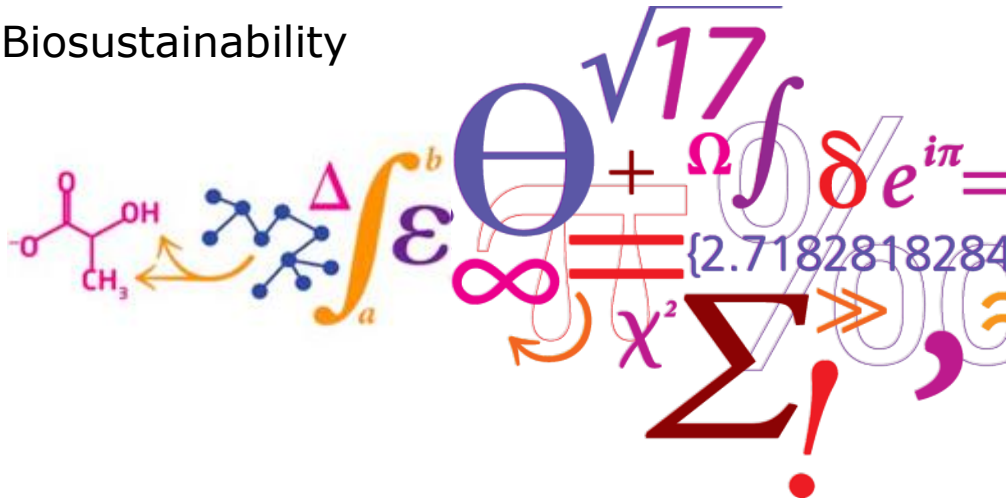
Computational tools for metabolic pathways prediction

Marina Fedorova
Computational Chemistry Group
Novo Nordisk Foundation Center for Biosustainability
Technical University of Denmark



DTU Biosustain
The Novo Nordisk Foundation Center for Biosustainability

Introduction

- The Novo Nordisk Foundation Center for Biosustainability (DTU Biosustain) aims to be a worldwide leader in cell factories design, construction and deployment, and thus be a driver for change towards a biosustainable society.

Center for Biosustainability

Introduction

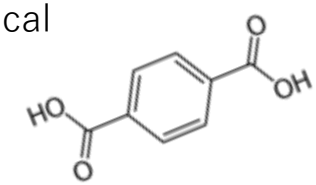
- Computational Biology division:
  - In silico design of cell factories
  - Automation group
  - Lab support (LIMS)
  - Genome scale modelling
  - Integrated software platform

- Computational Chemistry group
  - Metabolic pathways prediction tool – GemPath
  - Genome Scale Model Based
  - Used to be in Matlab, now moved to Python
  - Further extensions

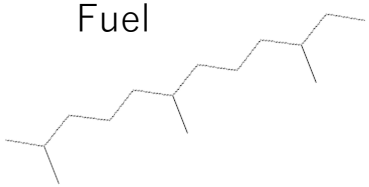
Motivation

Want this:

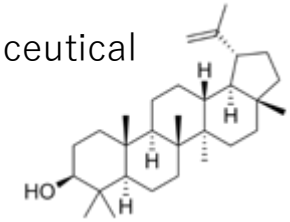
Commodity chemical



Fuel

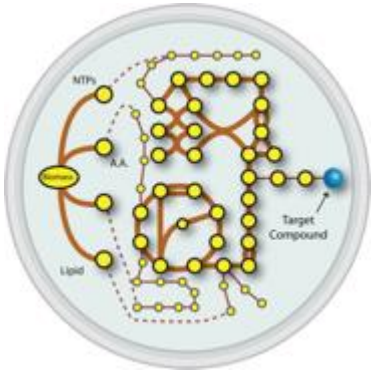


Nutraceutical

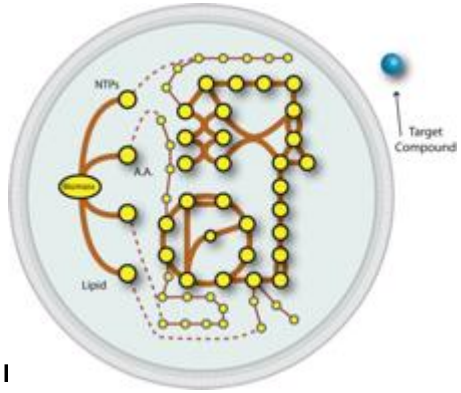


But there is a problem:

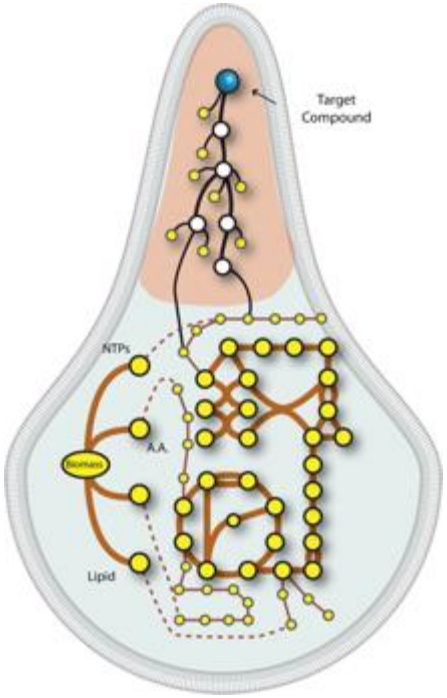
Target Compound has low Yield in the host organim



Target Compound is not in the host organim

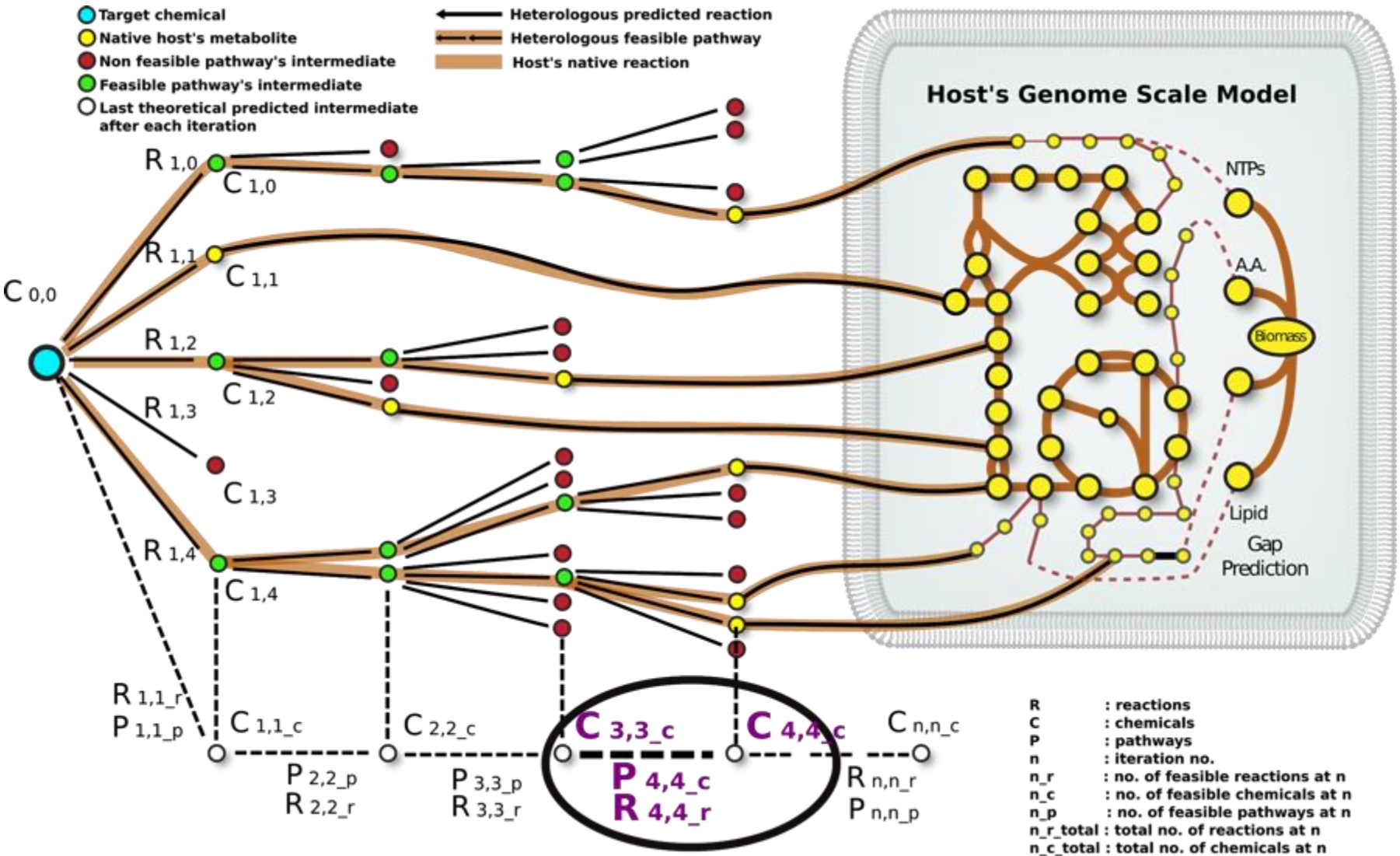


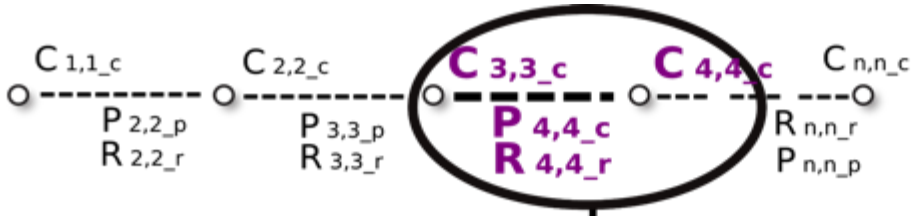
So you need this:



Synthetic Pathway Calculation

GEM-Path


Target chemical
Native host's metabolite
Non feasible pathway's intermediate
Feasible pathway's intermediate
Last theoretical predicted intermediate after each iteration
Heterologous predicted reaction
Heterologous feasible pathway
Host's native reaction
Host's Genome Scale Model
NTPs
A.A.
Biomass
Lipid
Gap Prediction
C 0,0
R 1,0
C 1,0
R 1,1
C 1,1
R 1,2
C 1,2
R 1,3
C 1,3
R 1,4
C 1,4
R 1,1_r
P 1,1_p
C 1,1_c
C 2,2_c
C 3,3_c
C 4,4_c
C n,n_c
P 2,2_p
R 2,2_r
P 3,3_p
R 3,3_r
P 4,4_c
R 4,4_r
R n,n_r
P n,n_p
R : reactions
C : chemicals
P : pathways
n : iteration no.
n_r : no. of feasible reactions at n
n_c : no. of feasible chemicals at n
n_p : no. of feasible pathways at n
n_r_total : total no. of reactions at n
n_c_total : total no. of chemicals at n
n_p_total : total no. of pathways at n



R : reactions
C : chemicals
P : pathways
n : iteration no.
n_r : no. of feasible reactions at n
n_c : no. of feasible chemicals at n
n_p : no. of feasible pathways at n
n_r_total : total no. of reactions at n
n_c_total : total no. of chemicals at n
n_p_total : total no. of pathways at n

GEMPath's individual iteration workflow

C 3,3_c | **C 4,4_c_total** **R 4,4_r_total**

Apply BROs and build reactions

- **443 BROs**
- **Cofactor and 3 digits EC no. assignation**

Additional Inputs
- max. pathway length
- Specify biosynthetic or degradation pathway type of predictions

Calculate reactions and metabolites chemical properties

- **Reactions**
  - reaction promiscuity score (PS)
  - reaction/gene/protein/patent association
  - reaction existence
  - mass balance
  - ΔrG' and ΔrG'°
- **Metabolites**
  - ΔfG'°
  - Toxicity (IC50)
  - Metabolome based Tanimoto Coefficient
  - Concentration

Additional Inputs
- Host metabolome structural information (GEM)
- Databases (BRENDA, KEGG, UNIPROT and LENS)
- Ionic strength, temperature and pH

R 4,4_r

Filter non feasible reactions out

- **Predicted reaction filters**
  - mass balance reactions
  - max reaction PS
  - max ΔrG'
  - BRO output (allow all or only the closest metabolite from the BRO simulation)

Additional Inputs

P 4,4_p_total **P 4,4_c**

Determine reaction metabolite conections and feasible pathways

- **Build heterologous pathways**
- **Theoretical heterologous pathways phenotipyc simulations:** Perform FBA to calculate the max. theoretical yield for all pathways. Pathways with yields greatyer than 0 are define as feasible.

Additional Inputs
- Host's genome scale model
- Media specification (substrate utilization, oxigenation and minerals among others)

C 4,4_c

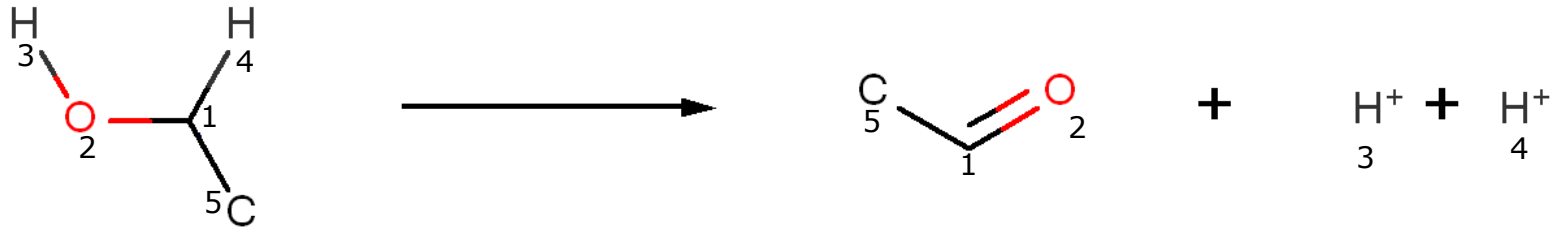
Node selection for next iteration

- **Predicted metabolites filters**
  - max. no. of metabolites per iteration
  - select the host's metabolome closest metabolites
  - allow/discard nodes that lead to feasible pathways
  - allow/discard nodes with the min. carbon content
  - prioriatize nodes generated from known reactions
  - focus the search towards a subset of predefine metabolites

Additional Inputs

Biochemical Reaction Operators (BROs)

SMIRKS: [C:5][C:1]([H:4])([H])[O:2][H:3] >> [C:5][C:1]([H])=[O:2].[H:3].[H:4]



CoSubstrate
fad
nad
nadp
o2
(0.5) o2

EC Number
1.1.2
1.1.1
1.1.1
1.1.3
1.1.3

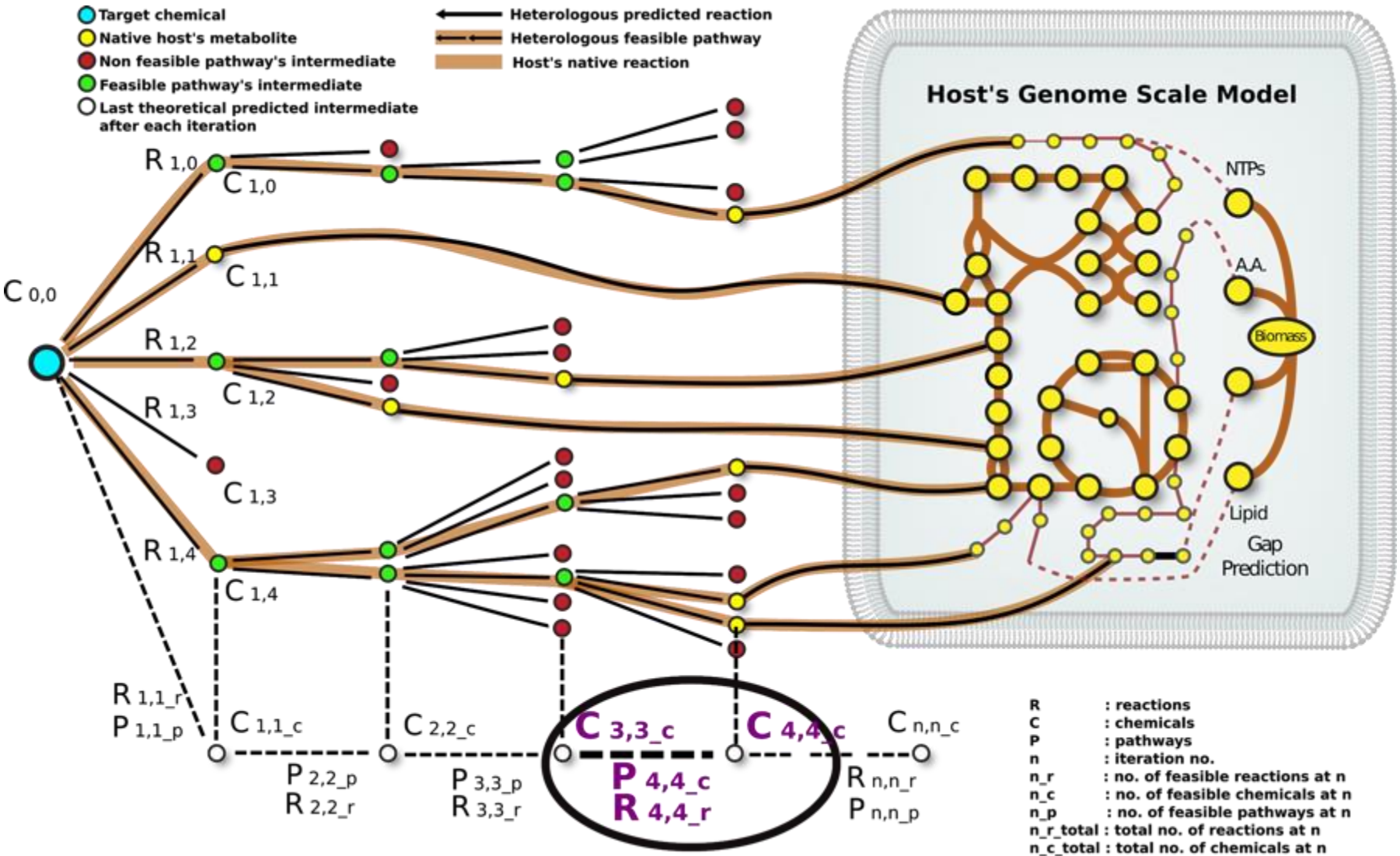
CoProduct
fadh2
nadh + h
nadph + h
o2
(0.5) o2

RDKit uses reaction SMARTS ☹

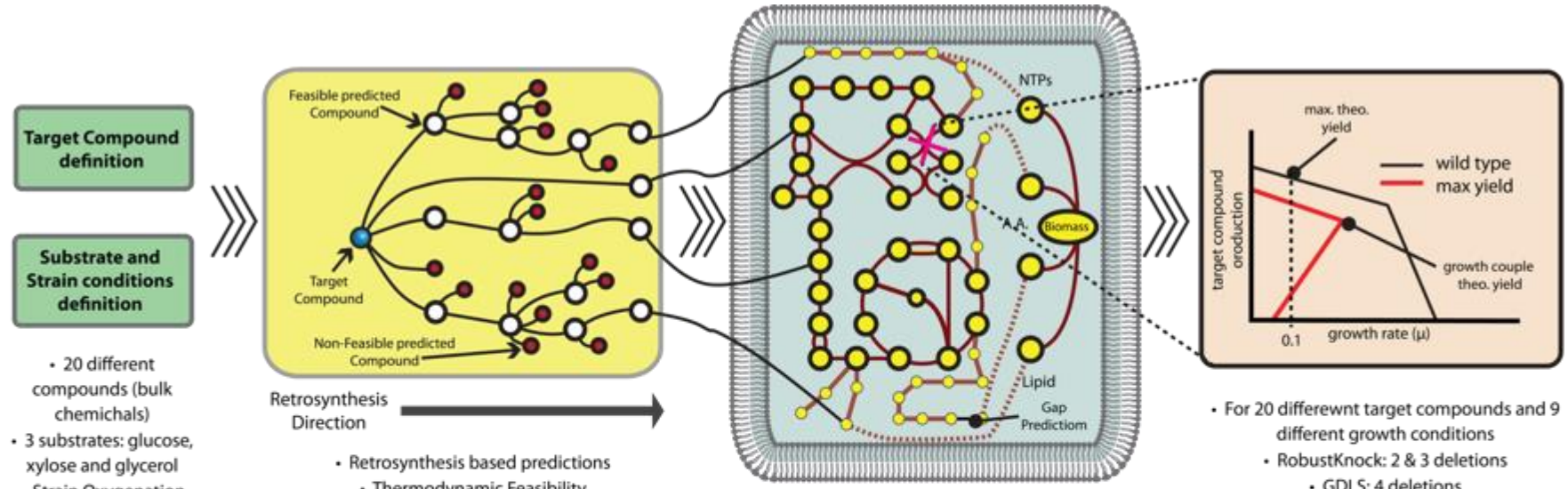
Chemical Structure Comparison

Tanimoto Coefficient & Fingerprints

- Fingerprints generation for the compounds in the metabolome and the target compound.
- Fingerprint generation for the current compound
- Tanimoto coefficient calculation between the current compound and metabolome
- Algorithm moves towards metabolome (looks for higher similarity score, but it is adjustable)

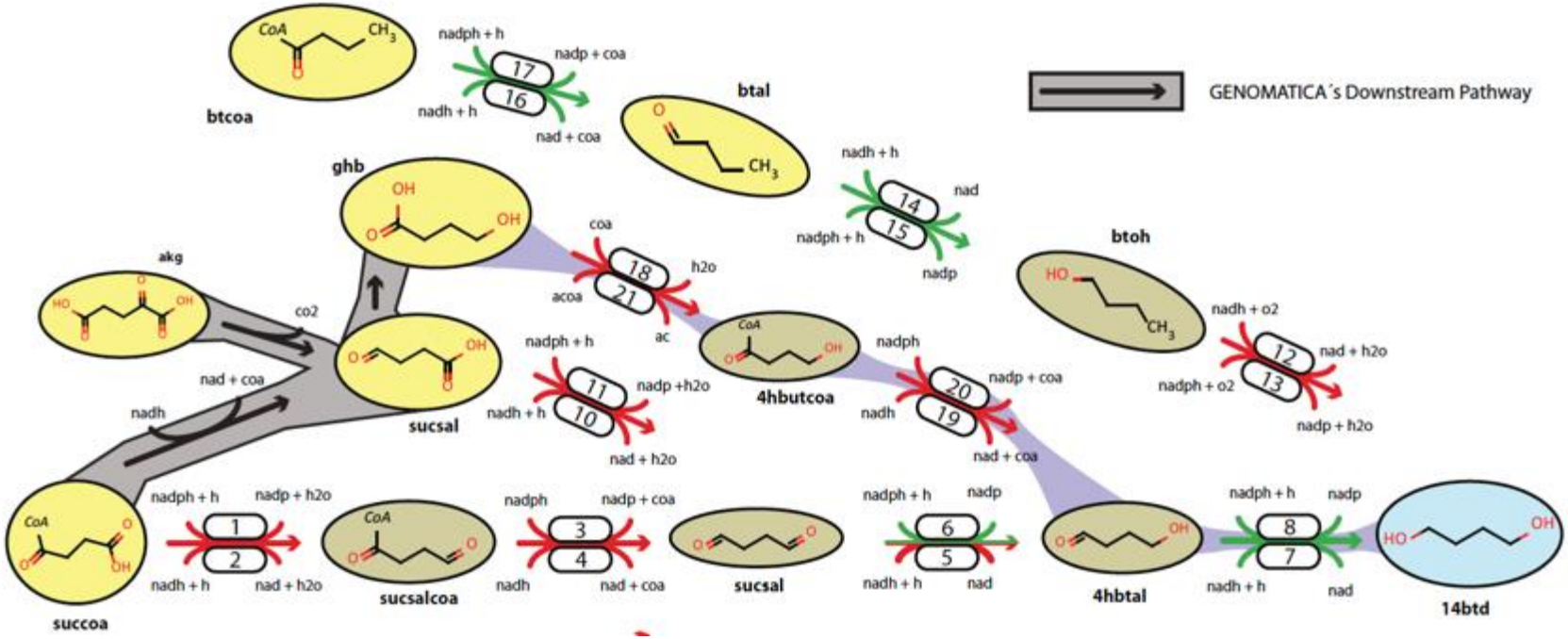

Target chemical
Native host's metabolite
Non feasible pathway's intermediate
Feasible pathway's intermediate
Last theoretical predicted intermediate after each iteration
Heterologous predicted reaction
Heterologous feasible pathway
Host's native reaction
Host's Genome Scale Model
NTPs
A.A.
Biomass
Lipid
Gap Prediction
C 0,0
R 1,0
C 1,0
R 1,1
C 1,1
R 1,2
C 1,2
R 1,3
C 1,3
R 1,4
C 1,4
R 1,1_r
P 1,1_p
C 1,1_c
C 2,2_c
C 3,3_c
C 4,4_c
C n,n_c
P 2,2_p
R 2,2_r
P 3,3_p
R 3,3_r
P 4,4_c
R 4,4_r
R n,n_r
P n,n_p
R : reactions
C : chemicals
P : pathways
n : iteration no.
n_r : no. of feasible reactions at n
n_c : no. of feasible chemicals at n
n_p : no. of feasible pathways at n
n_r_total : total no. of reactions at n
n_c_total : total no. of chemicals at n
n_p_total : total no. of pathways at n

Synthetic Pathway Prediction and Strain Design Pipeline



Center for Biosustainability

Validation of experimental pathways

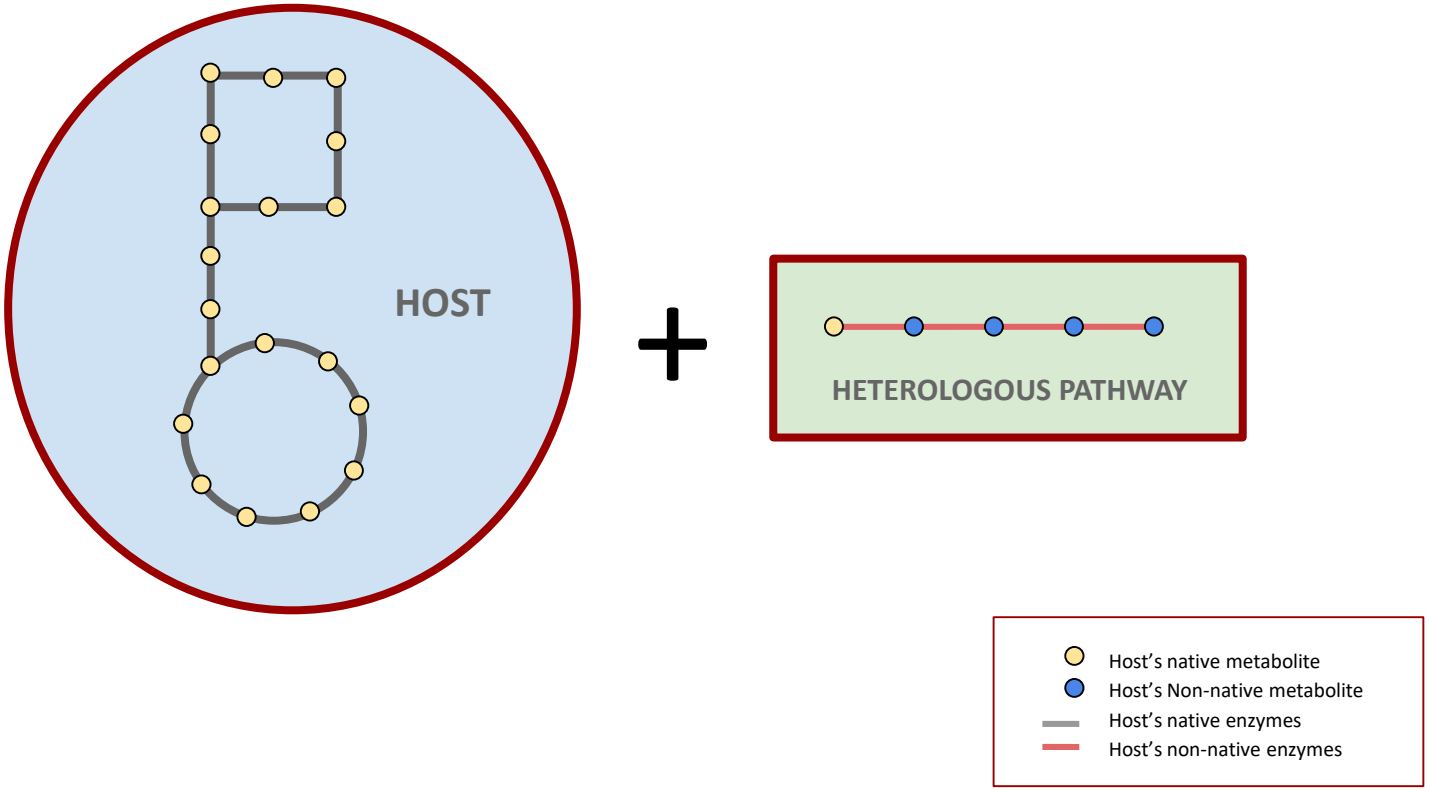


RDKit Usage

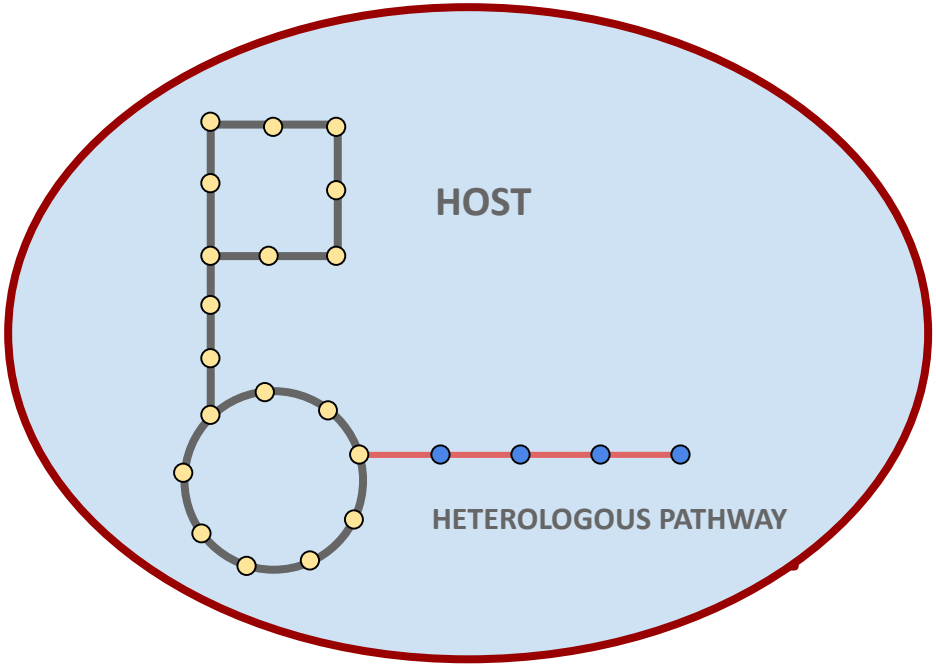
- Manipulations with molecules
  - Reading molecules from different formats (smiles, smarts, sdf, mol, inchi)
  - Getting molecular information (formula, charge)
  - Standartization of molecules
  - Molecular similarity (fingerprints, tanimoto coefficient)
  - Substructure match
- Manipulations with reactions
  - Reactor (for reaction SMARTs)
  - Custom SMIRK parser
  - Automatic BROs generation (in progress)
- Properties prediction
  - Qsar library
  - Toxicity IC50 (qsar model based on fingerprints)

Post-analysis

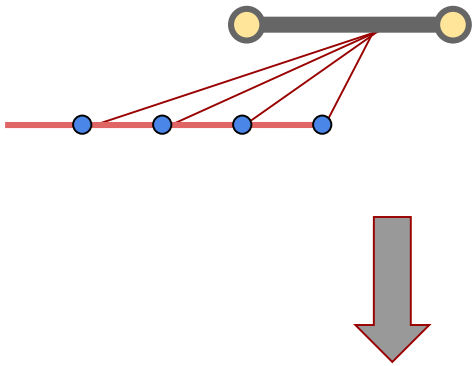
De-novo pathway design and Host integration



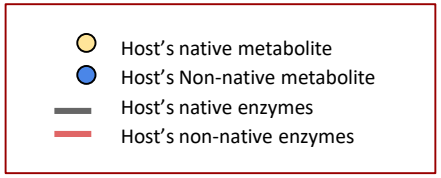
HOST METABOLISM OFF-TARGET EFFECTS



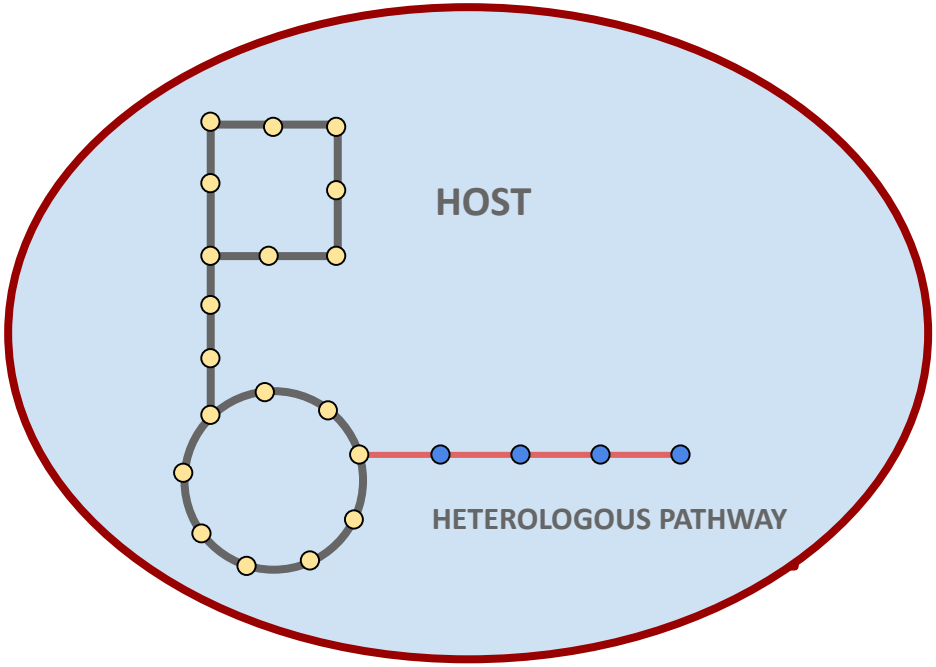
for native enzyme i in n:
apply BRO to all non-native metabolites



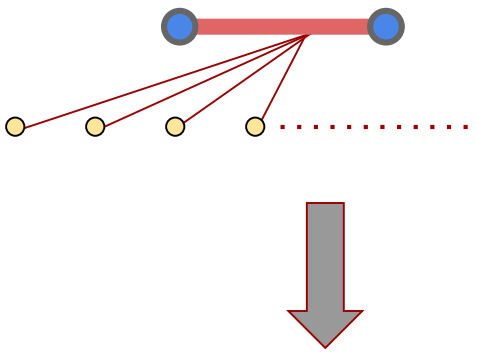
List of potential host metabolism off-target effects on heterologous pathway's non-native metabolites



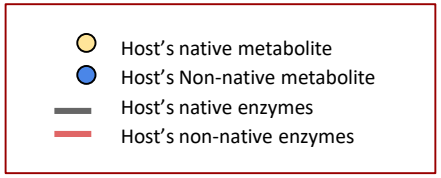
HETEROLOGOUS PATHWAY OFF-TARGET EFFECTS



for each non-native enzyme t in m:
apply BRO to all native metabolites



List of potential heterologous pathway's enzymes off-target effects on native metabolites



Summary

- Prediction of potential biochemical reactions pathways based on GEM
- Forward synthesis
- Further analysis
  - Theoretical yield
  - Off-target effect
  - Connection to other tools for further design
- Database storage
  - Graph database for pathway search
  - Sql database for additional information
- Documentation and release for user testing

The Novo Nordisk Foundation Center for Biosustainability

Thank you for your attention

Center for Biosustainability

Marina Fedorova

mfad@biosustain.dtu.dk

www.biosustain.dtu.dk