Active and Guided Learning of Enzyme Function

Luna De Ferrari
Stuart Aitken
John Mitchell

luna.deferrari@st-andrews.ac.uk

6th International Workshop
on Machine Learning in Systems Biology
@ECCB Basel, Switzerland

8 Sep 2012

University of St Andrews

BBSRC bioscience for the future

THE UNIVERSITY OF EDINBURGH

Predict enzyme function

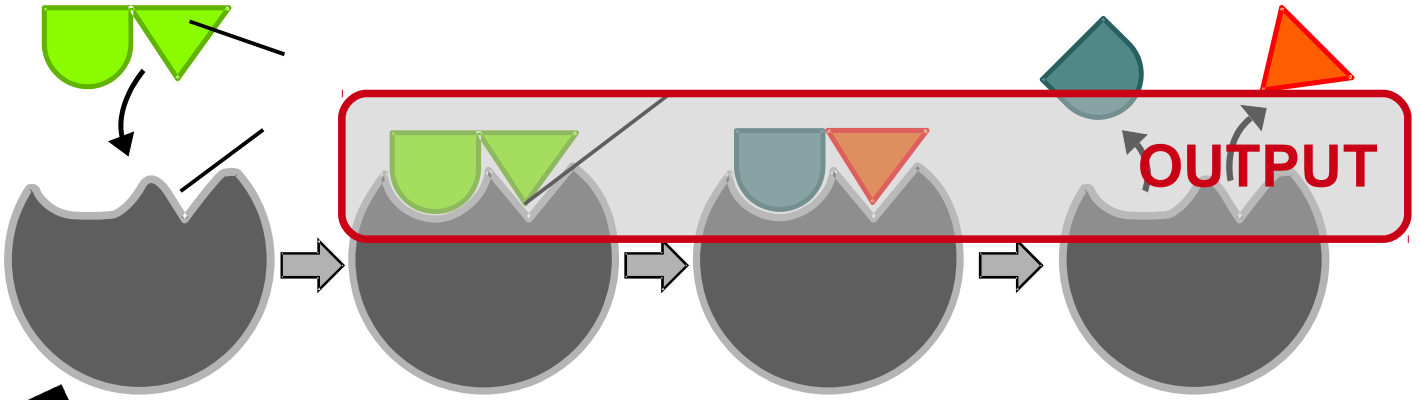


Fig: The induced fit model of enzyme activity. Provided by Tim Vickers. From: http://en.wikipedia.org/wiki/Enzyme

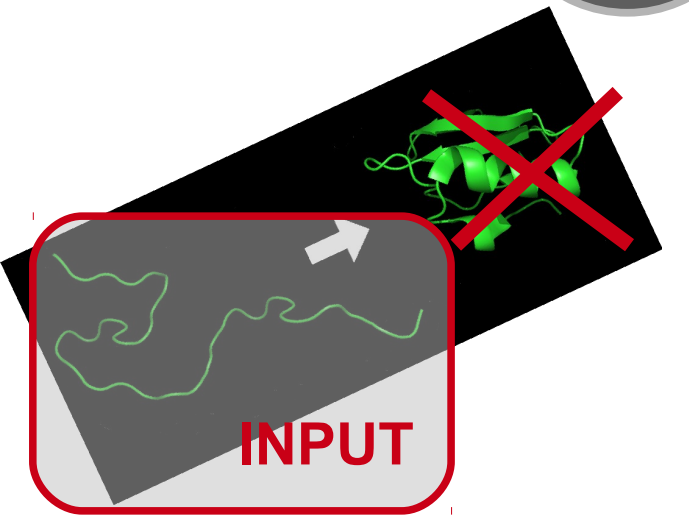


Fig: Protein folding. Provided by Dr Kjaergaard. From: http://en.wikipedia.org/wiki/Protein_folding

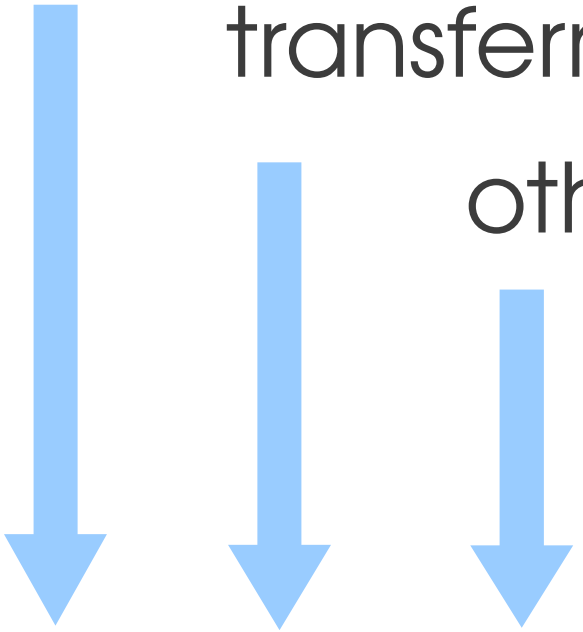
Enzyme Commission numbers

HOW

Transferases

transferring nitrogenous groups

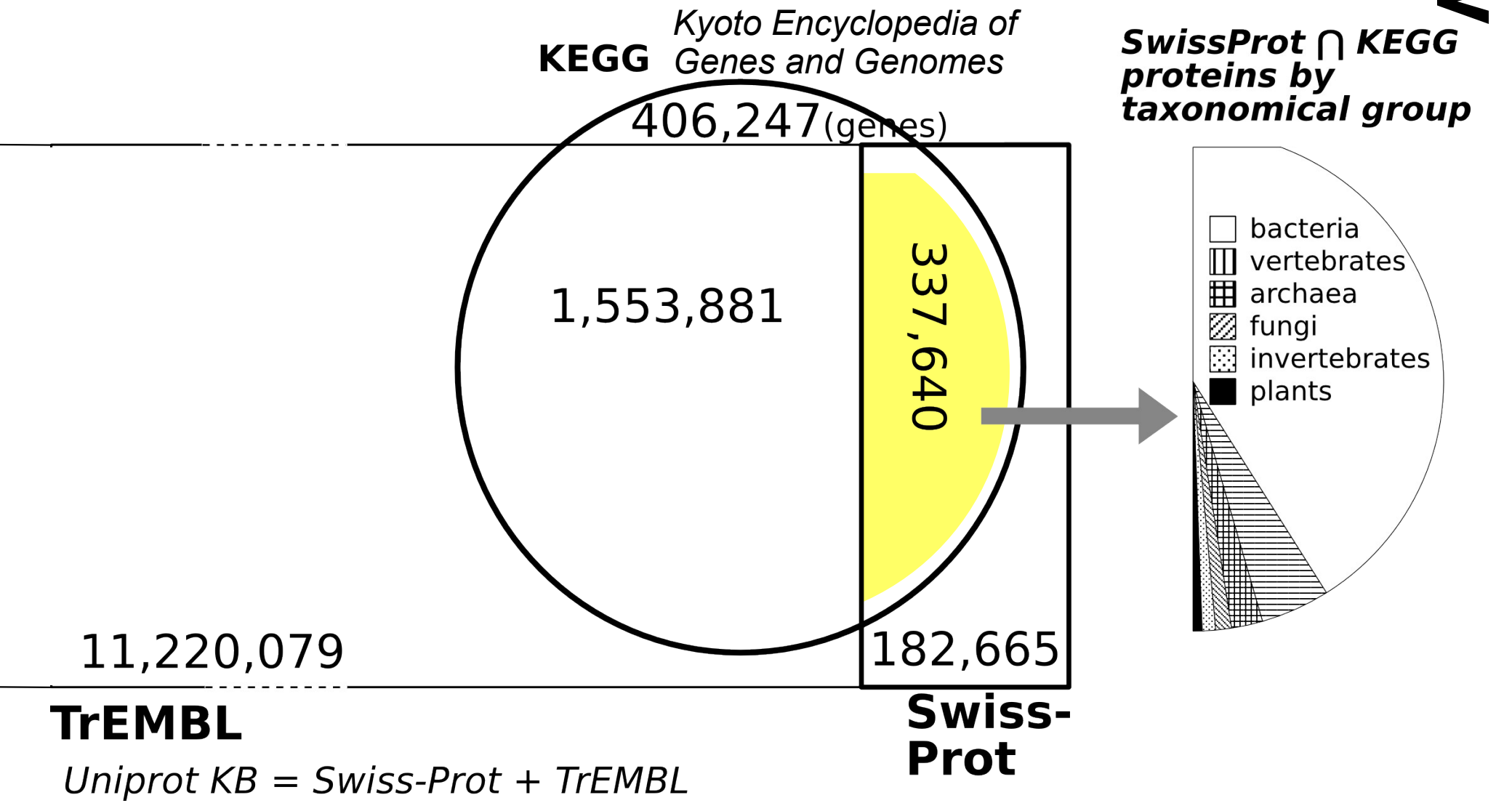
other nitrogenous groups



2.6.99.2

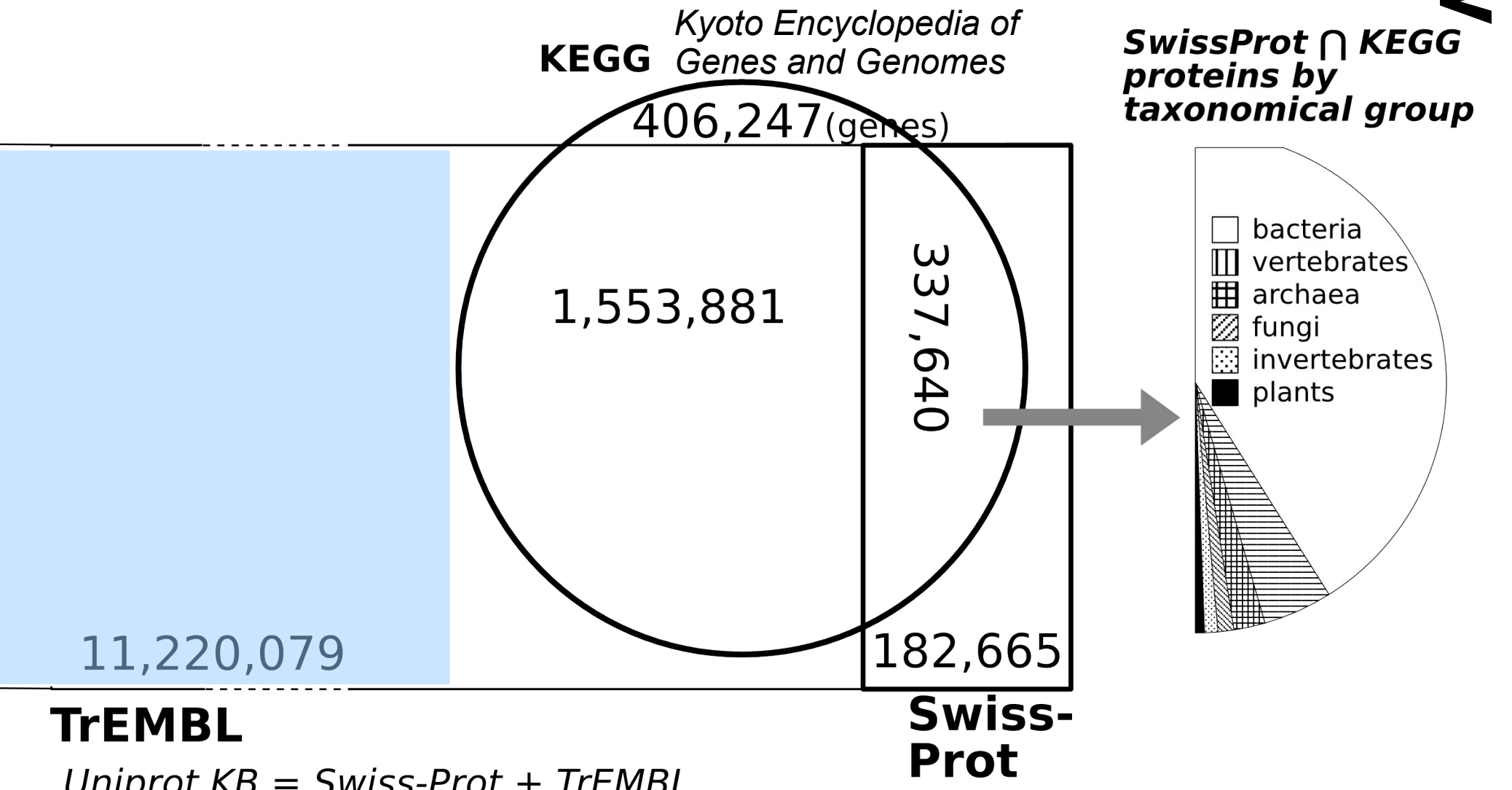
Pyridoxine 5'-phosphate synthase

Data sources

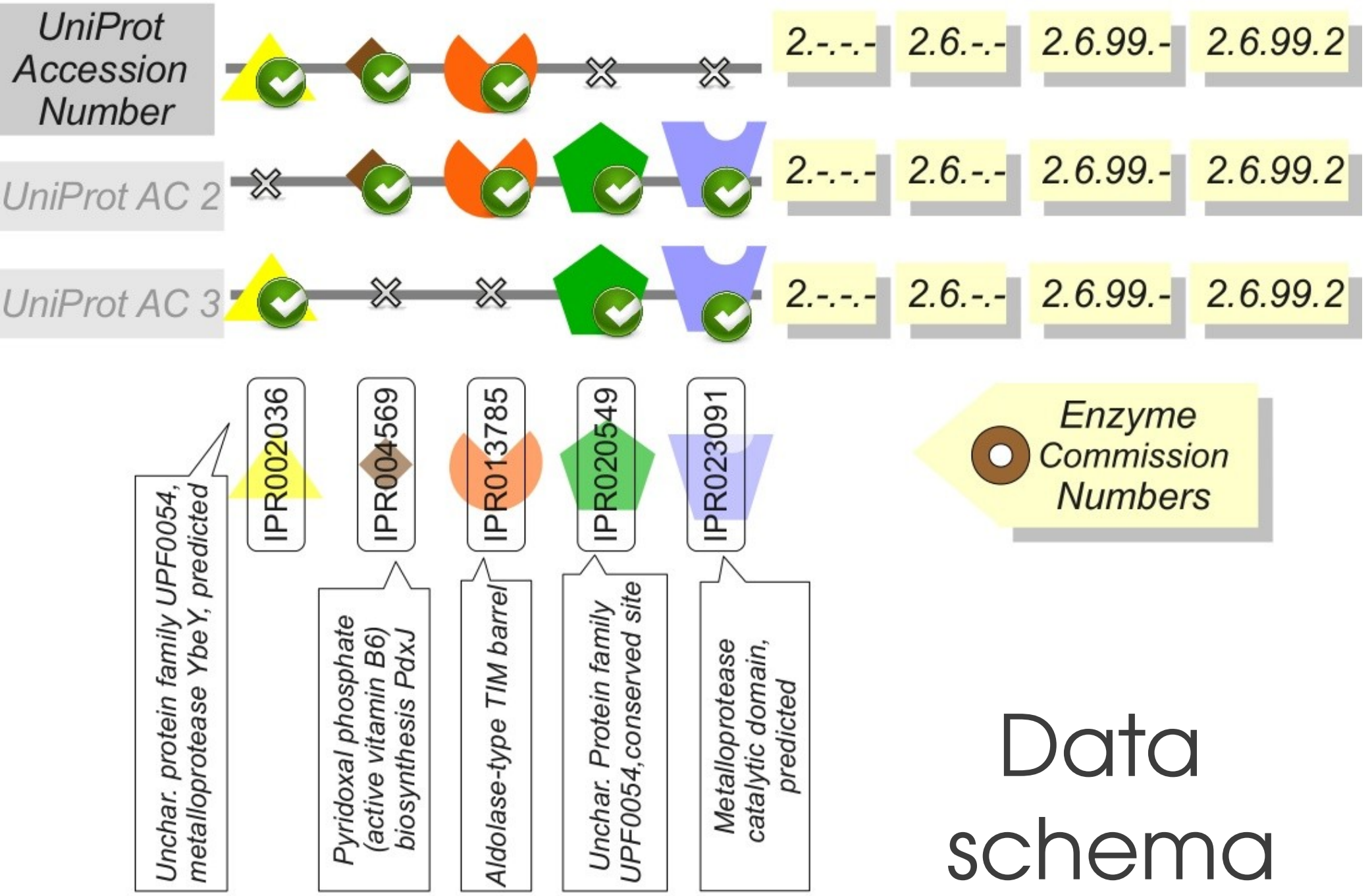


EnzML: Multi-label prediction of enzyme classes using InterPro signatures, L. De Ferrari et al. BMC Bioinformatics 2012, 13:61

Data sources



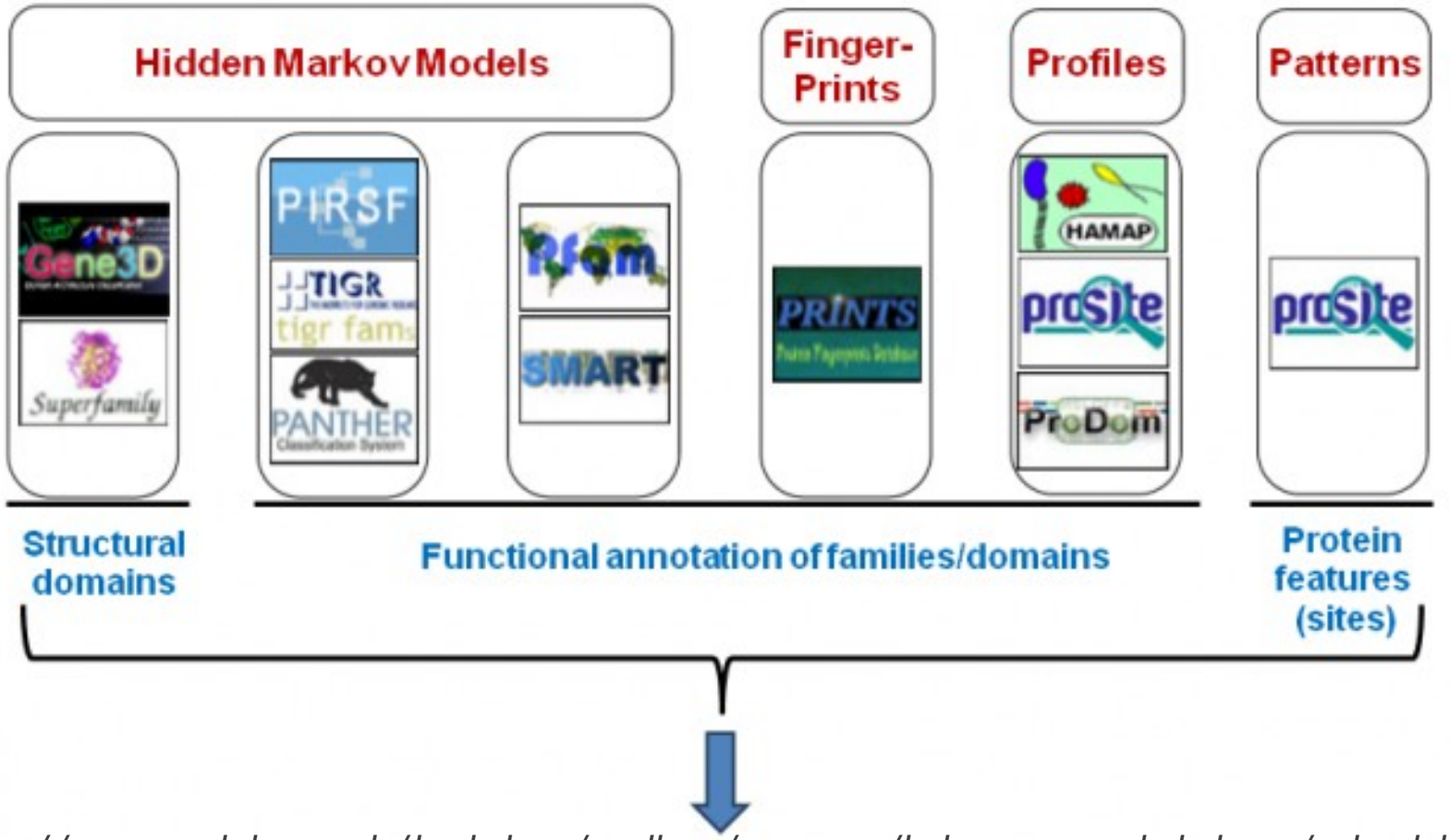
EnzML: Multi-label prediction of enzyme classes using InterPro signatures, L. De Ferrari et al. BMC Bioinformatics 2012, 13:61



Data schema

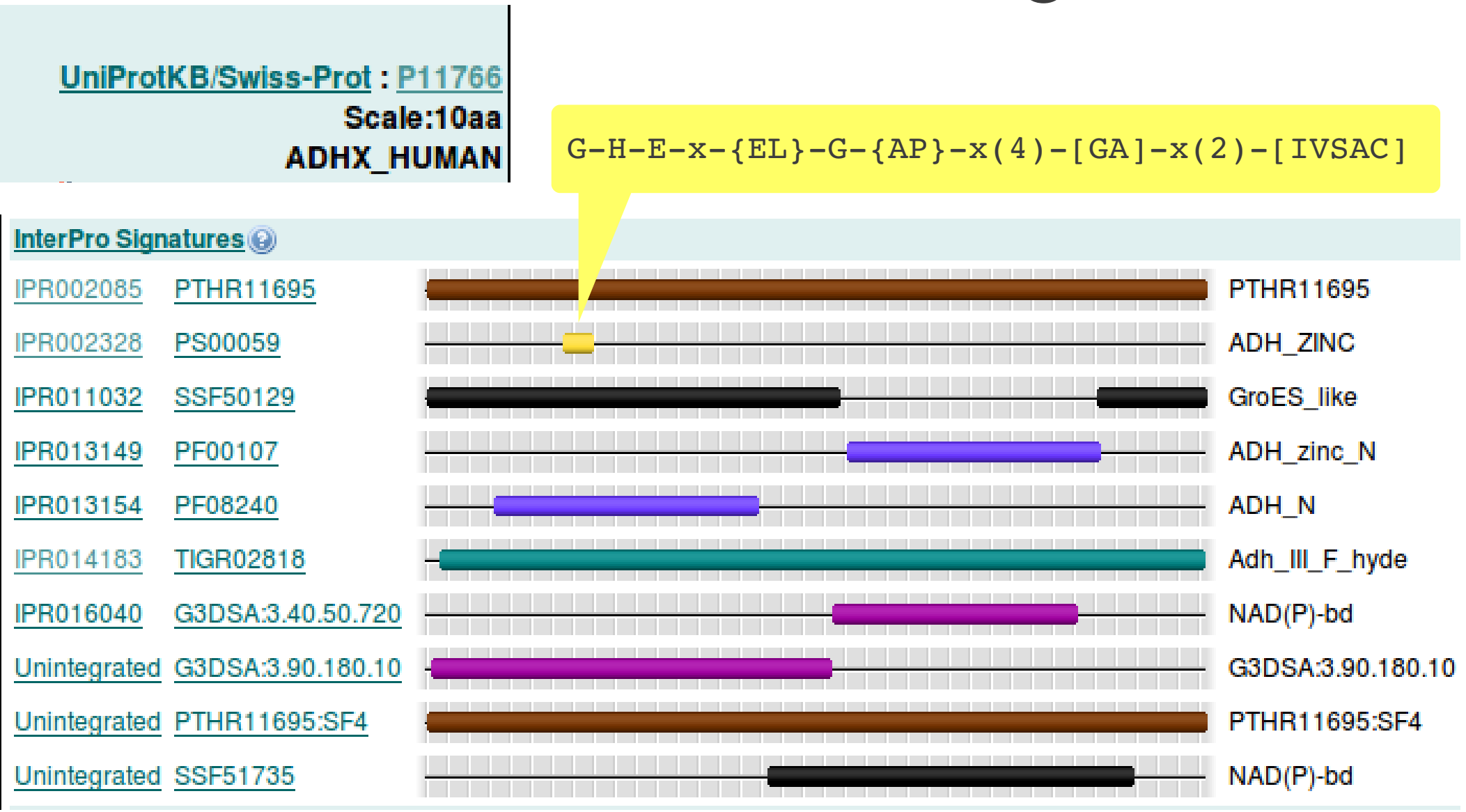
(1,108 Uniprot proteins)

InterPro sequence signatures

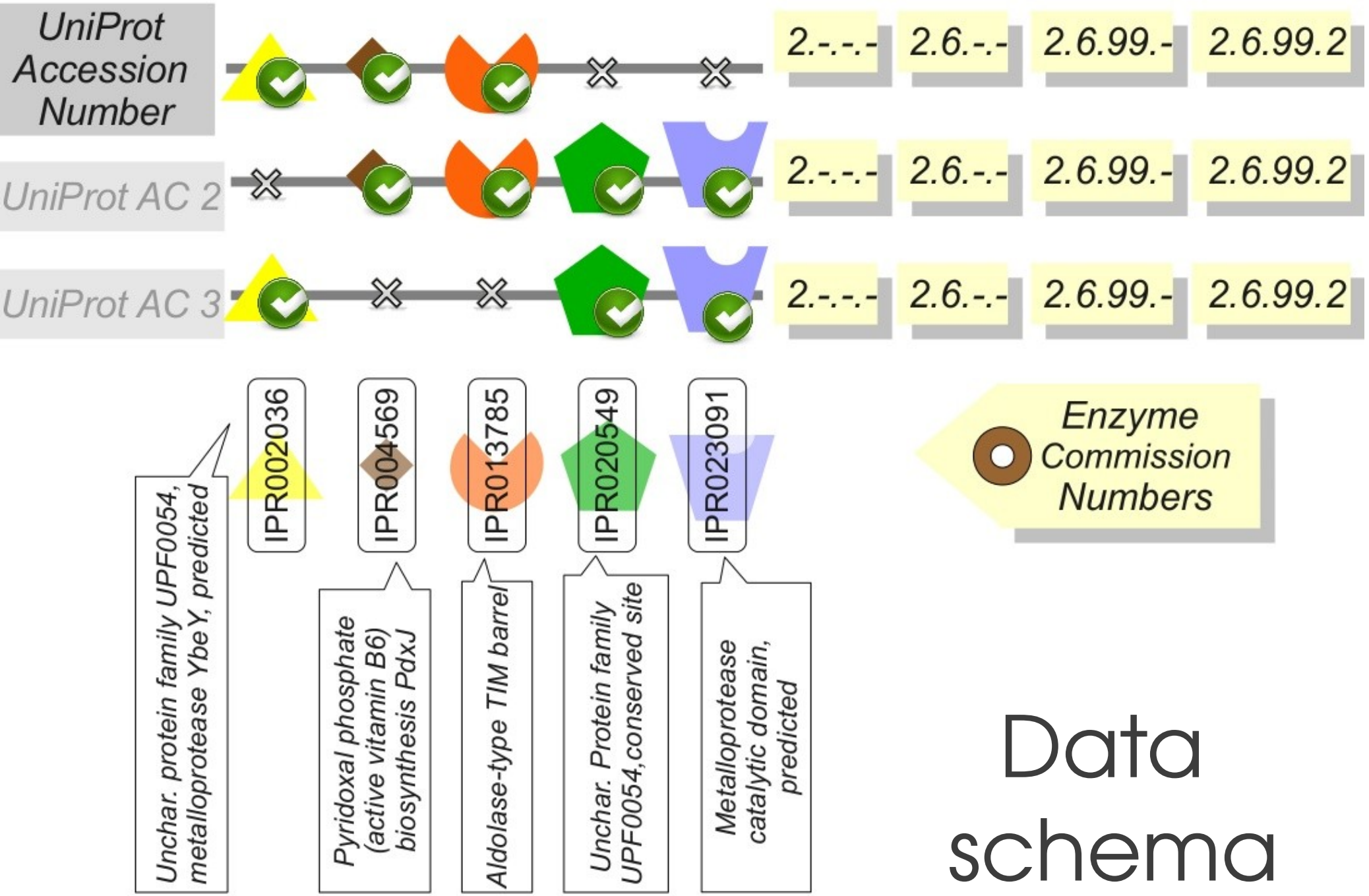


From:http://www.ebi.ac.uk/training/online/course/interpro-quick-tour/what-interpro

InterPro signatures



From: http://www.ebi.ac.uk/interpro/ISpy?ac=P11766



Data schema

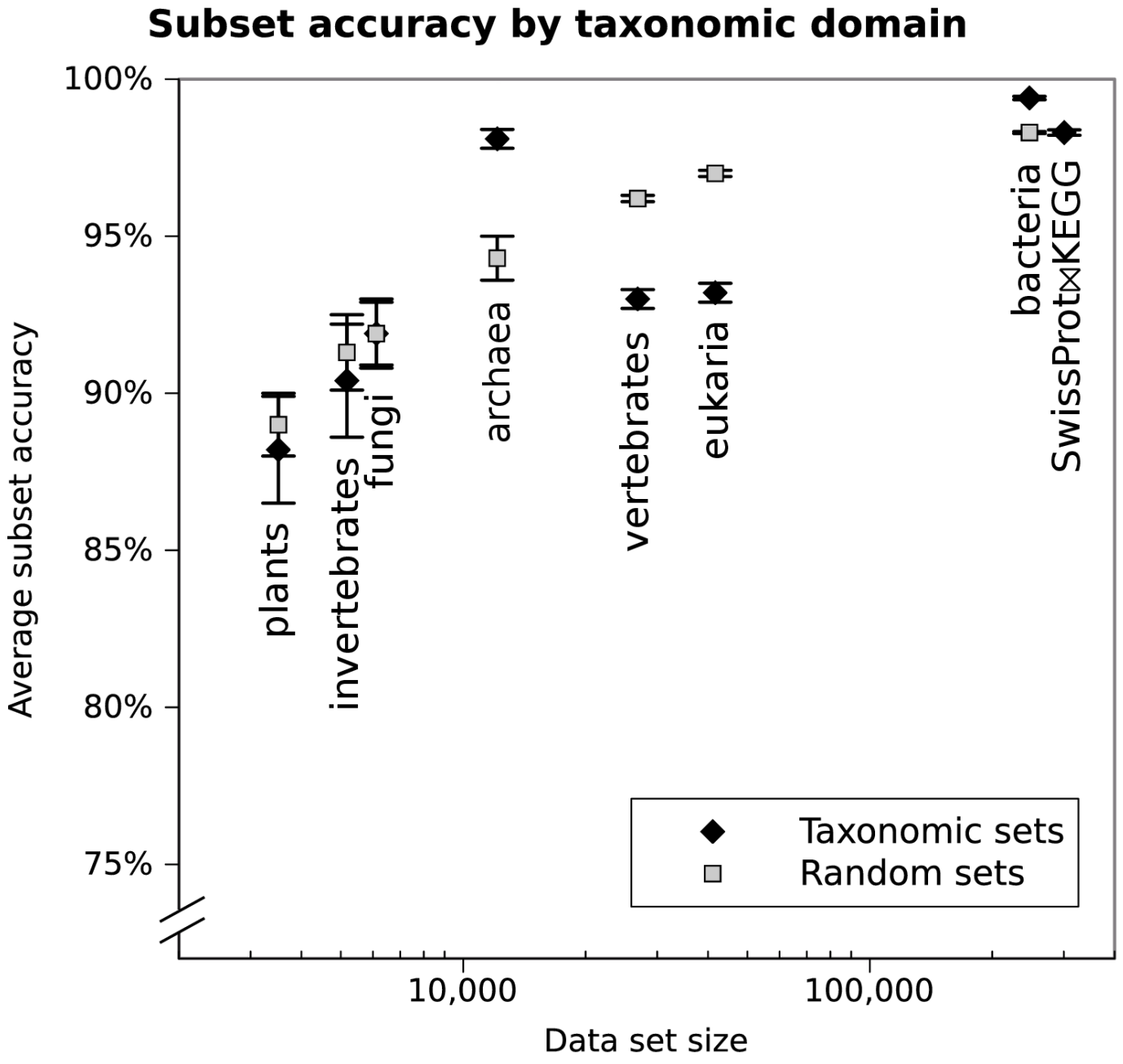
(1,108 Uniprot proteins)

Algorithm

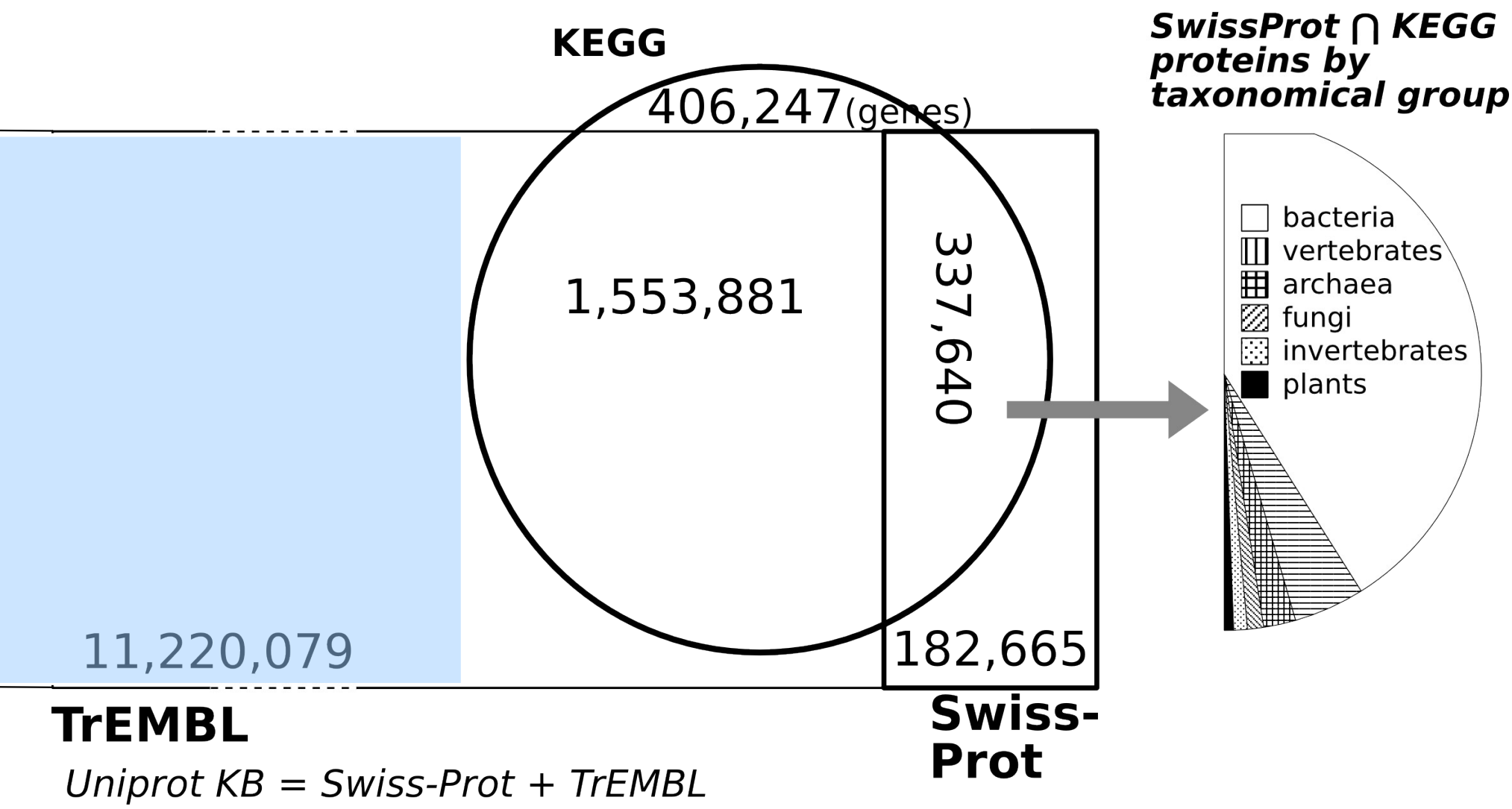
- Mulan http://mulan.sourceforge.net/
  - Multi-Label algorithms library, Weka compatible (Java)

- Fastest/best of the Mulan 1.2.0 algoritms
  - BR-kNN a Multi-Label **Nearest Neighbours** implementation
  - *(E. Spyromitros, et al. An Empirical Study of Lazy Multilabel Classification Algorithms, SETN 2008, Syros, Greece)*
  - k=1 Neighbour

SwissProt ⋈ KEGG cross-evaluation



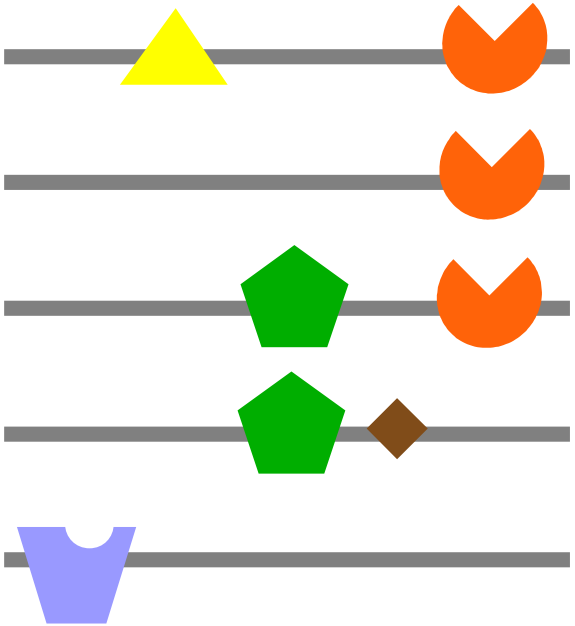
Biocuration



Uniprot KB = Swiss-Prot + TrEMBL

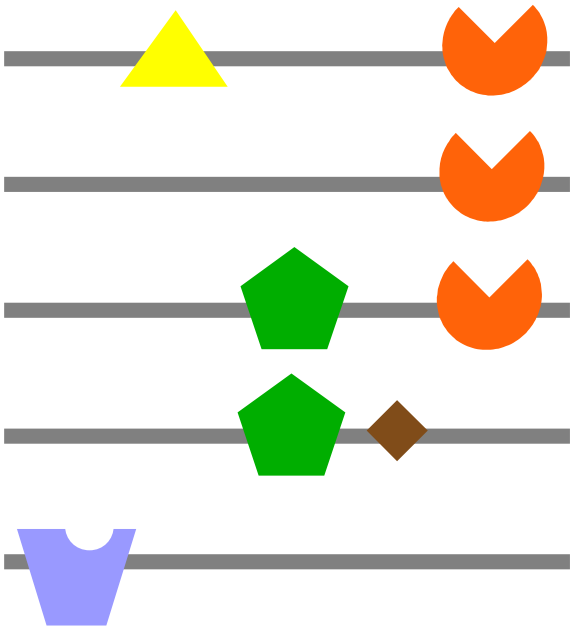
EnzML: Multi-label prediction of enzyme classes using InterPro signatures, L. De Ferrari et al. BMC Bioinformatics 2012, 13:61

Dynamics of collaborative bio-curation + machine learning



EC 1.2.3.4

Bio-curation + machine learning



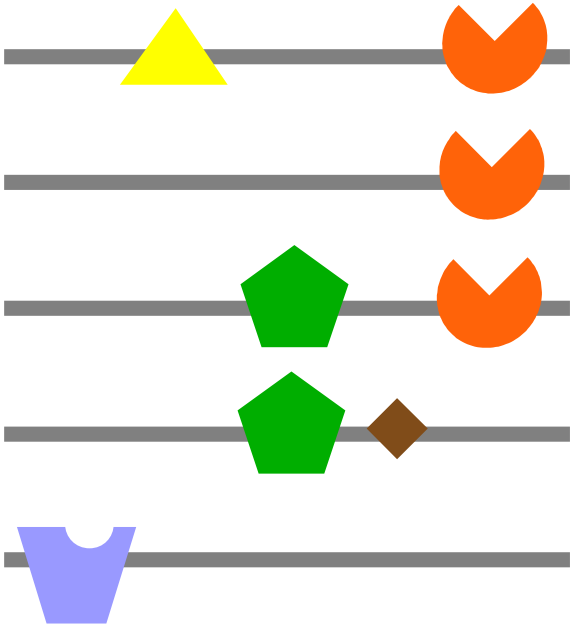
EC 1.2.3.4

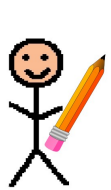
EC 1.2.3.4

EC 1.2.3.4

EC 1.2.3.4

Bio-curation + machine learning





EC 1.2.3.4

EC 1.2.3.4

EC 1.2.3.4

EC 1.2.3.4

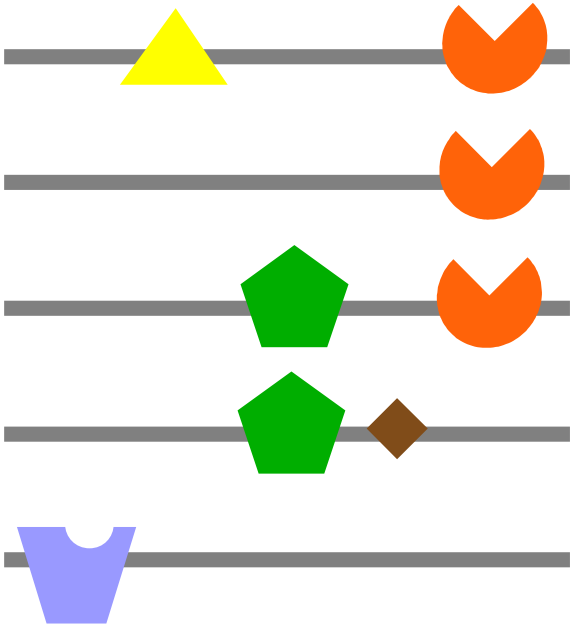
EC 1.2.3.4

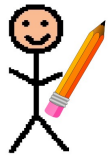
EC 1.2.3.4

EC 1.2.3.4

EC 6.7.8.9

Bio-curation + machine learning





EC 1.2.3.4

EC 1.2.3.4

EC 1.2.3.4

EC 1.2.3.4

EC 1.2.3.4

EC 1.2.3.4

EC 6.7.8.9

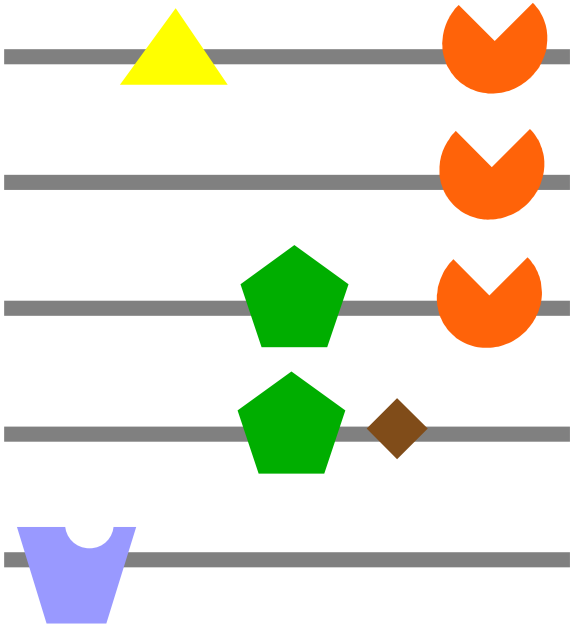
EC 1.2.3.4

EC 1.2.3.4

EC 6.7.8.9

EC 6.7.8.9

Bio-curation + machine learning



EC 1.2.3.4

EC 1.2.3.4

EC 1.2.3.4

EC 1.2.3.4

...

EC 1.2.3.4

EC 1.2.3.4

EC 6.7.8.9

EC 1.2.3.4

EC 1.2.3.4

EC 6.7.8.9

EC 6.7.8.9

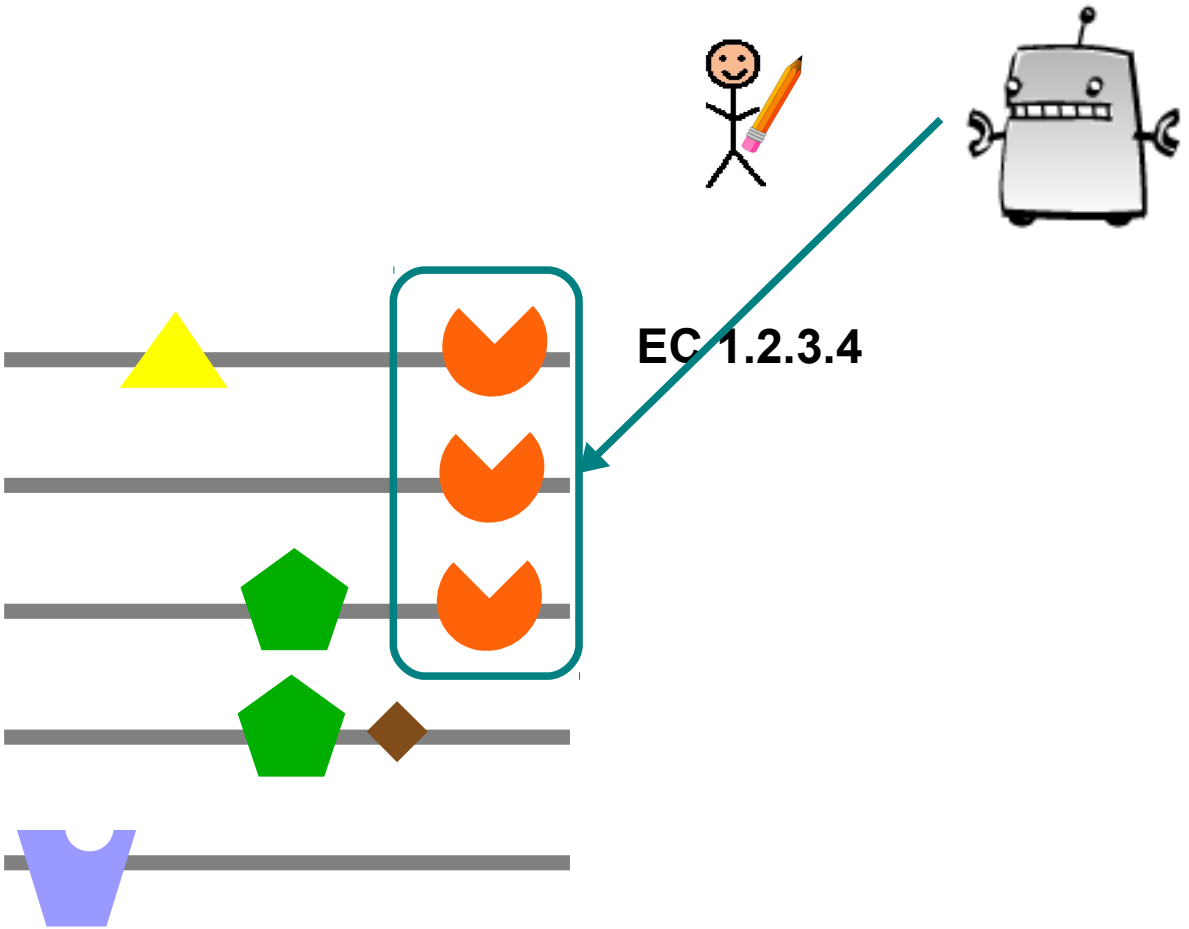
15 million

Active learning

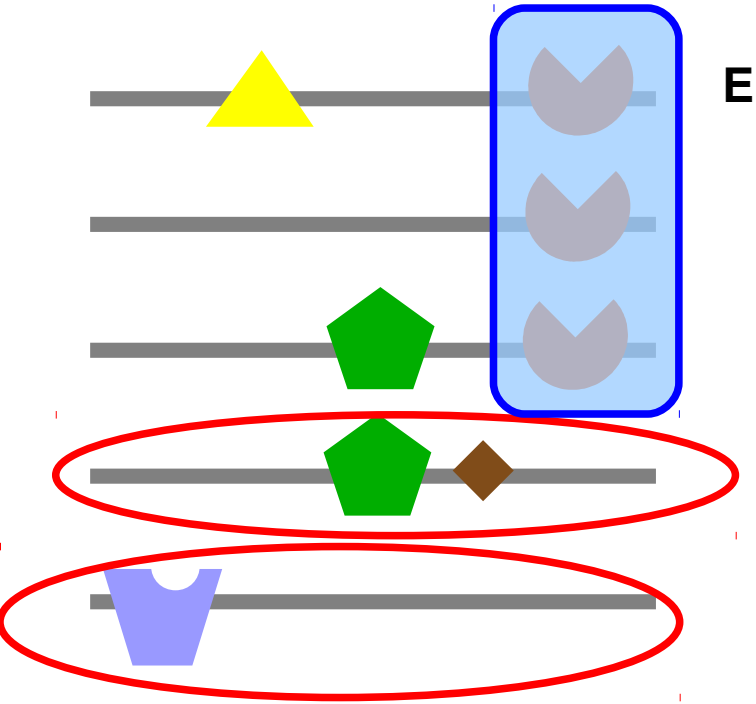
- Abundant unlabelled data
- Labels are expensive to obtain
- Allow the machine learning to **choose** the datum to be labelled next

Active Learning Literature Survey. Burr Settles. Computer Sciences Technical Report 1648, University of Wisconsin–Madison

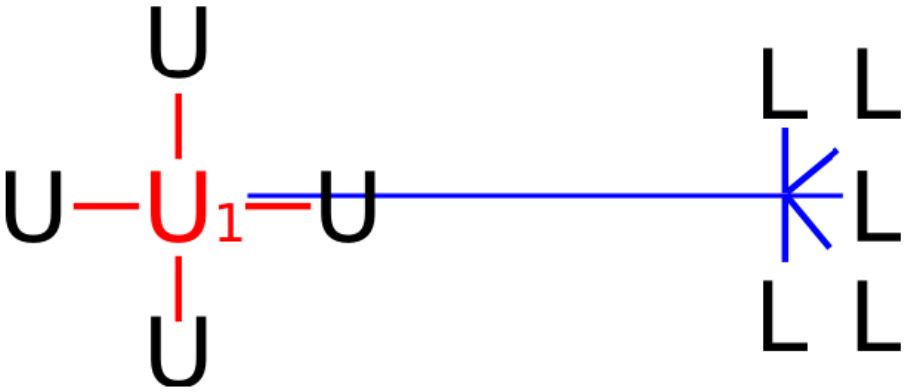
Active learning: different from labelled & similar to unlabelled



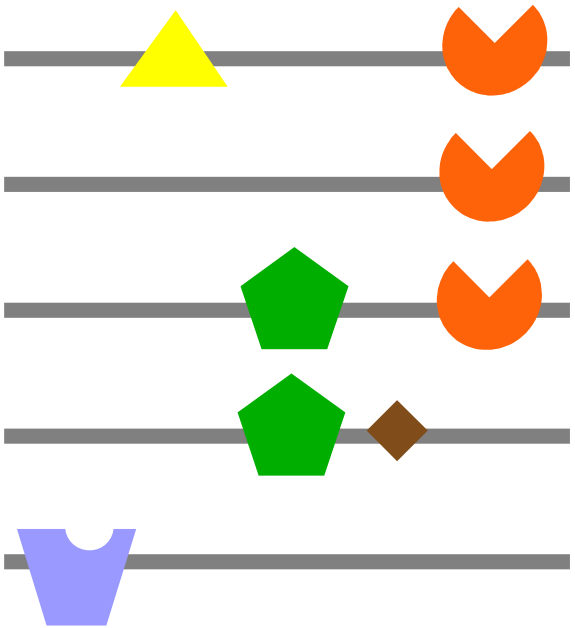
Active learning: different from labelled & similar to unlabelled



EC 1.2.3.4



Active learning: confidence



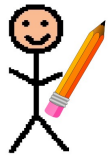
EC 1.2.3.4

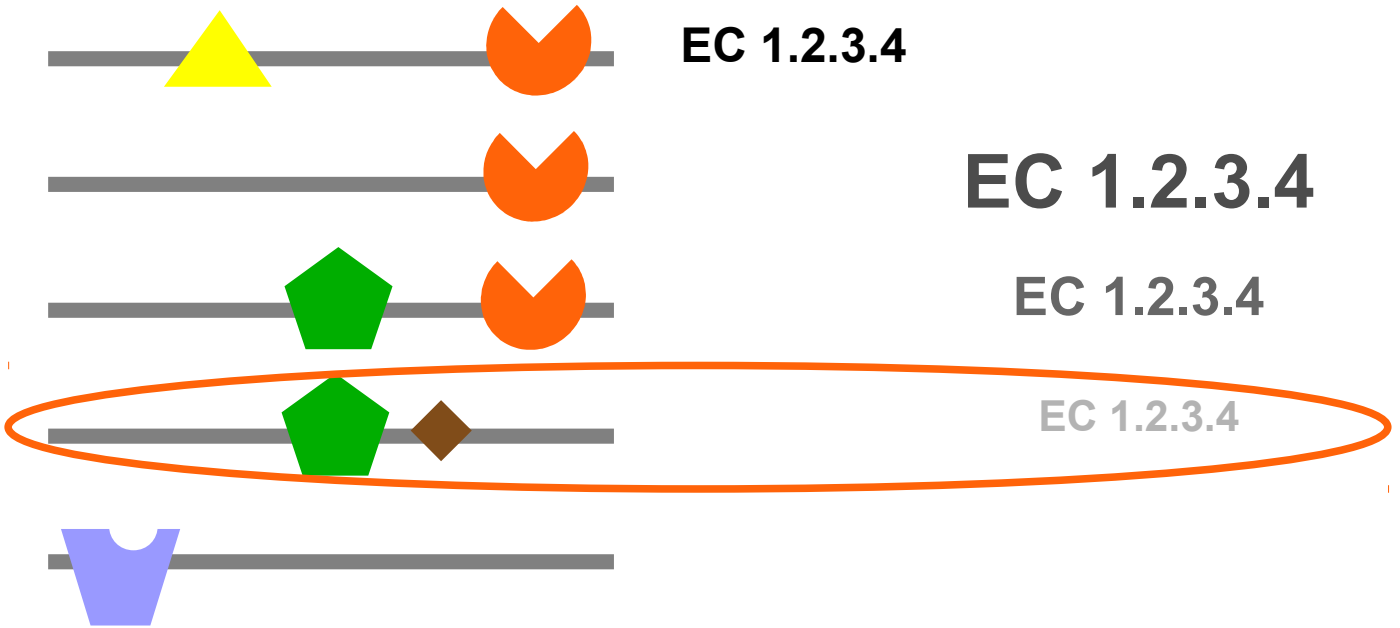
EC 1.2.3.4

EC 1.2.3.4

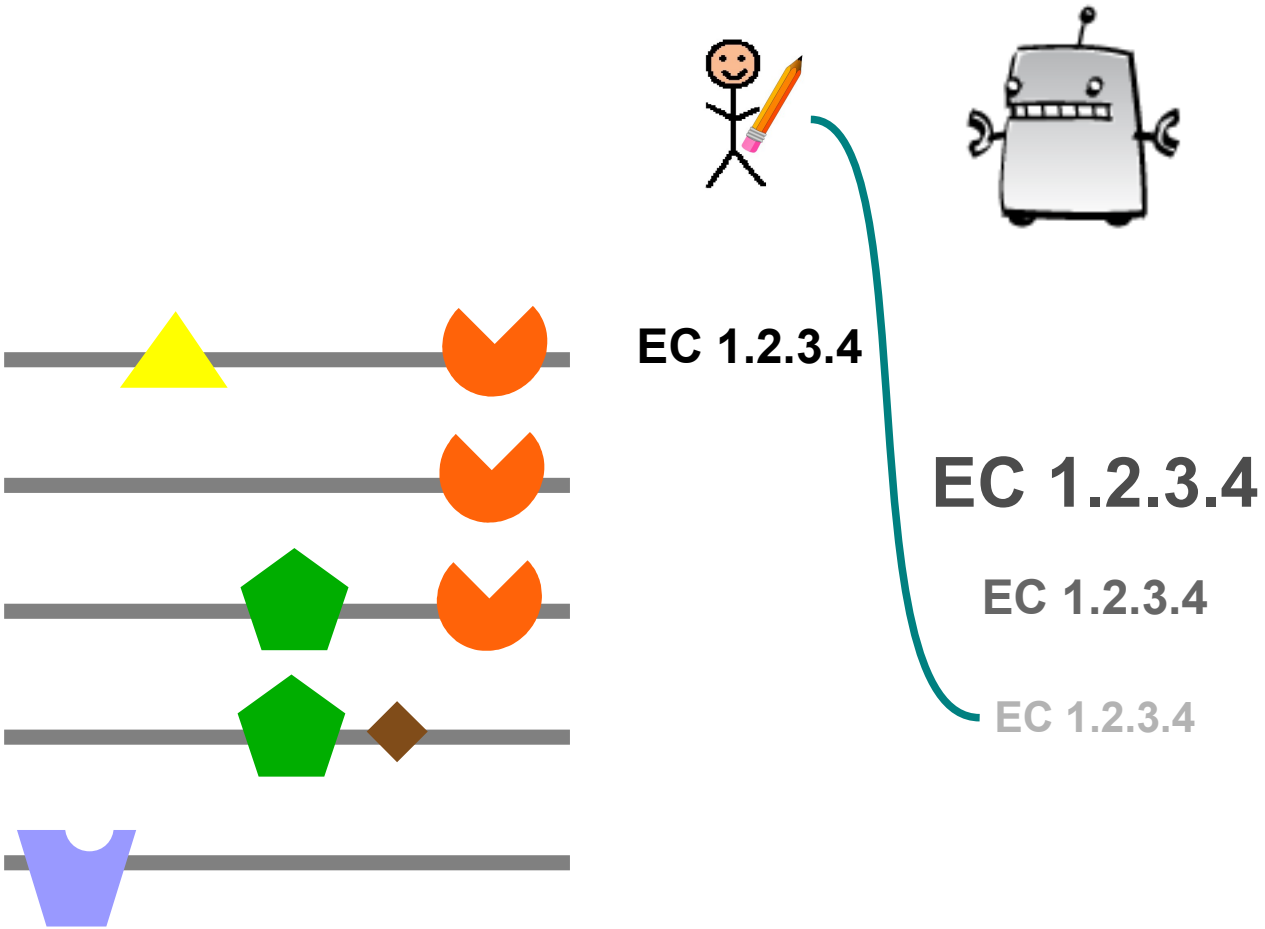
EC 1.2.3.4

Active learning: confidence

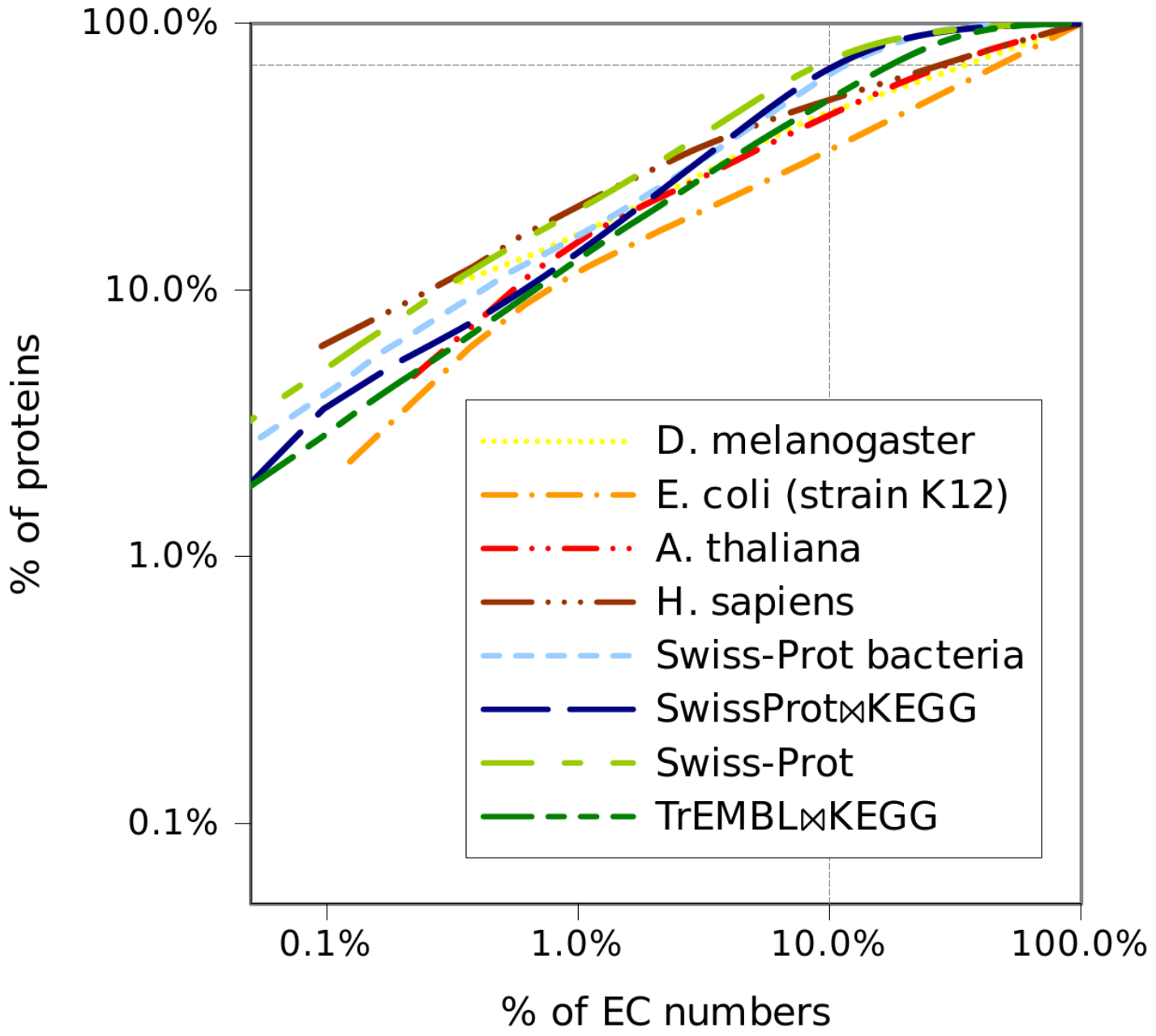




Active learning disadvantages: non-parallel & poor on rare classes

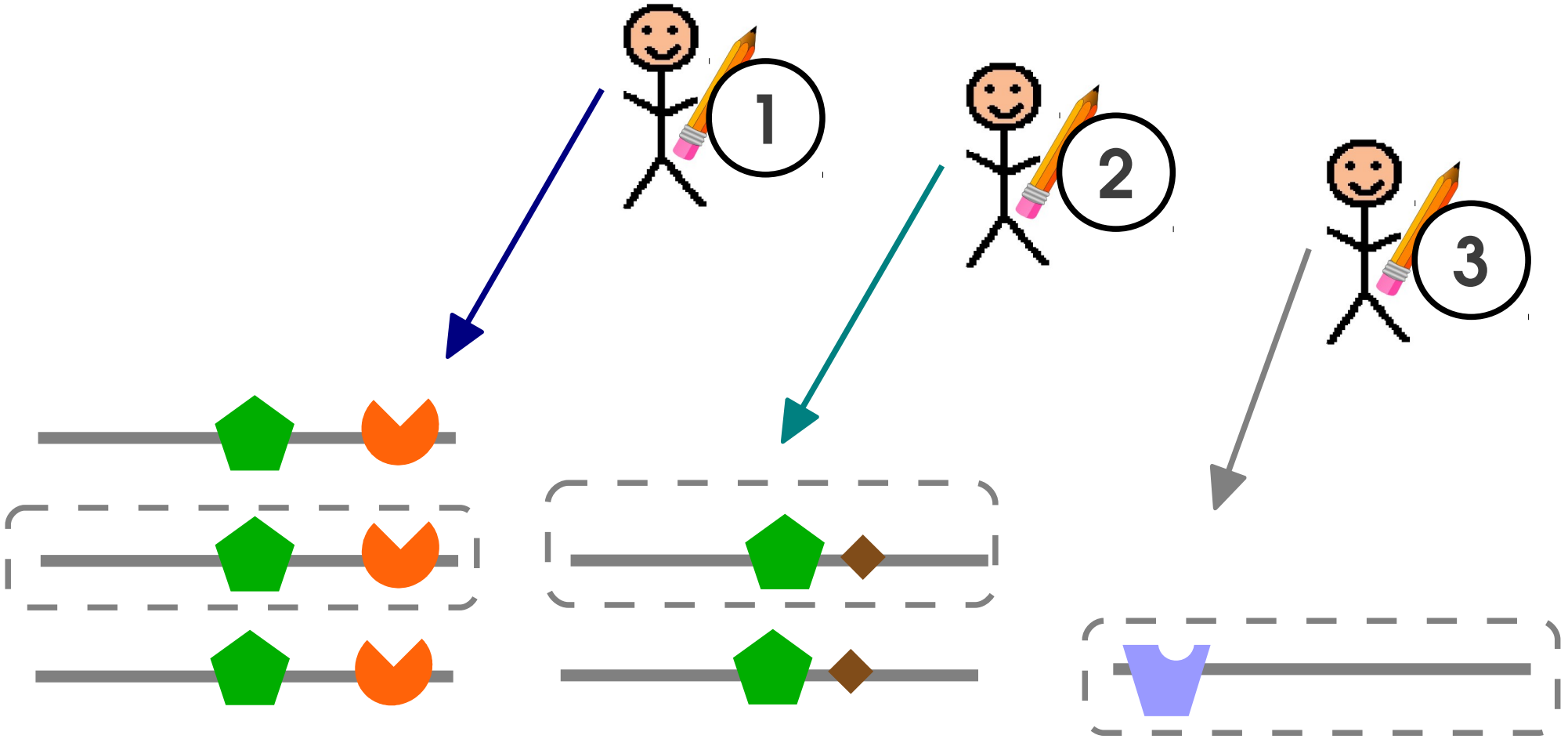


Distribution of Enzyme Commission numbers



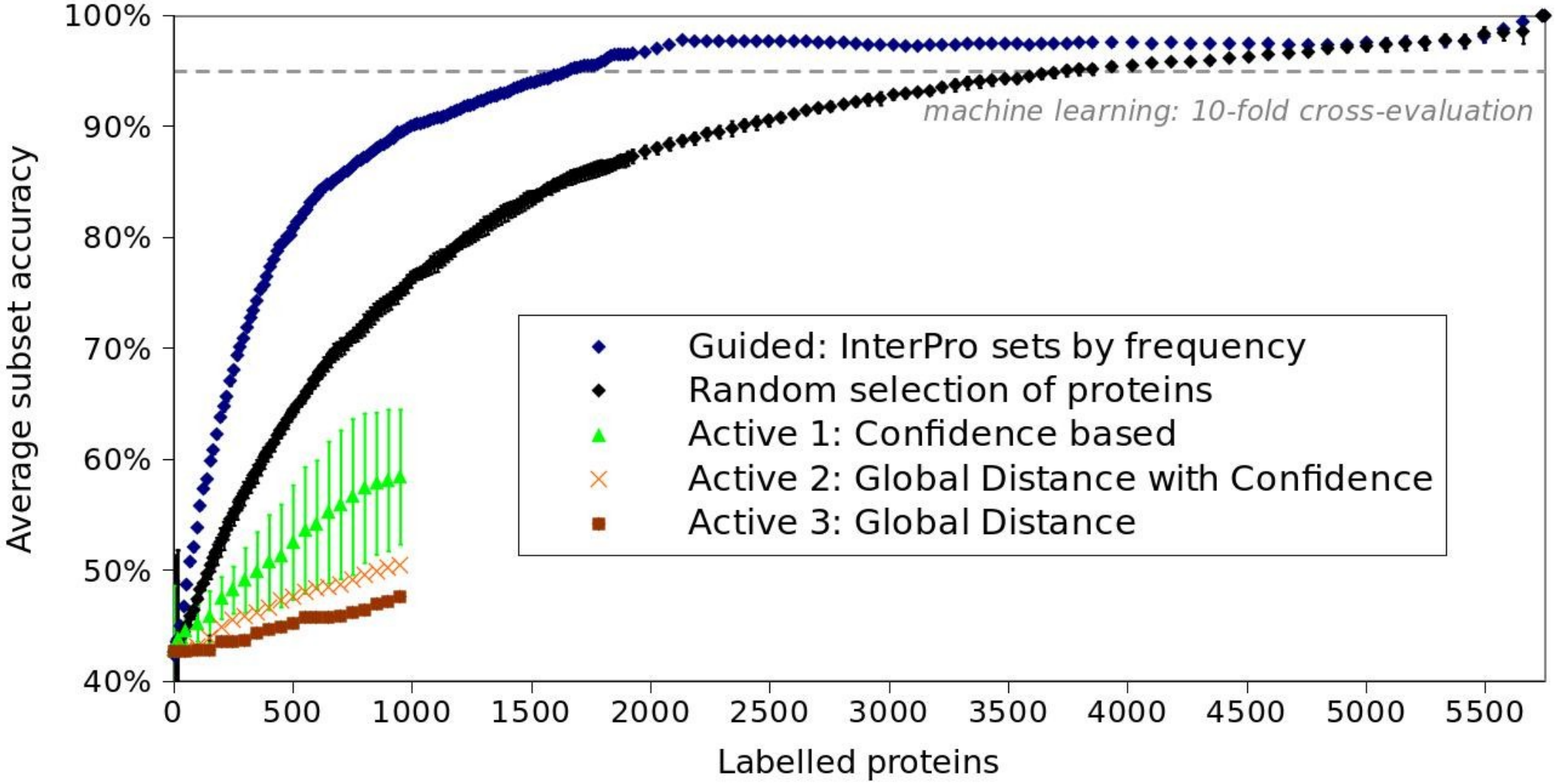
Guided learning

tackle InterPro signature sets by frequency



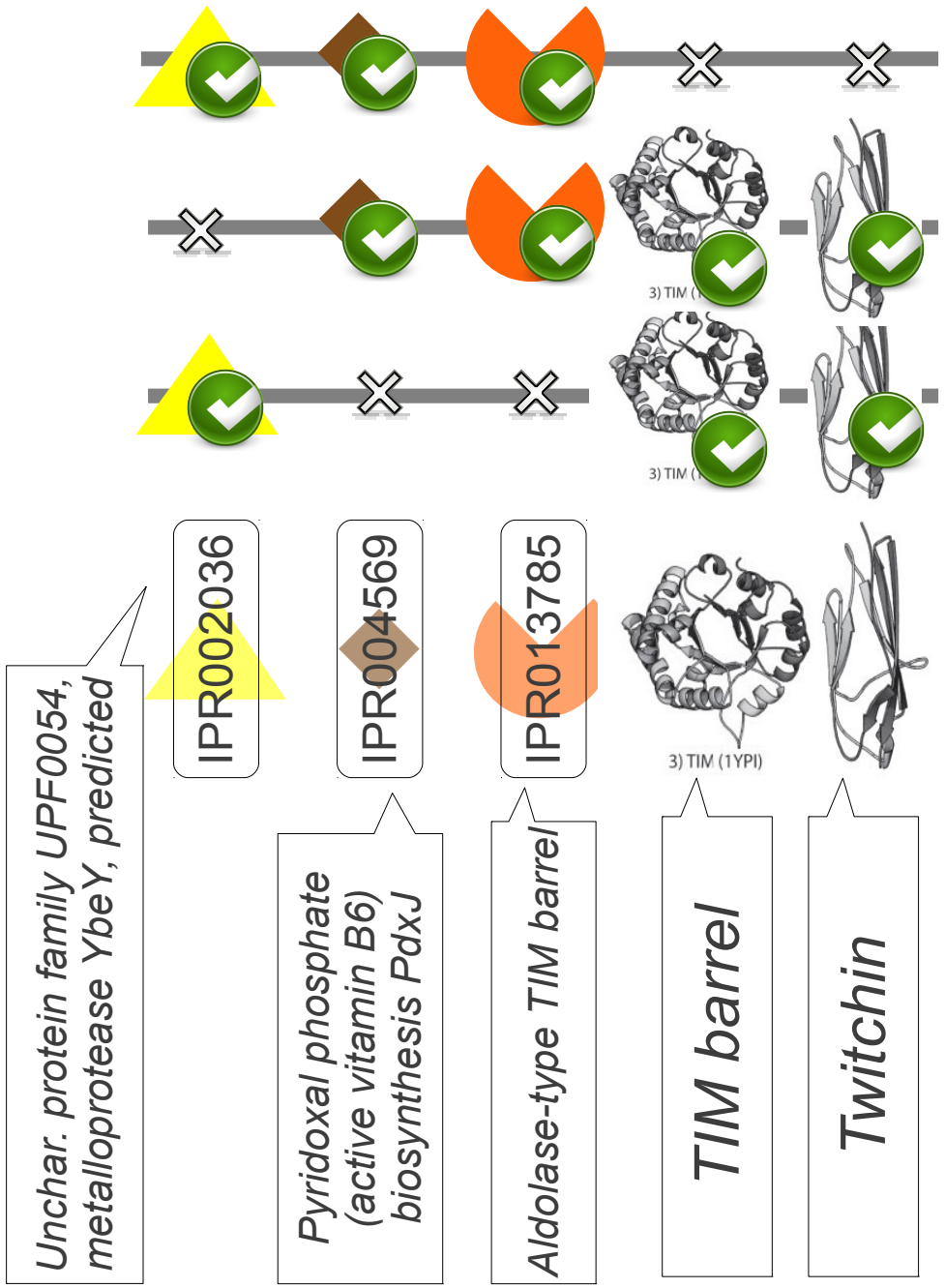
Inspired by: Attenberg, J. & Provost, F.
Why label when you can search?: alternatives to active learning for applying human resources to build classification models under extreme class imbalance. 16th ACM SIGKDD, ACM, 2010, 423-43

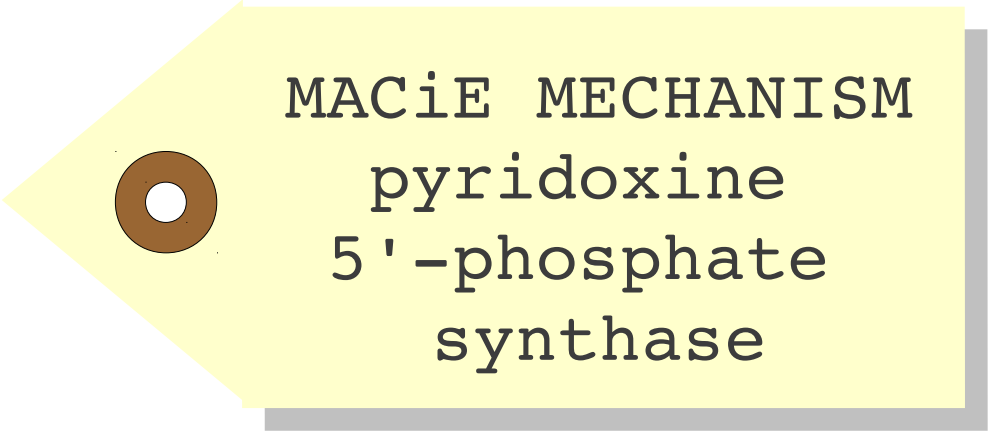
Guided vs. Active Learning



Attributes

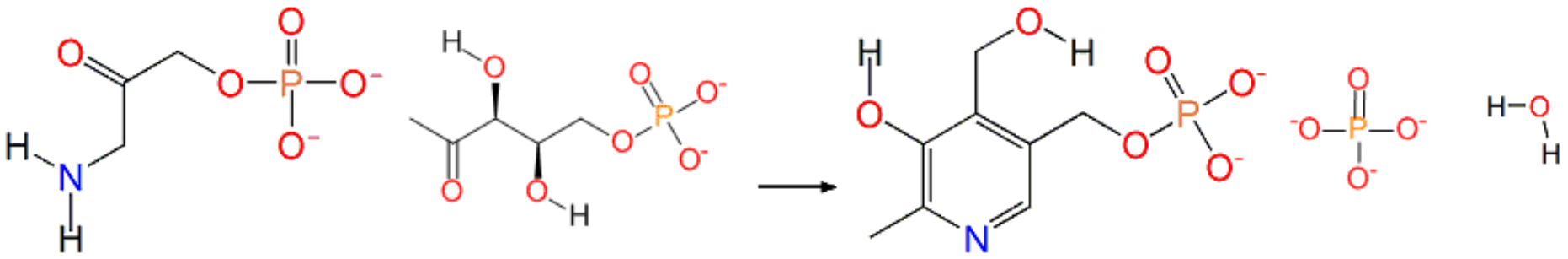
Class labels





Current work: predict enzyme **chemical mechanism**

MACiE: Mechanism, Annotation and Classification in Enzymes



3-amino-2-oxopropyl phosphate
C11638

1-deoxy-D-xylulose
5-phosphate
C11437
CHEBI:57792

pyridoxine 5'-phosphate
C00627
CHEBI:28803

phosphate
C00009
CHEBI:18367

2 water
C00001
CHEBI:15377

Overall Comment: This enzyme forms part of the vitamin B6 biosynthesis pathway in bacteria.

View similar reactions

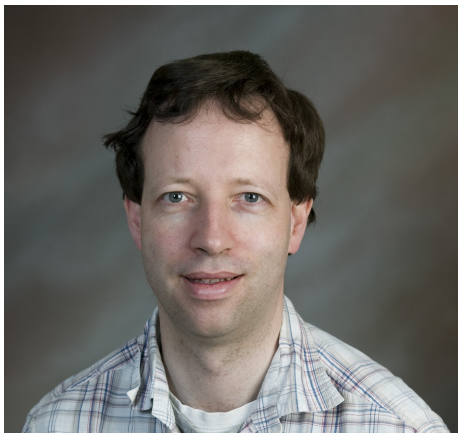
Stepwise Description of the Reaction

Step 1	The 3-amino-2-oxopropyl phosphate substrate initiates a nucleophilic attack upon the 1-deoxy-D-xylulose 5-phosphate in the first step of a Schiff base formation. The carbonyl oxygen is assumed to deprotonate the amine.
Step 2	The newly formed secondary amine initiates a intramolecular elimination of water, which obtains its proton first from Glu72, and then from water.
Step 3	His45 deprotonates the C3 of the intermediate, which initiates double bond rearrangement, with the Schiff base acting as an electron sink.
Step 4	The Schiff base reforms, initiating a double bond rearrangement and resulting in the elimination of water, which obtains it's proton from Glu72.
Step 5	Glu72 deprotonates the remaining hydroxide, which initiates a double bond rearrangement and eliminates phosphate.

From: http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/MACiE/entry/getPage.pl?id=M0243

Mitchell's group - School of Chemistry, University of St Andrews, United Kingdom

Lazaros Mavridis

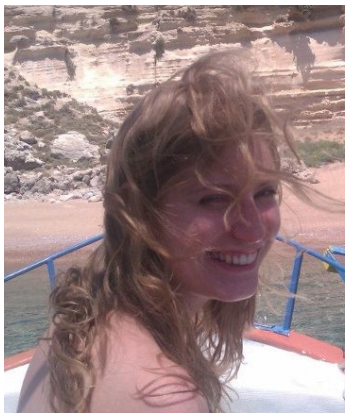


John Mitchell

Luna De Ferrari

Neetika Nath

James Mc Donagh



Rosanna Alderson

Stuart Aitken – Univ. of Edinburgh, School of Informatics & SynthSys

Thank you

Questions?

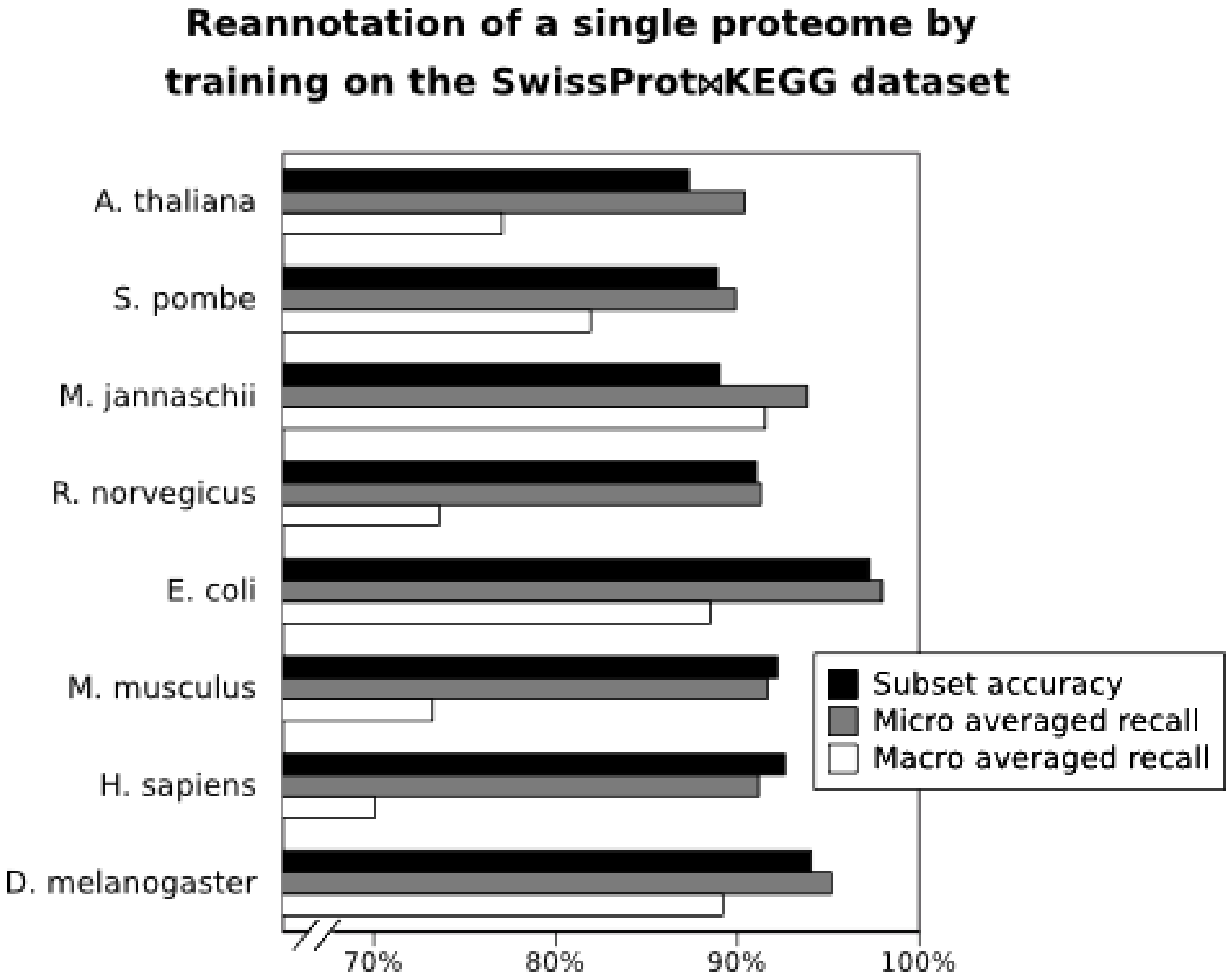
Luna De Ferrari
ldf@st-andrews.ac.uk

University of St Andrews

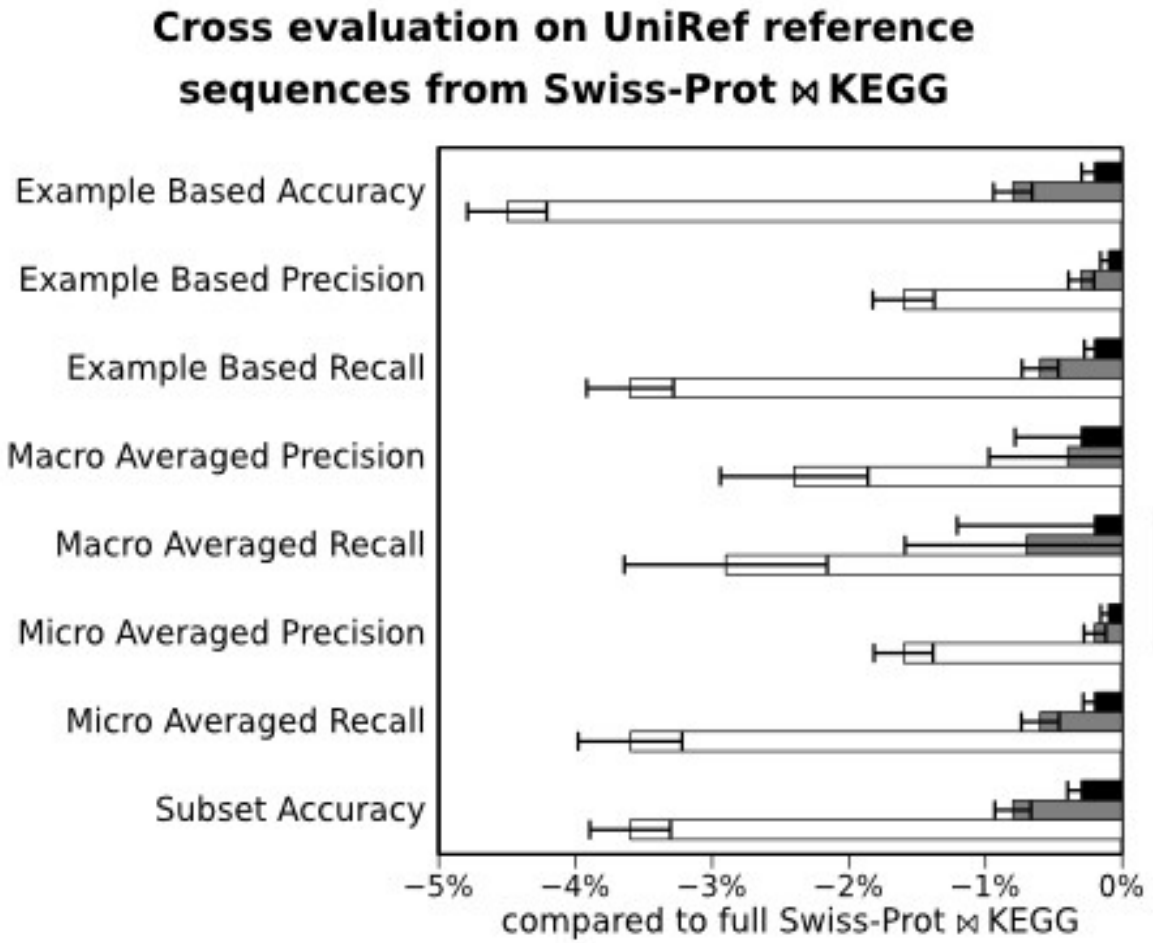
BBSRC bioscience for the future

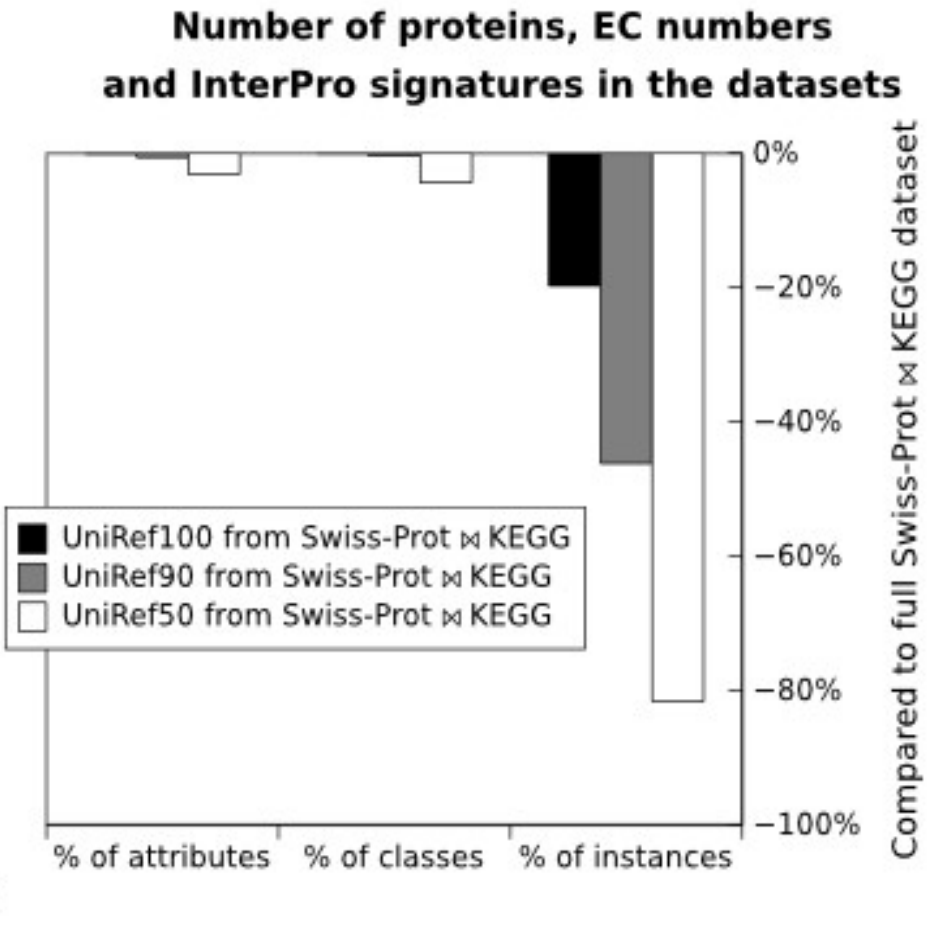
THE UNIVERSITY OF EDINBURGH

Proteome re-annotation



L. De Ferrari et al. BMC Bioinformatics 2012, 13:61





EnzML: Multi-label prediction of enzyme classes using InterPro signatures,
L. De Ferrari et al. BMC Bioinformatics 2012, 13:61

- Why label when you can search?: alternatives to active learning for applying human resources to build classification models under extreme class imbalance
- Authors: Josh Attenberg Polytechnic Institute of NYU, Brooklyn, NY, USA
- Foster Provost NYU Stern School of Business, New York, NY, USA
- KDD '10 Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining
- Pages 423-432
- ACM New York, NY, USA ©2010
- This paper analyses alternative techniques for deploying low-cost human resources for data acquisition for classifier induction in domains exhibiting extreme class imbalance - where traditional labeling strategies, such as active learning, can be ineffective. Consider the problem of building classifiers to help brands control the content adjacent to their on-line advertisements. Although frequent enough to worry advertisers, objectionable categories are rare in the distribution of impressions encountered by most on-line advertisers - so rare that traditional sampling techniques do not find enough positive examples to train effective models. An alternative way to deploy human resources for training-data acquisition is to have them "guide" the learning by searching explicitly for training examples of each class. We show that under extreme skew, even basic techniques for guided learning completely dominate smart (active) strategies for applying human resources to select cases for labeling. Therefore, it is critical to consider the relative cost of search versus labeling, and we demonstrate the tradeoffs for different relative costs. We show that in cost/skew settings where the choice between search and active labeling is equivocal, a hybrid strategy can combine the benefits.