MINING THE *MYCOBACTERIUM TUBERCULOSIS* FLAVOPROTEOME: A BIOINFORMATIC APPROACH.

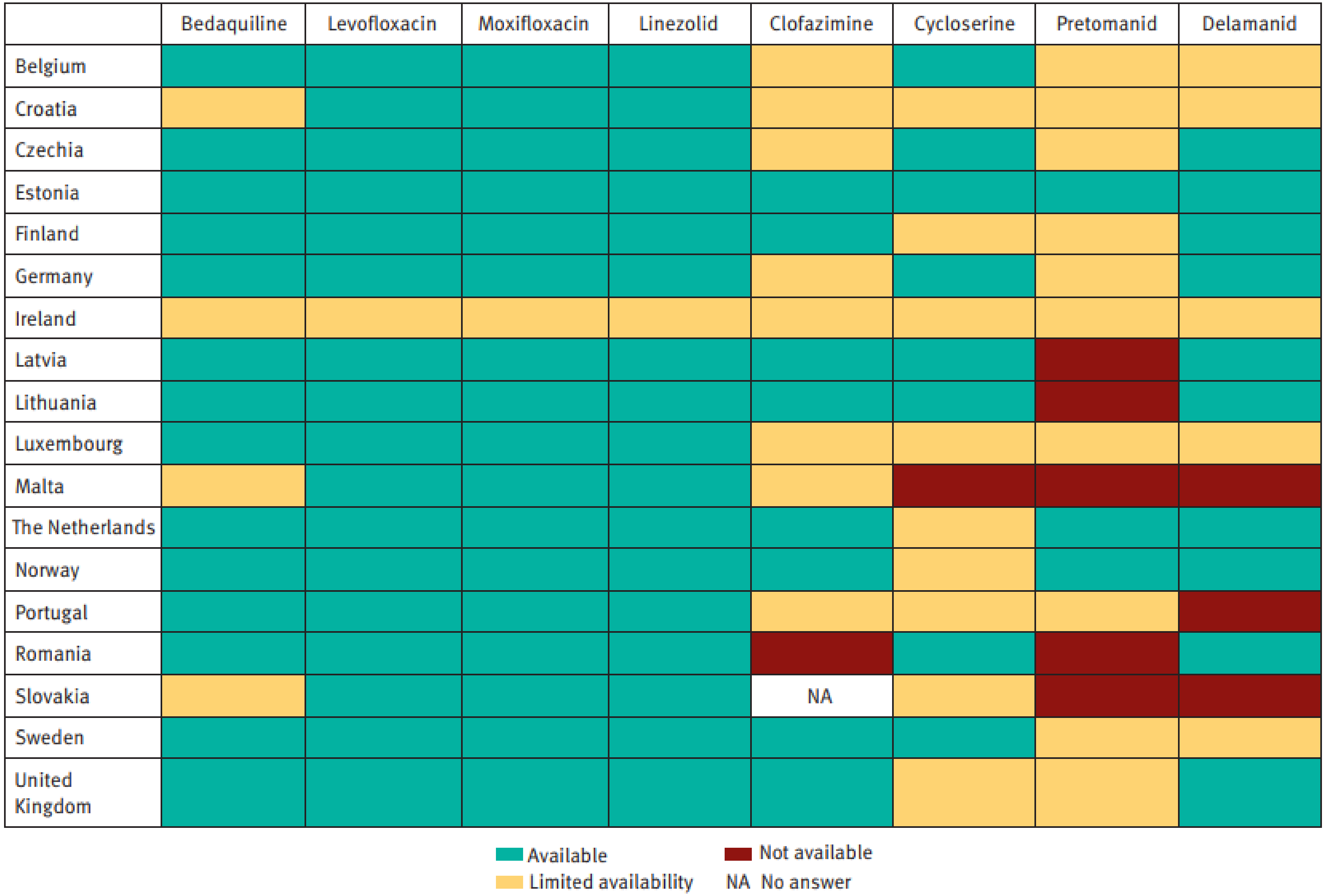
Speaker: Raquel Ventura Baños
Director: Milagros Medina Trullenque
Codirector: Marta Martínez Júlvez

UNIVERSIDAD DE ZARAGOZA – DPTO. BIOQUÍMICA – BIFI

Master in Biophysics and Quantitative Biotechnology

Pharmaceutical relevance of the study and drug target search

Availability of drugs in different countries, Europe survey, central and western Europe, October 2023 , Otto-Knapp et al., 2024

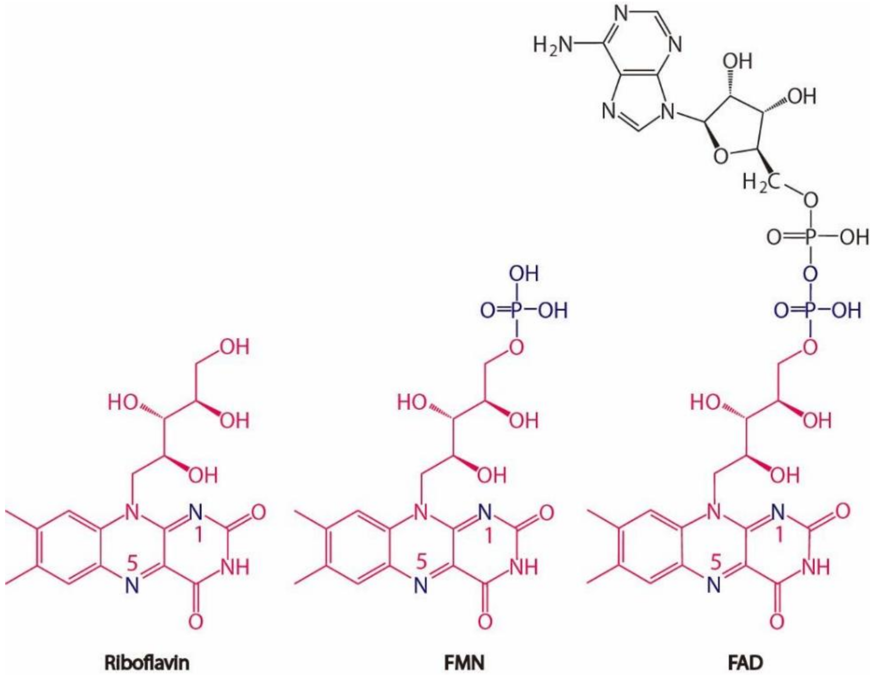


	Bedaquiline	Levofloxacin	Moxifloxacin	Linezolid	Clofazimine	Cycloserine	Pretomanid	Delamanid
Belgium	🟩	🟩	🟩	🟩	🟨	🟩	🟨	🟨
Croatia	🟨	🟩	🟩	🟩	🟨	🟨	🟨	🟨
Czechia	🟩	🟩	🟩	🟩	🟨	🟩	🟨	🟩
Estonia	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Finland	🟩	🟩	🟩	🟩	🟩	🟨	🟨	🟩
Germany	🟩	🟩	🟩	🟩	🟨	🟩	🟨	🟩
Ireland	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Latvia	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩
Lithuania	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩
Luxembourg	🟩	🟩	🟩	🟩	🟨	🟨	🟨	🟨
Malta	🟨	🟩	🟩	🟩	🟨	🟥	🟥	🟥
The Netherlands	🟩	🟩	🟩	🟩	🟩	🟨	🟩	🟩
Norway	🟩	🟩	🟩	🟩	🟩	🟨	🟩	🟩
Portugal	🟩	🟩	🟩	🟩	🟨	🟨	🟨	🟥
Romania	🟩	🟩	🟩	🟩	🟥	🟩	🟥	🟩
Slovakia	🟨	🟩	🟩	🟩	NA	🟨	🟥	🟥
Sweden	🟩	🟩	🟩	🟩	🟩	🟩	🟨	🟨
United Kingdom	🟩	🟩	🟩	🟩	🟩	🟨	🟨	🟩

🟩 Available
🟨 Limited availability
🟥 Not available
NA No answer

Previous knowledge about *M. tuberculosis* flavoproteome

Riboflavin derivatives, Zhang et al., 2020.

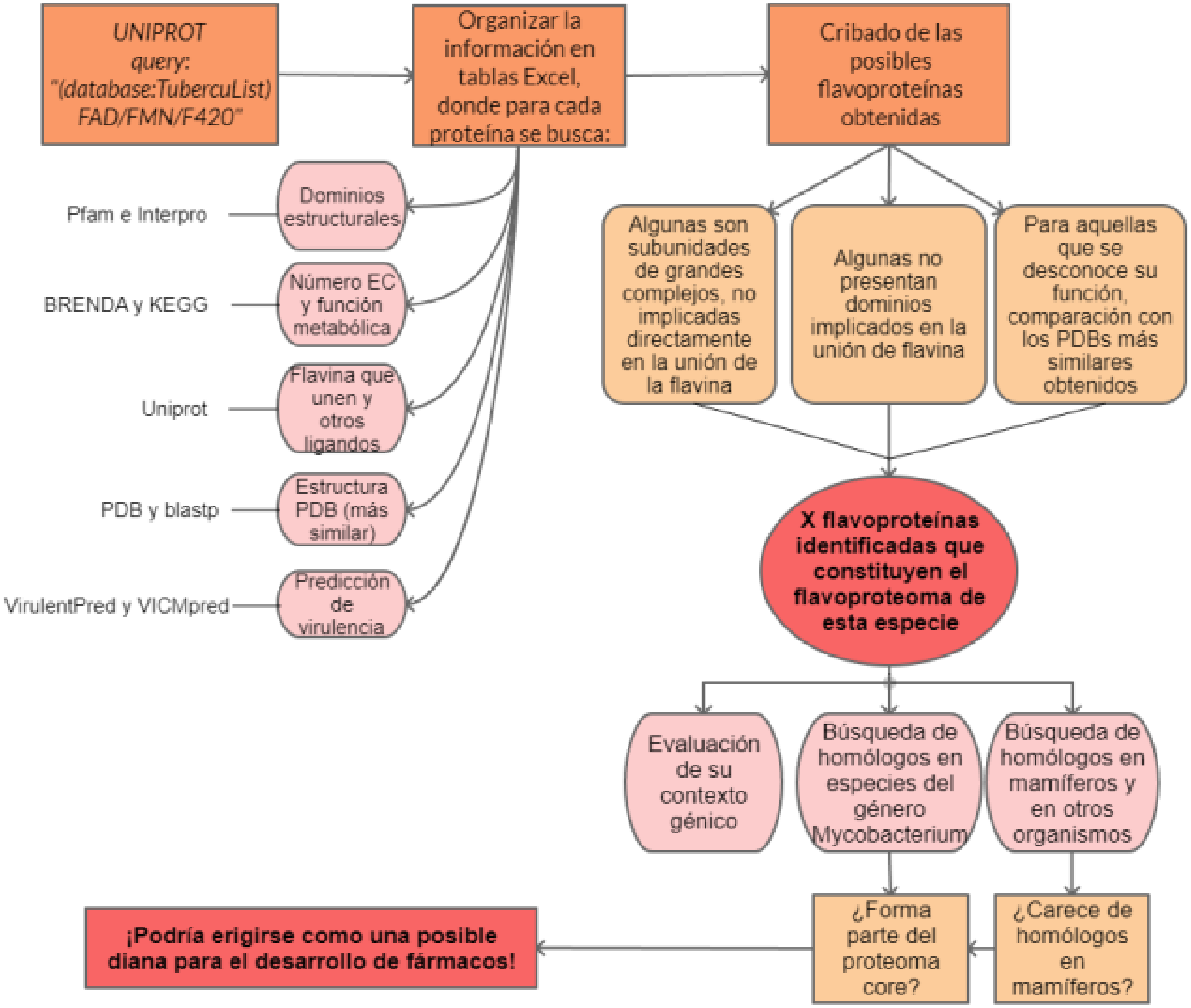


The flavoproteome content has only been reported for a small number of species with different protein diversity.

While human and *Saccharomyces cerevisiae* flavoproteomes contain 78 and 48 different proteins respectively, *Arabidopsis thaliana* has more than 200.

Montesa's study lacked the complete identification of the potential function of 133 out of the 184 envisaged as flavoproteins and flavoenzymes in *M. tuberculosis*.

Previous search workflow, Montesa et al., 2023 TFG.



UNIPROT query: "(database:TubercuList) FAD/FMN/F420"

Organizar la información en tablas Excel, donde para cada proteína se busca:

Cribado de las posibles flavoproteínas obtenidas

Pfam e Interpro – Dominios estructurales

BRENDA y KEGG – Número EC y función metabólica

Uniprot – Flavina que unen y otros ligandos

PDB y blastp – Estructura PDB (más similar)

VirulentPred y VICMpred – Predicción de virulencia

Algunas son subunidades de grandes complejos, no implicadas directamente en la unión de la flavina

Algunas no presentan dominios implicados en la unión de flavina

Para aquellas que se desconoce su función, comparación con los PDBs más similares obtenidos

X flavoproteínas identificadas que constituyen el flavoproteoma de esta especie

Evaluación de su contexto génico

Búsqueda de homólogos en especies del género Mycobacterium

Búsqueda de homólogos en mamíferos y en otros organismos

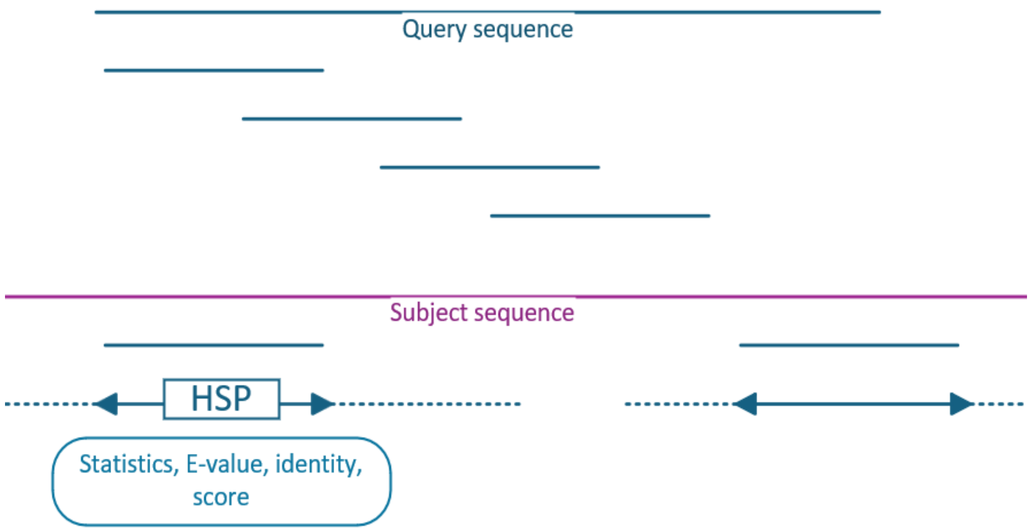
¿Forma parte del proteoma core?

¿Carece de homólogos en mamíferos?

¡Podría erigirse como una posible diana para el desarrollo de fármacos!

Similarity search

BLAST sequence similarity search algorithm diagram.



HSP: high segment scoring pair

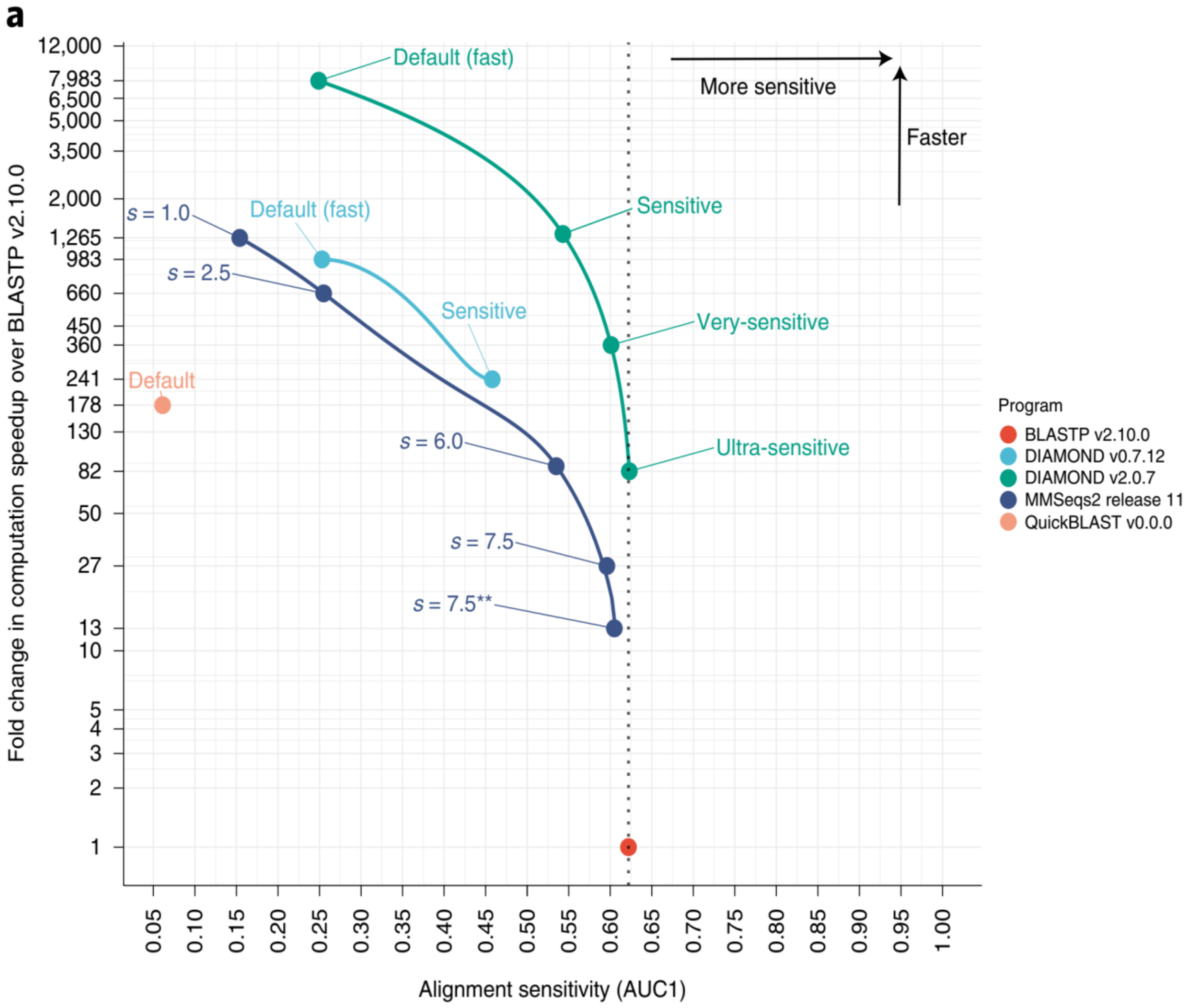
MMSeqs

- **K-mers** extension with inexact matches and **multiple processors (servers)**

Diamond (faster, blast sensitivity)

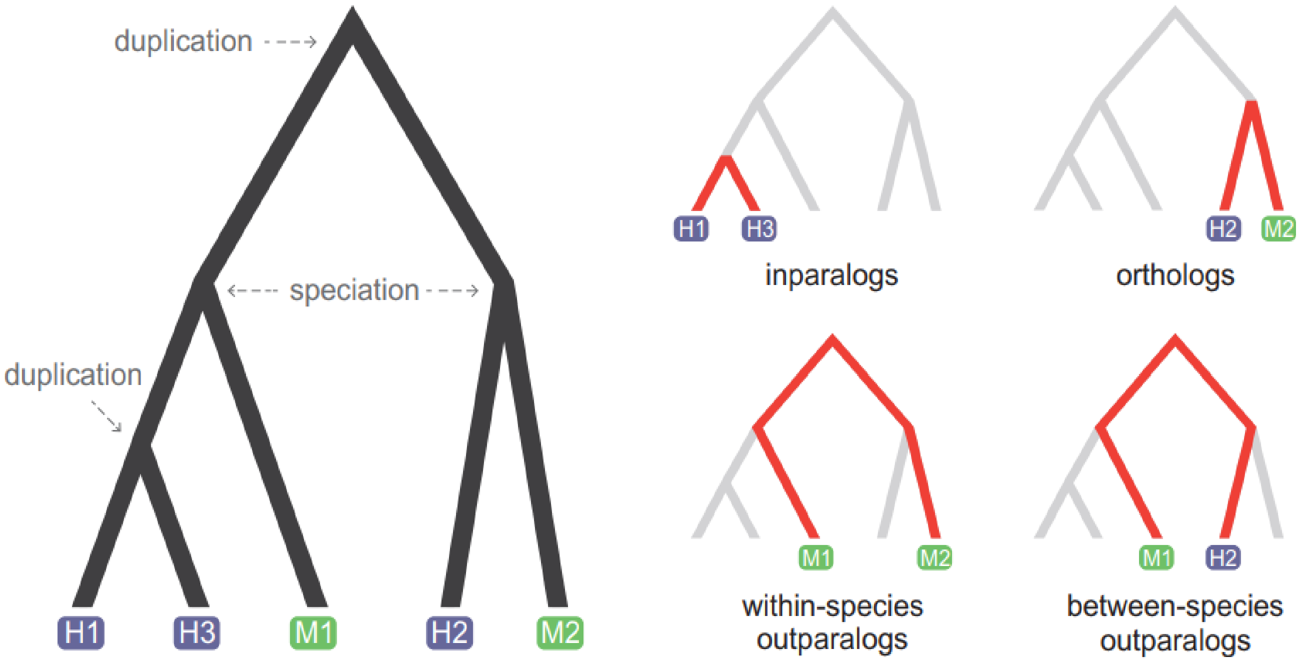
- **Reduced alphabet** and database **blocks**
- **Larger and spaced seeds** in different shapes
- **Double indexing**

Benchmark of DIAMND, MMSeqs2 and BLASTP Buchfink et al., 2021.



Orthology

Four types of homology relations, Stamboulian et al., 2020



- **Orthologs** are similar sequences that share a **common ancestor**, while **paralogs** are generated by **duplication**.

- **Ortholog conjecture** states that proteins from orthologs have higher chances to share functionality, orthologs are often used to predict function based on similarity.

- There is **ongoing debate about the terms** that are often used to describe other evolutionary relationships. This has led to the development of related terms to duplication and speciation.

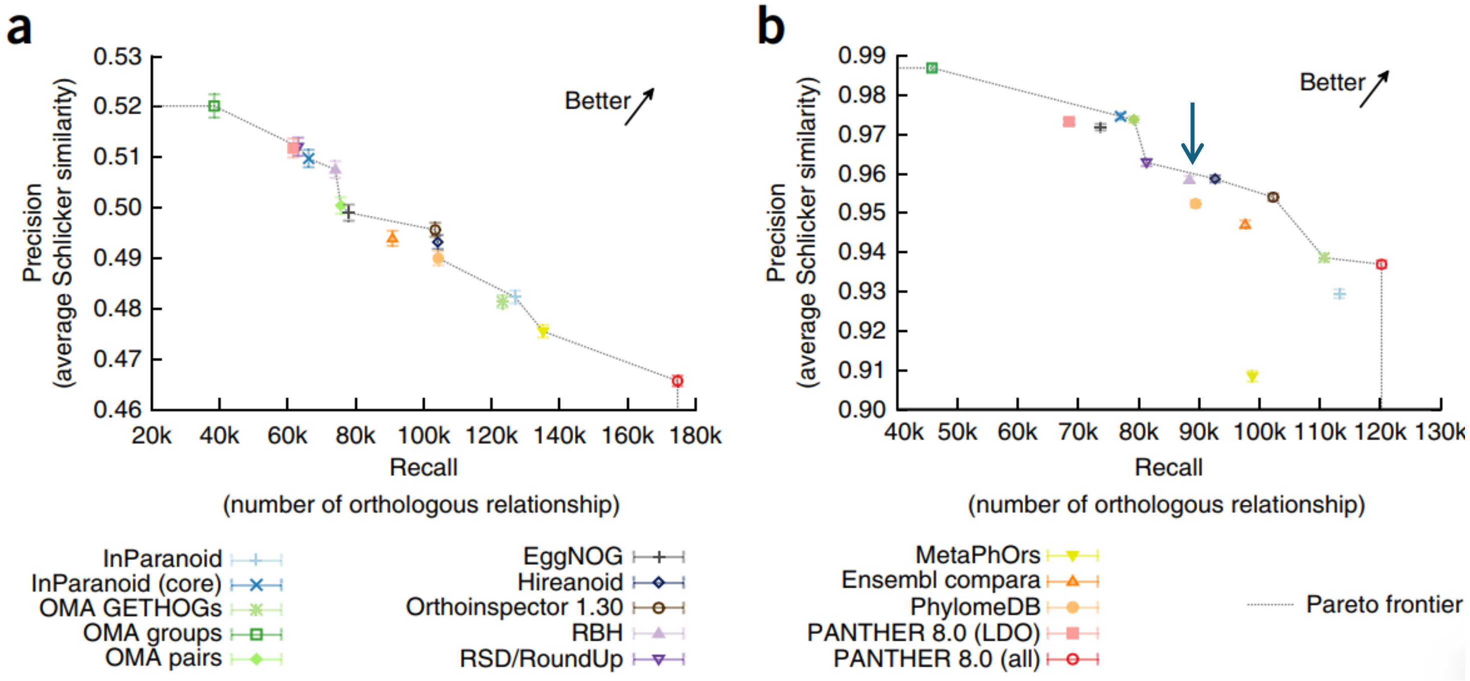
Objective

To deepen into our knowledge of the flavoproteome of *M. tuberculosis* by particularly:

- creating an efficient method of searching for orthologs.
- predicting potential catalytic activity based on protein sequences.

Experimental design; precision and annotations

Benchmarking of different ortholog search methods, Altenhoff et al., 2016



a. Experimentally supported **GO annotations**: sometimes only annotate some functional characteristics that are often related but are not necessarily indicative of similar catalytic activity.

b. Enzyme Commission **(EC) numbers**: data provided by spanning archaeal, bacterial and eukaryotic proteomes support the idea that orthologs can be a highly accurate predictor of enzyme functions in the way of enzyme commission numbers. (around 95% percent)

Experimental design; pipeline

- **Data**: all proteins detected as flavoproteins with catalytic activity records in UniProt.

- **Algorithm**: Diamond presents the fastest algorithm for similarity search that allows local running.

- **Complexity** (efficiency) and other methods: accurate methods rely on complex network approaches taking account structure, ppi apart from sequence.

- Taking some **genetic considerations** approach complexity is lower: **ortholog conjecture**.

  - **RBH**: Reciprocal best hits, best hits for the query in the whole filtered database represent best hits in the whole species genome, having best hits in both species, orthologs.

  - **UBH:** Unidirectional best hits are best hits from each taxonomic species resulting from first similarity search (more diverse homologs, with a high ortholog content)

Experimental design; orthology

RBH:

- **Gene loss is** not fully contemplated in this approach. Therefore, after such an event occurs, pairs of genes that are caused by duplication and subsequent speciation can be detected as orthologs and *vice versa*, causing FP (false positive) and FN (false negative) respectively (cases c and d in figure).

- In the presence of a **different number of duplications in each species**, only detects the most similar pair.

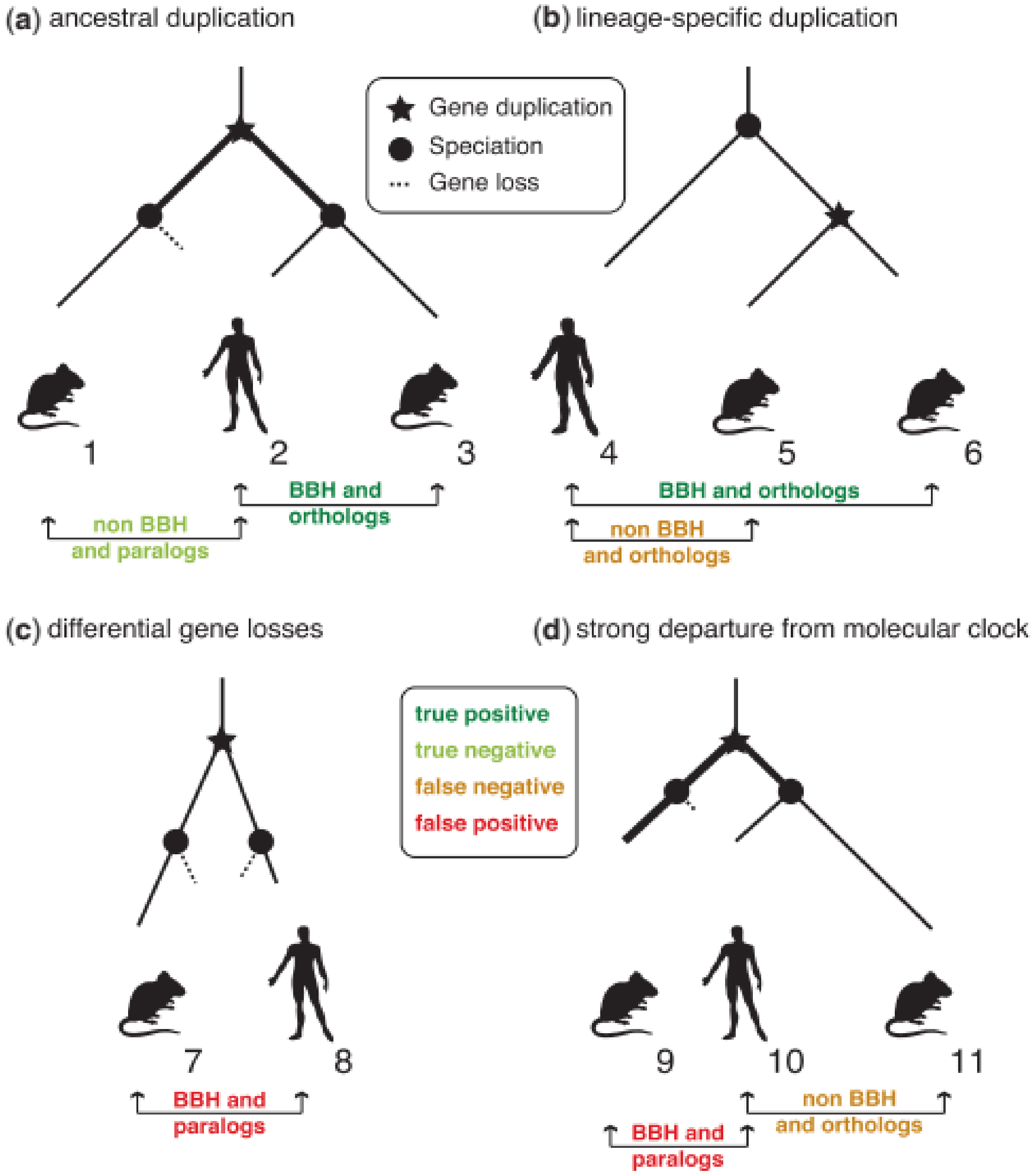
UBH:

- **May include gene loss** and **speciation after duplication** orthologs not contemplated by RBH.

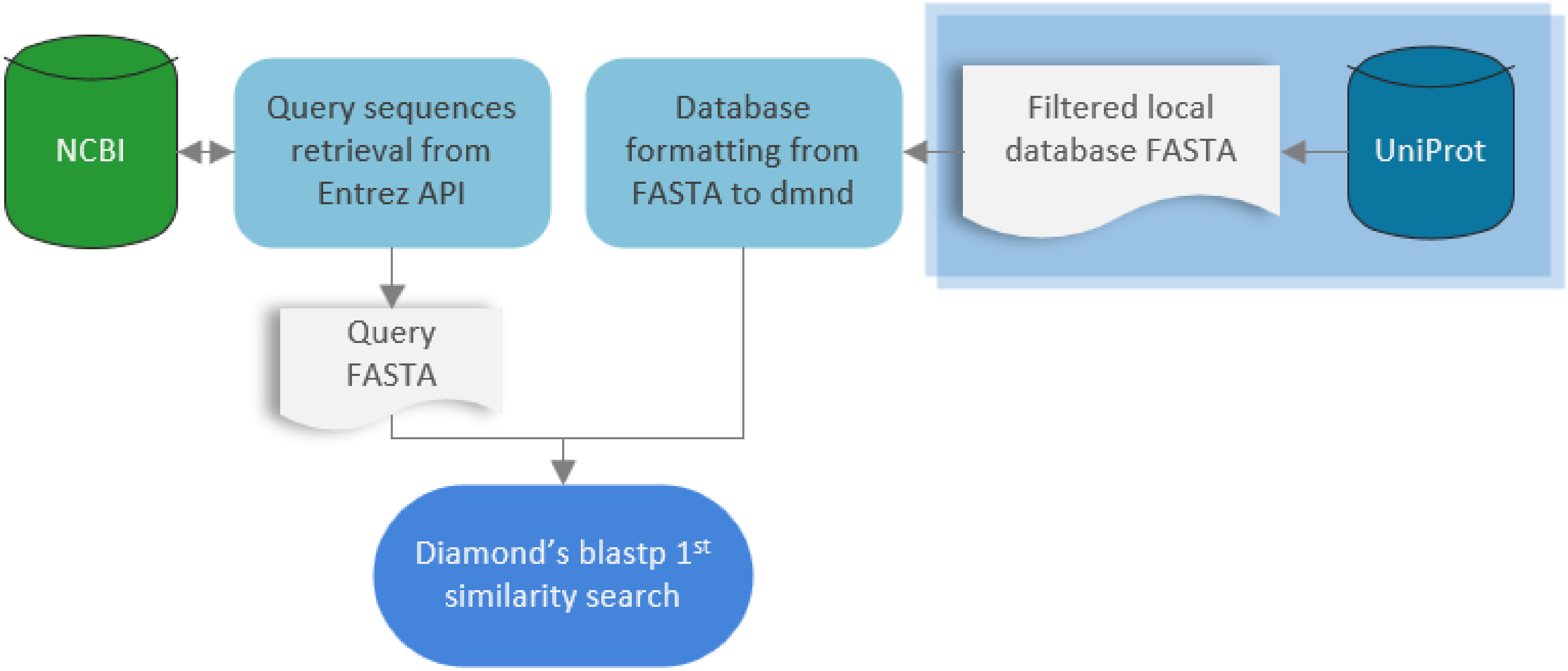
- Might **mislead more paralogs as orthologs** (non-reciprocity).

- Are **restricted to the best hits annotated** in the database.

Performance of RBH in conceptual examples, Dalquen and Dessimoz, 2013.

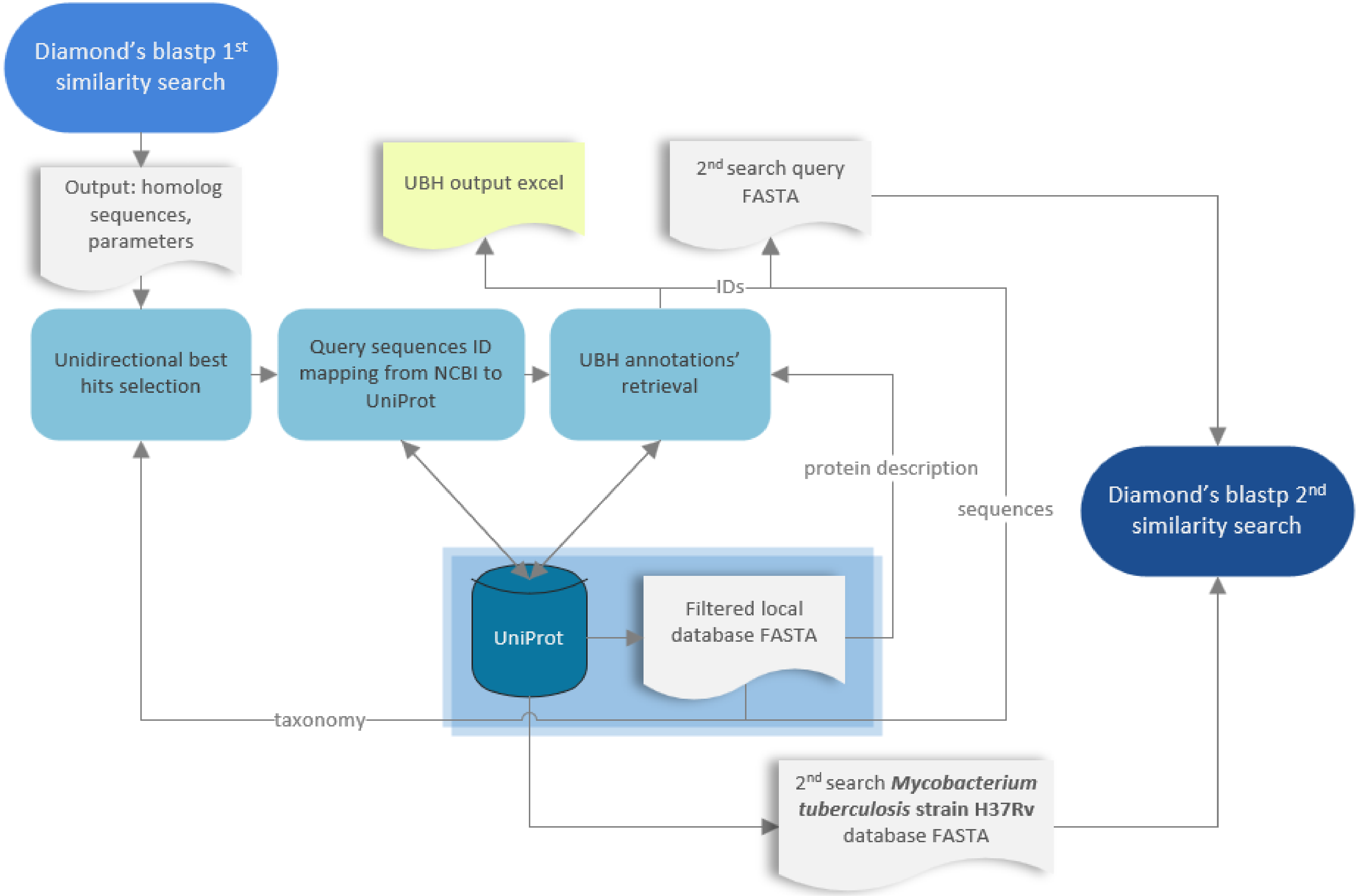


Workflow

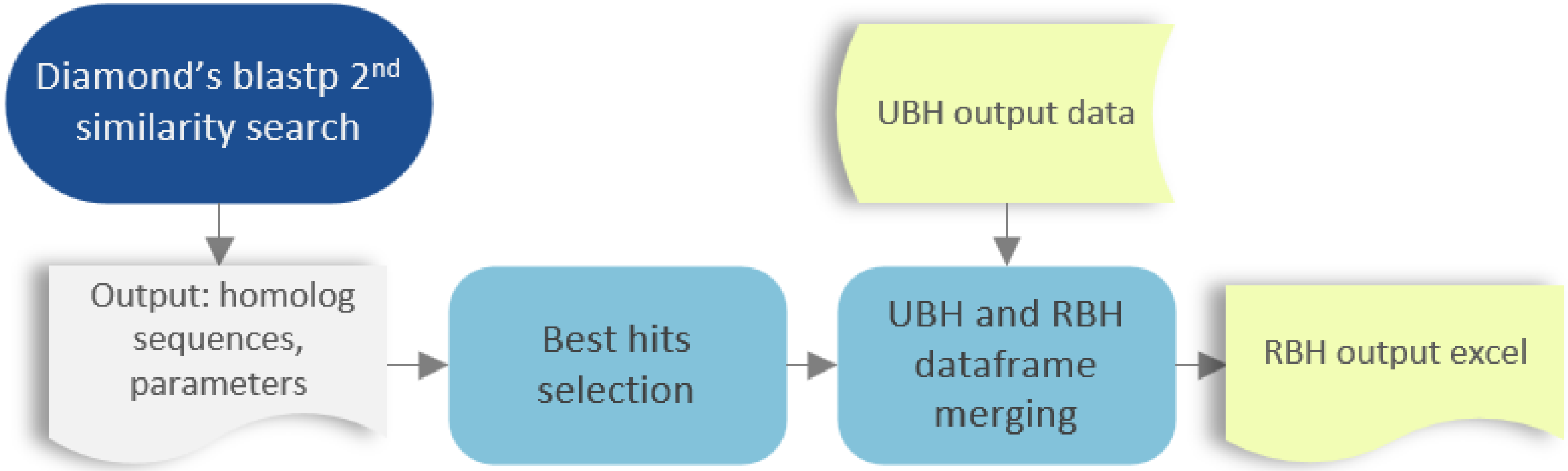


NCBI
Query sequences retrieval from Entrez API
Database formatting from FASTA to dmnd
Filtered local database FASTA
UniProt
Query FASTA
Diamond's blastp 1st similarity search

Workflow


Diamond's blastp 1st similarity search
Output: homolog sequences, parameters
Unidirectional best hits selection
Query sequences ID mapping from NCBI to UniProt
UBH annotations' retrieval
UBH output excel
2nd search query FASTA
IDs
protein description
sequences
Diamond's blastp 2nd similarity search
UniProt
Filtered local database FASTA
taxonomy
2nd search *Mycobacterium tuberculosis* strain H37Rv database FASTA

Workflow



Diamond's blastp 2nd similarity search

Output: homolog sequences, parameters

Best hits selection

UBH output data

UBH and RBH dataframe merging

RBH output excel

Statistical assessment

TP (true positives) = number of orthologs matching one or two annotations with its query annotation

FP = number of orthologs all of whose annotations do not match the query annotation

FN = number of orthologs of an annotated query with no annotations

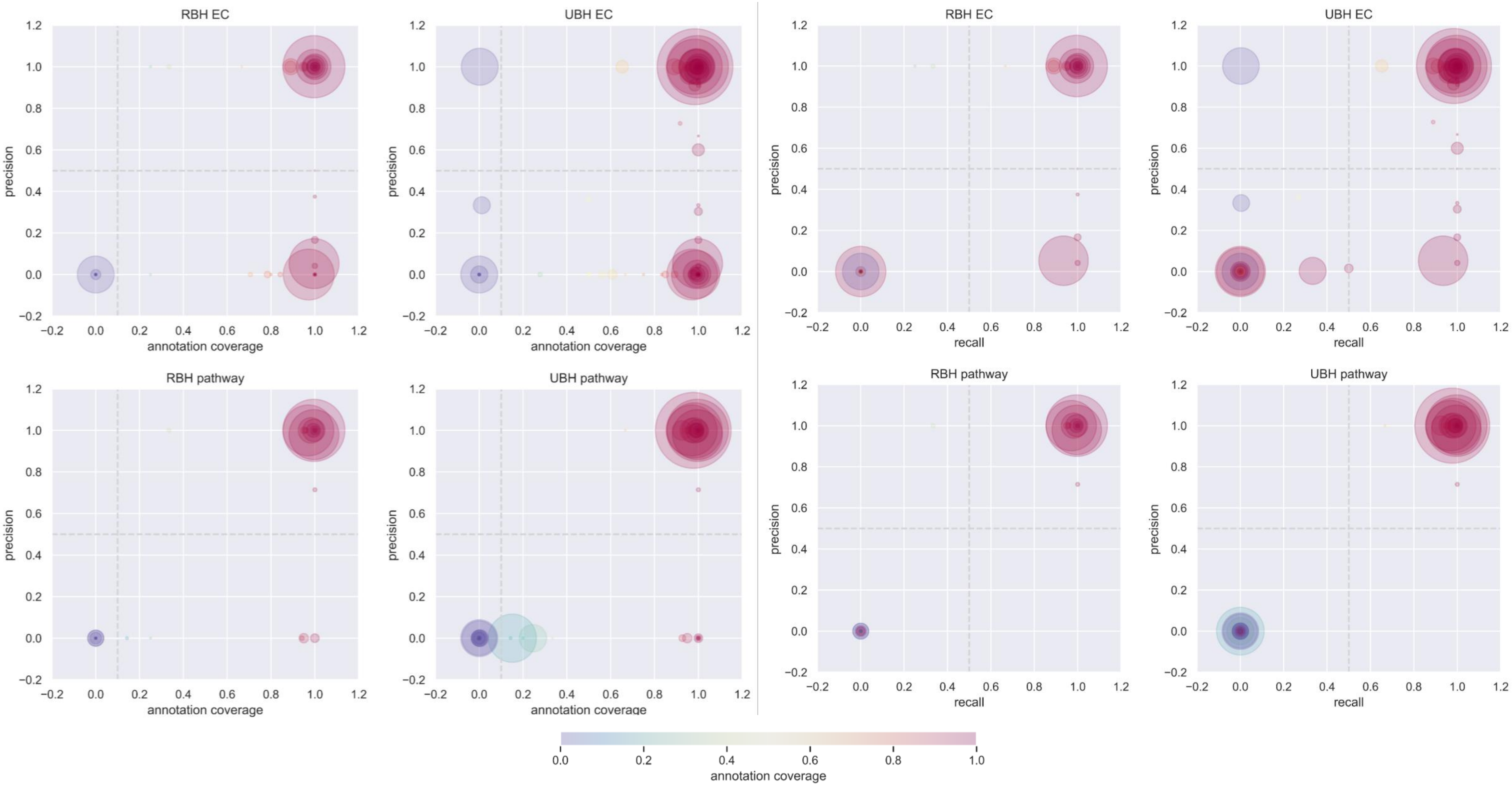
$$\text{precision} = \frac{\text{TP}}{\text{FP} + \text{TP}}$$

$$\text{recall} = \frac{\text{TP}}{\text{FN} + \text{TP}}$$

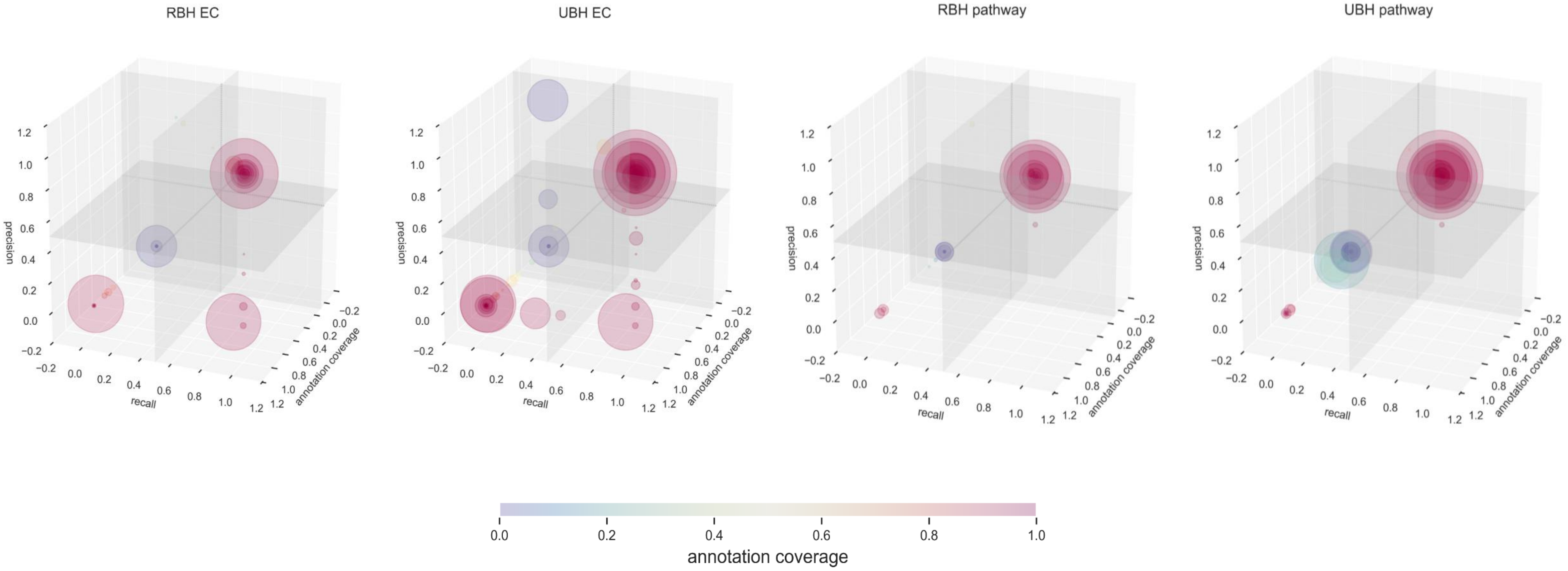
$$\text{ortholog number} = \text{FP} + \text{TP} + \text{FN}$$

$$\text{annotation coverage} = \frac{\text{TP} + \text{FP}}{\text{FP} + \text{TP} + \text{FN}}$$

Statistical assessment, recall, precision and annotation coverage

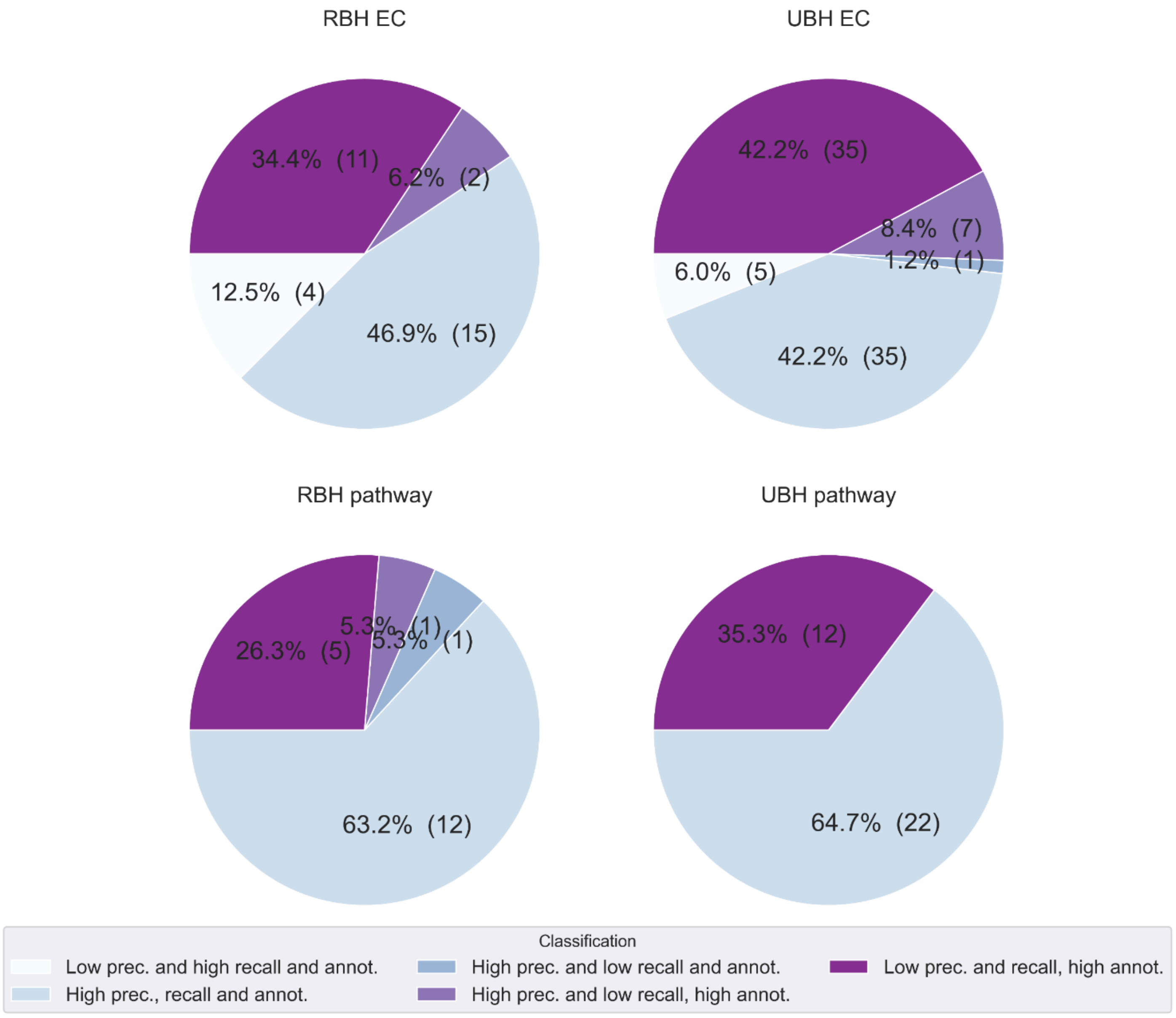


Statistical assessment, recall, precision and annotation coverage



Recall, precision and annotation coverage classifications population

- Queries with low annotation coverage, recall, and precision are those with no annotated hits (**FNs**), due to the stringent annotation threshold that excludes all unannotated orthologs.
- Low predictive precision, high annotation coverage and high recall mean numerous **FPs and TP presence**.
- Flavoproteins with high ortholog predictive precision, annotation coverage and recall are those with a high number of **TPs** and a low count of FN and FP.
- High precision, low recall and low or high annotation coverage means a strong presence of not annotated proteins (**FNs**) with a low count of **TPs**, that is, poorly annotated orthologs coincide with previous annotations.
- Predictions for queries with low recall and precision and high annotation coverage have **FNs and high FP count**.



E-value threshold

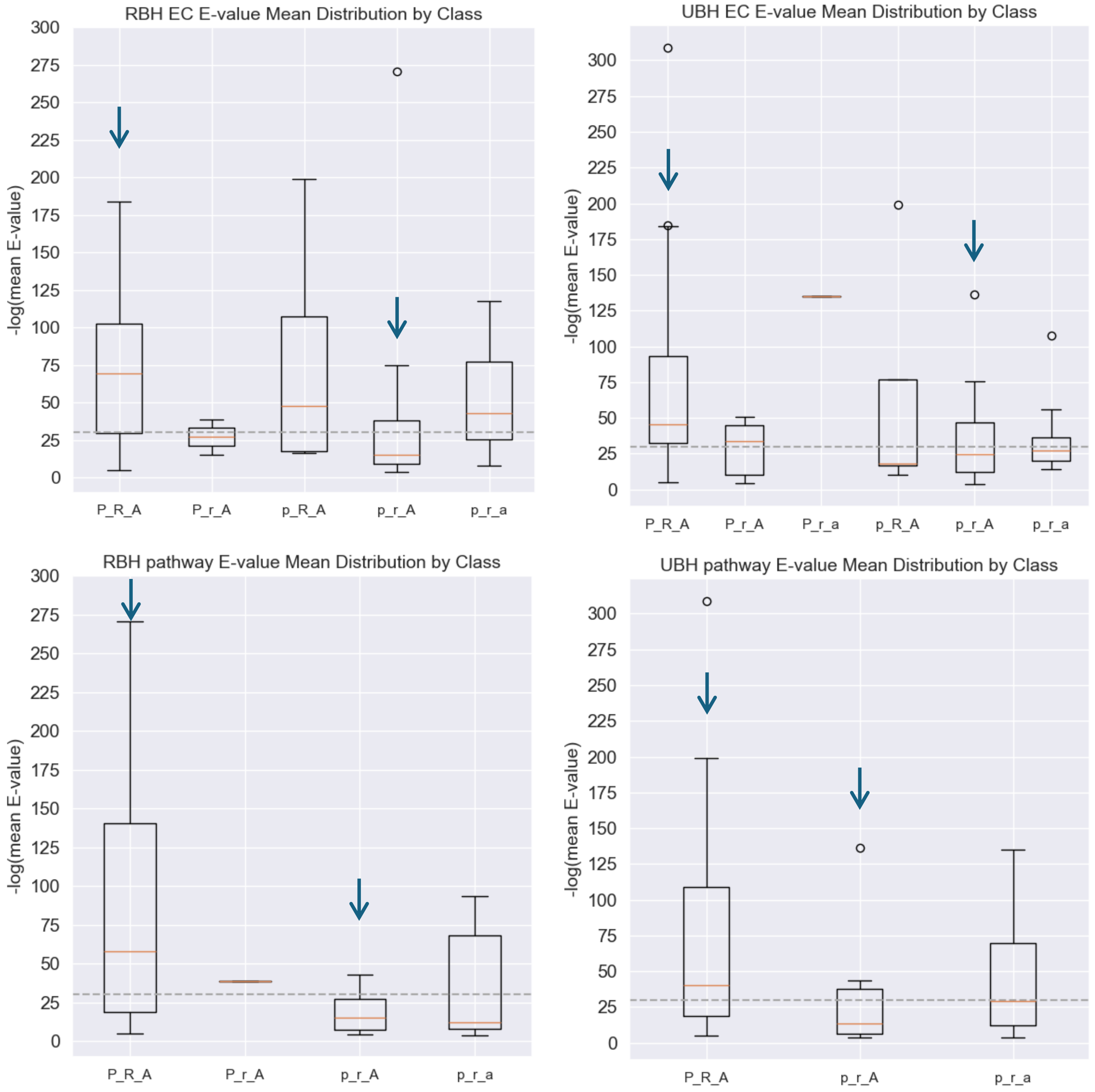
- Low predictive precision, high annotation coverage and high recall mean numerous FPs and TP presence.

- **Flavoproteins with high ortholog predictive precision, annotation coverage and recall are those with a high number of TPs** and a low count of FN and FP.

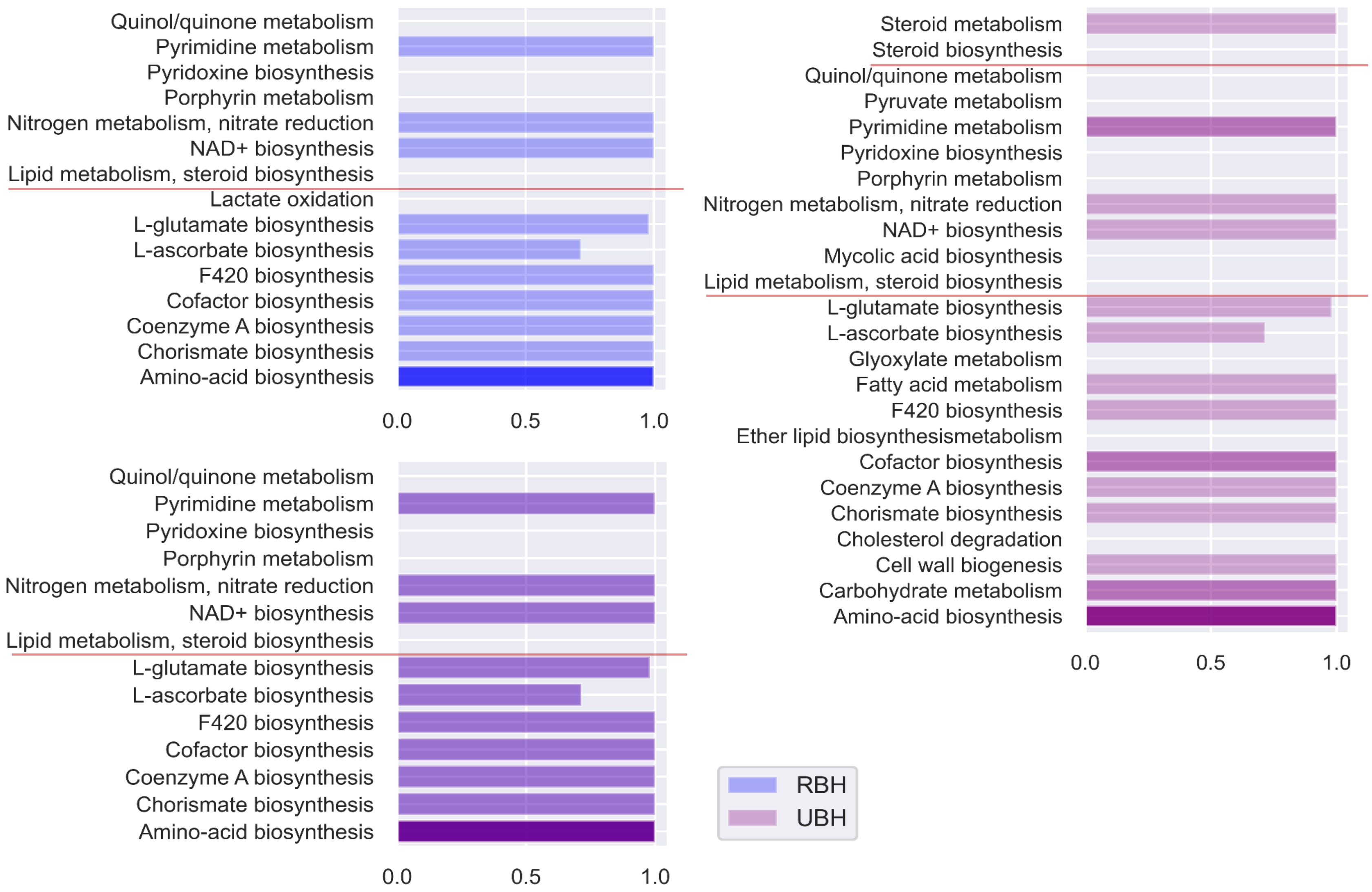
- High precision, low recall and low or high annotation coverage means a strong presence of not annotated proteins (FNs) with a low count of TPs, that is, poorly annotated orthologs coincide with previous annotations.

- **Predictions for queries with low recall and precision and high annotation coverage have FNs and high FP count.**

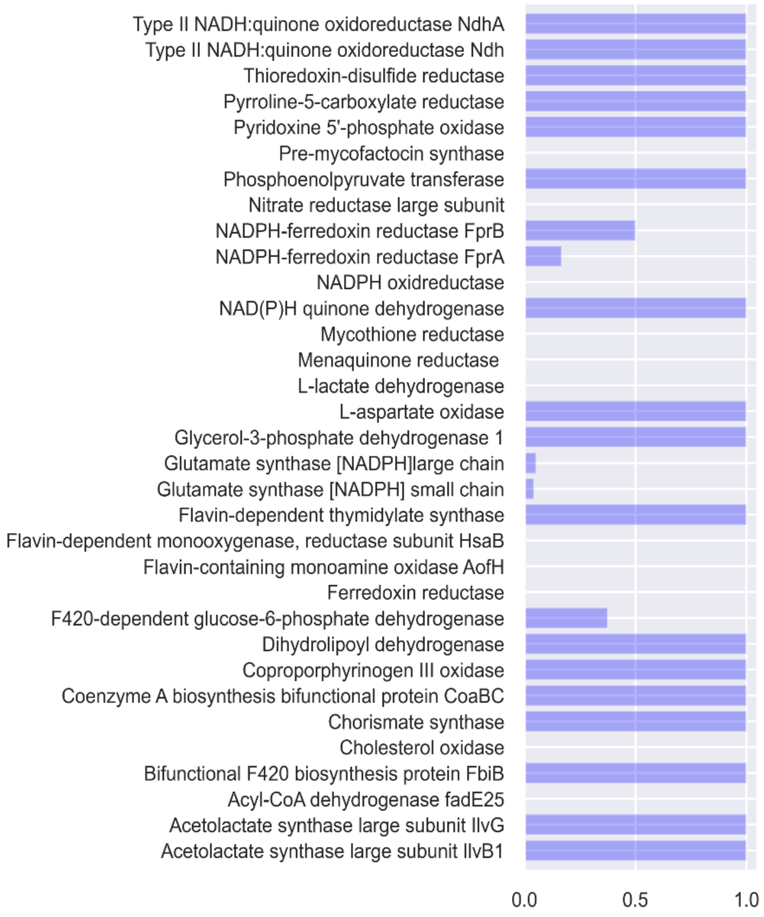
Negative base-10 logarithm of E-value means for each query for the different classifications

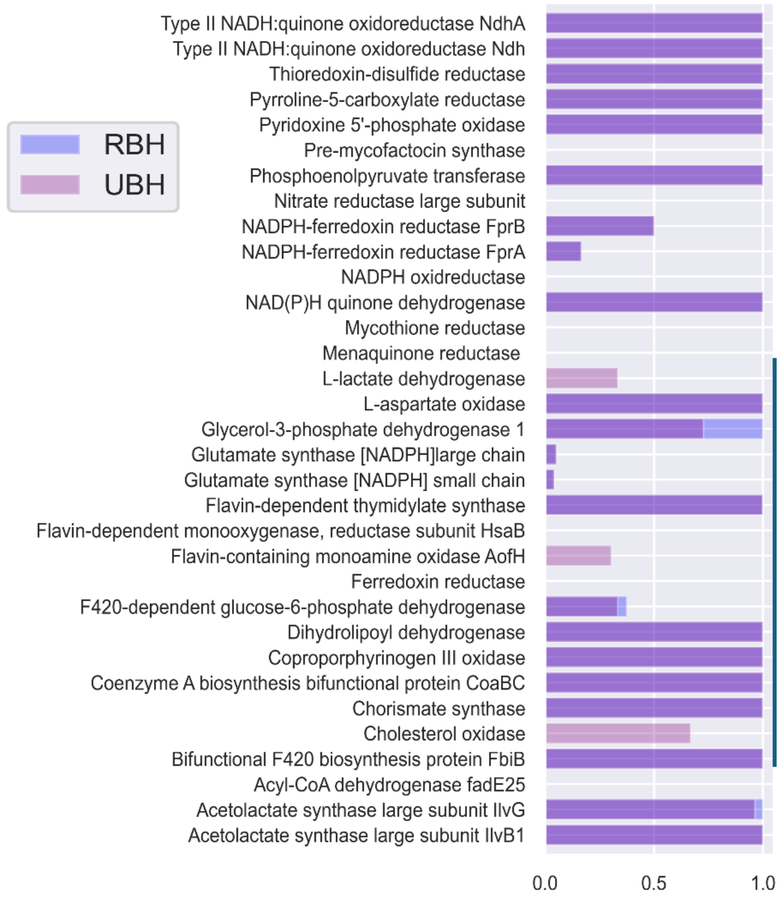


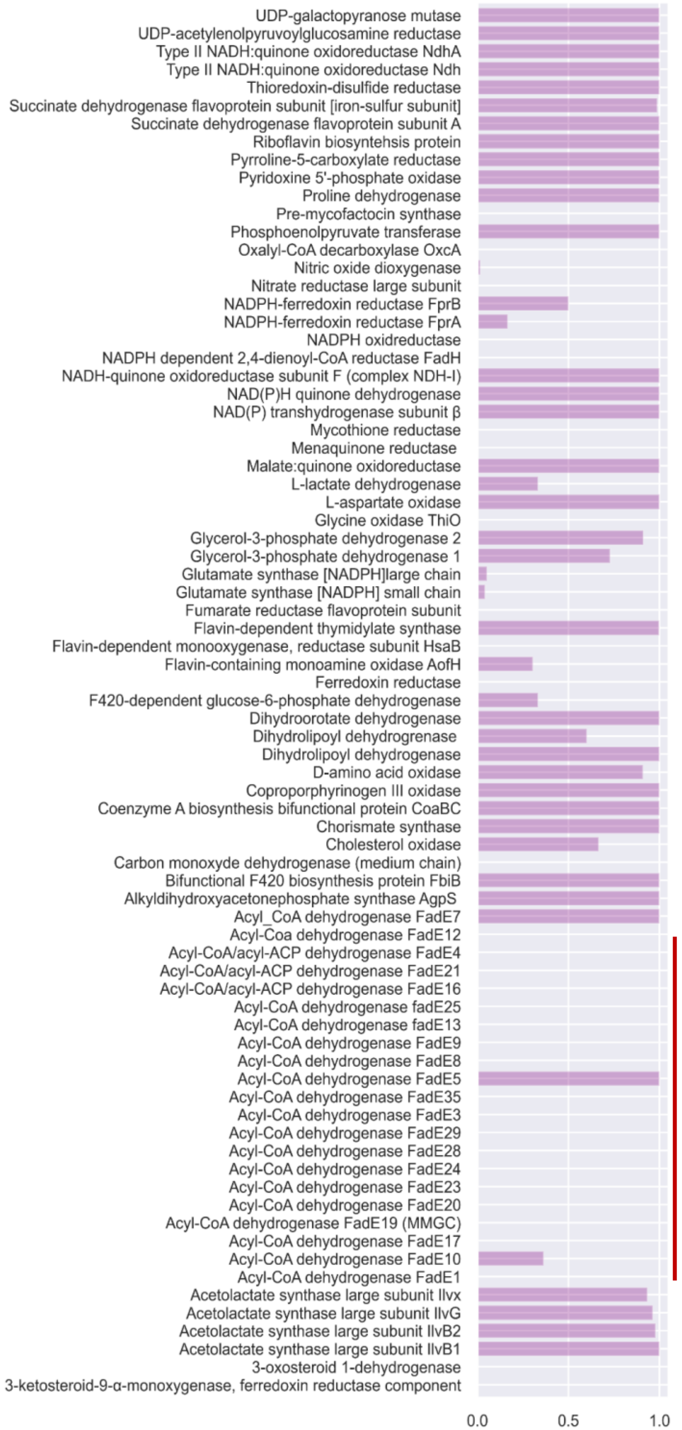
Precision for metabolic pathway prediction



Precision for catalytic activity prediction







Indentified Functions in *M. tuberculosis* Flavoenzymes

Pharmaceutical target present in 29 species of Mycobacterium and not present in the 8 mammal species revised including homo sapiens (Montesa et al., 2023 TFG)

Query code	Description	Pathway	EC
CCP42857.1	F420-dependent hydroxymycolic acid dehydrogenase	Mycolic acid biosynthesis	1.1.98.-
CCP43226.1	Long-chain-alcohol oxidase		1.1.3.20
CCP43640.1	NAD(P)/FAD-dependent oxidoreductase		1.14.13.-
CCP44538.1	L-gulonolactone oxidase	L-ascorbate biosynthesis	1.1.3.8
CCP45062.1	D-lactate dehydrogenase		1.1.99.6

Possible cholesterol side chain degrader that would lead to a potential drug target, an acute immune response of TB host is lipid metabolism gene up-regulation, mycobacteria use steroids as a primary source for energy and thus, survival (Wilburn *et al.*, 2018; Wipperman *et al.*, 2014).

Query code	Description	Pathways	EC
CCP42879.1	Acyl-CoA dehydrogenase FadE2	Lipid metabolism; fatty acid beta-oxidation	1.3.8.7
CCP46093.1	Acyl-CoA dehydrogenase fadE25 Short/branched chain specific acyl-CoA dehydrogenase	Lipid metabolism; fatty acid beta-oxidation / isoleucine, leucine and valine degradation	1.3.8.5
CCP45916.1	NADPH-ferredoxin reductase FprA	Cholesterol metabolism	1.18.1.2

Useful for ethanol synthesis (Tian et al., 2017).

Query code	Description	Pathway	EC
CCP43601.1	pyruvate decarboxylase	Carbohydrate metabolism; pyruvate metabolism	4.1.1.1
CCP44035.1	Oxigen dependent choline dehydrogenase	Amine and polyamine biosynthesis; betaine biosynthesis via choline pathway; betaine aldehyde from choline (cytochrome c reductase route): step 1/1	1.1.99.1
CCP45032.1	Alkylglycerone-phosphate synthase	Glycerolipid metabolism; ether lipid biosynthesis	2.5.1.26

Orthologs found with an alignment below the E-value threshold (<1E-30) are considered for function prediction.

Lowest E-value, higher identity percentage together with species and alignment coverage and posible annotation experimental assertion are considered to select an ortholog closer function.

Indentified Functions in *M. tuberculosis* Flavoenzymes

Query code	Description	Pathway	EC
CCP42887.1	D-2-hydroxyglutarate dehydrogenase		1.1.99.39
CCP44704.1	FAD-binding oxidoreductase		1.-.-.-
CCP46393.1	Flavin-dependent monooxygenase, oxygenase subunit HsaA	Steroid biosynthesis	1.14.14.12*
CCP43539.1	LLM class F420-dependent oxidoreductase		1.-.-.-
CCP44118.1	LLM class F420-dependent oxidoreductase		1.14.-.-
CCP45888.1	LLM class F420-dependent oxidoreductase		1.-.-.-
CCP45863.1	NAD(P)H-dependent oxidoreductase		1.-.-.-
CCP46180.1	NADPH dehydrogenase		1.6.99.1

Pharmaceutical target present in 29 species of Mycobacterium and not present in the 8 mammal species revised including homo sapiens (Montesa et al., 2023 TFG)

Indentified Functions in *M. tuberculosis* Flavoenzymes

Query code	Description	Pathway	EC
CCP43131.1	Acyl_CoA dehydrogenase FadE7 glutaryl-CoA dehydrogenase (ETF)	Amino-acid metabolism; lysine and tryptophan degradation	1.3.8.6
CCP42856.1	Acyl-CoA dehydrogenase FadE1 (R)-benzylsuccinyl-CoA dehydrogenase	Xenobiotic degradation; toluene degradation	1.3.8.3
CCP43621.1	Acyl-CoA dehydrogenase FadE10	Lipid metabolism; fatty acid beta-oxidation	1.3.8.7
CCP43721.1	Acyl-Coa dehydrogenase FadE12 Isovaleryl-CoA dehydrogenase	Amino-acid degradation; L-leucine degradation; (S)-3-hydroxy-3-methylglutaryl-CoA from 3-isovaleryl-CoA: step 1/3	1.3.8.4
CCP43724.1	Acyl-CoA dehydrogenase fadE13		1.3.8.7
CCP44701.1	Acyl-CoA dehydrogenase FadE17		1.3.99.3
CCP45294.1	Acyl-CoA dehydrogenase FadE19 (MMGC) short-chain 2-methylacyl-CoA dehydrogenase	Amino-acid degradation; L-isoleucine degradation/ Lipid metabolism; fatty acid beta-oxidation	1.3.8.5
CCP42879.1	Acyl-CoA dehydrogenase FadE2		1.3.8.7
CCP45522.1	Acyl-CoA dehydrogenase FadE20 Long-chain specific acyl-CoA dehydrogenase	Lipid metabolism; mitochondrial fatty acid beta-oxidation	1.3.8.8
CCP45951.1	Acyl-CoA dehydrogenase FadE23 short-chain 2-methylacyl-CoA dehydrogenase	Amino-acid degradation; L-isoleucine degradation/ Lipid metabolism; fatty acid beta-oxidation	1.3.8.5
CCP45950.1	Acyl-CoA dehydrogenase FadE24		1.3.99.3
CCP46093.1	Acyl-CoA dehydrogenase fadE25		1.3.99.3
CCP42943.1	Acyl-CoA dehydrogenase FadE3 Isovaleryl-CoA dehydrogenase	Amino-acid degradation; L-leucine degradation; (S)-3-hydroxy-3-methylglutaryl-CoA from 3-isovaleryl-CoA: step 1/3	1.3.8.4
CCP46626.1	Acyl-CoA dehydrogenase FadE35		1.3.99.3
CCP43415.1	Acyl-CoA dehydrogenase FadE8		1.3.99.3
CCP43498.1	Acyl-CoA dehydrogenase FadE9 short-chain 2-methylacyl-CoA dehydrogenase	Amino-acid degradation; L-isoleucine degradation/ Lipid metabolism; fatty acid beta-oxidation	1.3.8.5
CCP44444.1	Acyl-CoA/acyl-ACP dehydrogenase FadE16		1.3.99.3
CCP45588.1	Acyl-CoA/acyl-ACP dehydrogenase FadE21 Isovaleryl-CoA dehydrogenase	Amino-acid degradation; L-leucine degradation; (S)-3-hydroxy-3-methylglutaryl-CoA from 3-isovaleryl-CoA: step 1/3	1.3.8.4
CCP42959.1	Acyl-CoA/acyl-ACP dehydrogenase FadE4		1.3.8.7
CCP44635.1	Ferredoxin reductase	Aromatic compound metabolism	1.18.1.2/1.18.1.3
CCP45029.1	Glycerol-3-phosphate dehydrogenase 1	Polyol metabolism; glycerol degradation via glycerol kinase pathway; glycerone phosphate from sn-glycerol 3-phosphate (aerobic route): step 1/1	1.1.5.3
CCP46121.1	Glycerol-3-phosphate dehydrogenase 2	Polyol metabolism; glycerol degradation	1.1.5.3
CCP43146.1	Glycine oxidase ThiO		1.4.3.19
CCP45916.1	NADPH-ferredoxin reductase FprA		1.18.1.2
CCP43115.1	Nitric oxide dioxygenase		1.14.12.17
CCP43944.1	Proline dehydrogenase	Amino-acid degradation; L-proline degradation into L-glutamate; L-glutamate from L-proline: step 1/2	1.5.5.2
CCP42977.1	Succinate dehydrogenase flavoprotein subunit [iron-sulfur subunit]	Carbohydrate metabolism; tricarboxylic acid cycle; fumarate from succinate (bacterial route): step 1/1	1.3.5.1 / 1.3.99.1

Useful for toluene degradation since anaerobic oxidation of petroleum hydrocarbons can be coupled to the reduction of metals, this will accelerate the removal of pollutants(Tremblay and Zhang, 2017)

Indentified Functions in *M. tuberculosis* Flavoenzymes

Query code	Description	Pathway	EC
CCP42785.1	Carbohydrate oxidase	Energy metabolism	(1.1.3.5)*
CCP43180.1	NAD(P)/FAD-dependent oxidoreductase		1.-.-.-
CCP43226.1	Long-chain-alcohol oxidase	Energy metabolism	1.1.3.20/1.1.3.13
CCP43303.1	NAD(P)/FAD-dependent oxidoreductase		1.-.-.- (1.13.12.-)*
CCP43313.1	Oxidoreductase		1.-.-.-
CCP43441.1	Mycofactocin system GMC family oxidoreductase		1.-.-.-
CCP44016.1	flavin-dependent monooxygenase (7-chlorotetracycline and tetracycline oxidation)		1.-.-.-
CCP44152.1	Cyclopentanone 1,2-monooxygenase	Alcohol metabolism; cyclopentanol degradation; 5-valerolactone from cyclopentanol: step 2/2	1.14.13.16
CCP44492.1	FAD-binding oxidoreductase (R)-6-hydroxynicotine oxidase	Alkaloid degradation; nicotine degradation; 6-hydroxypseudooxynicotine from nicotine (R-isomer route): step 2/2	1.5.3.6
CCP44517.1	6-methylpretetramide 4-monooxygenase	Antibiotic biosynthesis; oxytetracycline biosynthesis	1.14.13.232
CCP45575.1	Carnitine monooxygenase reductase subunit	Amine and polyamine metabolism; carnitine metabolism	1.14.13.239
CCP45858.1	L-ornithine N(5)-monooxygenase	Siderophore biosynthesis	1.14.13.196
CCP46342.1	LLM class F420-dependent oxidoreductase		1.14.-.-
CCP46658.1	NAD(P)/FAD-dependent oxidoreductase		1.-.-.-

Pharmaceutical target present in 29 species of Mycobacterium and not present in the 8 mammal species revised including homo sapiens (Montesa et al., 2023 TFG)

Conclusions

- A novel computational method has been developed to effectively search for catalytic activity identification in flavoenzymes within the *Mycobacterium tuberculosis* flavoproteome.

- The method predicted 33 flavoprotein new complete functions leading to more accurate and metabolically contextualized descriptions, spanning diverse pathways and leading to 2 potential applications in biocatalysis and 3 in drug targeting.

- The method achieved approximately 60% agreement with previous annotations and successfully analysed 184 proteins in around 50 minutes.

- Out of the 133 unknown proteins, 54 were found to have similarity with fully annotated flavoproteins from all available species comprehending all kingdoms. 33 queries were found to have significance in high scoring segment pairs alignments.

- The flavoenzymes found are involved in amino-acid metabolism, lipid metabolism, xenobiotic and aromatic compound degradation, antibiotic biosynthesis biosynthesis, energy metabolism and amine synthesis.