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

Lizhi Liu ***(Presenter)***

Xiaodi Huang, Hiroshi Mamitsuka, Shanfeng Zhu

Email: *liulizhi1996@gmail.com*

School of Computer Science
Shanghai Key Lab of Intelligent Information Processing
Fudan University, Shanghai, China

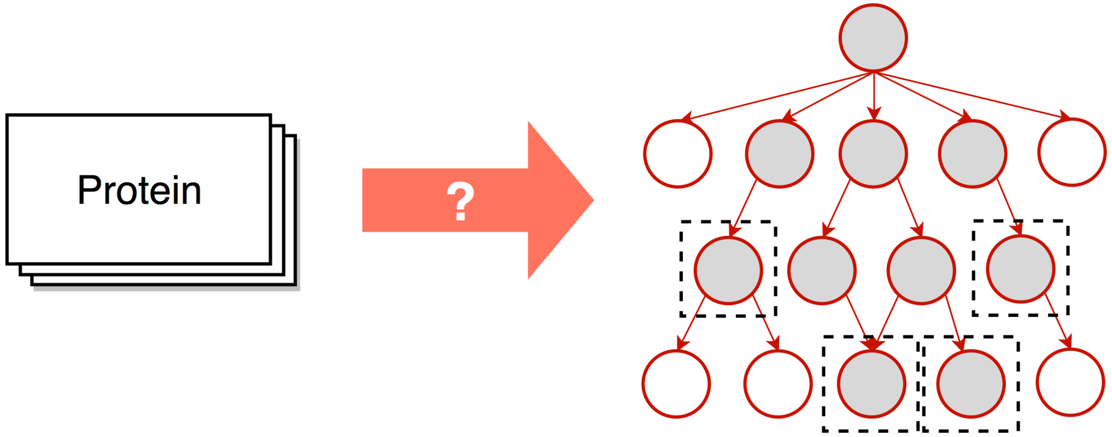
FUDAN UNIVERSITY 1905 復旦大學

montréal ISMB 2020 iSCB

HPOLabeler

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

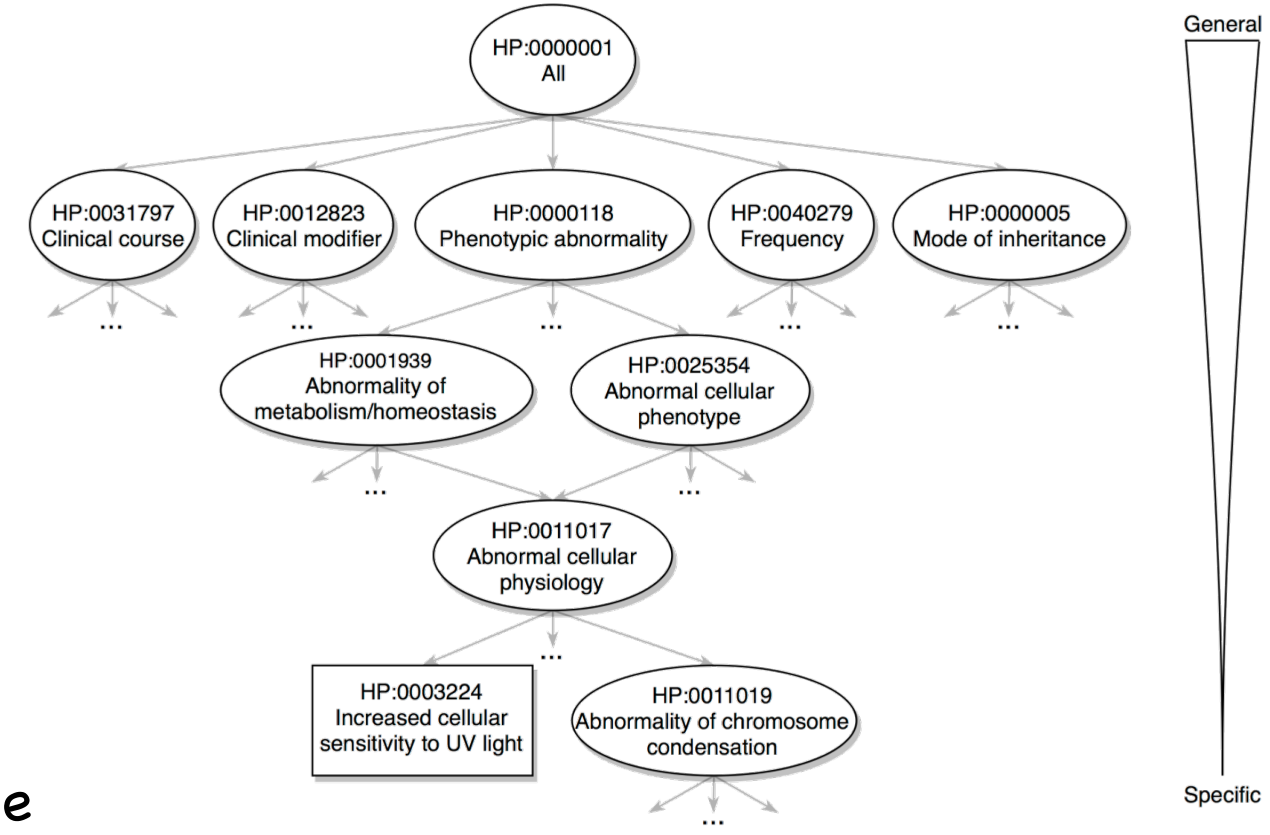
Problem Statement



HPO

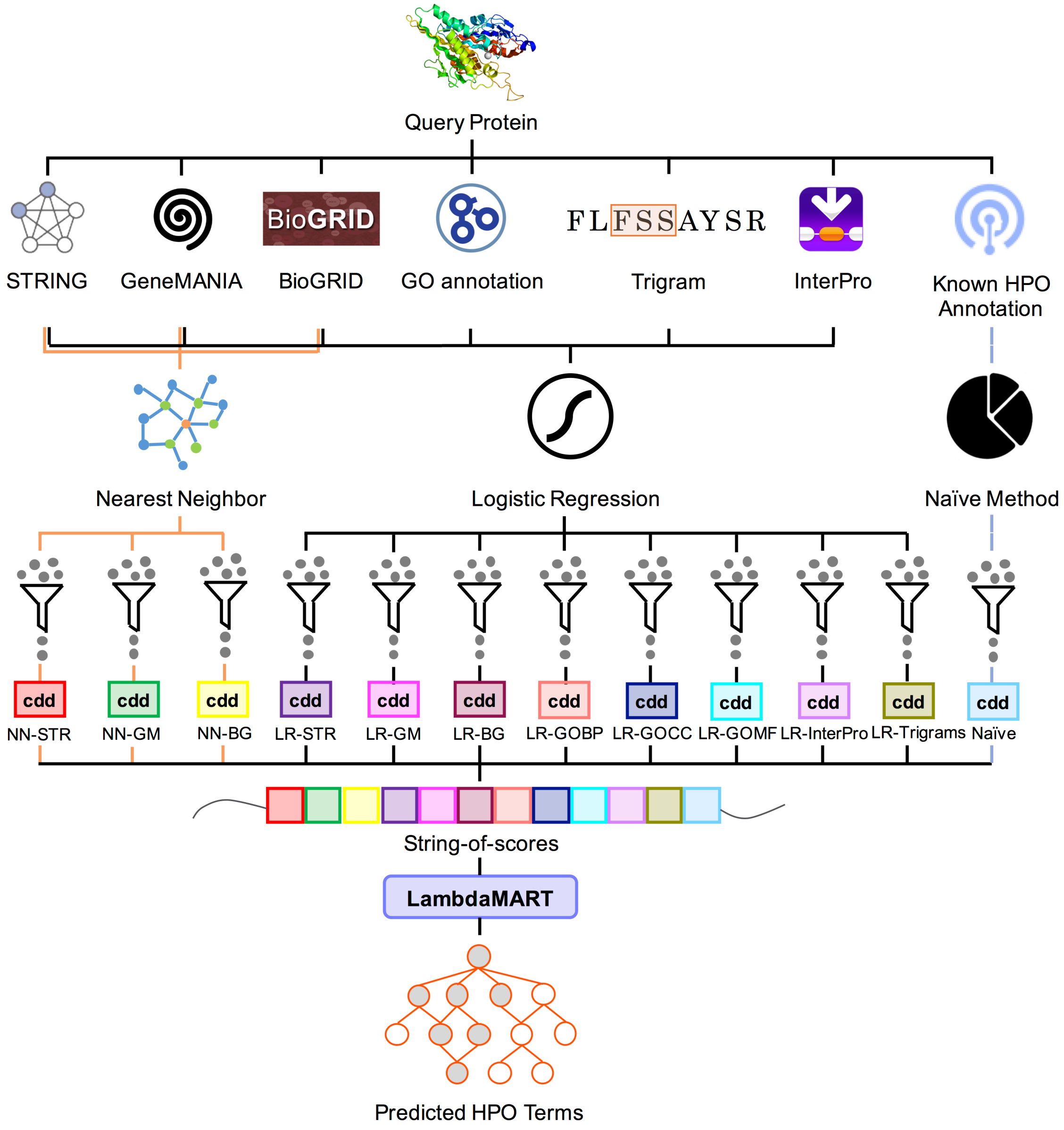
Predicting HPO annotations of Human Proteins

Our goal: Using machine learning techniques to integrate multiple data sources and improve the performance



HPO (Human Phenotype Ontology)

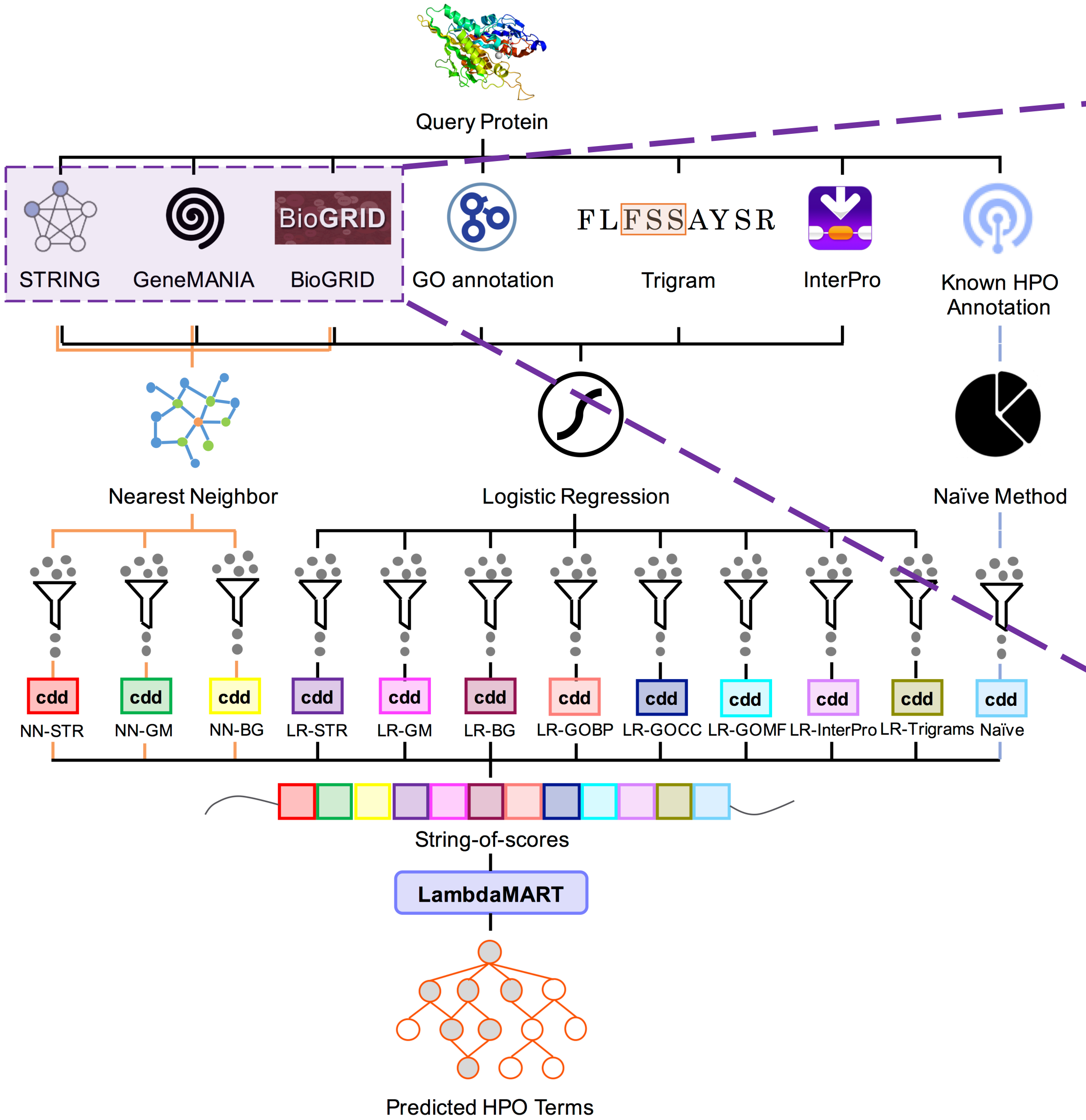
Our Proposal — HPOLabeler



Key Points

- **Ensemble learning : Stacking**
- **Learning to Rank to integrate multiple basic models to further improve the performance**
- **Only one better than Naïve method in temporal validation**

Feature Extraction – PPI Networks



STRING

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

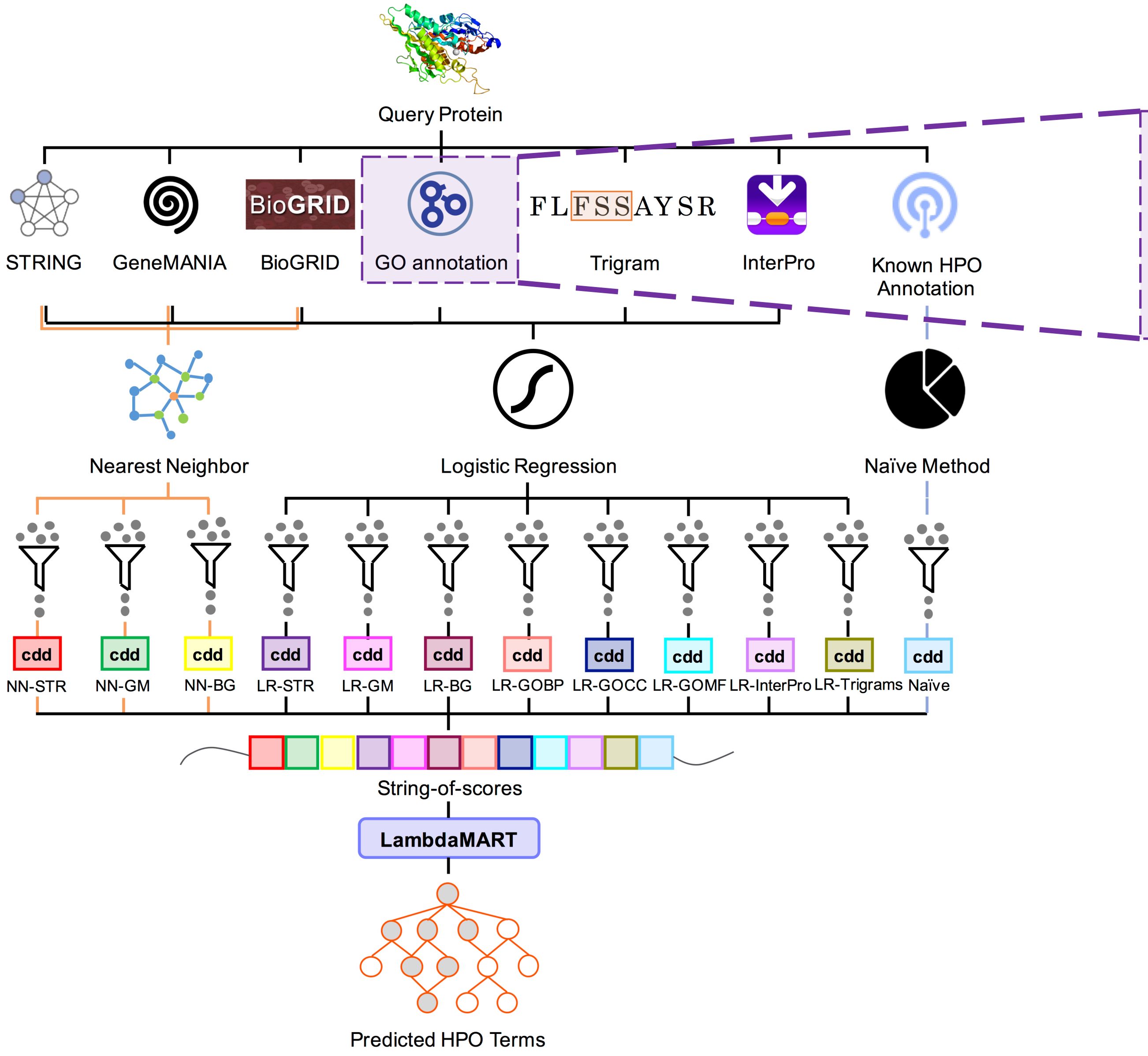
GeneMANIA

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

BioGRID

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

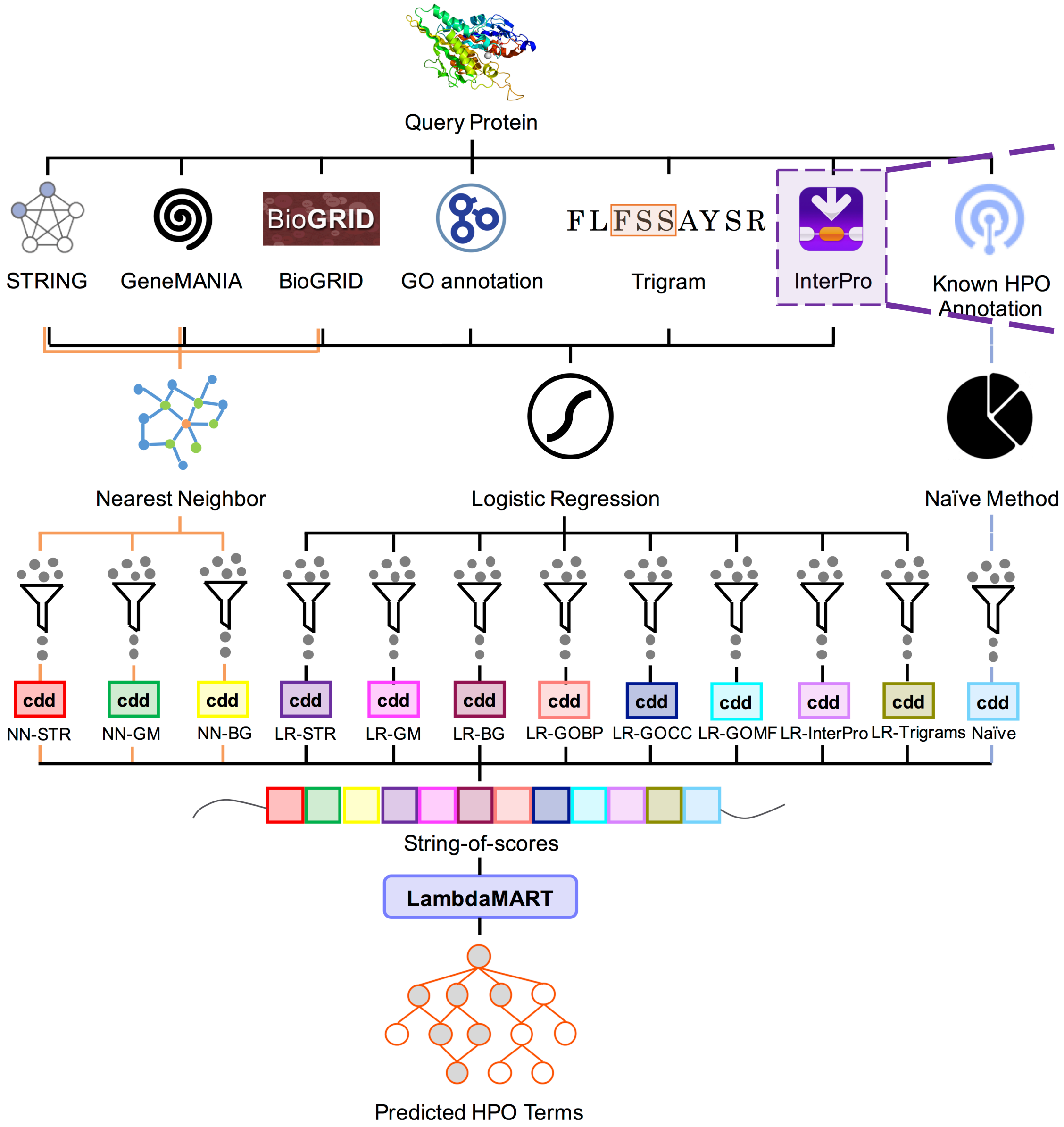
Feature Extraction – GO annotations



GO BP/CC/MF

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

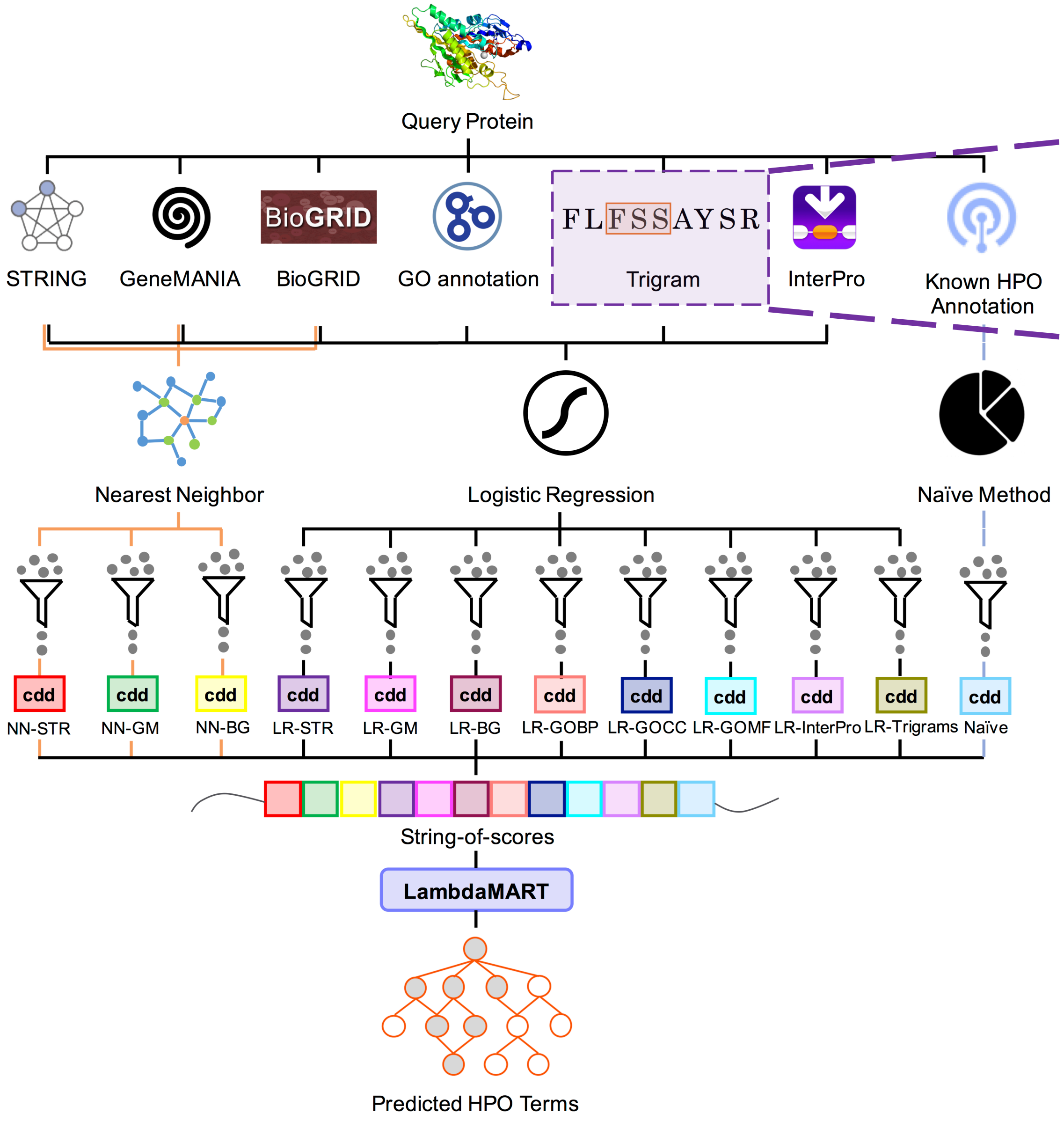
Feature Extraction – InterPro



InterPro signatures

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

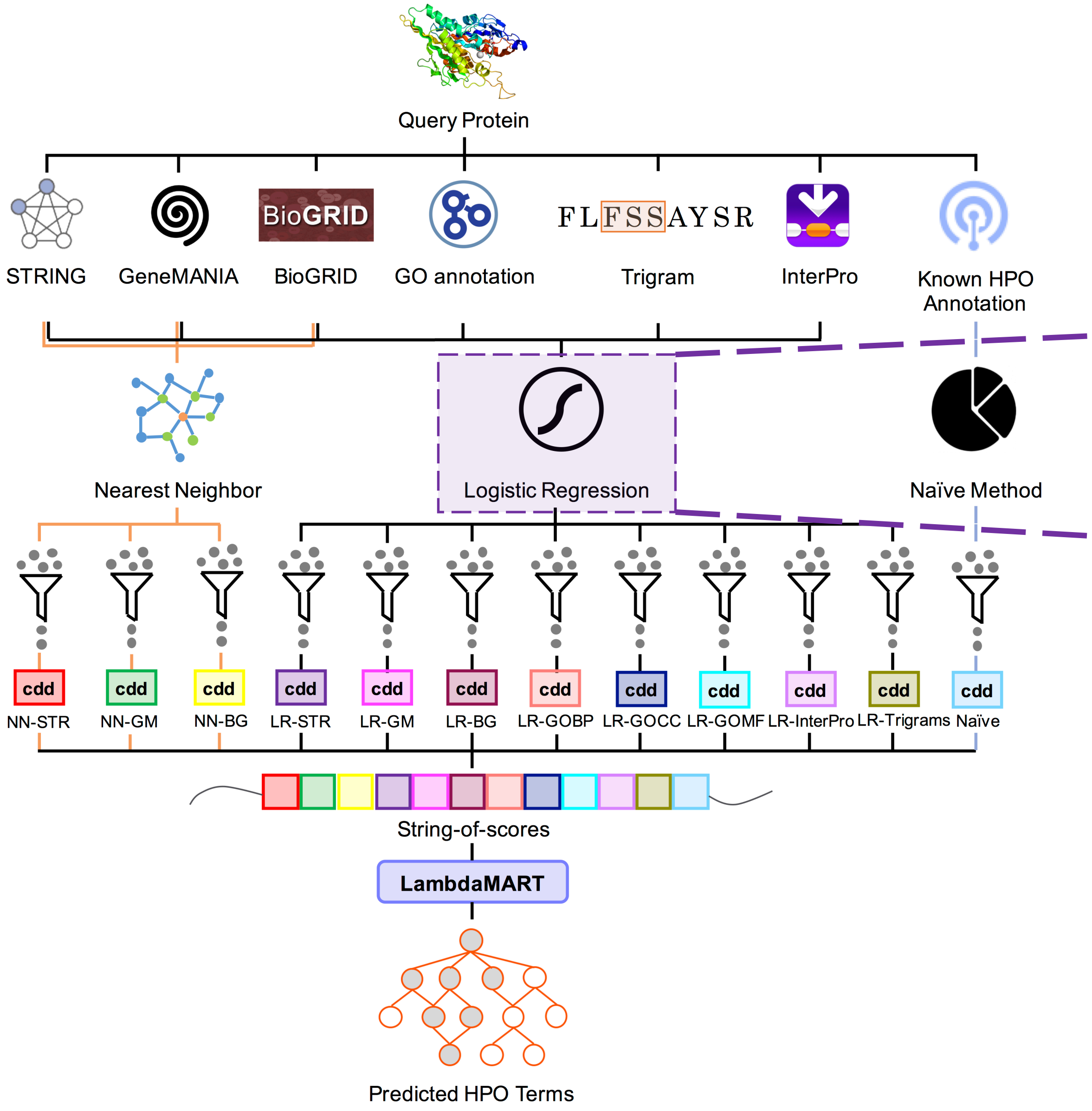
Feature Extraction – Amino Acid Sequences



Trigrams

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

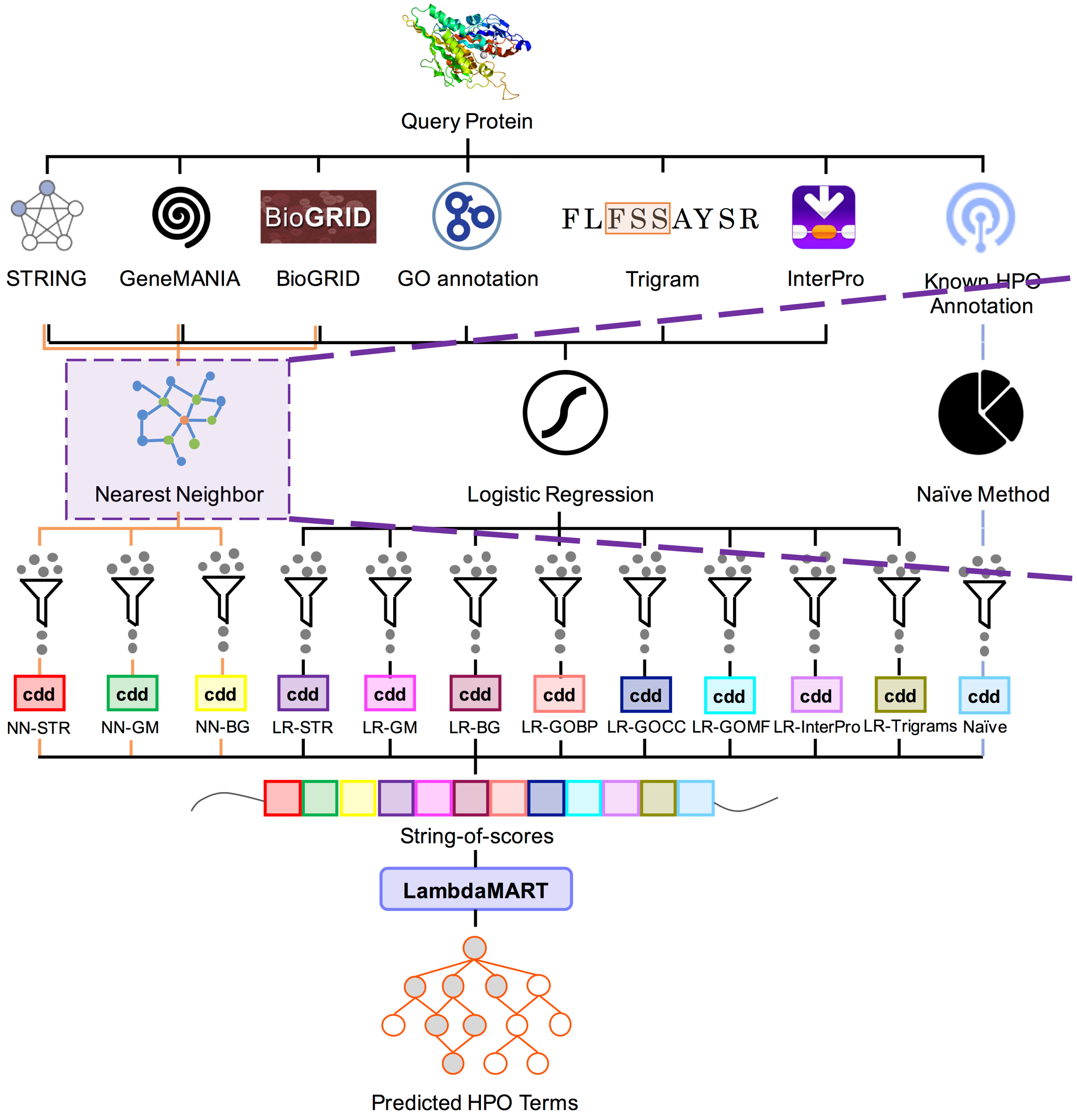
Basic Model – Logistic Regression



LR model for each HPO term

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

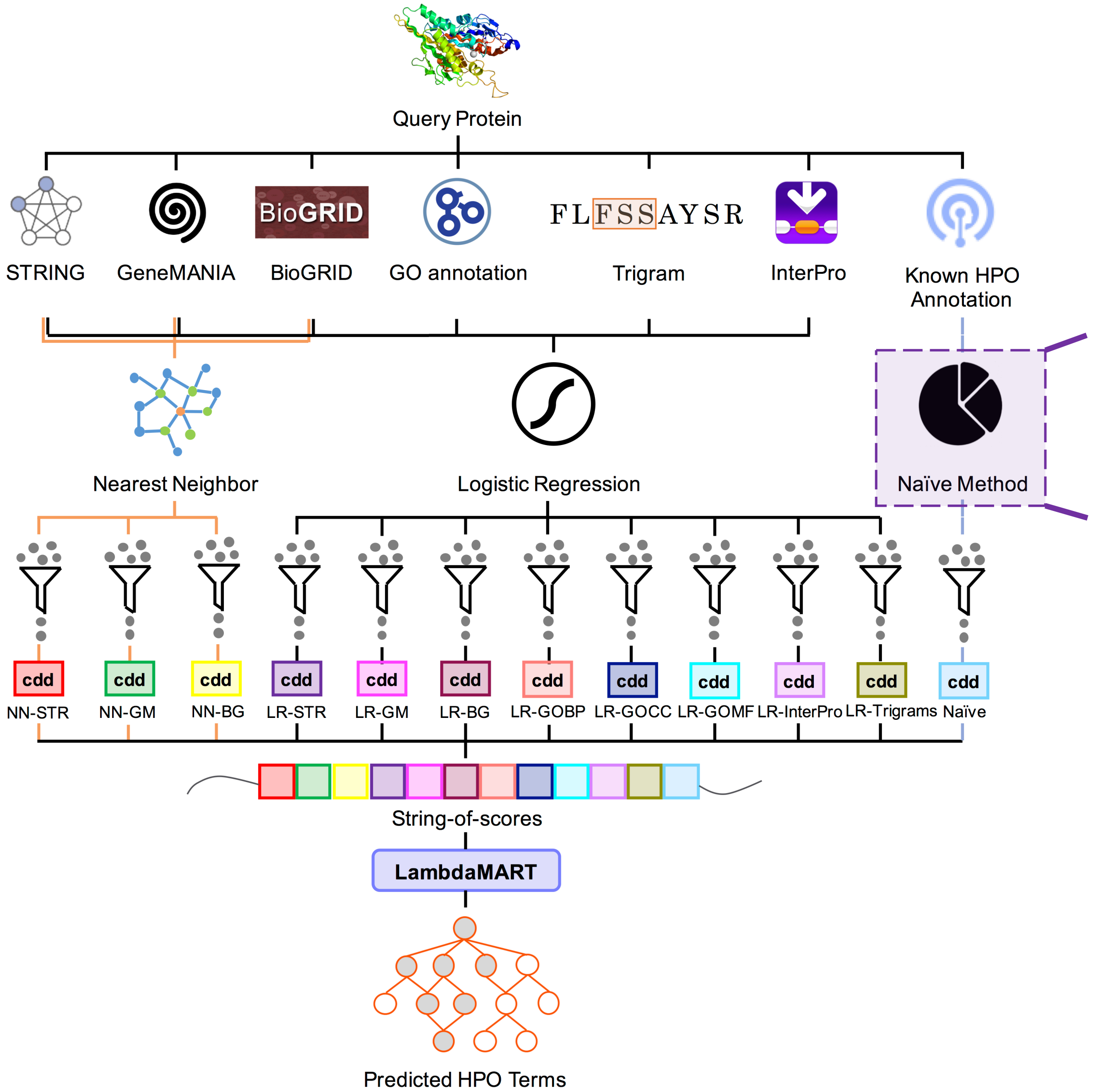
Basic Model – Nearest Neighbor



Nearest Neighbor on STRING, GeneMANIA and BioGRID

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

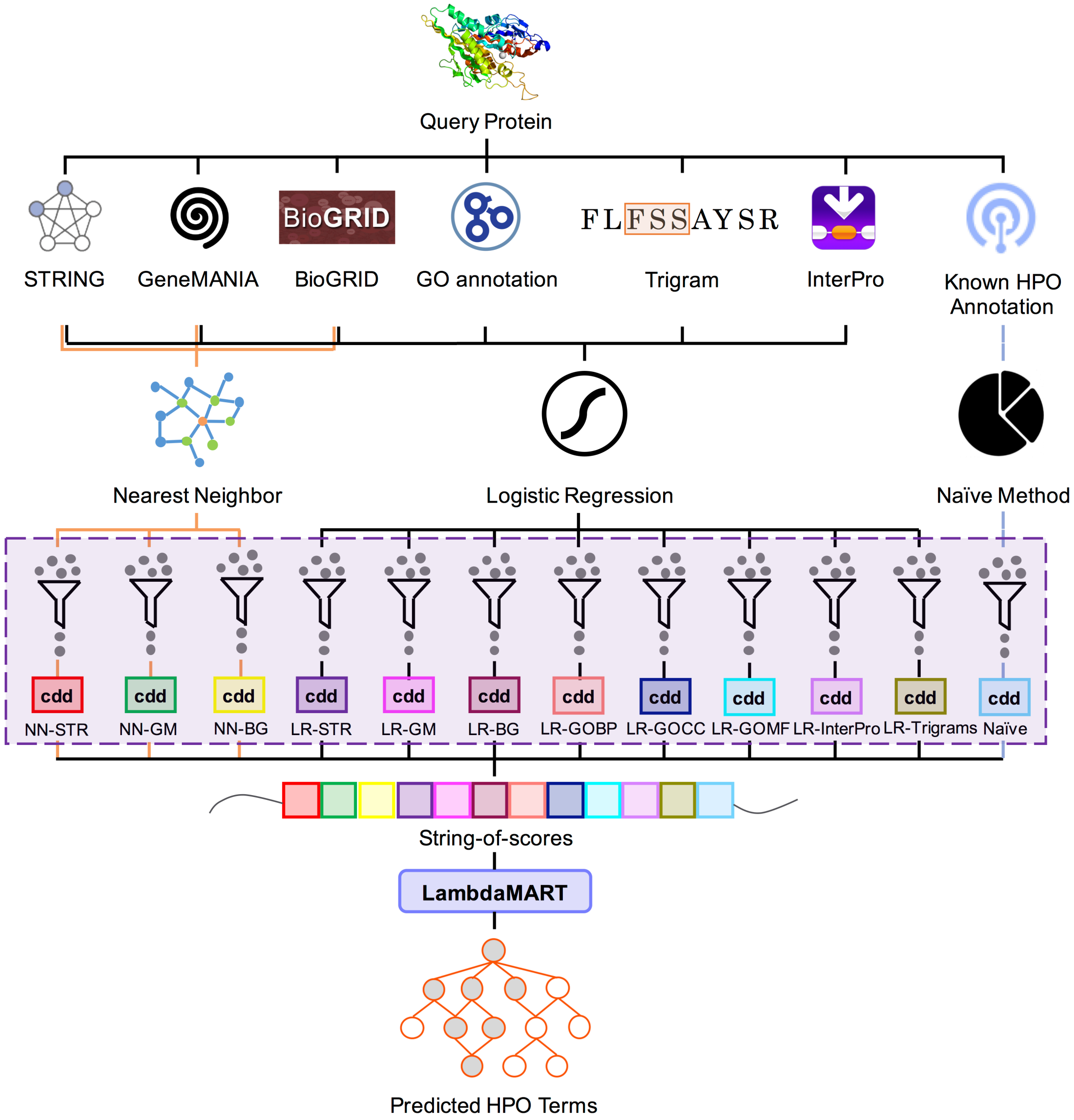
Basic Model – Naïve



Naïve

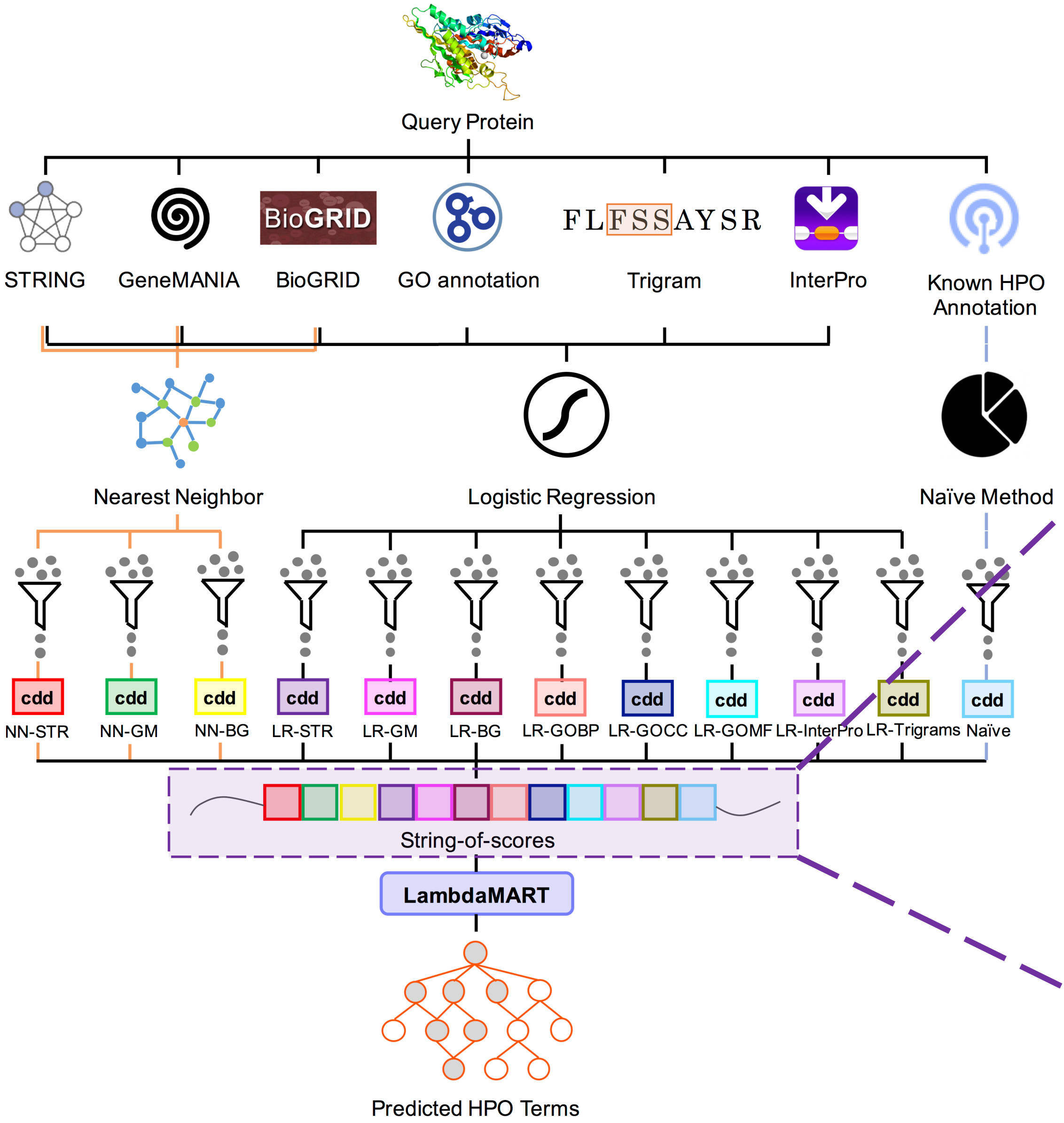
$$S^{(\text{Naïve})}(p_i, t) = \frac{|\{p_j \in \mathcal{P}_S | y_{j,t} = 1\}|}{m_S} \tag{9}$$

HPOLabeler – Step 1: Candidate Generation



- Top-*k* of HPO terms on each of basic models are selected
- Take the union of these subsets as the finalized candidates

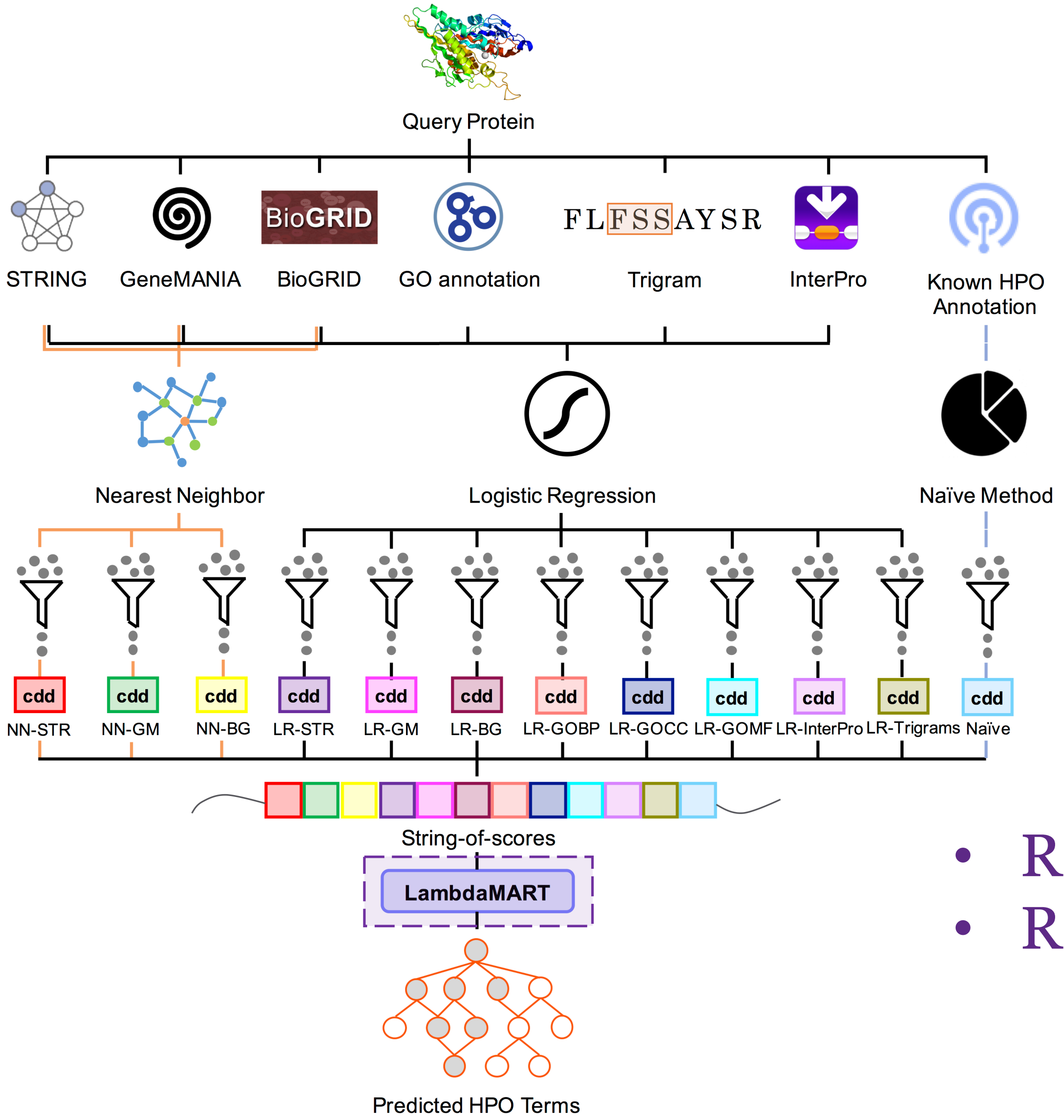
HPOLabeler – Step 2: Feature Generation for LTR



String-of-scores

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

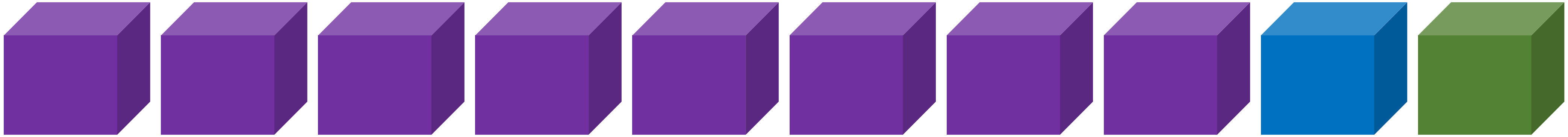
HPOLabeler – Step 3: Ranking



- Re-rank candidates based on **LambdaMART**
- Receive a ranked list of predictive scores

Evaluation 1: Cross-validation

2018-07-27

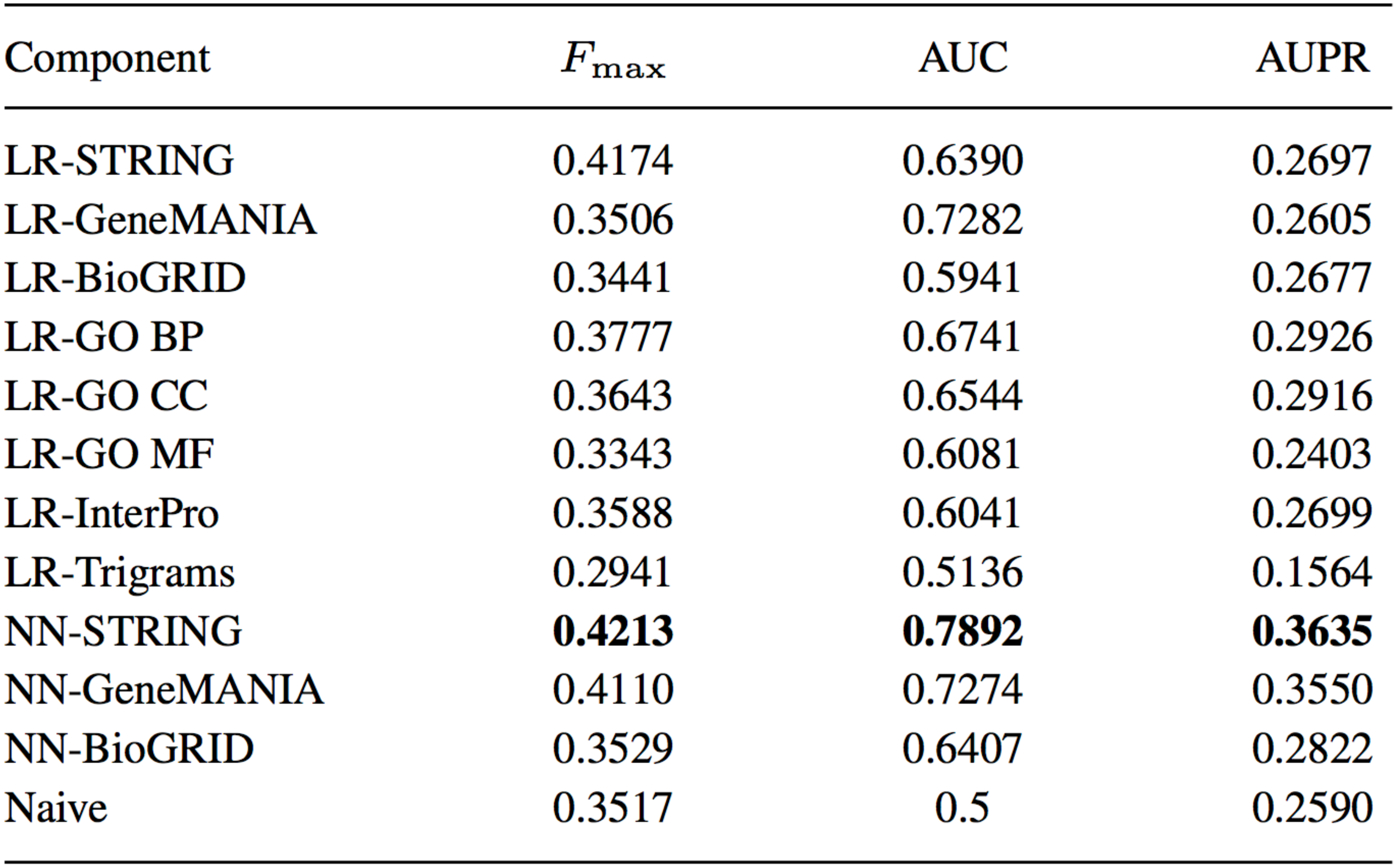


3,722 proteins

8,067 HPO terms

Avg. 119.4 annotations

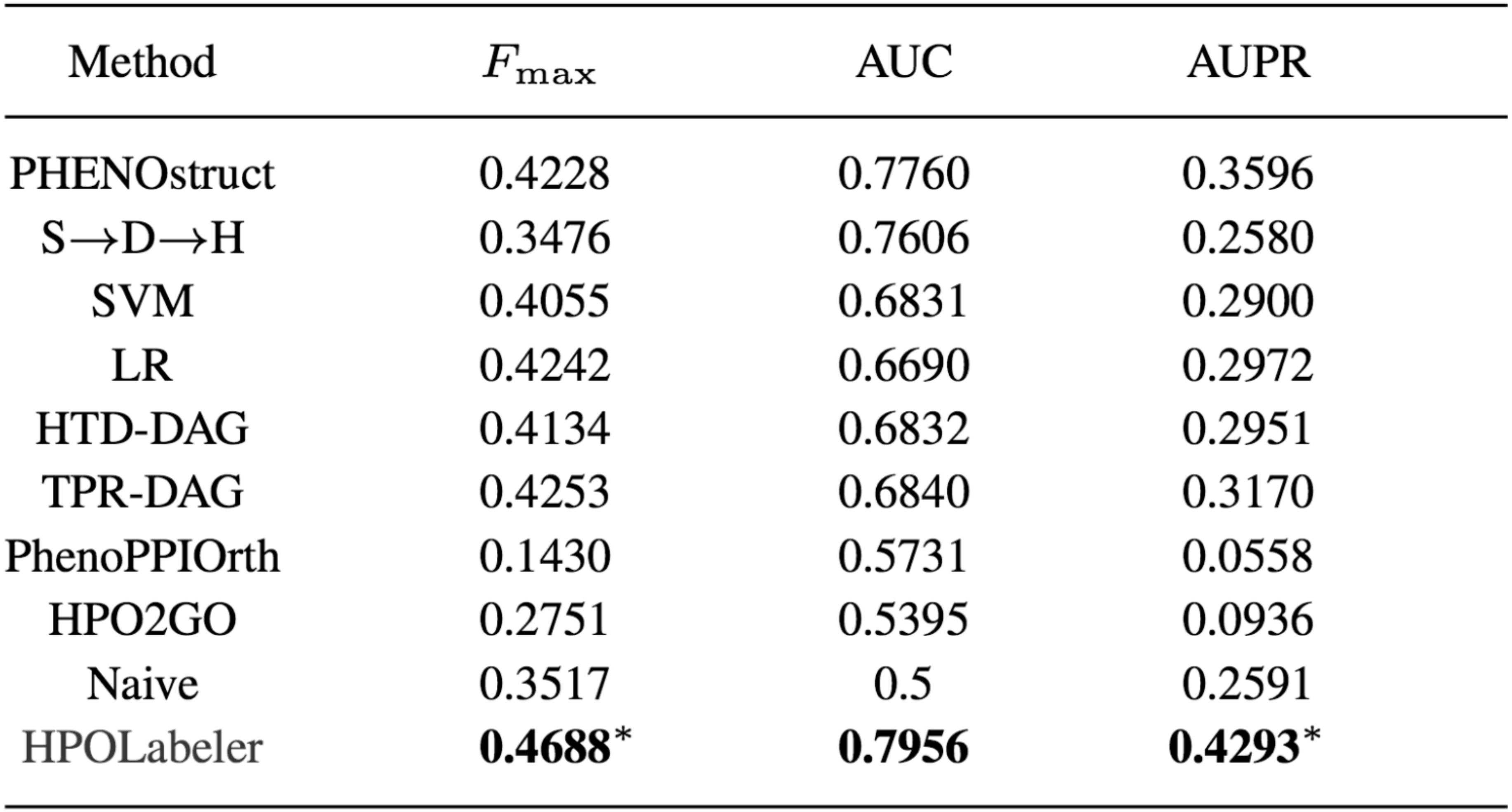
Results of Cross-validation – Basic Models



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
Naive	0.3517	0.5	0.2590

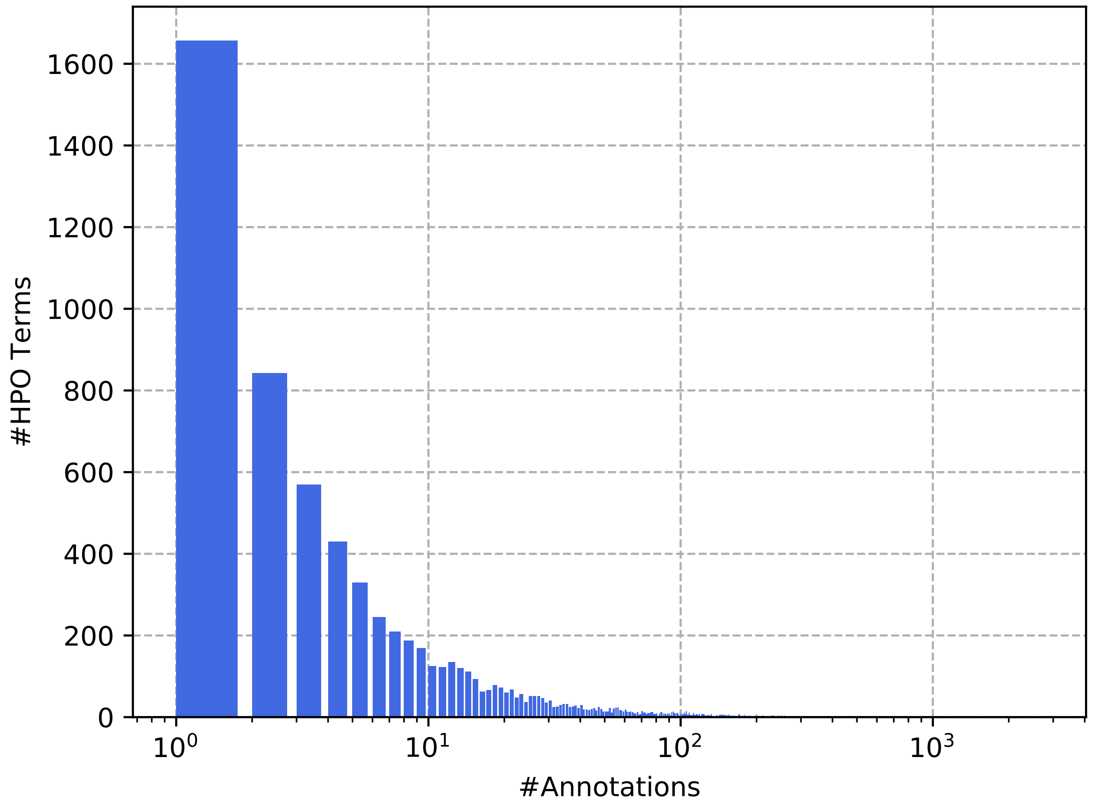
- Nearest Neighbor 👍
- PPI 👍
- NN > LR

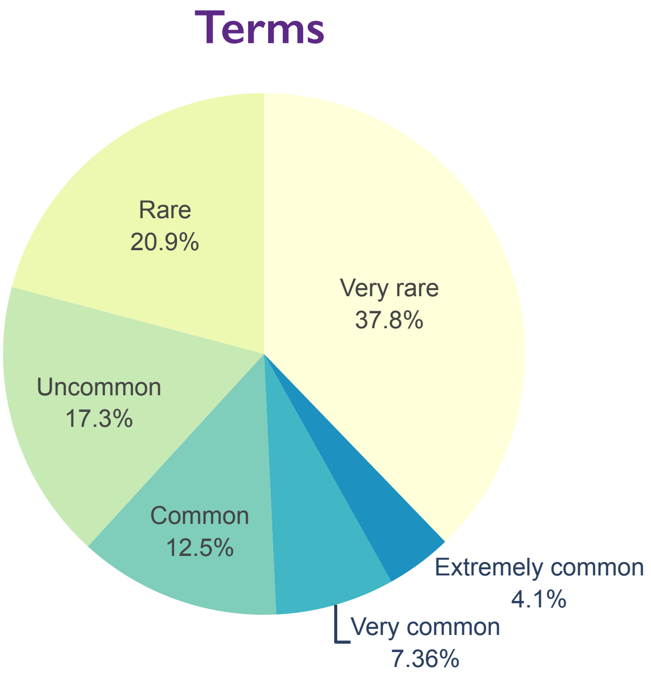
Results of Cross-validation – Comparison

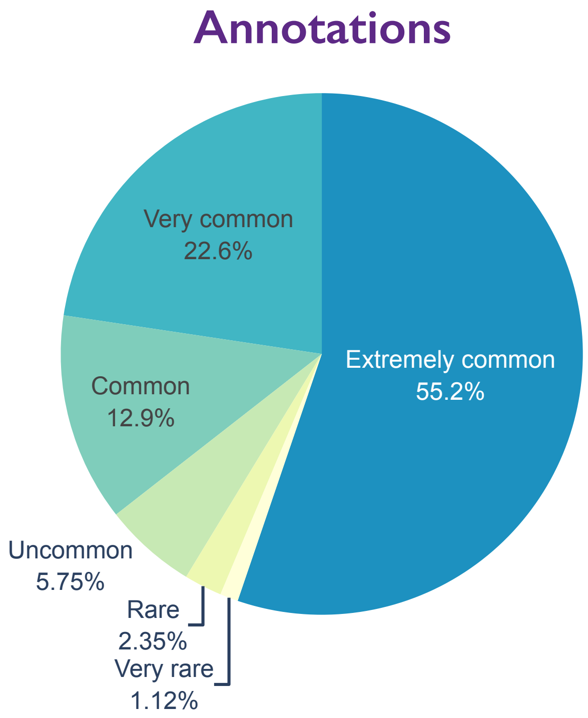


Method	$F_{\max}$	AUC	AUPR
PHENOstruct	0.4228	0.7760	0.3596
S→D→H	0.3476	0.7606	0.2580
SVM	0.4055	0.6831	0.2900
LR	0.4242	0.6690	0.2972
HTD-DAG	0.4134	0.6832	0.2951
TPR-DAG	0.4253	0.6840	0.3170
PhenoPPIOrth	0.1430	0.5731	0.0558
HPO2GO	0.2751	0.5395	0.0936
Naive	0.3517	0.5	0.2591
HPOLabeler	**0.4688***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

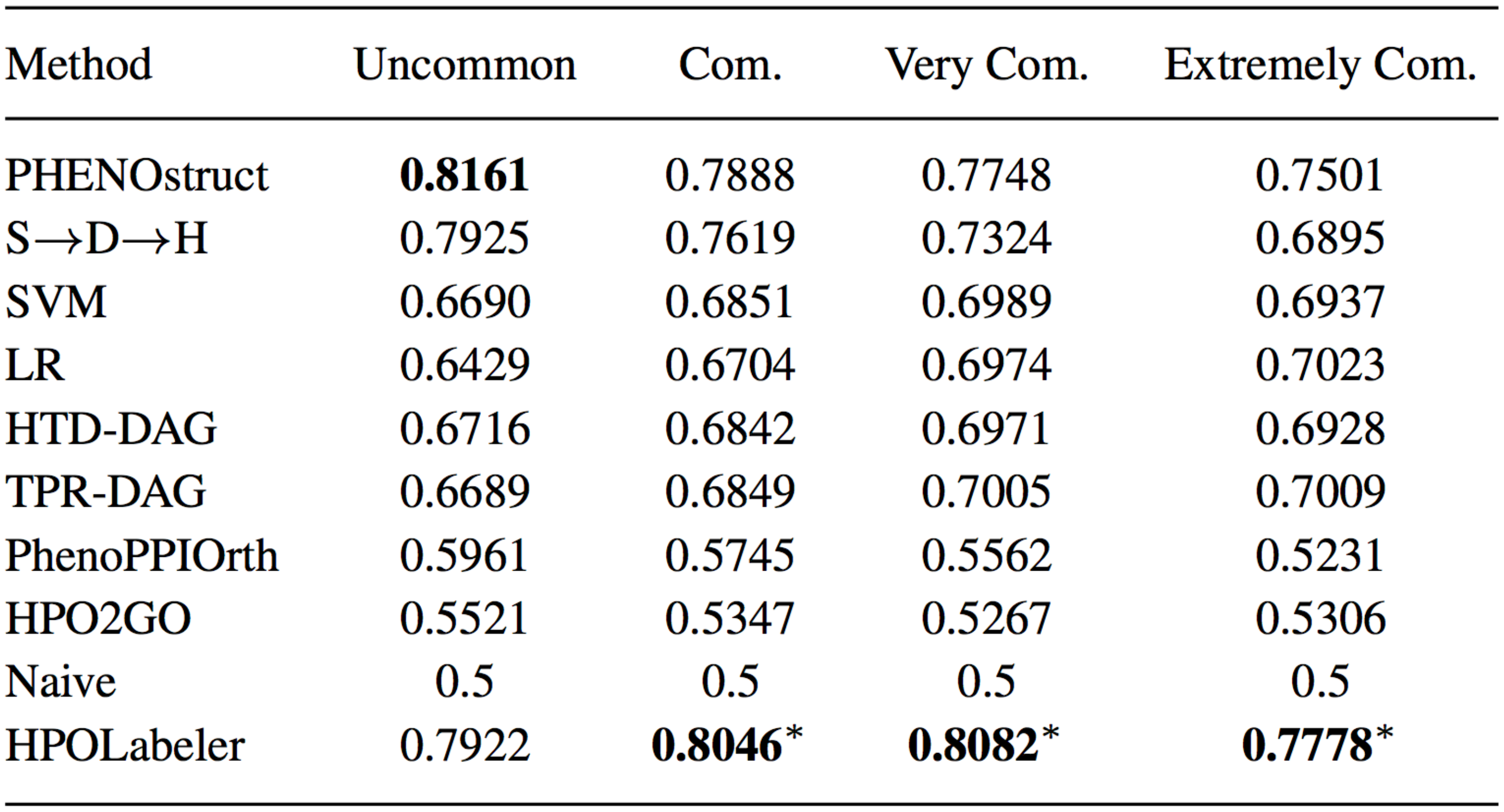
Facts: HPO and Annotations are unbalance







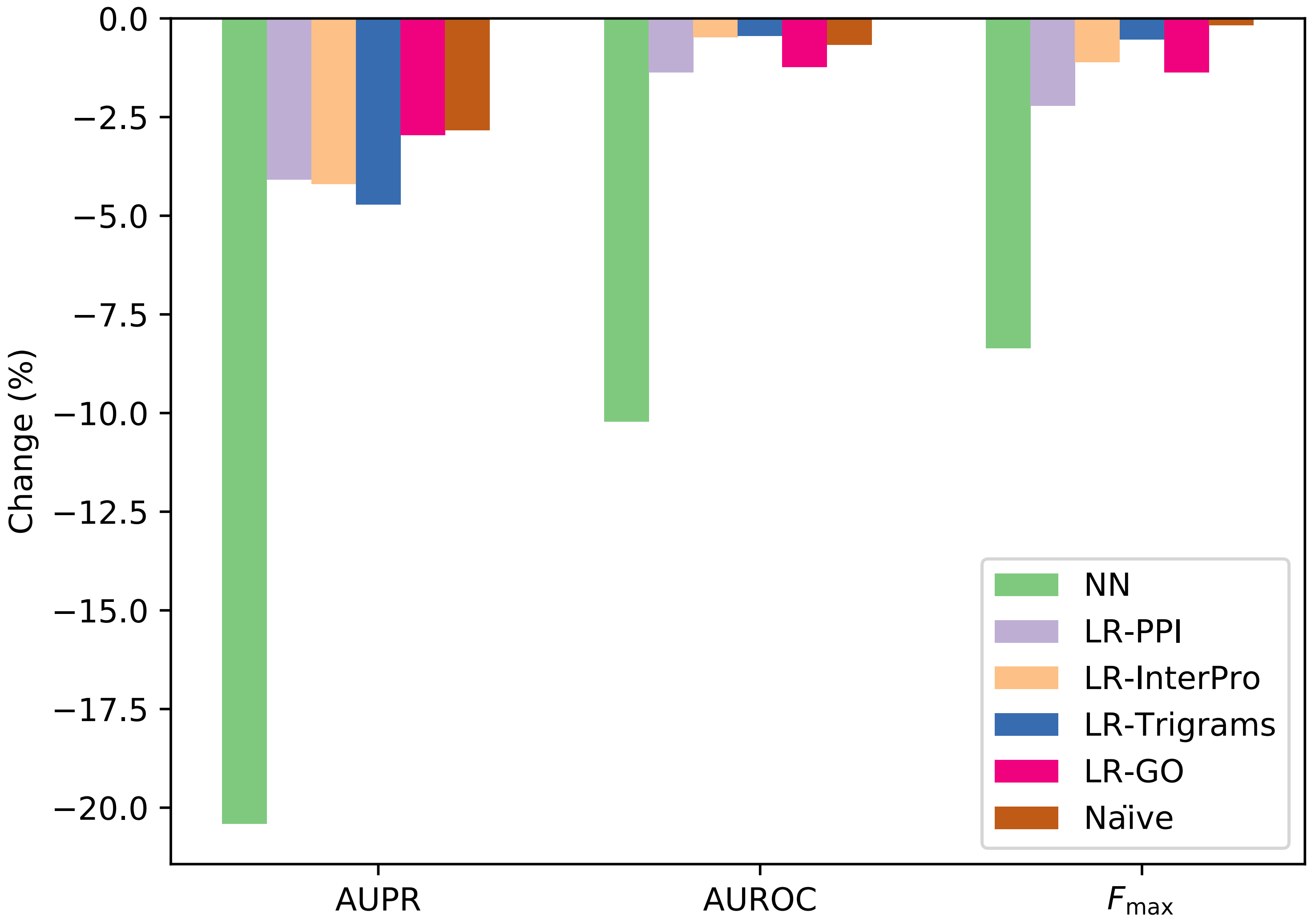
Results of CV – Avg. AUC group by frequency



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***

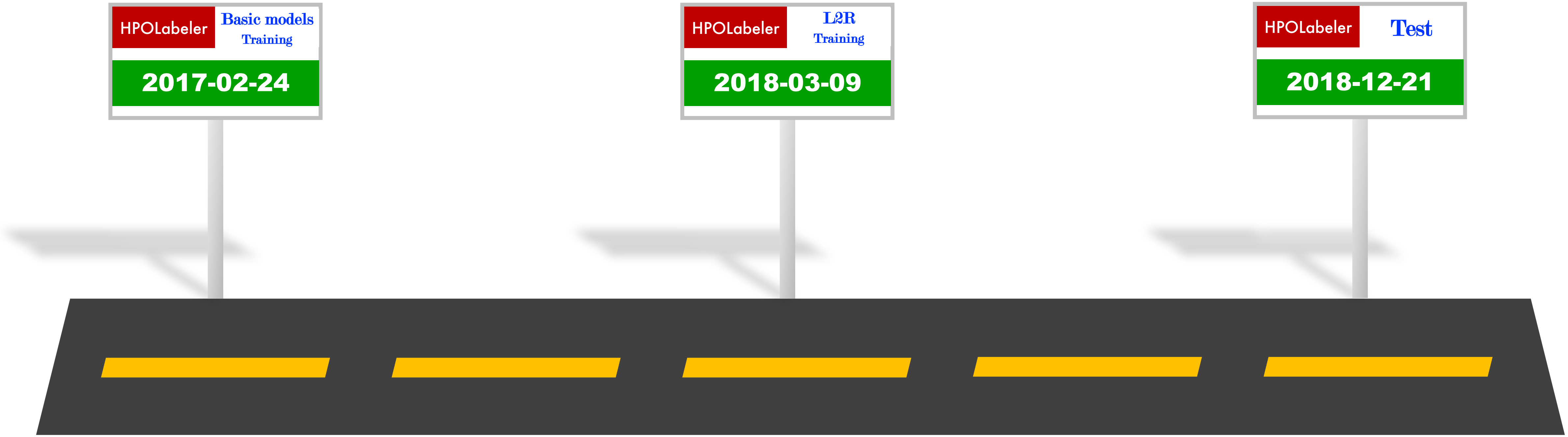
- High-frequency groups 😊
- Low-frequency groups 😐

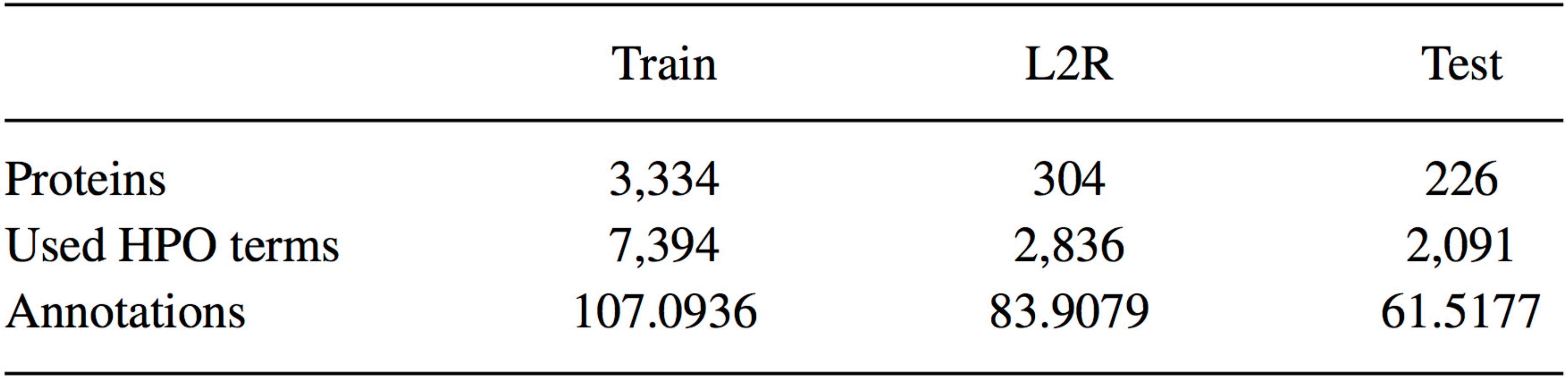
Results of CV – Leave-one-source-out



- PPI: most informative
- NN: best performing
- All changes < 0: indispensable

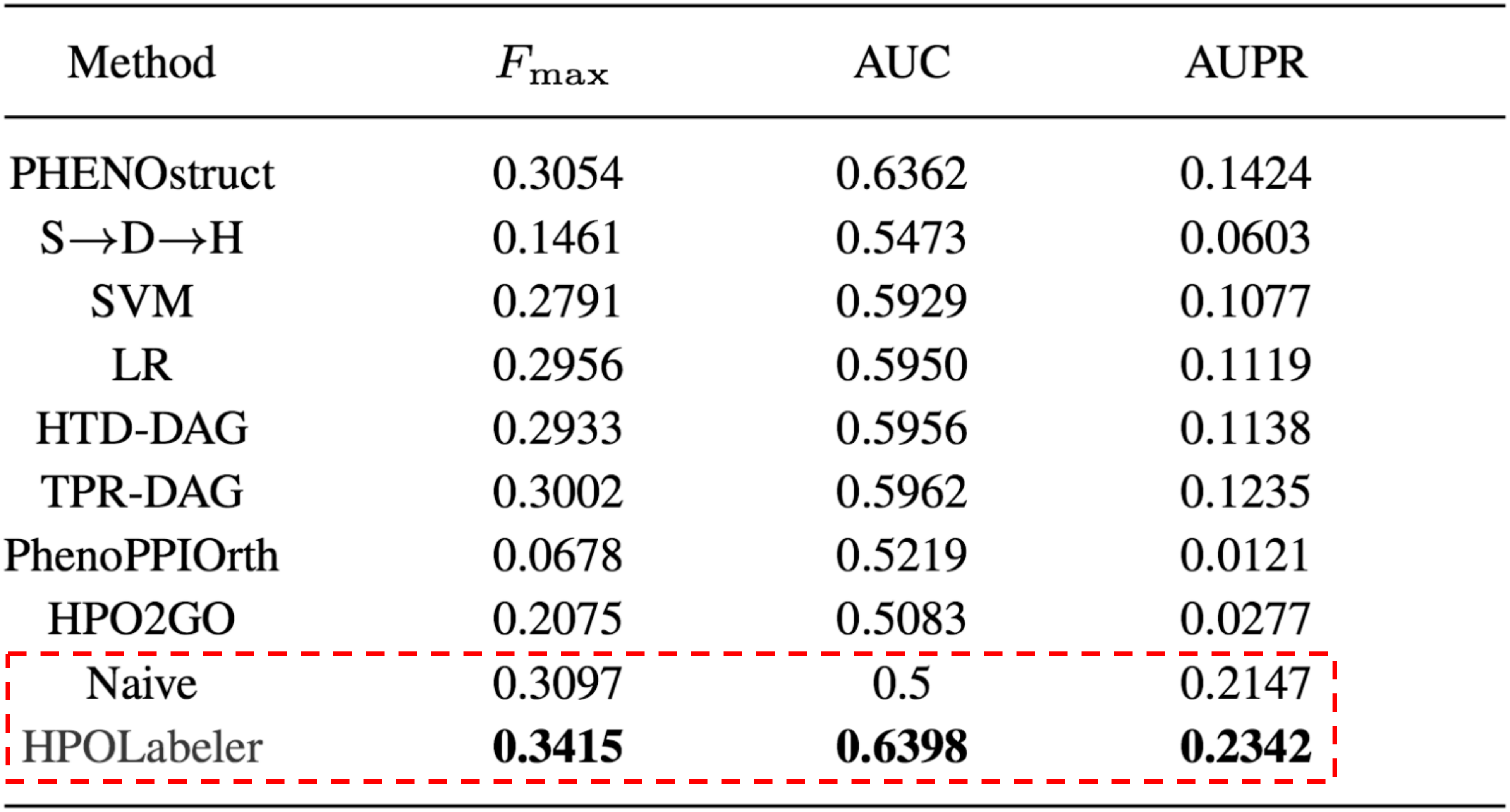
Evaluation 2: Temporal Validation





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

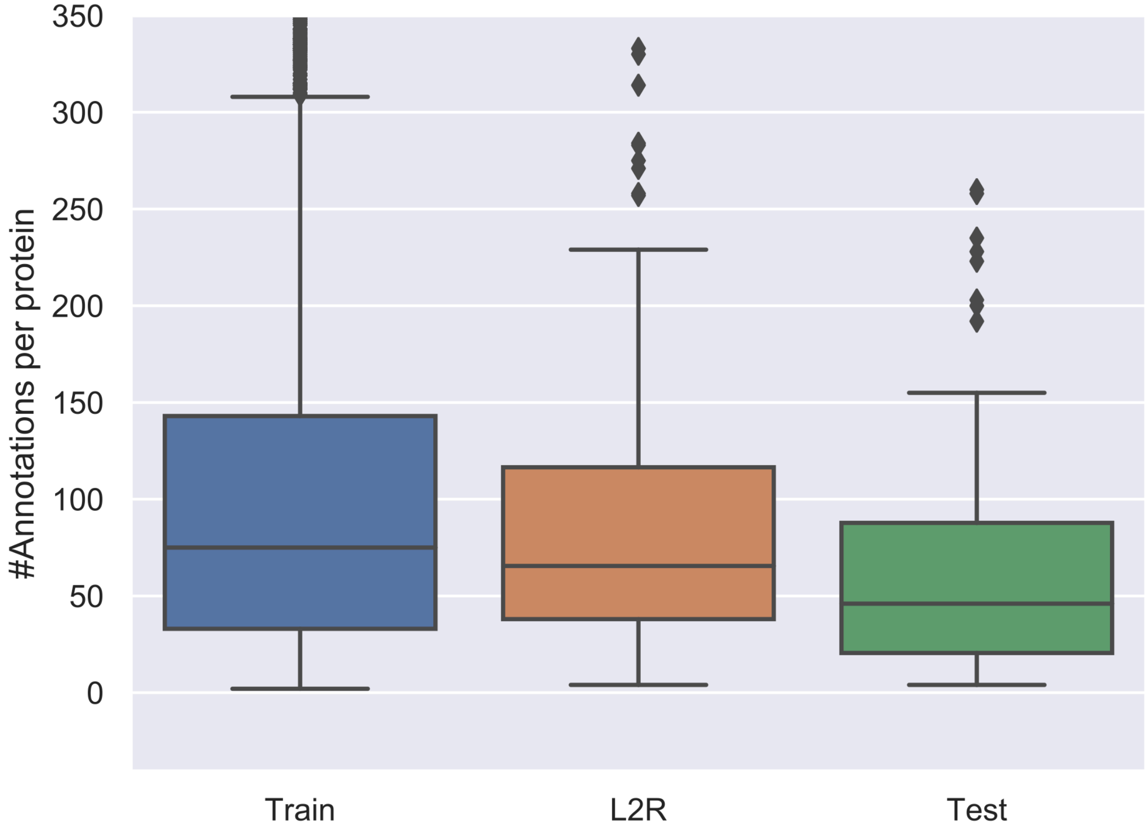
Results of Temporal Validation – Comparison



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
Naive	0.3097	0.5	0.2147
HPOLabeler	**0.3415**	**0.6398**	**0.2342**

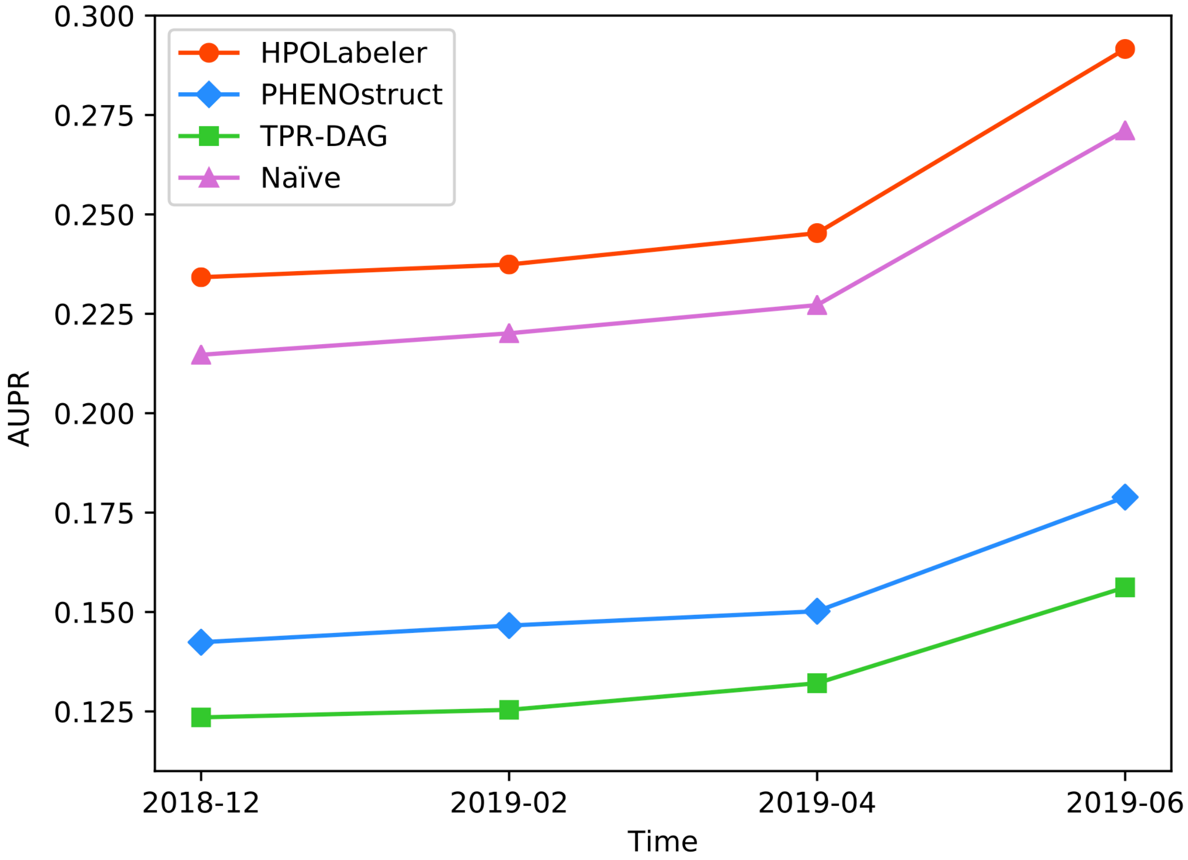
Findings: HPO annotations are incomplete

A



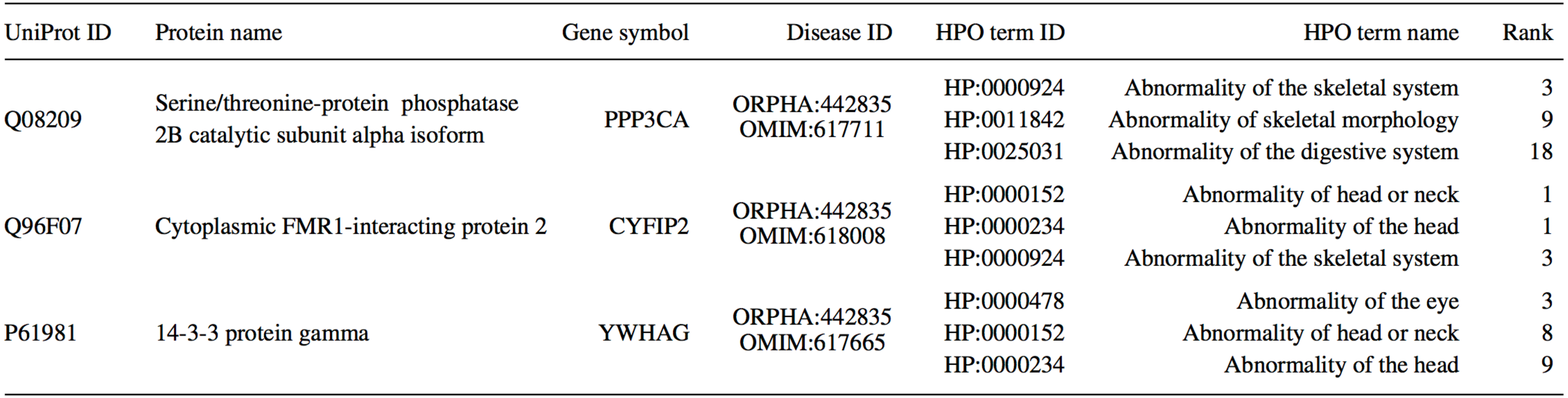
#HPO terms associated with a protein in each dataset

B



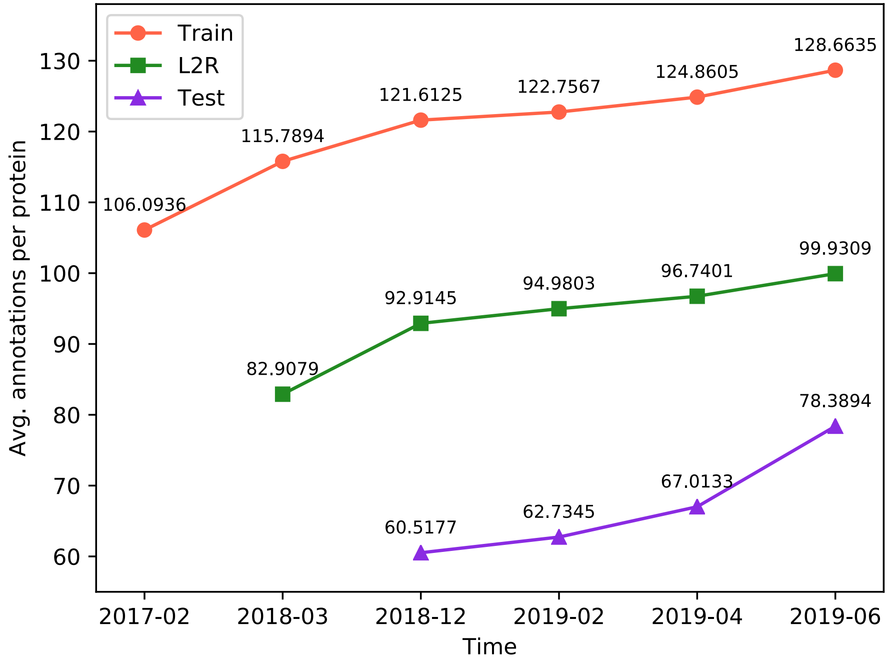
AUPRs evaluated by HPO annotations released at different times

Findings: HPO annotations are incomplete



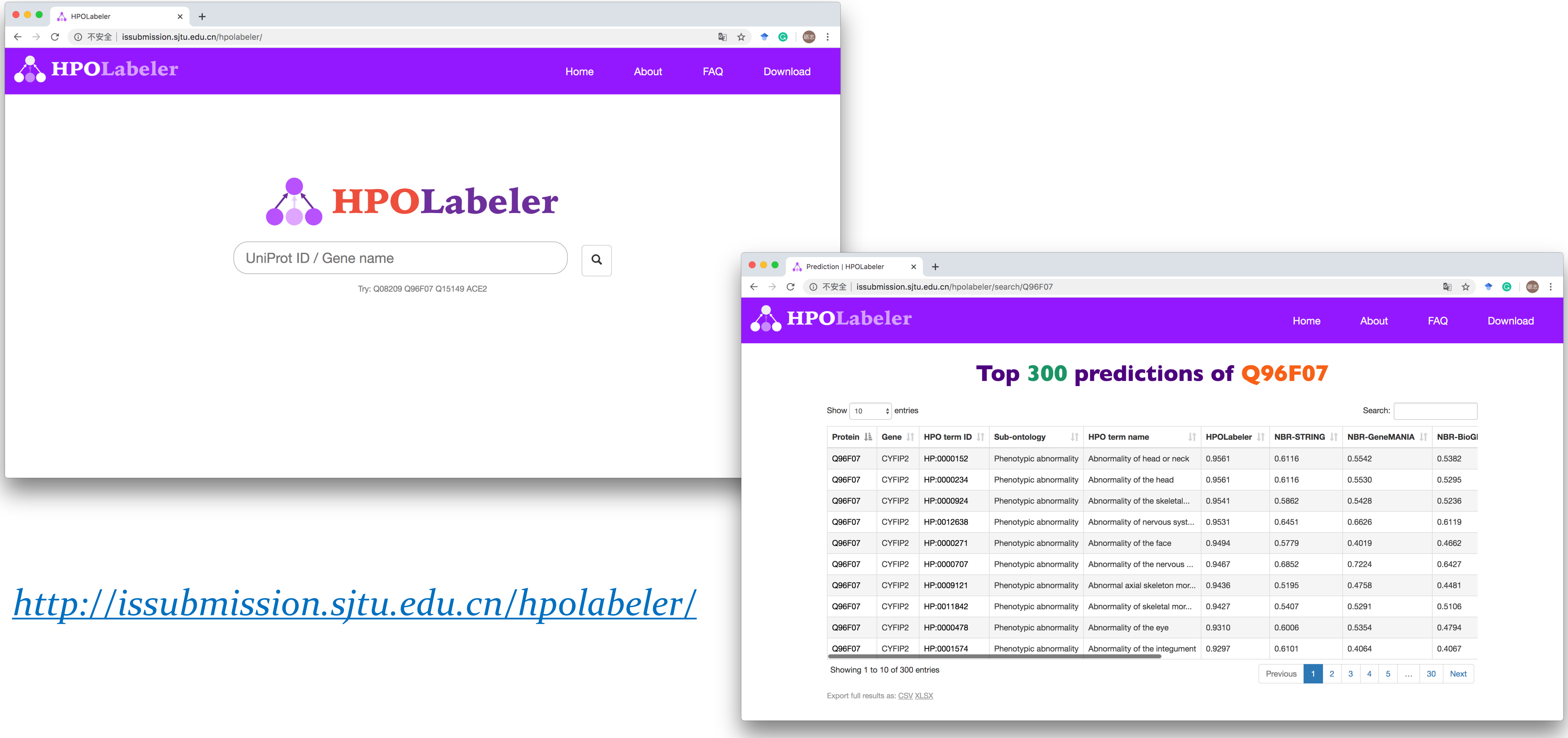
UniProt ID	Protein name	Gene symbol	Disease ID	HPO term ID	HPO term name	Rank
Q08209	Serine/threonine-protein phosphatase 2B catalytic subunit alpha isoform	PPP3CA	ORPHA:442835 OMIM:617711	HP:0000924 HP:0011842 HP:0025031	Abnormality of the skeletal system Abnormality of skeletal morphology Abnormality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

Predicted associations *(Excerpt)* which were evaluated as negatives by old annotations but appeared in the latest release in Feb. 2019, meaning that all are actually positives



Avg. #HPO annotations of newly added proteins keep increasing with time

Online Platform



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/

Conclusions

- We propose HPOLabeler, which is able to integrate diverse types of evidences including PPI, GO, InterPro and trigrams, in the framework of Learning to Rank.

- We empirically validated the performance of HPOLabeler, which significantly outperformed all competing methods.

- Further examinations of the experimental results indicate that:
  - PPI is the most informative data source;
  - lower predictive performance in temporal validation might be caused by incomplete annotations of new proteins.

- We developed an online platform:
  http://issubmission.sjtu.edu.cn/hpolabeler/

THANK YOU