# Hierarchical Ensemble Methods for Structured Prediction with Applications in Computational Biology

Computer Science Department

UNIVERSITÀ DEGLI STUDI DI MILANO

Anacleto Lab

Computational Biology and Bioinformatics

**Marco Notaro**

https://homes.di.unimi.it/notaro/

# Bioinformatics vs Computational Biology

• **Computational Biology**: is the study of Biology using computational techniques. The main goal of a computational biologist is to make new insights about Biology and living system. Then **Computational Biology** is about **Science**.

• **Bioinformatics**: is about the creation of new algorithms able to solve problems. The main goal of a bioinformatician is to build tools that can work on biological, medical and pharmaceutical data. Then **Bioinformatics** is about **Computer Science**.

- Bio-Ontologies (e.g. HPO, GO, DO): what are and why are useful in a bio-medical contest;

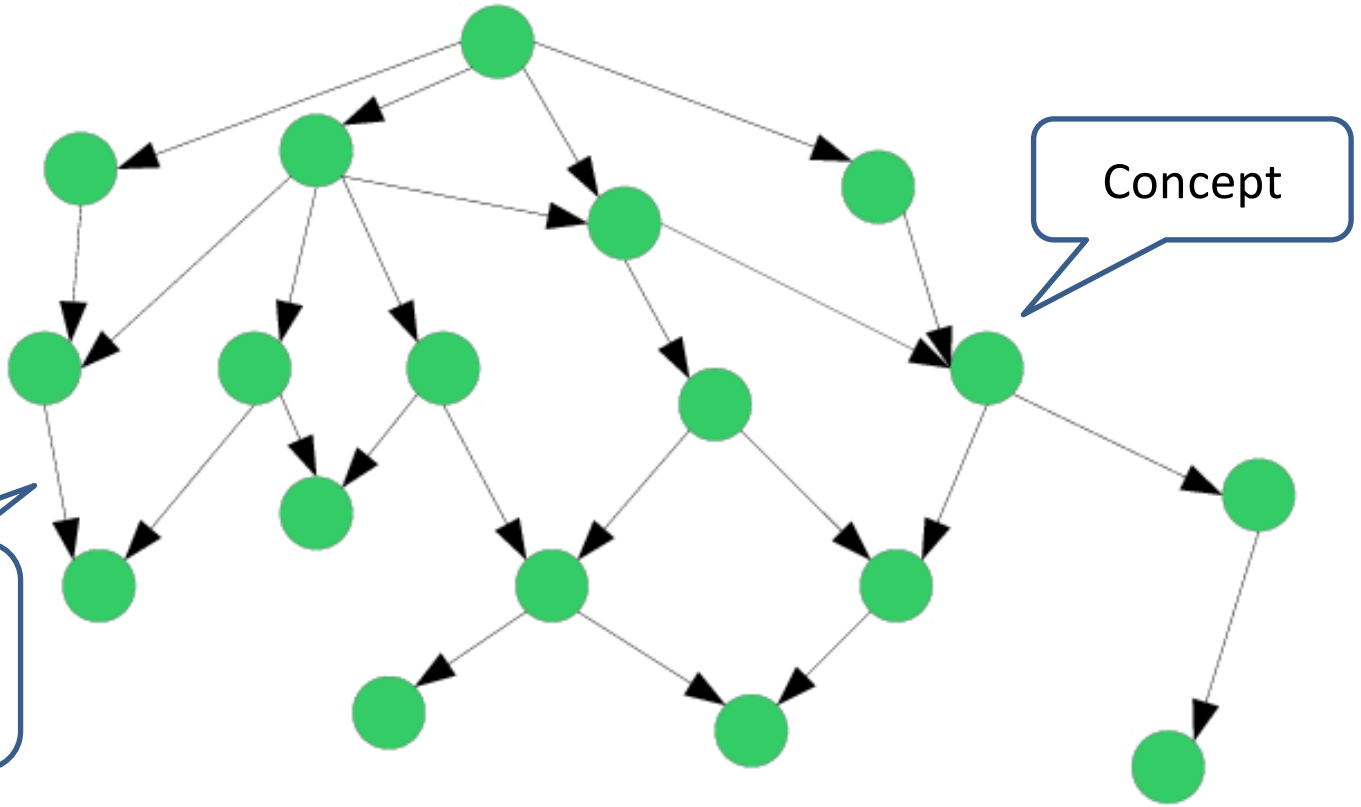
- State-of-the-art approaches ontology-based: Flat vs Hierarchy-aware learning methods;

- Proposed approaches: Hierarchical Ensemble Methods (HEM);

- Behavior of HEM in a state-of-the-art scenarios: HEM vs joint-kernel structured output method;

- Ongoing and Future Developments;

An ontology is an high-level representation of a domain of knowledge that describes concepts and semantic relationships between them in a form of Directed Acyclic Graph (DAG).



- **Human Phenotype Ontology** (HPO): provides a standardized categorization of the abnormalities associated to human diseases;
- **Disease Ontology** (DO): describes the classification of human diseases organized by etiology;
- **Gene Ontology** (GO): describes the function of genes and gene products;
- **Chemical Entities of Biological Interest** (ChEBI): structured dictionary of molecular entities focused on 'samall' chemical compound;
- **MErged Disease voCabulary** (MEDIC): map the flat list of OMIM disease terms into the hierarchical nature of the MeSH vocabulary;
- **Anatomical Ontologies** : structured controlled vocabulary of the anatomy and development of the Zebrafish (ZFO), Xenopus (XAO), Mouse (MA);

More at OBO Foundry (Open Biological and Biomedical Ontologies): http://www.obofoundry.org/

**OBO-EDIT** (http://oboedit.org/): open source ontology editor

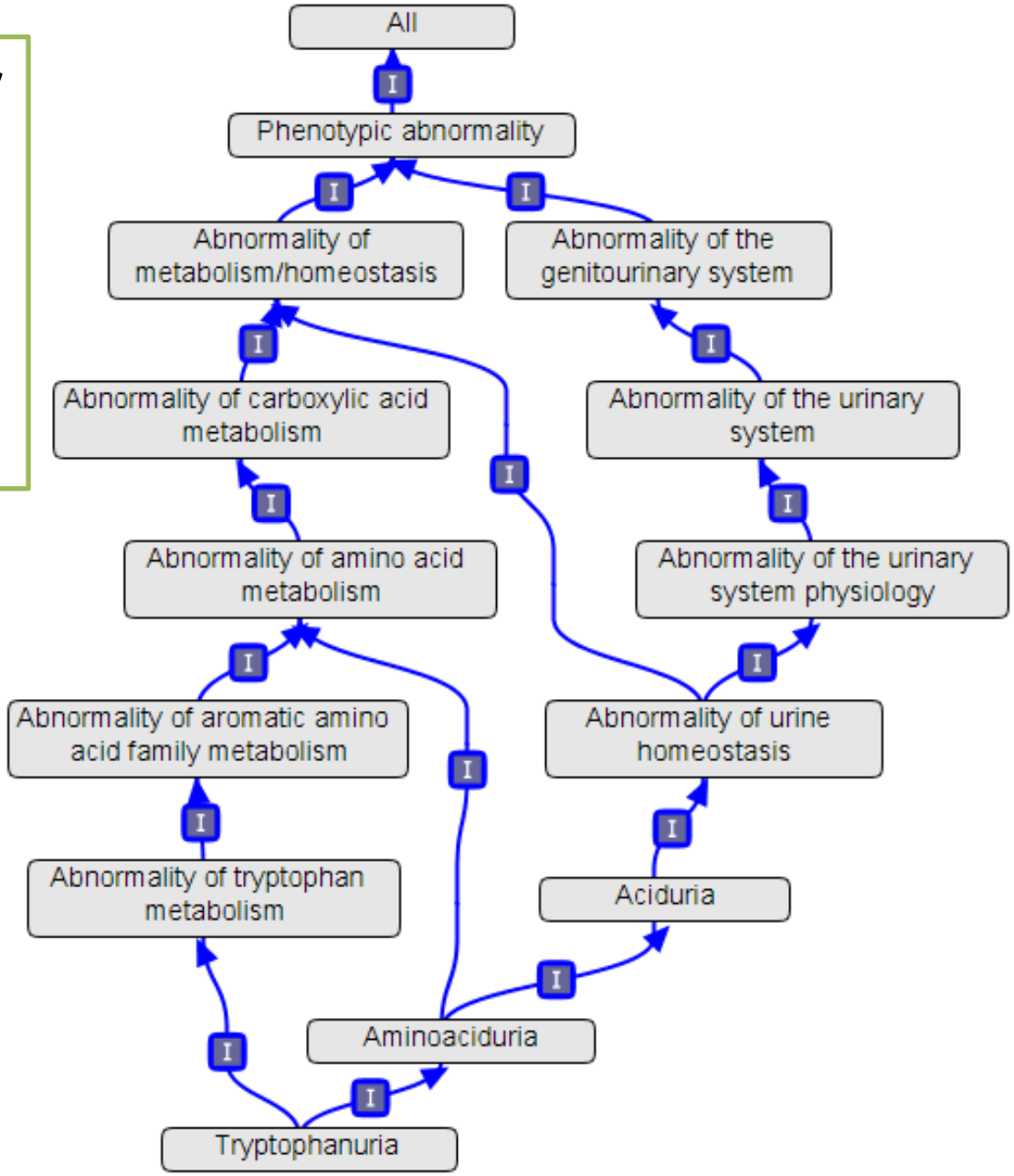
## Human Phenotype Ontology (HPO) (Köhler et al., 2017)

Link: http://human-phenotype-ontology.github.io/

What is: standardized categorization of the phenotypic abnormalities associated to human diseases

I : all relationships in the HPO are **is-a** relationships, i.e. simple class-subclass relationships

HPO Easter Release:
Tot. Number of Nodes: 12,226
Tot. Number of Edges: 16,044

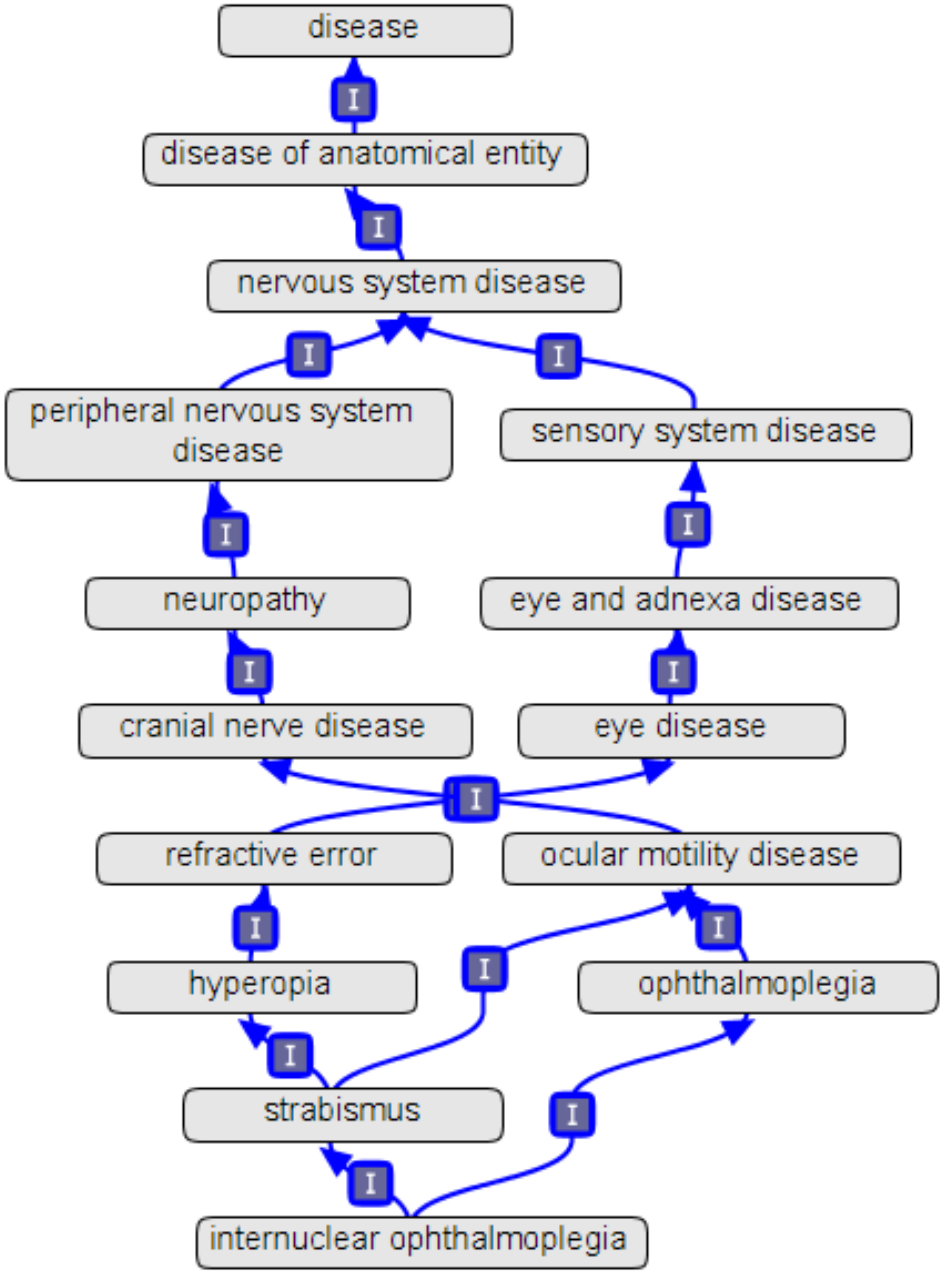


**Human Disease Ontology** (DO)
(Schriml et al., 2015)
Link: http://disease-ontology.org/

What is: controlled classification of human diseases

I : "is-a" relationships are transitive, meaning that they are inherited up all paths to the root

DO March Release:
Tot. Number of Nodes: 8,240
Tot. Number of Edges: 8,437

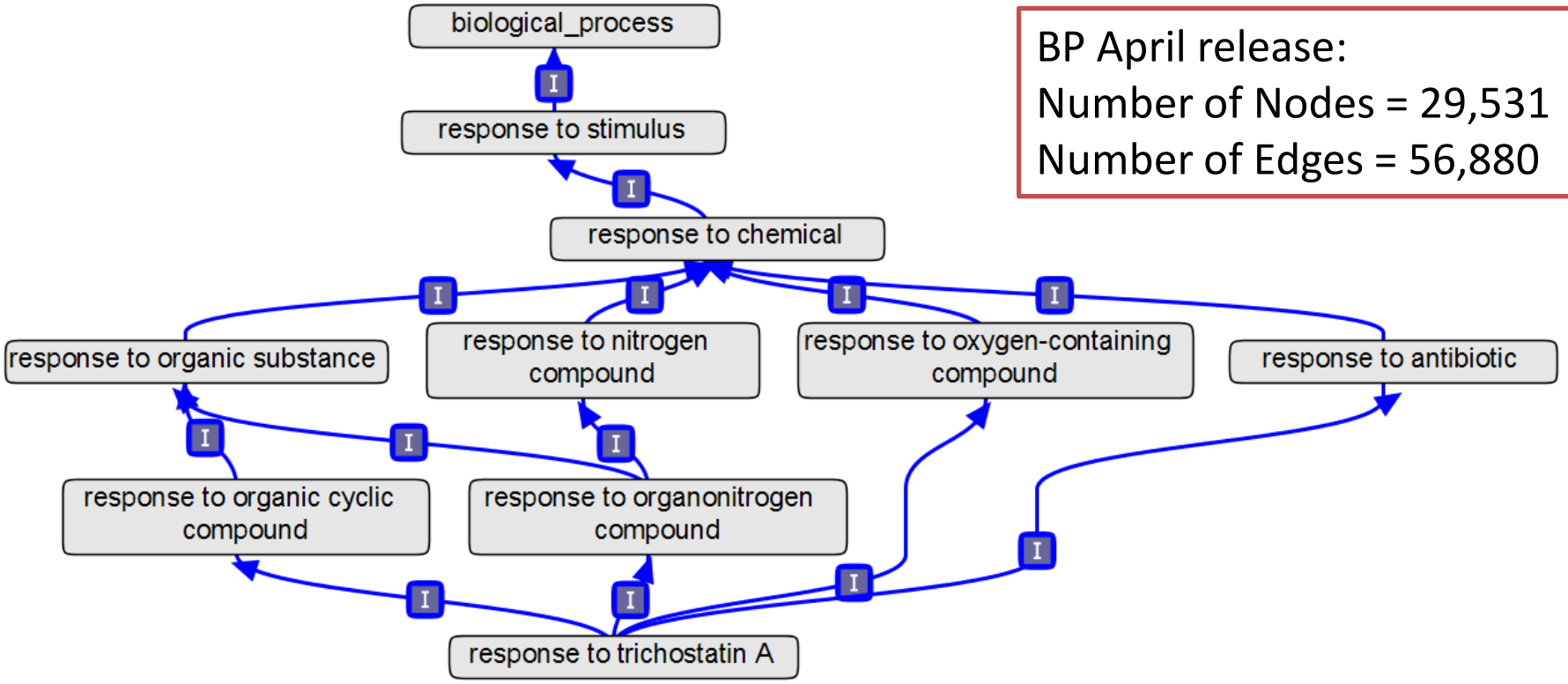


## **Gene Ontology** (GO) (Ashburner et al., 2000)

Link: http://www.geneontology.org/

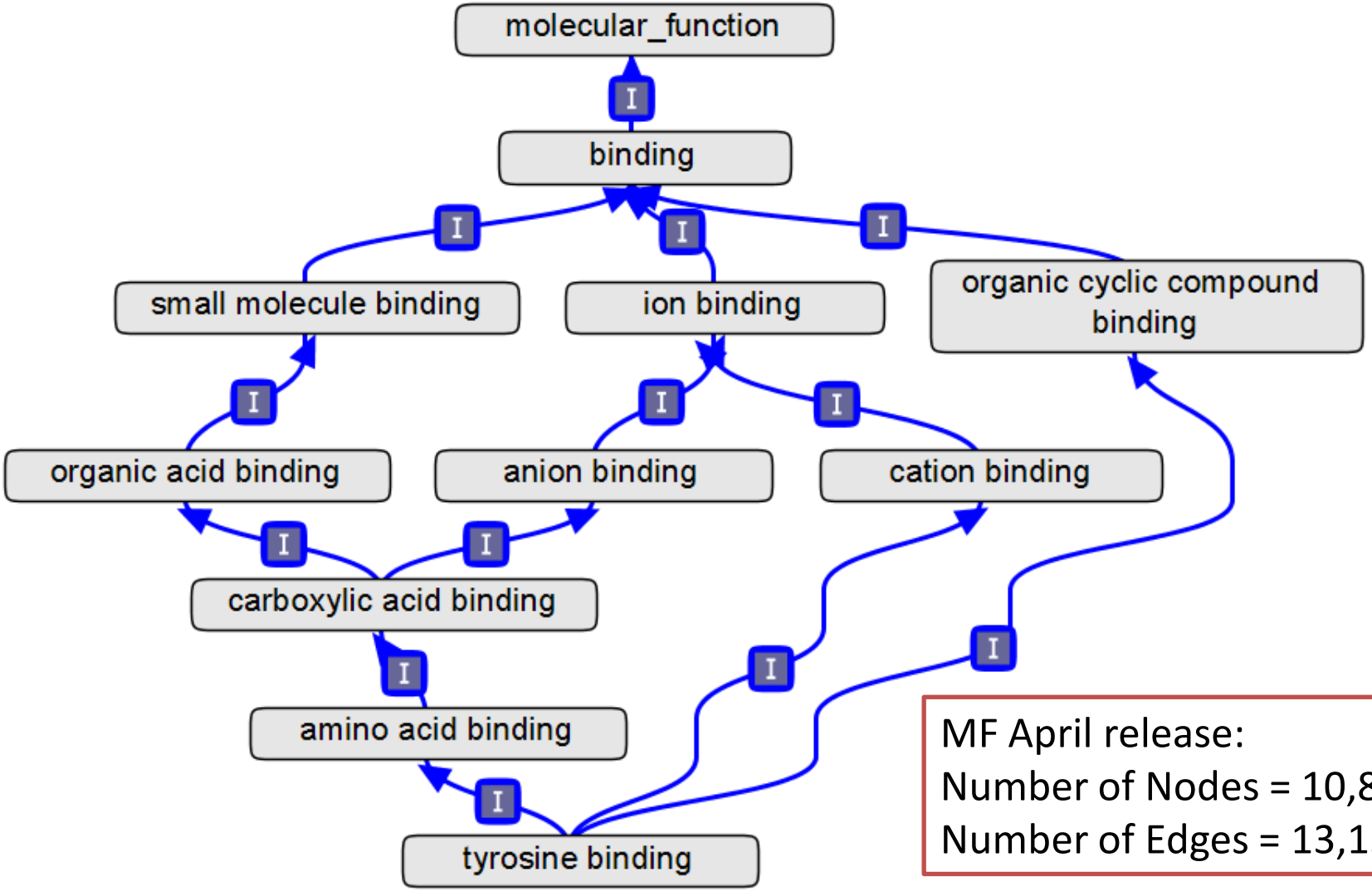
What is: three structured ontologies that describe gene products in terms of their association with BP, MF and CC in a species-independent manner.

**Biological Process** (BP) describes a collection of events carried out by one or more molecular functions (lipid metabolic process, Krebs acid cycle, antibiotic response).



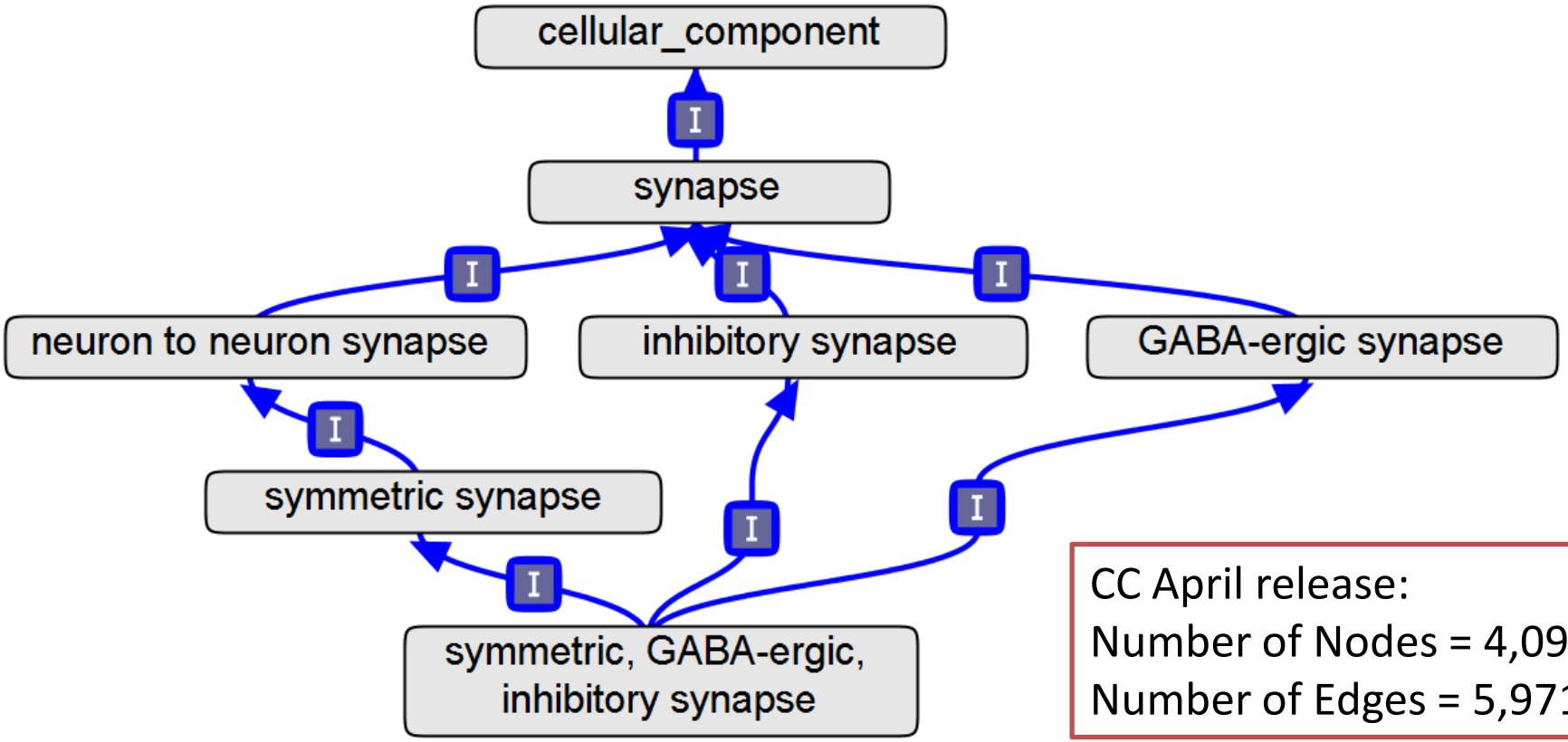

BP April release:
Number of Nodes = 29,531
Number of Edges = 56,880

**Molecular function** (MF) describes activities that occur at molecular level, such as catalytic or binding activities.




MF April release:
Number of Nodes = 10,895
Number of Edges = 13,156

**Cellular component** (CC) ontology describes locations, at the levels of subcellular structures or macromolecular complexes, in which a specific gene product is located (e.g. nucleus, nuclear inner membrane, ribosome, synapse).




CC April release:
Number of Nodes = 4,096
Number of Edges = 5,971

**Ontologies** provide predefined taxonomies for solving several relevant computational biology problems as:

- Protein Function prediction (GO);
- Prediction of human gene – abnormal phenotype associations (HP);
- Prediction of gene – disease associations (DO);

***In silico*** methods unlike ***in vitro*** methods are not costly in time and in money and can support the molecular biologist in solving several bio-medical problem:

- understanding the role of a protein in a BP;
- annotating a new gene/protein at high level of accuracy;
- solving a functional genomics problem;

# AFP: complex prediction problem characterized by several issues

- **Data Preparation**: construction, selection and normalization of the input data are complex and time-consuming. *Data preparation is relevant as algorithm design*;
- **Data-Fusion methods**: integration of multiple heterogeneous sources of data;
- **Unbalanced classification**: low number of positive and large number of negative examples;
- **Labels at different level of reliability**: each annotation is labeled with an *evidence code that* indicates how the annotation to a particular term is supported:
  - IPI/IGI: Inferred from Physical/Genetic Interaction (Experimental Evidence);
  - ISS: Inferred from Sequence Similarity (Computational Analysis Evidence)
  - TAS: Traceable Author Statement (annotation made on the basis of a statement made by the authors in the reference cited)
  - … and much more. Full set of available evidence codes at GO website;
- **Multi-class and multi-label**: thousand functional classes and multiple annotations for each gene/protein;
- **Structured multi-label classification**: terms are structured in a hierarchy ;

# AFP: complex prediction problem characterized by several issues

- **Data Preparation**: construction, selection and normalization of the input data are complex and time-consuming. *Data preparation is relevant as algorithm design*;
- **Data-Fusion methods**: integration of multiple heterogeneous sources of data;
- **Unbalanced classification**: positive examples usually largely lower than negatives;
- **Structured multi-label classification**: hierarchical relationship between classes;

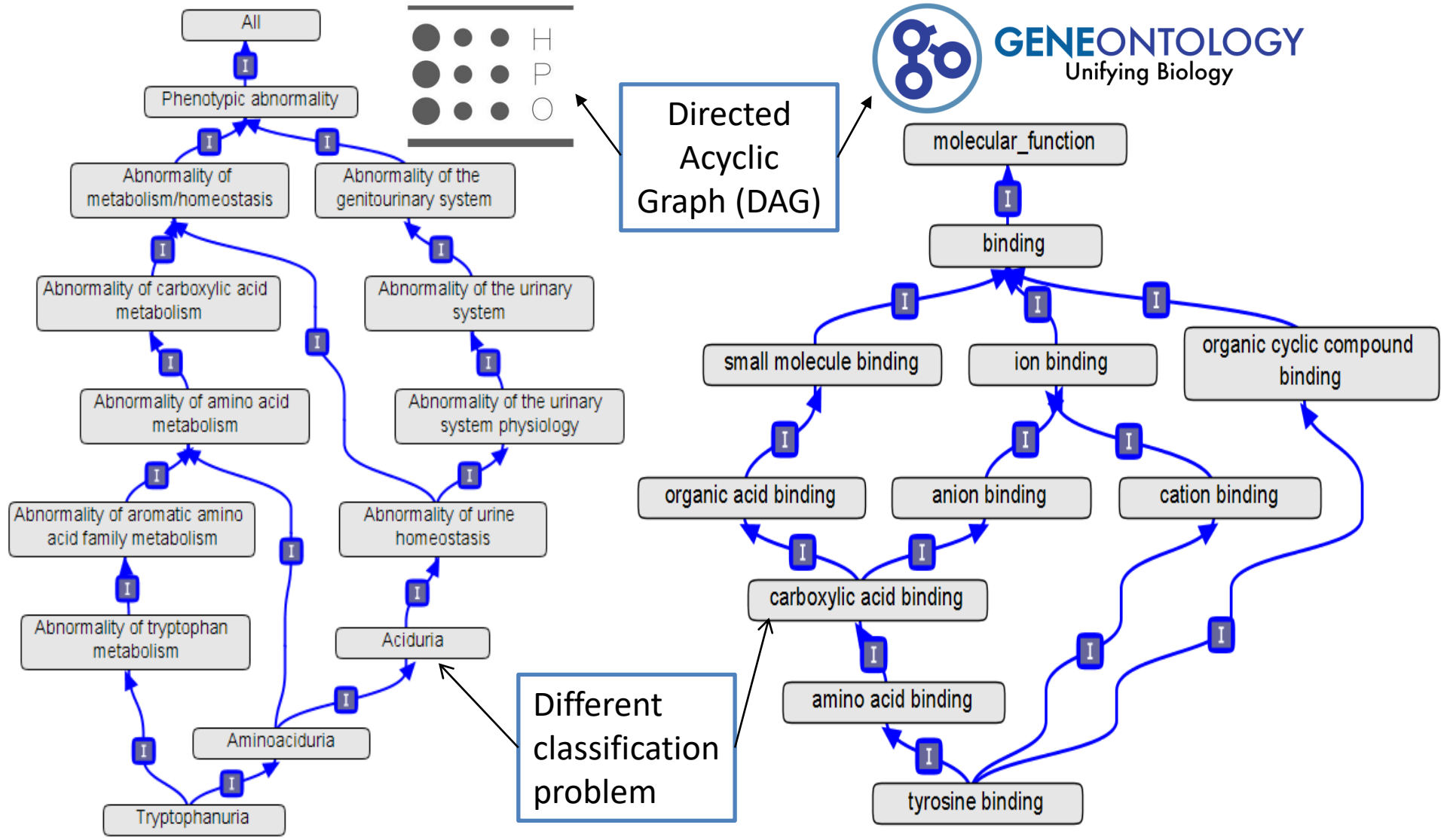
Can we design computational methods able to exploit the hierarchical relationships between ontology terms to provide biologically consistent predictions?

  - TAS: Traceable Author Statement (annotation made on the basis of a statement made by the authors in the reference cited)
  - … and much more. Full set of available evidence codes at GO website;

- **Multi-class and multi-label**: thousand of functional classes and multiple annotations for each gene/protein;

**Problem**: Hierarchical Prediction of Phenotypic Abnormalities associated to human diseases

**Problem**: Hierarchical Prediction of Protein Functions



**Flat Classifier**: predict each class separately

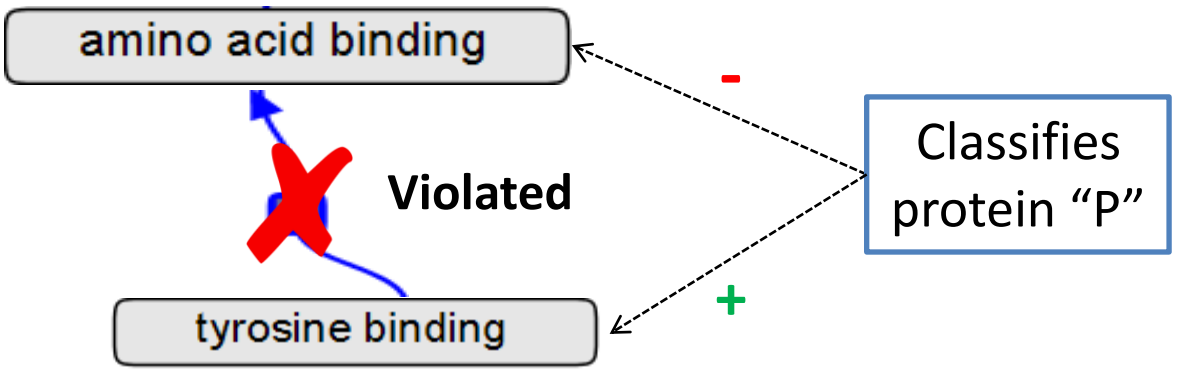
Advantage: simplicity⟶ makes prediction just for one class/term

Drawbacks:
- a priori loss of information
- neglect structural information between classes

## A Toy Example: Flat Classification

Hierarchical constraint:

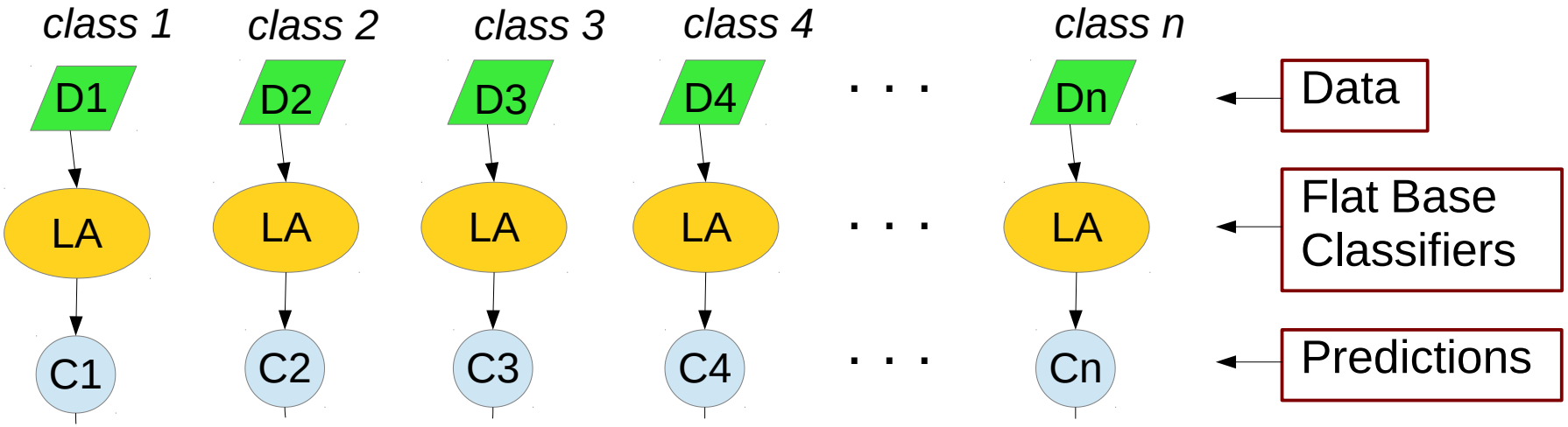
positive instance "P" for a class **implies** positive instance for all ancestors of that class



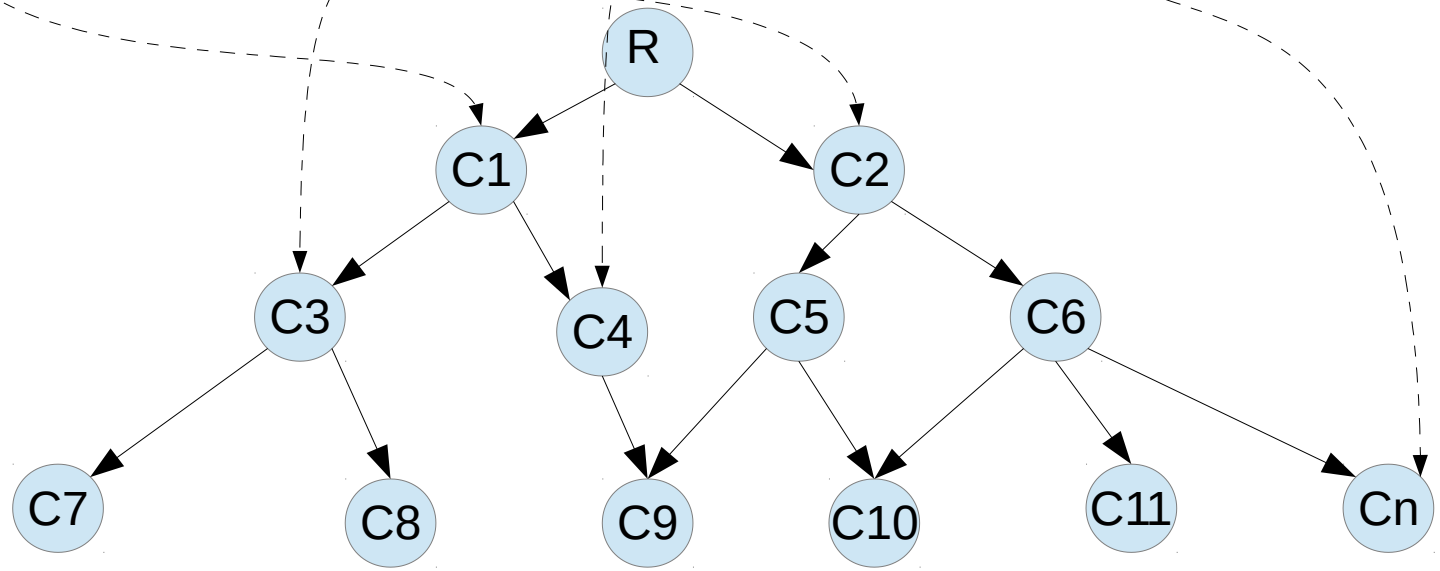
**Hierarchy-aware approaches proposed in literature:**

- Kernel-based structured output methods (Sokolov and Benhur 2010, Kahanda et al. 2015);
- Hierarchical Ensemble Methods (Silla et al. 2011, *Valentini 2014*);

# Step 1: Flat Learning of the Ontology Terms




# Step 2: Flat Predictions are Hierarchical Combined

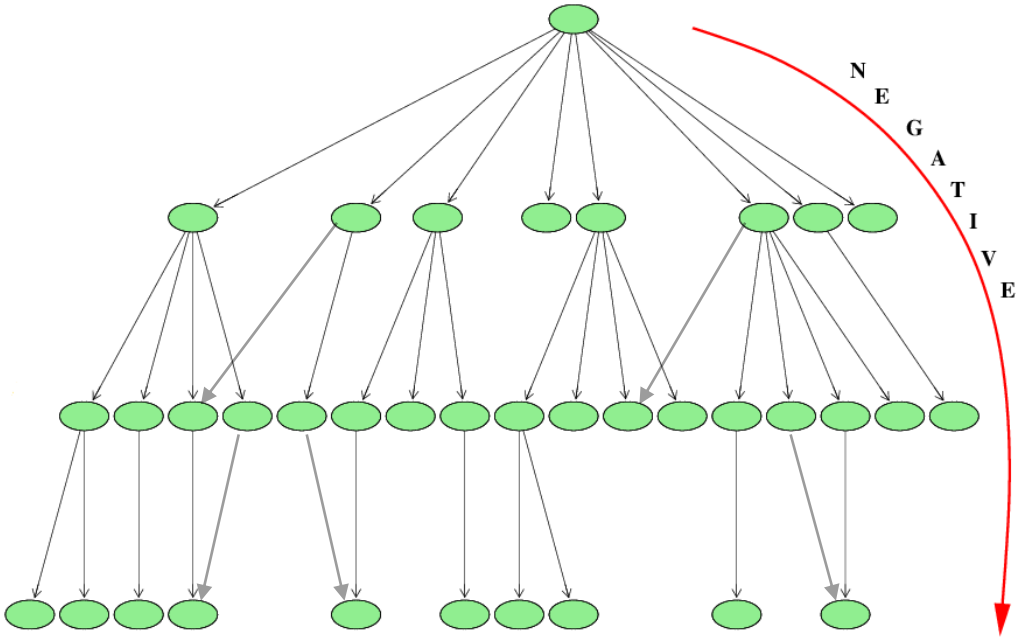


# State-of-the-art Hierarchical ensemble methods

- Most ensembles are conceived for tree-structured taxonomies (*Valentini 2011, Cesa-Bianchi et al. 2012, Paes et al. 2012, Hernandez et al. 2013*);
- Only a few for DAG-structured taxonomies (*Obozinski et al. 2008, Schietgat et al. 2010*);
- With DAG-structured taxonomies it is difficult to achieve results comparable with flat methods (*Obozinski et al. 2008*);
- DAGs are more complex than trees:
  - more parents;
  - more edges;
  - multiple paths;
  - nodes may belong to multiple levels;

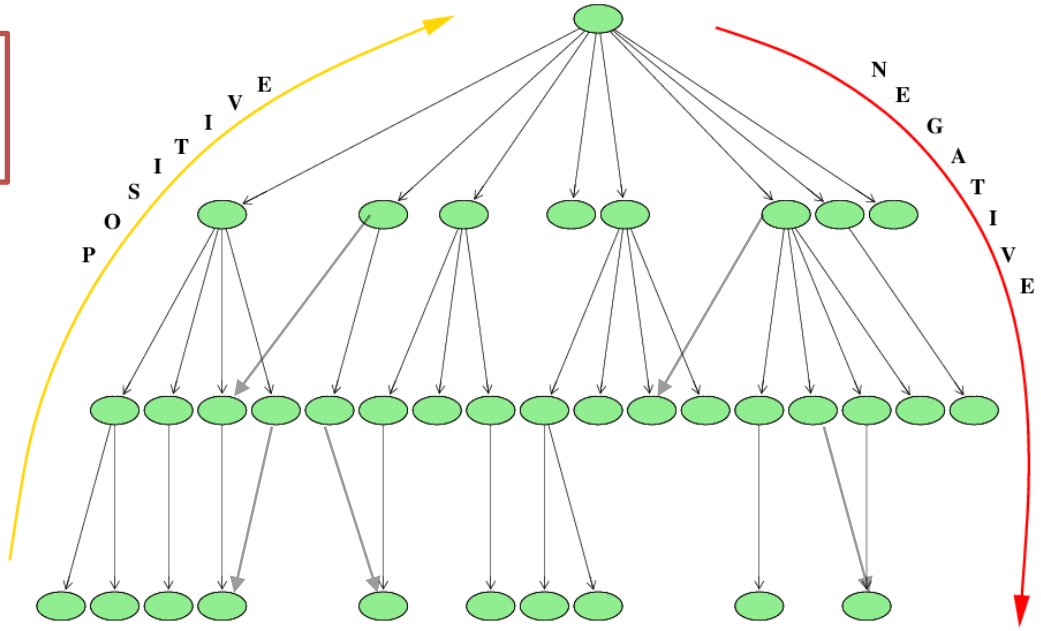
**HTD-DAG**: Hierarchical Top Down for DAGs

Just Top-Down Step



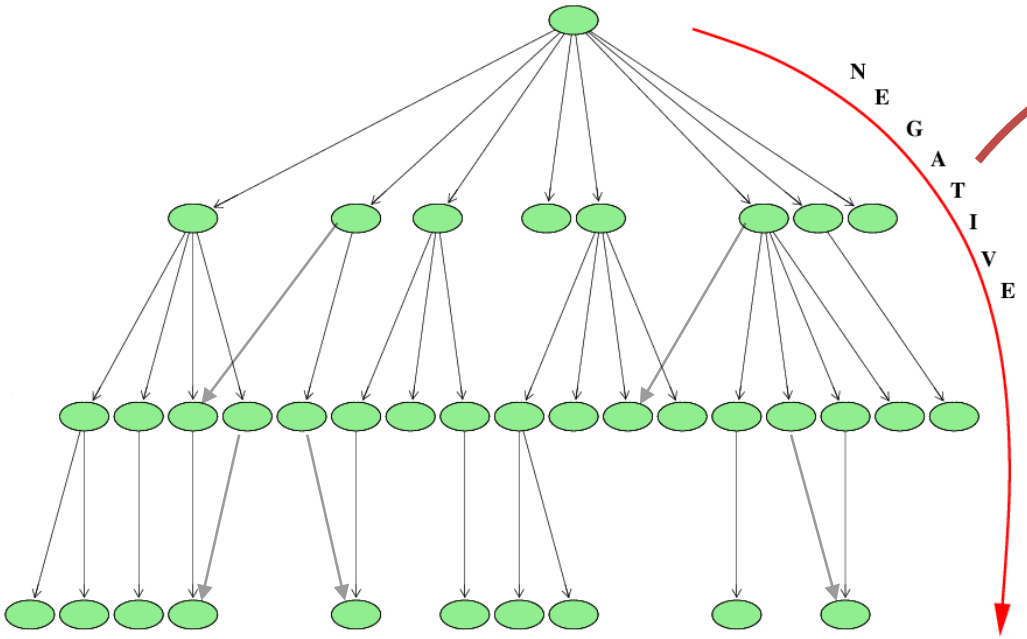

**TPR-DAG**: True Path Rule for DAGs

1. Bottom-Up Step
2. Top-Down Step



**HTD-DAG** is a two-step learning strategy:

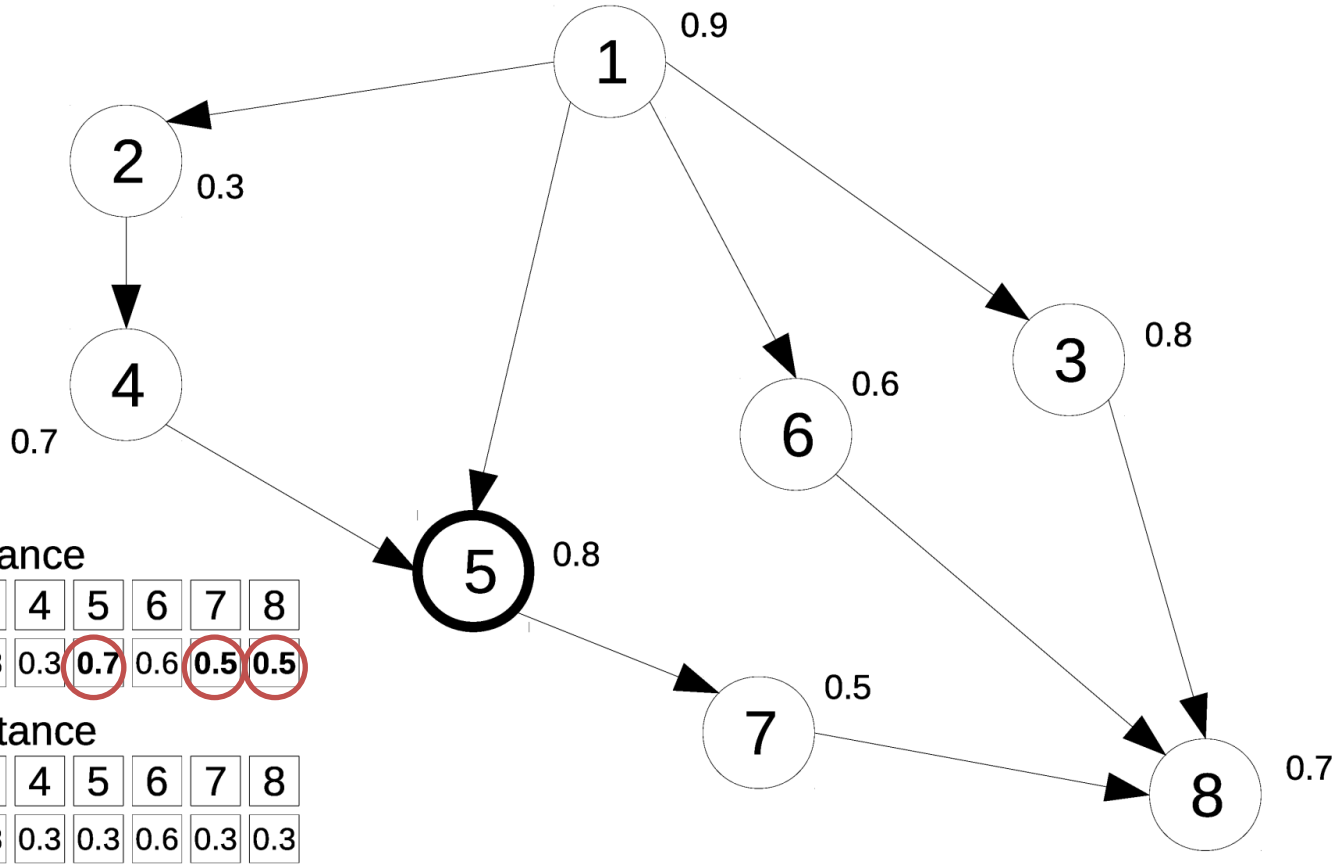
1) **Flat learning phase**: a base learner learns a specific class on a per-term basis (a set of independent classification problem);
2) **Top-Down step**: traversing the DAG by a per-level top-down visit to propagate the *negative predictions* towards the bottom of the hierarchy




- Remove constraint violations;
- Improvement of precision of the predictions;

To preserve the consistency of the predictions the levels must be defined according to the maximum distance from the root :

$$\boldsymbol{y} \text{ is consistent} \iff \forall i \in V, j \in par(i) \Rightarrow y_j \geq y_i$$




inconsistent predictions

Min. distance

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| 0.9 | 0.3 | 0.8 | 0.3 | **0.7** | 0.6 | **0.5** | **0.5** |

consistent predictions

Max. distance

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| 0.9 | 0.3 | 0.8 | 0.3 | 0.3 | 0.6 | 0.3 | 0.3 |

Input:

- $G =< V, E >$
- $\hat{\boldsymbol{y}} =< \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{|V|} >$ (flat predictions)

begin algorithm

01: A. $dist :=$ ComputeMaxDist $(G, root(G))$
02: B. Per-level top-down visit of $G$:
03: $\bar{y}_{root(G)} := \hat{y}_{root(G)}$
04: for each $d$ from 1 to $\xi$ do
05: $N_d := \{i | dist(i) = d\}$
06: for each $i \in N_d$ do
07: $x := \min_{j \in par(i)} \bar{y}_j$
08: if $(x < \hat{y}_i)$
09: $\bar{y}_i := x$
10: else
11: $\bar{y}_i := \hat{y}_i$
12: end for
13: end for

end algorithm

Output:

- $\bar{\boldsymbol{y}} =< \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{|V|} >$

Max.Dist.: Bellman-Ford or Topological sort algorithm

HTD: the nodes are processed by level in an increasing order and the HTD ensemble predictions are returned

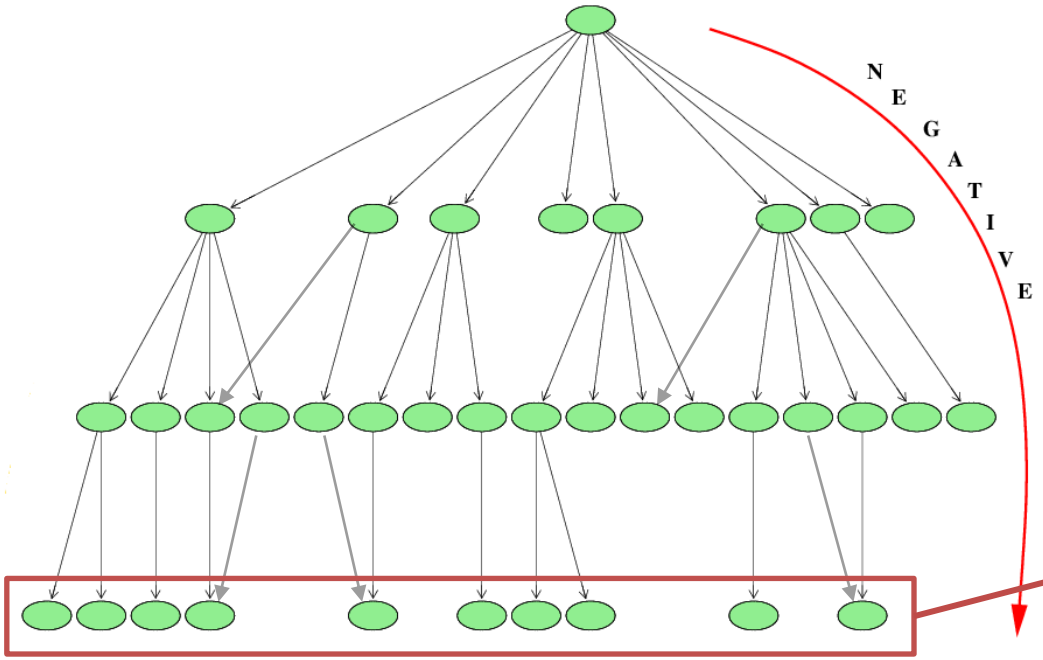
HTD-DAG Computational Complexity:

$$\mathcal{O}(|V| + |E|)$$

**HTD-DAG:**

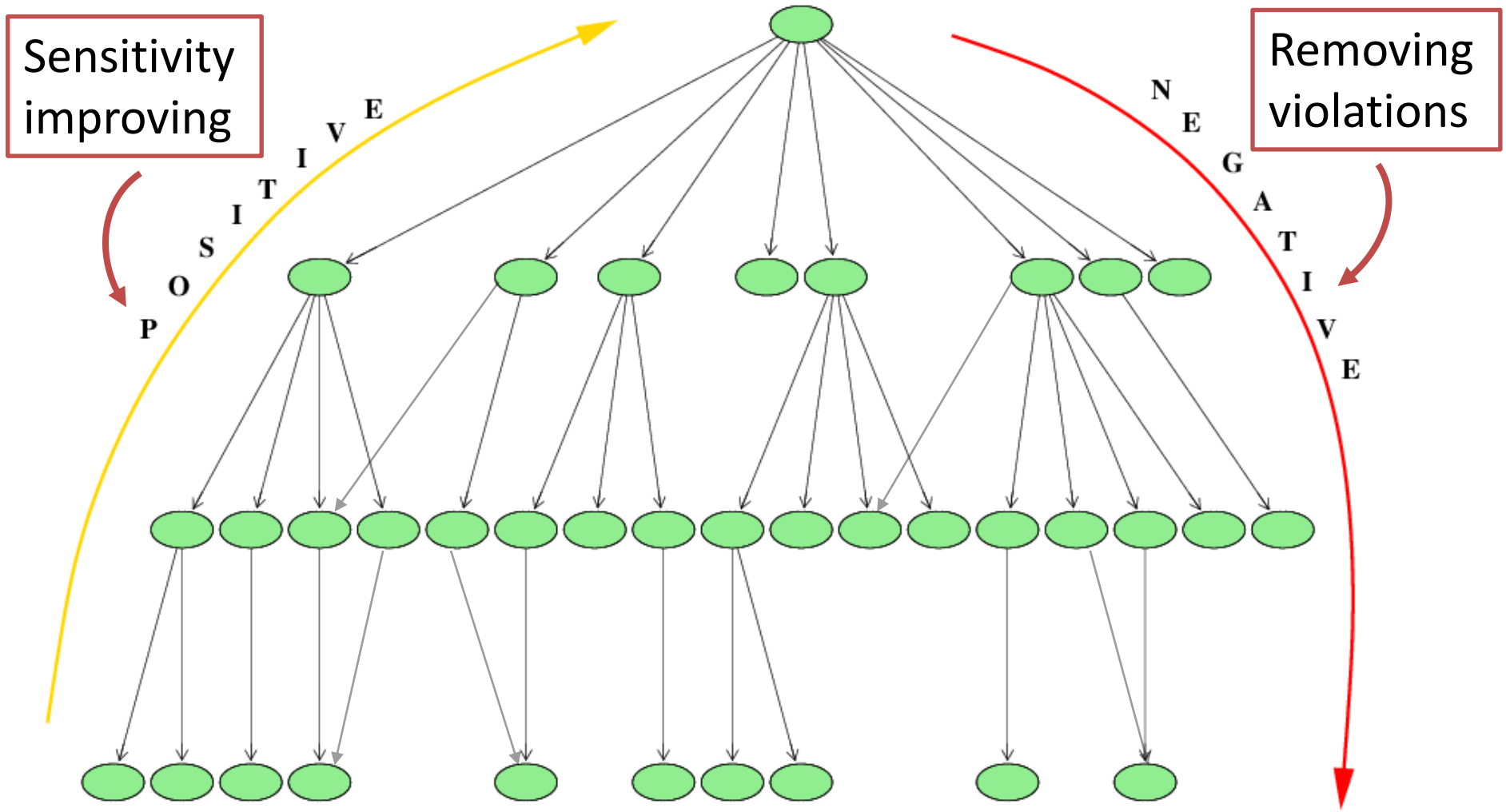
Flat scores $\hat{y}_i$ are hierarchically corrected to $\bar{y}_i$ according to this simple rule:

$$\bar{y}_i := \begin{cases} \hat{y}_i & \text{if} \quad i \in root(G) \\ \min_{j \in par(i)} \bar{y}_j & \text{if} \quad \min_{j \in par(i)} \bar{y}_j < \hat{y}_i \\ \hat{y}_i & \text{otherwise} \end{cases}$$




Limit case: predictions at leaves nodes are negatives

# TPR ensemble for DAGs: double flow of information



In the Bottom-Up Step the ensemble decision is modified in according to this simple rule:

$$\bar{y}_i := \frac{1}{1+|\phi_i|}(\hat{y}_i + \sum_{j \in \phi_i} \bar{y}_j) \qquad \textbf{1)}$$

Different strategy can be used to define the positive $\phi_i$ children of class *i*:

**A) Adaptive Threshold Strategy**: maximize $\mathcal{M}$ on training data by internal cv

$$\phi_i := \{j \in child(i) | \bar{y}_j > t_j^*, t_j^* = \arg\max_t \mathcal{M}(j,t)\}$$

**B) Threshold Free Strategy**: positive nodes those that achieve a score higher than that of their parents

$$\phi_i := \{j \in child(i) | \bar{y}_j > \hat{y}_i\}$$

# TPR-DAG is a family of algorithms

**2)** **Weighted TPR**: $w \in [0,1]$ to balance the contribution between node *i* and that of its 'positive' children

$$\bar{y}_i := w\hat{y}_i + \frac{(1-w)}{|\phi_i|}\sum_{j\in\phi_i}\bar{y}_j$$

**3)** **Descendants TPR**: to enhance the contribution of the most specific nodes we can consider the descendants instead of children

$$\bar{y}_i := \frac{1}{1+|\Delta_i|}(\hat{y}_i + \sum_{j\in\Delta_i}\bar{y}_j) \qquad \Delta_i = \{j \in desc(i) | \bar{y}_j > t_j\}$$

**4)** **Descendants-TAU**: $\tau \in [0,1]$ to balance the contribution between $\phi_i$ e $\delta_i$

$$\bar{y}_i := \frac{\tau}{1+|\phi_i|}(\hat{y}_i + \sum_{j\in\phi_i}\bar{y}_j) + \frac{1-\tau}{1+|\delta_i|}(\hat{y}_i + \sum_{j\in\delta_i}\bar{y}_j) \qquad \delta_i = \Delta_i \setminus \phi_i$$

Input:
- $G =< V, E >$
- $V = \{1, 2, \ldots, |V|\}$
- $\hat{\boldsymbol{y}} =< \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{|V|} >, \quad \hat{y}_i \in [0, 1]$

begin algorithm

A. Compute $\forall i \in V$ the max distance from $root(G)$:
  $E' := \{e' | e \in E, e' = -e\}$
  $G' :=< V, E' >$
  $dist :=$ Bellman.Ford$(G', root(G'))$

B. Per-level bottom-up visit of $G$:
  for each $d$ from $\max(dist)$ to 0 do
    $N_d := \{i | dist(i) = d\}$
    for each $i \in N_d$ do
      Select the set $\phi_i$ of "positive" children
      $\bar{y}_i := \frac{1}{1+|\phi_i|}(\hat{y}_i + \sum_{j \in \phi_i} \bar{y}_j)$
    end for
  end for

C. Per-level top-down visit of $G$:
  $\hat{\boldsymbol{y}} := \bar{\boldsymbol{y}}$
  for each $d$ from 1 to $\max(dist)$ do
    $N_d := \{i | dist(i) = d\}$
    for each $i \in N_d$ do
      $x := \min_{j \in par(i)} \bar{y}_j$
      if $(x < \hat{y}_i)$
        $\bar{y}_i := x$
      else
        $\bar{y}_i := \hat{y}_i$
    end for
  end for

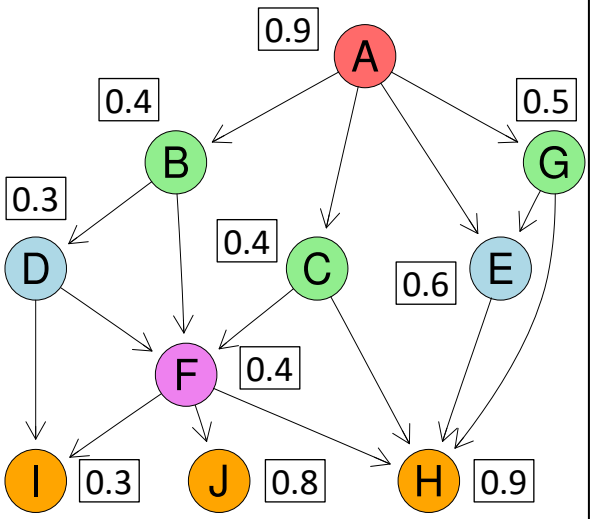
end algorithm

Output:
- $\bar{\boldsymbol{y}} =< \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{|V|} >$

Propagation of the positive predictions towards the top of the hierarchy in order to enhance the sensitivity of the predictions

TPR-DAG scales linearly with the number of classes (since the graph is sparse)

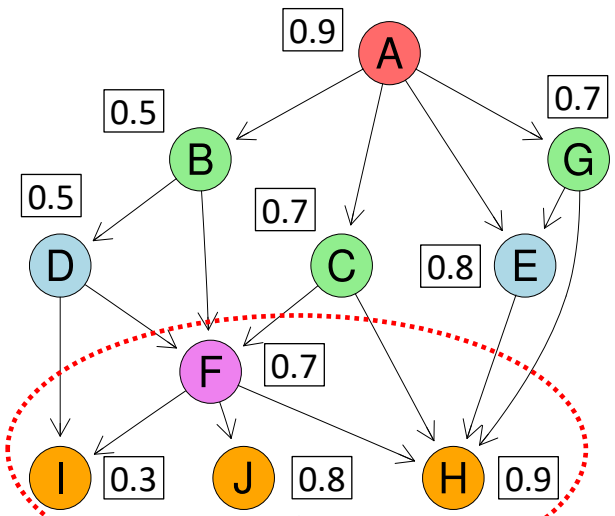
## FLAT SCORES



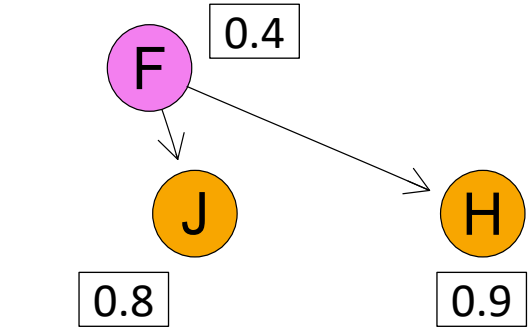

### Legend

- Node at Level 0
- Nodes at Level 1
- Nodes at Level 2
- Nodes at Level 3
- Nodes at Level 4

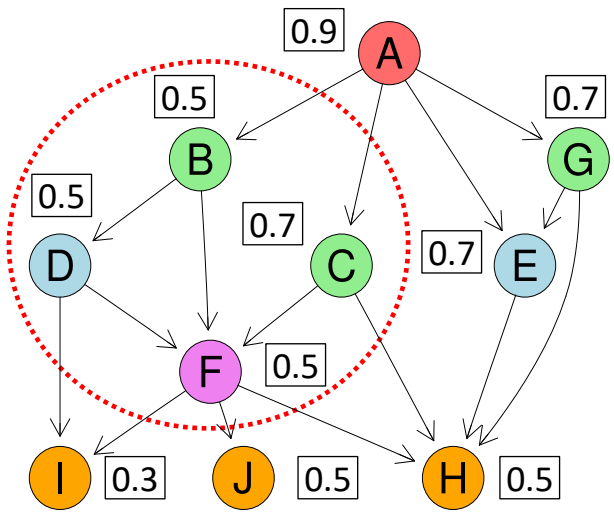
## BOTTOM UP STEP

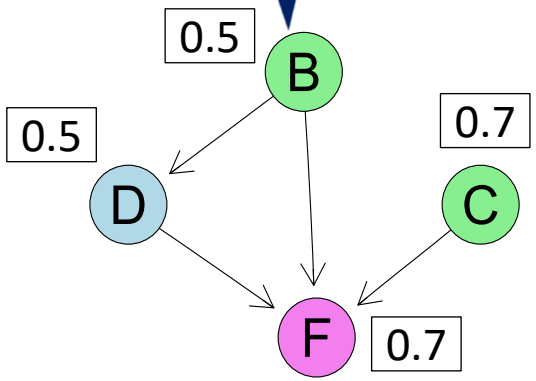



0.7 = (0.4 + 0.8 + 0.9)/3




## TOP DOWN STEP




0.5 = min (0.7, 0.5, 0.7, 0.5)

## HTD-DAG provides always biologically consistent predictions:

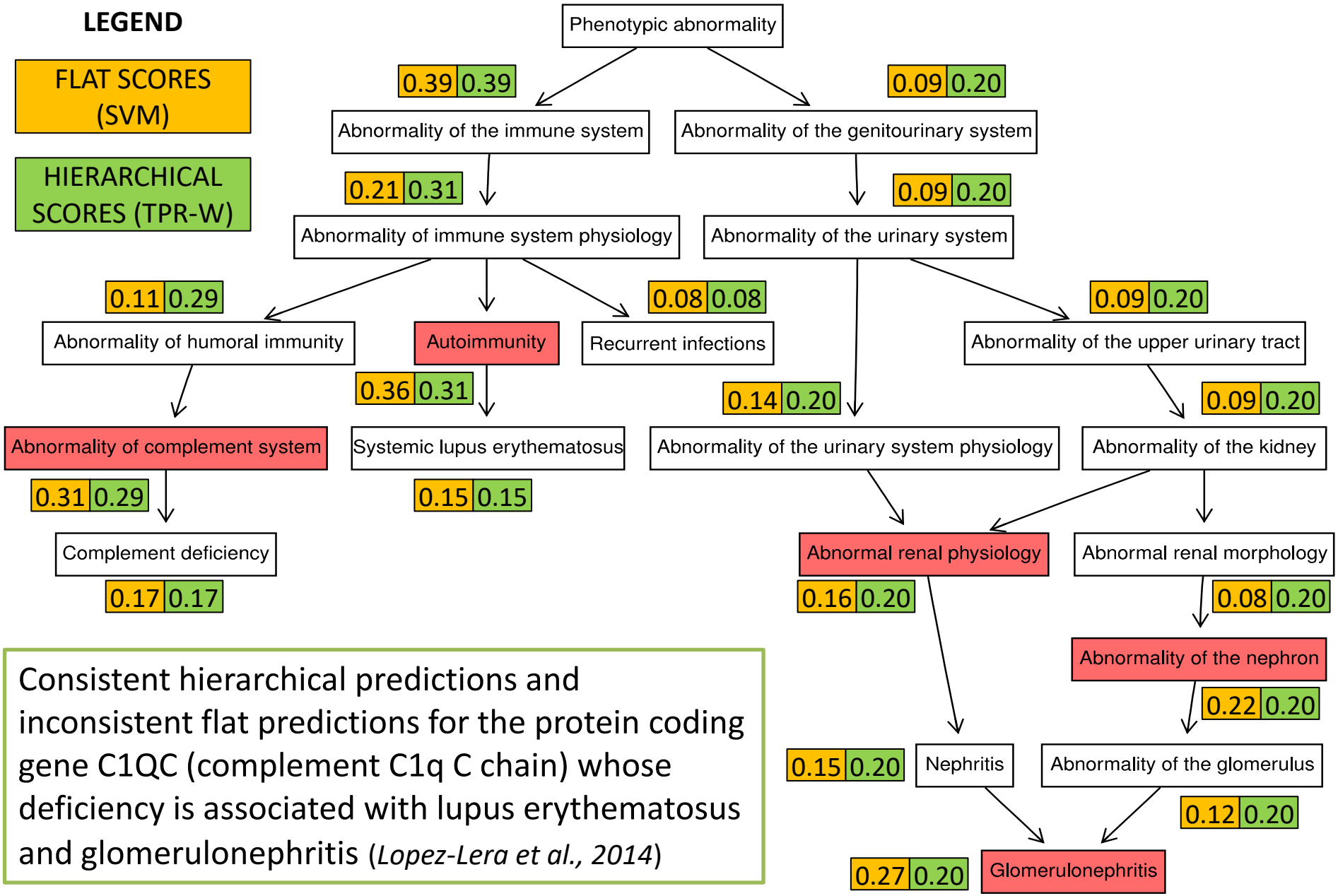
**Theorem 1.** *Given a DAG* $G =< V, E >$, *a level function* $\psi$ *that assigns to each node its maximum path length from the root and the set of* HTD-DAG *flat predictions* $\hat{\boldsymbol{y}} =< \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{|V|} >$, *the top-down hierarchical correction of the* HTD-DAG *algorithm assures that the set of ensemble predictions* $\bar{\boldsymbol{y}} =< \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{|V|} >$ *satisfies the following property:*

$$\forall i \in V,\ j \in par(i) \Rightarrow \bar{y}_j \geq \bar{y}_i$$

## TPR-DAG provides always biologically consistent predictions:

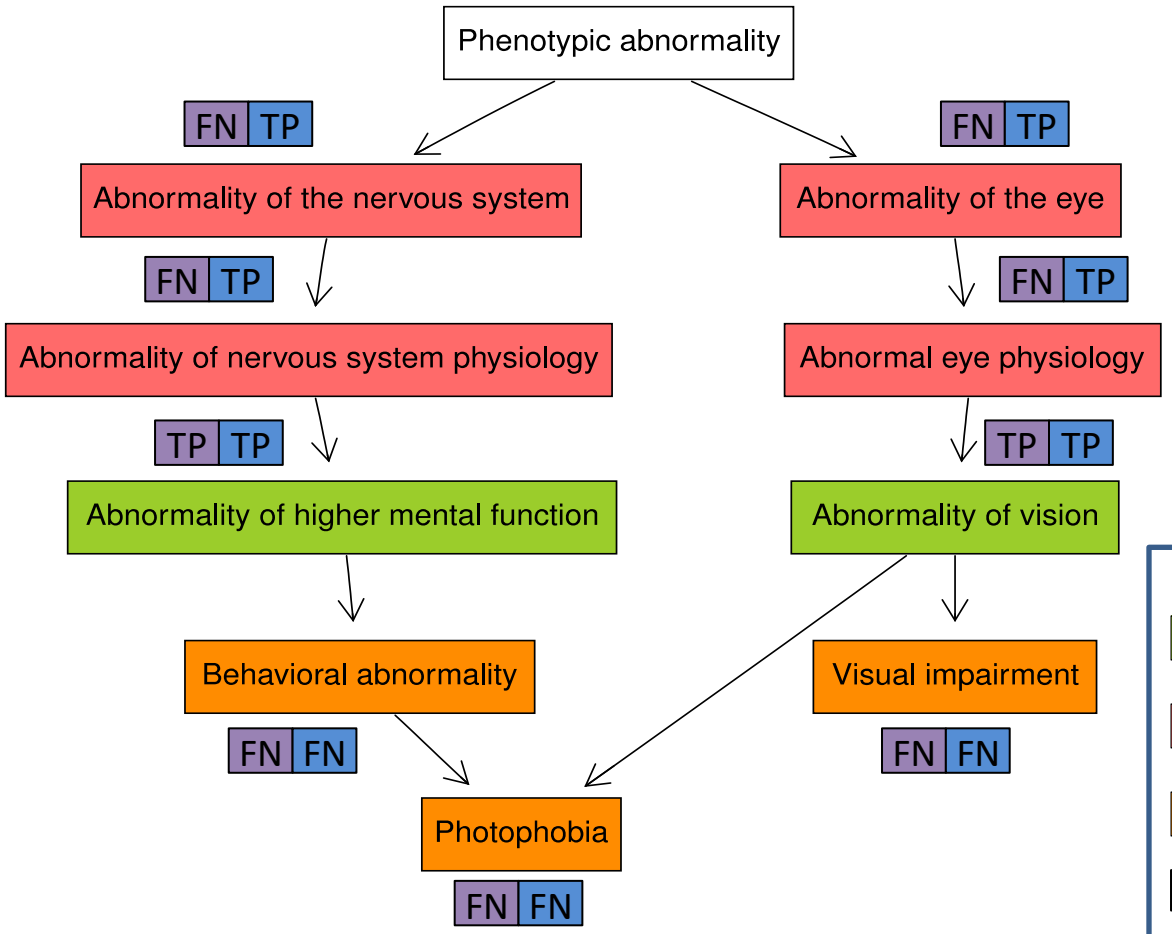
**Theorem 2.** *Given a DAG* $G =< V, E >$, *a level function* $\psi$ *that assigns to each node its maximum path length from the root, a set of predictions* $\tilde{\boldsymbol{y}} =< \tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_{|V|} >$ *generated by the bottom-up step of the* TPR *algorithm for each class associated with its corresponding node* $i \in \{1, \ldots, |V|\}$, *the top-down step of the* TPR *algorithm assures that for the set of ensemble predictions* $\bar{\boldsymbol{y}} =< \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{|V|} >$ *the following property holds:*

$$\forall i \in V,\ j \in par(i) \Rightarrow \bar{y}_j \geq \bar{y}_i$$




Consistent hierarchical predictions and inconsistent flat predictions for the protein coding gene C1QC (complement C1q C chain) whose deficiency is associated with lupus erythematosus and glomerulonephritis (*Lopez-Lera et al., 2014*)

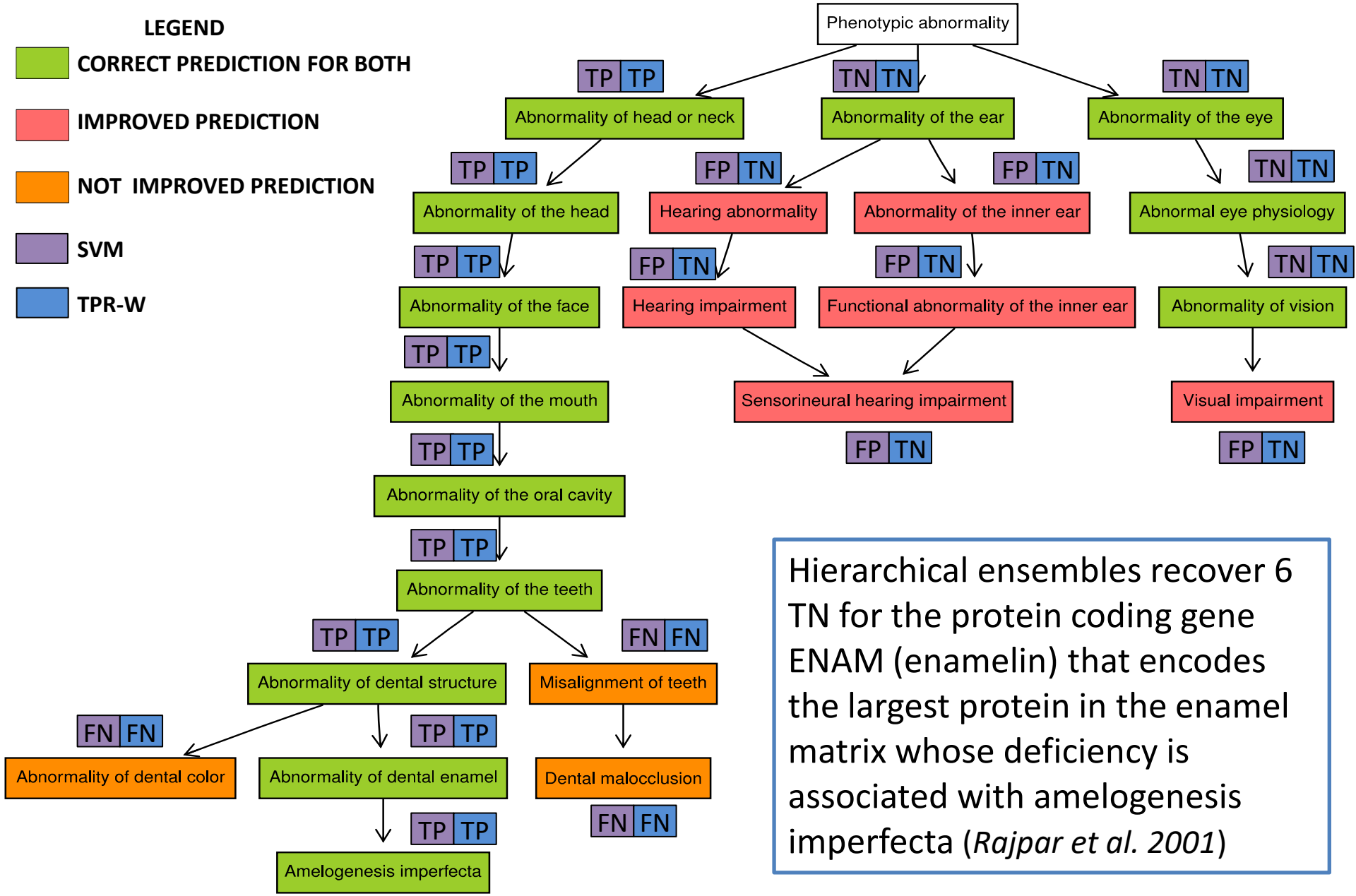
Hierarchical ensemble methods can improve flat predictions by reducing the number of FN and FP.




Hierarchical ensembles recover 4 TP for the protein coding gene RGS9 (regulator of G-protein signalling 9) whose mutations cause bradyopsia (*Michaelides et al. 2010*)

**LEGEND**

- **CORRECT PREDICTION FOR BOTH**
- **IMPROVED PREDICTION**
- **NOT IMPROVED PREDICTION**
- **SVM**
- **TPR-W**

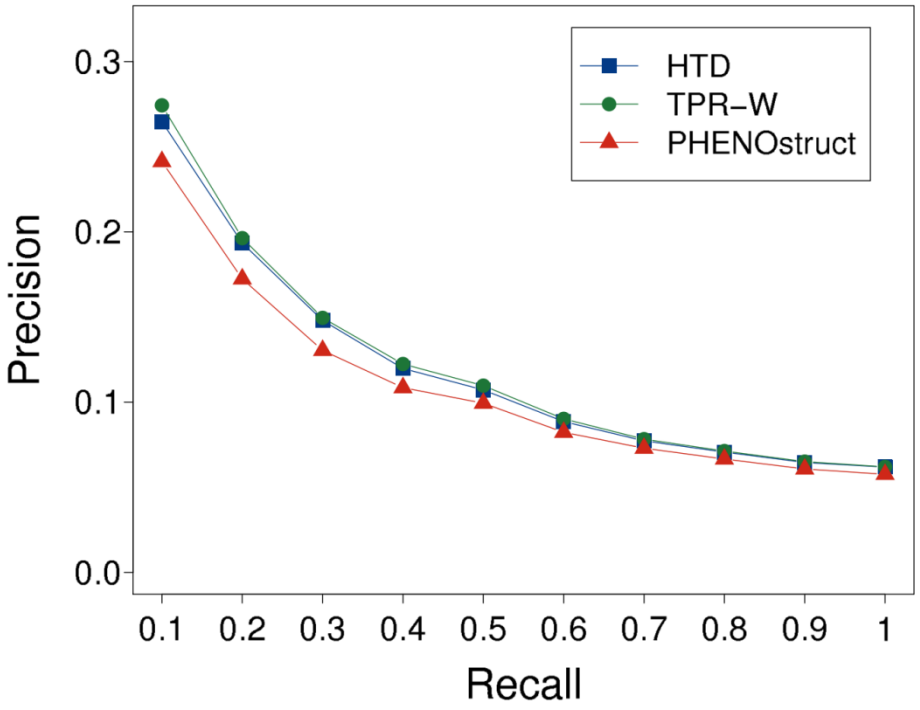



Hierarchical ensembles recover 6 TN for the protein coding gene ENAM (enamelin) that encodes the largest protein in the enamel matrix whose deficiency is associated with amelogenesis imperfecta (*Rajpar et al. 2001*)

How is the behaviour of our ensemble methods in a state-of-the-art scenario?

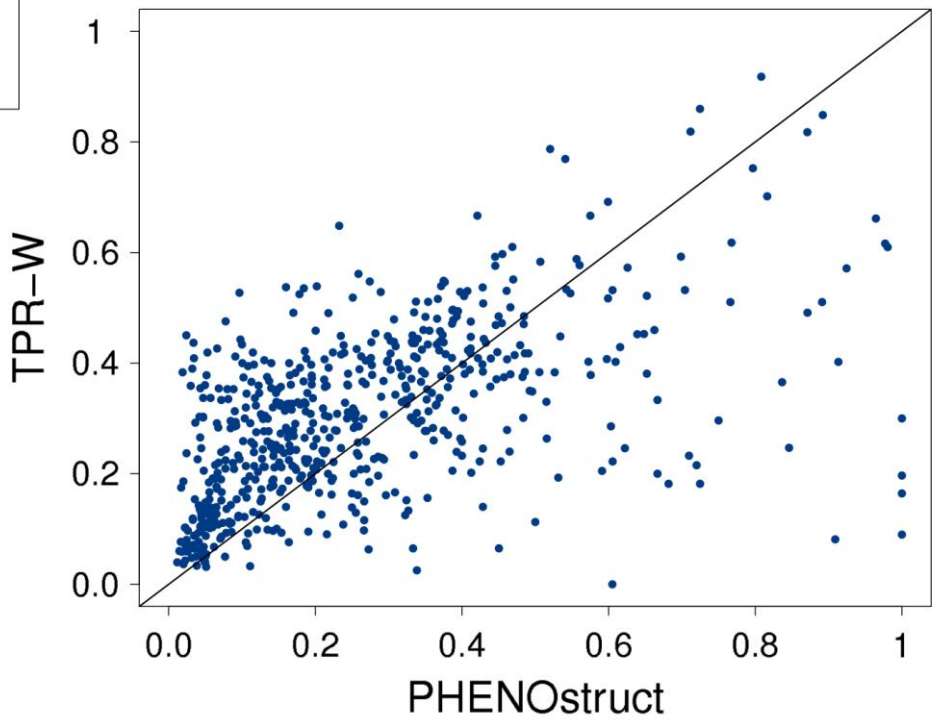
1. Comparison with PHENOstruct: a state-of-the-art joint-kernel structured output approach (*Kahanda et. al 2015*)

2. Assess the capacity to predict novel HPO annotations for human genes: we used the annotation of an old HPO release (January 2014) to predict the newly annotated genes of a recent HPO release (April 2016).
   In other words we applied a classical *hold-out* procedure between two different HPO releases.

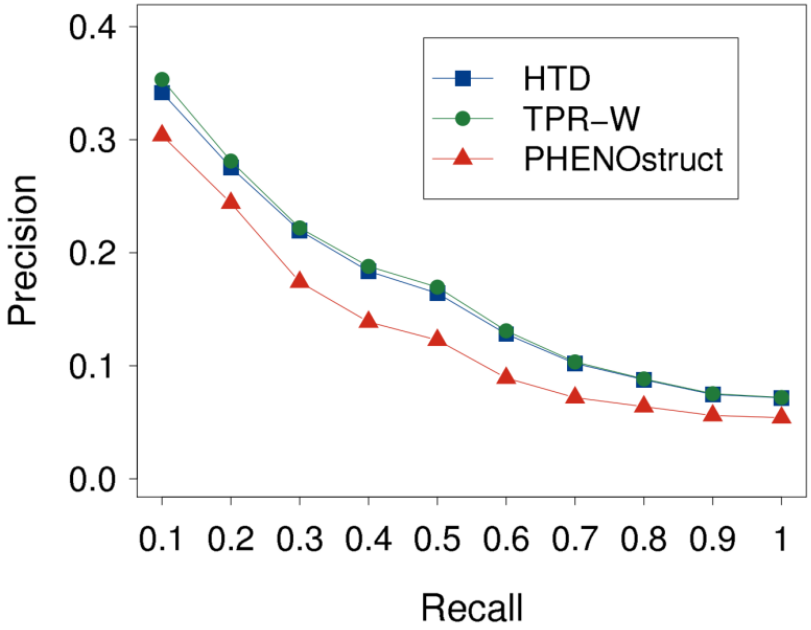


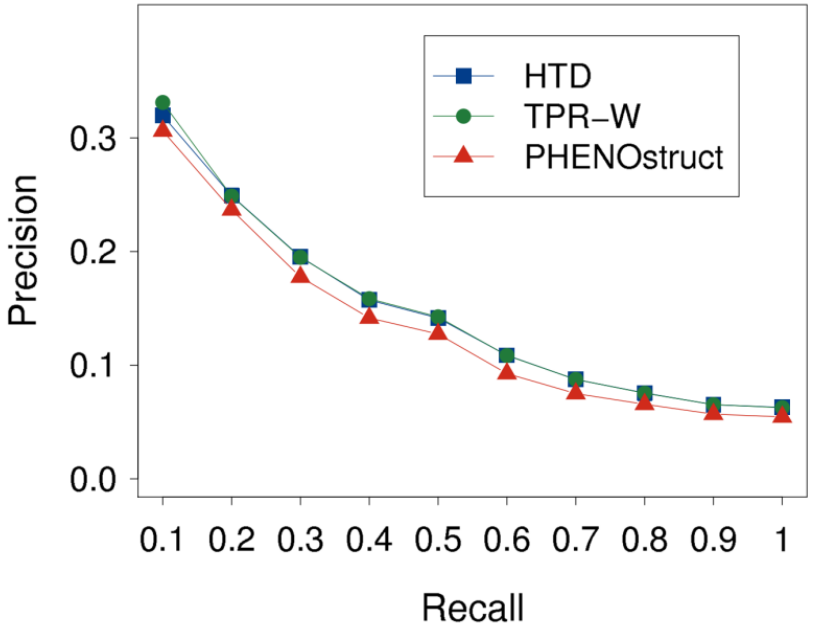

Precision-Recall curves across 2444 HPO terms: HEMs significantly improve PHENOstruct in according to Wilcoxon Sum Rank test

Scatter plot of hierarchical $F_{max}$ values: TPR-W "wins" with 431 gene and "loses" with 177 genes (Tot. genes test set: 608)



Precision-Recall curves of the newly annotated genes considering only the best predicted terms: AUROC > 0.7

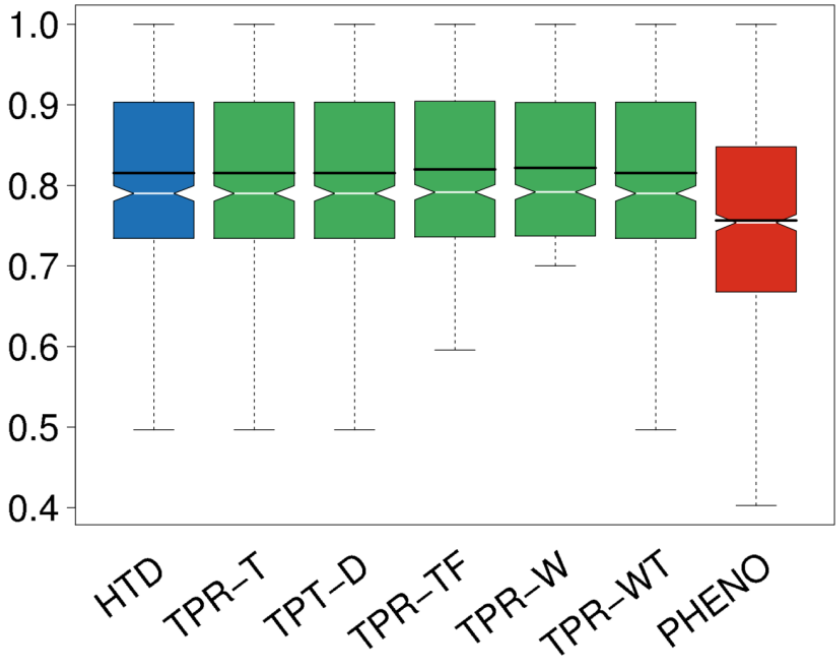


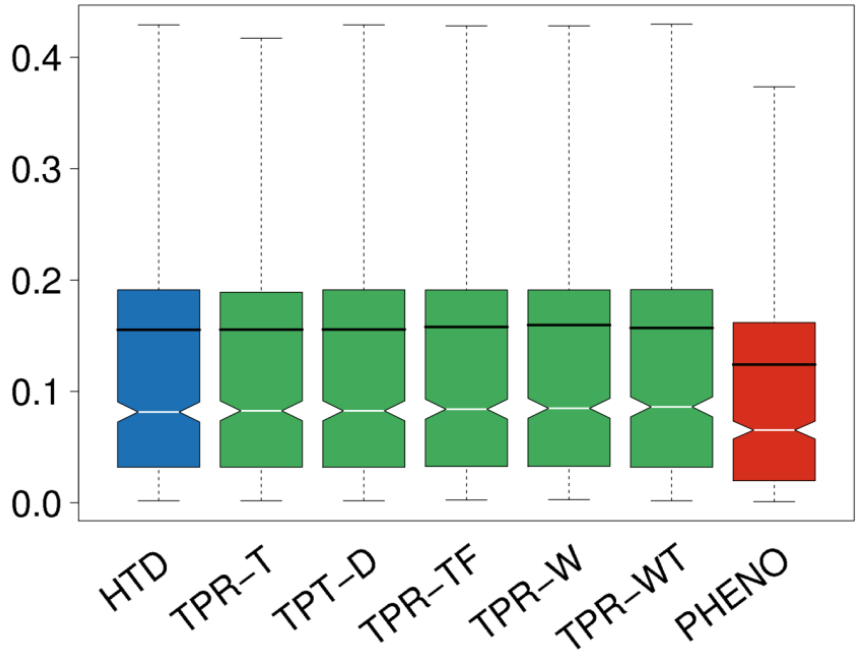

Results considering only the HPO terms predicted with AUROC>0.7 by TPR-W (778 terms)

Results considering only the HPO terms predicted with AUROC>0.7 by PHENOstruct (852 terms)

Distribution of AUROC and AUPRC values across HPO terms between our ensemble methods and PHENOstruct




AUROC distribution




AUPRC distribution

List of **novel gene-abnormal phenotype associations** predicted by our HEM and confirmed in the HPO release of March 2017 and in the most recent literature

| Gene Symbol | HPO Term | AUROC | Depth | Distance from Leaves | Evidence |
|---|---|---|---|---|---|
| XRCC2 | Clubbing of Toes | 1.000 | 9 | 0 | HPO March 2017 Release |
| LIPE | Insulin-Resistant Diabetes Mellitus | 0.9934 | 6 | 0 | HPO March 2017 Release |
| IGF2 | Neoplasm of the Adrenal Gland | 0.9781 | 5 | 0 | HPO March 2017 Release |
| ECHS1 | Abnormality of Fatty-Acid Metabolism | 0.9753 | 4 | 0 | Chika et al. 2015 |
| CFB | Systemic Lupus Erythematosus | 0.9967 | 5 | 0 | Grossman et al. 2016 |
| TGFBR3 | Emphysema | 0.9785 | 5 | 0 | Hersh et al. 2009 |
| BARD1 | Nephroblastoma aka Wilms Tumor | 0.9615 | 8 | 0 | Fu et al. 2017 |
| MSH3 | Breast Carcinoma | 0.9723 | 5 | 0 | Miao et al. 2015 |
| CAD | Abnormality of Pyrimidine Metabolism | 0.9951 | 4 | 0 | Bobby et al. 2015 |
| COX10 | Abnormal Mitochondria in Muscle Tissue | 0.9967 | 6 | 0 | Pitceathly et al. 2013 |

Since for about half of Mendelian disease no causative genes are know (*Chong et.al 2015*), our ensemble methods may contribute to the discovery such genes and to unveil the full spectrum of phenotypes associated with them

- Genome-wide and ontology-wide experimental results show that our hierarchical ensemble method are competitive with PHENOStruct;
- Computational time significantly lower than state-of-the-art joint-kernel structured output methods (training + test of hold-out):
  - HTD: 12 minutes;
  - TPR-W: 3 hours (tuning of *w* parameter by 5-fold-cross-validation);
  - PHENOstruct: 18 hours;

Using an Intel Xeon CPU E5-2630 2.6Hz with 128 GB of RAM

HTD-DAG and TPR-DAG:

a) scale linearly with number of the class;
b) can be applied to big data;
c) provide biologically consistent predictions;
d) improve upon flat predictions;

1. In principle the HEM can significantly improve any flat prediction independently of which flat approach we used: I plan to do more experiments with a large set of base learners (GO ontology and in multi-species environment)

2. Top-Down step: instead of HTD strategy → Isotonic Regression

Input:
- $G =< V, E >$
- $V = \{1, 2, \ldots, |V|\}$
- $\hat{\boldsymbol{y}} =< \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{|V|} >, \quad \hat{y}_i \in [0, 1]$

begin algorithm

01: A. Isotonic correction:

02: $$\bar{\boldsymbol{y}} = \begin{cases} \min_{\bar{\boldsymbol{y}}} \sum_{i \in V} (\hat{y}_i - \bar{y}_i)^2 \\ \forall i, \quad j \in par(i) \Rightarrow \bar{y}_j \geq \bar{y}_i \end{cases}$$

end algorithm

Output:
- $\bar{\boldsymbol{y}} =< \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{|V|} >$

→

- select the closest solution to flat predictions in the sense of the squared error
- constraints are maintained by construction

3. TPR-DAG is a family of algorithm: I plan to design novel TPR-D variants
   i. *linear decay* of weights with respect to the levels;
   ii. *linear increment* of the weights (from bottom to top) in order to put more weight on predictions made on the most specific terms;

4. Prediction of coding (mRNA) and non coding (lncRNA, miRNA) RNA interactions inferred from an integrated multi-ontology environment (GO, HPO, DO)
   i. Hierarchical prediction of GO, DO, HPO terms: prediction of PCGs and NCGs performed across multi-ontology;
   ii. Discover novel interaction between ncRNA and PCGs using known and predicted annotation → will help us to better understand complex genetic diseases and cancer

(Joint Collaboration with Dept. of Computer Engineering of the University of Granada)

# Acknowledgments

Giorgio Valentini

https://anacletolab.di.unimi.it

Peter N. Robinson

The Jackson Laboratory
*Leading the search for tomorrow's cures*

https://www.jax.org/

Max Schubach

CHARITÉ
UNIVERSITÄTSMEDIZIN BERLIN

https://www.charite.de/

I. M. Notaro, M. Schubach, P.N. Robinson and G. Valentini, *Prediction of Human Phenotype Ontology terms by means of Hierarchical Ensemble methods*, submitted to BMC Bioinformatics.

II. M. Notaro, M. Frasca, M. Mesiti, M. Schubach, P.N. Robinson and G. Valentini, *Ensembling Descendant Term Classifiers to Improve Gene – Abnormal Phenotype Prediction***s**, submitted to *Computational Intelligence methods for Bioinformatics and Biostatistics (CIBB 2017)*

III. P.N. Robinson, M.Frasca, S. Köhler, M. Notaro, M. Re, G. Valentini, *A Hierarchical Ensemble Method for DAG-Structured Taxonomies*, *Lecture Notes in Computer Science*, vol. 9132, pp. 15-26. Berlin: Springer, 2015.

IV. G. Valentini, S. Köhler, M. Re, M. Notaro, P.N. Robinson, *Prediction of Human Gene-Phenotype Associations by Exploiting the Hierarchical Structure of the Human Phenotype Ontology*, *Lecture Notes in Computer Science*, vol. 9043, pp. 66–77. Cham: Springer, 2015.

V. G. Valentini, Hierarchical Ensemble Methods for Protein Function Prediction, *ISRN Bioinformatics, vol. 2014 (2014).*

VI. Kahanda, I., Funk, C., Verspoor, K., Ben-Hur, A.: *PHENOstruct: Prediction of human phenotype ontology terms using heterogeneous data sources.* F1000Research 4, 259 (2015).

VII. Liang Cheng, Jie Sun, Wanying Xu, Lixiang Dong, Yang Hu, Meng Zhou, OAGH: an integrated resource for annotating human genes with multi-level ontologies, Scientific Reports, 6:34820, 2016

# THANKS for YOUR ATTENTION!

# ANY QUESTIONS?