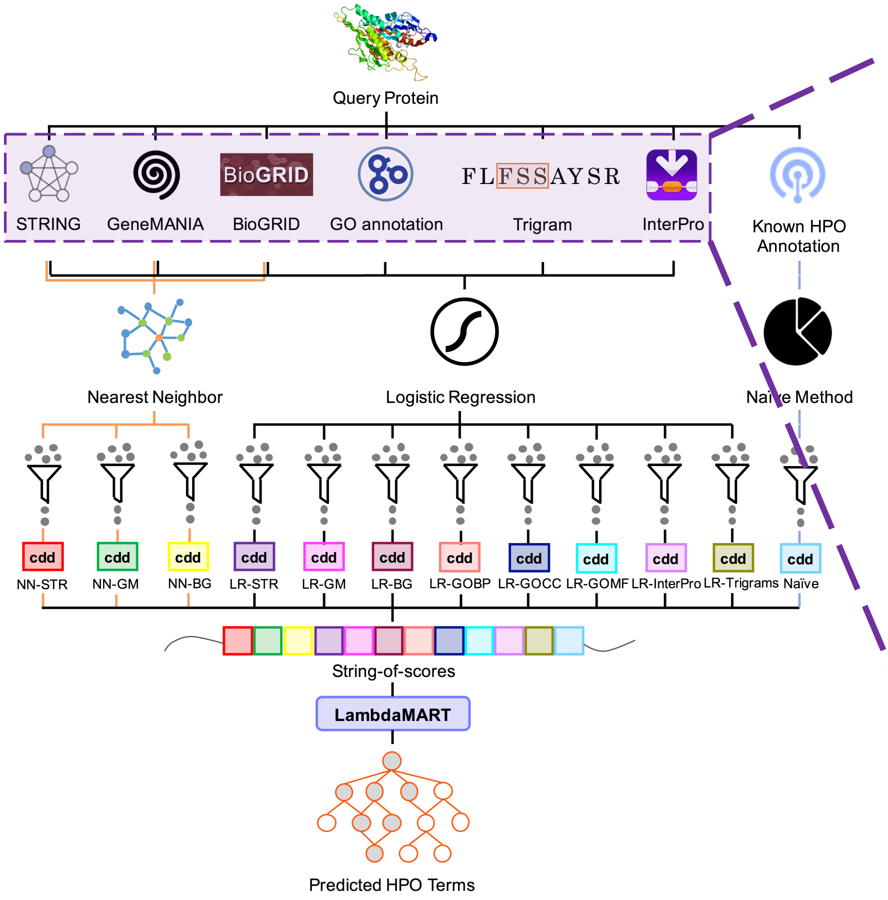
HPOLabeler — 使用排序学习提升预测效果



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

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

GeneMANIA

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

BioGRID

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

GO BP/CC/MF

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

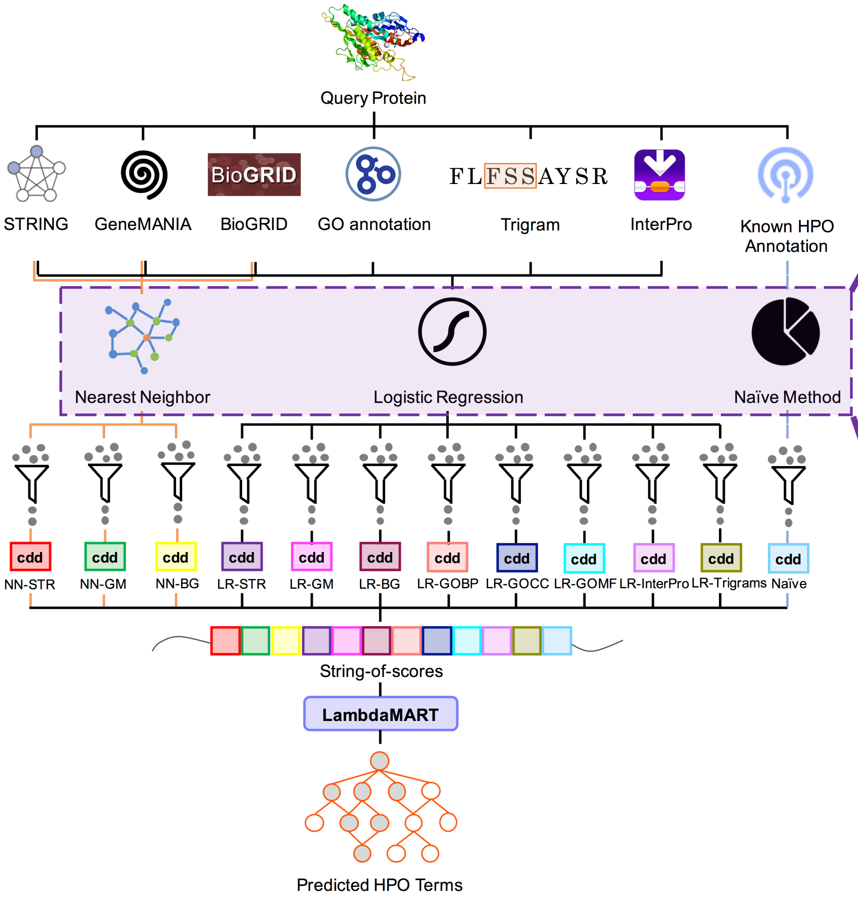
InterPro signatures

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

Trigrams

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

基础模型



LR model for each HPO term

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

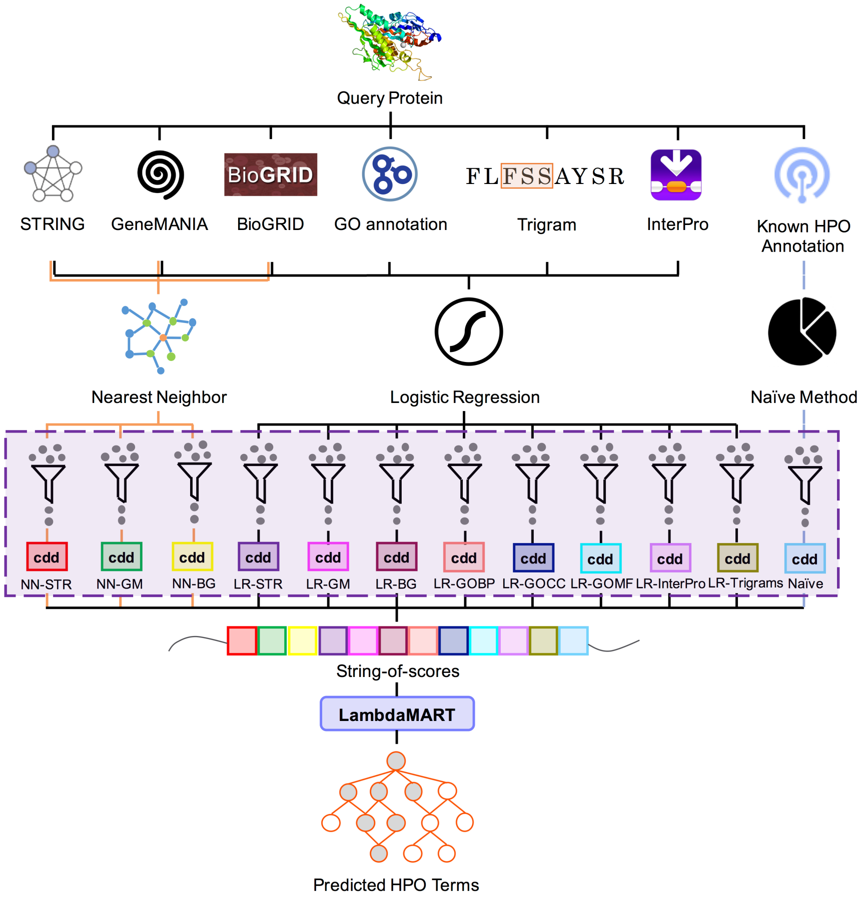
Nearest Neighbor on STRING, GeneMANIA and BioGRID

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

Naïve

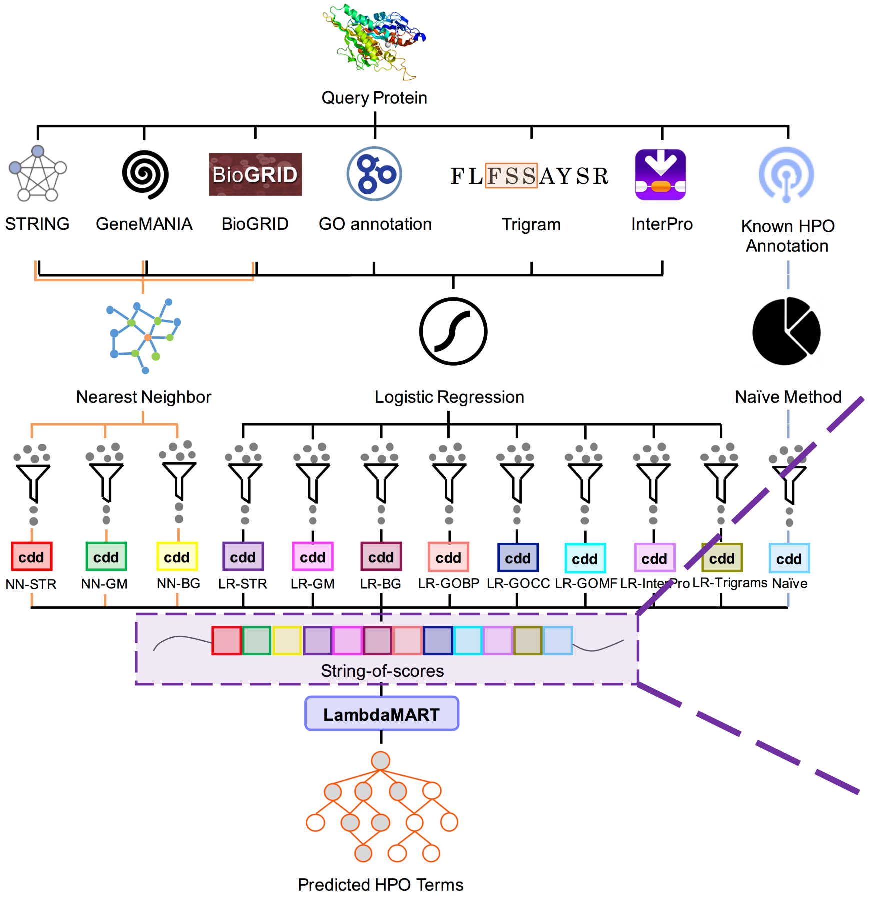
$$S^{(\text{Naïve})}(p_i,t) = \frac{|\{p_j \in \mathcal{P}_S | y_{j,t} = 1\}|}{m_S} \tag{9}$$

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

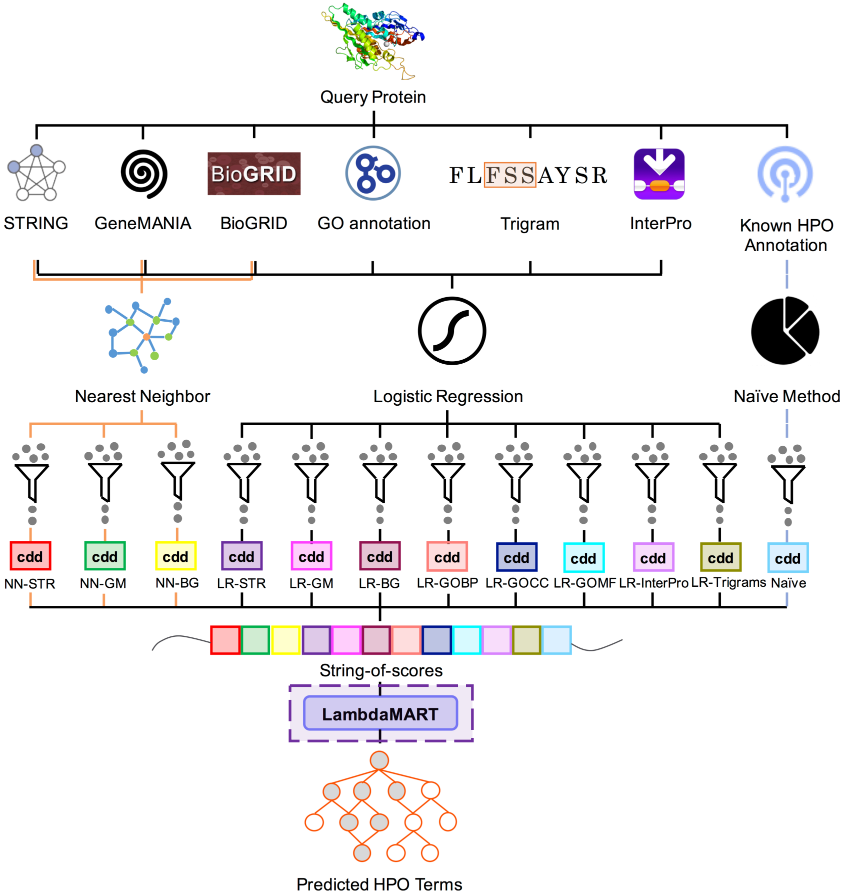
HPOLabeler — 第二步：为排序学习生成特征



String-of-scores

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

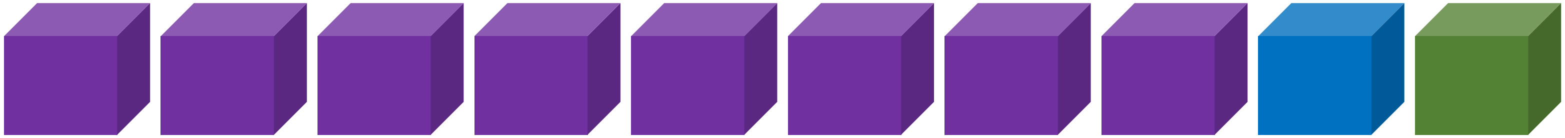
HPOLabeler — 第三步：排序



- 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

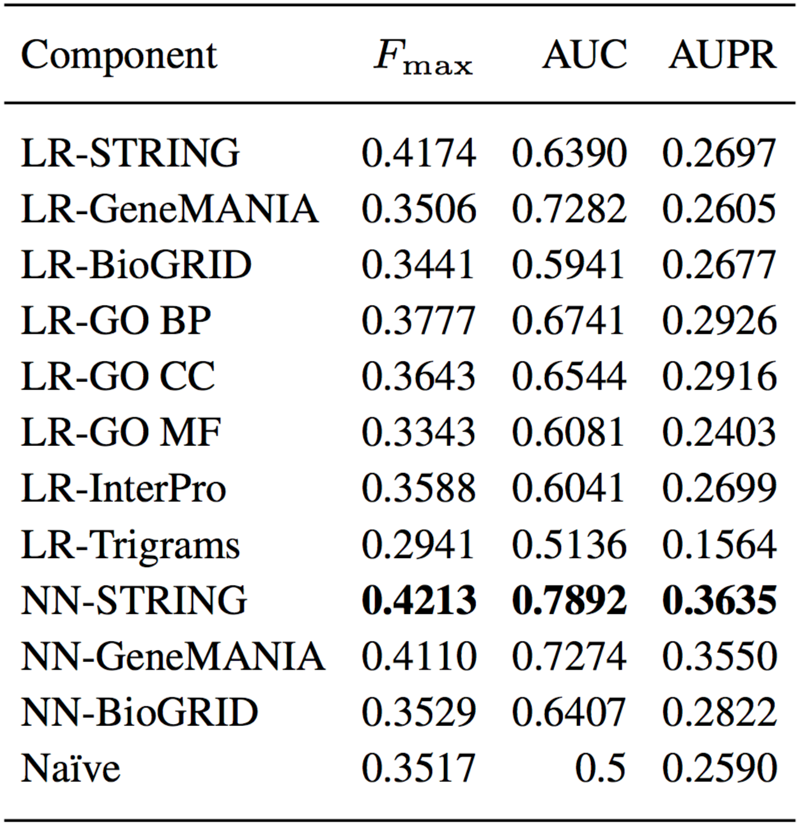
2018-07-27



3,722 proteins 8,067 HPO terms Avg. 119.4 annotations

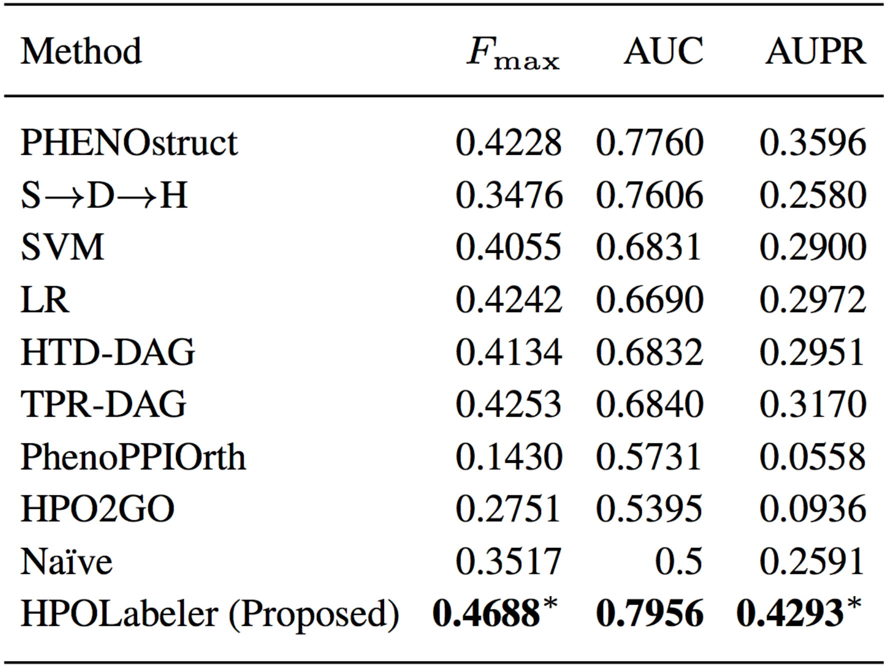
实验结果之交叉验证 — 对比

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

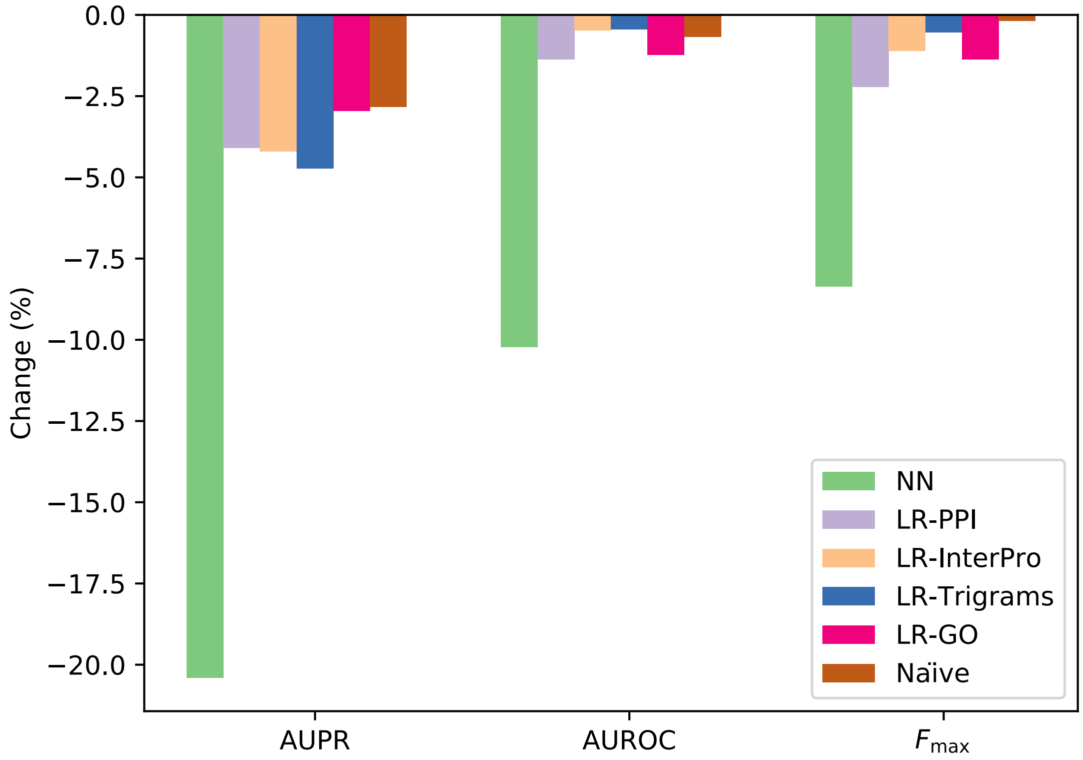
整体模型同对比方法的性能



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***

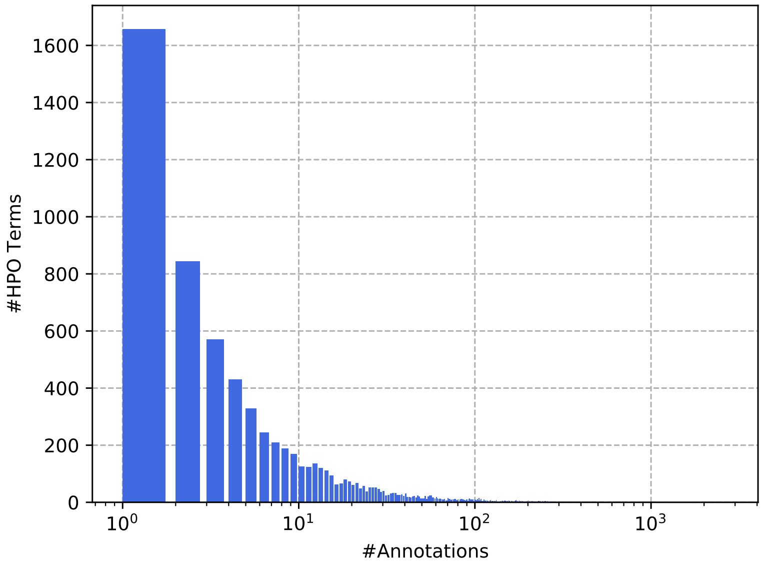
注：$F_{\max}$是基于蛋白质计算的
AUC是基于HPO术语计算的
AUPR是就整体结果而言的

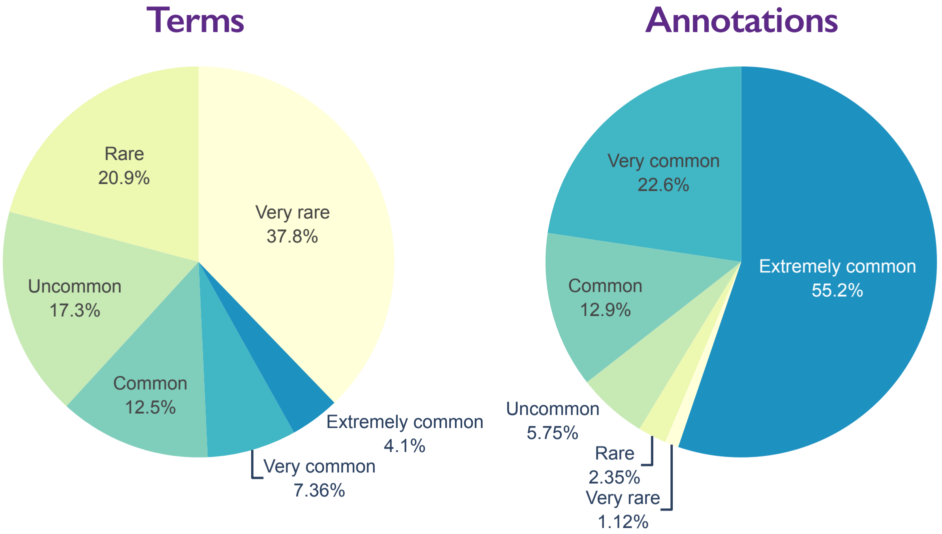
实验结果之交叉验证 — Leave-one-source-out



- **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



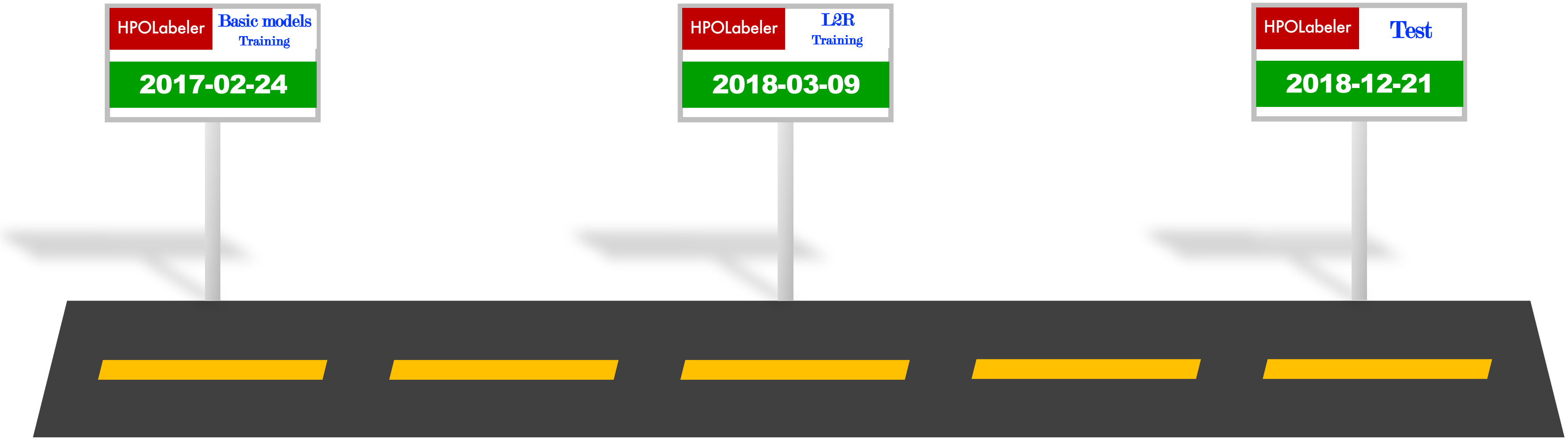


HPO及其注释是不均衡的

- 高频率小组 ^_^
- 低频率小组 -_-

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***

评估之二：依时间划分验证

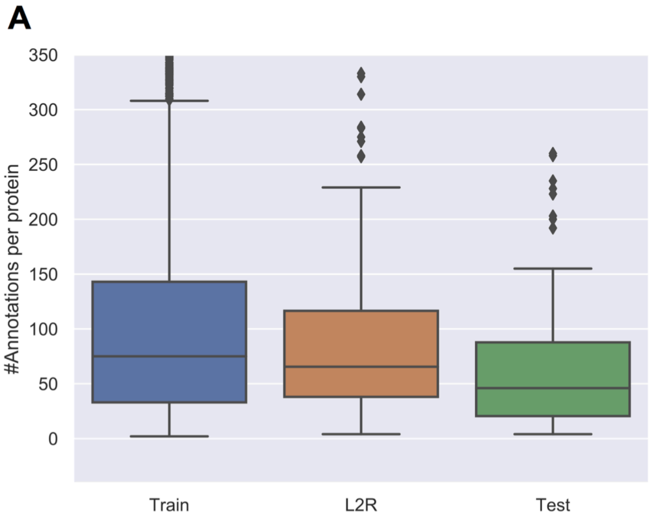


	Train	L2R	Test
Proteins	3,334	304	226
Used HPO terms	7,394	2,836	2,091
Annotations	107.0936	83.9079	61.5177

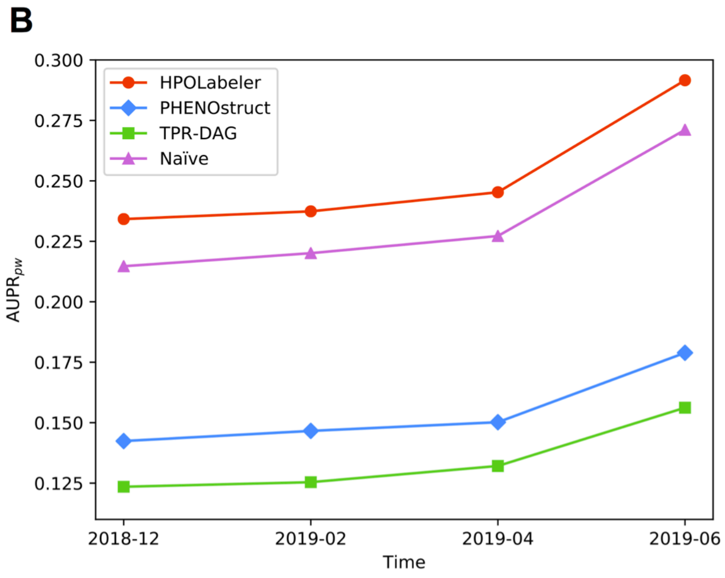
实验结果之依时间划分验证

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

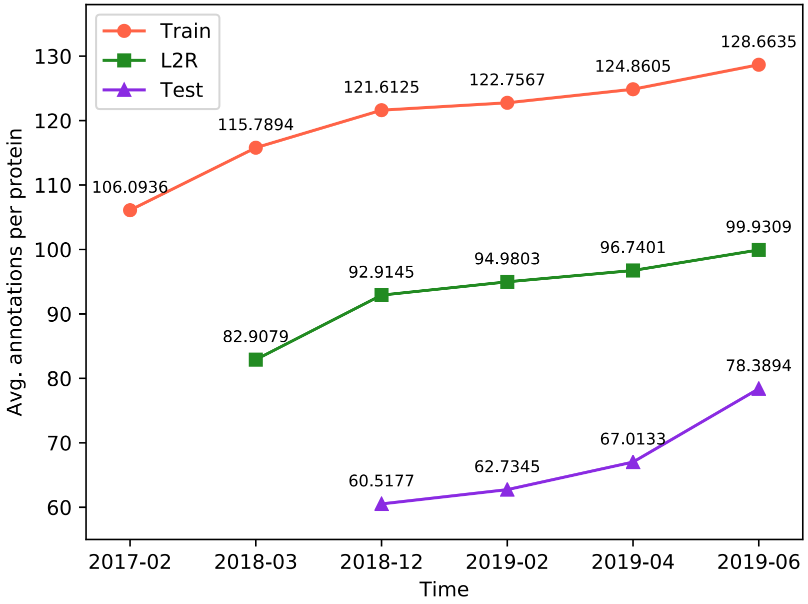


使用不同时间发布的标注文件对预测结果进行评估

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

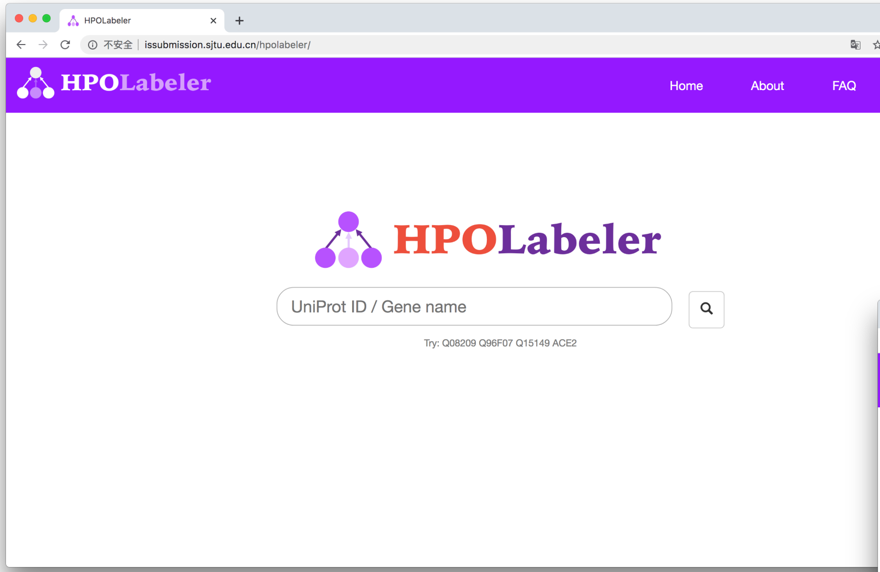
小结

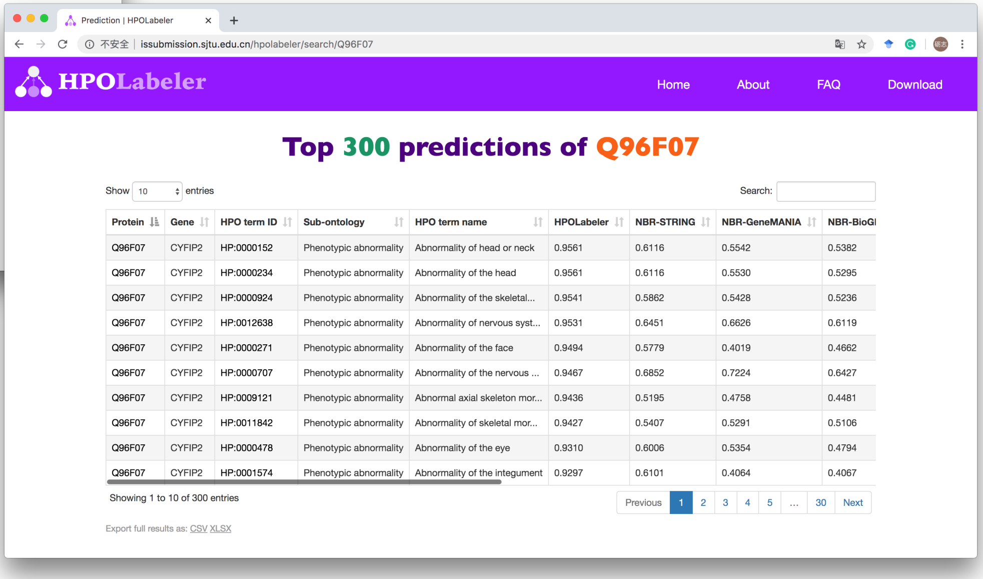
- 我们提出了预测人类蛋白质的HPO标注的算法HPOLabeler，其在排序学习的框架下整合了包括PPI、GO、InterPro和序列信息等在内的多种信息源。

- 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。

在线平台


HPOLabeler
不安全 | issubmission.sjtu.edu.cn/hpolabeler/
HPOLabeler
Home
About
FAQ
Download
HPOLabeler
UniProt ID / Gene name
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-BioGI
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

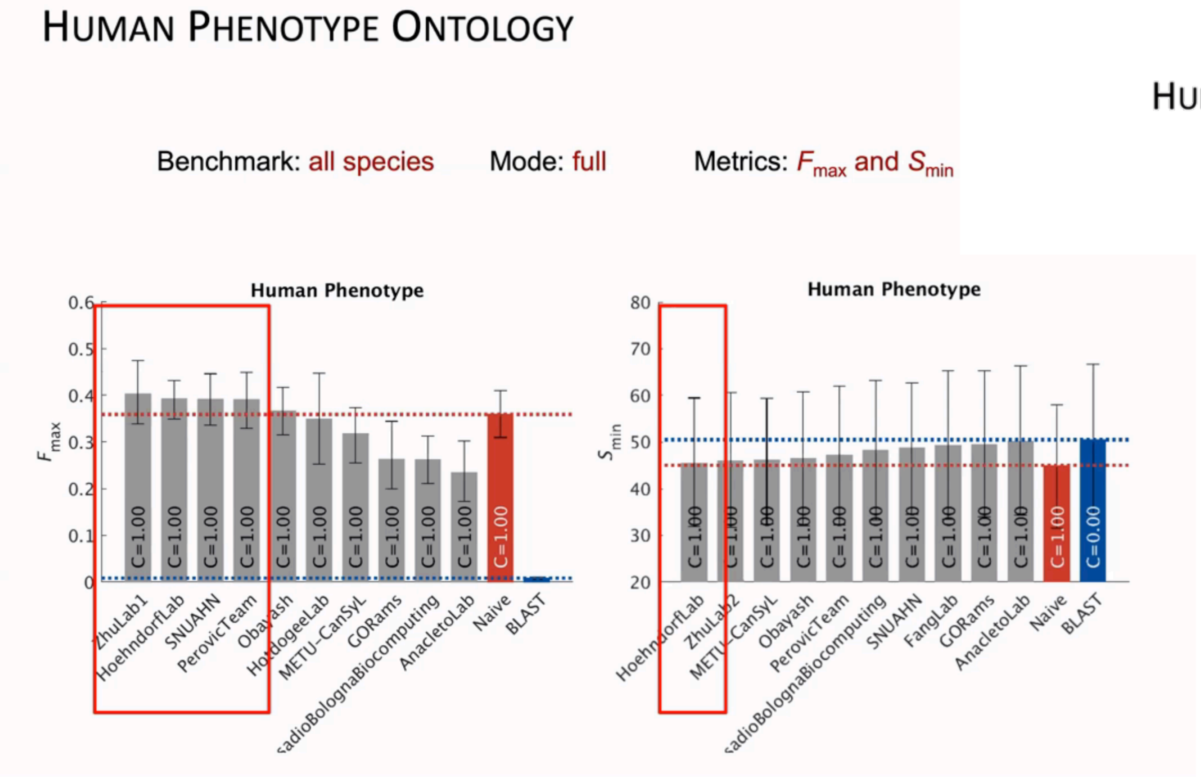
Showing 1 to 10 of 300 entries

Previous 1 2 3 4 5 … 30 Next

Export full results as: CSV XLSX

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

CAFA4竞赛初步评估结果



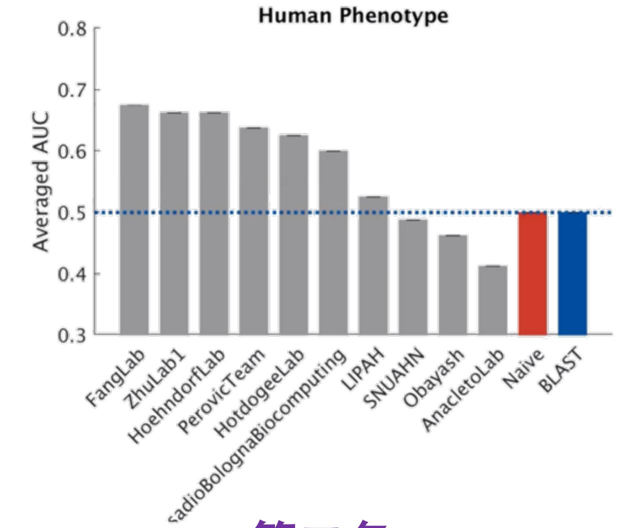
Human Phenotype Ontology

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

第一名　　第二名

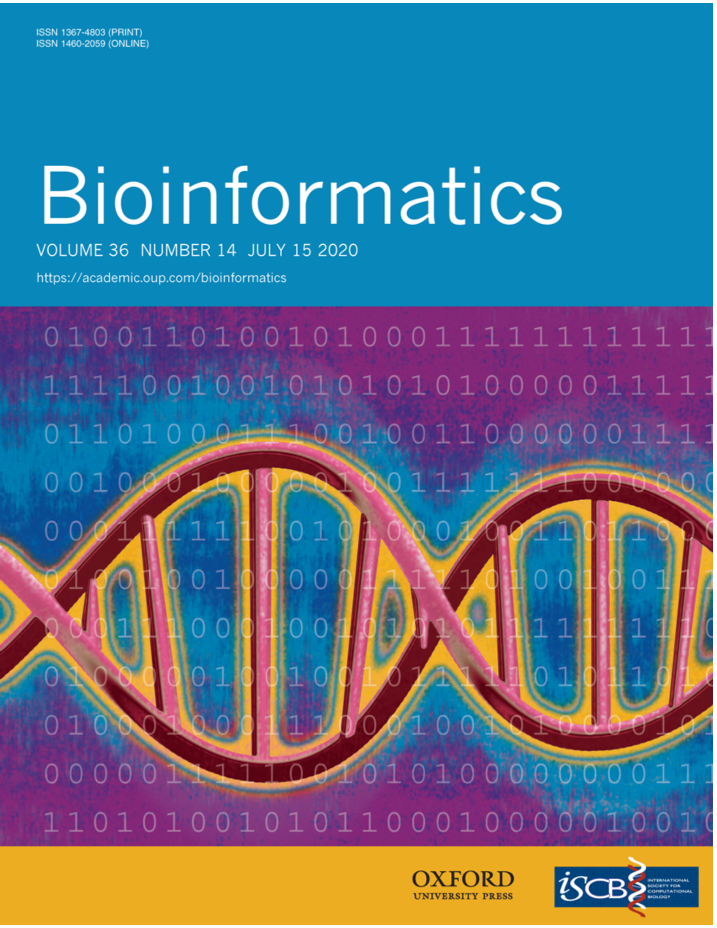
Human Phenotype, Term-Centric Evaluation

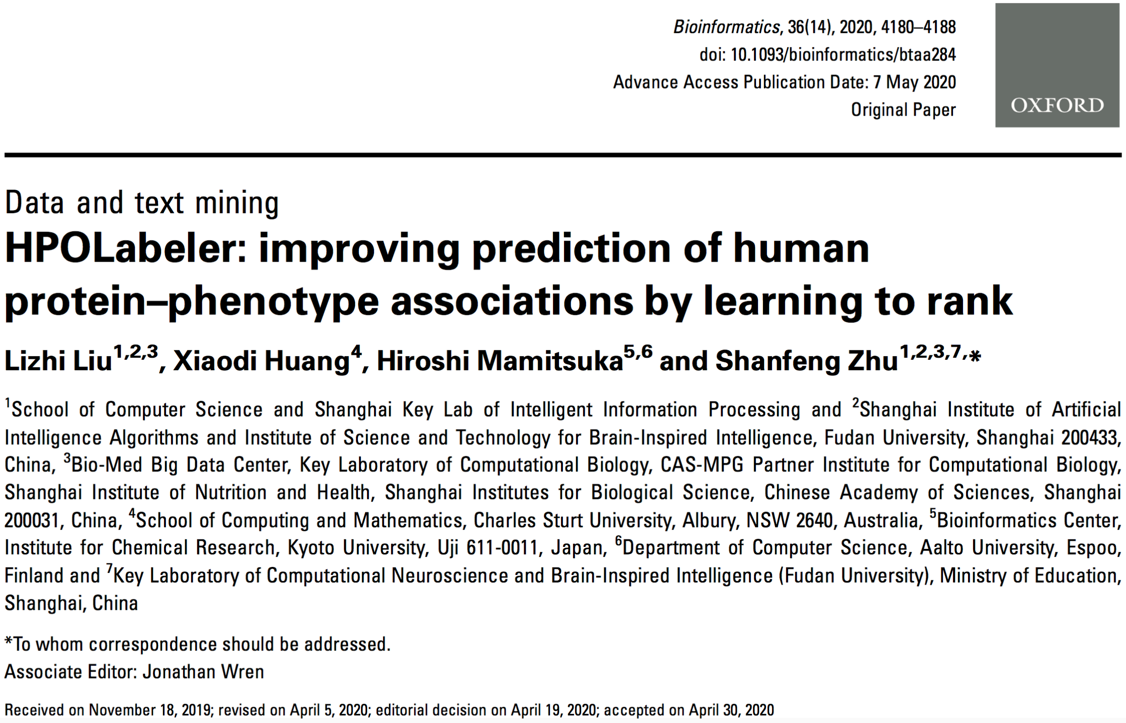
Terms: 4



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

OXFORD

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

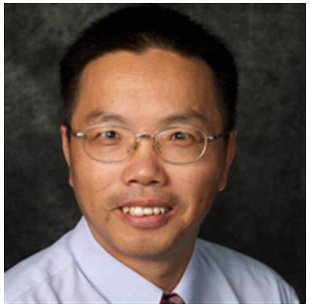
致谢

復旦大學 FUDAN UNIVERSITY

智能信息处理上海市重点实验室

FUNDING

上海市表型组
重大研究计划



Xiaodi Huang
@USC

Hiroshi Mamitsuka
@Kyoto U

Shanfeng Zhu
@Fudan

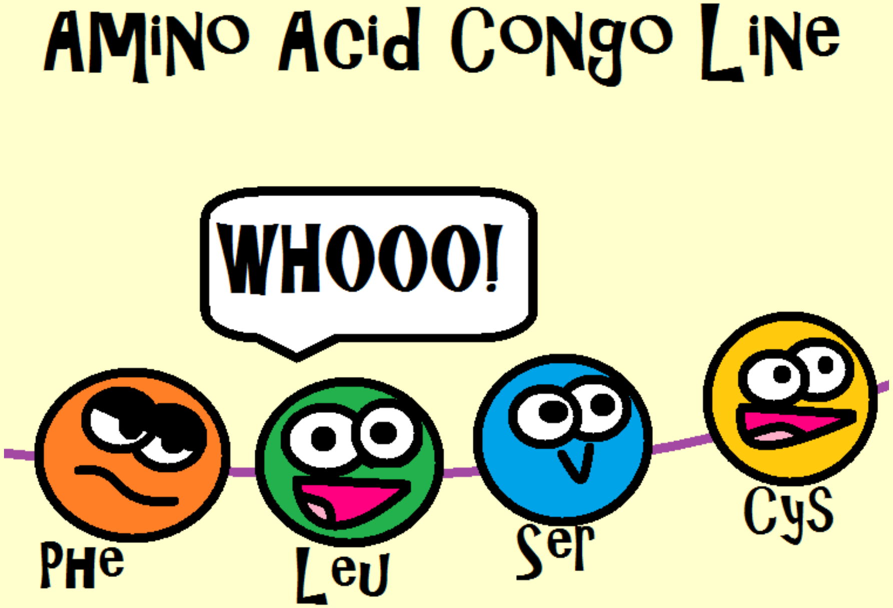
欢迎海内外英才加入我们

复旦大学数据挖掘与智能信息处理实验室

研究主题：人工智能与生物医学大数据挖掘

硕士生 | 博士生 | 博士后

联系方式：zhusf@fudan.edu.cn



谢谢大家

欢迎提问

- 邮箱：liulizhi1996@gmail.com
- 主页：liulizhi1996.github.io

日月光华
旦复旦兮