universität wien

Metabolism prediction

Johannes Kirchmair

Slides available from:

31-Jan-23 Johannes Kirchmair

universität wien

The Computational Drug Discovery and Design Group (COMP3D), Christian-Doppler Laboratory for Molecular Informatics in the Biosciences

- Monika Babicki
- Malena Brenek
- **Christina Brenner**
- **Ya “Anya” Chen**
- Ningning Fan
- Hosein Fooladi
- Marina Garcia de Lomana
- Steffen Hirte
- **Roxane Jacob**
- Judith Maaß
- Tian-Yu Niu
- Vincenzo Palmacci
- **Wojtek Plonka**
- **Vincent-Alexander Scholz**
- Nicole Schuldhaus
- Axinya Tokareva
- Thi Ngoc Lan Vu
- **Huanni Zhang**

Christian Doppler Forschungsgesellschaft

FTE NATIONALSTIFTUNG FORSCHUNG | TECHNOLOGIE | ENTWICKLUNG

Federal Ministry Republic of Austria Digital and Economic Affairs

CHINA SCHOLARSHIP COUNCIL 国家留学基金管理委员会

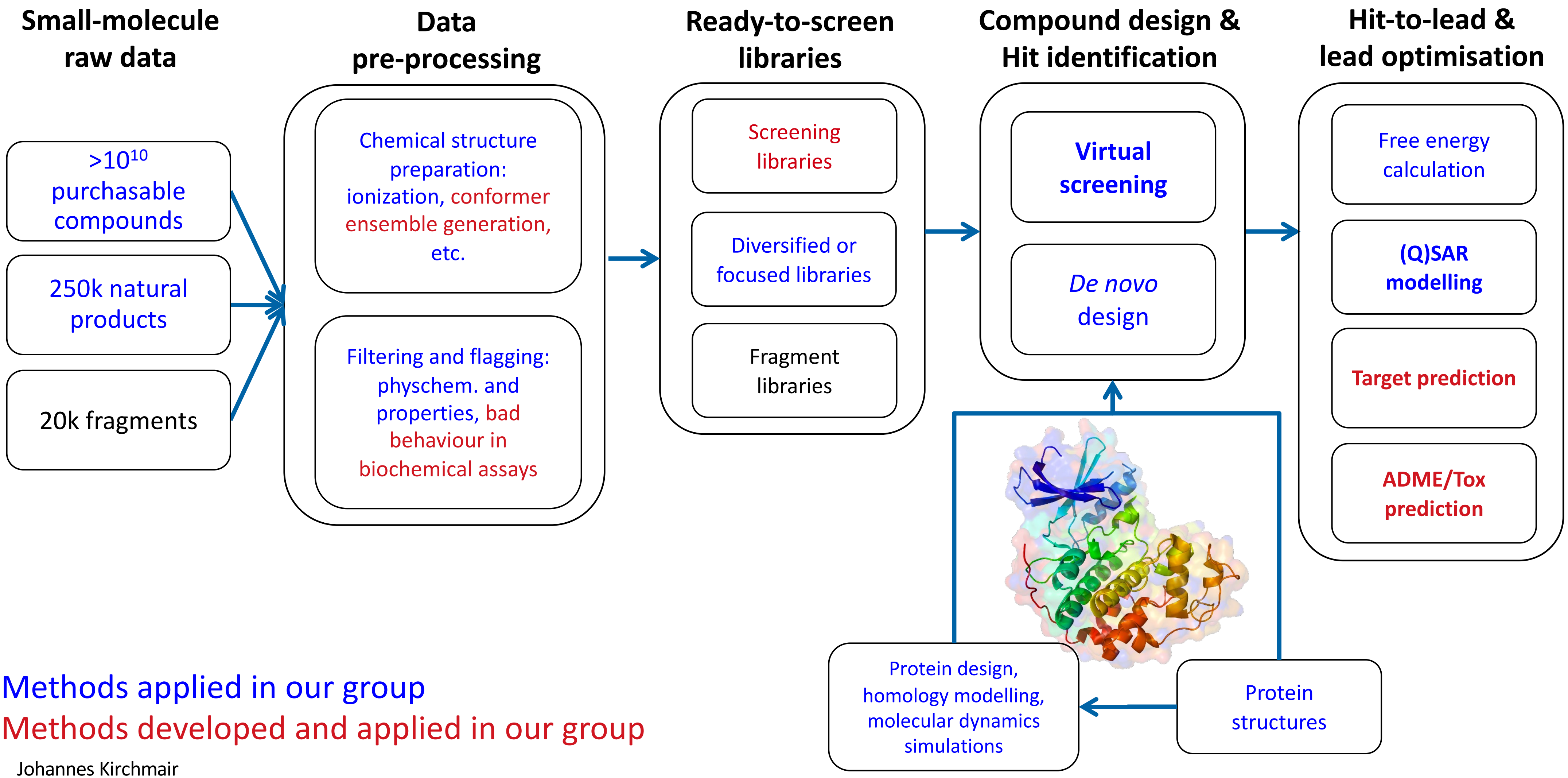
BASF We create chemistry

Beiersdorf

Boehringer Ingelheim

syngenta

Core research topics:
Machine learning – Bioactivity prediction – ADME/T prediction – natural products



Small-molecule raw data

- $>10^{10}$ purchasable compounds
- 250k natural products
- 20k fragments

Data pre-processing

- Chemical structure preparation: ionization, conformer ensemble generation, etc.
- Filtering and flagging: physchem. and properties, bad behaviour in biochemical assays

Ready-to-screen libraries

- Screening libraries
- Diversified or focused libraries
- Fragment libraries

Compound design & Hit identification

- **Virtual screening**
- *De novo* design

Hit-to-lead & lead optimisation

- Free energy calculation
- **(Q)SAR modelling**
- **Target prediction**
- **ADME/Tox prediction**

Protein structures

Protein design, homology modelling, molecular dynamics simulations

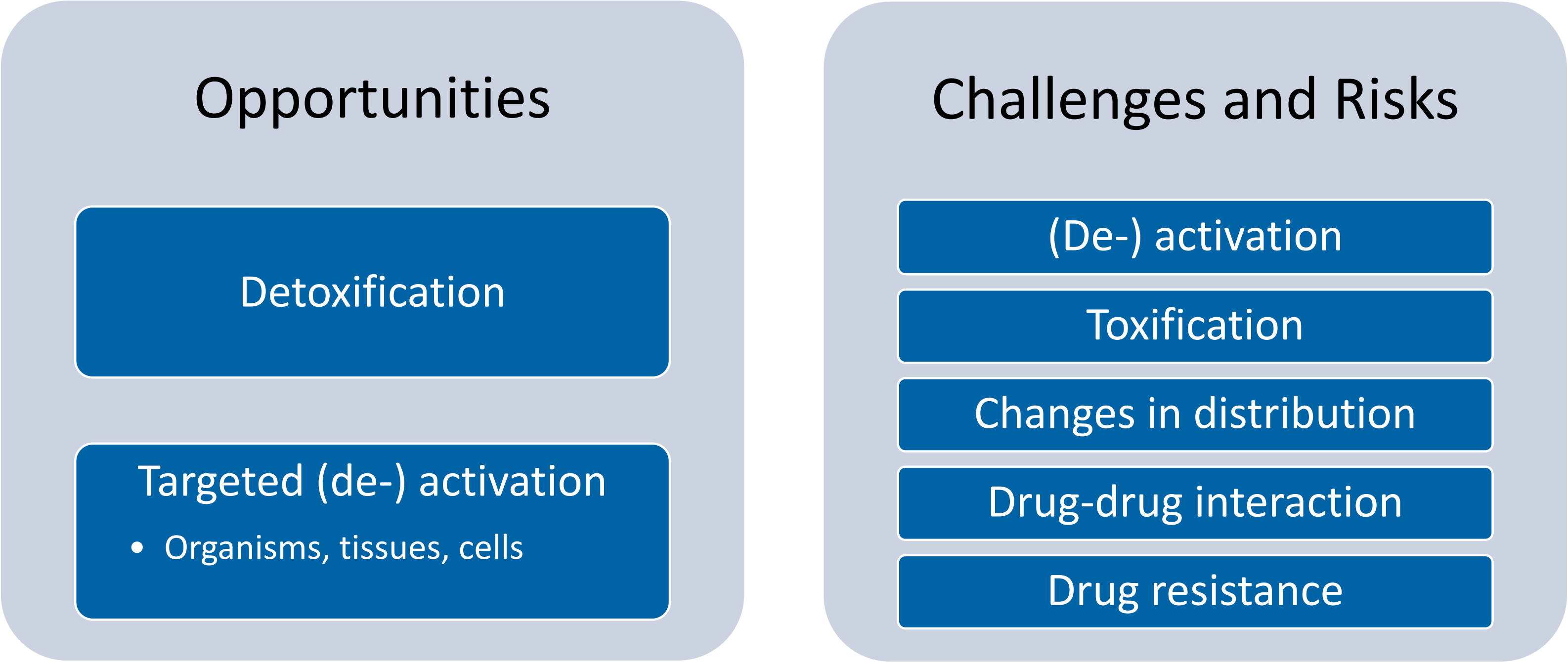
Methods applied in our group

Methods developed and applied in our group

Understanding xenobiotic metabolism is key to the design of safe and efficacious small molecules

- Metabolism is the main clearance pathway of 75 to 90% of all drugs
- Drugs and drug-like compounds have, on average, metabolites[1]

- Only 3% of metabolites are confirmed to maintain their pharmacological activity[1]
- **At least 7% of metabolites are known to be reactive and/or toxic[1]**



Opportunities

Detoxification

Targeted (de-) activation

- Organisms, tissues, cells

Challenges and Risks

(De-) activation

Toxification

Changes in distribution

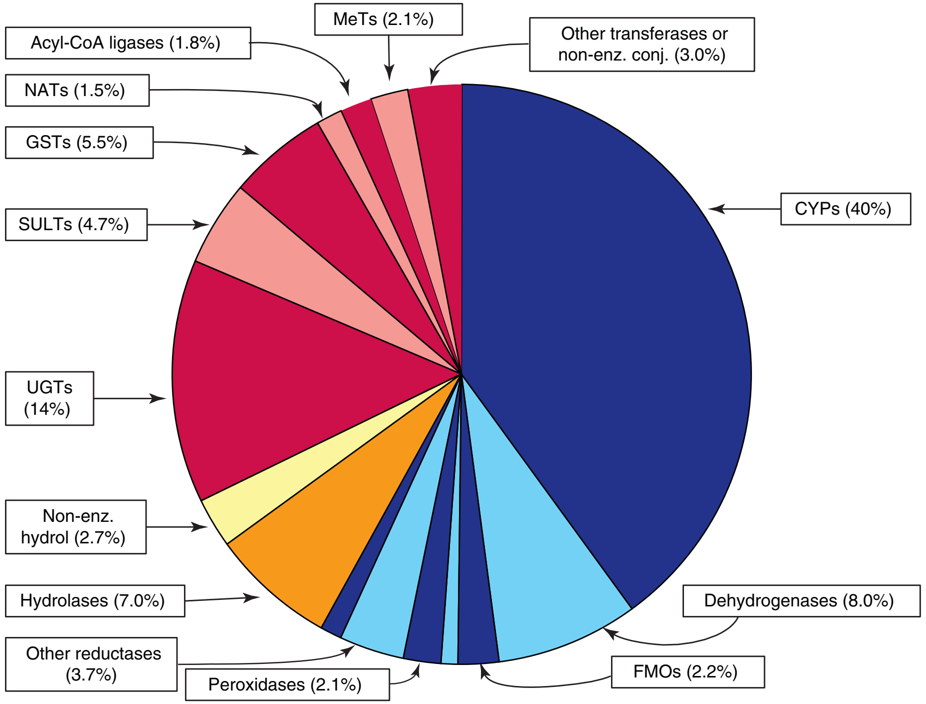
Drug-drug interaction

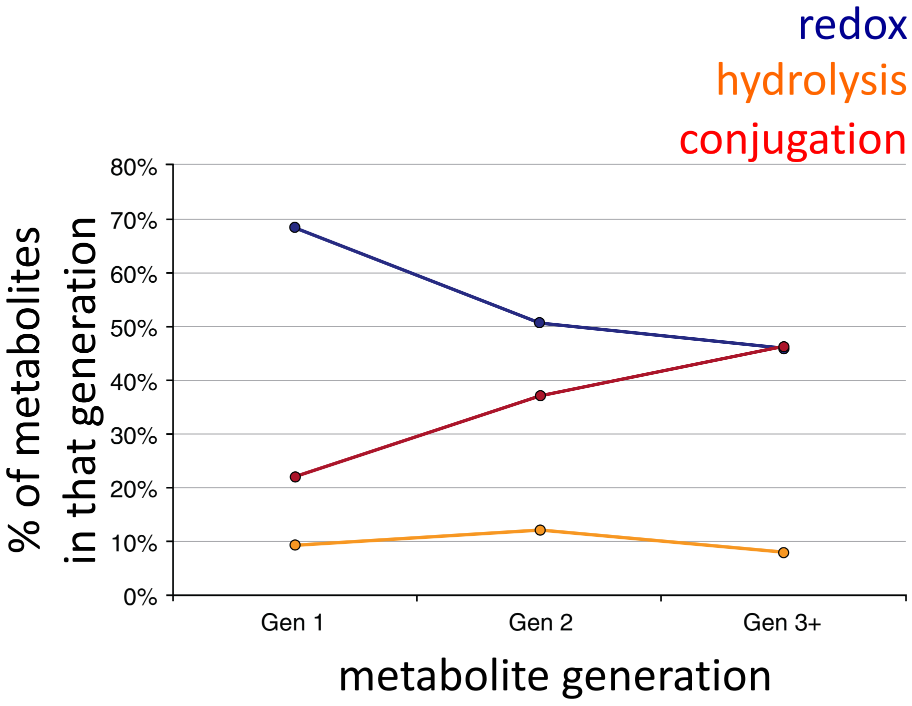
Drug resistance

[1] Testa et al, Drug Discov Today 2012, 17, 549-560. doi: 10.1016/j.drudis.2012.01.017
Kirchmair et al., Nat Rev Drug Discov 2015, 14, 387-404. doi: 10.1038/nrd4581

The metabolic system is highly complex and adaptive

- Diverse and complex families of enzymes
- Varying expression patterns among different species, organs and tissues
- Inter-individual factors: genetic differences, polymorphisms
- Intra-individual factors: age, pregnancy, disease, stress, diet, etc.
- Synergistic collaborations with transporters
- Important but weakly understood role of gut microbiota in metabolism

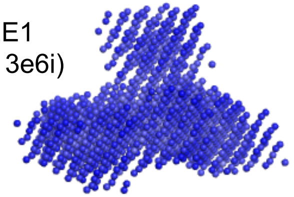




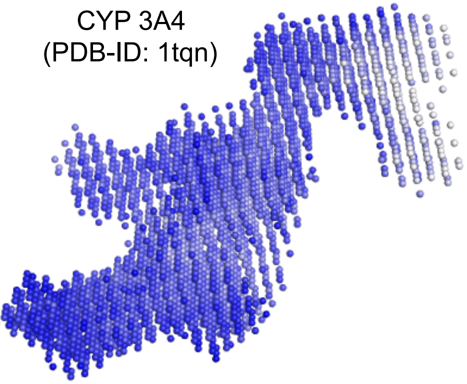
Testa et al, Drug Discov Today 2012, 17, 549-560. doi: 10.1016/j.drudis.2012.01.017

CYPs are highly malleable and promiscuous

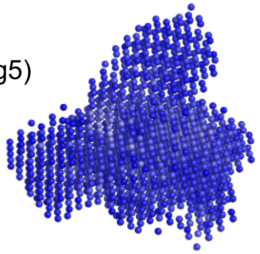
CYP 2E1
(PDB-ID: 3e6i)

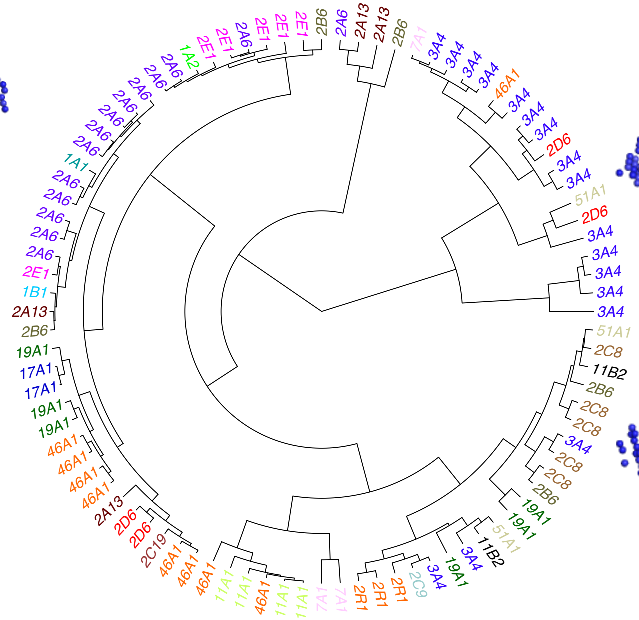


CYP 3A4
(PDB-ID: 1tqn)

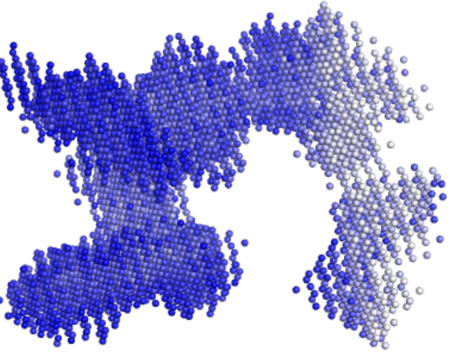


CYP 2A6
(PDB-ID: 2pg5)

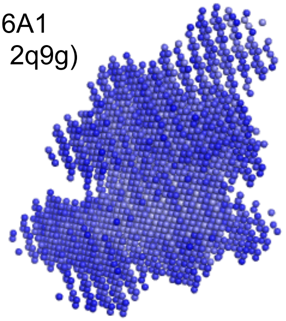




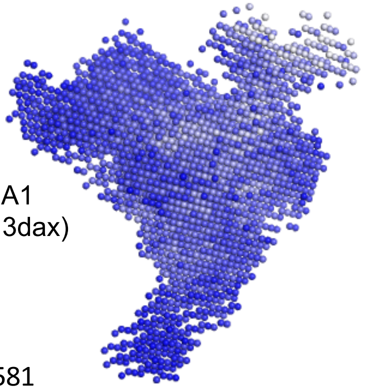
CYP 2C8
(PDB-ID: 1pq2)



CYP 46A1
(PDB-ID: 2q9g)



CYP 7A1
(PDB-ID: 3dax)



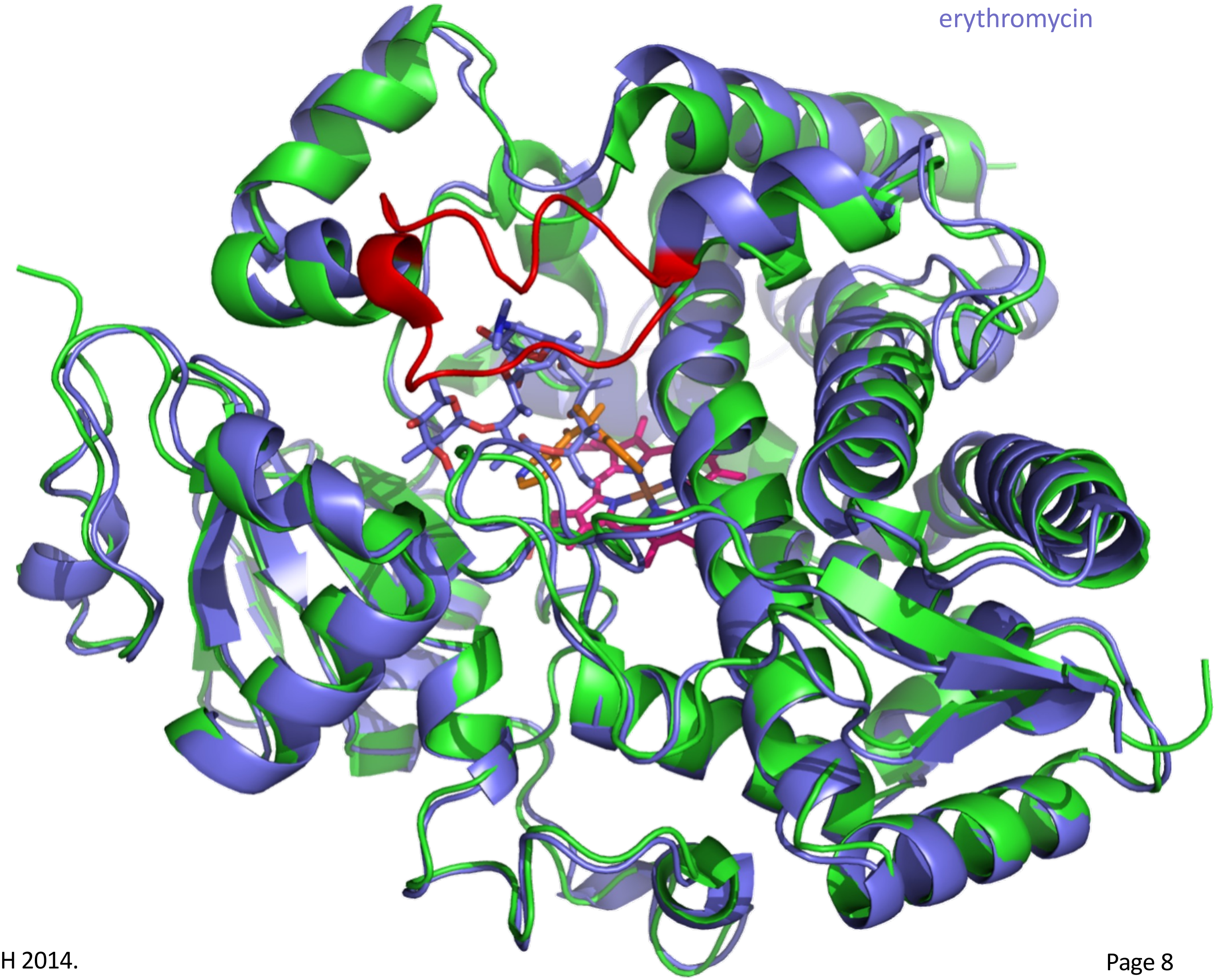
Kirchmair et al., Nat Rev Drug Discov 2015, 14, 387-404. doi: 10.1038/nrd4581

Structural data on CYPs have become available but enzyme malleability remains challenging for drug design

Coverage human CYPs with X-ray structures

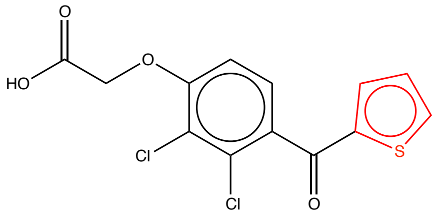
Sterols	Xenobiotics	Fatty acids	Eicosanoids	Vitamins	Unknown
🟩 1B1	🟩 1A1	2J2	4F2	🟩 2R1	2A7
7A1	🟩 1A2	4A11	4F3	24A1	2S1
7B1	🟩 2A6	4B1	4F8	26A1	2U1
8B1	🟩 2A13	4F12	5A1	26B1	2W1
🟩 11A1	🟩 2B6		🟩 8A1	26C1	3A43
11B1	🟩 2C8			27B1	4A22
11B2	🟩 2C9				4F11
🟩 17A1	2C18				4F22
🟩 19A1	🟩 2C19				4V2
21A2	🟩 2D6				4X1
27A1	🟩 2E1				4Z1
39A1	2F1				20A1
🟩 46A1	🟩 3A4				27C1
🟩 51A1	3A5				
	3A7				

CYP3A4 structures bound with
metyrapone
erythromycin

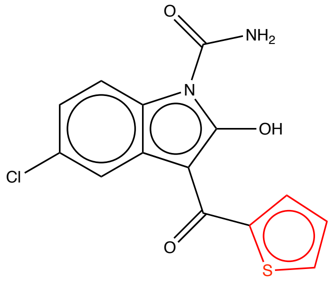


31-Jan-23 Table based on the work of Guengerich, In *Drug Metabolism Prediction*, Wiley-VCH 2014.

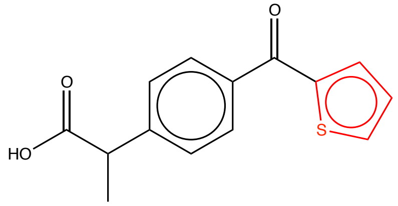
Thiophene is a safety risk



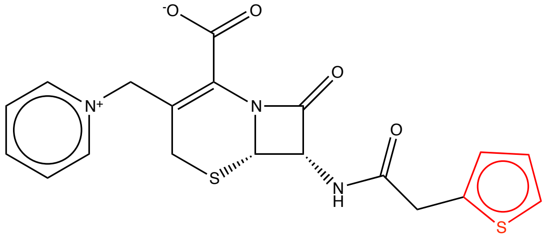
Tienilic acid
- idiosyncratic toxicity
- hepatotoxicity
- withdrawn after launch



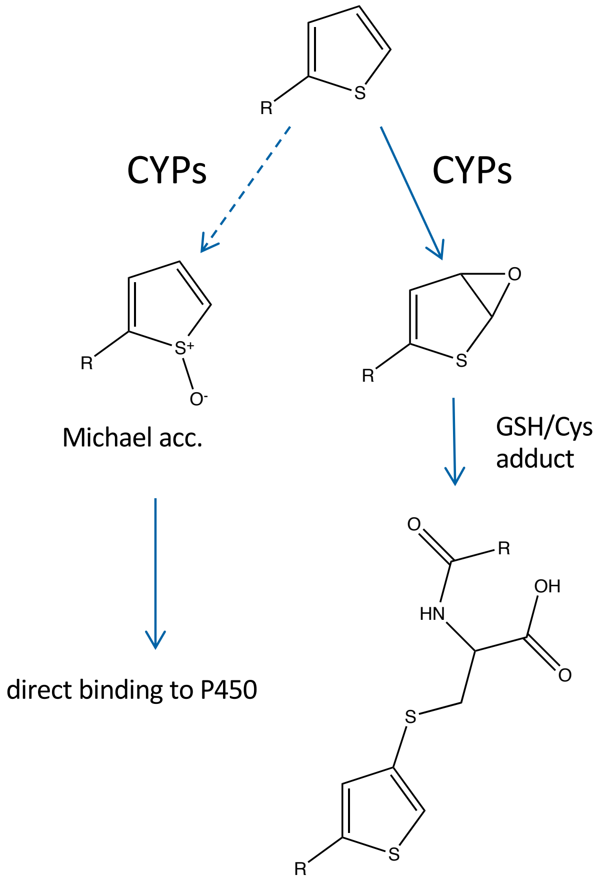
Tenidap
- hepatotoxicity
- immunotoxicity
- development discontinued



Suprofen
- idiosyncratic toxicity
- nephrotoxicity
- withdrawn after launch



Cephaloridine
- nephrotoxicity
- development discontinued



CYPs

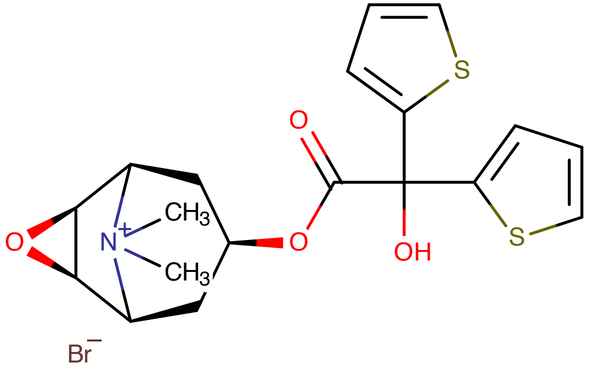
CYPs

Michael acc.

GSH/Cys adduct

direct binding to P450

As always, there are exceptions…

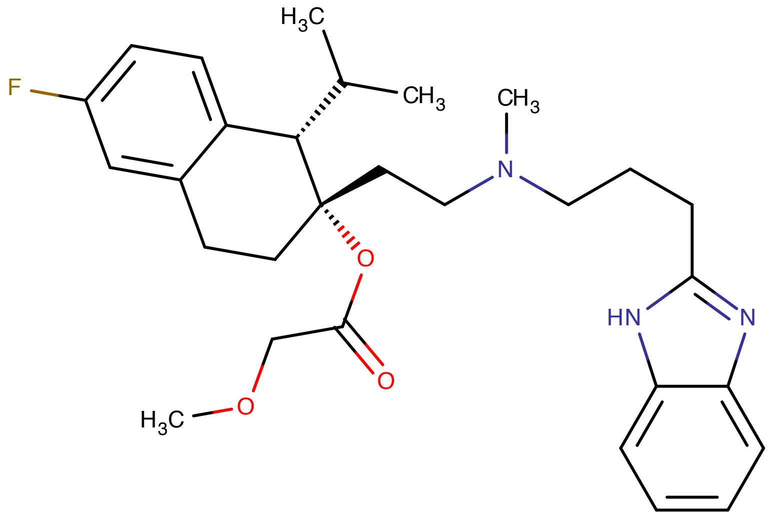


Tiotropium bromide:
no liver toxicity observed

What makes the difference?

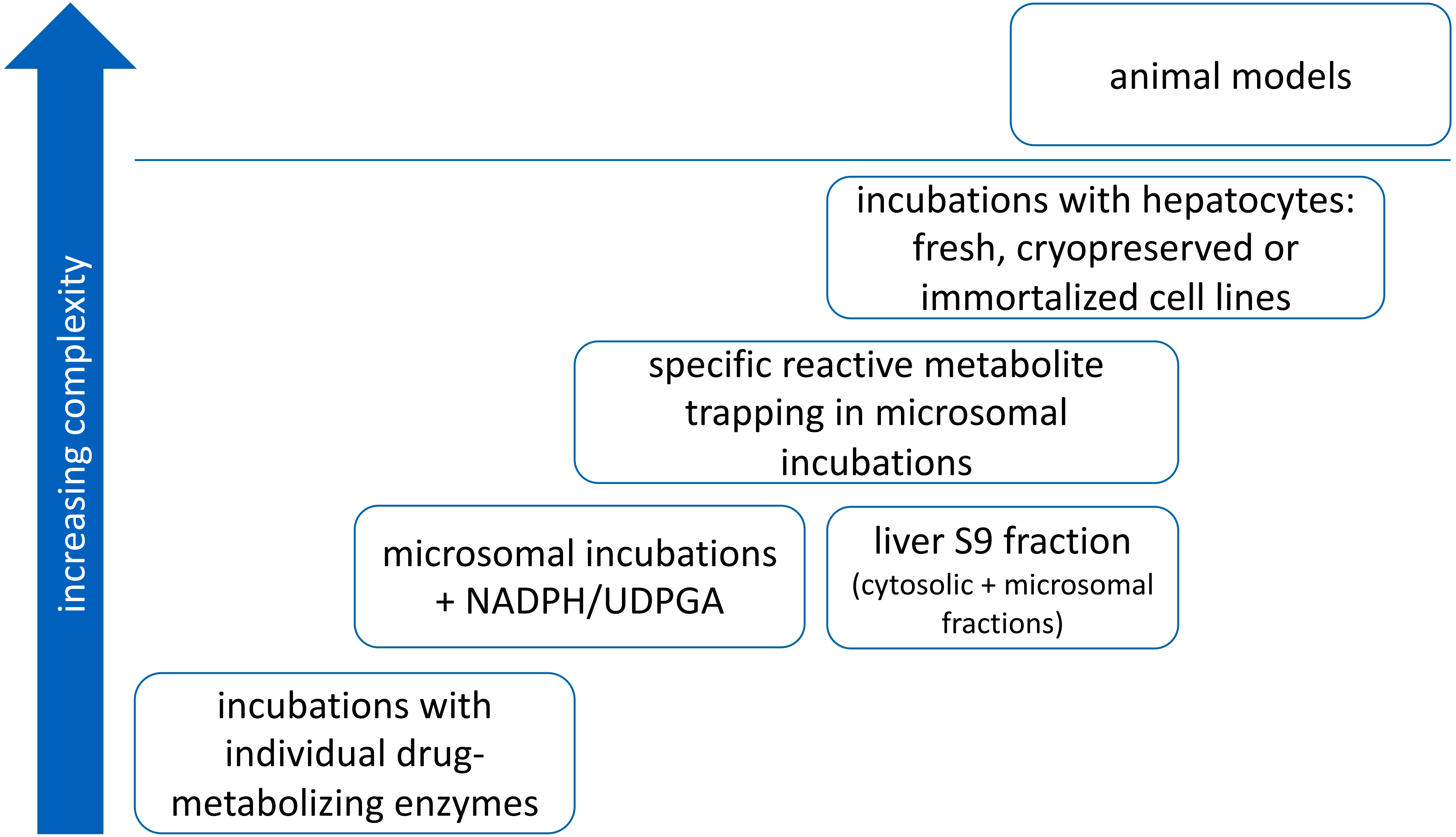
Drug-Drug Interactions (DDIs)

- Block/induction of a specific metabolic enzyme causes substantial (>10-fold) shift in pharmacokinetics of another drug
- Particularly problematic if a drug is metabolized via
  - a single enzyme
  - polymorphous enzymes (i.e. enzymes with genetic variants; e.g. CYP2D6, 2C19, and 2C9)
- Mibefradil
  - T-type Ca^{2+} channel blocker for treatment of hypertension
  - Withdrawn 1997 due drug-drug interactions with 3A4 substrates such as simvastatin
  - ~70% of CYP3A4 activity is lost in the first minute of incubation with mibefradil[1]



[1] Foti, Drug Metab Dispos 2011, doi:10.1124/dmd.111.038505

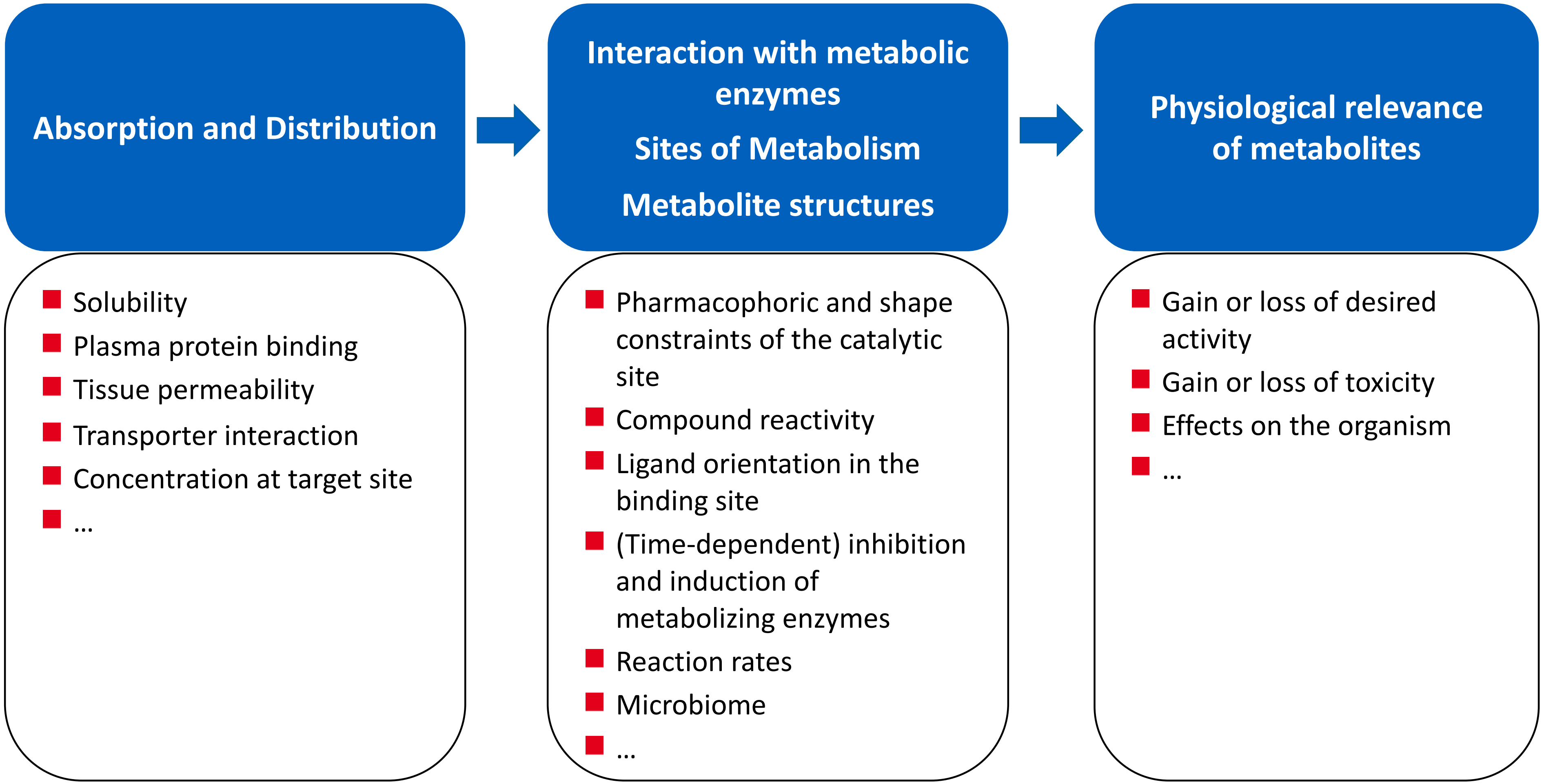
Modern analytical methods and biosystems for metabolism research are very powerful but resource-demanding



increasing complexity
animal models
incubations with hepatocytes: fresh, cryopreserved or immortalized cell lines
specific reactive metabolite trapping in microsomal incubations
microsomal incubations + NADPH/UDPGA
liver S9 fraction (cytosolic + microsomal fractions)
incubations with individual drug-metabolizing enzymes

Kirchmair et al., Nat Rev Drug Discov 2015, 14, 387-404. doi: 10.1038/nrd4581

Simulation of metabolism requires the consideration of many components but current *in silico* models consider only a single one or a few



Absorption and Distribution

- Solubility
- Plasma protein binding
- Tissue permeability
- Transporter interaction
- Concentration at target site
- …

Interaction with metabolic enzymes

Sites of Metabolism

Metabolite structures

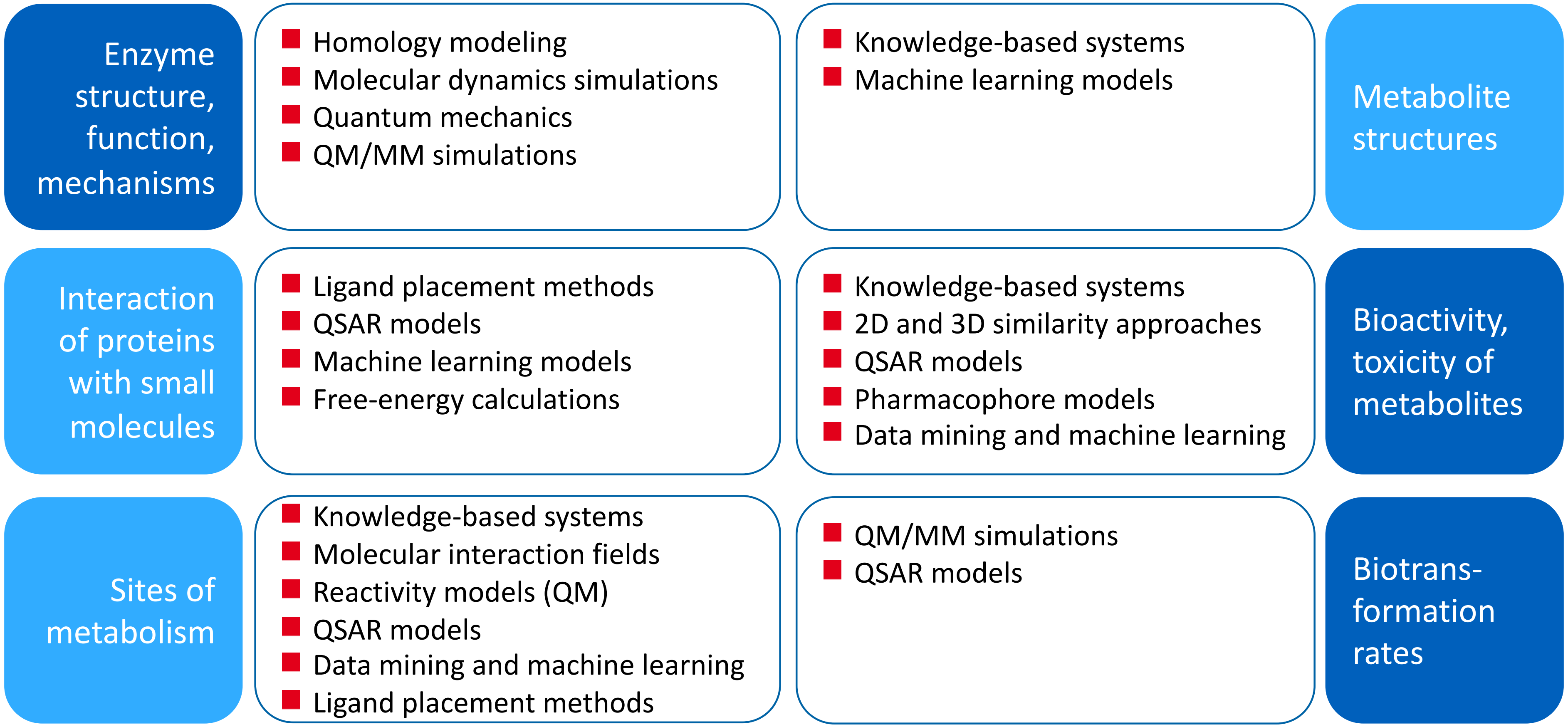
- Pharmacophoric and shape constraints of the catalytic site
- Compound reactivity
- Ligand orientation in the binding site
- (Time-dependent) inhibition and induction of metabolizing enzymes
- Reaction rates
- Microbiome
- …

Physiological relevance of metabolites

- Gain or loss of desired activity
- Gain or loss of toxicity
- Effects on the organism
- …

Kirchmair et al., Nat Rev Drug Discov 2015, 14, 387-404. doi: 10.1038/nrd4581

Computational approaches to the prediction of xenobiotic metabolism



Enzyme structure, function, mechanisms	- Homology modeling - Molecular dynamics simulations - Quantum mechanics - QM/MM simulations	- Knowledge-based systems - Machine learning models	Metabolite structures
Interaction of proteins with small molecules	- Ligand placement methods - QSAR models - Machine learning models - Free-energy calculations	- Knowledge-based systems - 2D and 3D similarity approaches - QSAR models - Pharmacophore models - Data mining and machine learning	Bioactivity, toxicity of metabolites
Sites of metabolism	- Knowledge-based systems - Molecular interaction fields - Reactivity models (QM) - QSAR models - Data mining and machine learning - Ligand placement methods	- QM/MM simulations - QSAR models	Biotrans-formation rates

Tyzack and Kirchmair, Chem Biol Drug Des 2019, 93, 377–386. doi: 10.1111/cbdd.13445
Kirchmair et al., Nat Rev Drug Discov 2015, 14, 387-404. doi: 10.1038/nrd4581
Kirchmair (Ed.), Methods and Principles in Medicinal Chemistry: Drug Metabolism Prediction. Wiley-VCH, 2014

Available data on xenobiotic metabolism

Data on	Resources
Interaction of small molecules with metabolizing enzymes	Zaretzki dataset ADMEDB (Fujitsu) BindingDB ChEMBL DrugBank (Univ. Alberta) MetraBase (Cambridge Univ.) PubChem SuperCyp (Charité)
Metabolites	EAWAG-BBD GOSTAR Drug Database (GVK BIO) HMDB KEGG MetaBase (MetaDrug) ~~Metabolite~~ METLIN MetXBioDB
Sites of metabolism (SoMs)	Zaretzki dataset MetaQSAR
Drug-drug interactions	DIDB (Drug Interaction Database)
Biomolecular structures of metabolic enzymes	PDB

Challenges and limitations:

Limited quantity and coverage

Limited comparability and relevance

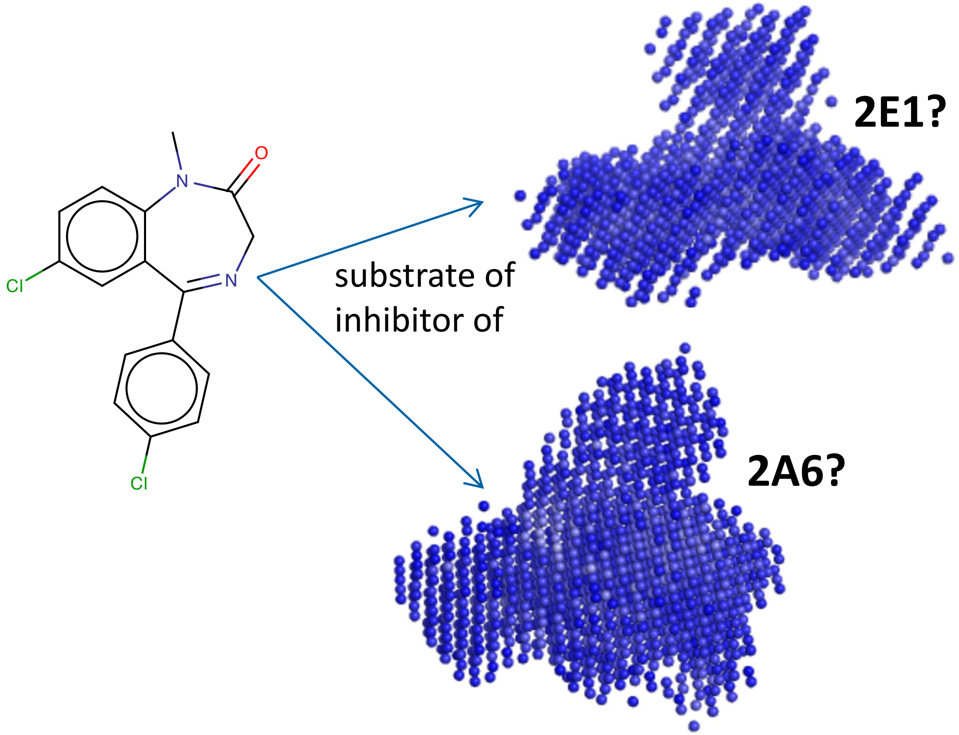
Incomplete, inaccurate, inconclusive

Not stored in a machine-readable format

~130k	Biotransformations
~1200	Parent molecules annotated with ~2000 metabolites
~700	Molecules with annotated SoMs (CYPs only)
~2300	Molecules with annotated SoMs (phase I and II)
~120	X-ray structures of CYPs

Q1: What metabolic enzymes is my small-molecule likely to interact with?

- Several good models available for predicting CYP inhibition and substrate selectivity
- Predictors dominated by **machine learning models**



substrate of
inhibitor of

2E1?

2A6?

+ Good classification accuracy within the applicability domain

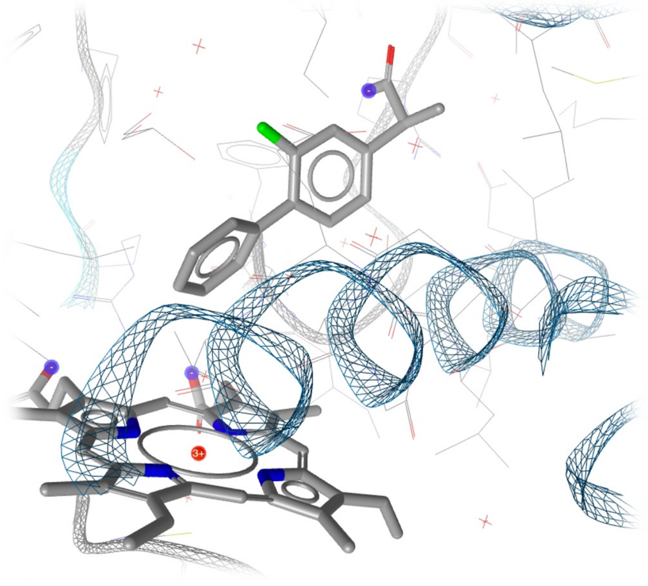
– Many models lack definition of applicability domain and indicators of prediction confidence

– Applicability domain quite narrow (due to lack of data for training)

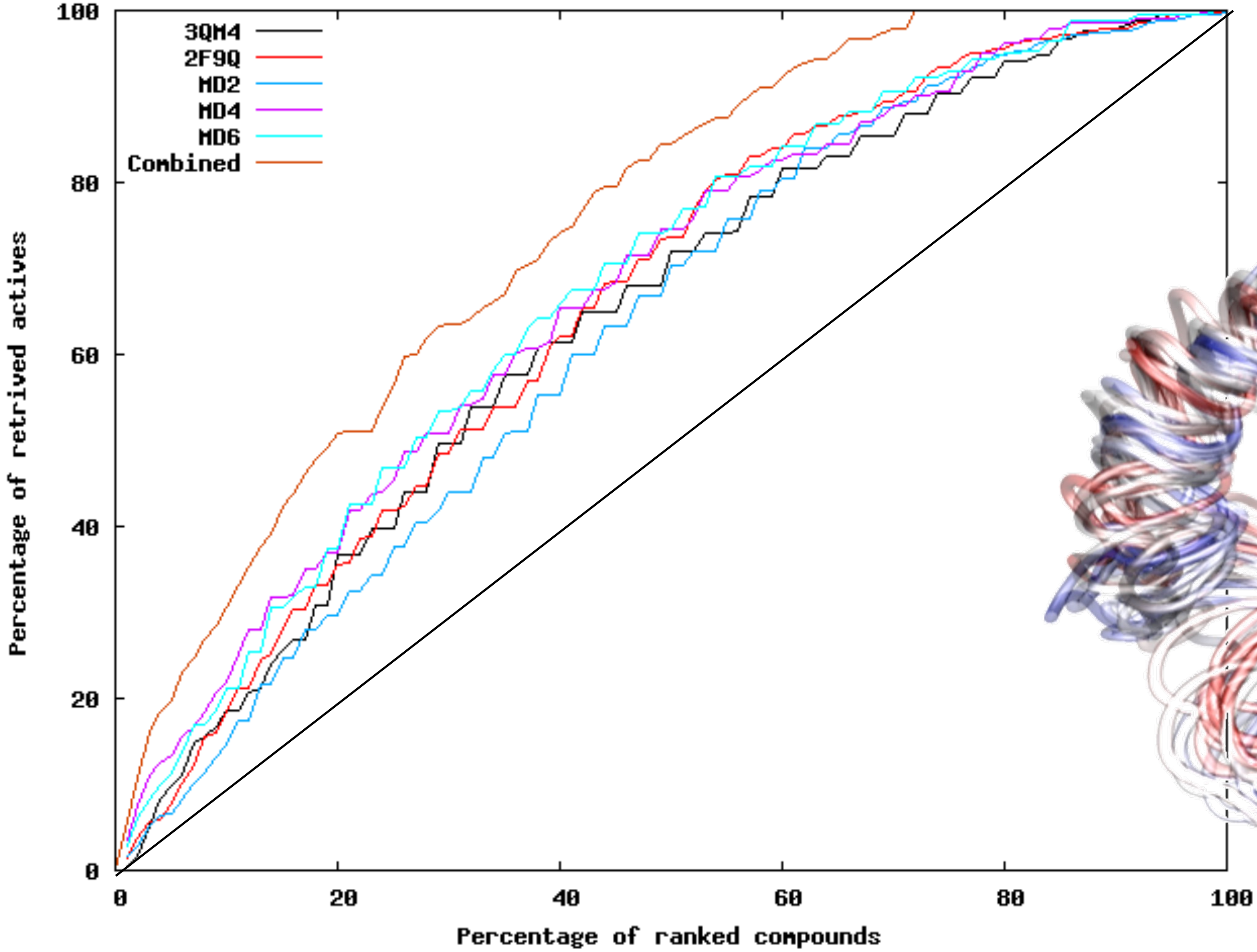
Name	Scope	Core components	Description	Licence	Exec.
VirtualToxLab (Biographics Laboratory 3R)	• Binder-nonbinder classification for 5 CYPs	Docking + QSAR	Uses flexible docking in combination with a multi-dimensional QSAR approach	Comm.	Local
Percepta P450 Specificity module (ACD/Labs)	• Substrate-nonsubstrate classification for 5 CYPs • Inhibitor-noninhibitor classification for 5 CYPs	PLS	Collection of models for predicting CYP inhibitors and substrates	Comm.	Local
ADMEWORKS Predictor (Fujitsu)	• Substrate-nonsubstrate and inhibitor-noninhibitor classification for 2 CYPs	Multiple linear regression	Collection of QSAR models for the prediction of K_i and K_m values	Comm.	Local
ADMET Predictor Metabolism module (Simulations Plus)	• Substrate-nonsubstrate classification for 9 CYPs • Inhibitor-noninhibitor classification for 5 CYPs	Artificial neural network ensemble	Predictor based on a large, curated data set. Also predicts K_m and V_{max} values for hydroxylation reactions, and Cl_{int} resulting from the action of 5 CYPs	Comm.	Local
WhichCYP	• Inhibitor-noninhibitor classification for 5 CYPs	SVM	Trained on the PubChem Bioassay 1851 dataset. AUCs between 0.88 and 0.95	Free	Web
SwissADME	• Inhibitor-noninhibitor classification for 5 CYPs	SVM	Trained on the PubChem Bioassay 1851 dataset. AUCs between 0.81 and 0.91	Free	Web
CypRules	• Inhibitor-noninhibitor classification for 5 CYPs	Decision trees	Trained on the PubChem Bioassay 1851 dataset. Classification accuracies > 90%	Free	Web
CYPlebrity	• **Inhibitor-noninhibitor classification for 5 CYPs**	**Random forest**	**Trained on PubChem Bioassay, ChEMBL and ADMEDB data. Trained on up to 18815 known inhibitors and noninhibitors. MCCs of up to 0.70.**	**Free**	**Web**
WhichP450 (Optibrium)	• Substrate-nonsubstrate classification for 7 CYPs	Multi-class random forest model	Trained on measured data for 465 compounds. Average AUC = 0.89 (5-fold CV)	Comm.	Local
CypReact	• Substrate-nonsubstrate classification for 9 CYPs	Machine learning	Trained on small dataset of approx. 1600 compounds	Free	Web
CYPstrate	• **Substrate-nonsubstrate classification for 9 CYPs**	**Random forest**	**Trained on approx. 1800 confirmed substrates and non-substrates. MCCs up to 0.85**	**Free**	**Web**

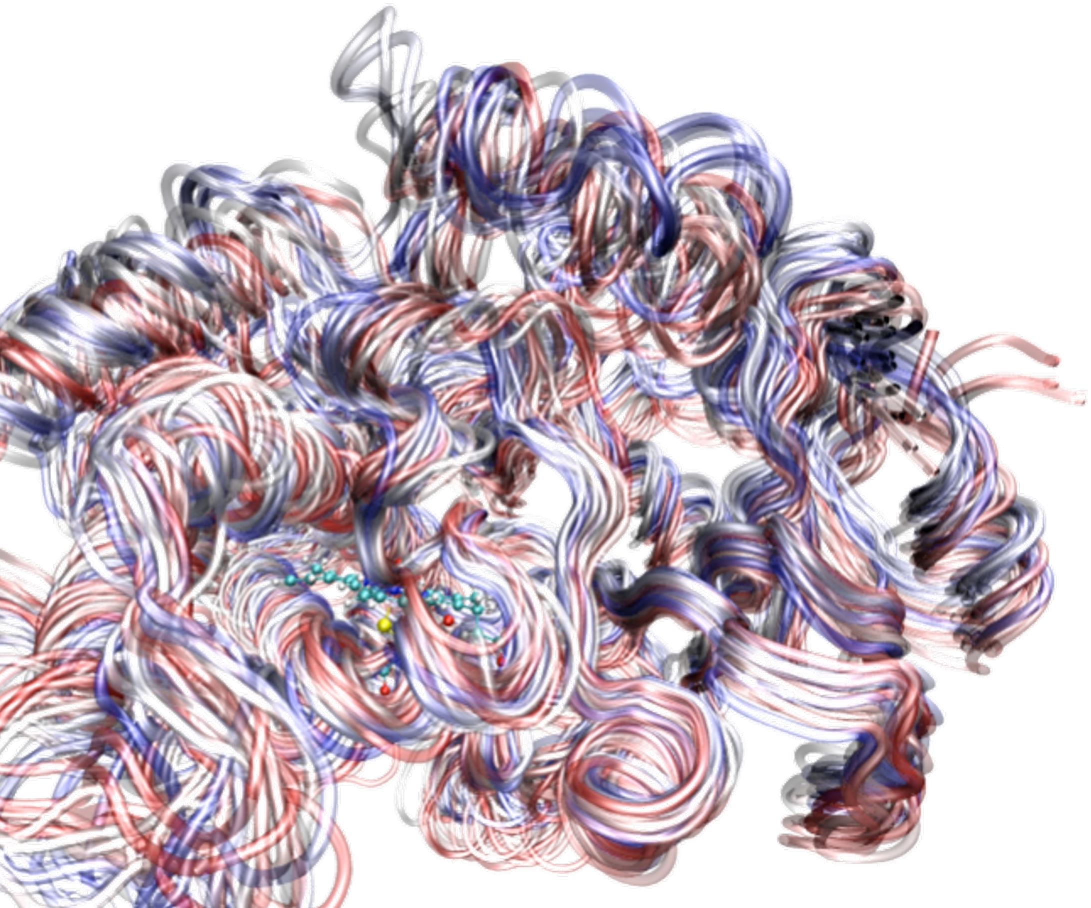
Ligand specificity prediction: Structure-based approaches

- Advantages
  - More insight into the orientation of a ligand at the binding site
  - Understand stereoselectivity in metabolism
- Disadvantages
  - The usual docking problems, but CYPs are particularly challenging because of protein flexibility and lack of a defined pharmacophore
  - Requires expert knowledge and only is usable with individual protein-ligand pairs



CYPs are highly malleable and promiscuous: docking approaches face challenges





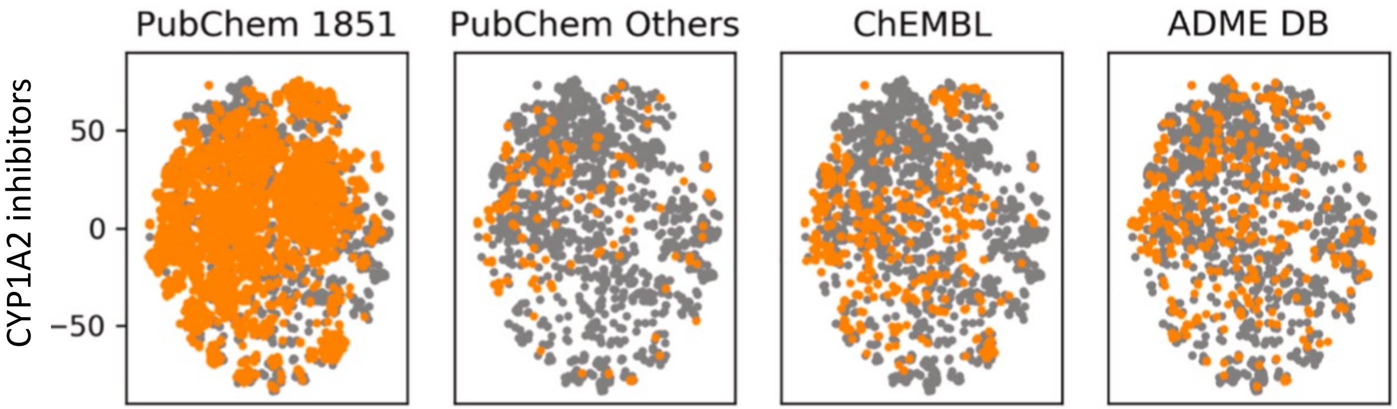
100 ns full-atom MD simulation courtesy of Mark J. Williamson

Martiny et al., Bioinformatics 2015, 31, 3930–3937. doi: 10.1093/bioinformatics/btv486

universität wien

CYPlebrity: Machine learning models for the prediction of CYP 1A2, 2C9, 2C19, 2D6 and 3A4 inhibition

Wojtek Plonka



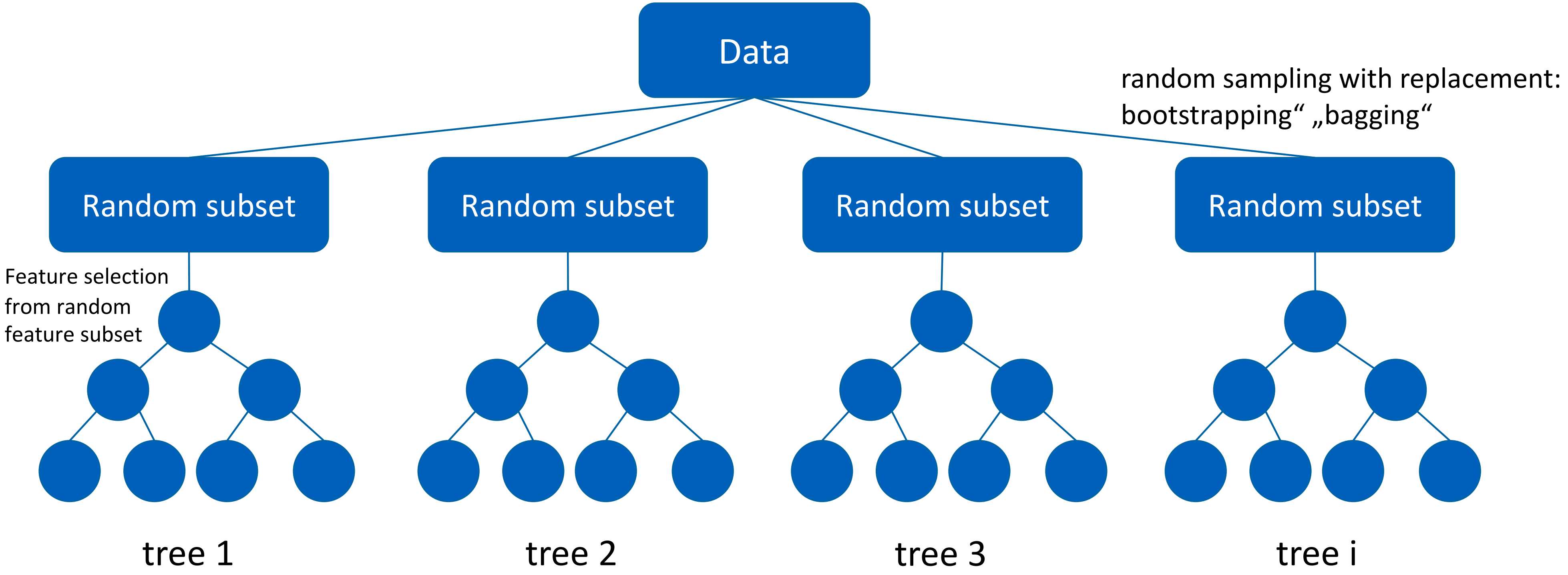
Approved drugs (from DrugBank)

CYP isozyme	Inhibitors total	Inhibitors exclusively from ADMEDB	Noninhibitors total
1A2	7391	693	7868
2C9	5033	741	9784
2C19	6235	534	8094
2D6	3711	708	12694
3A4	7763	1158	11052

Plonka et al., Bioorg Med Chem 2021, 46, 116388. doi: 10.1016/j.bmc.2021.116388

CYPlebrity: Machine learning models for the prediction of CYP 1A2, 2C9, 2C19, 2D6 and 3A4 inhibition

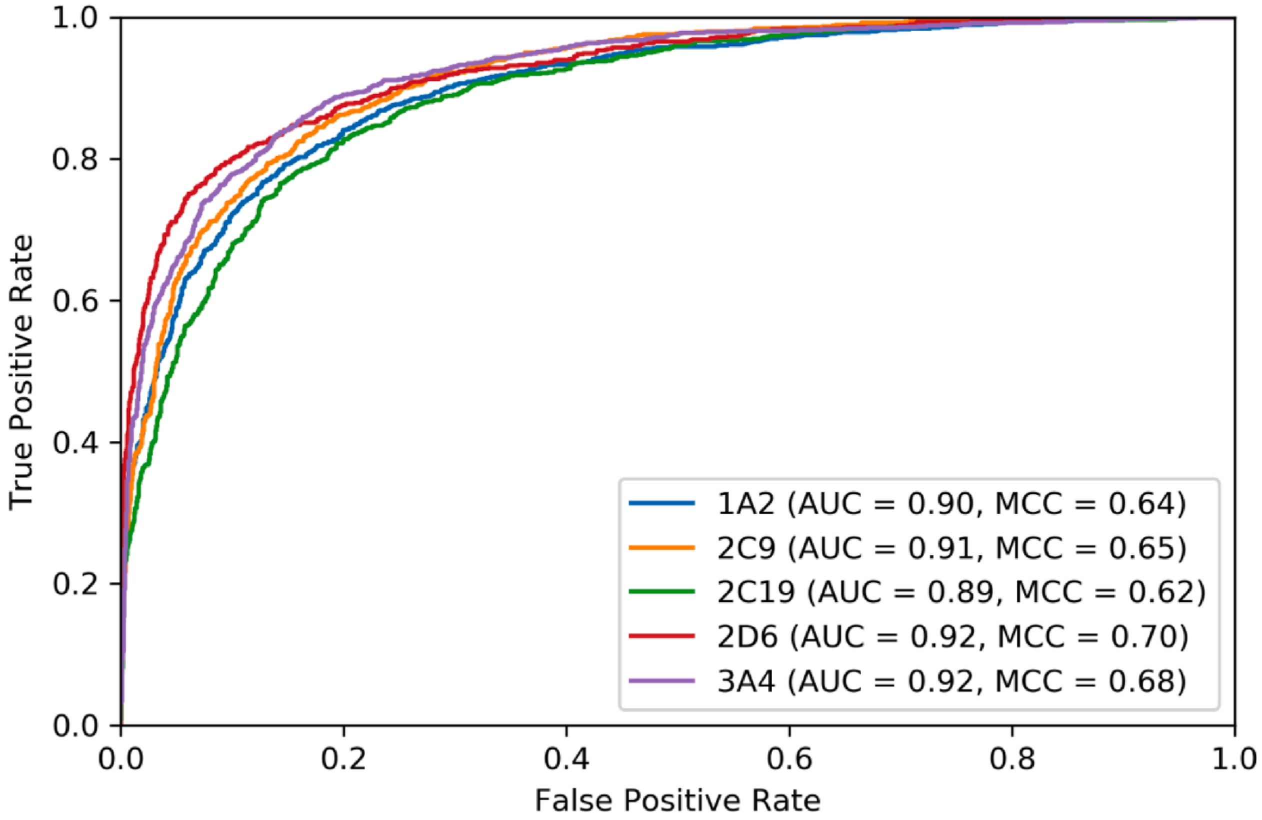
- Modeling approach:
  - Random forest
  - Morgan 3 fingerprints, 2048 bits (feature reduction method applied)

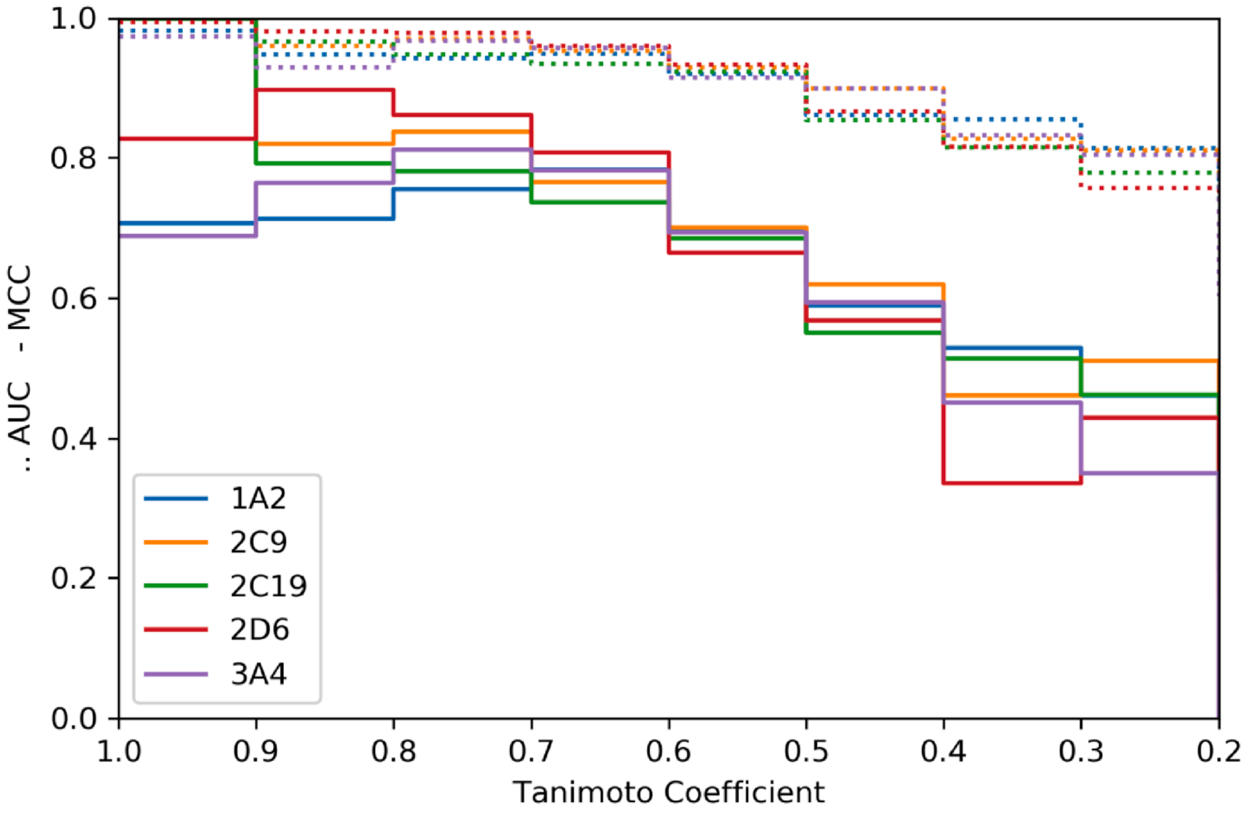


Plonka et al., Bioorg Med Chem 2021, 46, 116388. doi: 10.1016/j.bmc.2021.116388

CYPlebrity: Performance of the final models on an independent test set

- Modeling approach:
  - Random forest
  - Morgan 3 fingerprints, 2048 bits (feature reduction method applied)





Plonka et al., Bioorg Med Chem 2021, 46, 116388. doi: 10.1016/j.bmc.2021.116388

Q2: What atoms of my small molecule are susceptible to metabolism?

- Knowing the SoMs in a molecule can aid the derivation of likely metabolites and hence, optimisation strategies
- Models based on diverse approaches

+ Several good models available for CYPs, few for other metabolizing enzymes

+ Some models cover different mammalian species

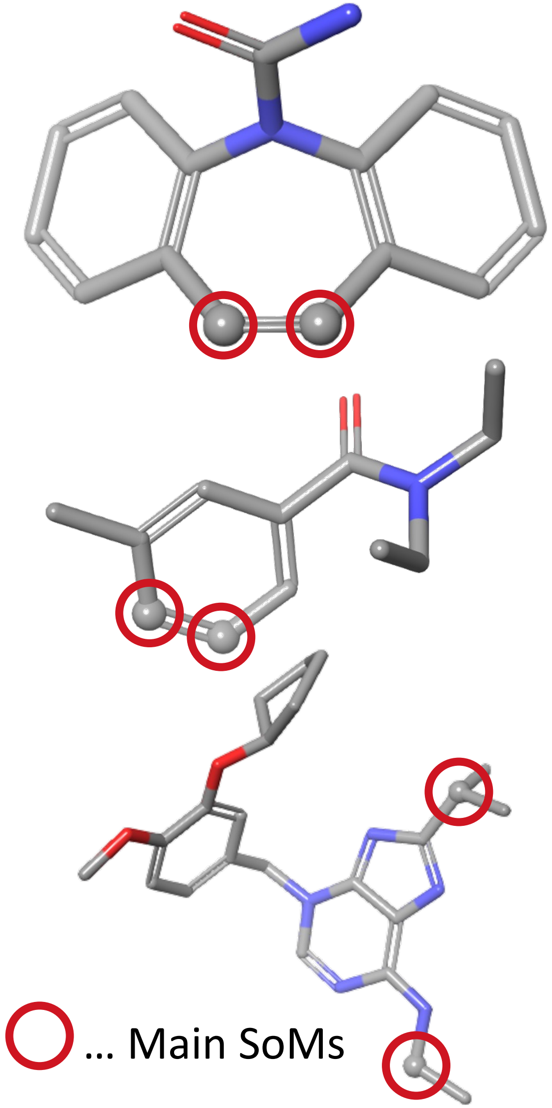
+ **Accuracy: At least one known SoM among the top-2 ranked atom positions in a molecule in >85% of all cases**

+ **Large applicability domain**

– Most models limited to CYPs

– Most models lack definition of applicability domain and error estimation

– Models able to discriminate major and minor metabolites at best



Prediction of sites of metabolism (SoMs) I

Name	Scope	Core components	Description	License	Exec.
MetaSite (Molecular Discovery)	CYPs and FMOs	Molecular interaction fields + reactivity model	Molecular interaction fields derived from protein structures plus molecular orbital calculations to identify likely SoMs	Comm.	Local
StarDrop P450 Metabolism Prediction (Optibrium)	3 CYPs	Reactivity model + ligand-based model	Combines quantum chemical analysis with a ligand-based model of CYP substrates to identify SoMs	Comm.	Local
ADMET Predictor Metabolism module (Simulations Plus)	3 CYPs	Artificial neural network ensemble	Derives likelihoods of metabolic reactions using artificial neural network ensembles on a large, curated dataset	Comm.	Local
Percepta P450 Regioselectivity module (ACD/Labs)	3 CYPs	Partial least squares	Global partial least squares-based QSAR model for calculating baseline regioselectivity; local corrections according to training data. Predicts and ranks major reaction types	Comm.	Local
P450 SoM Predictor (Schrödinger)	3 CYPs	Induced fit docking + reactivity model	Induced fit docking in combination with a quantum chemical model	Comm.	Local

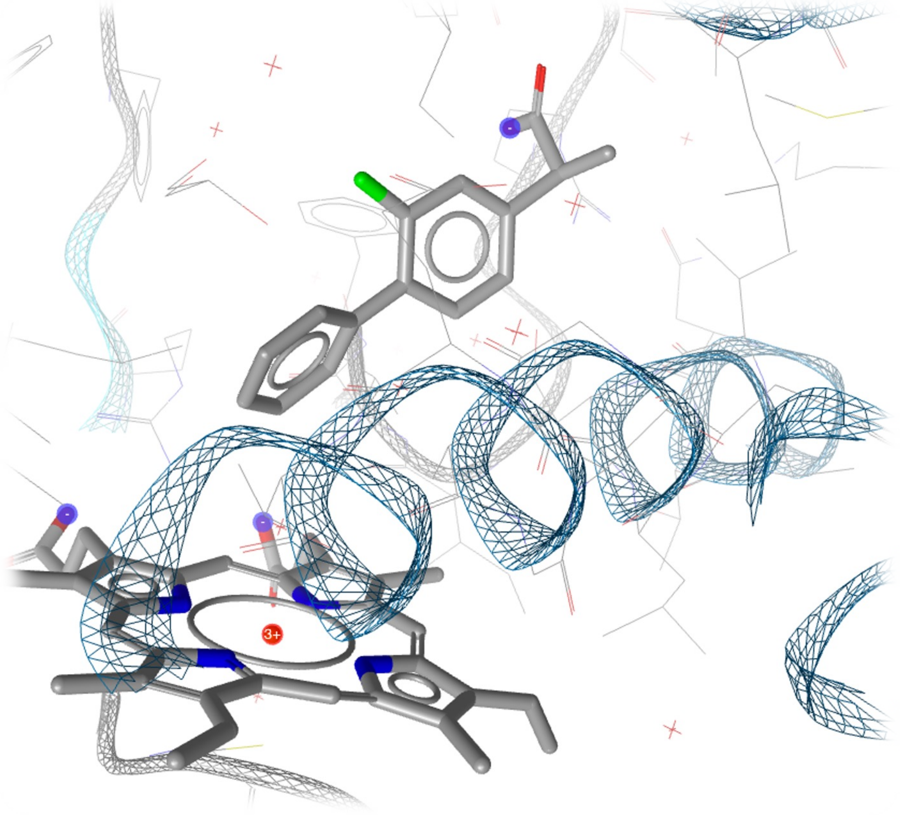
Prediction of sites of metabolism (SoMs) II

Name	Scope	Core components	Description	License	Exec.
~~MetaPrint2D~~	~~Any~~	~~Atom mapping + statistical model~~	~~Derives likelihoods of metabolic transformation for atoms with a defined atom environment by mining large biotransformation databases.~~	No longer available	
SMARTCyp	7 CYPs	Reactivity model derived from DFT calculations	Lookup table of DFT-derived activation energies for fragments	Free	Web, local
Xenosite	9 CYPs	Artificial neural network	Machine learning model for SoM prediction	Free	Web
SOMP	5 CYPs + UGTs	PASS algorithm	Combination of the PASS algorithm with labeled multilevel neighborhoods of atom (LMNA descriptors)	Free	Web
FAME (3rd generation)	**Any**	**Random forest**	**Machine learning model for SoM prediction**	**Free**	**Web, local**

Approaches to SoM prediction:

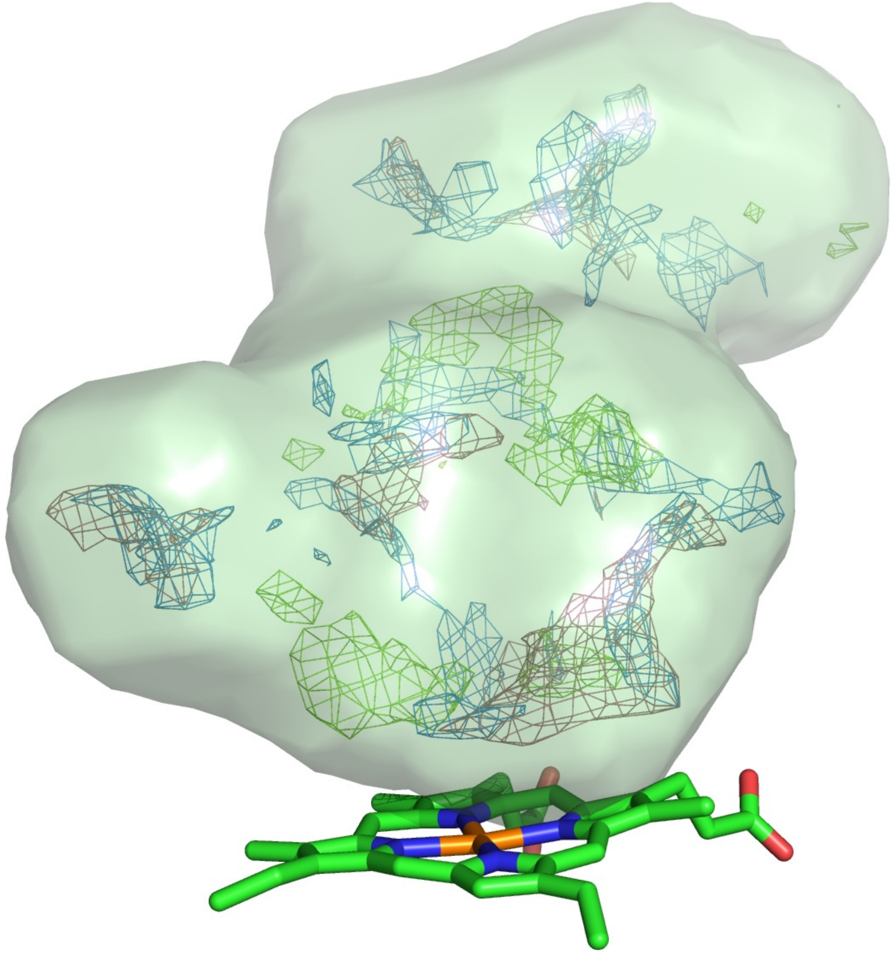
Structure-based approaches

- Focus on geometrical aspects
- Mostly automated ligand docking approaches
- Advantages
  - More insight into the orientation of a ligand at the binding site
  - Understand stereoselectivity in metabolism
- Limitations and challenges
  - The usual docking problems, but CYPs are particularly challenging because of protein flexibility and lack of a defined pharmacophore
  - No consideration of chemical reactivity
  - Requires expert knowledge and only is usable with individual protein-ligand pairs



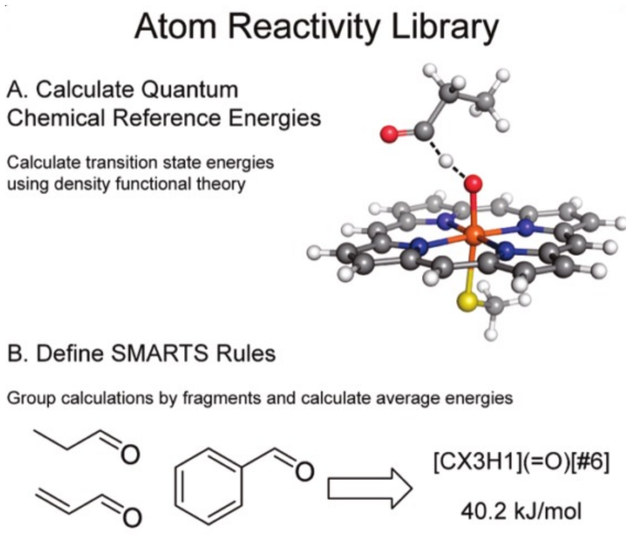
Approaches to SoM prediction:
Molecular interaction fields

- Structure-based approach
- Probes representing a specific chemical property (e.g. a carbor oxygen, representing H-bond acceptor functionality) are move a grid to identify favorable interaction spots and derive *grid mc*
- Consideration of side chain flexibility
- Usually combined with reactivity models



Reactivity models for SoM prediction

- Identification of SoMs based on reaction barriers (activation energies of carbon sites)
- SMARTCyp: Look-up table of hydrogen abstraction energies
- Usually combined with a method to take steric accessibility into account
- Advantages
    - Good accuracy
- Limitations and challenges
    - Limited coverage of reaction types and atom environments
    - No explicit consideration of protein structure



Atom Reactivity Library

A. Calculate Quantum Chemical Reference Energies

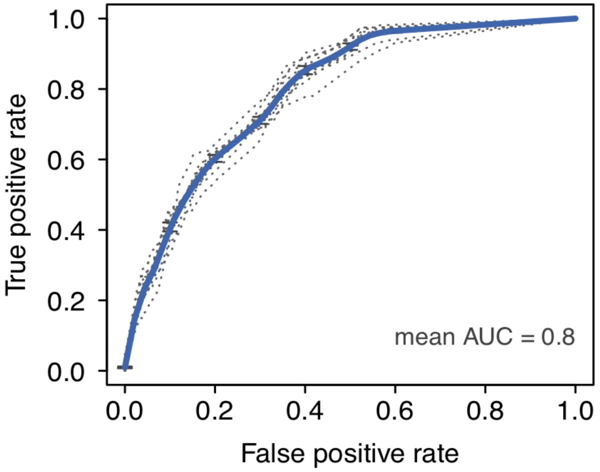
Calculate transition state energies using density functional theory

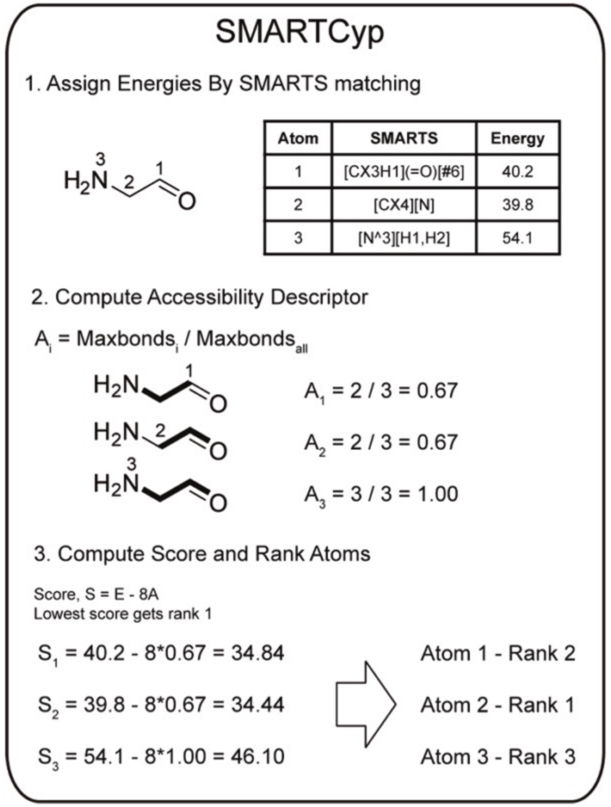
B. Define SMARTS Rules

Group calculations by fragments and calculate average energies

[CX3H1](=O)[#6]

40.2 kJ/mol





SMARTCyp

1. Assign Energies By SMARTS matching

Atom	SMARTS	Energy
1	[CX3H1](=O)[#6]	40.2
2	[CX4][N]	39.8
3	[N^3][H1,H2]	54.1

2. Compute Accessibility Descriptor

A_i = Maxbonds$_i$ / Maxbonds$_{all}$

$A_1 = 2 / 3 = 0.67$

$A_2 = 2 / 3 = 0.67$

$A_3 = 3 / 3 = 1.00$

3. Compute Score and Rank Atoms

Score, S = E - 8A
Lowest score gets rank 1

$S_1 = 40.2 - 8*0.67 = 34.84$

$S_2 = 39.8 - 8*0.67 = 34.44$

$S_3 = 54.1 - 8*1.00 = 46.10$

Atom 1 - Rank 2

Atom 2 - Rank 1

Atom 3 - Rank 3

Rydberg et al., ACS Med. Chem. Lett. 2010, 1, 96-100. doi: 10.1021/ml100016x

Development of FAst MEtabolizer (FAME)

Martin Sicho

UNIL | Université de Lausanne

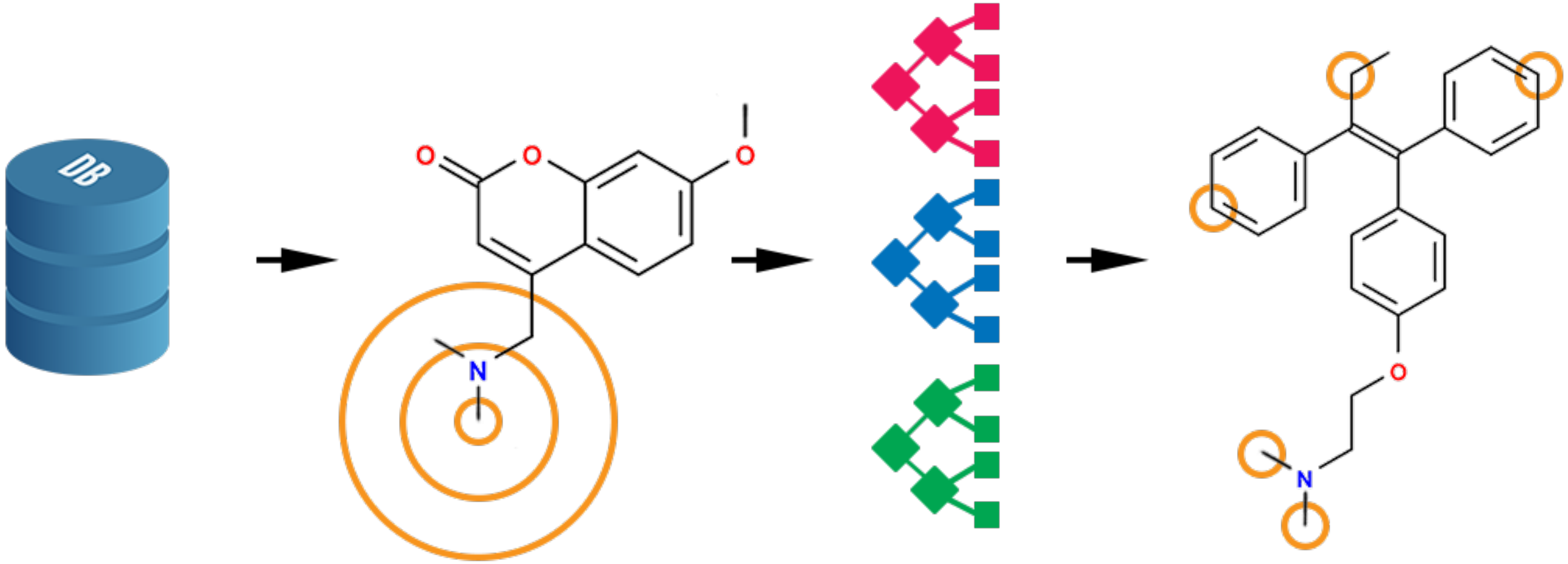
UNIVERSITÀ DEGLI STUDI DI MILANO

MetaQSAR Database

Atom environment encoding

Machine learning

Site of metabolism prediction



Sicho et al., J Chem Inf Model 2019, 59, 3400-3412. doi: 10.1021/acs.jcim.9b00376

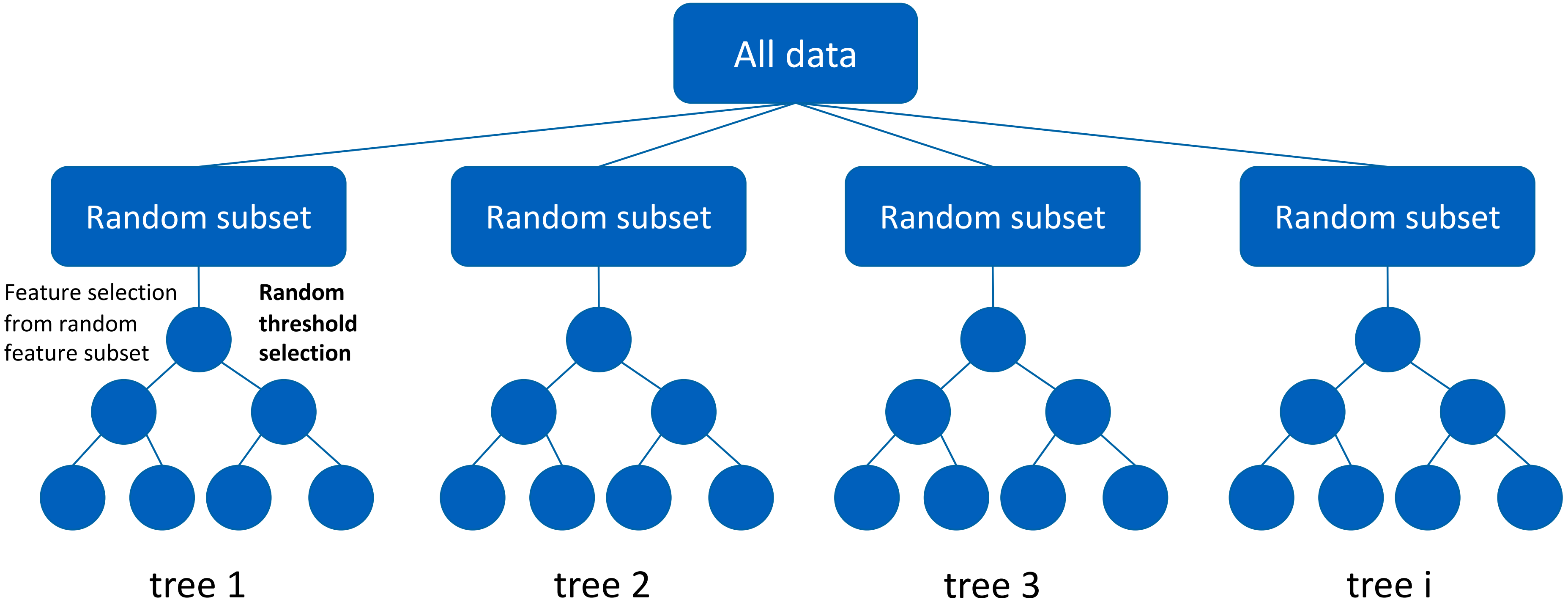
FAst Metabolizer (FAME)

	FAME 1 (2013)	FAME 2 (2017)	FAME 3 (2019)
Training set source	Metabolite DB (proprietary, discontinued)	Zaretzki Dataset	MetaQSAR DB
Training set size	Up to ~21,000 substrates	Up to ~540 substrates	Up to ~2150 substrates
CYP P450 enzymes	Yes	Yes	Yes
Phase 1 metabolism	Yes	CYPs only	Yes
Phase 2 metabolism	Yes	No	Yes
SoM quality	Automated assignment based on substructure matching	Expert-curated but some quality issues	**Expert-curated**
Machine learning approach	Random forest	Extremely randomized trees	
Descriptors	15 2D-descriptors including Sybyl atom types	Circular fingerprints encoding Sybyl atom types plus 15 2D-descriptors	
Applicability domain definition and error estimation	No	No	**Yes**
Prediction accuracy	Mediocre	High	High
Availability	Discontinued	Software package	Software package and web service

Sicho et al., J Chem Inf Model 2019, 59, 3400-3412. doi: 10.1021/acs.jcim.9b00376

FAME 3: Model development

- MetaQSAR database split into training set (80%) and test set (20%)
- Four different sets of descriptors (ATF, CDK, circCDK and QC) explored
- Feature reduction down to max. of 400 by ANOVA F-Test
- Model generation: Extremely randomized trees
- Hyperparameters derived by grid search with 10-fold cross-validation

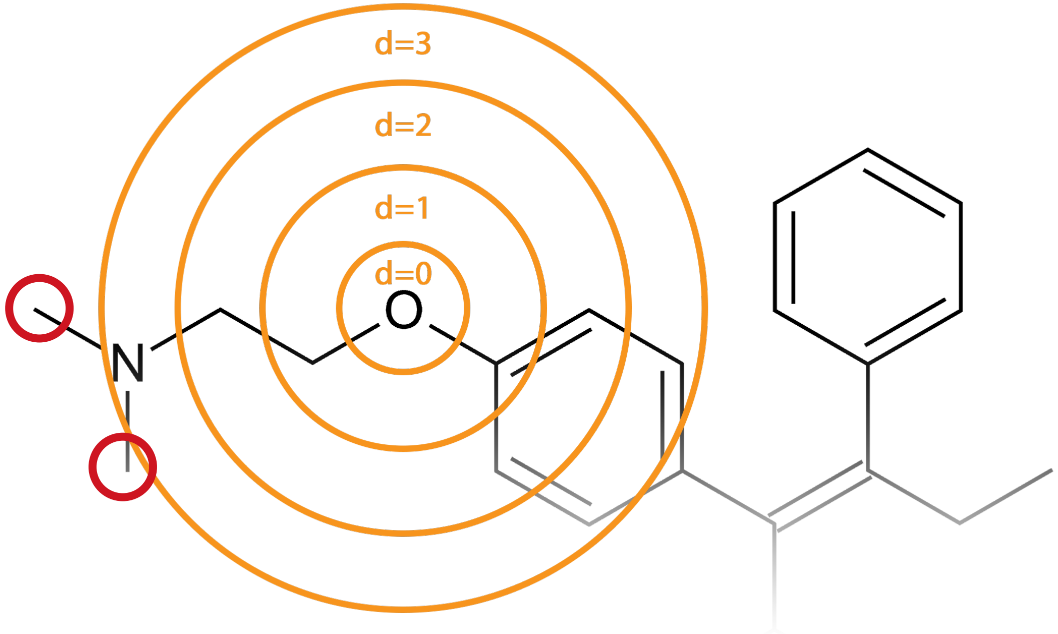


Sicho et al., J Chem Inf Model 2019, 59, 3400-3412. doi: 10.1021/acs.jcim.9b00376

FAME 3: Atom descriptors

- Four sets of descriptors have been explored
- Combination of ATFs with circCDK descriptors identified as most suitable descriptors set

Acronym	Description
ATF	Circular fingerprint based on Sybyl atom types
CDK	15 Basic 2D descriptors implemented in CDK
circCDK	Circular descriptors derived from the CDK descriptor set
QC	10 AM1-based descriptors calculated with MOPAC



Atom	AtomType_O.3_0	...	AtomType_C.ar_3	AtomType_N.3_3	...
...	...	...	...	...	...
O	1	...	2	1	...
...	...	...	...	...	...

fingerprints

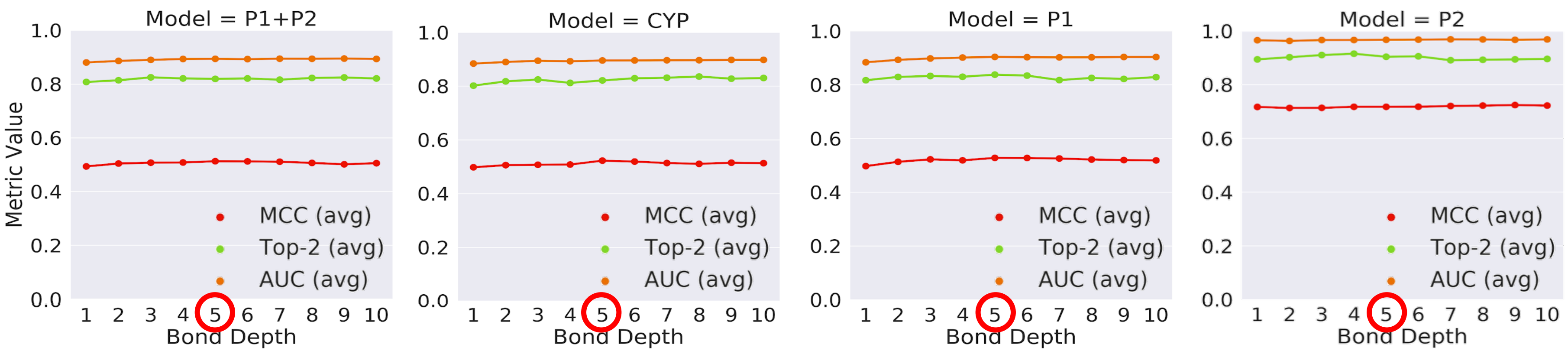
Atom	sigmaEng_O.3_0	...	sigmaEng_C.ar_3	sigmaEng_N.3_3	...
...	...	...	...	...	...
O	9.91	...	8.31	8.35	...
...	...	...	...	...	...

descriptors

Sicho et al., J Chem Inf Model 2019, 59, 3400-3412. doi: 10.1021/acs.jcim.9b00376

FAME 3: Performance of “circCDK+ATF” models

10-fold cross-validation



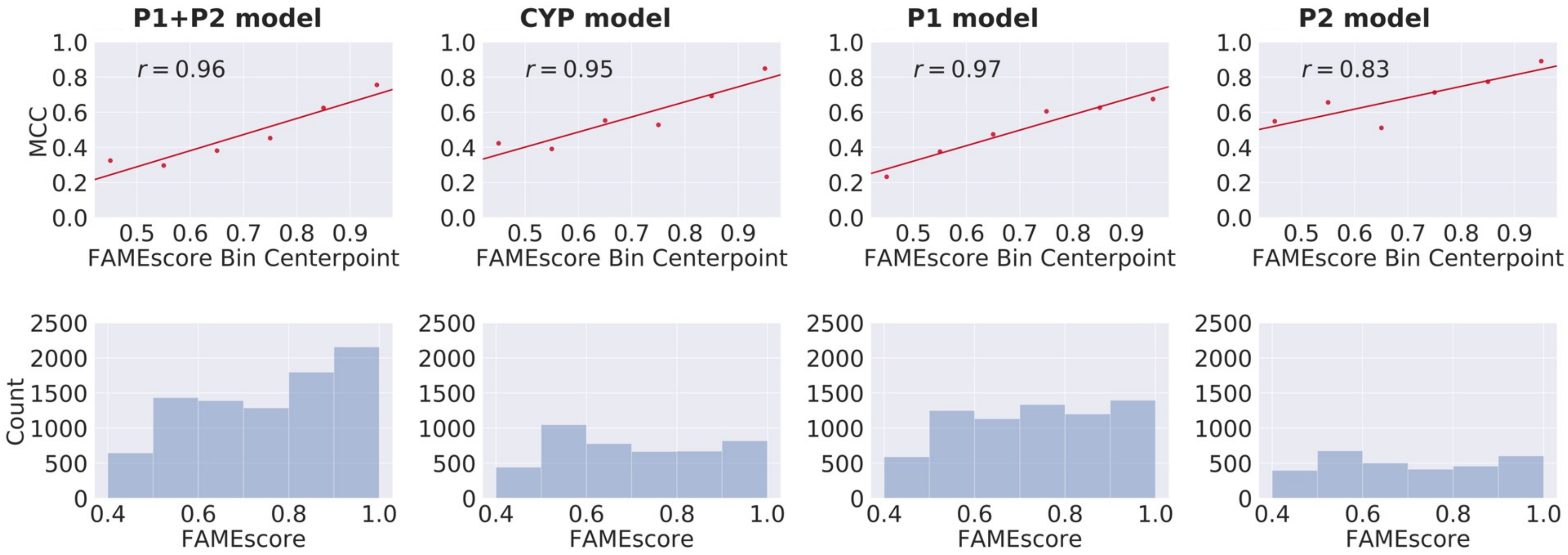
test on holdout data

bond depth=5

Model	MCC	AUC	Top-2
P1+P2	0.50	0.90	82%
P1+P2 100+	0.55	0.92	87%
CYP	0.57	0.92	90%
CYP 100+	0.63	0.94	86%
P1	0.53	0.88	83%
P1 100+	0.52	0.92	80%
P2	0.71	0.97	92%
P2 100+	0.75	0.97	91%

Sicho et al., J Chem Inf Model 2019, 59, 3400-3412. doi: 10.1021/acs.jcim.9b00376

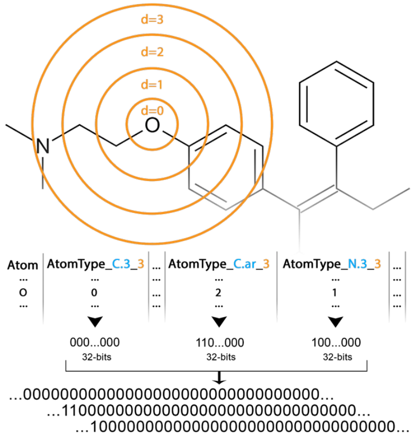
FAME 3: Performance of the final models on holdout data



$$FAMEscore = 1 - \frac{\sum_{i=1}^{k} d_i}{k}$$

d… distance (Tanimoto coefficient)

k… number of nearest neighbours (we use *k*=3)



Atom	AtomType_C.3_3	…	AtomType_C.ar_3	AtomType_N.3_3	…
…	…	…	…	…	…
O	0	…	2	1	…
…	…	…	…	…	…

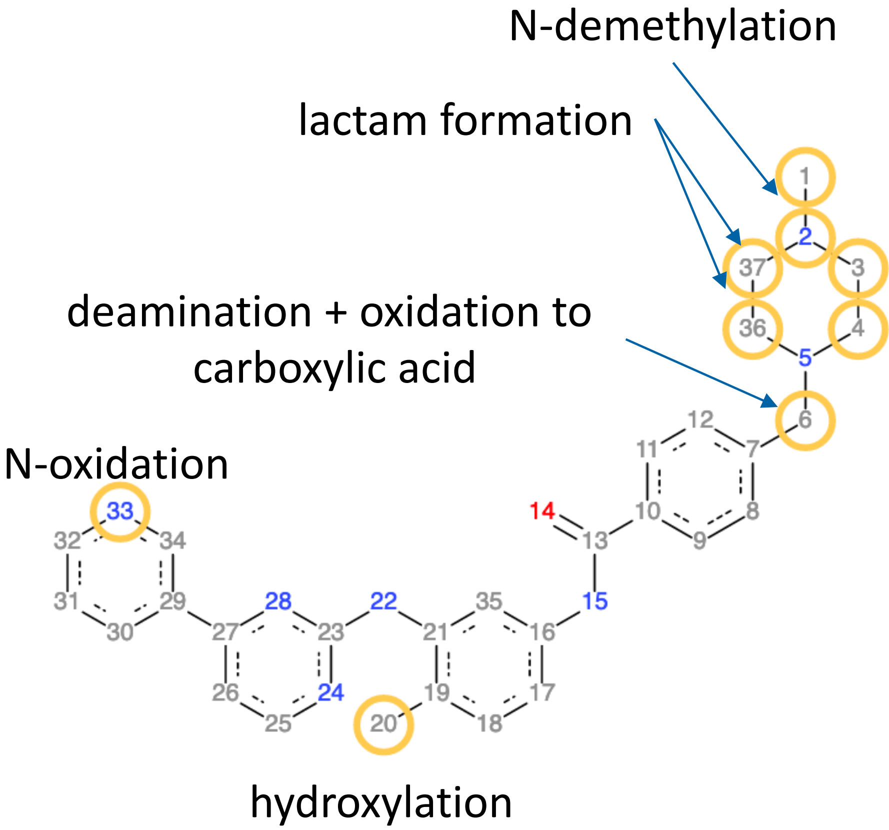
000…000 32-bits

110…000 32-bits

100…000 32-bits

…00000000000000000000000000000000…
…11000000000000000000000000000000…
…10000000000000000000000000000000…

FAME 3: Prediction of the sites of metabolism of imatinib



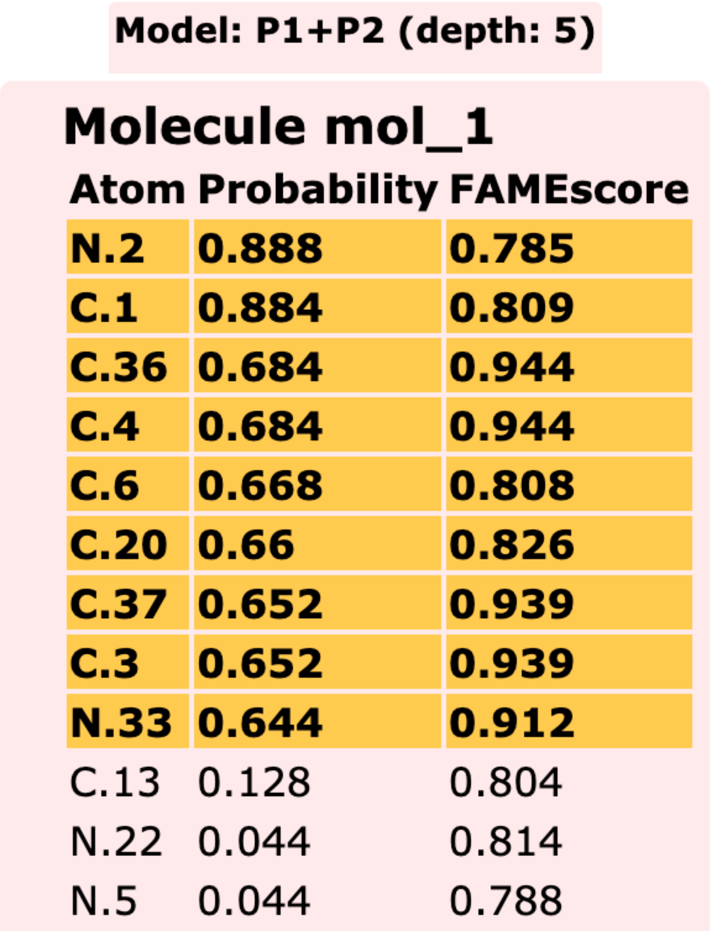
N-demethylation

lactam formation

deamination + oxidation to carboxylic acid

N-oxidation

hydroxylation



Model: P1+P2 (depth: 5)

Molecule mol_1

Atom	Probability	FAMEscore
N.2	**0.888**	**0.785**
C.1	**0.884**	**0.809**
C.36	**0.684**	**0.944**
C.4	**0.684**	**0.944**
C.6	**0.668**	**0.808**
C.20	**0.66**	**0.826**
C.37	**0.652**	**0.939**
C.3	**0.652**	**0.939**
N.33	**0.644**	**0.912**
C.13	0.128	0.804
N.22	0.044	0.814
N.5	0.044	0.788

Q3: What are the likely metabolites of my compound?

- **Dominated by rule-based (expert) systems**
- Include knowledge-bases that are enormously useful for the interpretation of predictions
- Increasingly combined with site-of-metabolism prediction models
- **Latest development: transformers** trained on chemical reaction data and fine-tuned on metabolic reaction data[1]

+ Several good models available for phase I and II metabolism (mostly commercial)

+ Several models cover different (mammalian) species

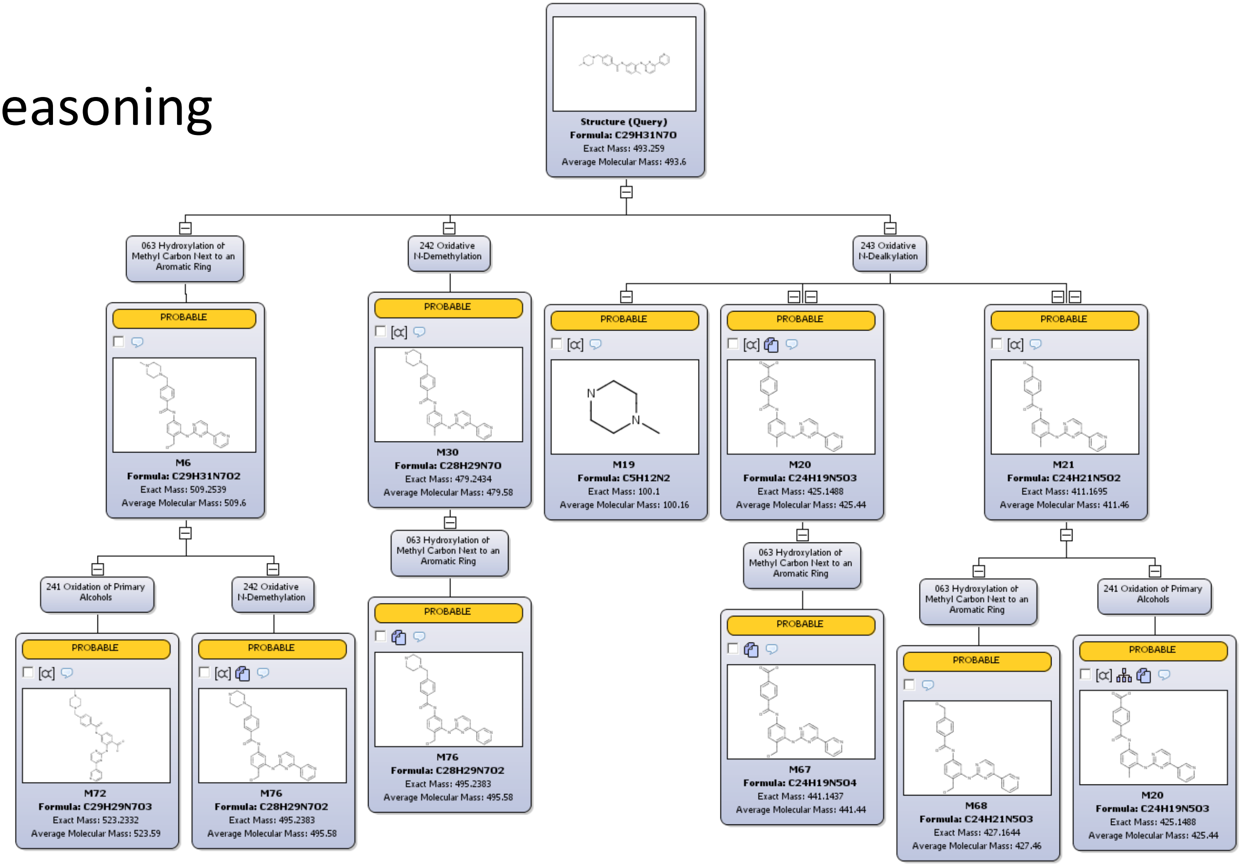
– Limited accuracy: very high number of predicted metabolites

– Ranking the likelihood of metabolites is a major challenge and bottleneck

[1] Litsa et al., Chemical Science 2020, doi: 10.1039/D0SC02639E.

Prediction of metabolite structures: Expert-curated biotransformation dictionaries (expert systems/knowledge-based systems)

- A set of (expert-) curated biotransformation rules (“Dictionary”) is applied to predict likely metabolites
  - Rules encode fragments and their associated biotransformations
  - Transformations are applied to any molecules containing any such fragments
- Advantages
  - Knowledge base provides rational basis for reasoning
  - Emulation of an expert panel
- Limitations and challenges
  - **Combinatorial explosion problem:** Very large number of metabolites may be generated → increasingly combined with other approaches in an attempt to overcome this problem
  - Metabolite ranking is insufficient
  - Lack of effective visualization
- Leading software: Derek Nexus (Lhasa Ltd.)



Structure (Query)
Formula: $C_{29}H_{31}N_7O$
Exact Mass: 493.259
Average Molecular Mass: 493.6

063 Hydroxylation of Methyl Carbon Next to an Aromatic Ring

PROBABLE
M6
Formula: $C_{29}H_{31}N_7O_2$
Exact Mass: 509.2539
Average Molecular Mass: 509.6

241 Oxidation of Primary Alcohols

PROBABLE
M72
Formula: $C_{29}H_{29}N_7O_3$
Exact Mass: 523.2332
Average Molecular Mass: 523.59

242 Oxidative N-Demethylation

PROBABLE
M76
Formula: $C_{28}H_{29}N_7O_2$
Exact Mass: 495.2383
Average Molecular Mass: 495.58

242 Oxidative N-Demethylation

PROBABLE
M30
Formula: $C_{28}H_{29}N_7O$
Exact Mass: 479.2434
Average Molecular Mass: 479.58

063 Hydroxylation of Methyl Carbon Next to an Aromatic Ring

PROBABLE
M76
Formula: $C_{28}H_{29}N_7O_2$
Exact Mass: 495.2383
Average Molecular Mass: 495.58

243 Oxidative N-Dealkylation

PROBABLE
M19
Formula: $C_5H_{12}N_2$
Exact Mass: 100.1
Average Molecular Mass: 100.16

PROBABLE
M20
Formula: $C_{24}H_{19}N_5O_3$
Exact Mass: 425.1488
Average Molecular Mass: 425.44

063 Hydroxylation of Methyl Carbon Next to an Aromatic Ring

PROBABLE
M67
Formula: $C_{24}H_{19}N_5O_4$
Exact Mass: 441.1437
Average Molecular Mass: 441.44

PROBABLE
M21
Formula: $C_{24}H_{21}N_5O_2$
Exact Mass: 411.1695
Average Molecular Mass: 411.46

063 Hydroxylation of Methyl Carbon Next to an Aromatic Ring

PROBABLE
M68
Formula: $C_{24}H_{21}N_5O_3$
Exact Mass: 427.1644
Average Molecular Mass: 427.46

241 Oxidation of Primary Alcohols

PROBABLE
M20
Formula: $C_{24}H_{19}N_5O_3$
Exact Mass: 425.1488
Average Molecular Mass: 425.44

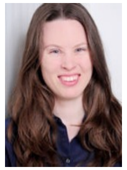
Prediction of metabolite structures I

Name	Coverage	Core components	Description	License	Exec.
Meteor Nexus (Lhasa)	Any	Knowledge-based system + SoM predictor	Contains three different methodologies for assessing the likelihood of metabolites. Toxicity of metabolites can be directly assessed	Comm.	Local
TIMES (LMC, Oasis)	Any	Knowledge-based system	Utilizes a biotransformation library and a heuristic algorithm to generate metabolic maps	Comm.	Local
MetaSite (Molecular Discovery)	CYPs and FMOs	Molecular interaction fields	Produces a comprehensive set of likely metabolites from a set of metabolic reactions. Connection to Mass-MetaSite for Metabolite-ID	Comm.	Local
MetaDrug (Thomson Reuters)	Any	Knowledge-based system	Generates metabolites from a biotransformation dictionary. Toxicity of metabolites can be directly assessed	Comm.	Web
SyGMa	Any	Rule-based system	Generates structures of likely metabolites based on rules derived from Biovia’s Metabolite database	Free	Local

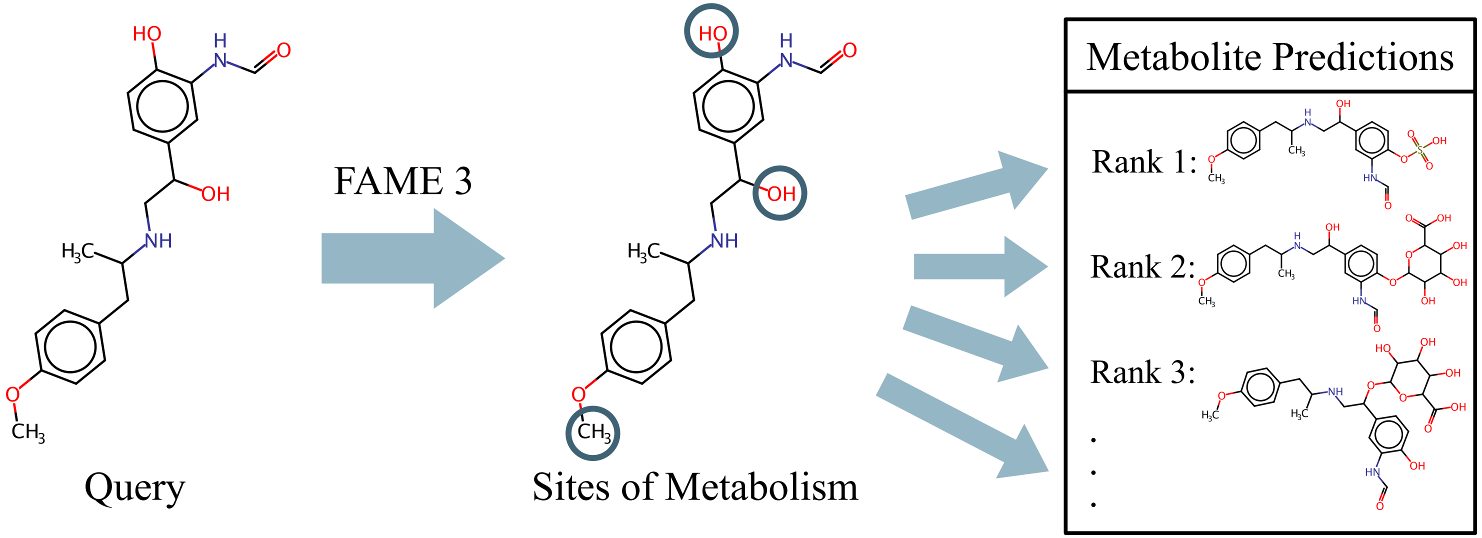
Prediction of metabolite structures II

Name	Coverage	Core components	Description	License	Exec.
EAWAG-BBD Pathway Prediction System	Any	Knowledge-based system	Rule-based system specialized in microbial catabolic metabolism of environmental pollutants. Classification of metabolites with respect to their likelihood	Free	Web
~~MetaPrint2D-React~~	~~Any~~	~~Atom mapping + statistical model~~	~~Generates structures of likely metabolites based on the MetaPrint2D data mining approach~~	~~Free~~	~~No longer available~~
SMARTCyp + Toxtree	7 CYPs	SMARTCyp + rule-based system	Uses a set of rules to generate metabolites on sites of metabolism predicted by SMARTCyp	Free	Local
OECD Toolbox	Liver metab.	Rule-based approach similar to the one implemented in TIMES	Various different models for predicting likely metabolites	Free	Local
GLORYx	**Any**	**Rule-based approach**	**Combines SOM prediction with rule-based metabolite prediction for enhanced metabolite ranking**	**Free for academic use**	**Web and local**
MetaTrans	**Any**	**Deep learning transformer approach**	**Trained on chemical reaction data and fine-tuned on metabolism data**	**Free**	**Local**

GLORYx: Predictor of likely metabolites



Christina
de Bruyn Kops



Query

FAME 3

Sites of Metabolism

Metabolite Predictions

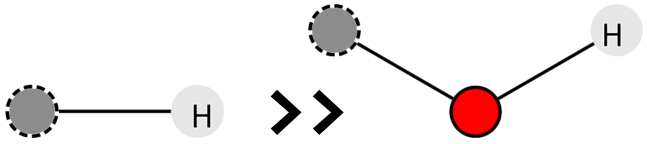
Rank 1:

Rank 2:

Rank 3:

.
.
.

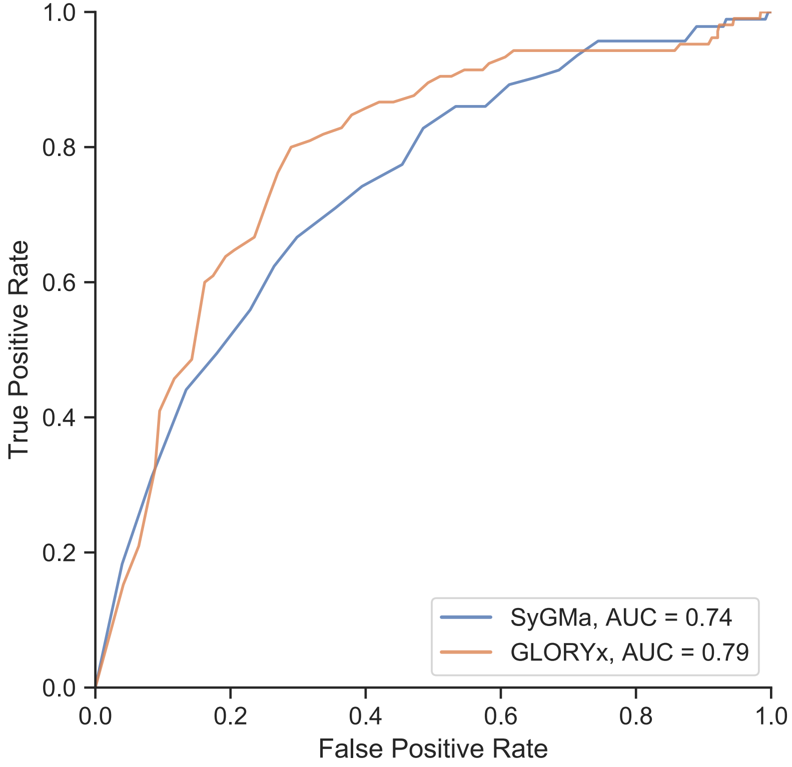
1. Extracted reaction types for phase I and phase II enzymes from the literature
2. Represented reaction types by SMIRKS:
   - e.g. "[c:1][H:2]>>[c:1][O][H:2]"
3. Applied transformations using AMBIT SMIRKS
   - Open-source Java library (IdeaConsult Ltd)
4. The transformations are only applied at those positions



de Bruyn Kops et al., Chem Res Toxicol 2020. doi: 10.1021/acs.chemrestox.0c00224

GLORYx: Performance on an external test set

	GLORYx	SyGMa
Recall	0.77	0.68
Precision	0.06	0.12
Total no. predictions (metabolites)	1724	800
No. true positives	105	93



de Bruyn Kops et al., Chem Res Toxicol 2020. doi: 10.1021/acs.chemrestox.0c00224

Integration of metabolism prediction in toxicity prediction

Study	Endpoint(s)	Modeling approach	Integration of metabolism	Performance of the metabolism-aware approach as compared to the baseline models
Dimitriev et al. 2017	Rat acute toxicity	Linear regression models trained on LD_{50} values for 3000 **parent compounds**	Predictions for measured metabolites integrated by, e.g., averaging predicted LD_{50} values	R^2 increased by 0.03 (from 0.78 to 0.81)
Filimonov et al. 2020	28 endpoints	Bayesian classification trained on up to 5583 **parent compounds** per endpoint	Predictions for measured metabolites integrated by max fusion	Precision increased by up to 0.14 Recall increased by up to 0.16
Mekenyan et al. 2004	In vitro mutagenicity (AMES assay)	Decision trees	Predictions for predicted metabolites integrated by max fusion	Performance dropped but some toxic compounds were identified correctly via their mutagenic metabolites
Further works from the LMC	Skin sensitization, respiratory sensitization, liver genotoxicity, etc.	Decision trees	Predictions for predicted metabolites	No comparison to baseline approach was performed

Garcia de Lomana et al., Digital Discov 2022. doi: 10.1039/d1dd00018g

Integration of metabolism prediction in toxicity prediction

Marina Garcia de Lomana

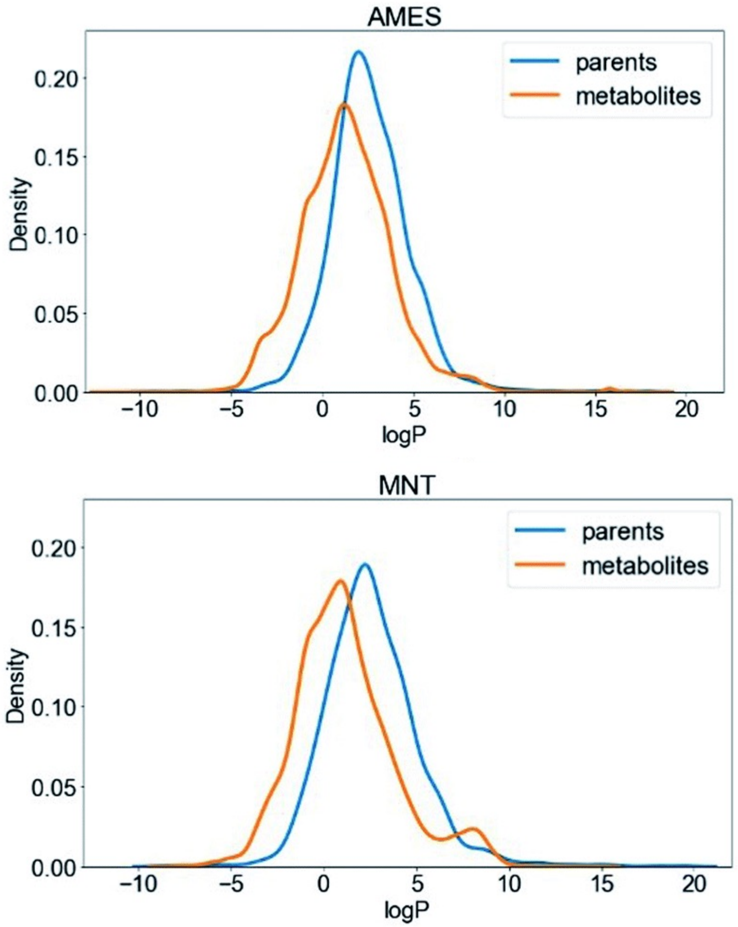
Endpoint/testing system	No. toxic compounds	No. non-toxic compounds	Ratio
Ames mutagenicity (considering metabolic activation with S-9 liver extract)	1908	3153	1 : 2
Micronucleus test (MNT) for assessing genotoxicity	315	1460	1 : 5
Drug induced liver injury (DILI)	435	226	2 : 1
Drug-induced cardiological complications (DICC)	965	2243	1 : 2
Murine local lymph node assay (LLNA)	521	749	1 : 1

- **Metabolites predicted** with Meteor:
  - Leading software for metabolite prediction
  - Use of the recommended “SOM scoring method”
  - Distinguishes ~500 types of biotransformations (phase 1 and 2)
- Descriptors: count-based Morgan2 fingerprints, physicochemical properties, CDDD descriptors
- Machine learning algorithm: random forest (other algorithms were also explored)
  - +/-feature selection (LASSO), +/- data balancing with SMOTENC, +/- filtering of certain metabolites

Garcia de Lomana et al., Digital Discov 2022, 1, 158-172. doi: 10.1039/d1dd00018g

Analysis of the chemical space of the parent compounds and their predicted metabolites

- Metabolites predicted by Meteor:
  - Up to 828
  - Median: 8 to 12 (depending on the data set)
- Physicochemical properties of the metabolites of “toxic” and “non-toxic compounds” generally similar
  - Metabolites of “toxic compounds” have, on average, a higher Clog*P* (+0.8)
- Over-representation of certain types of biotransformations among “toxic compounds” observed; however, these observations universal



31.01.2023 Garcia de Lomana et al., Digital Discov 2022, 1, 158-172. doi: 10.1039/d1dd00018g

Experiment 1: Integration of metabolism information into model input

Random forest models	Parent encoding	Metabolite encoding	Performance during 5-fold CV
Baseline models	Morgan2 fingerprints and/or RDKit physchem properties	Not encoded	Mean F1 scores ranging from 0.64 (MNT) to 0.82 (Ames)
Type A Metabolism-aware models		Morgan2 fingerprints and/or RDKit physchem properties for the **five top-ranked metabolites**	Minor gains in performance which did not exceed +0.04 among the evaluated metrics
Type B Metabolism-aware models		**Biotransformation signature** encoding the no. occurrences of the individual types of biotransformations	No gain in performance, also not when applying (addn.) feature selection

Garcia de Lomana et al., Digital Discov 2022, 1, 158-172. doi: 10.1039/d1dd00018g

Experiment 2: Combination of the predictions obtained for parent compounds and predicted metabolites

Random forest models	Metabolite encoding	Combination of predicted probabilities of toxicity	Gains in performance over the baseline models
Baseline models	Not encoded	n/a	n/a
Type C metabolism-aware models	Dedicated models for the parent compounds plus dedicated models for the labelled, predicted metabolites	**Mean** predicted probability over **all** parent compounds and predicted metabolites	No gain
Type D metabolism-aware models		**Median** predicted probability over **all** parent compounds and predicted metabolites	No gain
Type E metabolism-aware models		**Maximum** predicted probability over **all** parent compounds and predicted metabolites	No gain
Type F metabolism-aware models		Mean between the predicted probabilities for the parent compound and the metabolite predicted as most likely toxic	F1 scores, on average, +0.03 (only few diffs. statistically significant)
Type F’ metabolism-aware models		Identical to Type F, with the additional filtering of metabolites with ClogP < 3 and phase II metabolites	F1 scores, on average, +0.06

Conclusions

- **Computational methods can make a significant contribution to understanding metabolism, yet global models for quantitative prediction are still out of reach:**
  - Small molecule-enzyme interaction (++)
  - Sites of metabolism (+++)
  - Structures of likely metabolites (+~)
- Integration of metabolism prediction in toxicity prediction is the logical next step
  - Limited success in integrating metabolism and toxicity prediction so far
  - Primary challenge: Scarcity of the available data, in particular of data on measured and labeled (i.e. toxic, non-toxic) metabolites

universität wien

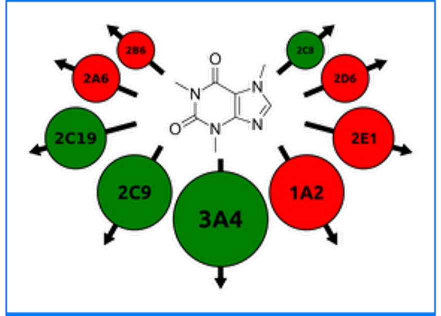
NERDD

New E-Resource for Drug Discovery

nerdd.univie.ac.at

Stork et al., *Bioinformatics* 2019, 36, 1291–1292

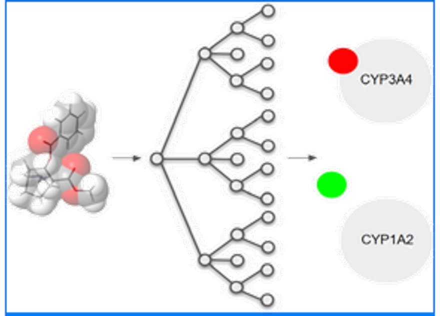
Cytochrome P450 substrates



CYPstrate

Prediction of Cytochrome P450 substrates

Cytochrome P450 inhibitors



CYPlebrity

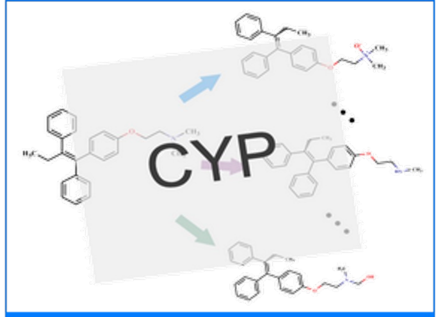
Prediction of Cytochrome P450 inhibitors

Sites of Metabolism

FAME 3

Regioselectivity prediction for phase 1 and phase 2 metabolism

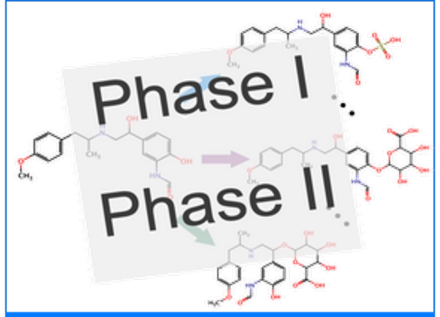
Metabolite Structures



GLORY

Metabolite structure prediction for cytochrome P450 metabolism

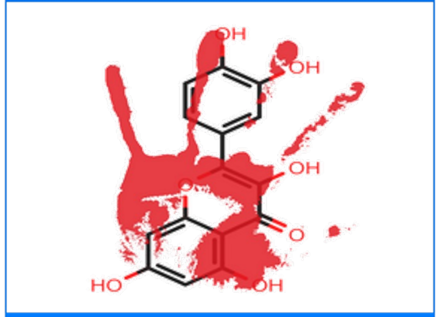
Metabolite Structures



GLORYx

Metabolite structure prediction for phase I and II metabolism

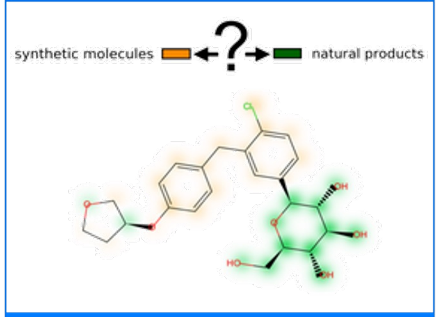
Frequent Hitters



Hit Dexter 3

Prediction of frequent hitters

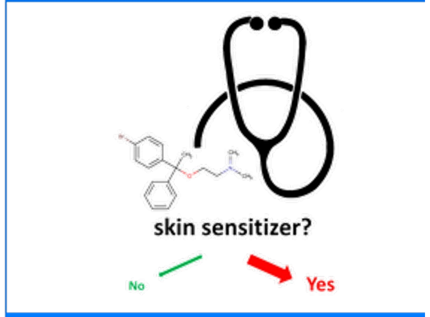
Natural Product-Likeness



NP-Scout

Identification and visualization of natural product-likeness

Skin Sensitization



Skin Doctor CP

Prediction of skin sensitization potential

Slides available from: