Why is predicting drug metabolism still so hard

Swapnil Chavan

RISE

What happens to drug?

After absorption, a drug enters a living chemical system.

It may be:

- transported
- bound to proteins
- distributed into tissues
- metabolized
- cleared

Defense Mechanism

The liver as a chemical filter

Most drugs are lipophilic enough to cross membranes.
But lipophilic molecules are hard to excrete.

So the body often modifies them to become:

- ❑ more polar
- ❑ more water-soluble
- ❑ easier to eliminate

Phases of metabolism

Phase I metabolism:

Phase II metabolism:

Functionalization

- Introduces or exposes functional groups (–OH, –NH_2, –SH)
- Main reactions:

  -Oxidation

  -Reduction

  -Hydrolysis
- Enzymes involved: Cytochrome P450 (CYP450)
- Purpose: Makes substances slightly polar or more reactive

What activates CYP enzymes?

CYP activity depends on two levels:

- **Immediate catalytic activation**
  - substrate binds CYP active site
  - oxygen and electrons enter the catalytic cycle
  - reactive iron–oxo species forms
- **Longer-term induction**
  - nuclear receptors sense xenobiotics
  - gene expression of CYPs increases

How CYPs chemically attack drugs

Basic catalytic logic:

- Drug binds near heme iron
- CYP receives electrons from NADPH
- O_2 binds and is activated
- Reactive iron–oxo species forms
- One oxygen atom is inserted into the drug
- Metabolite leaves active site

Conjugation

- Adds large polar molecules to compounds
- Common reactions:

-Glucuronidation

-Sulfation

-Glutathione conjugation

-Acetylation

- Purpose:

-Makes compounds highly water-soluble

-Usually inactivates toxins/drugs

How Phase II enzymes attack drugs

- Phase 2 enzymes transfer activated polar groups onto functional sites created in Phase 1, forming covalent bonds that increase water solubility.
- For example:

  Drug–OH + UDP-glucuronic acid → Drug–O–glucuronide

CHALLENGE 1:
Data Foundations: Quantity & Quality

THE DATASET

MetXBioDB

2,133

Biotransformation records

Curated parent→metabolite pairs from the published literature

1,252

Unique parent substrates

Across Phase I, Phase II, and gut-microbial biotransformation

777

Phase-I (CYP) substrates

The subset relevant to oxidative, enzyme-mediated metabolism

230

Distinct reaction-types

Phase-I reactions alone

Biosystem = Human; biotransformation_type = Human Phase I / Human Phase II / Human Gut Microbial.

Source: MetXBioDB

Data format

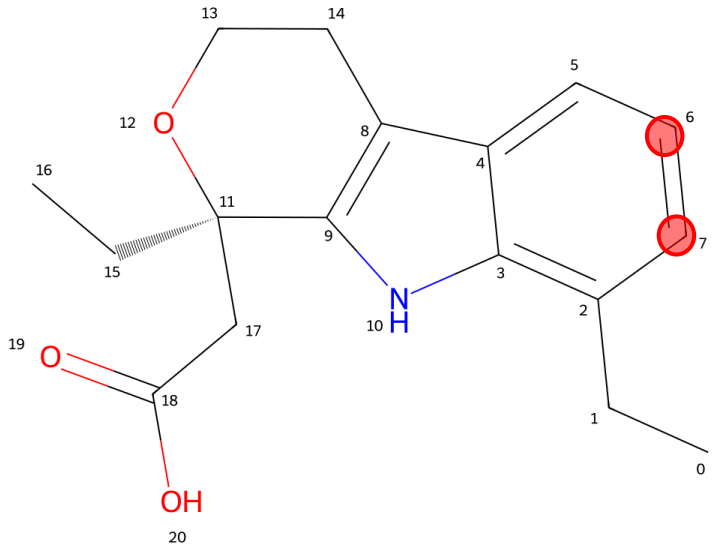
- Data looks like:

Predecessor_CID	Predecessor_Name	Predecessor_SMILES	Successor_CID	Successor_Name	Successor_SMILES
10892731	(-)-bisoprolol	CC(C)NC[C@@H](COC1=CC=C(C=C1)COCCOC(C)C)O	154699479	(-)-bisoprolol, O-deisopropyl	CC(C)NC[C@@H](COC1=CC=C(C=C1)COCCO)O
6603844	(-)-ethosuximide	CC[C@@]1(CC(=O)NC1=O)C	54520268	3-(1-Hydroxyethyl)-3-methylpyrrolidine-2,5-dione	CC(C1(CC(=O)NC1=O)C)O
667528	(-)-etodolac	CCC1=C2C(=CC=C1)C3=C(N2)[C@@](OCC3)(CC)CC(=O)O	14112821	7-Hydroxyetodolac	CCC1=C(C=CC2=C1NC3=C2CCOC3(CC)CC(=O)O)O
667528	(-)-etodolac	CCC1=C2C(=CC=C1)C3=C(N2)[C@@](OCC3)(CC)CC(=O)O	14146232	6-Hydroxyetodolac	CCC1=C2C(=CC(=C1)O)C3=C(N2)C(OCC3)(CC)CC(=O)O
439250	(-)-limonene	CC1=CC[C@H](CC1)C(=C)C	369312	Perillyl alcohol	CC(=C)[C@H]1CCC(=CC1)CO
439250	(-)-limonene	CC1=CC[C@H](CC1)C(=C)C	7438	Carveol	CC1=CCC(CC1O)C(=C)C
16666	(-)-Menthol	C[C@@H]1CC[C@H]([C@@H](C1)O)C(C)C	19100	p-Menthane-3,-8-diol	C[C@@H]1CC[C@H]([C@@H](C1)O)C(C)(C)O
644018	(-)-pseudococaine	CN1[C@@H]2CC[C@H]1[C@H]([C@@H](C2)OC(=O)C3=CC...	154699498	Norcocaine	COC(=O)[C@@H]1[C@@H]2CC[C@@H](N2)C[C@H]1OC(=...
92142	(-)-thalidomide	C1CC(=O)NC(=O)[C@H]1N2C(=O)C3=CC=CC=C3C2=O	154699639	5'-Hydroxythalidomide	C1[C@@H](C(=O)NC(=O)C1O)N2C(=O)C3=CC=CC=C3C2=O
92142	(-)-thalidomide	C1CC(=O)NC(=O)[C@H]1N2C(=O)C3=CC=CC=C3C2=O	44250311	5-hydroxy-thalidomide	C1CC(=O)NC(=O)[C@H]1N2C(=O)C3=C(C2=O)C=C(C=C3)O
92142	(-)-thalidomide	C1CC(=O)NC(=O)[C@H]1N2C(=O)C3=CC=CC=C3C2=O	154699796	(-)-thalidomide arene oxide	C1CC(=O)NC(=O)[C@H]1N2C(=O)C3=CC4C(O4)C=C3C2=O
10220327	(-)-zileuton	C[C@@H](C1=CC2=CC=CC=C2S1)N(C(=O)N)O	95162932	Zileuton sulfoxide	C[C@@H](C1=CC2=CC=CC=C2S1=O)N(C(=O)N)O
10220327	(-)-zileuton	C[C@@H](C1=CC2=CC=CC=C2S1)N(C(=O)N)O	154700148	Hydroxyzileuton	C[C@@H](C1=CC2=C(S1)C=C(C=C2)O)N(C(=O)N)O

CHALLENGE 2:
The Numbering Problem

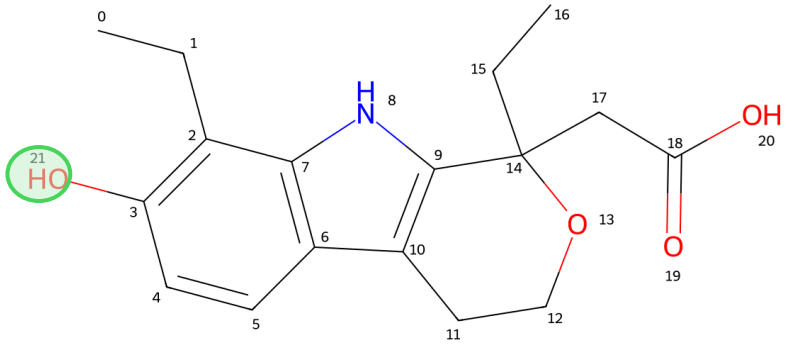
Label extraction challenge

- RdKit numbering issue



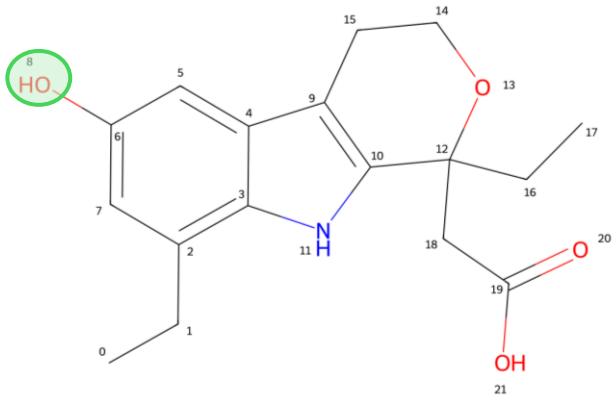
Etodolac

CN(C)CCOC1=CC2=CC=CC=C2SC3=C1C=C(C=C3)Cl



Metabolite 1

CN(C)CCOC1=CC2=CC=CC=C2S(=O)C3=C1C=C(C=C3)Cl

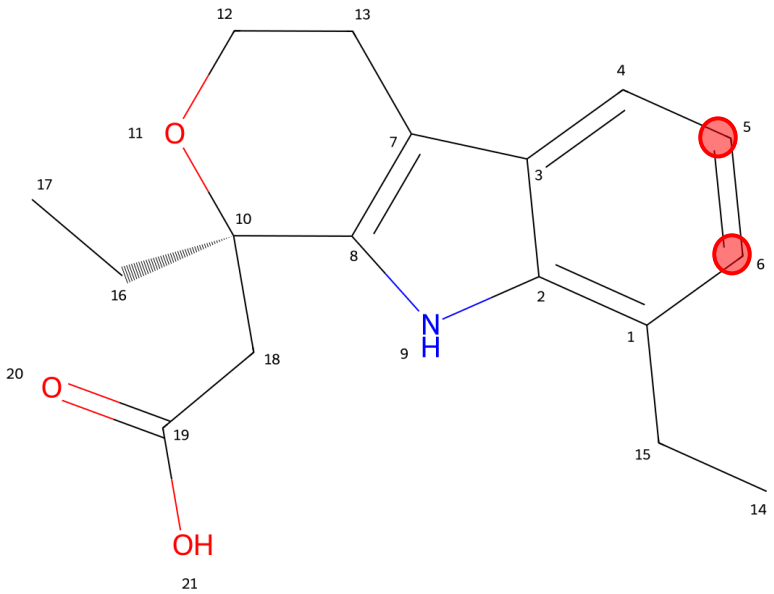


Metabolite 2

C[N+](C)(CCOC1=CC2=CC=CC=C2SC3=C1C=C(C=C3)Cl)[O-]

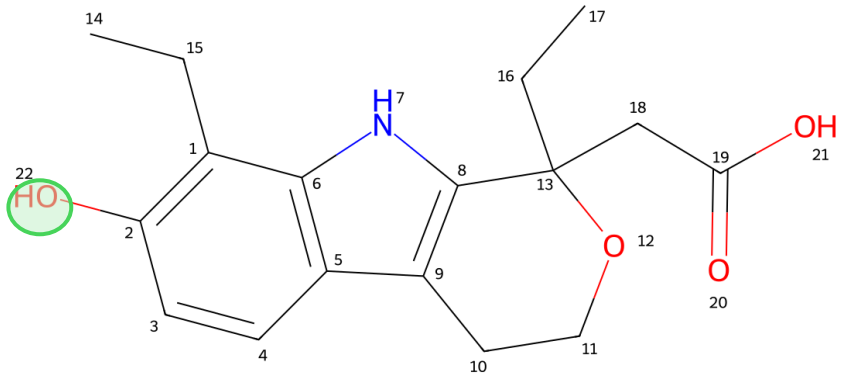
Label extraction challenge

- Scaffold-based numbering issue



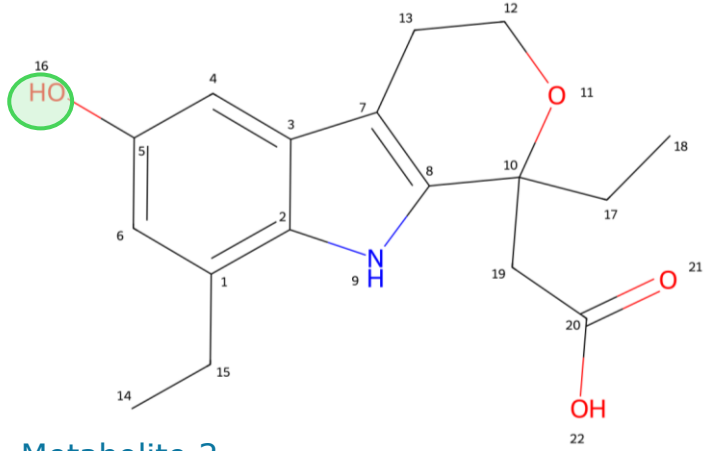
Etodolac

CN(C)CCOC1=CC2=CC=CC=C2SC3=C1C=C(C=C3)Cl



Metabolite 1

CN(C)CCOC1=CC2=CC=CC=C2S(=O)C3=C1C=C(C=C3)Cl

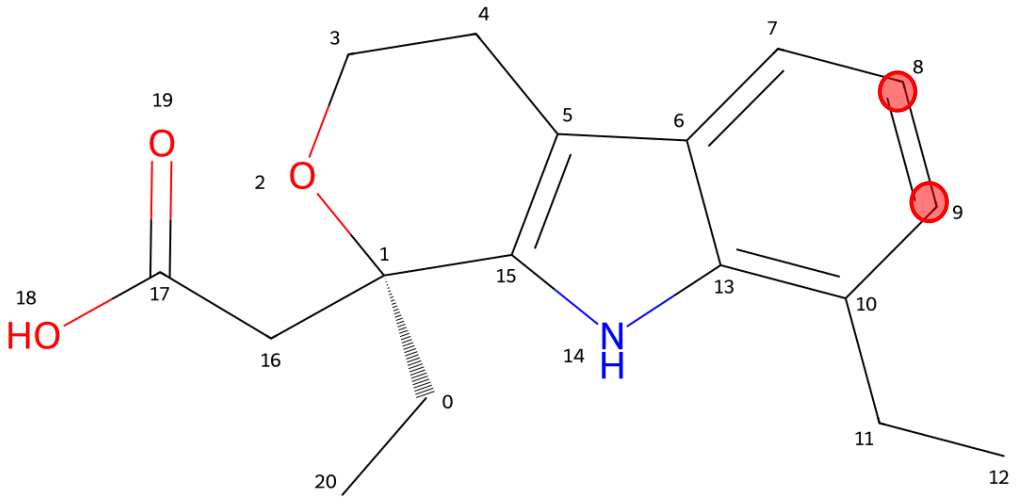


Metabolite 2

C[N+](C)(CCOC1=CC2=CC=CC=C2SC3=C1C=C(C=C3)Cl)[O-]

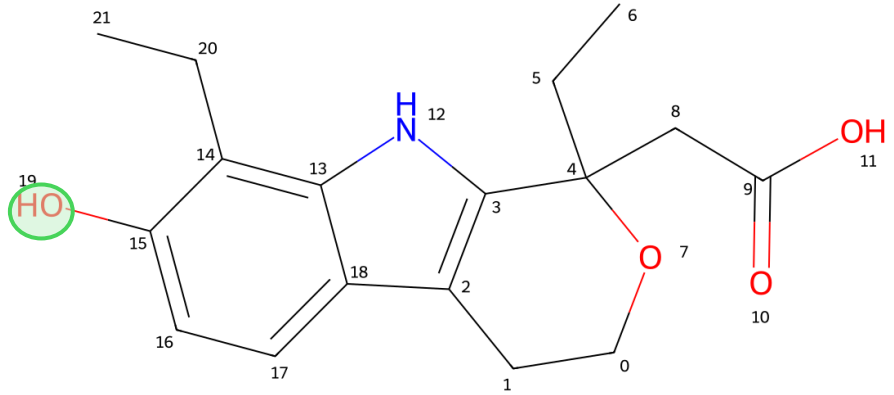
Label extraction challenge

- Enumerated SMILES - RdKit numbering issue



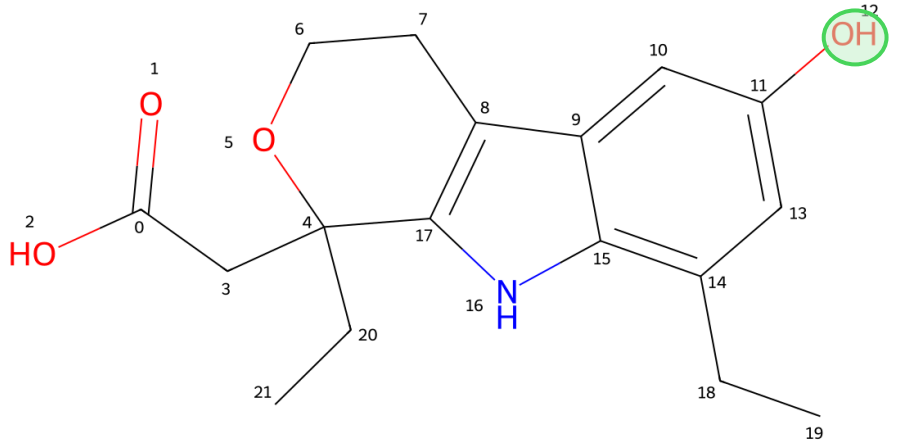
Etodolac

C([C@@]1(OCCc2c3cccc(CC)c3[nH]c21)CC(O)=O)C



Metabolite 1

C1Cc2c(C(CC)(O1)CC(=O)O)[nH]c1c(c(ccc12)O)CC

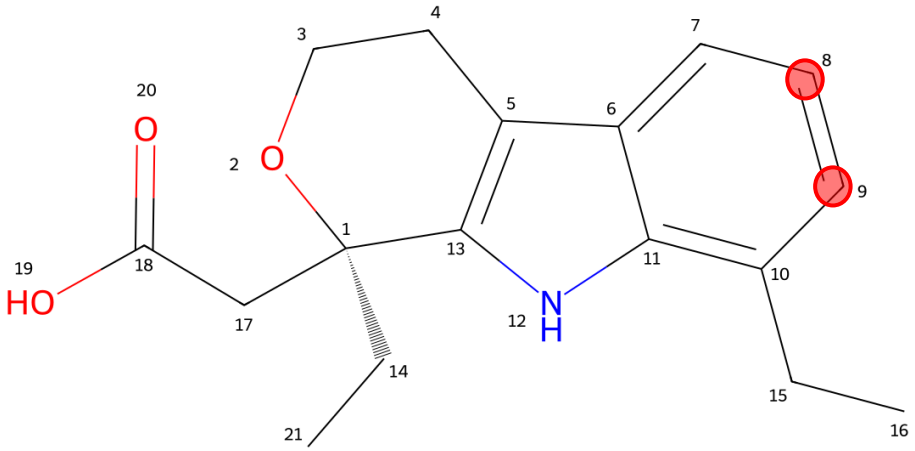


Metabolite 2

C(=O)(O)CC1(OCCc2c3cc(O)cc(c3[nH]c12)CC)CC

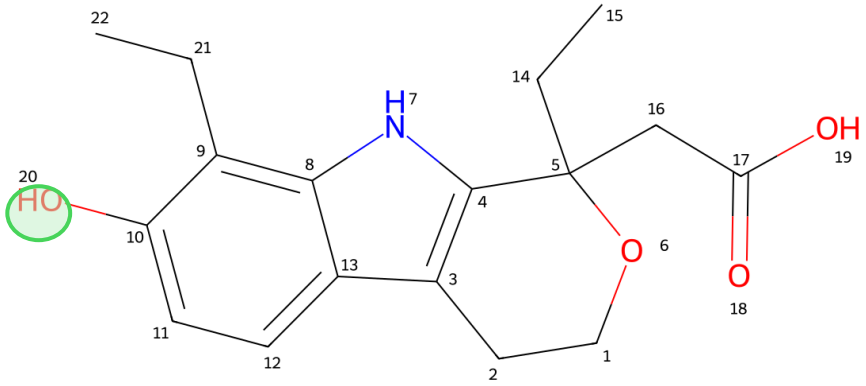
Label extraction challenge

- Enumerated SMILES- Scaffold-based numbering issue



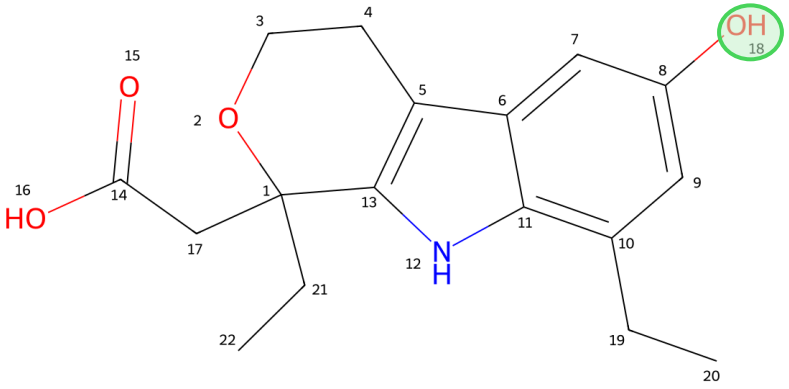
Etodolac

C([C@@]1(OCCc2c3cccc(CC)c3[nH]c21)CC(O)=O)C



Metabolite 1

C1Cc2c(C(CC)(O1)CC(=O)O)[nH]c1c(c(ccc12)O)CC

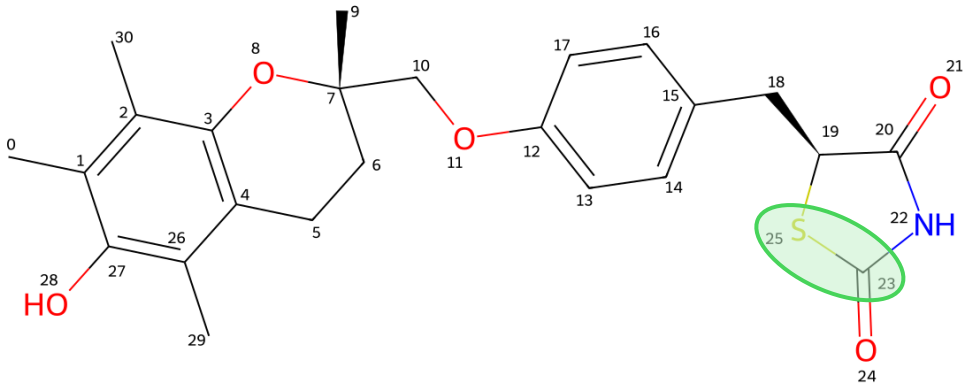


Metabolite 2

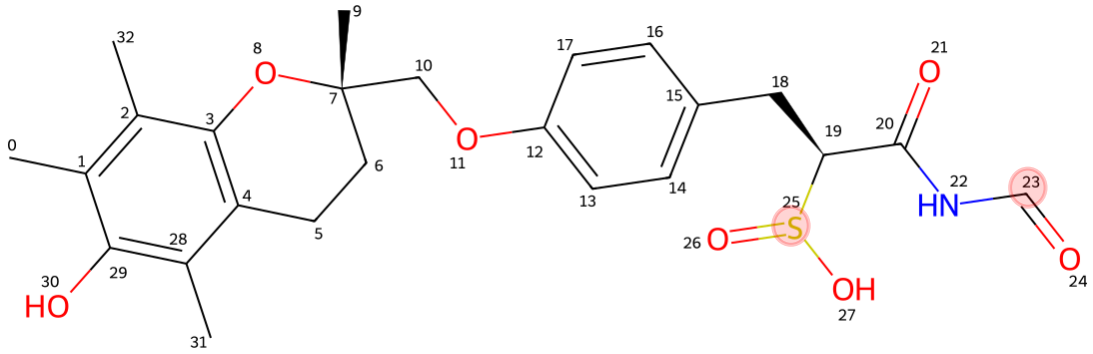
C(=O)(O)CC1(OCCc2c3cc(O)cc(c3[nH]c12)CC)CC

Label extraction challenge

- Bond or atom? Ring cleavage scenario



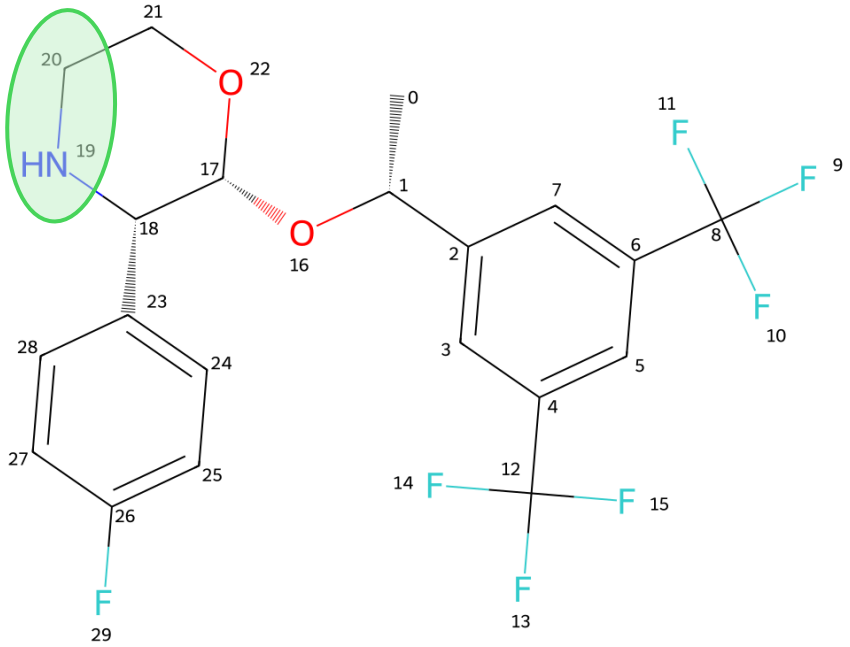
(2s,5s)-troglitazone



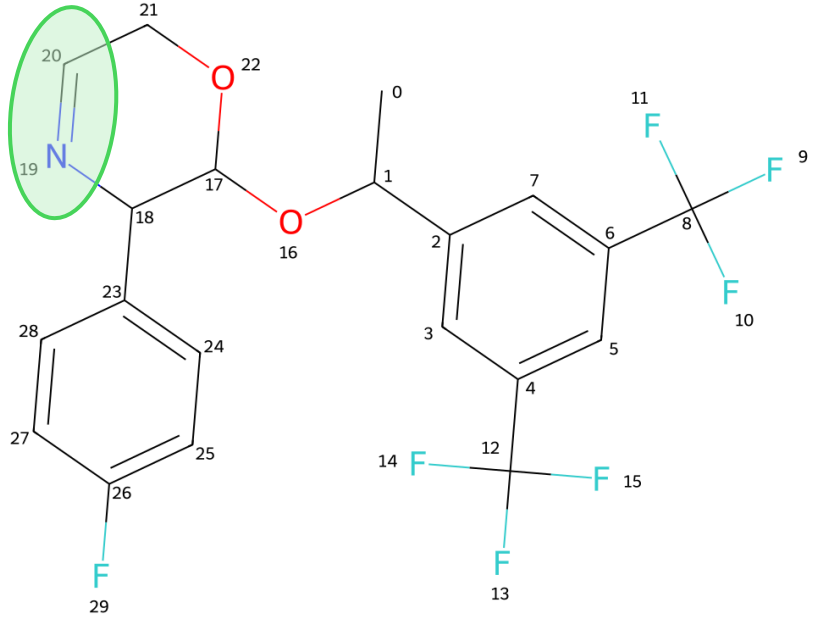
(1S)-1-(formylcarbamoyl)-2-(4-{[(2S)-6-hydroxy-2,5,7,8-tetramethyl-3,4-dihydro-2H-1-benzopyran-2-yl]methoxy}phenyl)ethane-1-sulfinic acid

Label extraction challenge

- Dehydrogenation Cases



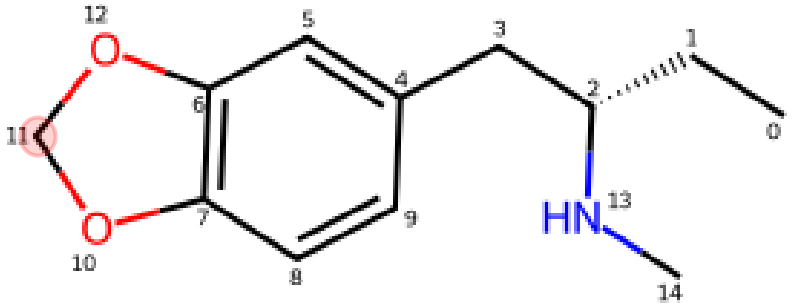
(2*R*,3*S*)-2-[(1*R*)-1-[3,5-bis(trifluoromethyl)phenyl]ethoxy]-3-(4-fluorophenyl)morpholine



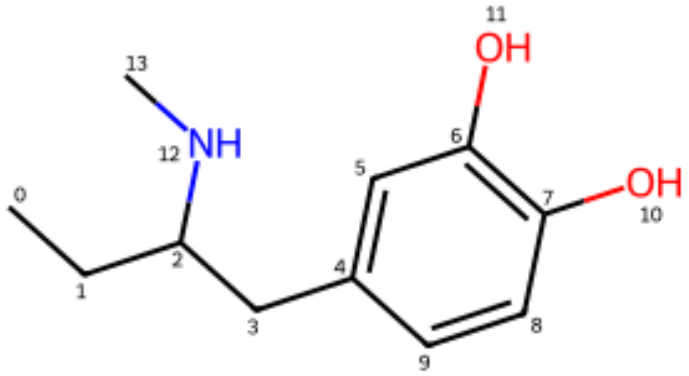
2-[1-[3,5-bis(trifluoromethyl)phenyl]ethoxy]-3-(4-fluorophenyl)-3,6-dihydro-2H-1,4-oxazine

Label extraction challenge

- Ring Cleavage & atom removed scenario



(2S)-1-(1,3-Benzodioxol-5-yl)-N-methyl-2-butanamine

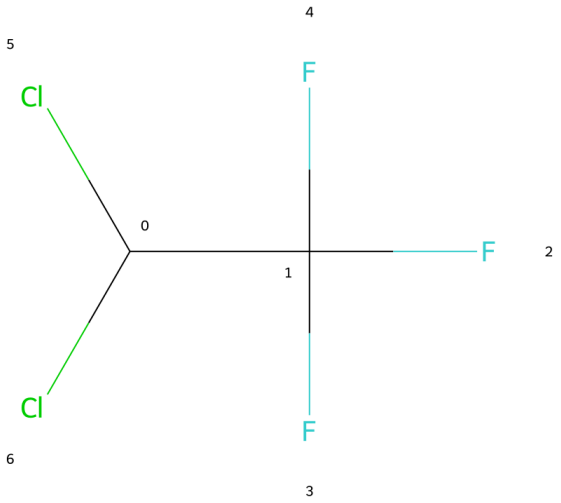


4-[2-(Methylamino)butyl]benzene-1,2-diol

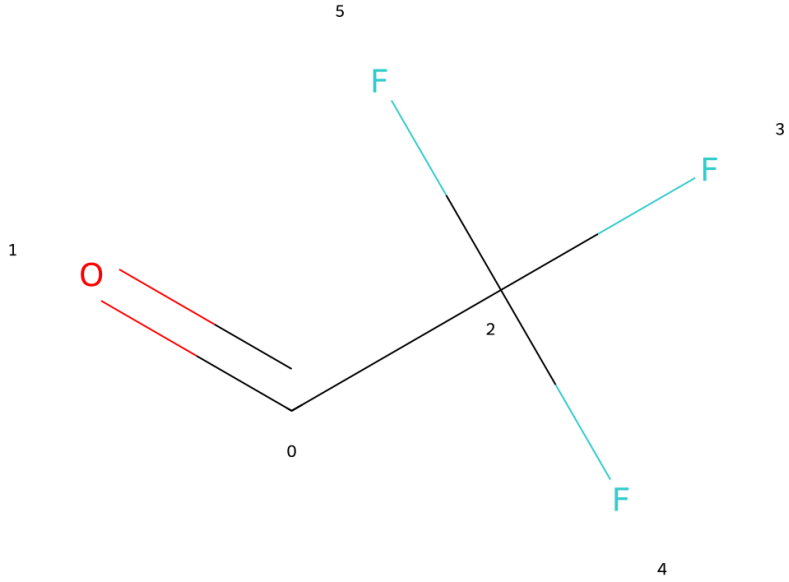
Decrement!

Label extraction challenge

- Multiple atoms removed scenario



2,2-dichloro-1,1,1-trifluoroethane

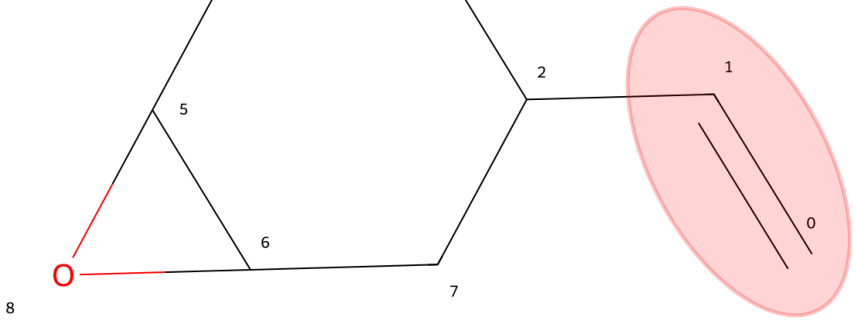


Trifluoroacetyl chloride

Decrement! Large difference in Mwt & formula

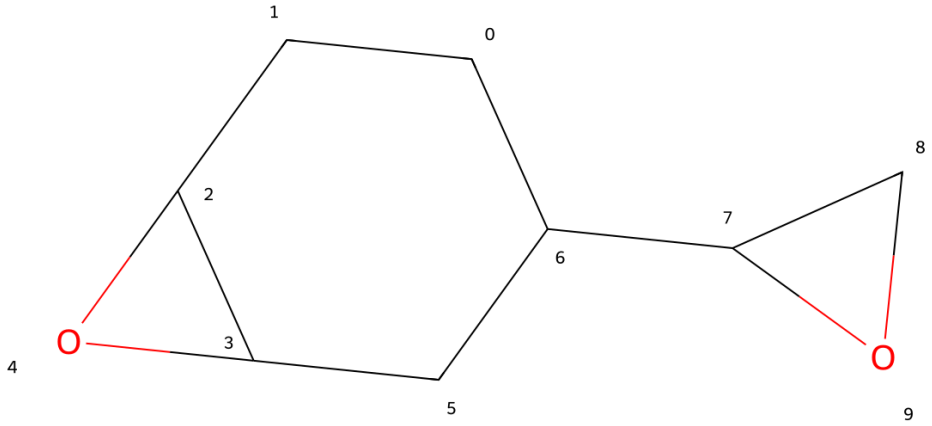
Label extraction challenge

- Bond formation scenario (rdkit numbering)



Vinylcyclohexene monoxide

IUPAC: 3-ethenyl-7-oxabicyclo[4.1.0]heptane

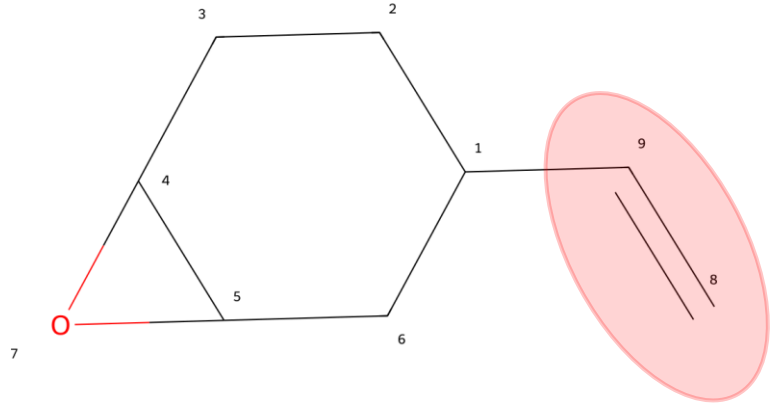


Vinyl cyclohexene dioxide

IUPAC: 3-(oxiran-2-yl)-7-oxabicyclo[4.1.0]heptane

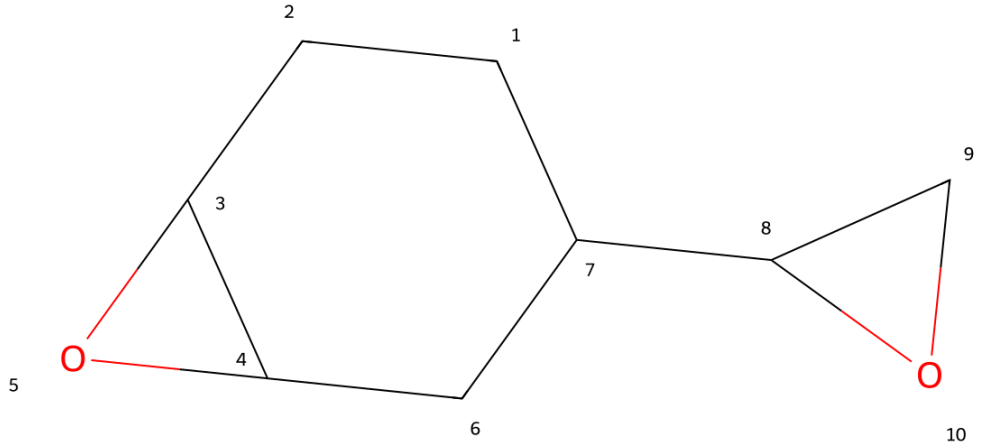
Label extraction challenge

- Bond formation scenario (scaffold-based numbering)



Vinylcyclohexene monoxide

IUPAC: 3-ethenyl-7-oxabicyclo[4.1.0]heptane

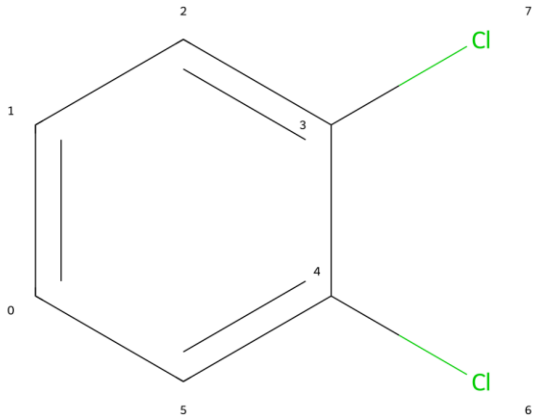


Vinyl cyclohexene dioxide

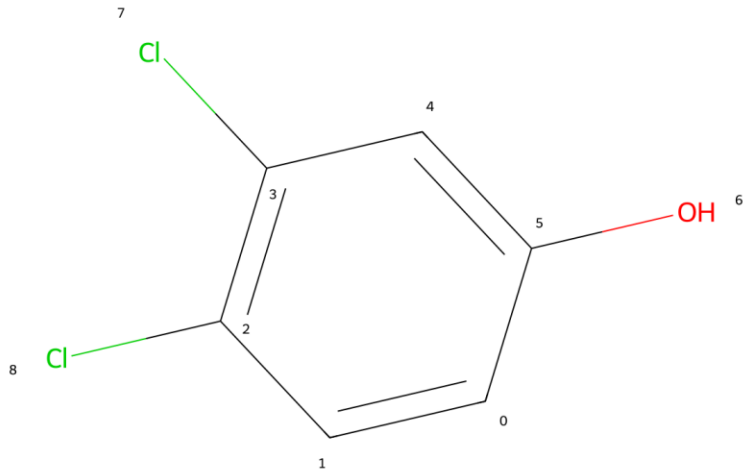
IUPAC: 3-(oxiran-2-yl)-7-oxabicyclo[4.1.0]heptane

Symmetry Challenge

- **Symmetry-Induced Ambiguity**



1,2-Dichlorobenzene

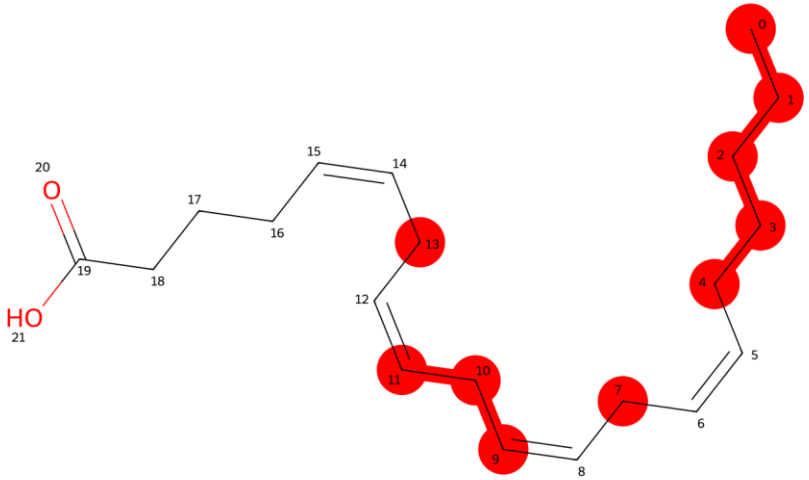


3,4-dichlorophenol

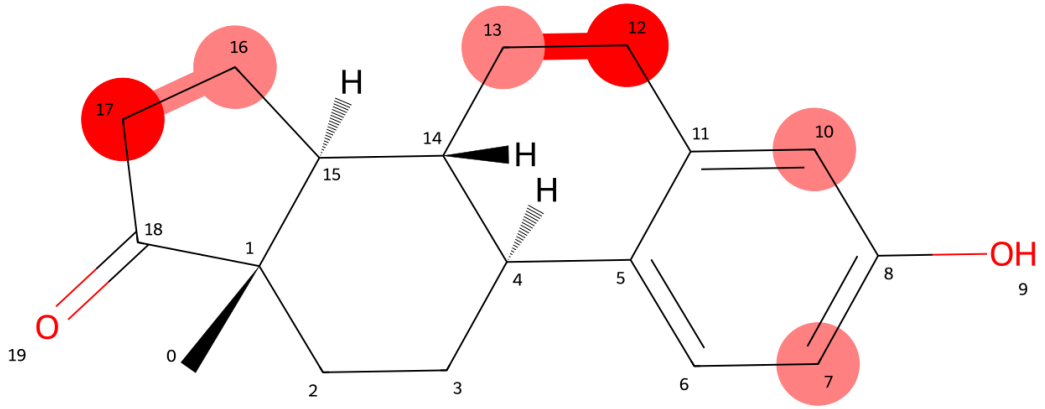
CHALLENGE 3:
Multiple labels

Phase I Metabolism

Multiple SOMs



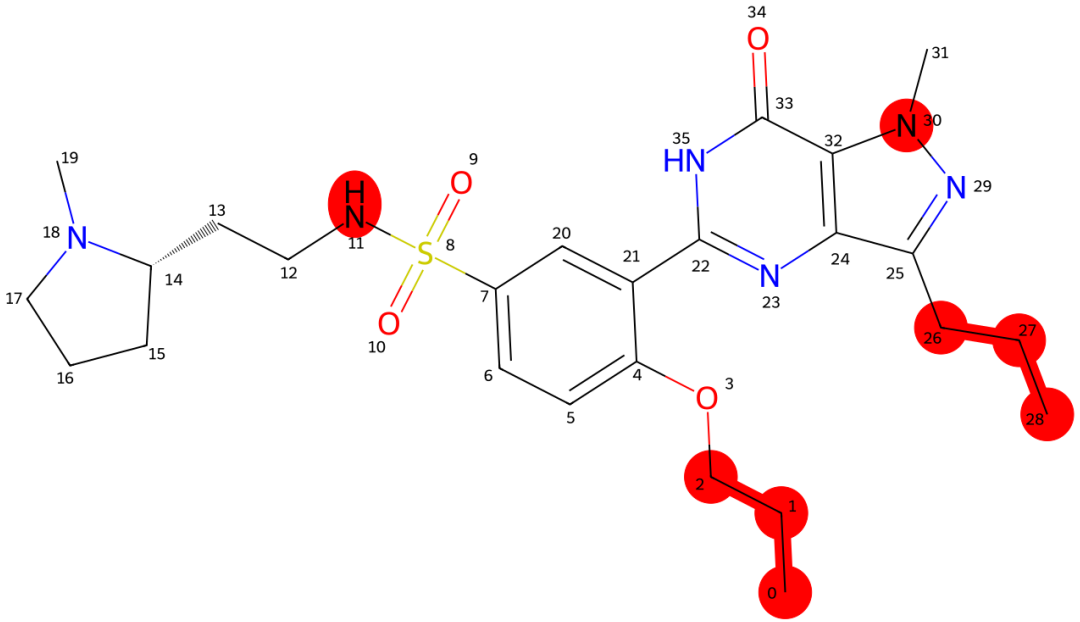
Arachidonic Acid



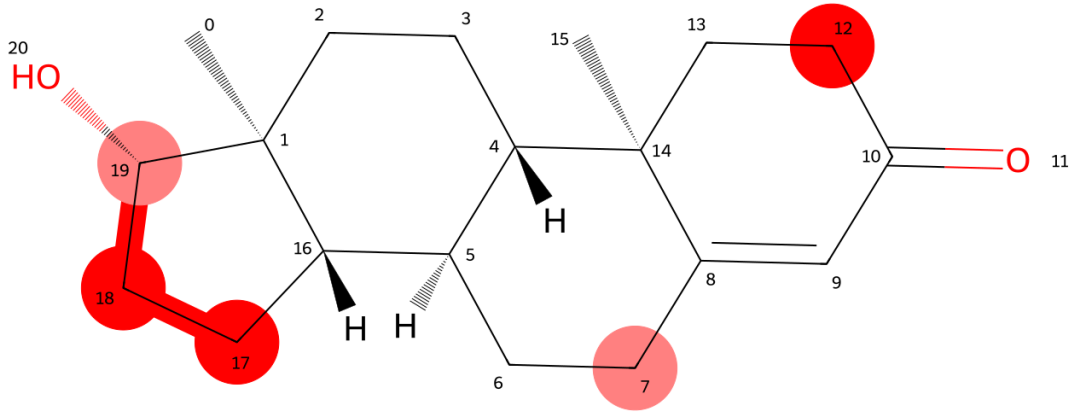
Estrone

Phase I Metabolism

Multiple SOMs

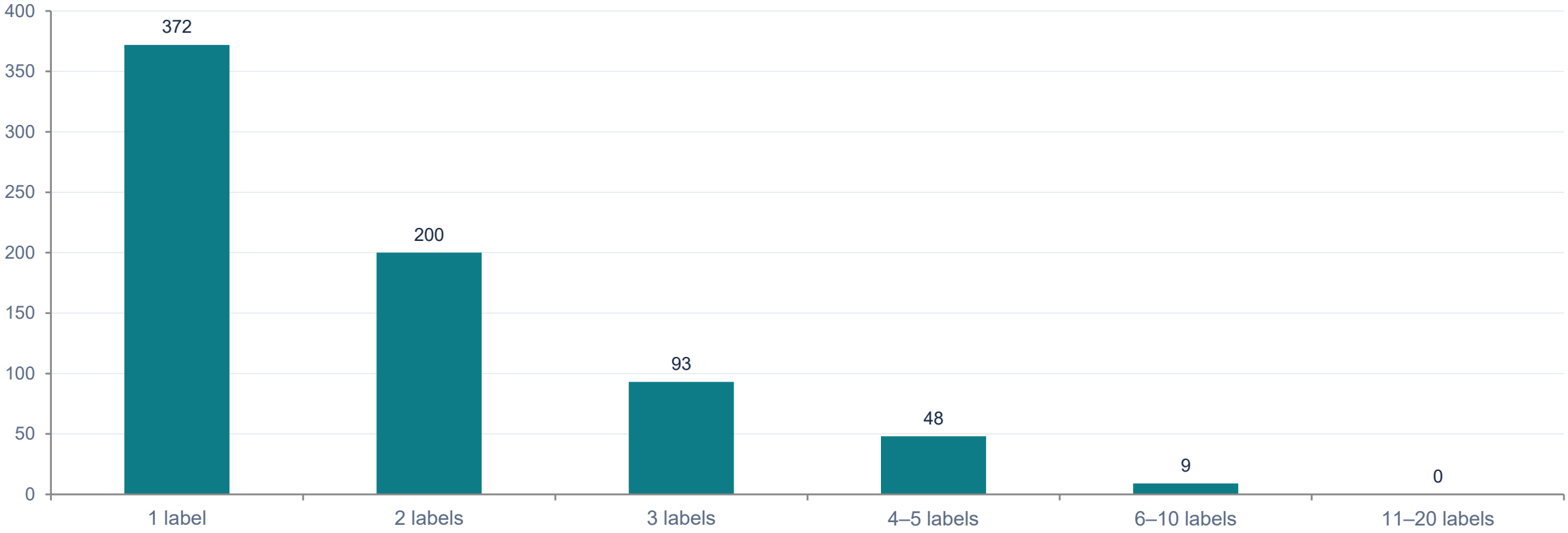


Udenafil



Testosterone

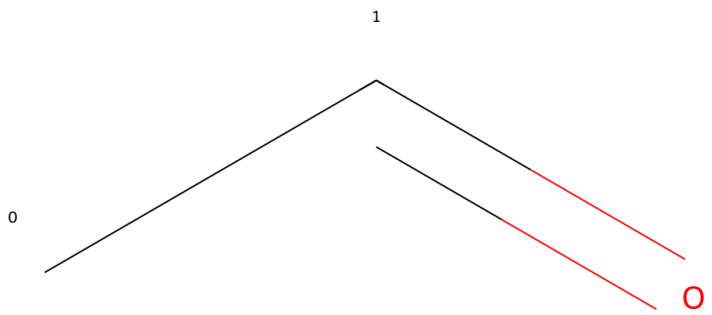
Single label compounds are 'easy'; a long tail is genuinely hard



CHALLENGE 4:
Label variation

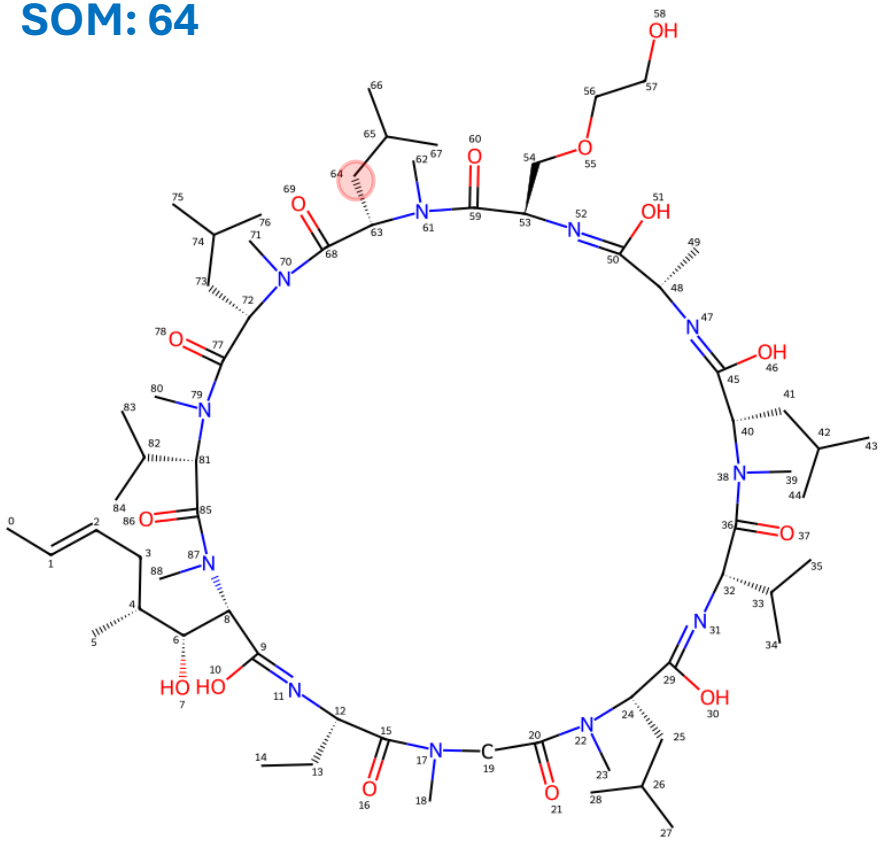
Size variation

SOM: 1



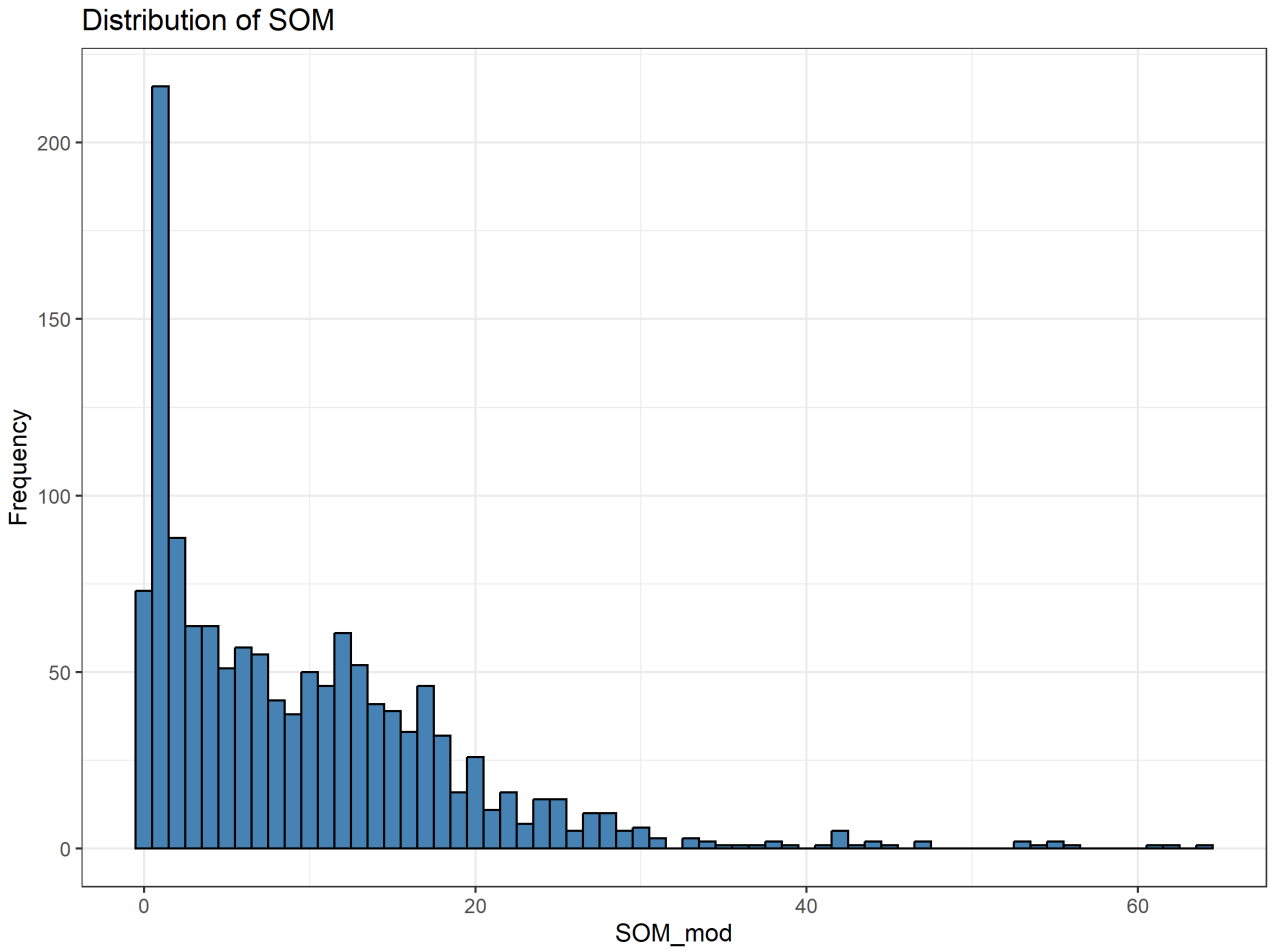
Mwt. = 44.053

SOM: 64



Mwt. = 1262.687

Histogram of SOM index based on rdkit numbering



CHALLENGE 5:
One-to-many relation

Phase I Metabolism

How Many Metabolites Does One Compound Have?

372

Compounds, 1 known metabolite

372 records

350

Compounds, 2+ known metabolites

949 records

10

Maximum metabolites for one compound

Arachidonic Acid

WHY CARDINALITY SKEW BREAKS NAIVE EVALUATION

Single-label and multi-label compounds need different scoring

The 1-label majority

- Easy to frame as single-class classification
- A model that always predicts the single most common SOM already scores deceptively well here
- Dominates an unweighted average accuracy number

The multi-label tail

- Requires set prediction — precision and recall behave completely differently than for single-label classification
- Class imbalance in label count biases models toward predicting too few sites

CHALLENGE 6:
Multiple CYP Enzymes

MULTIPLE ENZYMES

More than half of records list more than one CYP

274

Compounds flagging a single enzyme

502

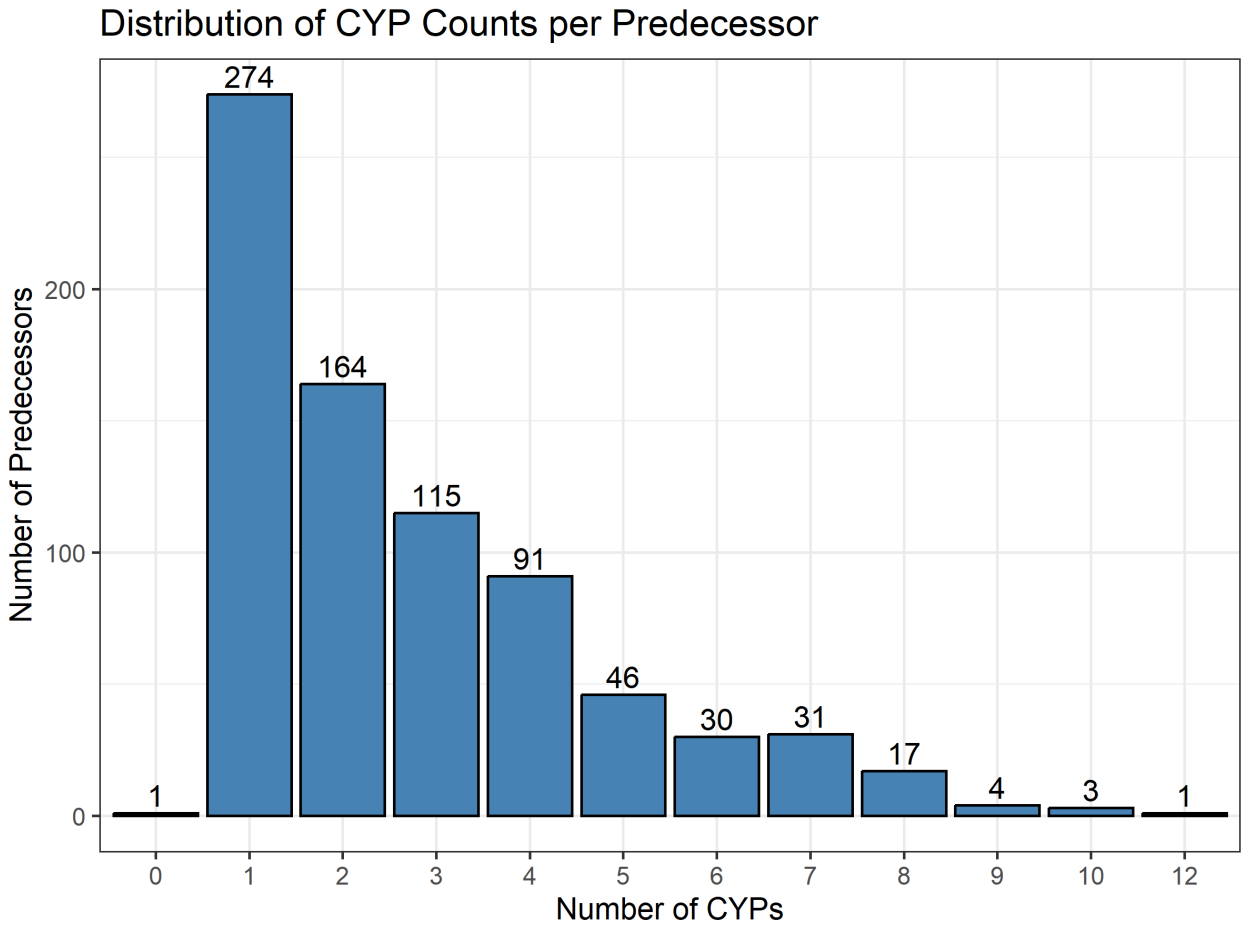
Compounds listing 2 or more enzyme

1

Compound listing 12 Distinct CYPs

MULTIPLE ENZYMES

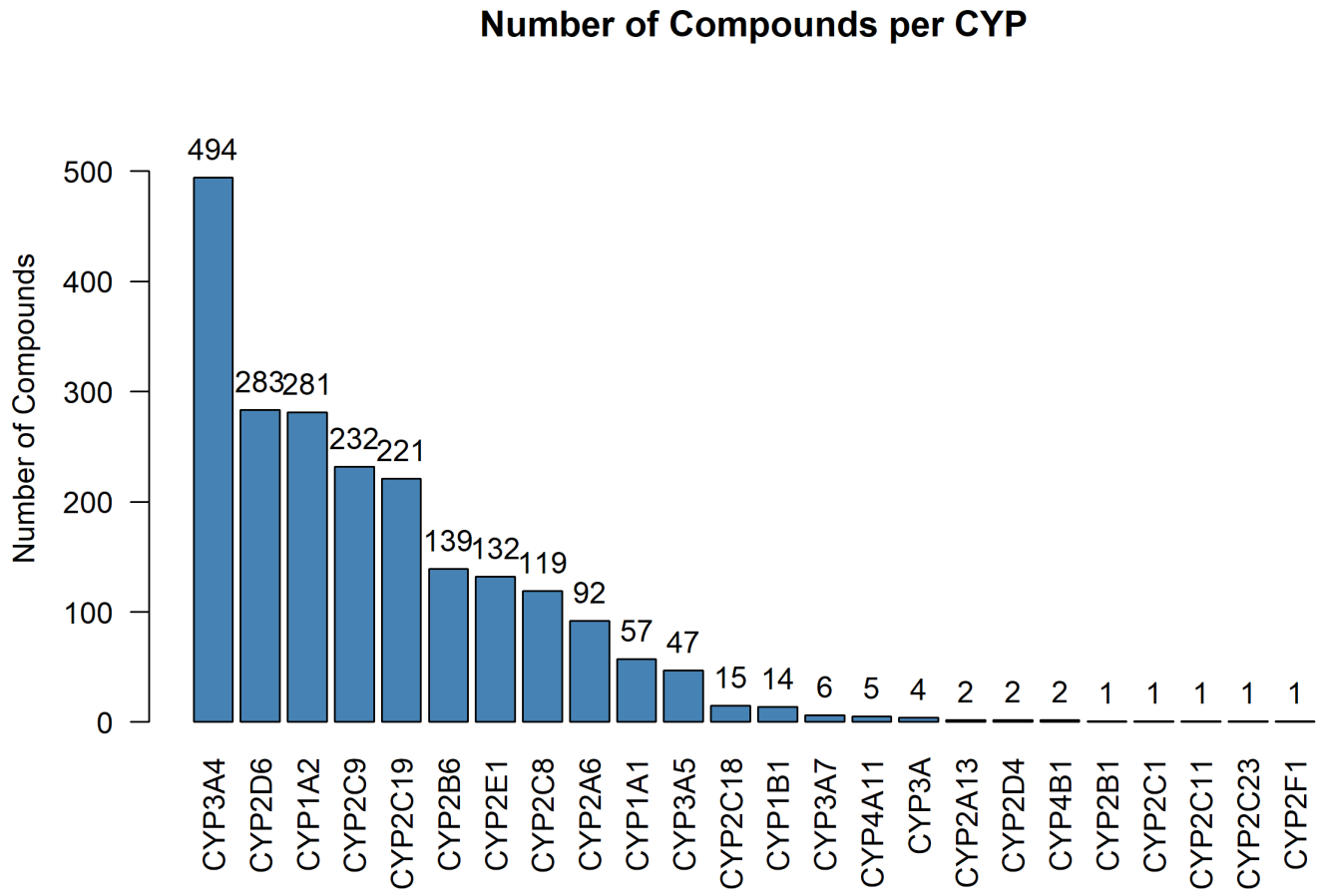
CYP Frequency Per Compound



CHALLENGE 7:
Multiple Reaction Classes

MULTIPLE ENZYMES

Number of Compounds Associated with Each CYP Enzyme



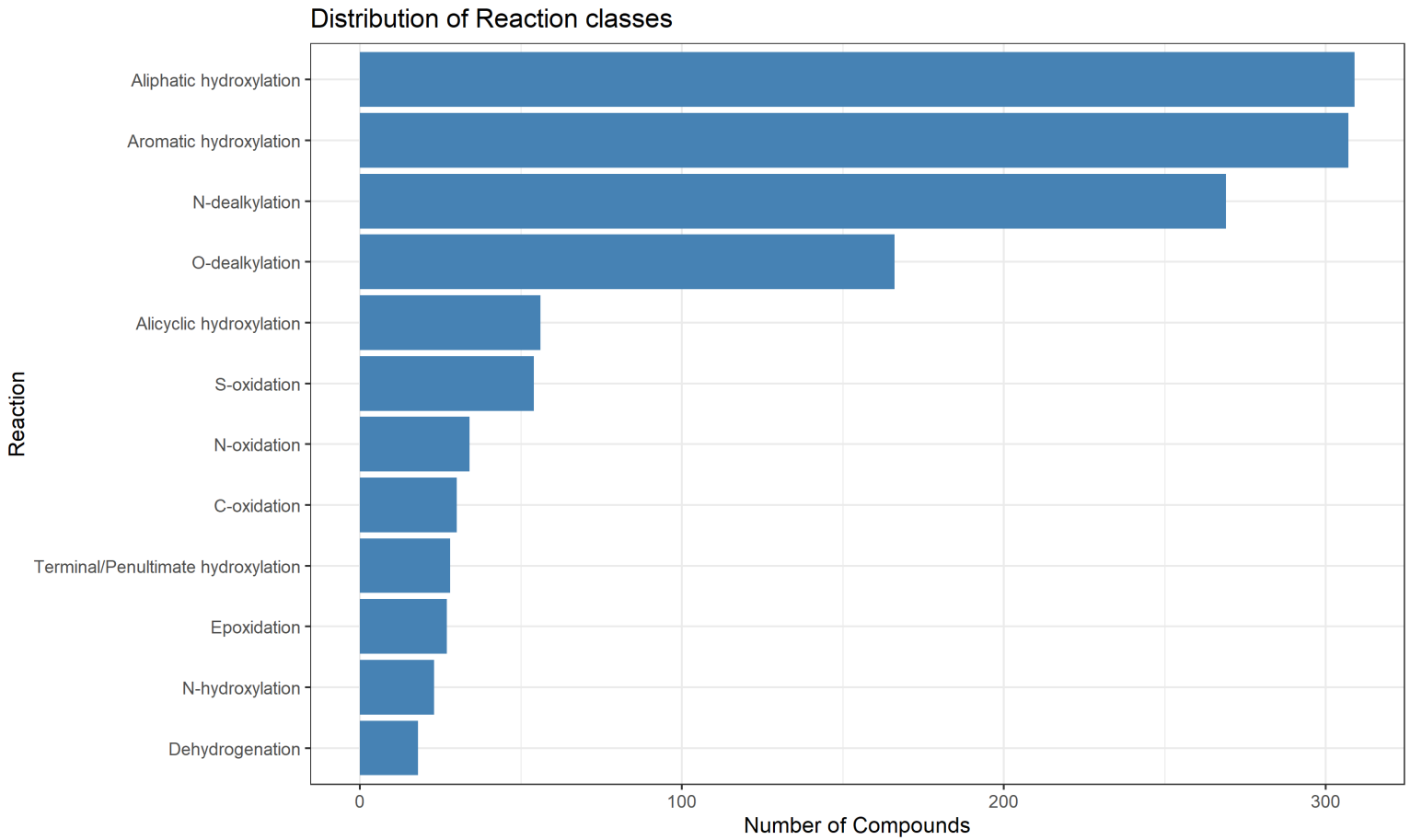
CYP3A4

One enzyme multiple MOA

- Midazolam : Aliphatic hydroxylation
- Trazodone : Aromatic hydroxylations
- Diazepam : N-demethylation
- Nimodipine : O-dealkylation
- Quetiapine : S-oxidation

MULTIPLE REACTIONS CLASSES

Reaction Class Frequency



Reduction , Cleavage, Desulfuration, Hydrolysis reactions had very few records and not shown here

THE REACTION CLASS VOCABULARY

The Most Occurring CYP-mediated reaction classes

- Aliphatic hydroxylation
- Aromatic hydroxylation
- C-oxidation
- Dehydrogenation
- Epoxidation
- N-dealkylation
- N-hydroxylation
- N-oxidation
- O-dealkylation
- S-oxidation
- Terminal / penultimate hydroxylation
- Alicyclic hydroxylation

CHALLENGE 8:
Combinatorial Explosion in Metabolite Prediction

Many Possible Combinations

Combinatorial Explosion in Metabolite Prediction

- 1 SOM × 1 reaction = 1 candidate transformation
- 10 SOMs × 12 reactions = 120 candidates transformation

CHALLENGE 9:
Descriptors To Best Represent Problem

Variety of descriptors

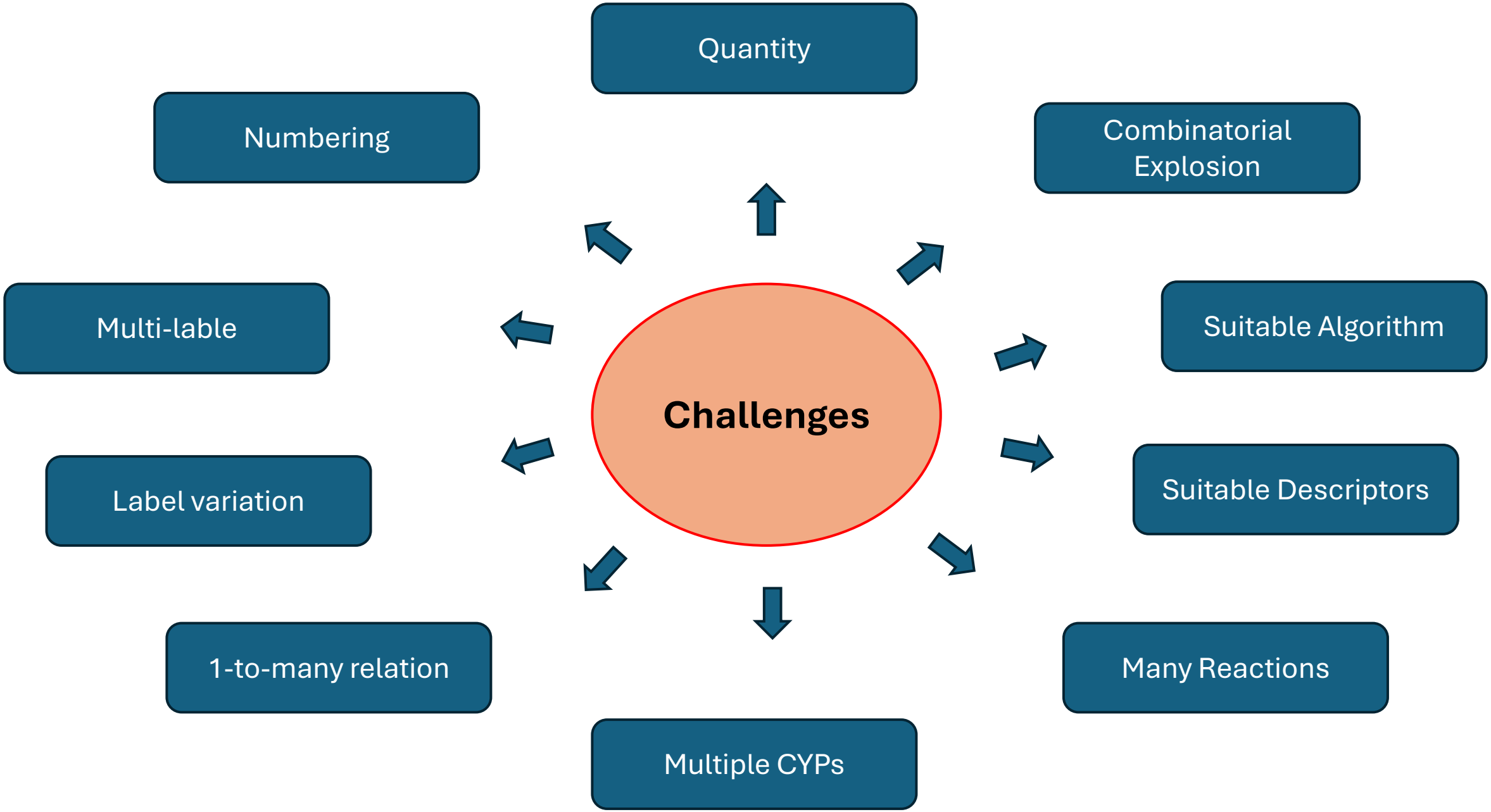
- Electronic /Reactivity Descriptors
- Topological / 2D Structural Descriptors
- Steric Descriptors
- 3D Binding Descriptors
- Fingerprints
- Descriptor-free

CHALLENGE 10:
Algorithm To Handle Several Tasks

Commonly Used Algorithm

Variety of Algorithms

- Rule based expert systems
- Docking
- Structural alerts,
- Fragment-based QSAR,
- Bayesian/probabilistic reasoning
- Machine learning
- Deep learning


Quantity
Numbering
Combinatorial Explosion
Multi-lable
Challenges
Suitable Algorithm
Label variation
Suitable Descriptors
1-to-many relation
Many Reactions
Multiple CYPs

What I Built:
METABOLITE-GEN

A Graph-Edit Network that predicts site of metabolism, reaction class, and metabolite structure end to end.

AIM

What is the goal?

To predict metabolism, we need to answer three questions together:

- **Where** will metabolism occur?
  Site of metabolism
- **What reaction** will happen?
  Reaction class
- **What structure** will form?
  Metabolite generation

Phase I metabolism Prediction

End-to-end pipeline: from a predecessor record to a ranked metabolite list

Curate

722 parent molecules, 12 reaction classes, atom-level SOM labels

Graph-encode

atom features + bond features per molecule

Dual Tasks

SOM head + reaction-class head

Generator

Generate candidate metabolites

Re-ranker

score candidate metabolites, output top-5

DATA CURATION & SPLIT

722 molecules, split to preserve scaffold diversity

722

Curated parent molecules

Unambiguous metabolite annotations, atom-indexed SOM + reaction labels

578 / 144

Pre-training pool / held-out test

80% / 20% split — the test set is never touched during model selection

5

Scaffold-stratified CV folds

~462 training / ~ 116 validation pairs per fold, by Murcko scaffold group

CONTINUED

Reaction-class imbalance

Reaction class	Training (per fold)	Validation (per fold)	Test set
Aliphatic hydroxylation	**137**	**34**	**43**
N-dealkylation	**129**	**32**	**36**
Aromatic hydroxylation	**110**	**27**	**21**
O-dealkylation	**94**	**23**	**19**
S-oxidation	**26**	**6**	**7**
Alicyclic hydroxylation	**18**	**4**	**6**
C-oxidation	**18**	**5**	**3**
Epoxidation	**13**	**3**	**4**
Terminal / penultimate hydrox.	**9**	**2**	**3**
N-hydroxylation	**12**	**3**	**1**
Dehydrogenation	**10**	**2**	**1**
N-oxidation	**4**	**1**	**1**
Total (approx. per-fold)	**580**	**142**	**145**

MODEL ARCHITECTURE

GEN: a shared GATv2 encoder, two task-specific heads

GATv2 Encoder

- 4 graph-attention layers
- 4 attention heads
- DropEdge p = 0.1
- ELU activation
- 128 hidden (128 × 4 heads = 512 dim)

SOM head

sigmoid per atom
Binary focal loss

Reaction-class head

12-way softmax
Multi-class focal loss

Per-atom SOM score + per-molecule reaction-class distribution

Graph-Edir Networks (GEN)

A metabolic reaction is a minimal graph edit, not a new molecule from scratch

the smallest number of node/edge insertions, deletions, or relabelling operations needed to transform one graph into another.

- A biotransformation is, almost by definition, a small, local graph edit: a hydroxyl group added at one node, a bond broken in an N-dealkylation, an oxygen inserted across a double bond in an epoxidation
- Instead of asking the model to generate an entire product molecule from nothing, GEN asks two narrower questions:

  -where on the parent graph is the edit anchored (the SOM head),

  -what kind of edit is it (the reaction-class head)
- The candidate metabolite is then produced by applying a deterministic, chemistry-specific transform — the actual graph edit — at the predicted site, for the predicted reaction class
- This reframes "predict a new molecule" as "predict a small, well-defined edit to a molecule you already have" — a fundamentally easier and more data-efficient learning problem

Graph-edit framing turns an open-ended generation problem into a constrained where + what problem.

FROM PREDICTIONS TO RANKED METABOLITES

A learned reranker collapses the SOM × reaction combinatorics

- Top-5 SOM atoms × top-5 reaction classes are combined, and a deterministic, reaction-specific transform generates a candidate metabolite structure for each pairing — duplicates removed by canonical SMILES
- Each surviving candidate is scored by a 3-layer MLP, taking the parent-molecule embedding, a 2048-bit Morgan fingerprint of the candidate, and the joint SOM–reaction probability as input
- Trained with a diversity-aware listwise ranking loss, so the top-5 output isn't five near-duplicate variations of the same guess
- This is precisely the mechanism that addresses the combinatorial-explosion problem — instead of reporting one unranked atom and one unranked reaction class, the model reports a ranked, deduplicated shortlist of actual candidate molecules

Top-5

RANKED OUTPUT

Reported per parent molecule, scored by a dedicated reranking MLP — not just the raw SOM/reaction logits.

TRAINING PROTOCOL

Hyperparameters used across all folds

Hyperparameter	Value
GNN layers / attention heads	4 / 4
Hidden dim (SOM / Reaction heads)	128 / 128
Reranker hidden dims	256
DropEdge probability (train only)	0.1
SOM loss	Binary focal loss
Reaction-class loss	Multi-class focal loss
SMILES augmentation factor (minority classes, train only)	2×
Optimiser	Adam
Max epochs / early-stopping patience	300 / 30
Early-stopping criterion	Joint score = 0.6 × Top-1 + 0.4 × MRR
Cross-validation folds	5 (scaffold-stratified)
Batch size	32
seed	123

NOTE: val_top1 is the fraction of validation molecules for which the rank-1 predicted metabolite is a hit, val_MRR is the mean reciprocal rank of the first hit across all validation molecules

The joint score — a blend of metabolite top-1 hit rate and ranking quality — is used for model selection precisely because SOM accuracy alone does not reflect end-to-end usefulness.

RESULTS

5-fold cross-validation results for SOM and reaction prediction.

Fold	Molecules	SOM atom-ranking accuracy			MRR	Weighted F1
		Top-1 (%)	Top-3 (%)	Top-5 (%)		
1	116	49.14	74.14	85.34	0.6174	0.5563
2	116	50	81.9	91.38	0.6593	0.4582
3	116	15.52	64.66	80.17	0.4076	0.5184
4	115	57.39	77.39	82.61	0.6736	0.5446
5	115	58.26	78.26	92.17	0.7017	0.5179

RESULTS

5-fold cross-validation results for metabolite prediction.

Fold	Molecules	Top-k correct predictions			Top-k Accuracy (%)		
		Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
1	116	67	83	88	57.76	71.55	75.86
2	116	52	78	80	44.83	67.24	68.97
3	116	56	67	69	48.28	57.76	59.48
4	115	61	76	80	53.04	66.09	69.57
5	115	67	91	95	58.26	79.13	82.61

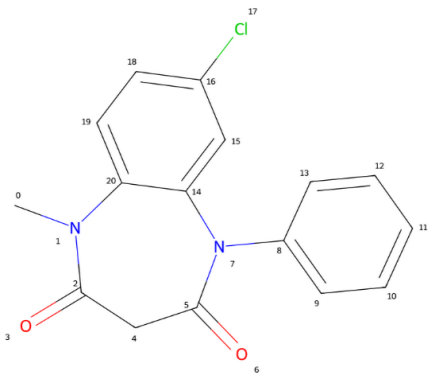
RESULTS

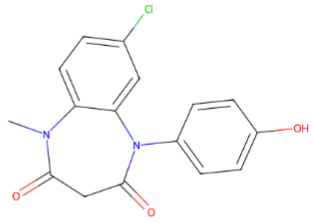
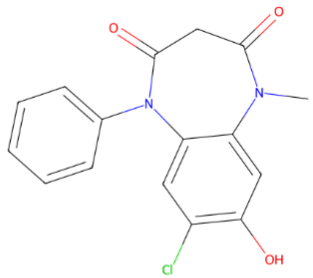
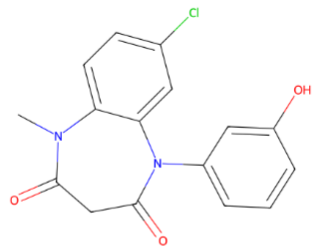
GEN performance on the held-out test set (n = 144 molecules)

SOM atom ranking accuracy (%)				Reaction prediction	Metabolite prediction accuracy (%)		
Top-1	Top-3	Top-5	MRR	Weighted F1	Top-1	Top-3	Top-5
53.5	75.7%	84.7%	0.66	0.51	46.53	64.58	69.44

RESULTS

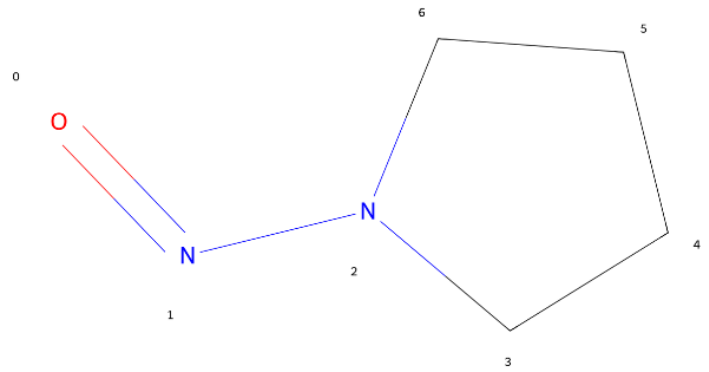
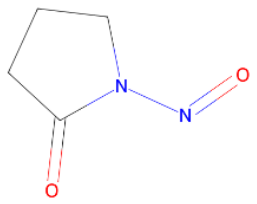
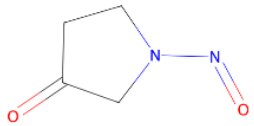
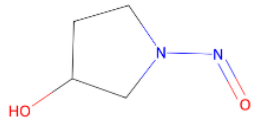
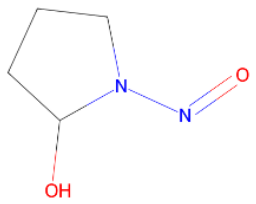
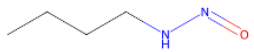
Correct Prediction

Parent		**GT-1** SOM : 11 (C) RXN : Aromatic hydroxylation

Predicted Metabolites			
			
	Rank : 1 Pred SOM : Atom 11 (C) SOM proba : 0.501 Pred RXN : Aromatic hydroxylation RXN proba : 0.828	**Rank :** 2 Pred SOM : Atom 18 (C) SOM proba : 0.443 Pred RXN : Aromatic hydroxylation RXN proba : 0.828	**Rank :** 3 Pred SOM : Atom 10 (C) SOM proba : 0.437 Pred RXN : Aromatic hydroxylation RXN proba : 0.828

RESULTS

Incorrect Prediction

Parent					**GT-1** SOM : 6 (C) RXN : Aliphatic hydroxylation
Predicted Metabolites					
	Rank : 1 Pred SOM : Atom 3 (C) SOM proba : 0.463 Pred RXN : C-oxidation RXN proba : 0.004	**Rank :** 2 Pred SOM : Atom 4 (C) SOM proba : 0.440 Pred RXN : C-oxidation RXN proba : 0.004	**Rank :** 3 Pred SOM : Atom 4 (C) SOM proba : 0.440 Pred RXN : Alicyclic hydroxylation RXN proba : 0.012	**Rank :** 4 Pred SOM : Atom 3 (C) SOM proba : 0.463 Pred RXN : Alicyclic hydroxylation RXN proba : 0.012	**Rank :** 5 Pred SOM : Atom 2 (N) SOM proba : 0.469 Pred RXN : N-dealkylation RXN proba : 0.096

KEY TAKEAWAYS

Metabolism prediction is hard for structural, not just statistical, reasons

- It is not just a "need more data" problem — the data we have is small, unevenly distributed, and the labels themselves are ambiguous by construction
- Site-of-metabolism, reaction class, and metabolite structure are three coupled but distinct prediction problems, and treating them as one collapses real uncertainty into false confidence
- Graph-based, relational modelling genuinely helps with the numbering problem — but it does not, by itself, fix data scarcity or enzyme-label ambiguity
- Honest evaluation — top-k metrics, scaffold-disjoint test sets, and reporting where the model fails — matters as much as the headline accuracy number
- GEN is a step forward on this problem, not a solved problem

THANK YOU!

- Questions: swapnil.chavan@ri.se