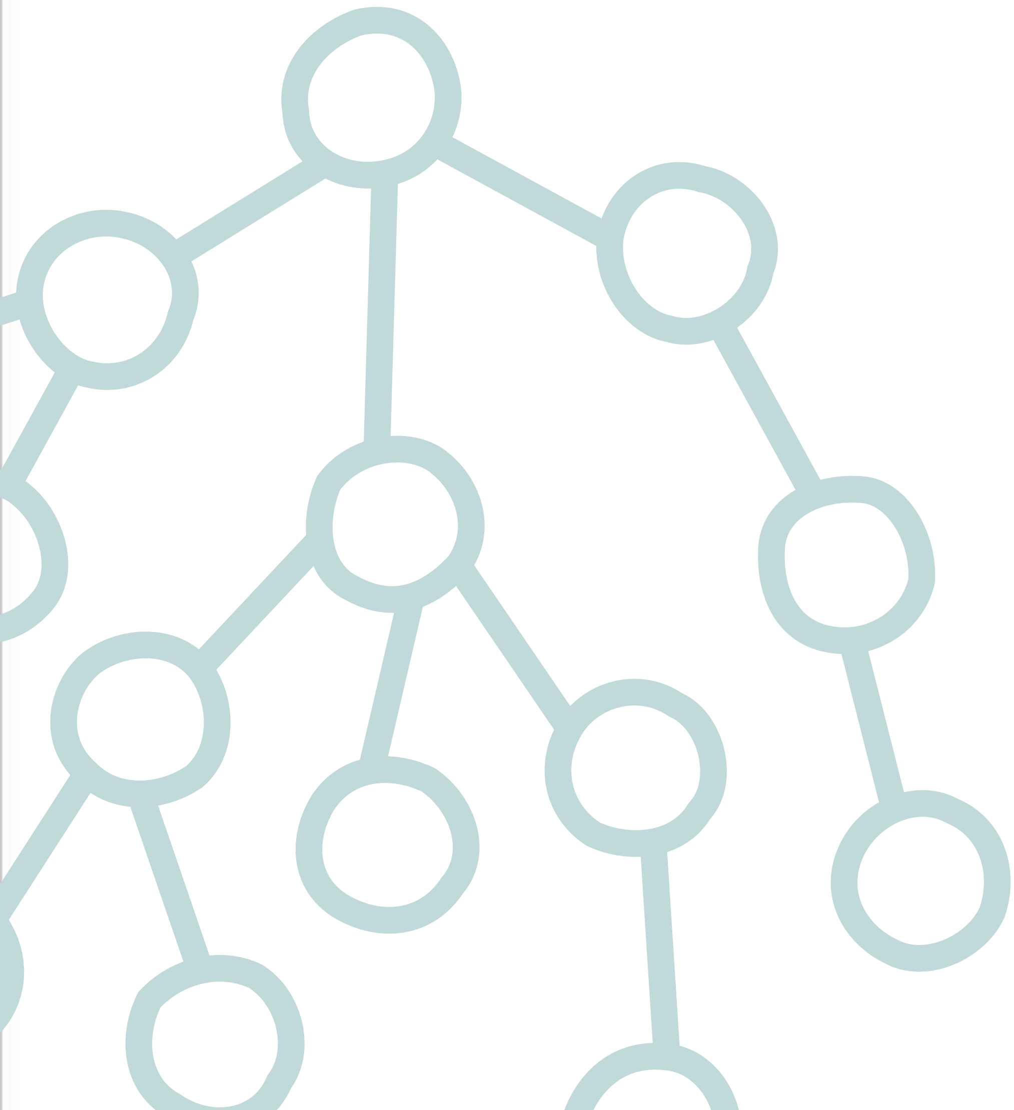
enviPath | THE ENVIRONMENTAL CONTAMINANT BIOTRANSFORMATION PATHWAY RESOURCE

Eawag
Swiss Federal Institute of Aquatic Science and Technology

eawag aquatic research

Universität Zürich UZH



Linking enviPath to systems biology and sequencing data

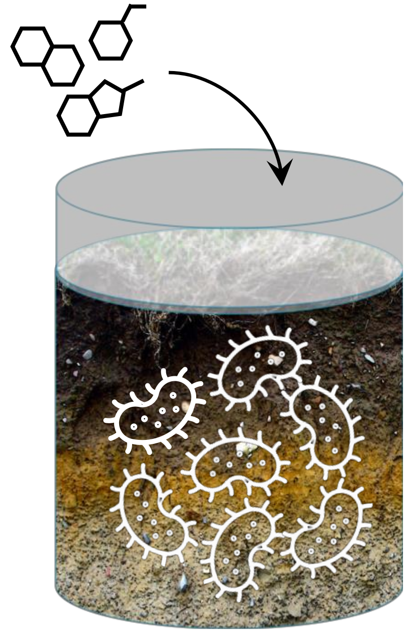
Jasmin Hafner, Albert Anguera Sempere, Kathrin Fenner

11 May 2025

Overview

1. Why **systems biology** and **sequencing** data?
2. Linking compounds and reactions to **external databases**
3. Briding the **Gene-Protein-Reaction** (GPR) knowledge gap
4. Outlook: Connecting to **metagenomes**

Why systems biology and sequencing data?

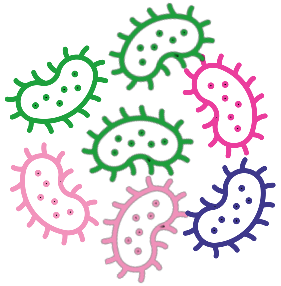


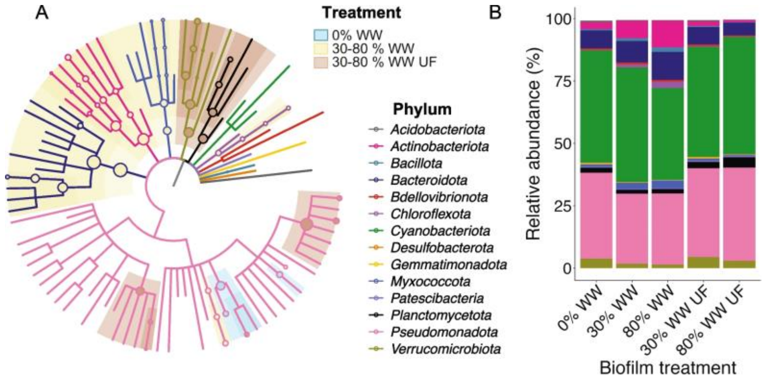
Microbial communities

WHO?

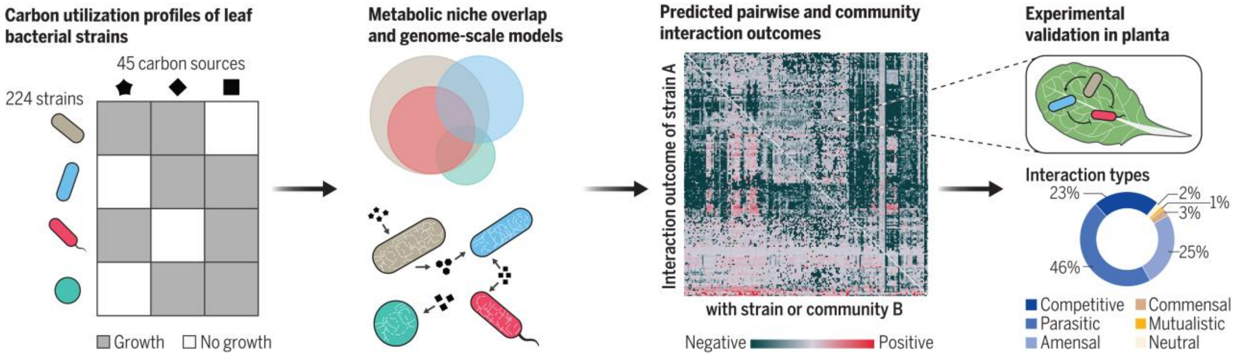
WHAT?

Taxonomic composition (sequencing data)



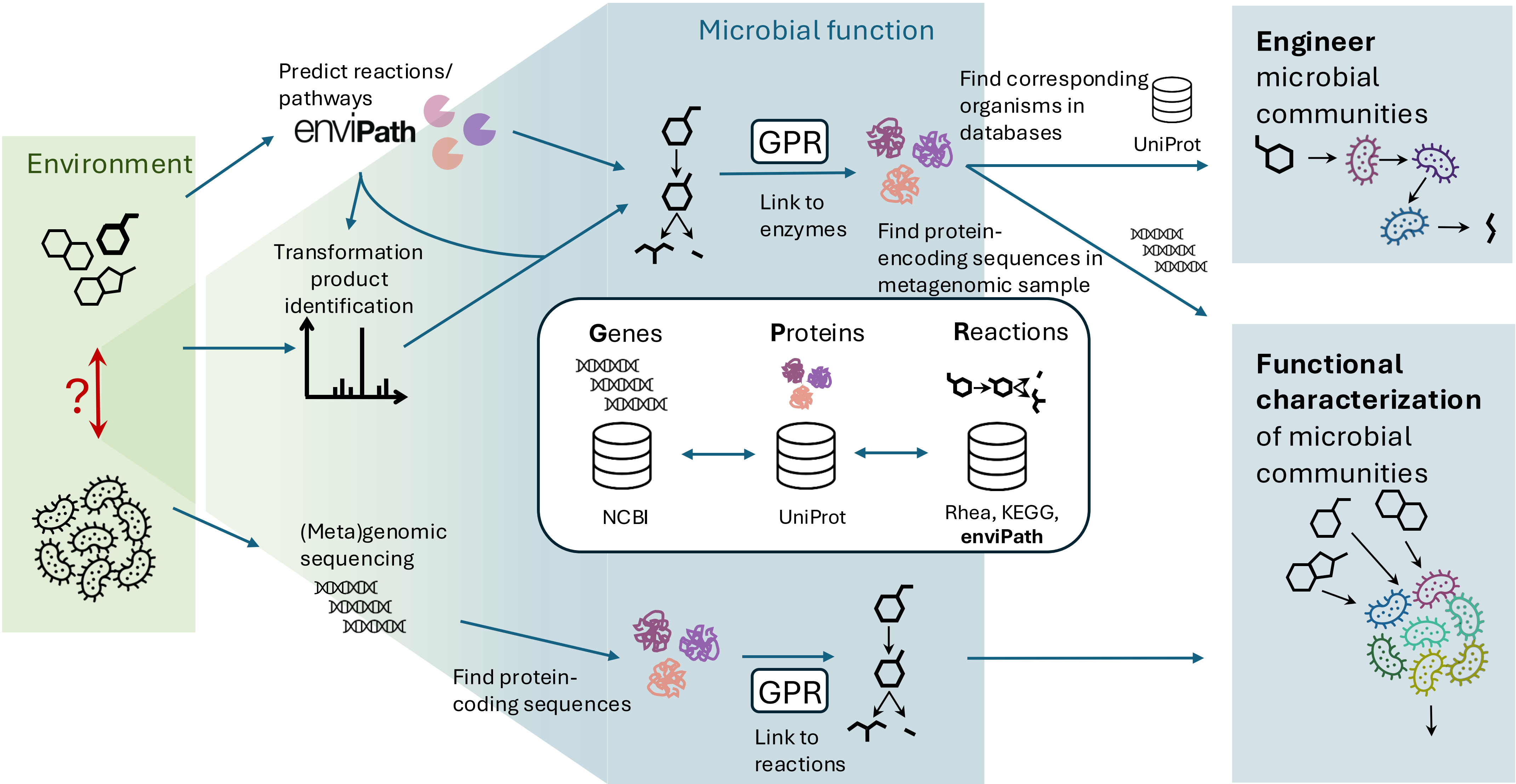


Enzymatic composition, community function
(metabolism / community interactions)


Carbon utilization profiles of leaf bacterial strains
45 carbon sources
224 strains
Growth
No growth
Metabolic niche overlap and genome-scale models
Predicted pairwise and community interaction outcomes
Interaction outcome of strain A
with strain or community B
Negative
Positive
Experimental validation in planta
Interaction types
23%
2%
1%
3%
25%
46%
Competitive
Commensal
Parasitic
Mutualistic
Amensal
Neutral

Figures: Attrah et al., *Microbiome*, 2024., Schäfer et al., *Science*, 2023

Different approaches – same challenge



Objectives

Environment

?

Predict reactions/ pathways

enviPath

Transformation product identification

(Meta)genomic sequencing

Microbial function

GPR

Link to enzymes

Find corresponding organisms in databases

UniProt

Find protein-encoding sequences in metagenomic sample

Genes

Proteins

Reactions

NCBI

UniProt

Rhea, KEGG, **enviPath**

Find protein-coding sequences

GPR

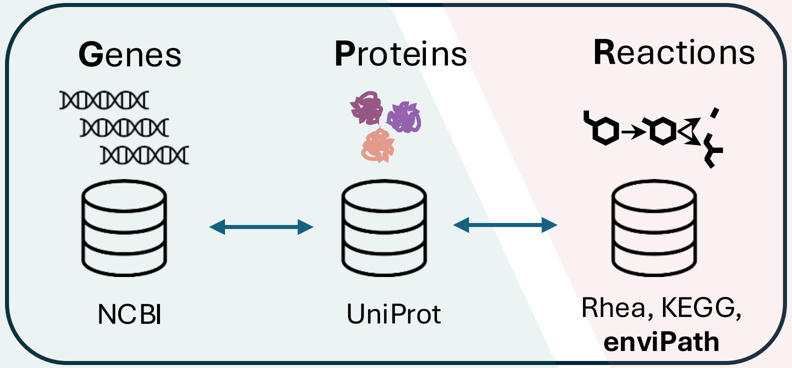
Link to reactions

Engineer microbial communities

Functional characterization of microbial communities

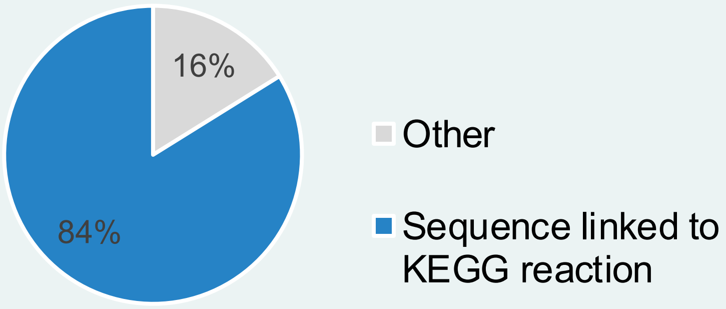
Challenge: Link function (reaction) **to sequence** (gene, protein)

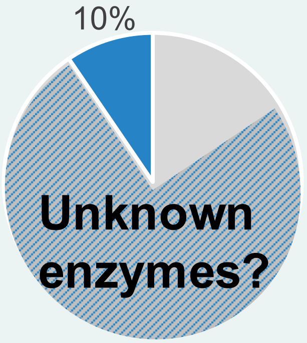
- **Orphan gene / protein sequences** (no function associated)
- Promiscuous enzyme activity



SwissProt (curated):
~ 500K proteins

Trembl:
> 200 M proteins





Orphan reactions
(no gene associated)

Rhea: 4,760 out of 17,614 reactions
→ 27% orphan

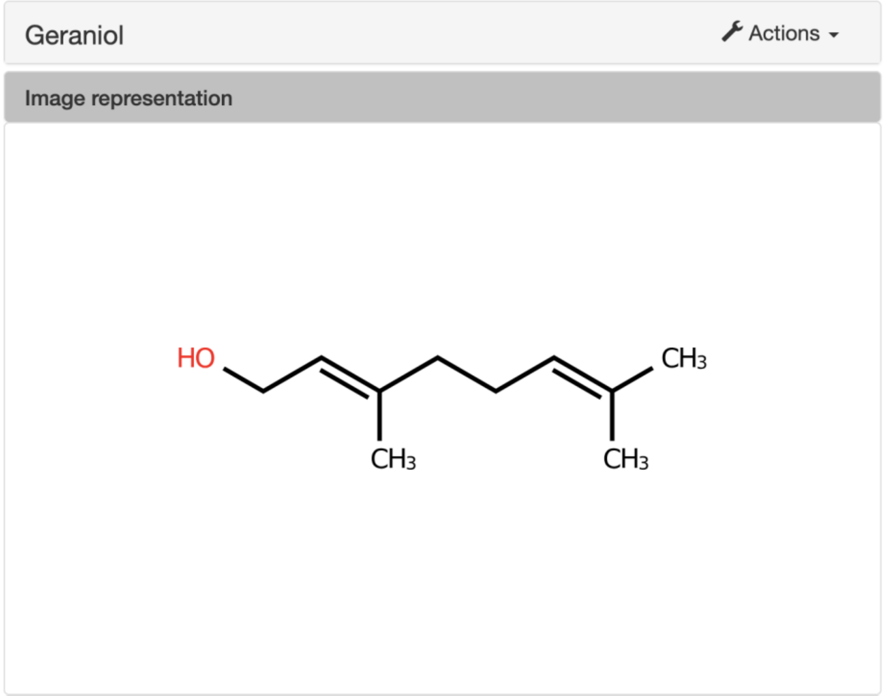
enviPath: ?

Rhea and UniProt consulted in April 2025

Overview

1. Why **systems biology** and **sequencing** data?
2. Linking compounds and reactions to **external databases**
3. Briding the **Gene-Protein-Reaction** (GPR) knowledge gap
4. Outlook: Connecting to **metagenomes**

Linking compounds to external databases



Geraniol

Actions

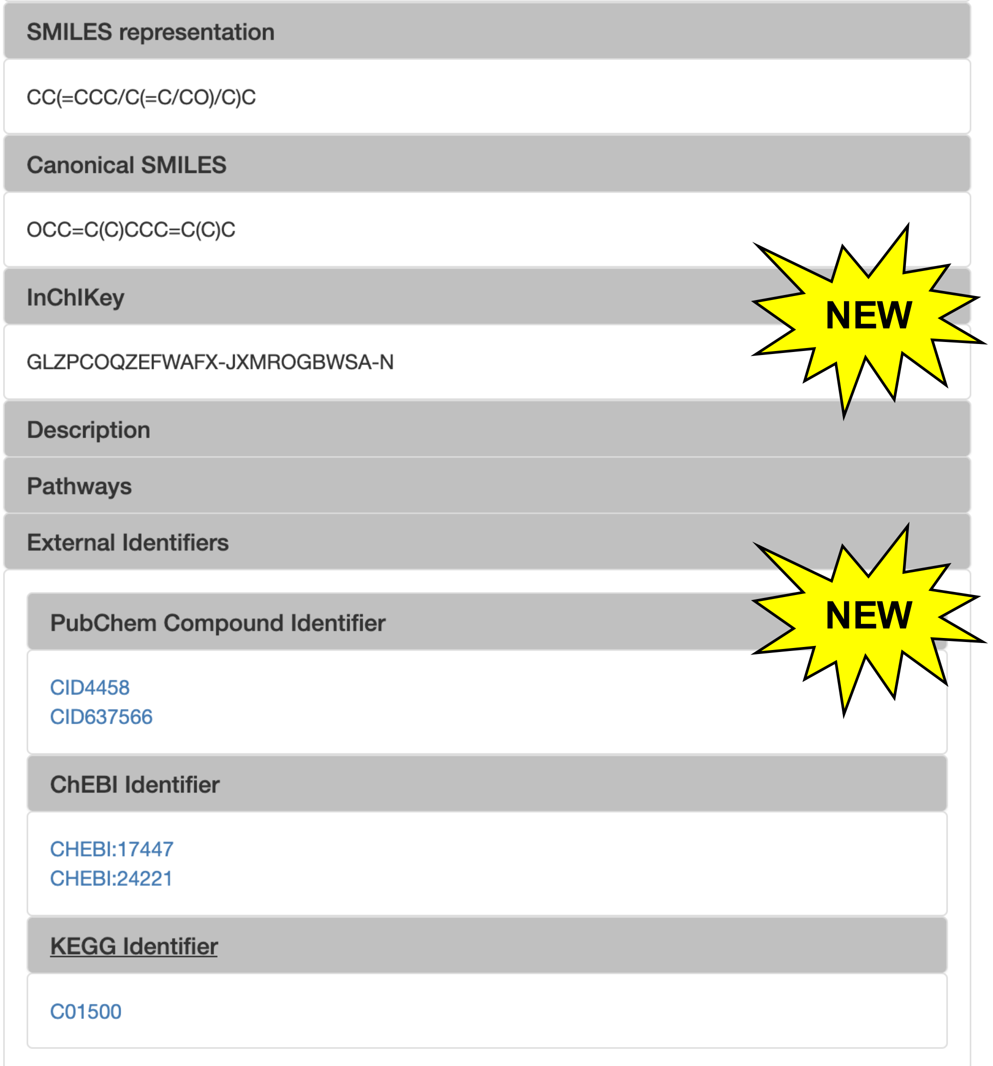
Image representation

Database look-up
via SMILES and
InChiKey

PubChem

ChEBI

KEGG Kyoto Encyclopedia of Genes and Genomes



SMILES representation

CC(=CCC/C(=C/CO)/C)C

Canonical SMILES

OCC=C(C)CCC=C(C)C

InChIKey

GLZPCOQZEFWAFX-JXMROGBWSA-N

Description

Pathways

External Identifiers

PubChem Compound Identifier

CID4458
CID637566

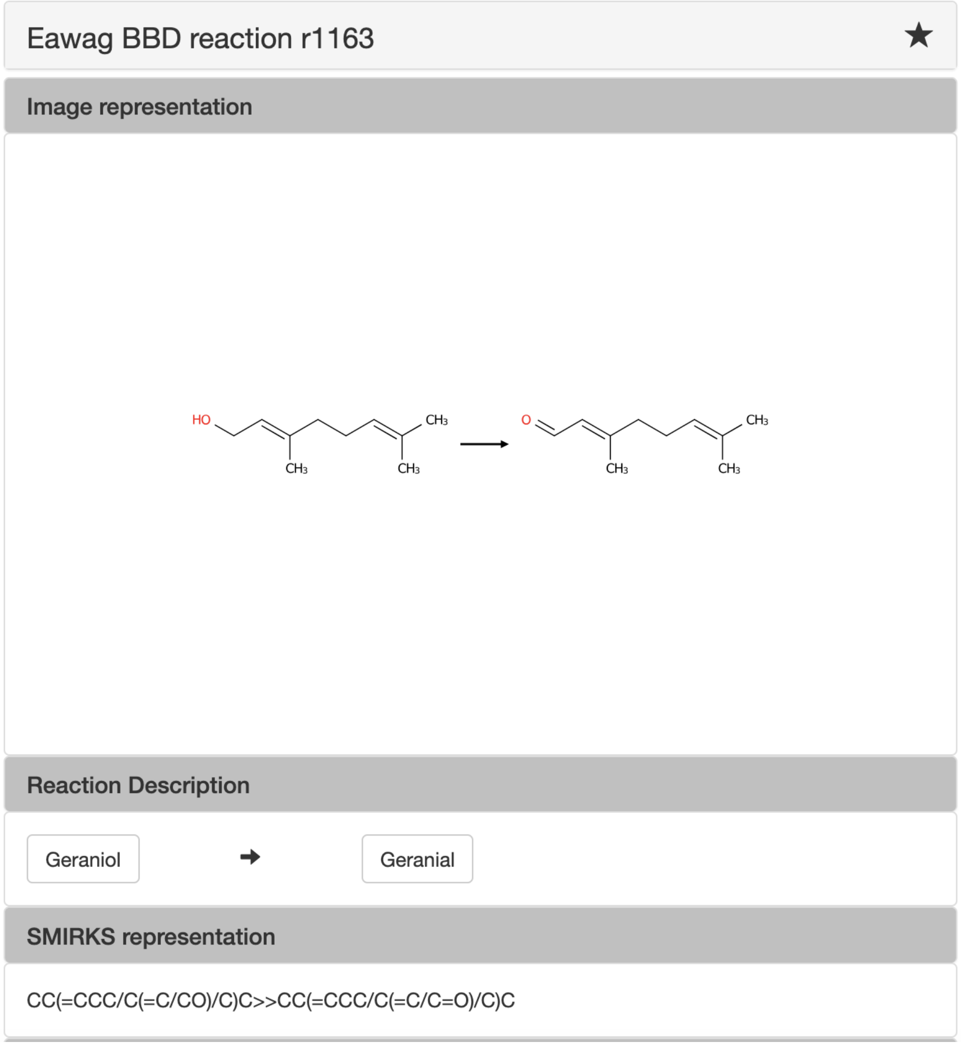
ChEBI Identifier

CHEBI:17447
CHEBI:24221

KEGG Identifier

C01500

Linking reactions to external databases



Eawag BBD reaction r1163 ★

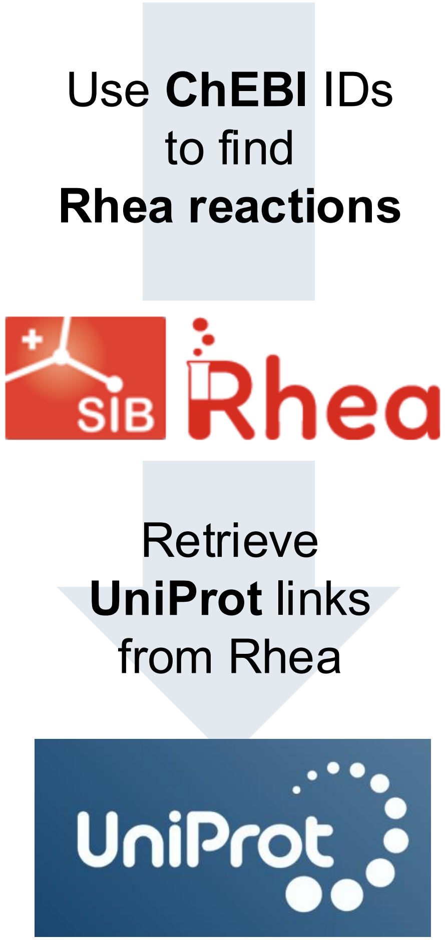
Image representation

Reaction Description

Geraniol ➜ Geranial

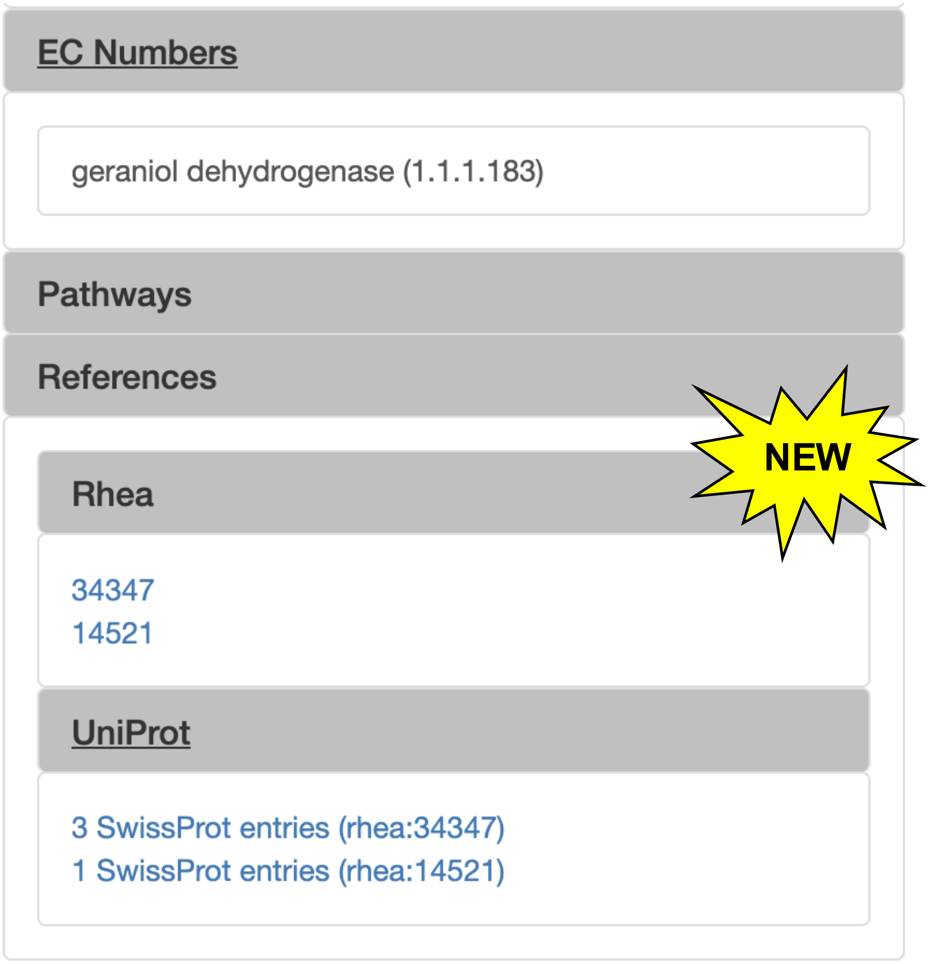
SMIRKS representation

CC(=CCC/C(=C/CO)/C)C>>CC(=CCC/C(=C/C=O)/C)C



Use **ChEBI** IDs to find **Rhea reactions**

Retrieve **UniProt** links from Rhea



EC Numbers

geraniol dehydrogenase (1.1.1.183)

Pathways

References

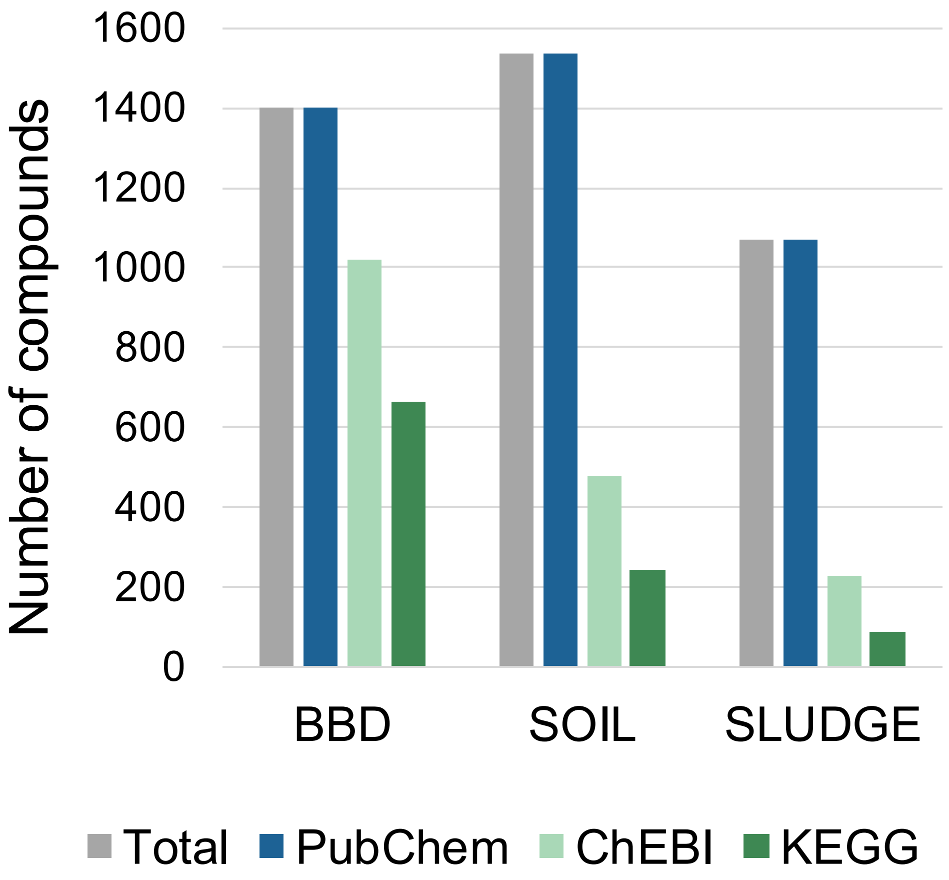
Rhea

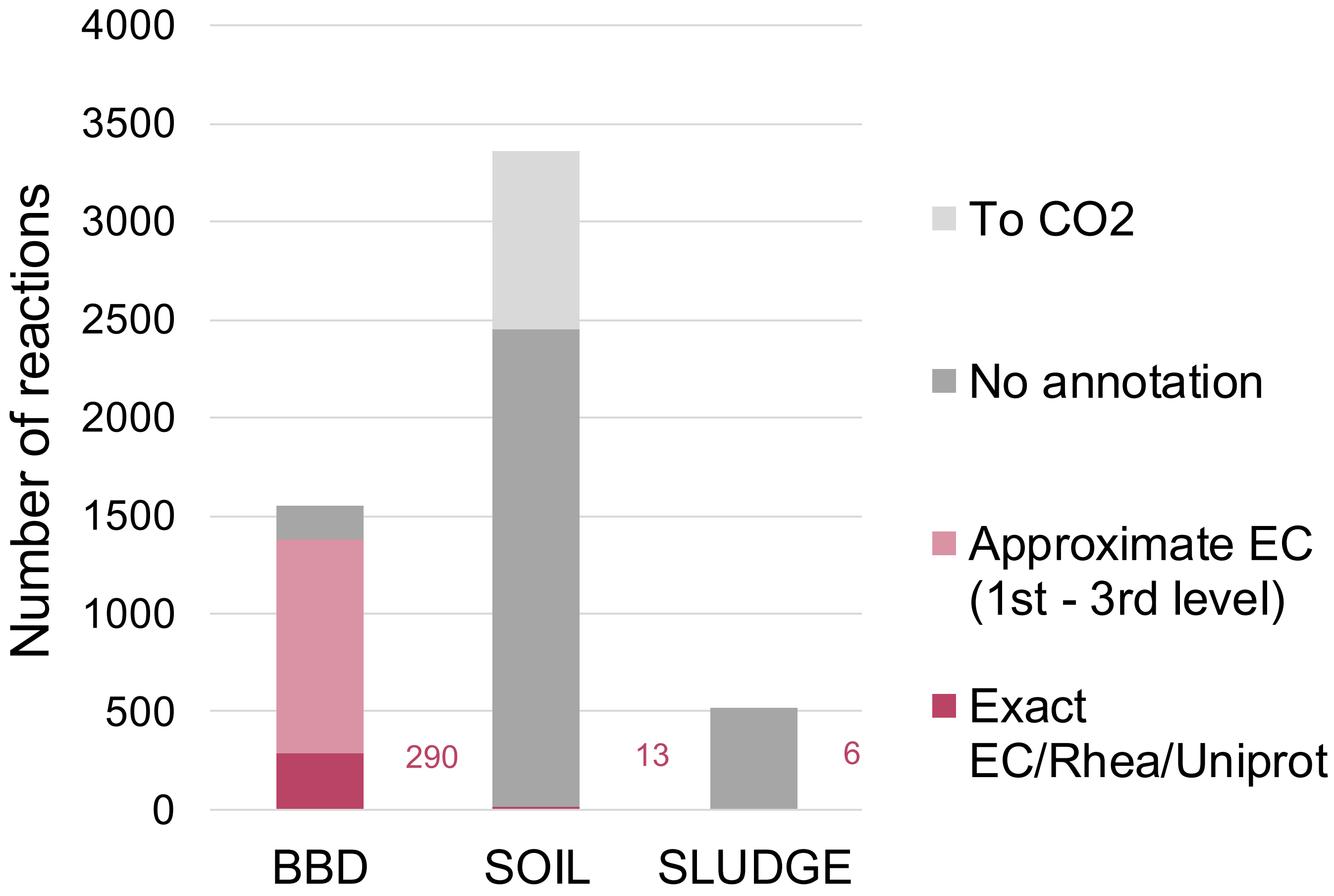
34347
14521

UniProt

3 SwissProt entries (rhea:34347)
1 SwissProt entries (rhea:14521)

How many compounds and reactions could be linked?



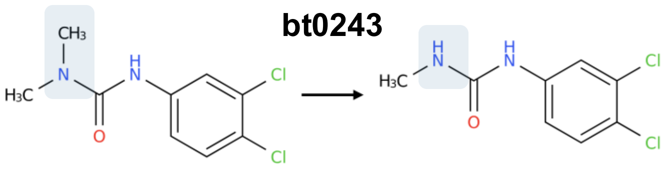


In BBD, only **15% of reactions** (223) are linked to **UniProt protein sequences – 85% are orphan!**

Why most reactions don't have known enzymes?

Example: EAWAG-SOIL (excluded: reactions leading to CO_2)

Known protein sequence

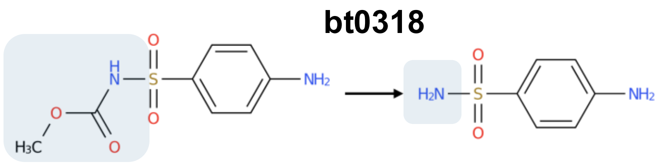


Diuron DCPMU

1% of SOIL reactions

Linked to Rhea: 13
EC numbers: 8

Mechanism known (bt rule)

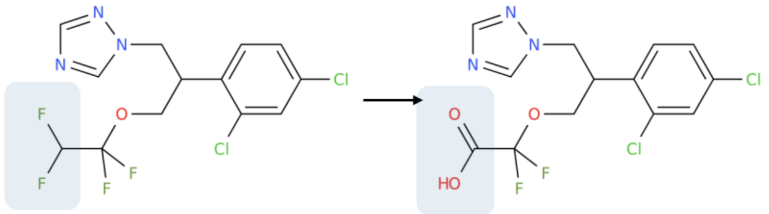


Asulam Sulfanilamide

50% of SOIL reactions

Potential to find sequences!

Mechanism unknown (no rule)



Tetraconazole M14360-DFA

49% of SOIL reactions

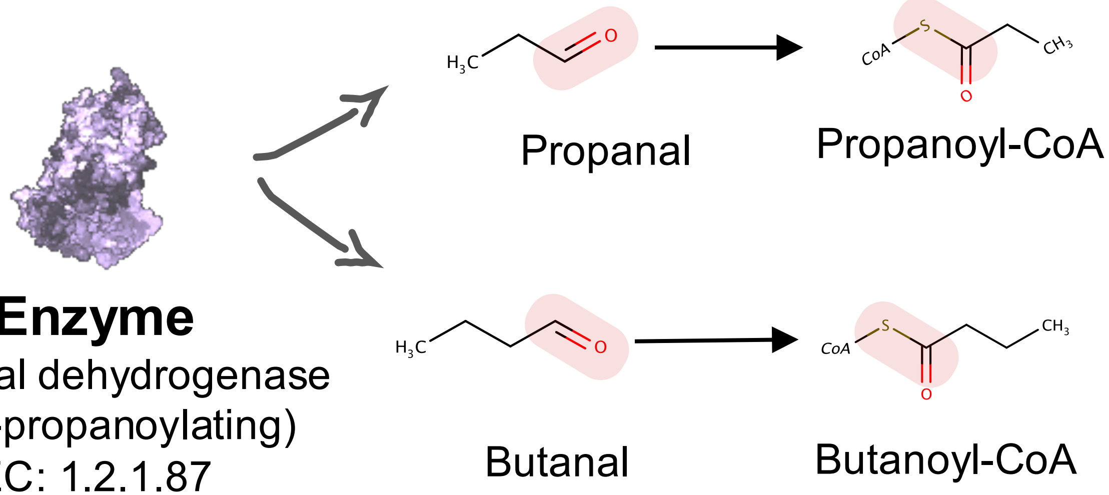
→ **Xenobiotic** metabolism not well covered by metabolic databases
→ **Co-metabolic** processes not well understood
→ Reported pathways in biodegradation studies **based on observed TPs** only

Overview

1. Why **systems biology** and **sequencing** data?
2. Linking compounds and reactions to **external databases**
3. Briding the **Gene-Protein-Reaction** (GPR) knowledge gap
4. Outlook: Connecting to **metagenomes**

Can we find potential protein sequences for orphan biotransformations?

Assumption: Similar reactions are catalyzed by similar/same enzymes

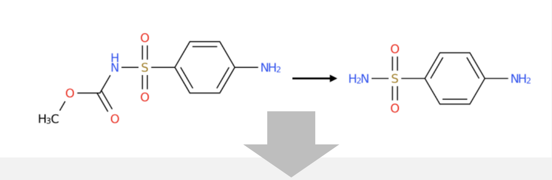


Enzyme
Propanal dehydrogenase
(CoA-propanoylating)
EC: 1.2.1.87

✓ Works well for **promiscuous** enzyme **activities**

✗ Difficult to find completely **new enzymes**

Orphan reaction



Create fingerprint

1001101001…

Compare to database of reactions linked to protein sequences

SIB Rhea   KEGG

Find most similar reaction

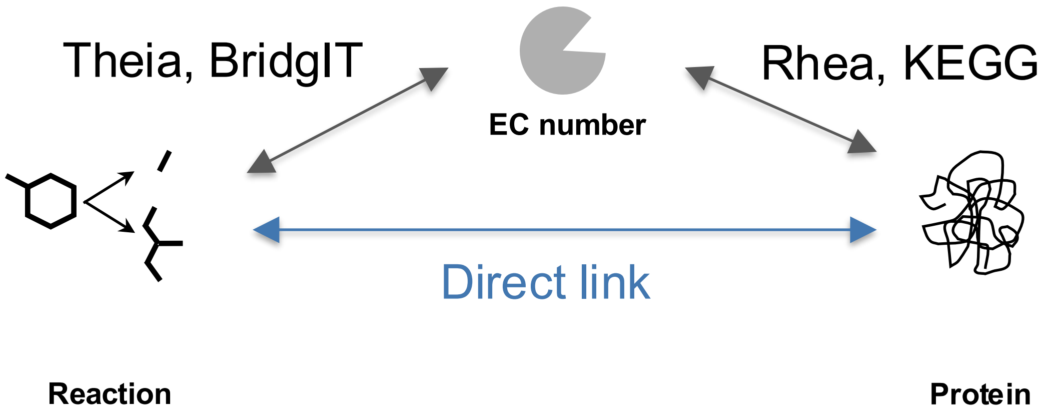
Potentially catalyzing enzyme

UniProt

How can we link orphan reactions to enzymes?

Kunyang Zhang

Tools to **predict EC** numbers: Theia, BridgIT

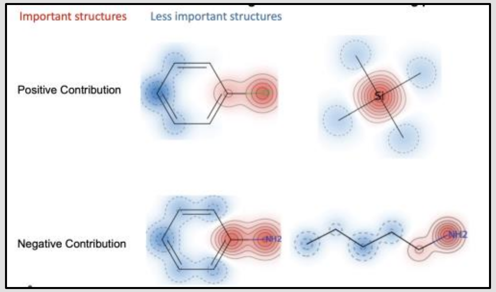


	Theia	BridgIT	***Ideally***
Considers reactive site	no	yes	yes
Need for mass balance	no	yes	no
Open source	yes	no	yes
Bypass EC	no	no	yes
Xenobiotics metabolism	no	no	yes

Pretraining
Chemical reactions

Fine-tuning
Biochemical & biodegradation reactions

AI-driven identifcation of relevant **functional groups**



Similarity metric for reaction comparison

Overview

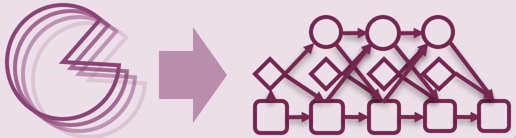
1. Why **systems biology** and **sequencing** data?
2. Linking compounds and reactions to **external databases**
3. Briding the **Gene-Protein-Reaction** (GPR) knowledge gap
4. Outlook: Connecting to **metagenomes**

New and ongoing developments

Repository of **HMM profiles** for **biotransformation** by

Dr. Serina Robinson, Victoria Poltorak

https://github.com/MSM-group/ContaHMM



Link to **metagenomic studies** in enviPath

Scenarios:
Add links to **ENA** and MG**nify** entries

ENA European Nucleotide Archive

MGnify

Bridging **to Systems Biology**

SBML

Export **pathways** in Systems Biology Markup Language

https://envipath-python.readthedocs.io/en/develop/tutorials/download_pathway_SBML.html

Take-home message

- Big **knowledge gaps** in linking enzymatic reactions to protein sequences, *especially* for xenobiotic biotransformations
- **Data curation** efforts needed to fill knowledge gaps
- In the future, **new ML/AI tools** may also help in filling those gaps

Acknowledgements

Prof. Kathrin Fenner

Albert Anguera Sempere

& the Fenner Team

Funding sources

PREMIER
PRIORITISATION AND RISK EVALUATION OF MEDICINES IN THE ENVIRONMENT

erc

European Commission | Horizon 2020 European Union funding for Research & Innovation

swissuniversities