Artificial Intelligence /Machine Learning for Secondary Metabolite Prediction

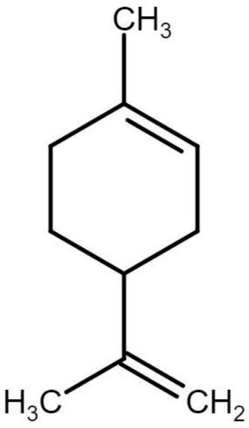
Yannick Djoumbou Feunang
Corteva Agriscience, Indianapolis, IN, US

Online workshop: Computational Applications in Secondary Metabolite Discovery
March 9, 2021

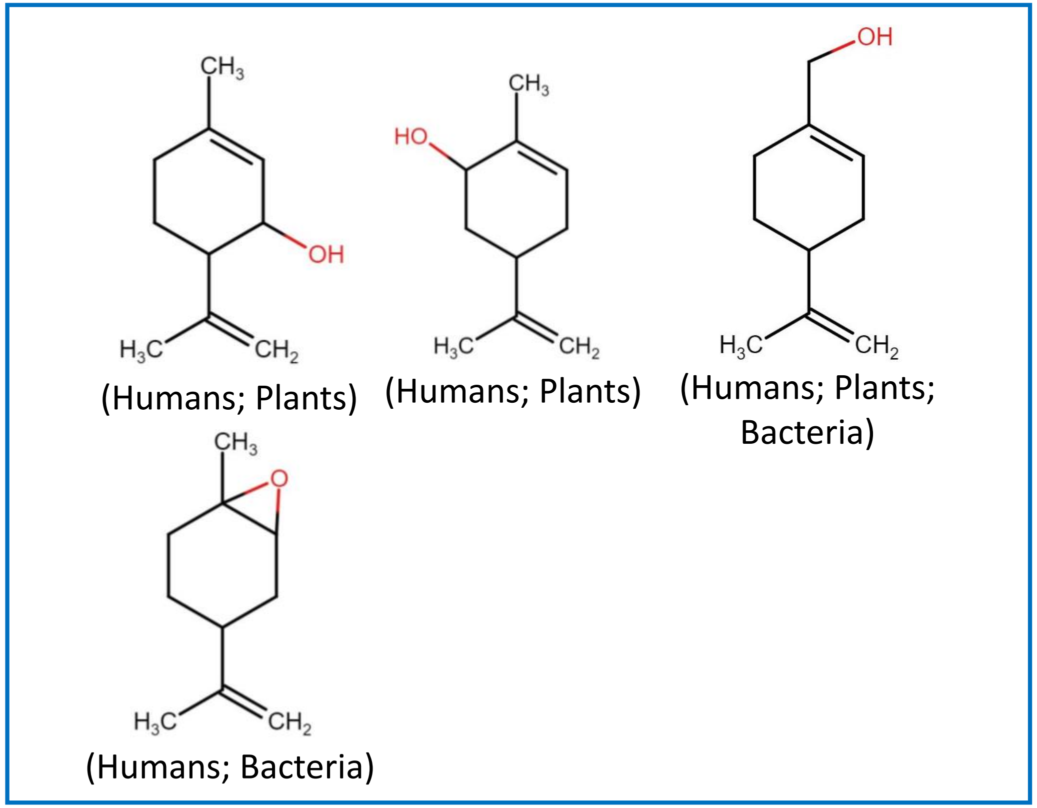
In Silico Metabolism Prediction

Task: Given a small molecule, use computational tools to predict the outcome of its interactions with metabolic enzymes.

Here, we will focus on the structural elucidation of potential metabolites.



Limonene
(Citrus)



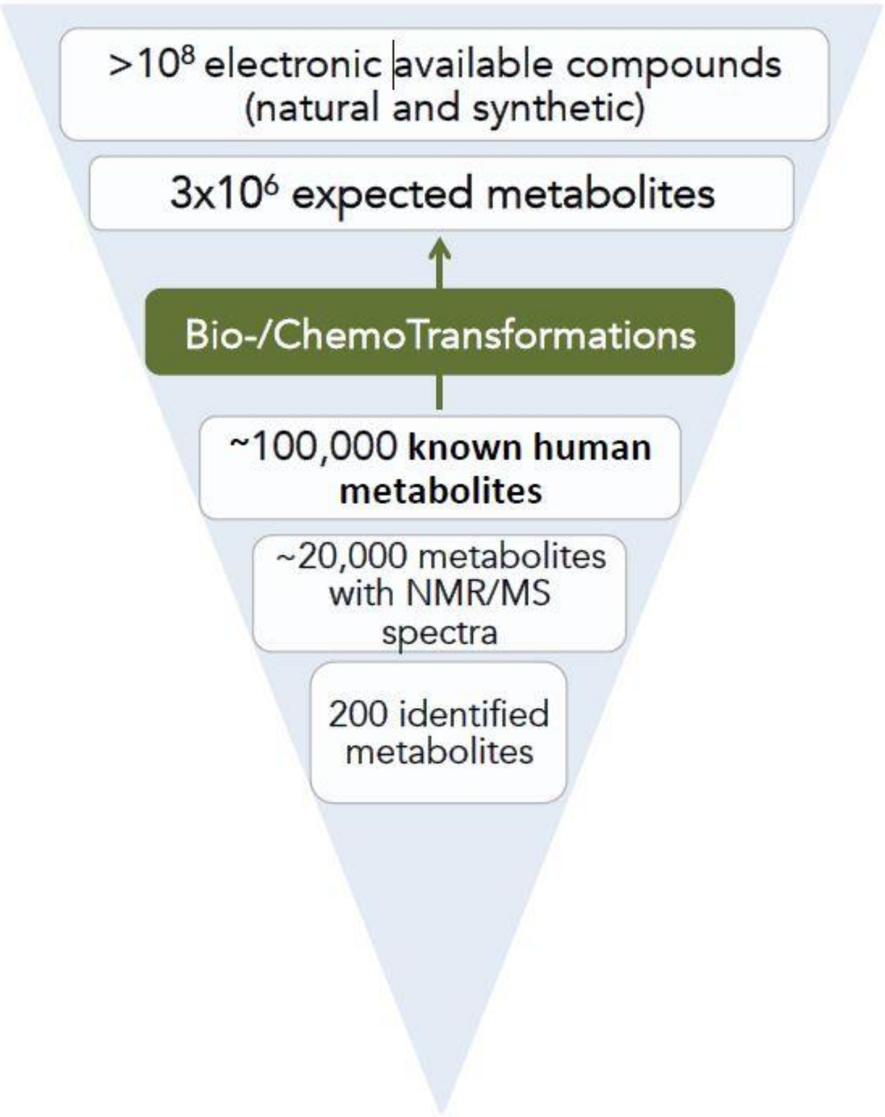
(Humans; Plants)

(Humans; Plants)

(Humans; Plants; Bacteria)

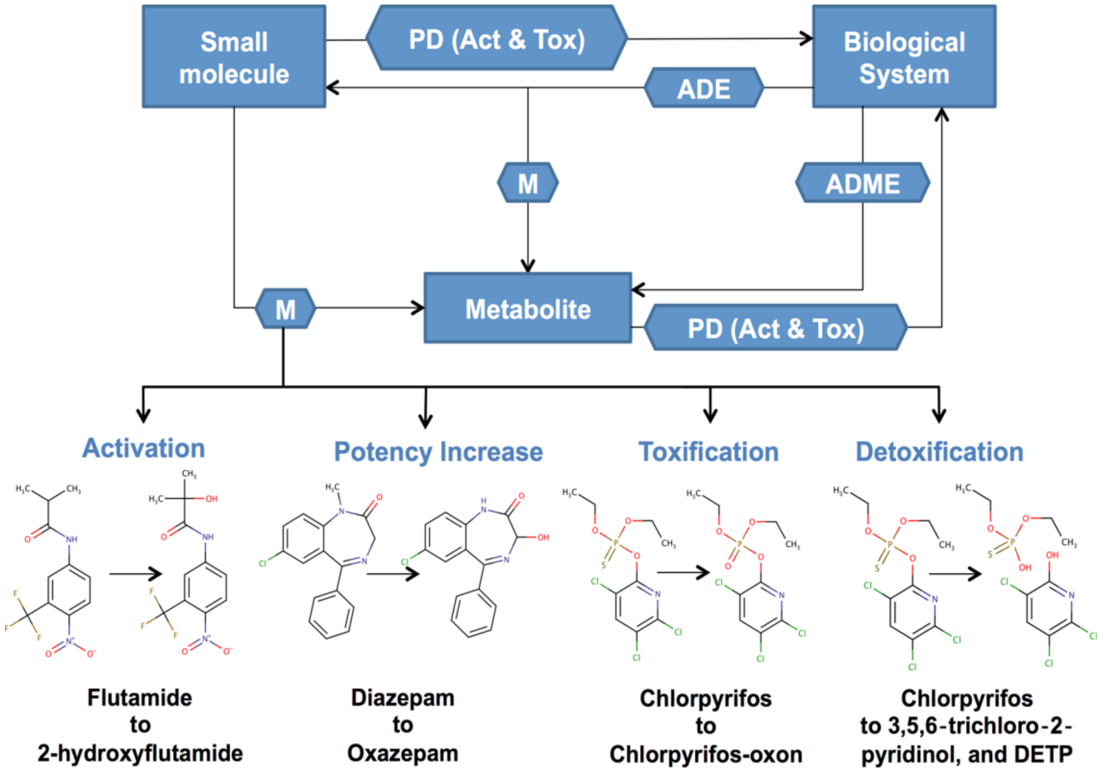
(Humans; Bacteria)

Why is Metabolism so Important?


$>10^8$ electronic available compounds (natural and synthetic)
$3x10^6$ expected metabolites
Bio-/ChemoTransformations
~100,000 **known human metabolites**
~20,000 metabolites with NMR/MS spectra
200 identified metabolites

- Bio/chemo-transformations influence changes of our chemical exposome
- <2% of detectable peaks identifiable in large-scale untargeted Metabolomics
- What are these unknowns?
- How to determine their structures?
- How to determine their activities?
- Cheminformatics and AI can be used to:
    - ➢ Detect common patterns
    - ➢ Predict enzyme/ligand interactions
    - ➢ Understand metabolism
    - ➢ Generate biologically feasible structures through simulated reactions

Why is Metabolism so Important?

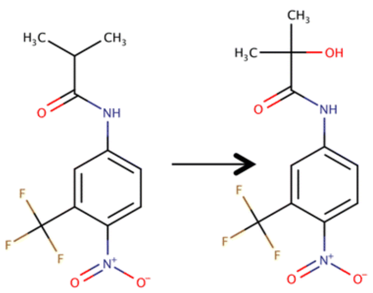


Djoumbou Feunang *et al.* (2019);

Influence of metabolism on a xenobiotic's pharmacodynamics (PD), including pharmacological activity (Act), and toxicological effects (Tox).

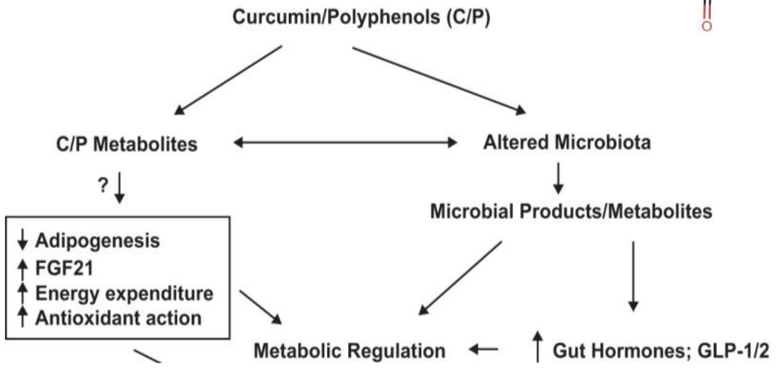
Pro/soft-drug design

Activation



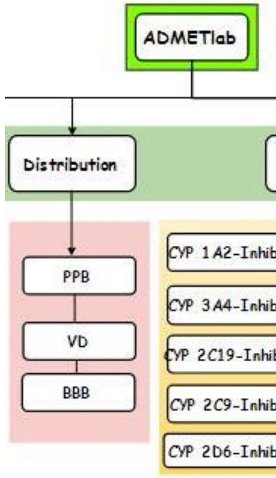
Exposure science

Nutrition science



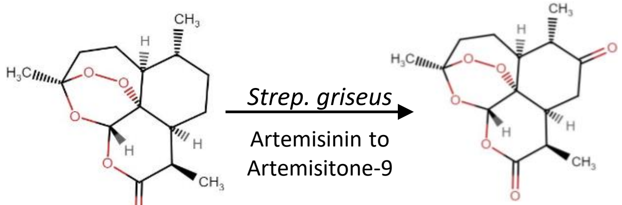
Jin *et al.* (2017)

ADMET profiling



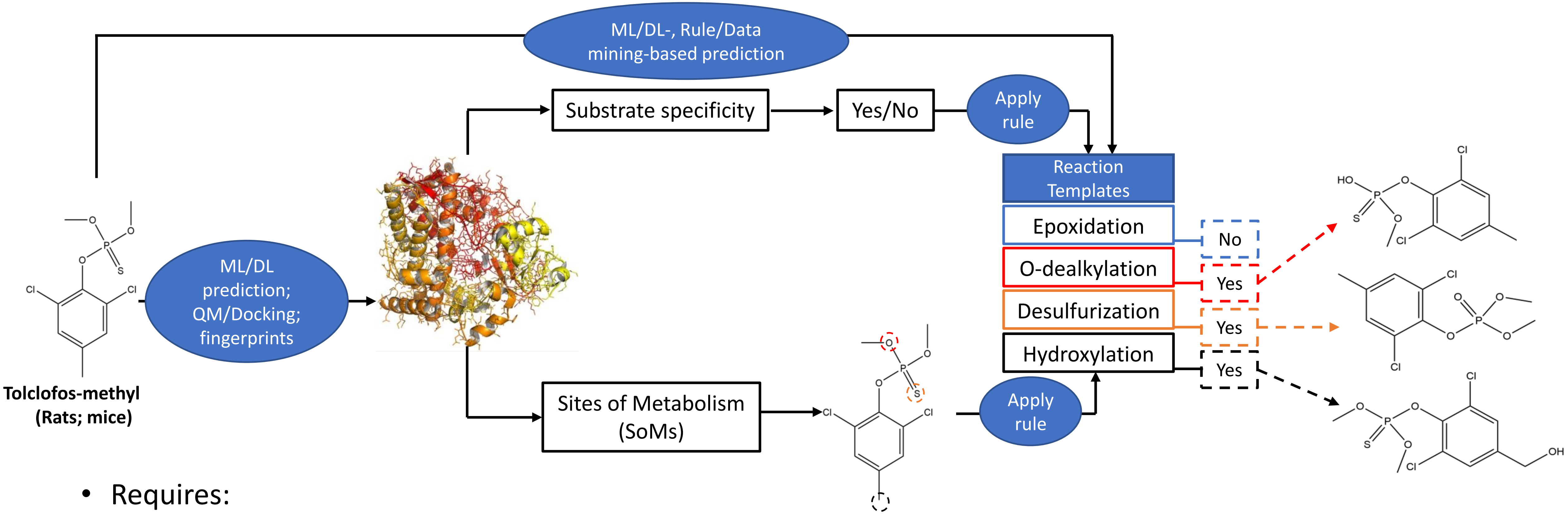
http://admet.scbdd.com/

Microbial biosynthesis



Liu *et al.* (2006)

Approaches for *In silico* Metabolism Prediction



ML/DL-, Rule/Data mining-based prediction

Substrate specificity

Yes/No

Apply rule

Reaction Templates

Epoxidation

O-dealkylation

Desulfurization

Hydroxylation

No

Yes

Yes

Yes

ML/DL prediction; QM/Docking; fingerprints

Sites of Metabolism (SoMs)

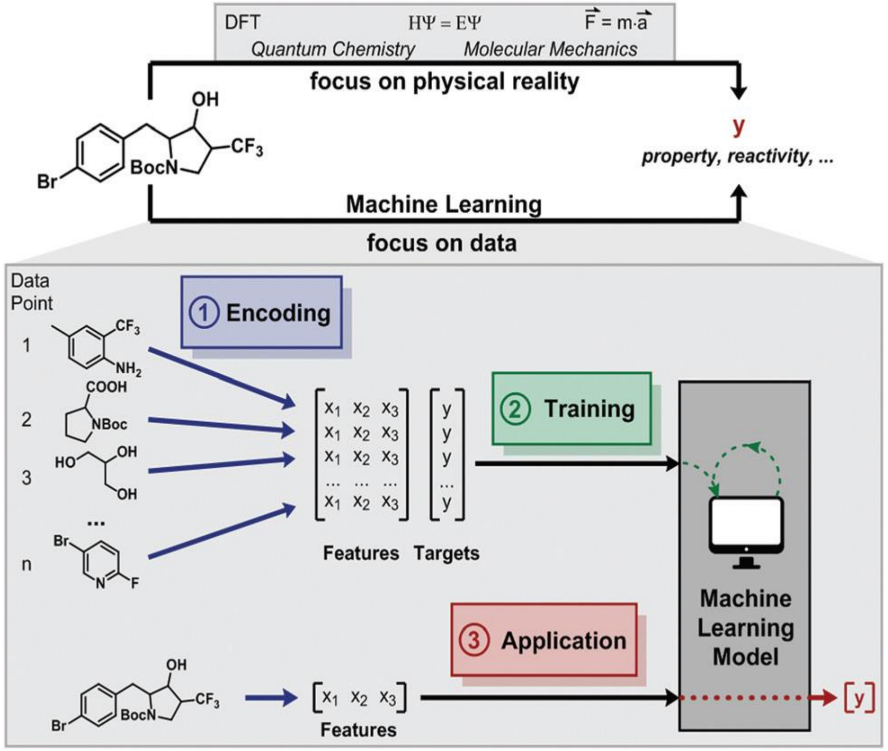
Apply rule

Tolclofos-methyl (Rats; mice)

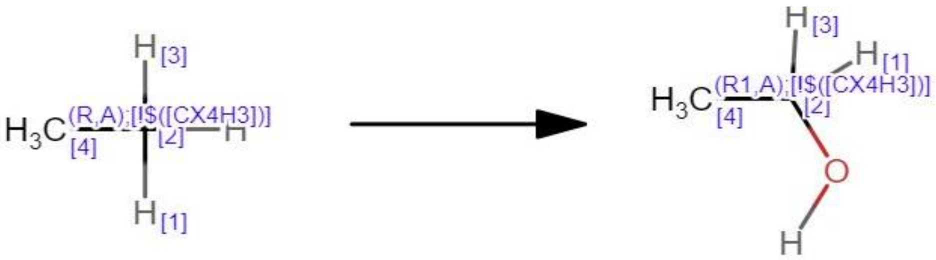
- Requires:
  - Module to predict (and rank/score) SoMs, enzyme-substrate selectivity, or reaction groups
  - Library of reaction templates to apply or select (via prediction) from
  - Modules are usually specific (chemical/enzyme classes), or comprehensive (whole species)
- Prediction approaches can be ligand- or structure-based
- Some tools include: ML/DL-based (MetaTrans, Glory), Rule-based (MetabolExpert), Hybrid (BioTransformer, Meteor Nexus)

Ligand-based Prediction

- Use structures of ligands/non-ligands to predict specificity, accessibility to enzymes
- Some methods include, among others:
  - QSAR/QSMR (SoM, BoM, ESS)
  - Data mining/fingerprints (SoM)
  - Substructures/Rules (ESS, Reaction)
- Advantages:
  - Speed
  - Seem to perform as well as structure-based methods
- Disadvantages:
  - Approaches/models/rule bases tend to not work on novel chemistries
  - Data quantity and quality is a limiting factor



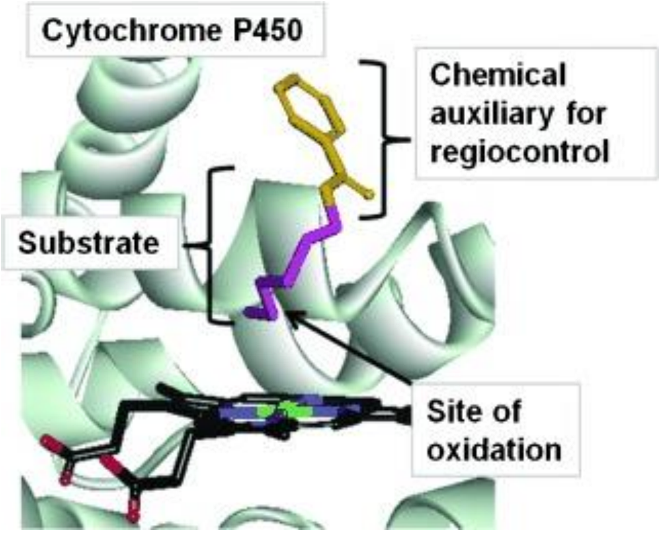
Strieth-Kalthoff et al. (2020)



Hydroxylation of methyl carbon adjacent to aliphatic ring

Structure-based Prediction

- Explicitly model the interaction within the enzyme’s binding pocket
- Help predicting sites of metabolism (SoMs)
- Some methods include, among others:
  - Protein-ligand docking
  - Molecular dynamics
- Advantages
  - Detailed study of the enzyme-substrate interaction
- Disadvantages
  - High computational power to model structural flexibility



Ménard *et al.* (2012)

Examples of *In Silico* Metabolism Prediction Tools

Based on their coverage, *in silico* metabolism prediction tools can be classified as:

1. Specific tools, which apply to simple biological systems (e.g. a small set of enzymes), and often, cover a rather small chemical space.
2. Comprehensive tools, which apply to a more diverse and larger set of enzymes, species, and cover a larger chemical space.

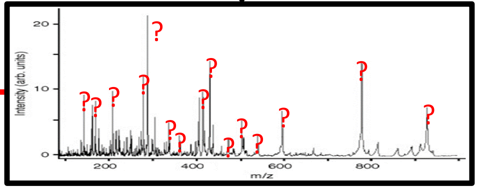
Prediction Tool	Accessibility	Spectrum	Approach	Methods	Availability
BioTransformer	Web server / Command line	Comprehensive	Ligand-based	Machine Learning Expert system	Open source
Meteor Nexus	GUI-based standalone	Comprehensive	Ligand-based	Machine Learning Expert system	Commercial
GLORYx	Web server	Comprehensive	Ligand-based	Machine Learning	Freely available
SmartCYP	Web server	Specific	Combined	Machine Learning	Freely available
MetaSite	GUI-based standalone	Specific	Combined	Machine Learning	Commercial
MetabolExpert	GUI-based standalone	Specific	Ligand-based	Expert system	Commercial

Examples of tools for metabolite prediction

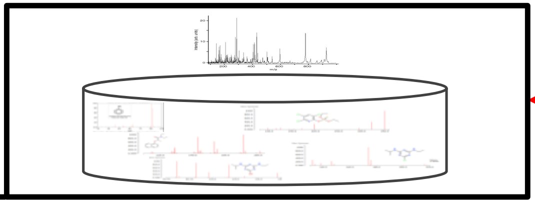
Enhancing Metabolite Discovery and Identification

- >2x10^5 known metabolites
- > 3x10^6 expected metabolites

Measuring the exposome

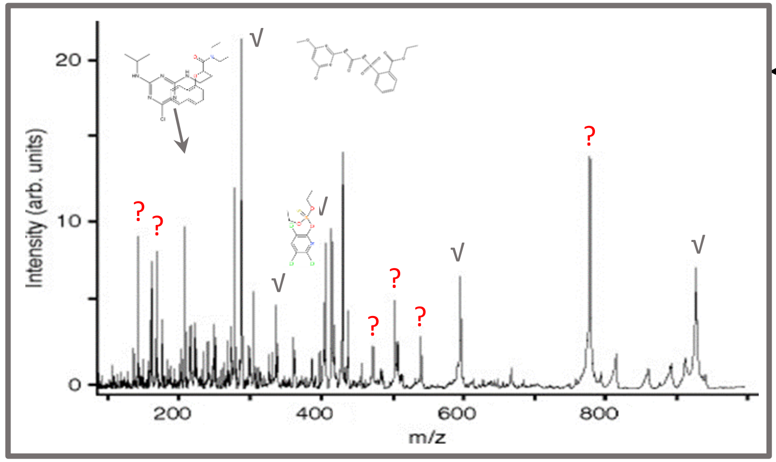


MS/NMR-based Metabolomics



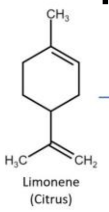
More structures and spectra available for database search

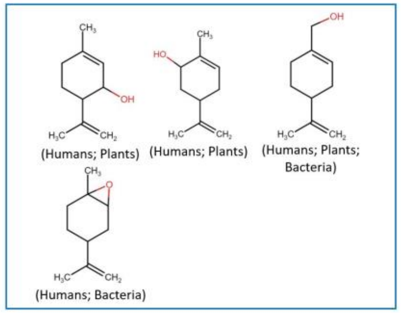
Higher identification rates



- Expand the databases with predicted structures
- Add predicted spectra for these metabolites

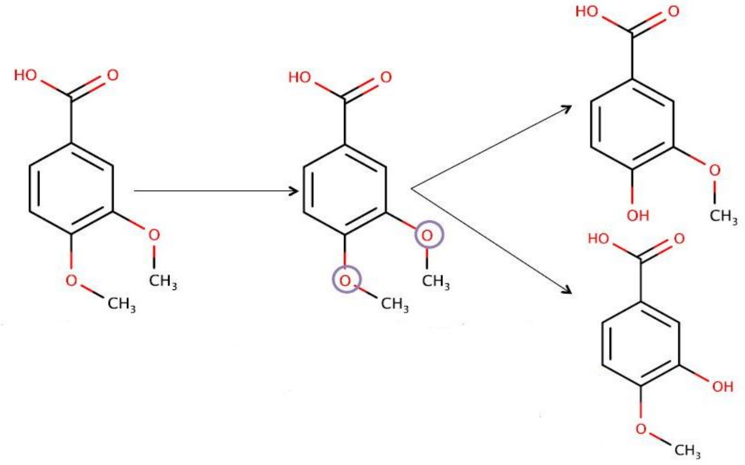
More structures available for mass/formula matching to annotate the peaks



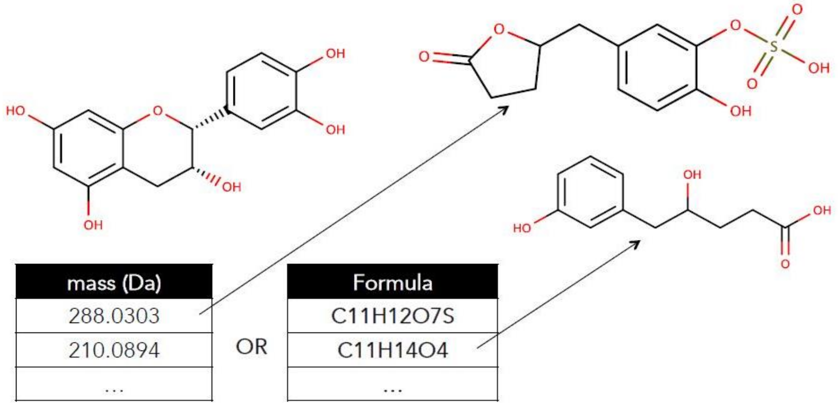


BioTransformer: Open Source for Metabolite Identification

BioTransformer Metabolism Prediction Tool



BioTransformer Metabolite Identification Tool

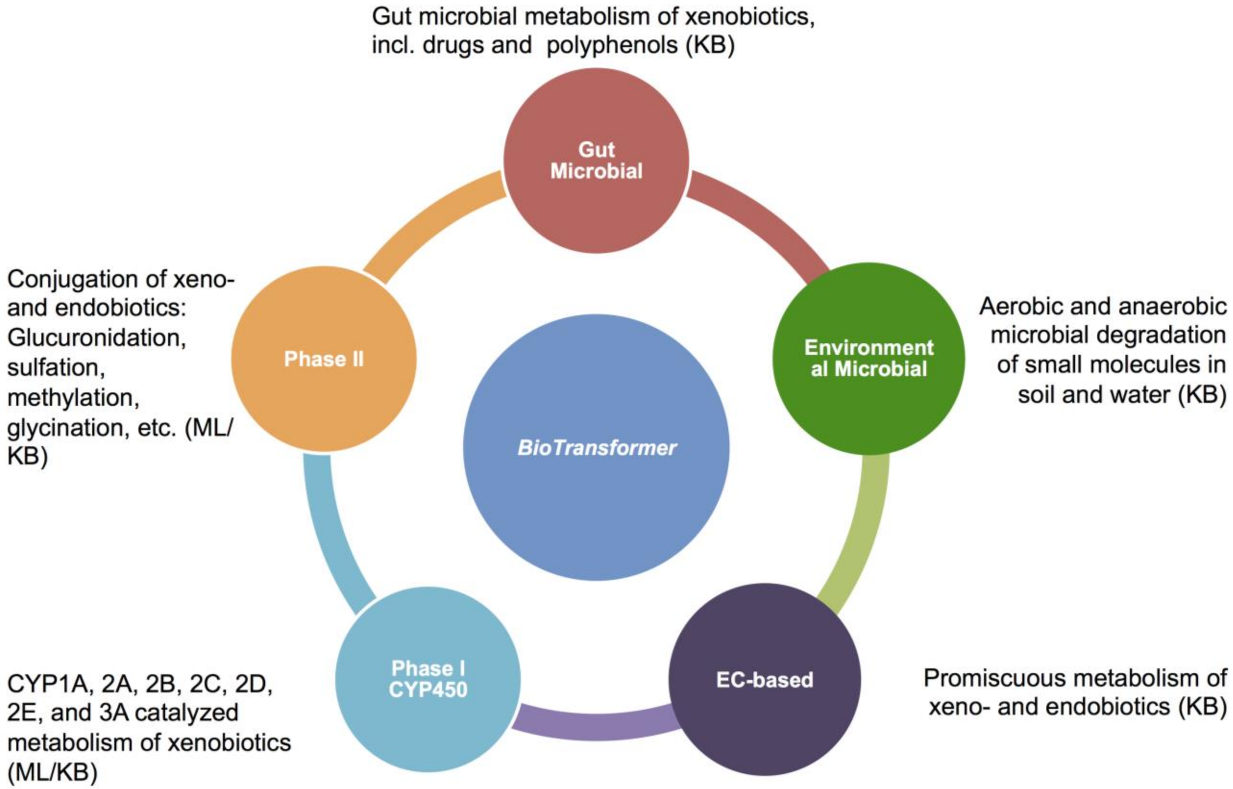


mass (Da)
288.0303
210.0894
…

OR

Formula
C11H12O7S
C11H14O4
…

Djoumbou Feunang *et al.* (2019); J. Cheminf.; DOI 10.1186/s13321-018-0324-5



Gut microbial metabolism of xenobiotics, incl. drugs and polyphenols (KB)

Aerobic and anaerobic microbial degradation of small molecules in soil and water (KB)

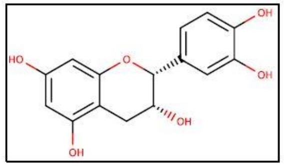
Promiscuous metabolism of xeno- and endobiotics (KB)

CYP1A, 2A, 2B, 2C, 2D, 2E, and 3A catalyzed metabolism of xenobiotics (ML/KB)

Conjugation of xeno- and endobiotics: Glucuronidation, sulfation, methylation, glycination, etc. (ML/KB)

- **Djoumbou-Feunang *et al.* J Cheminform. 2019 Jan 5;11(1):2**
- Open Source: bitbucket.org/djoumbou/biotransformer
- Docker: https://hub.docker.com/r/djoy2018/biotransformer-cl
- RESTful web server: www.biotransformer.ca

BioTransformer: Open source for Metabolite Identification



Epicatechin

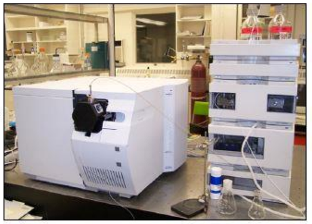
→

12-week old Wistar rats were fed epicatechin

→

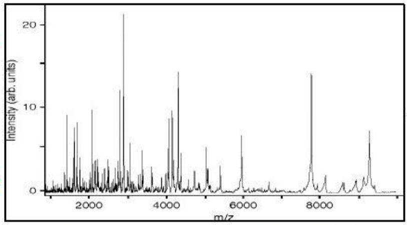
Urine collected before and after 5-day treatment

↓



Sample Analysis with UPHLC - QToF MS

←



Data Extraction (m/z and intensities)
260 neutral monoisotopic masses (53.48 – 969.86 Da)

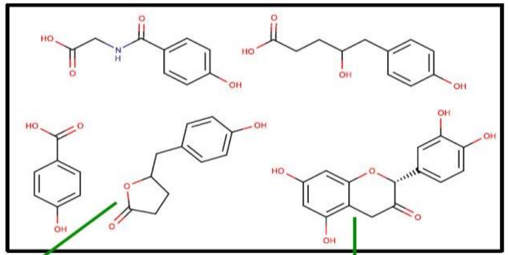
Use BioTransformer to identify/suggest metabolites

???

In collaboration with Manach C., and Fiamoncini J., INRA, France

m/z (Da)
122.0369
138.0319
176.0837
192.0786
195.0532
208.0736
210.0894
…

BioTransformer (BMIT) →



Metabolite never reported, but mass known

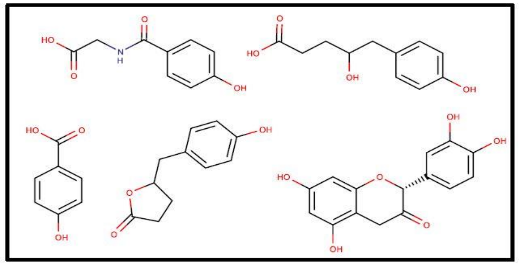
Metabolite and mass never reported before

- BioTransformer (BMIT) suggested 19 new compounds
  - 11 compounds match previously reported masses
  - 15 compounds match masses observed exclusively in our study
  - Experimental validation needed

- Literature mining revealed 56 single- and multi-step epicatechin metabolites corresponding to 37 unique masses
  - 11 out of 37 masses were measured in our experimental study

m/z (Da)
122.0369
138.0319
176.0837
192.0786
195.0532
208.0736
210.0894
…

BioTransformer (BMIT) →

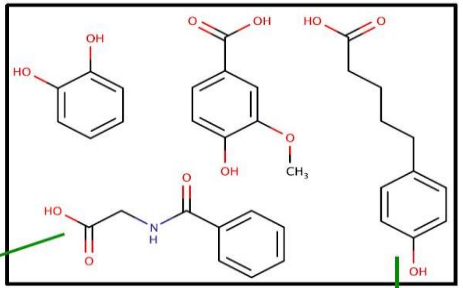


- BioTransformer (BMIT) suggested 37 compounds (20 masses)
  - 22/37 matched the 11 unique and previously reported masses
  - 18/22 compounds were confirmed in previous reports

- BioTransformer (BMIT) was used to identify metabolites matching the 26 masses not measured in our study

m/z (Da)
110.0367
126.0316
182.0579
194.0942
274.0511
276.0997
292.0946
306.0409
370.0358
…

BioTransformer (BMIT) →



Hippuric acid: a biomarker of tea consumption

Metabolite never reported before

- BioTransformer (BMIT) suggested 28 compounds (19 masses)
  - 21 previously reported metabolites matching 18 unique masses
  - Experimental validation needed for the 7 new compounds

Challenges in Metabolism Prediction

- Challenges in understanding metabolism lead to lower prediction accuracy
  - ➢ ML tools can rank SoMs with high accuracy, but not the resulting biotransformations
  - ➢ Knowledge-driven approaches achieve higher recall, but suffer from combinatorial explosion
- Limited accessibility/availability of tools
  - ➢ Most tools developed in academia are available only via web-sever, raising confidentiality issues for most potential users
  - ➢ Restricted shareability of data generated by commercial tools
- The limited availability of high-quality data
  - ➢ Crucial for achieving better accuracy, higher coverage, and larger applicability domains
  - ➢ Needed to derive more reaction rules
  - ➢ However, data curation (Sites of metabolism, reaction annotation) is tedious and expensive

Outlook

- Data sharing
  - ➢ Increase the amount of high-quality, publicly available, curated, and downloadable data (substrate-product, Sites of Metabolism). E.g.: MetXBioDB, PubChem, XMetDB, etc.
  - ➢ Develop publicly accessible databases of predicted metabolites (e.g.: BioTransformerDB)
  - ➢ Community-wide efforts needed
- Novel, and innovative AI approaches:
  - ➢ Seq2Seq transformer architectures have proven applicable for end-to-end learning-based method, with results comparable to existing tools (MetaTrans: Litsa *et al.* 2020)
    - ▪ Could improve accuracy, while bypassing manual rule design
  - ➢ Bonds of Metabolism (BoMs) seem to provide a better description of reaction centers, leading to higher accuracy, compared to SoMs (Upcoming - Tian, et al. (2021))
- Open source communities
  - ➢ Provide means for easier user feedback loops; valuable for improving prediction tools
  - ➢ Provide developers opportunities to improve software tools, in an agile, continuous manner
    - ▪ Successful projects include (BioTransformer, Chemistry Development Kit, Knime)

Thank You

- The organizing committee:
  - Fidele Ntie-Kang
  - J. Ludwig Muller
- The Wishart Lab @UofAlberta, Canada
- Corteva Agriscience
- The BioTransformer community
- The listeners

- To learn more about BioTransformer:
  - *HS03: In Silico Prediction and Identification of Metabolites with BioTransformer : Enabling Secondary Metabolite Discovery; March 10, 2021*

https://unlockinglifescode.org/education-resource-profile/resource-month-ask-biologist