

What are we missing to characterize small molecules as systematically as proteins?

Filip Jozefov



IOCB
PRAGUE



CZECH INSTITUTE
OF INFORMATICS
ROBOTICS AND
CYBERNETICS
CTU IN PRAGUE

M U N I

cesnet
.....



AI is changing how we study proteins

A protein's sequence can guide a prediction of its three-dimensional shape.

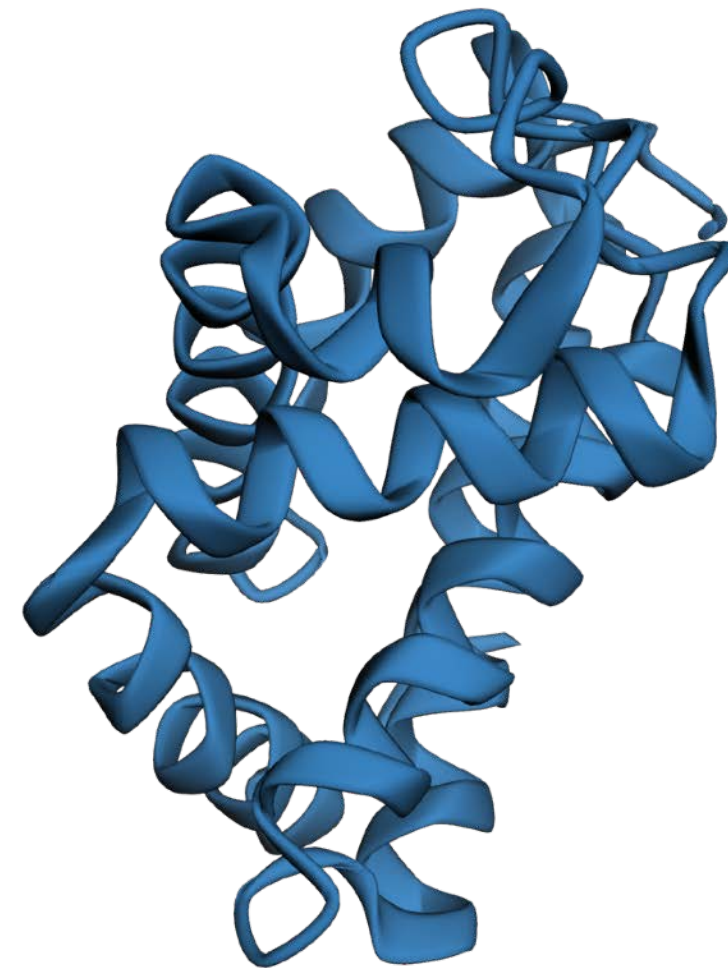
Amino-acid sequence

MGLSDGEWQLVLNVWGKVEA
DIPGHGQEVLIIRLFKGHPET
LEKFDKFKHLKSEDEMKASE...

Human myoglobin

A protein that stores oxygen in muscle

AlphaFold



Predicted 3D structure



Nobel Prize in Chemistry 2024

Protein design and structure prediction

200M+

predictions openly available

AlphaFold DB milestone · July 2022

Google DeepMind EMBL-EBI 

A structural model gives researchers a starting point for experiments.

What made this possible?

Protein research built a shared foundation that AI could learn from.

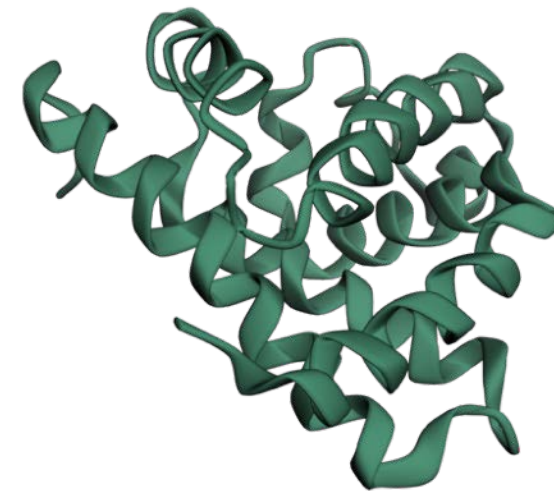
A common language

MGLSDGEWQLVL
NVWGKVEADIPG
HGQEVLIIRLFKG

Sequences we can compare



Shared experiments



Structures we can learn from



AI and computing

128

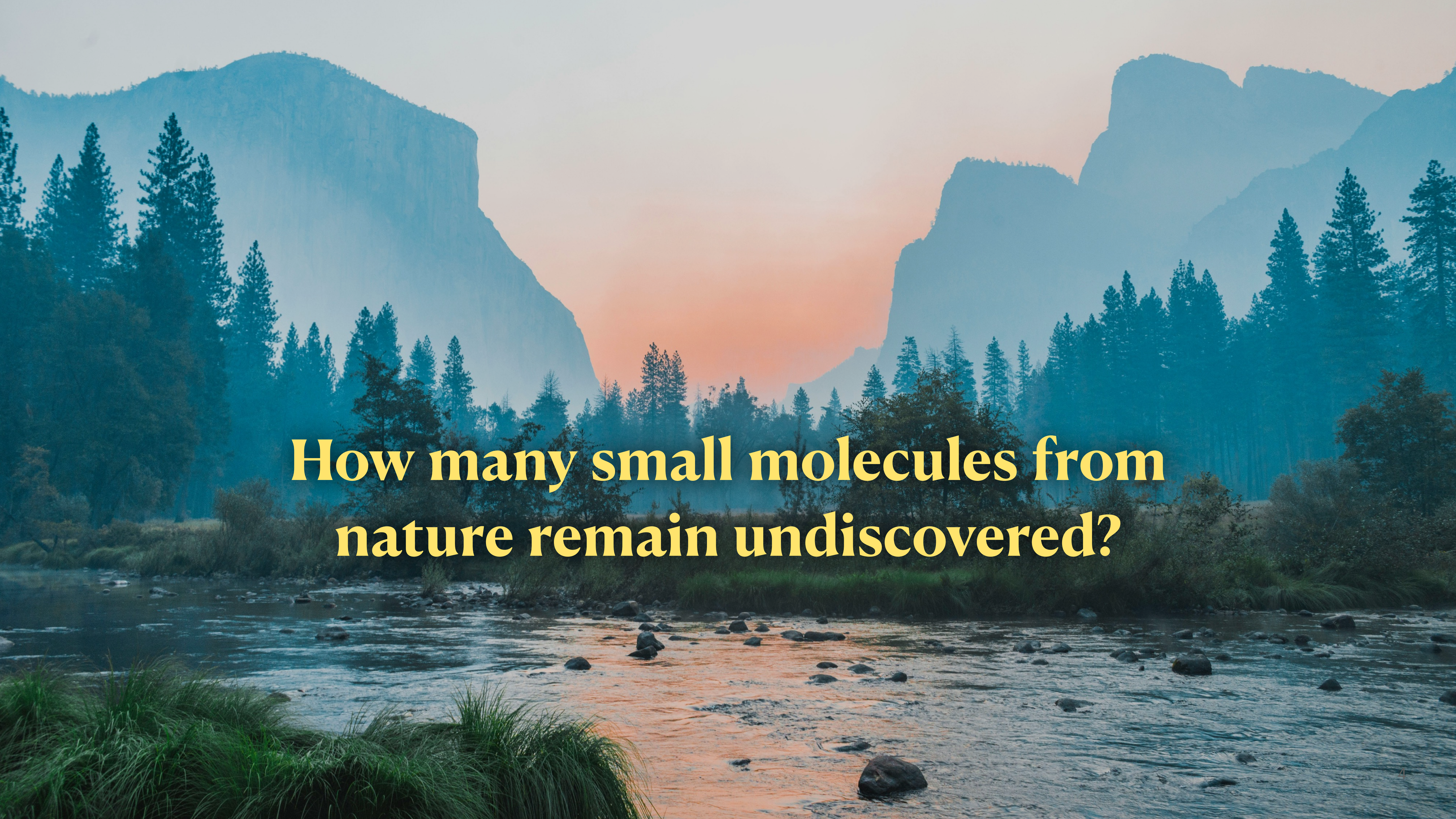
TPU v3 cores

Original AlphaFold 2 training

Then inference across millions of proteins

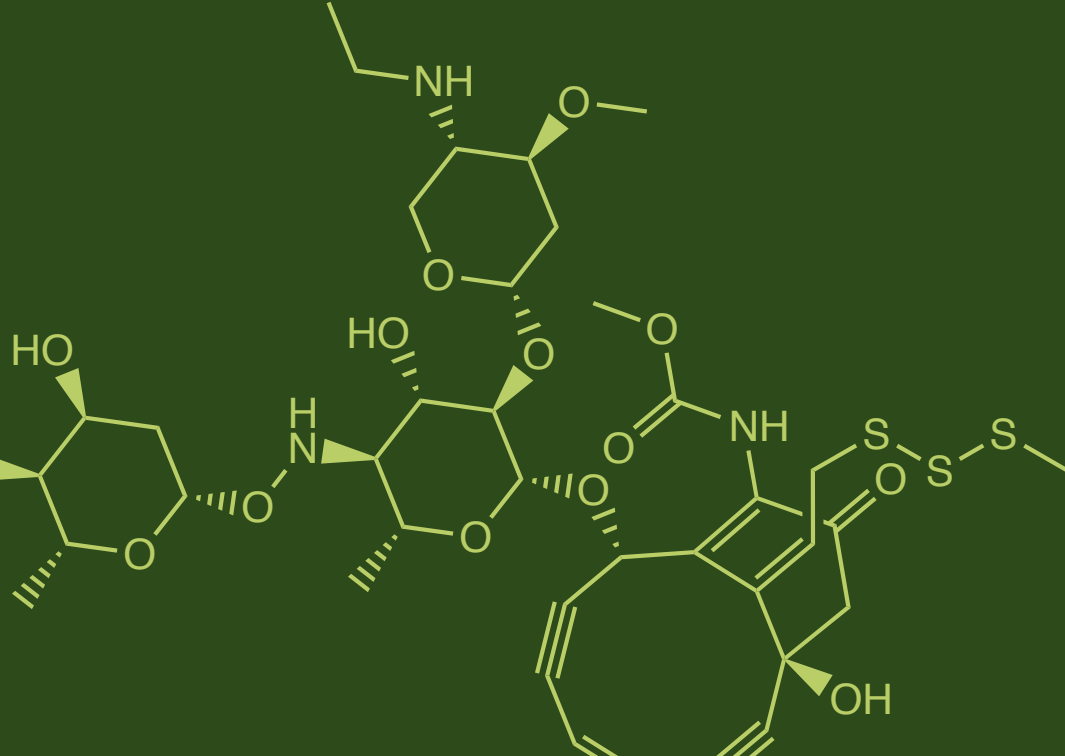
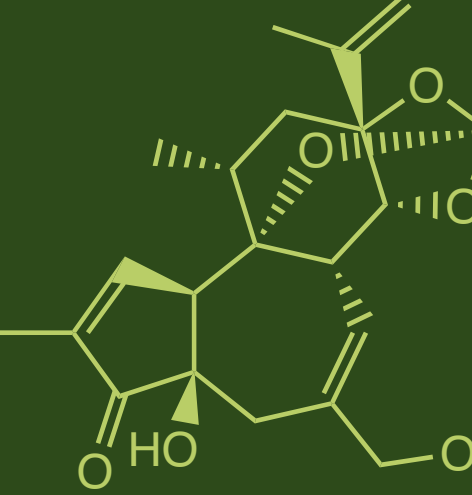
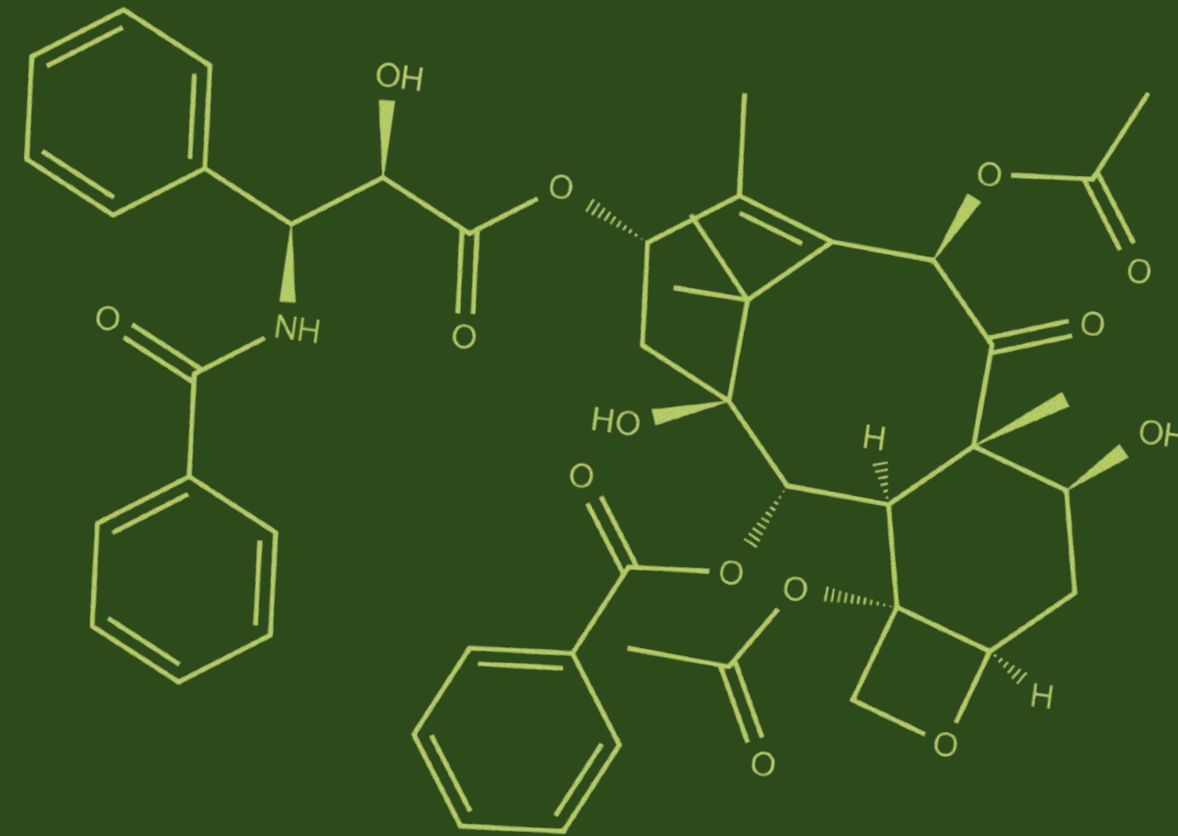
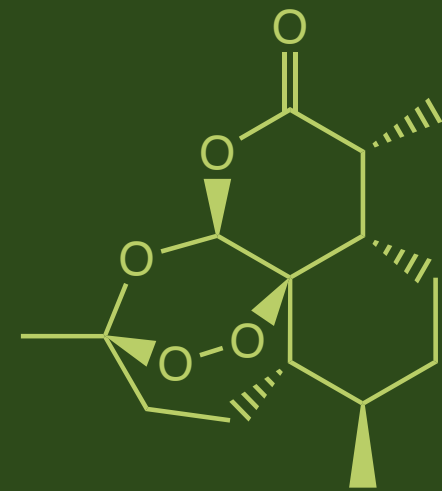
Infrastructure makes the data reusable: storage, access and search.

What would it take to do this for small molecules?

A scenic landscape photograph of a river flowing through a valley. In the foreground, there are green grasses and a river with rocks. The middle ground shows a dense forest of evergreen trees. In the background, there are large, rugged mountains under a sky with a warm, orange and pink sunset glow. The text "How many small molecules from nature remain undiscovered?" is overlaid in the center in a bold, yellow, serif font.

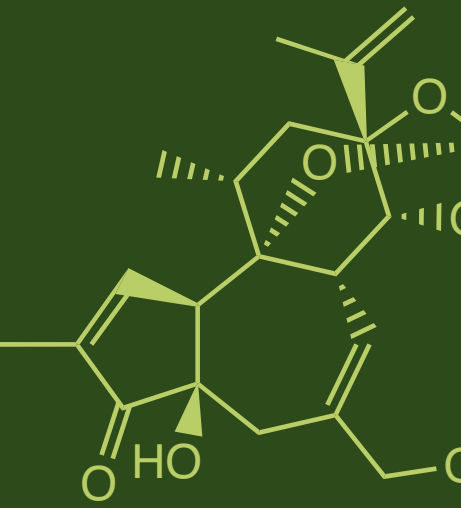
**How many small molecules from
nature remain undiscovered?**

>90% of natural products remain undiscovered



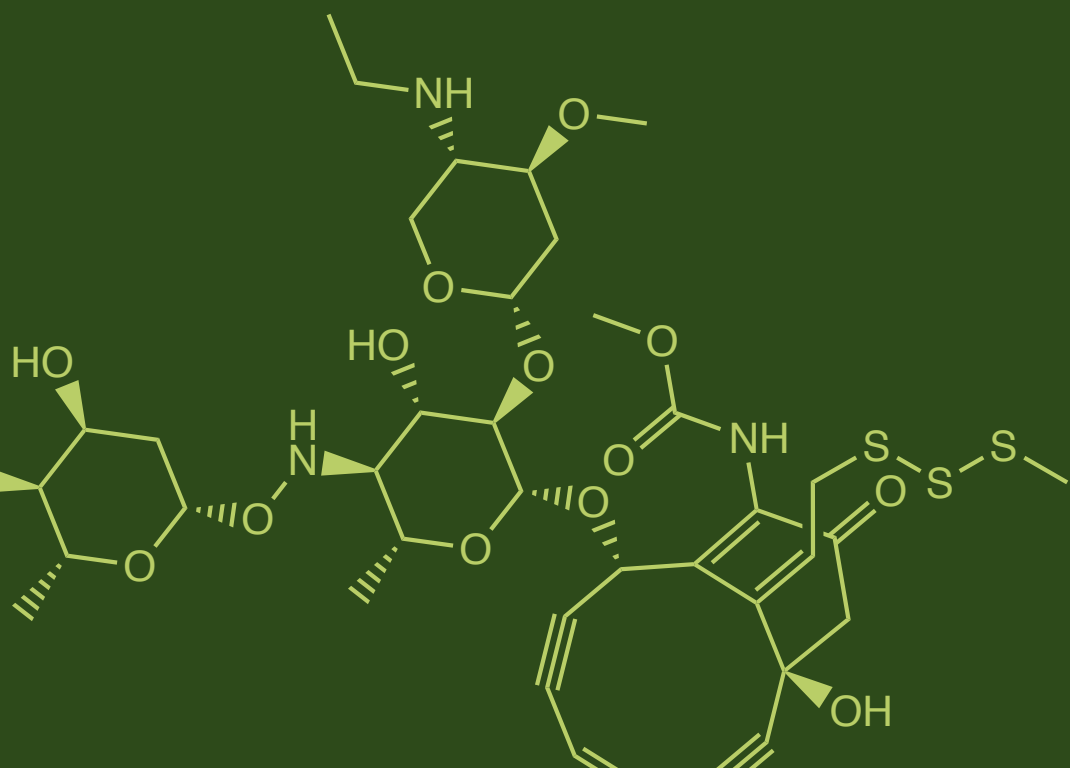


The image displays a complex chemical structure, likely a diterpene derivative, featuring a central polycyclic core with multiple stereocenters. The structure includes a carboxylic acid group (HO-C=O) and a side chain containing a benzamide moiety (NH-C(=O)-C₆H₅) and a hydroxyl group (OH). The molecule is shown in a 3D representation with wedged and dashed bonds indicating stereochemistry.



Taxol

Chemotherapy drug

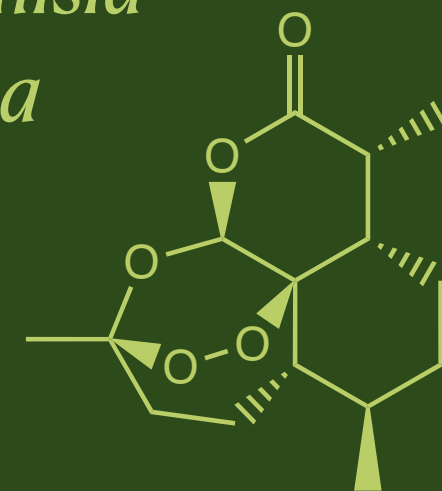


>90% of natural products remain undiscovered

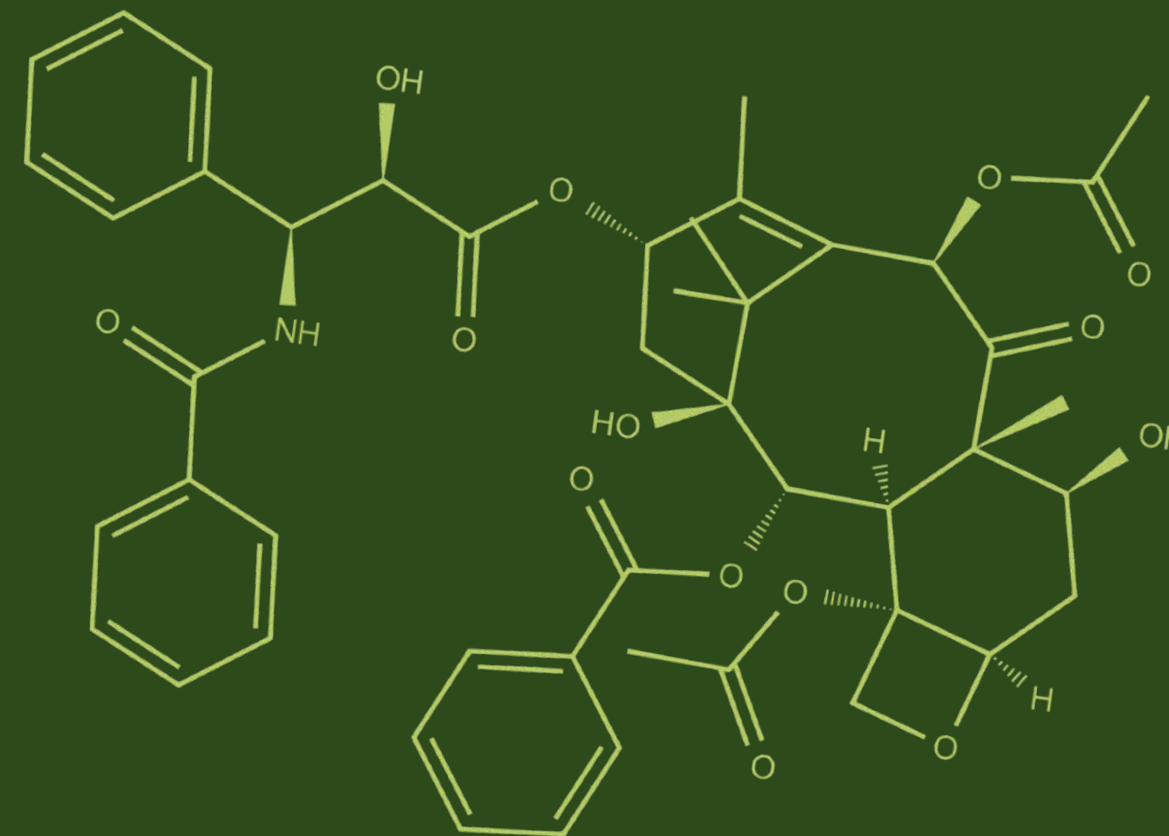
While most of the small-molecule drugs are derived/inspired by natural products



*Artemisia
annua*

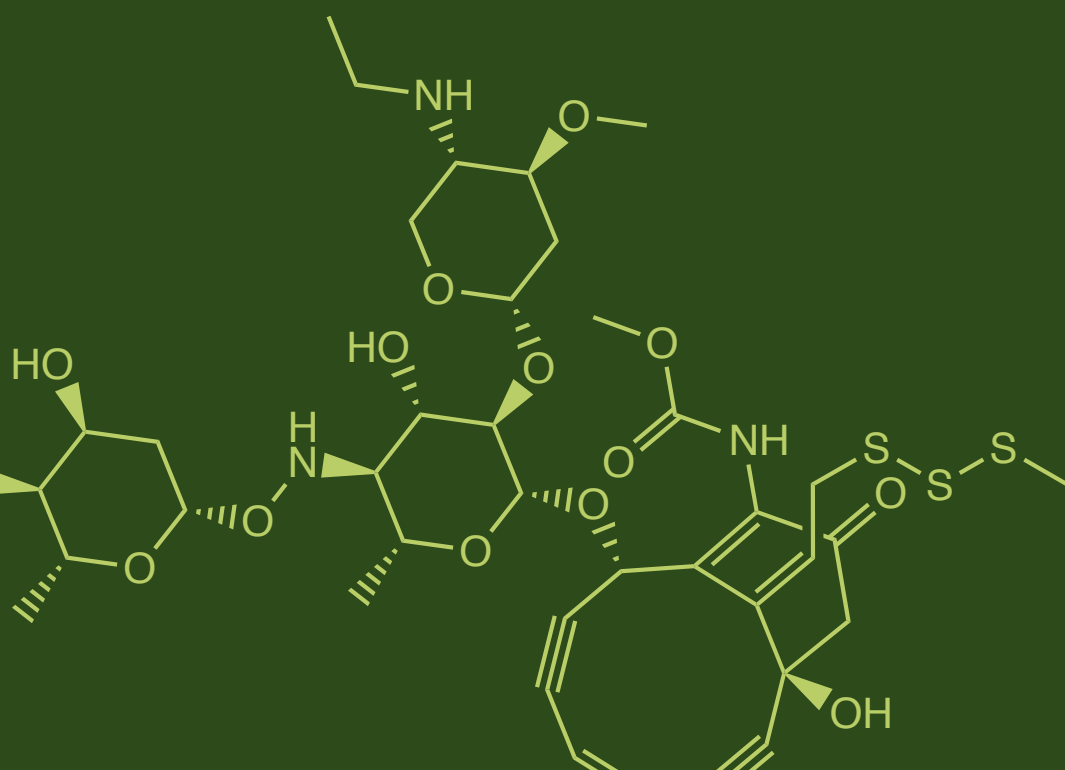
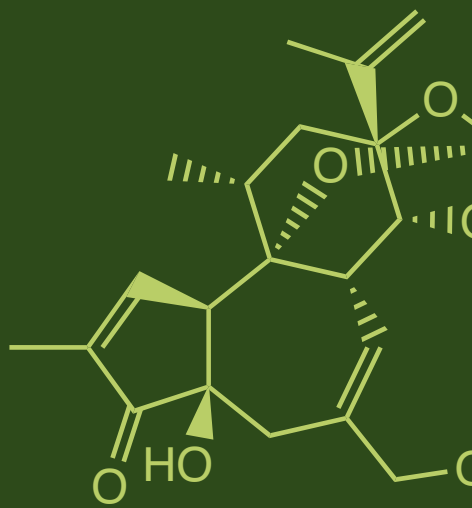


Artemisinin
Anti malaria



*Taxus
brevifolia*

Taxol
Chemotherapy
drug



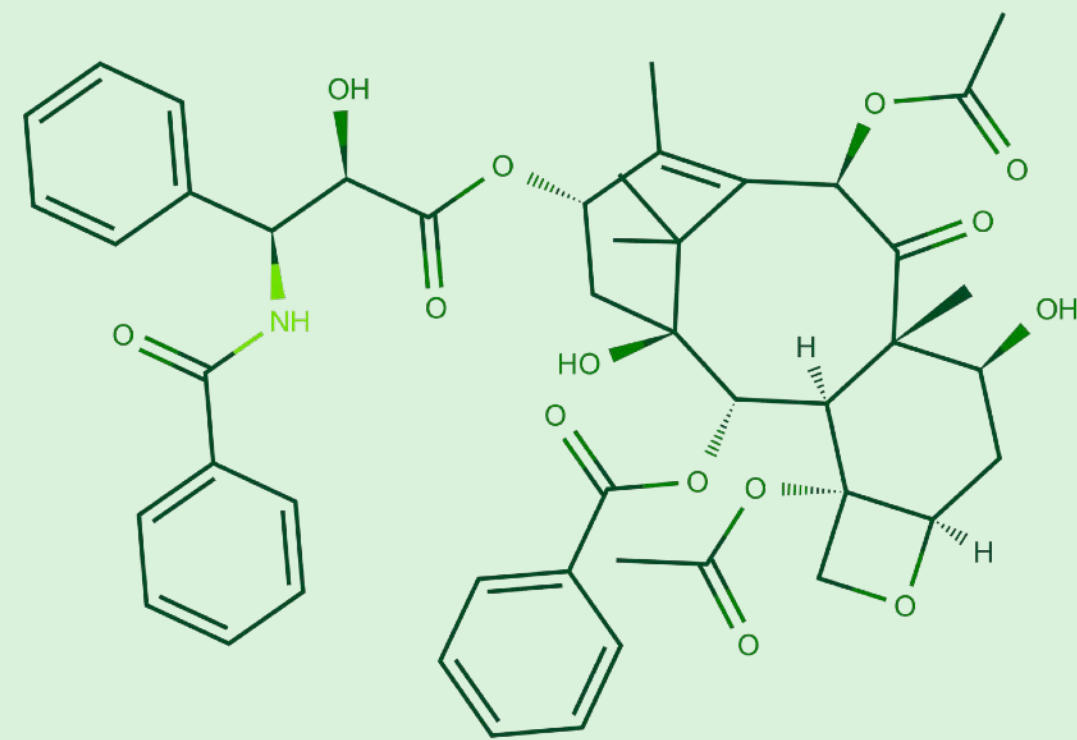
A garden yew offered another route to Taxol

A chemotherapy medicine

Yew bark



Taxus brevifolia bark



Paclitaxel (Taxol)

40–165 mg

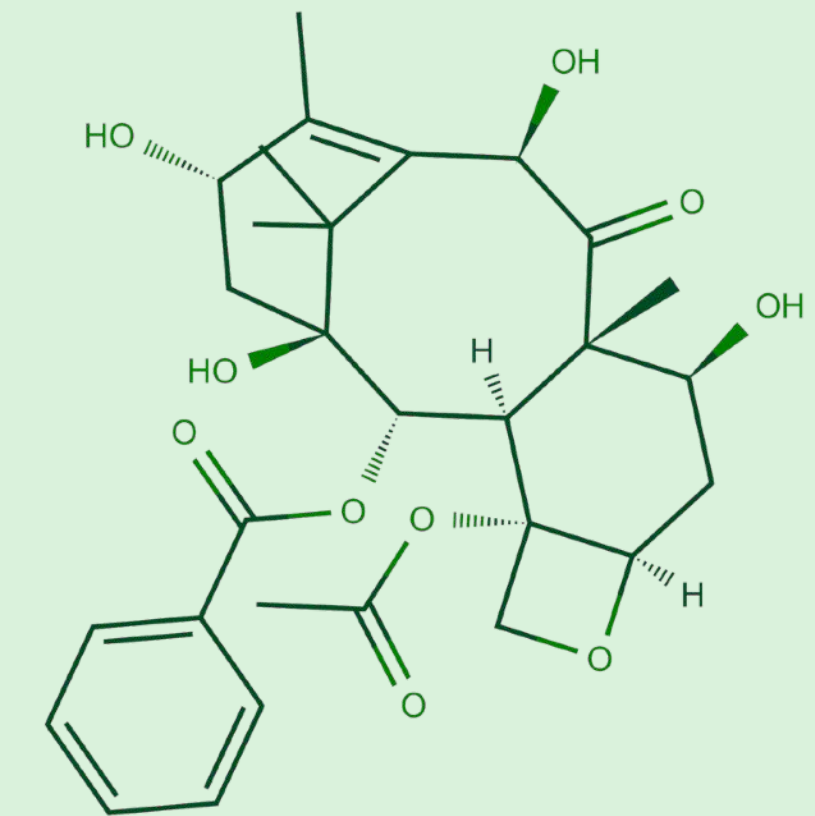
Taxol per kg of bark*

Bark harvesting could kill the tree.

English yew needles



Taxus baccata



10-deacetylbaccatin III

≈1,000 mg

precursor per kg of fresh needles

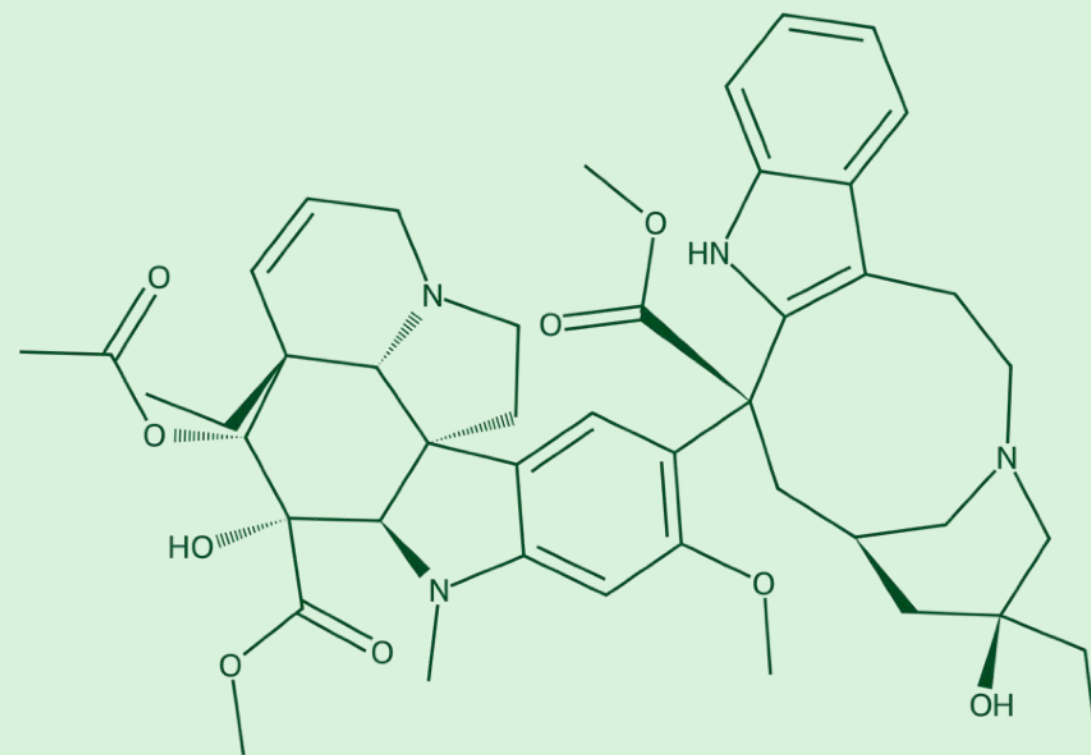
Needles regrow after careful harvesting.

Chemists could turn this garden-yew precursor into Taxol.

*Historical isolation yields, not a direct concentration comparison. Bark range spans yew species. Denis et al., JACS (1988).

Useful molecules can be hard to obtain from plants

Vinblastine



*Nothapodytes
nimmoniana*



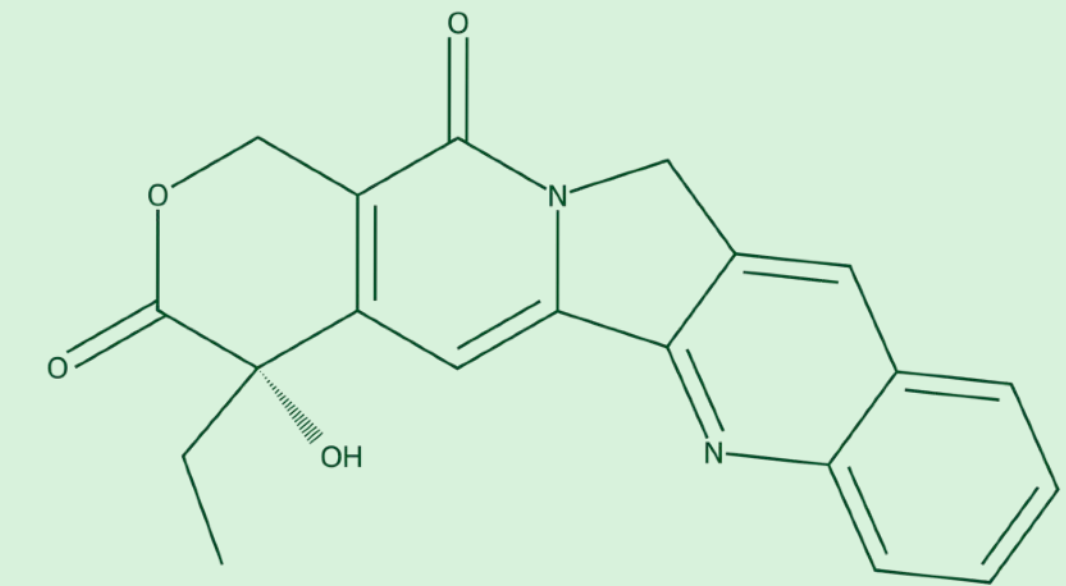
Anticancer medicine



Catharanthus
roseus

1,000 kg of dry leaves contain **1 g**
at that concentration

Camptothecin



Starting material for anticancer medicines



40 mg / kg



1,910 mg / kg

48× more recovered from roots



Could we find the important connections sooner?

Anticancer medicine

anticancer medicines

40 mg / kg

Mass spectrometry archive and search engine



*Taxus
brevifolia*

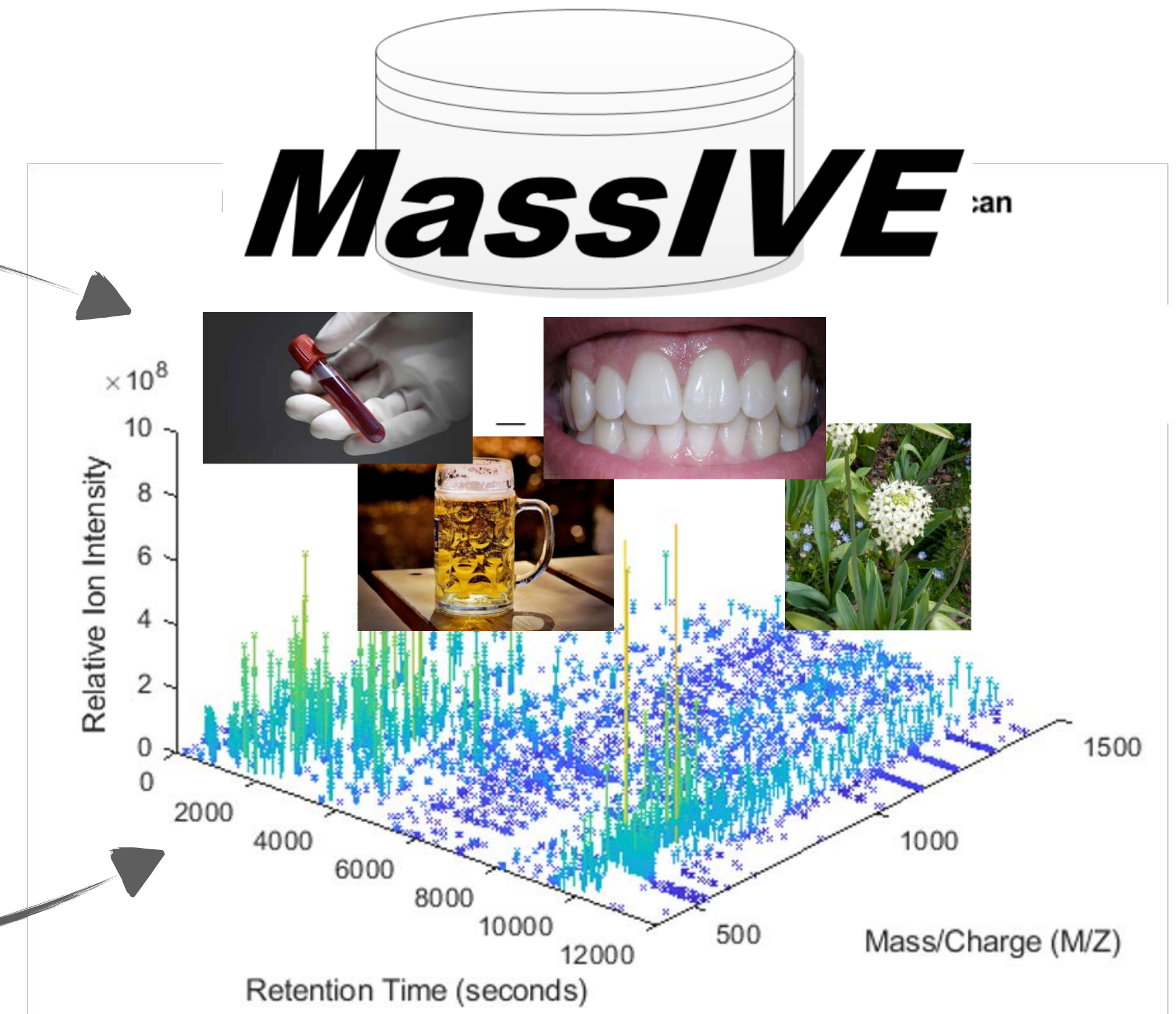


Taxus baccata



*Catharanthus
roseus*

Sampling and measurements

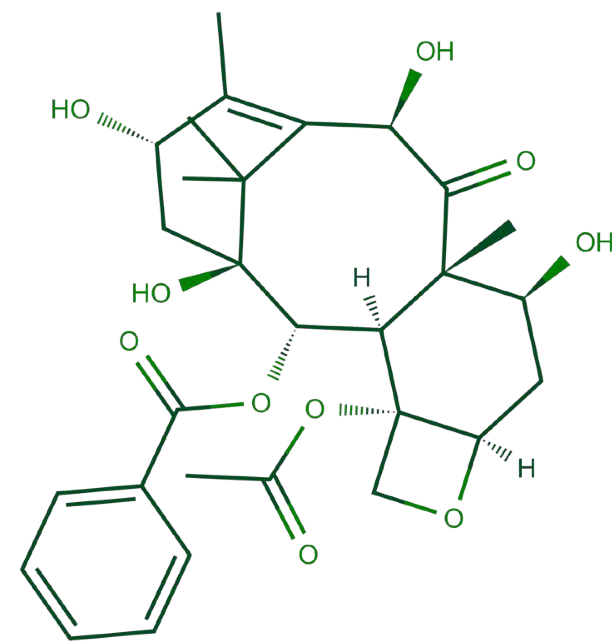


**Billions of MS measurements
from all around the world**

Different questions. A shared collection of measurements.

Plant chemistry

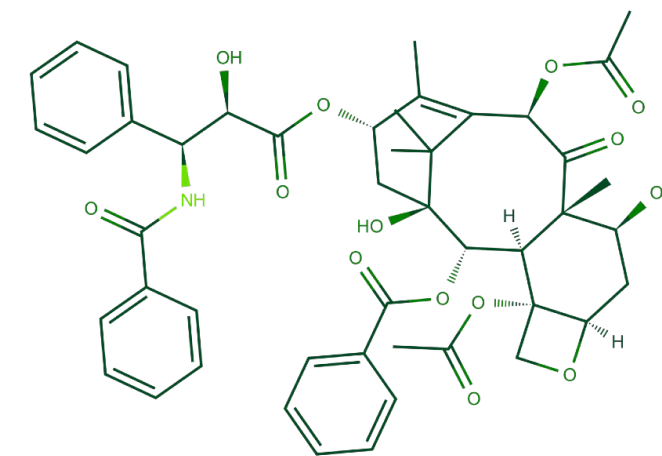
Which taxanes are present in yew needles?



10-deacetylbaccatin III

Drug discovery

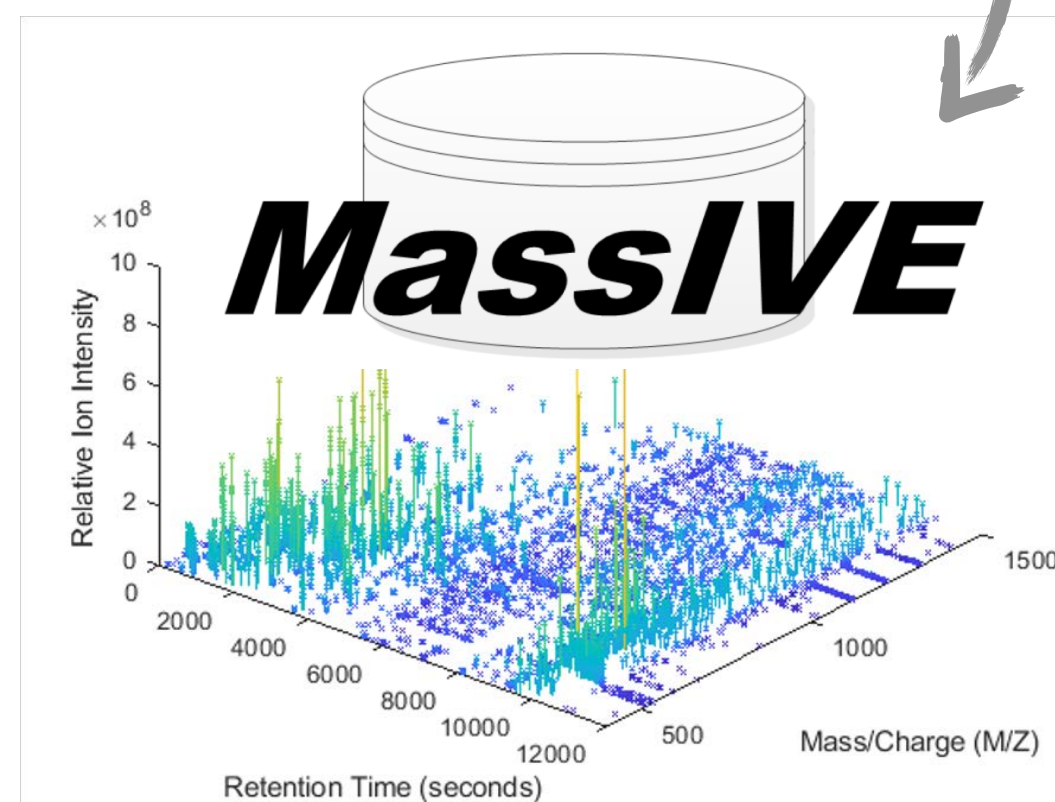
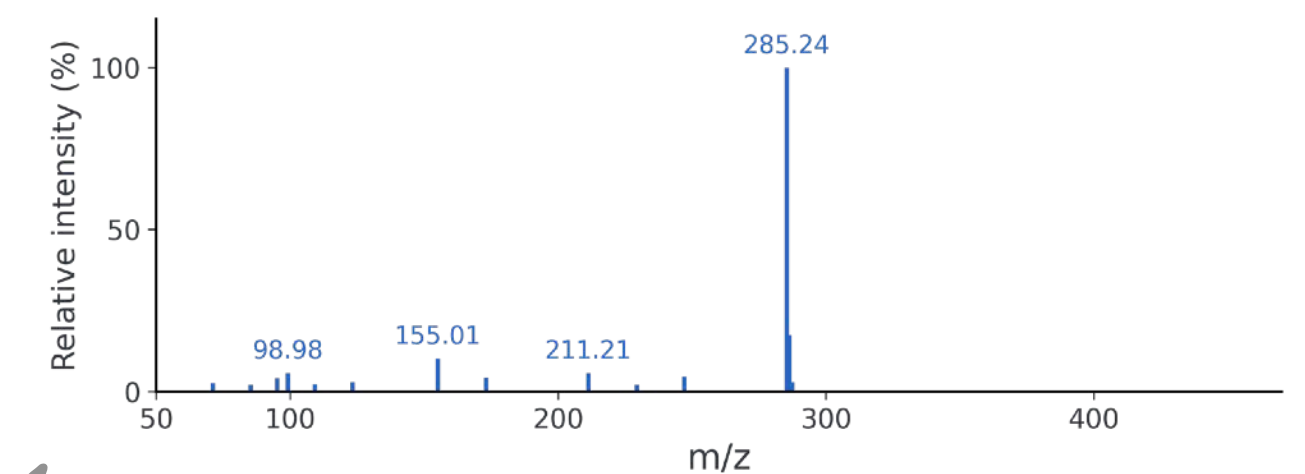
Which plant compounds have anticancer activity?



Paclitaxel (Taxol)

Food chemistry

What molecules are present in our food?



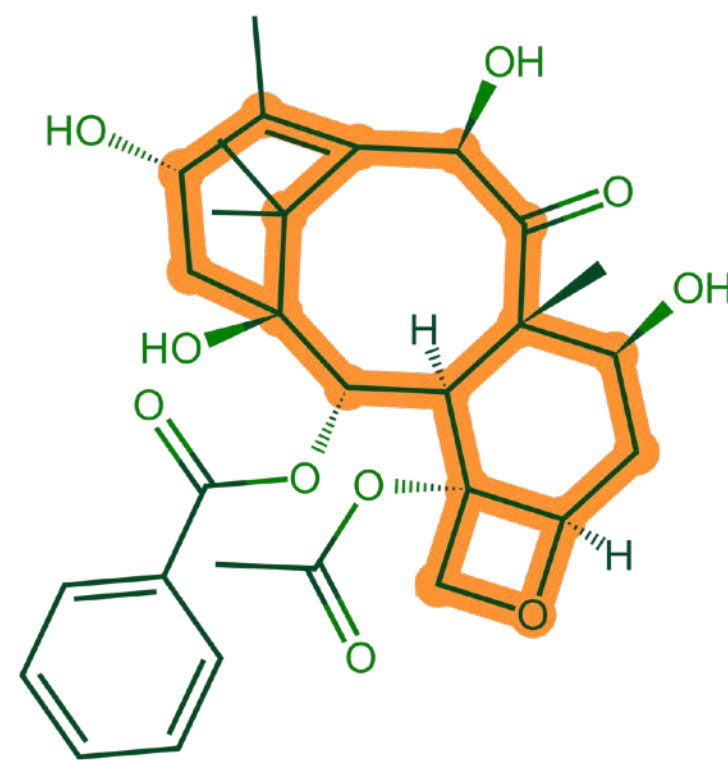
Researchers can share their measurements
in archives such as MassIVE.

Related molecules can connect different scientific questions

Which taxanes are present in yew needles?



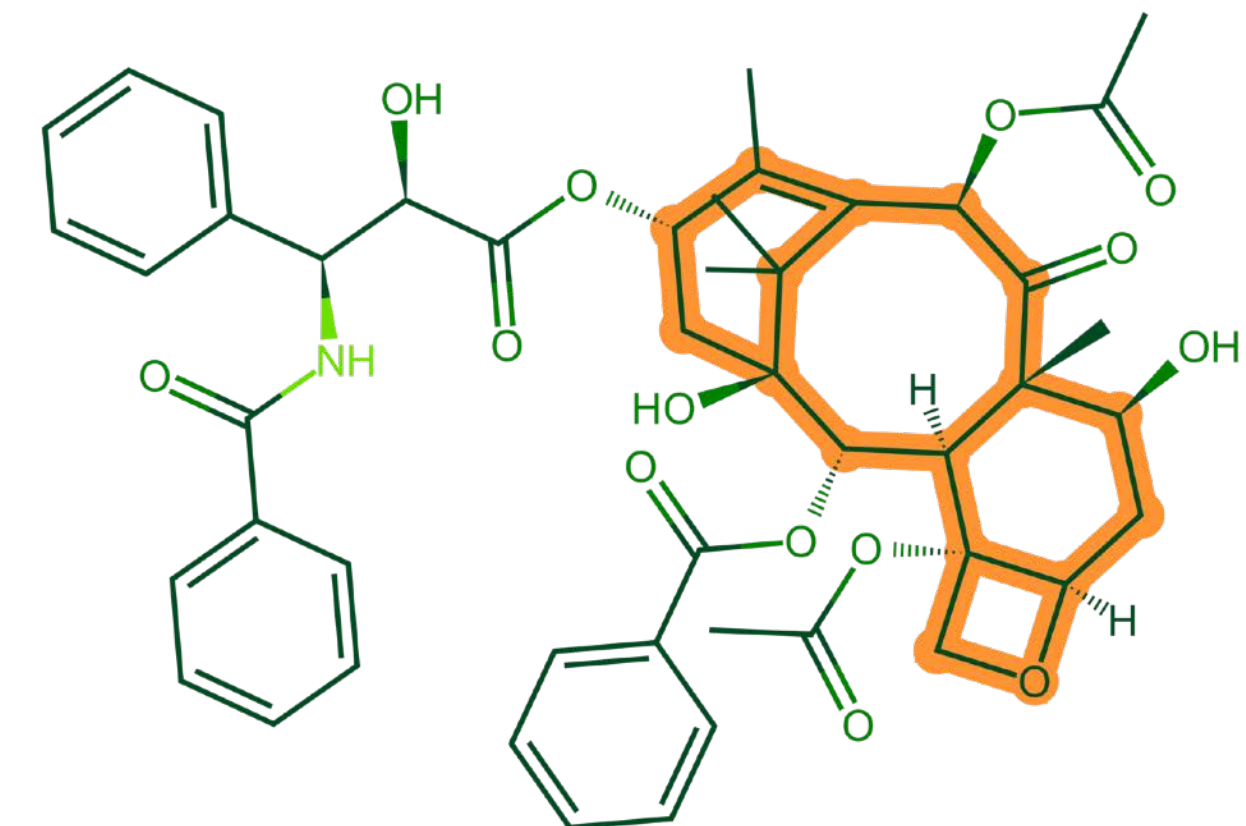
10-deacetylbaccatin III



Which plant compounds have anticancer activity?



Paclitaxel (Taxol)



Semisynthesis



Chemical steps turn
the precursor into Taxol

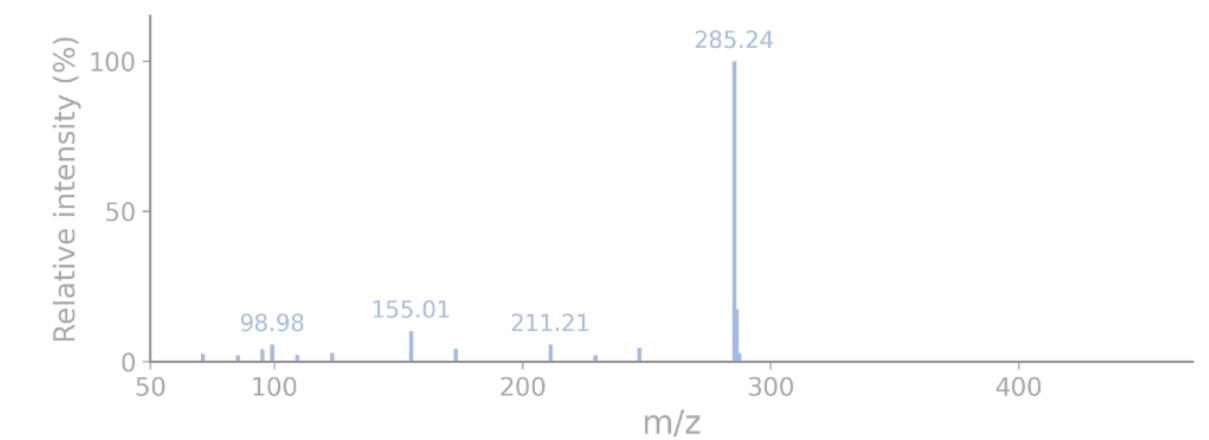
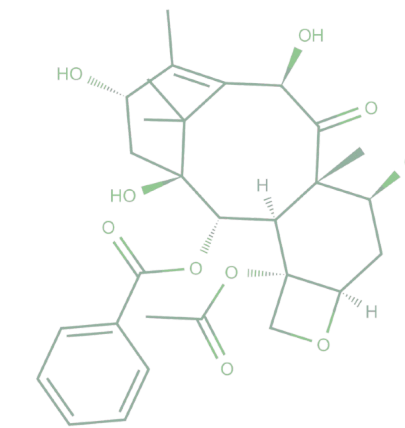
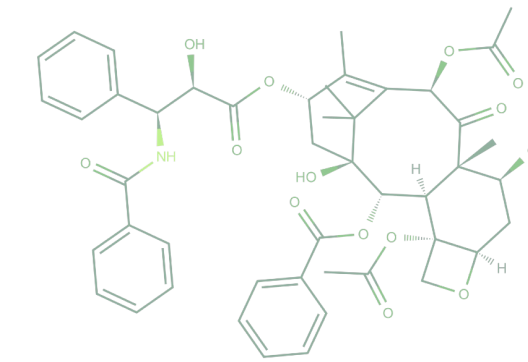
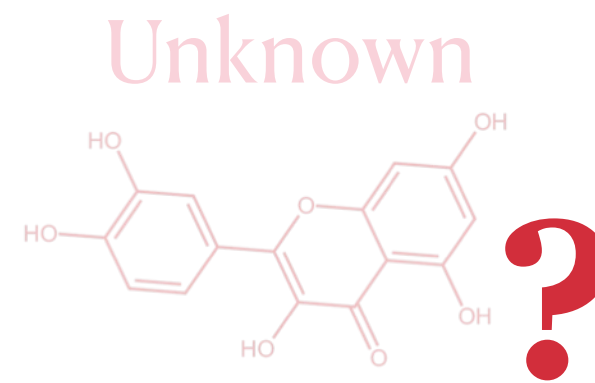
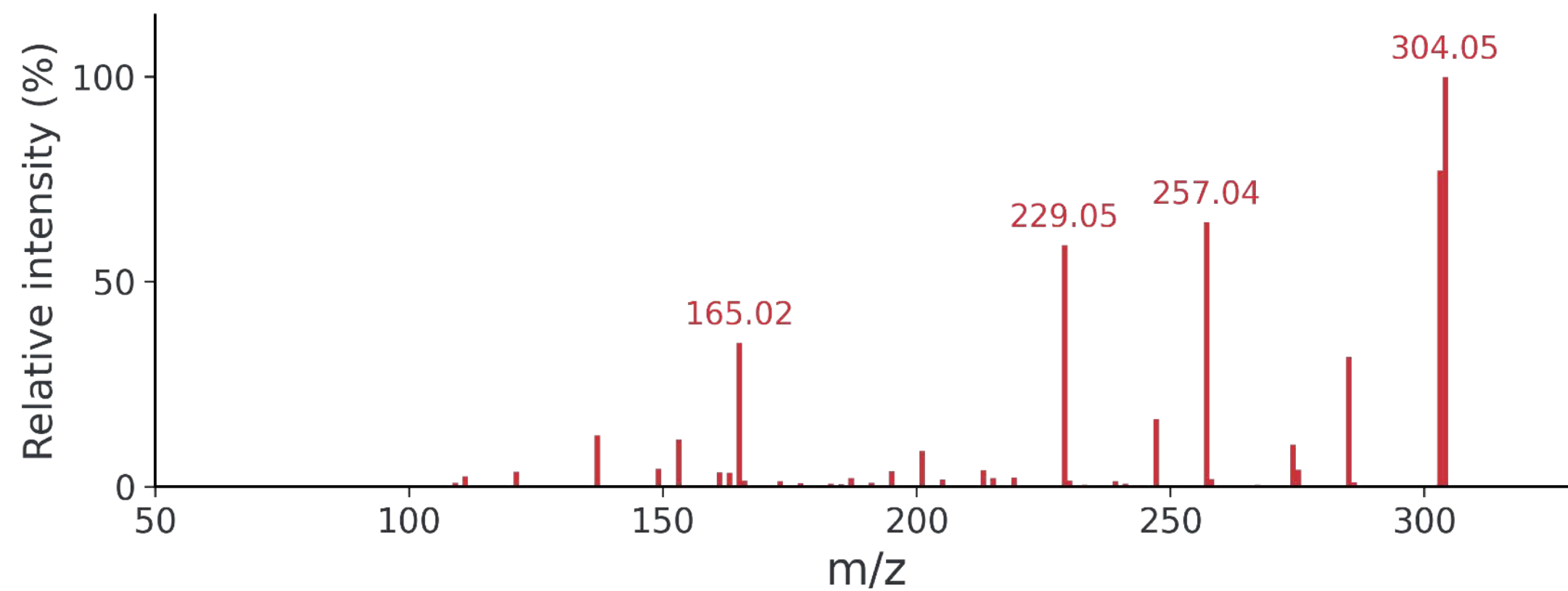
A related molecule provided another route to Taxol.

Could another study help explain your spectrum?

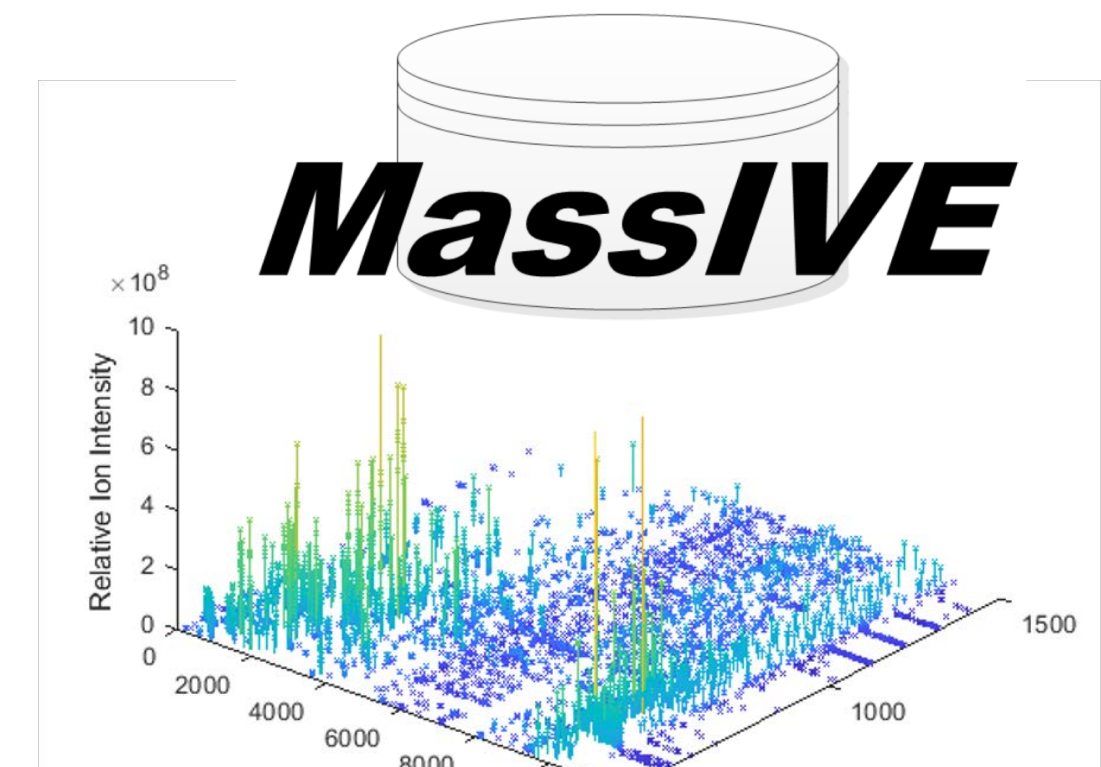


Your plant sample

An unfamiliar signal



Could your signal be related to something already measured?

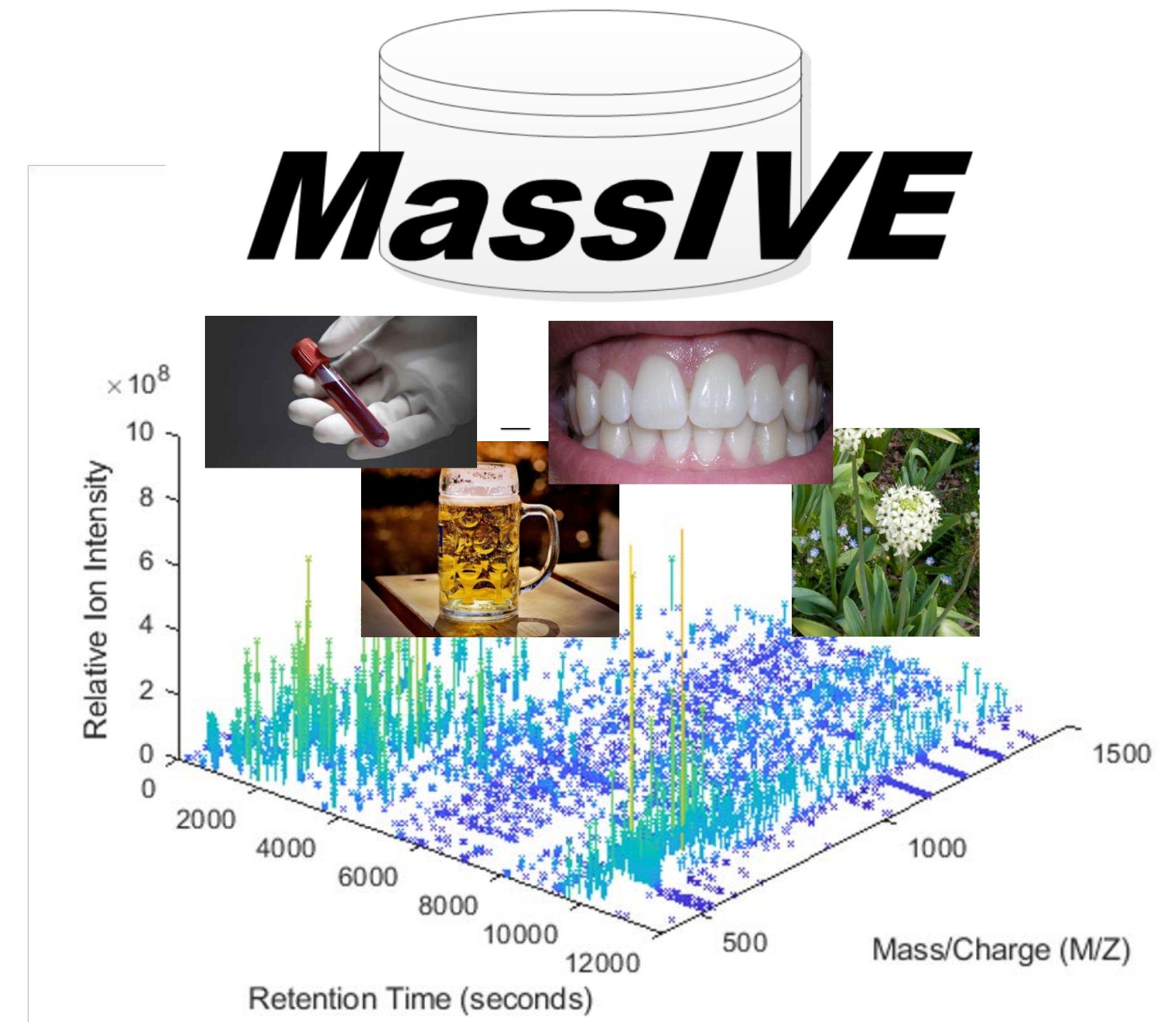


A vast collection of measurements

How do we find that connection at this scale?

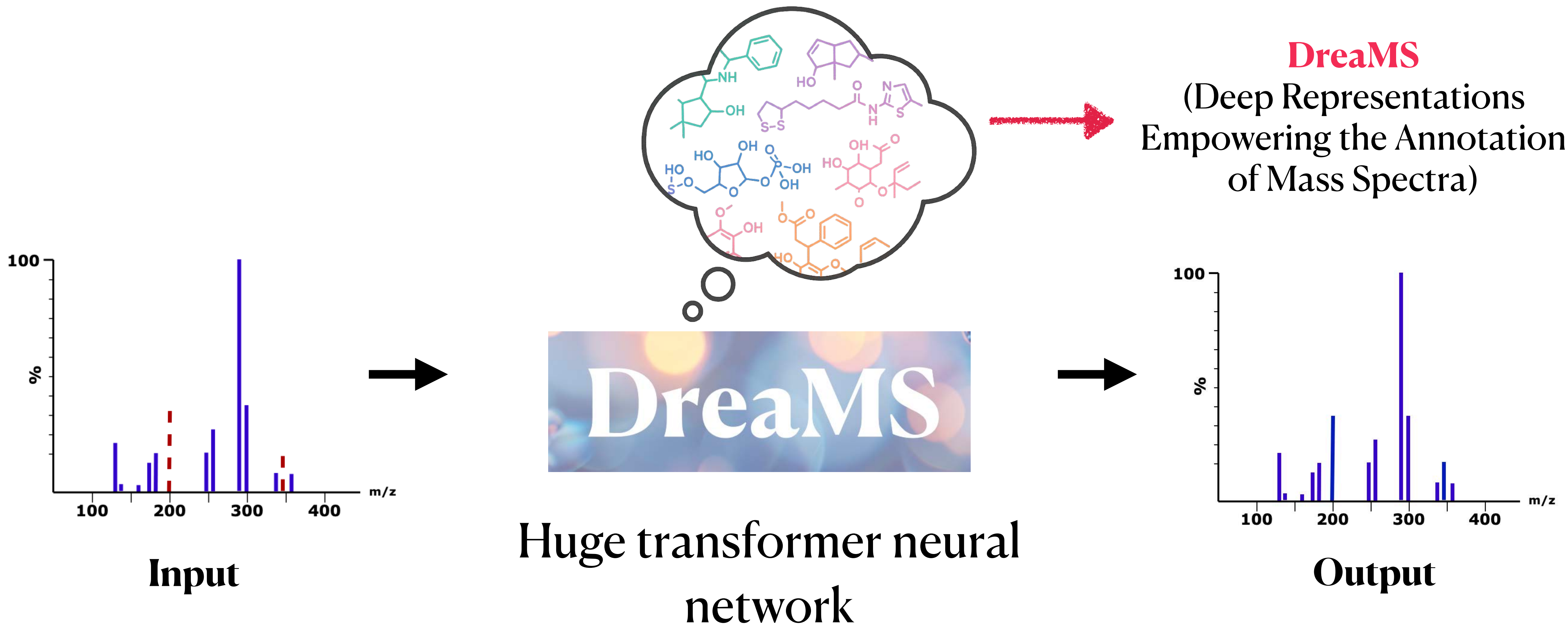
DreaMS

Huge transformer neural
network



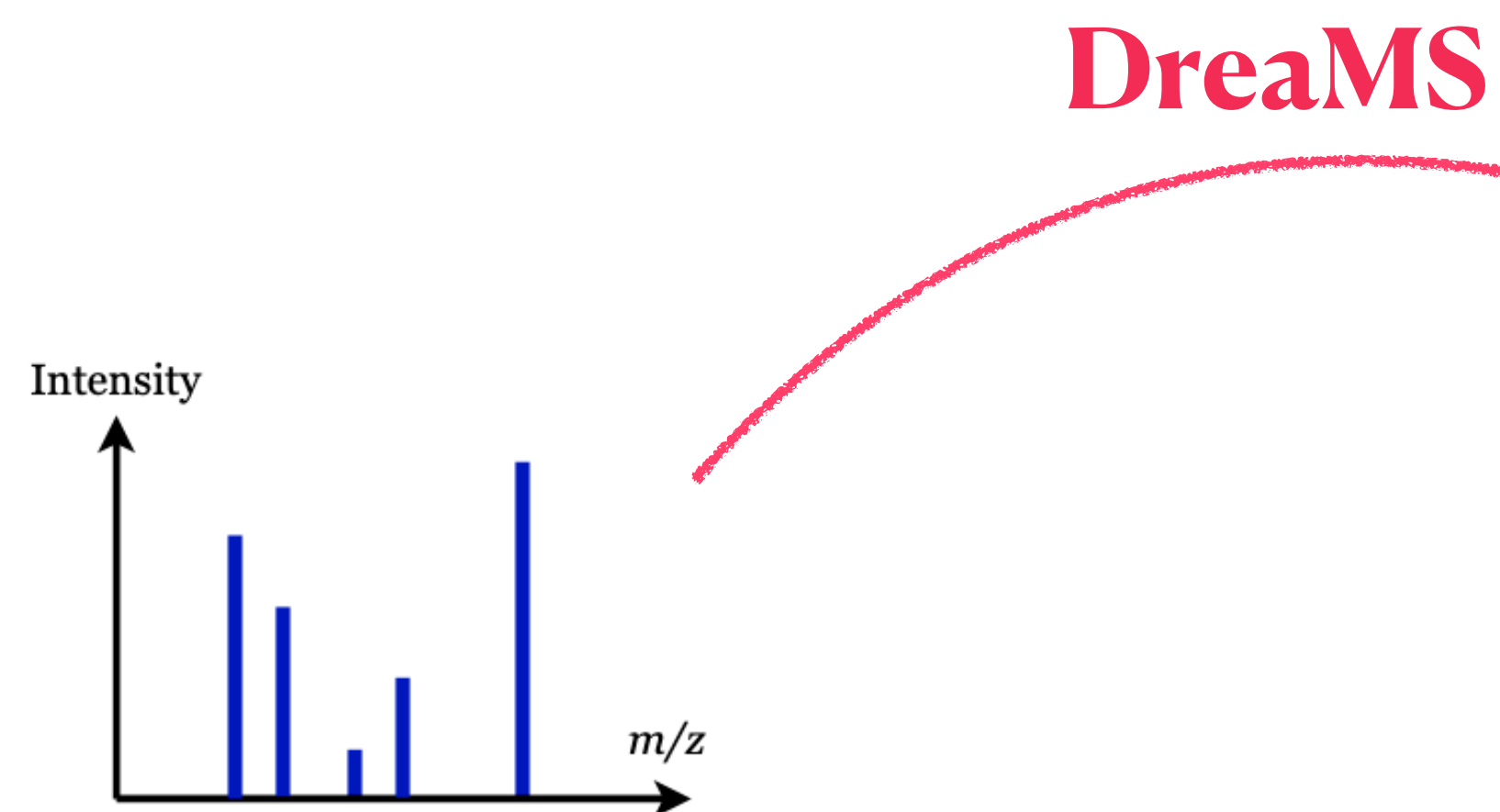
Huge amount of training
data

DreaMS learns by predicting hidden peaks



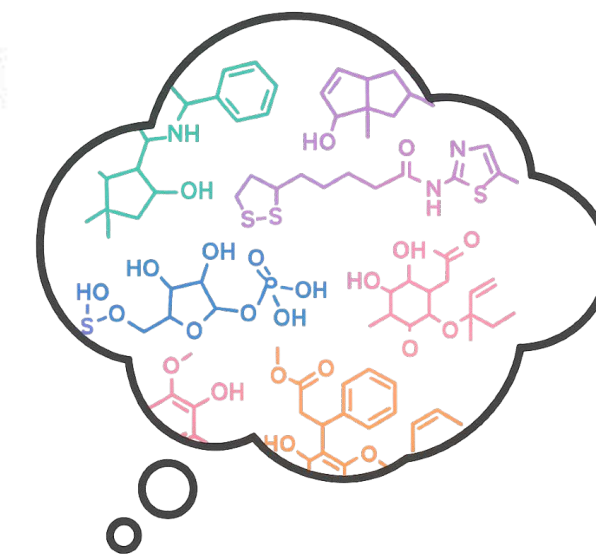
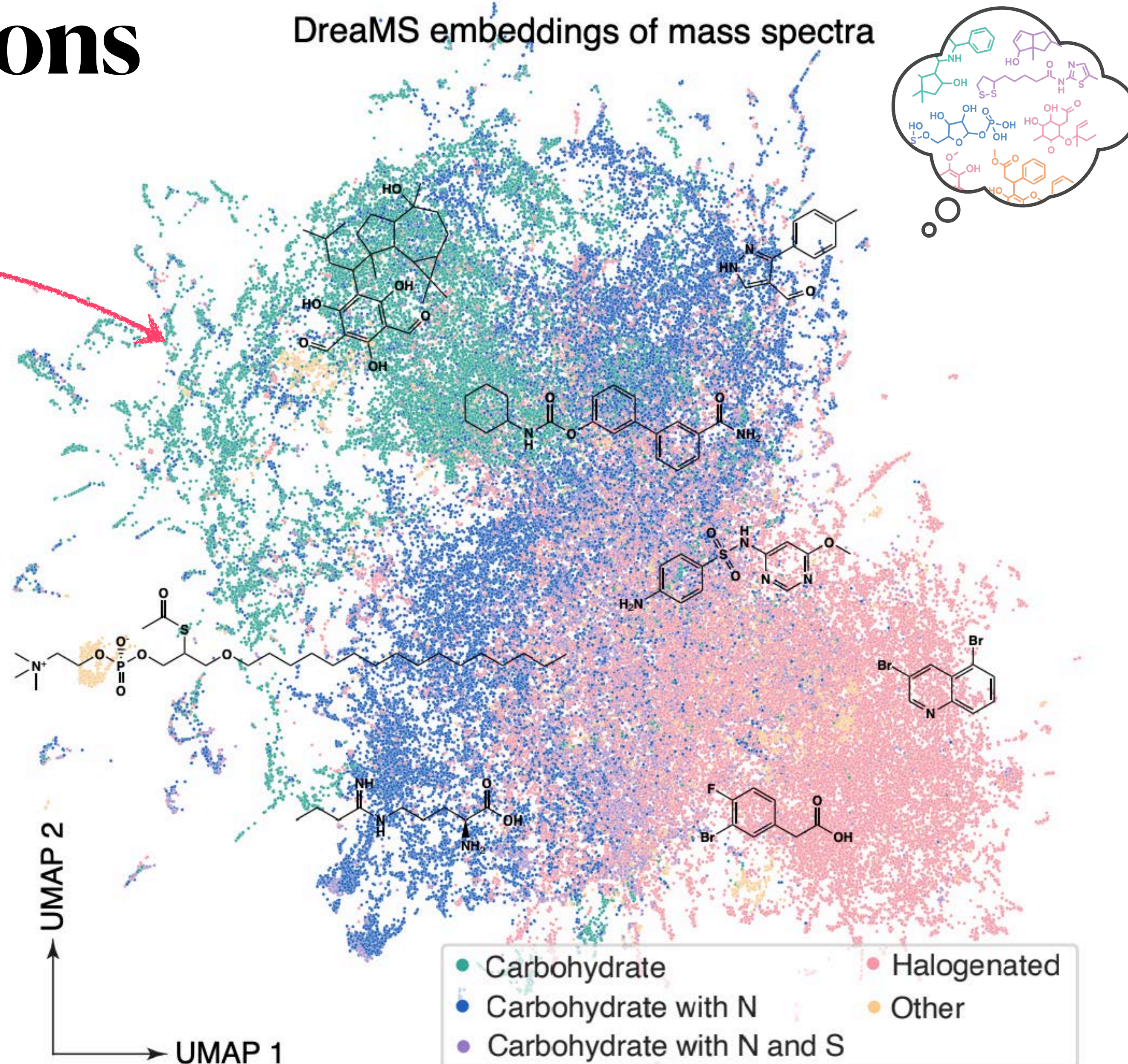
DreaMS representations

DreaMS embeddings of mass spectra

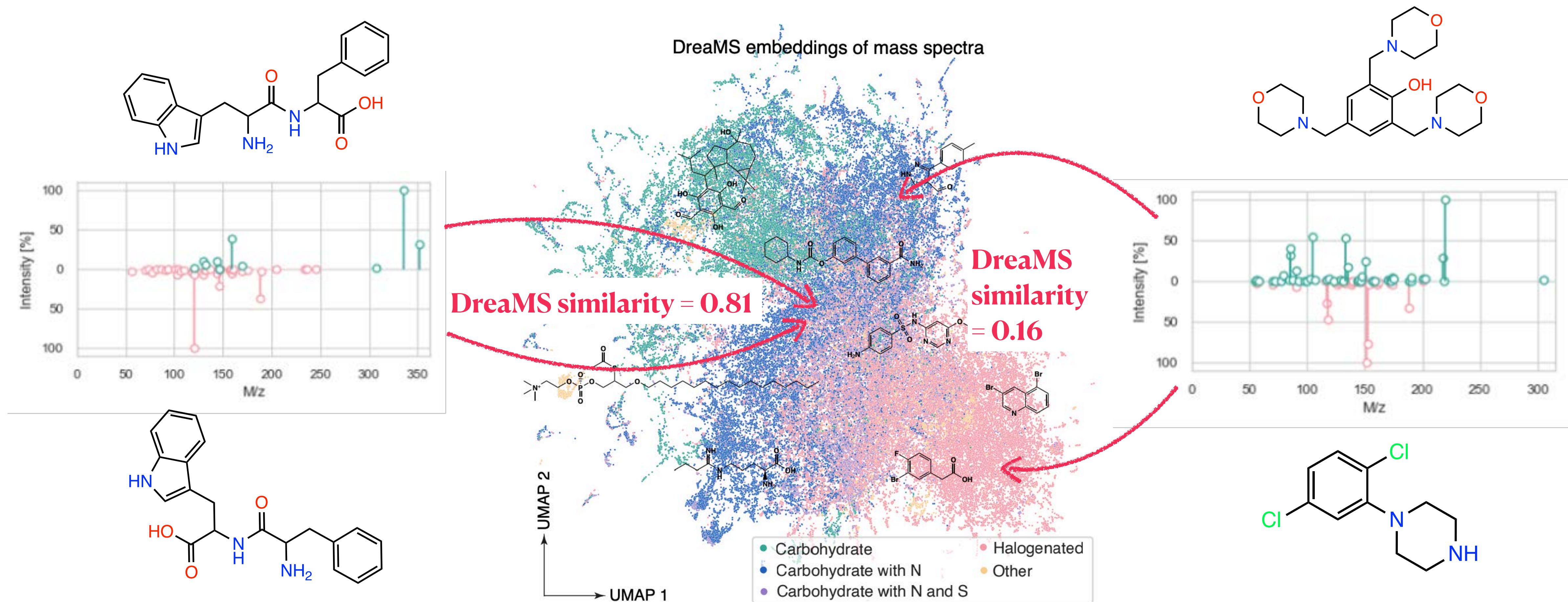


DreaMS
(Each spectrum becomes
a list of numbers)

1024-dimensional vectors

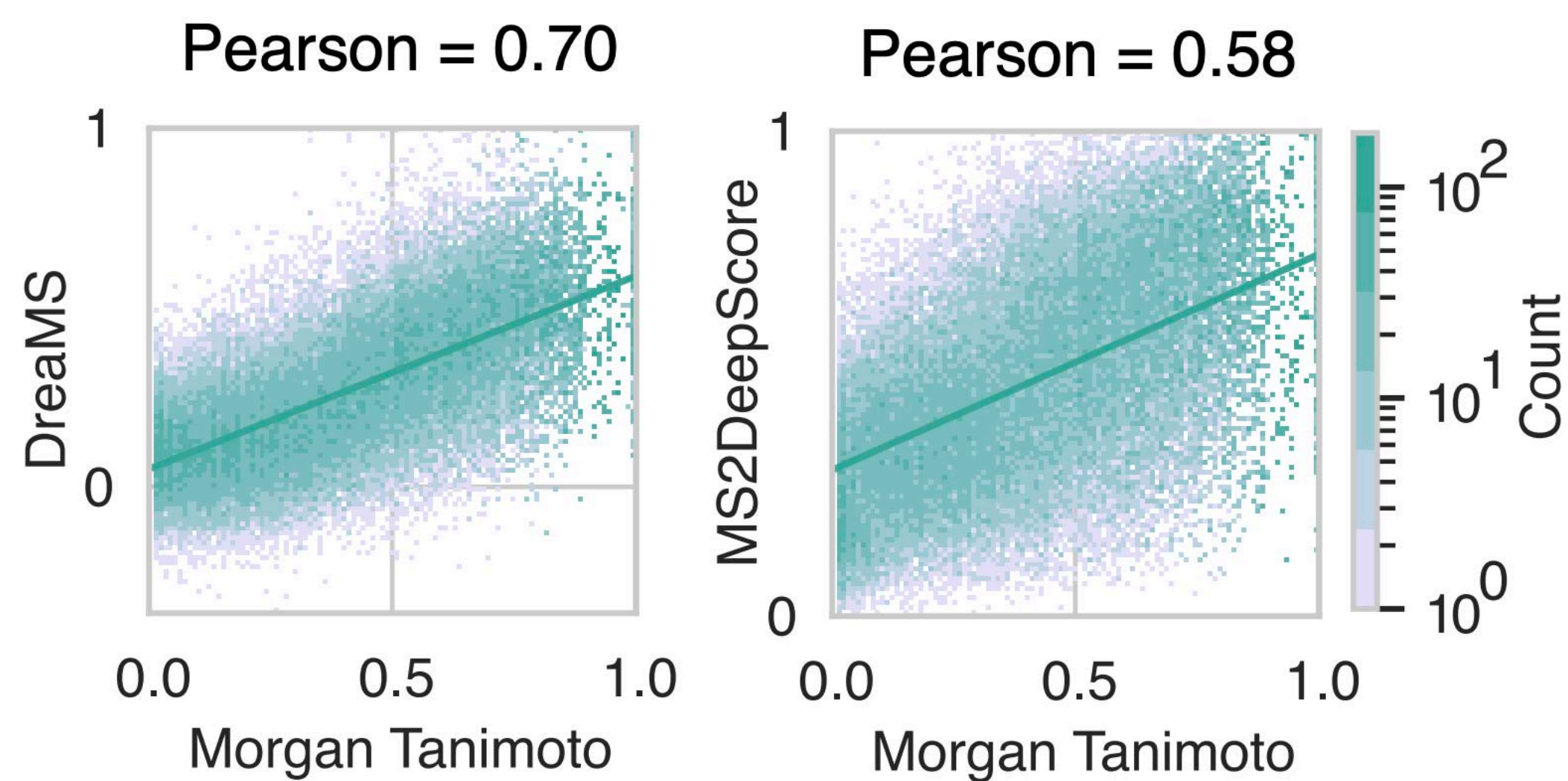


DreaMS mass spectral similarity

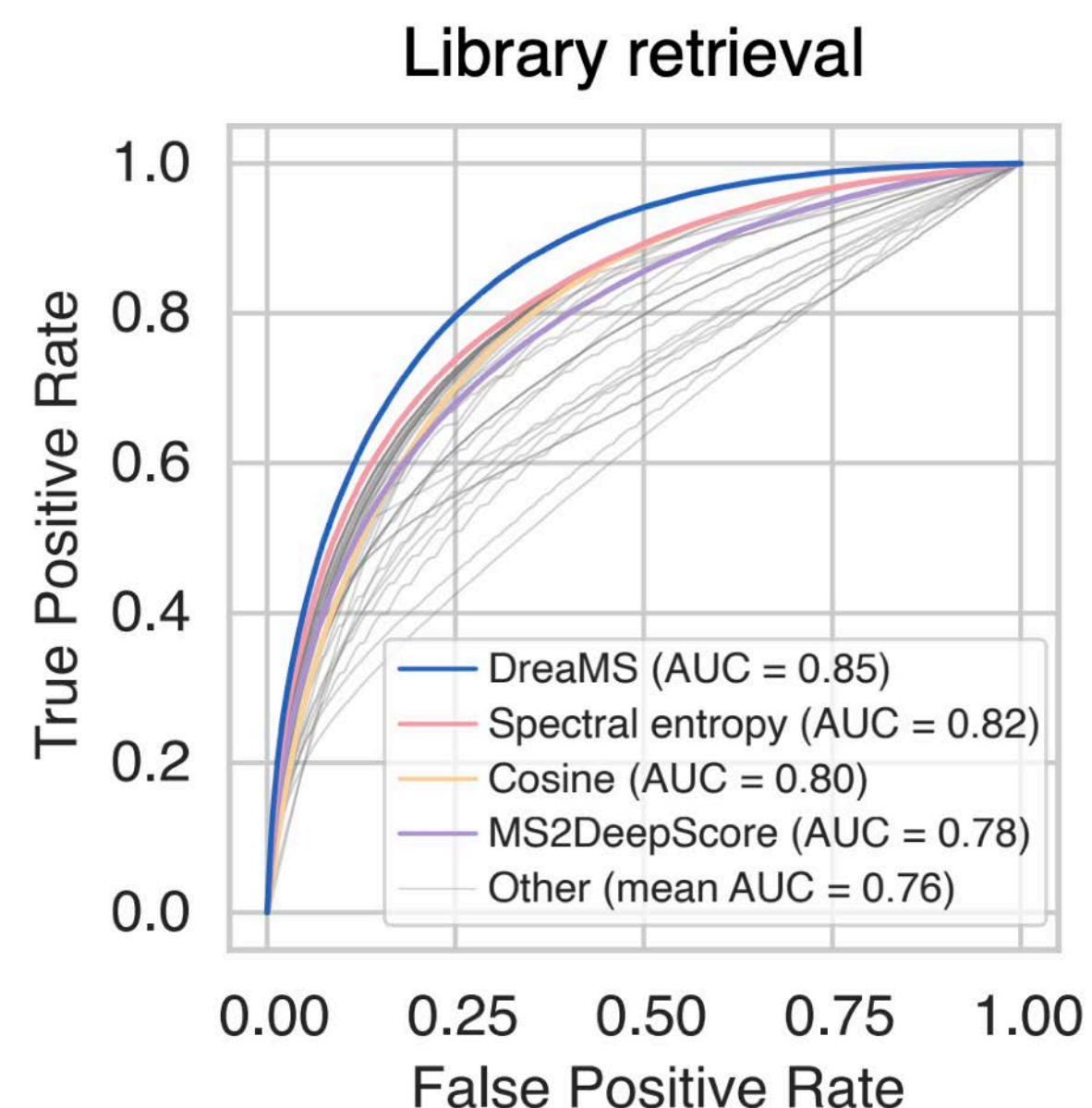


Similar DreaMS representations = similar molecular structures

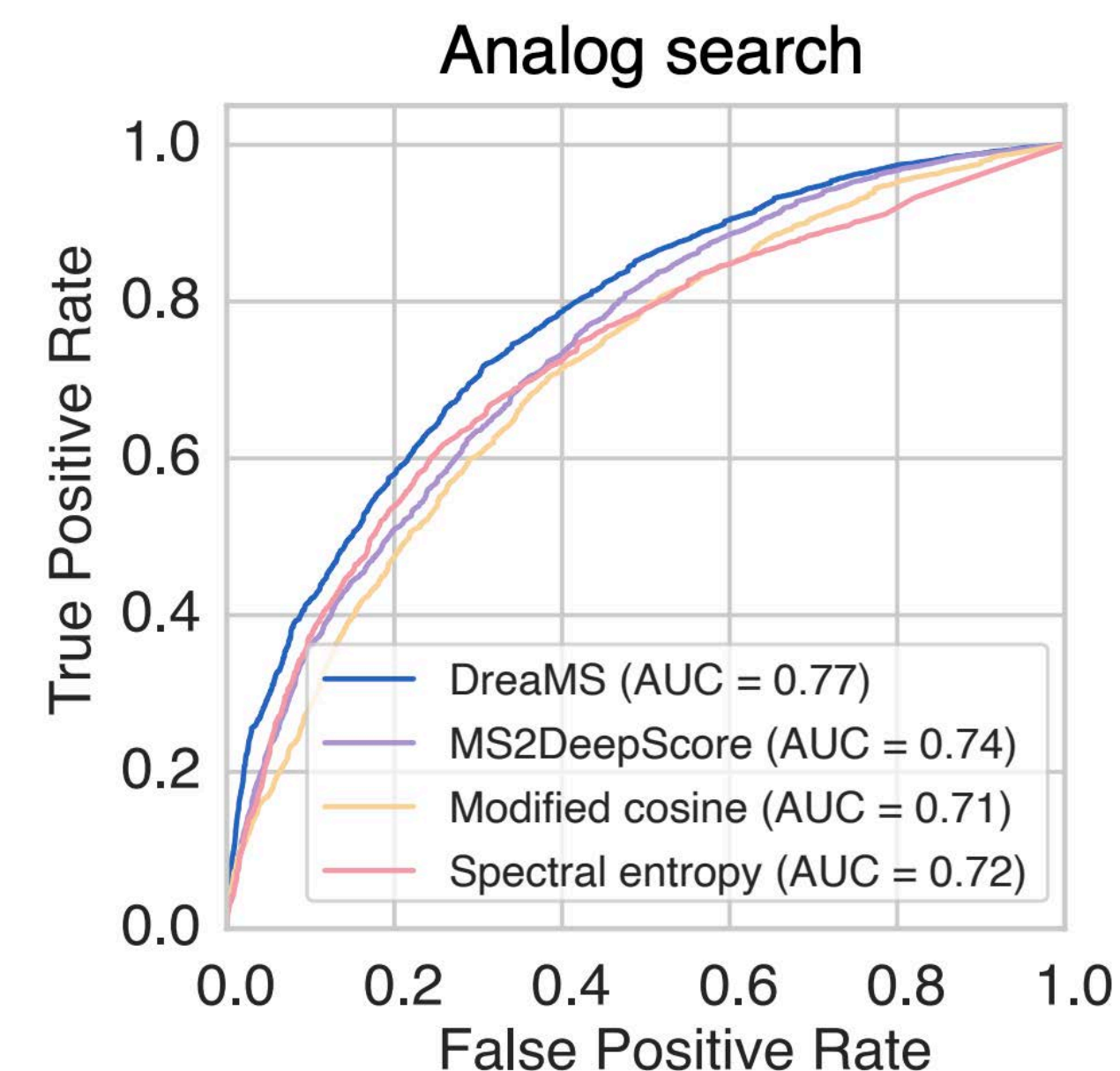
DreaMS is a state-of-the-art spectral similarity for ...



**... comparing spectra
based on molecular
similarity**



**... library
retrieval**

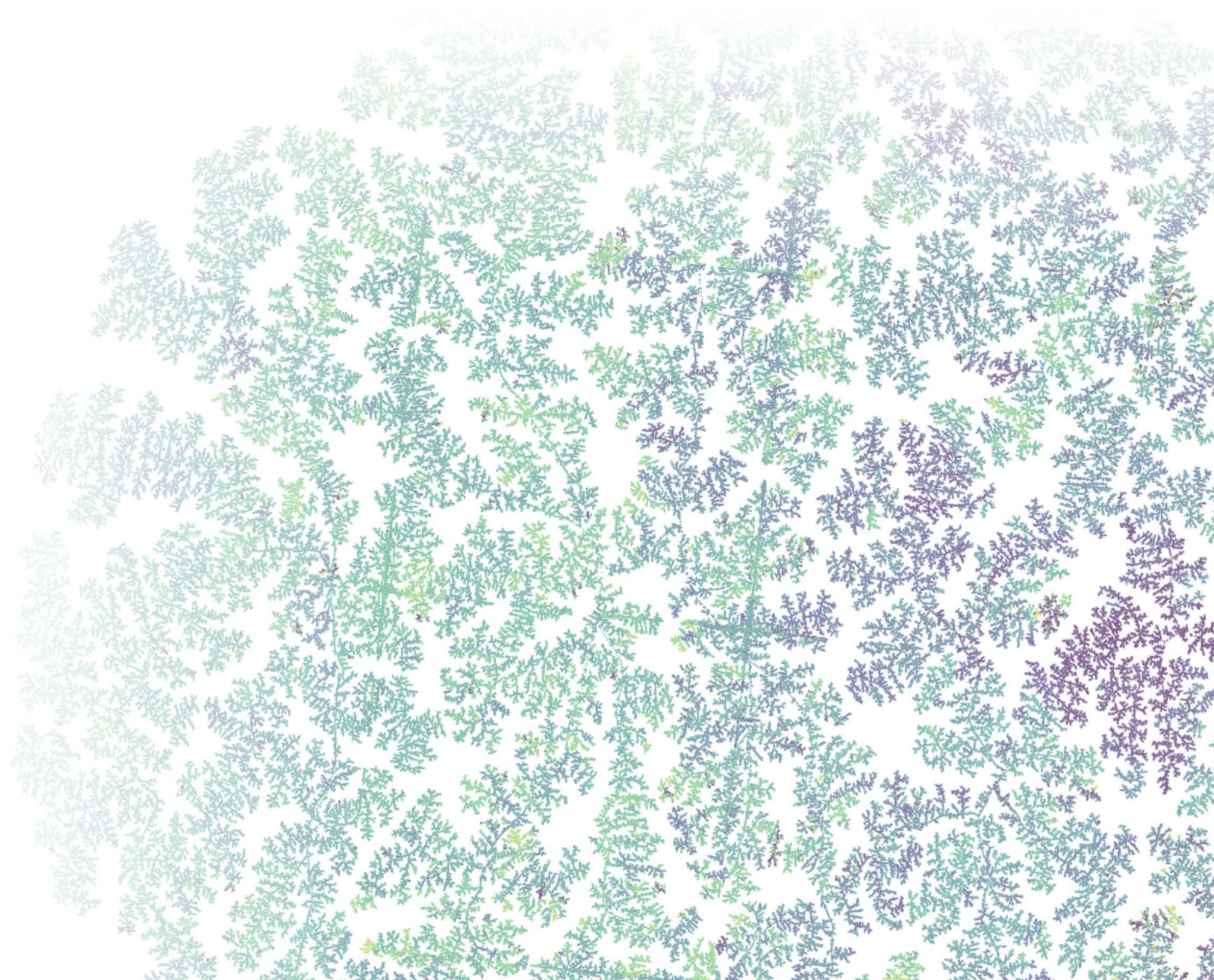


**... analog search /
molecular networking**

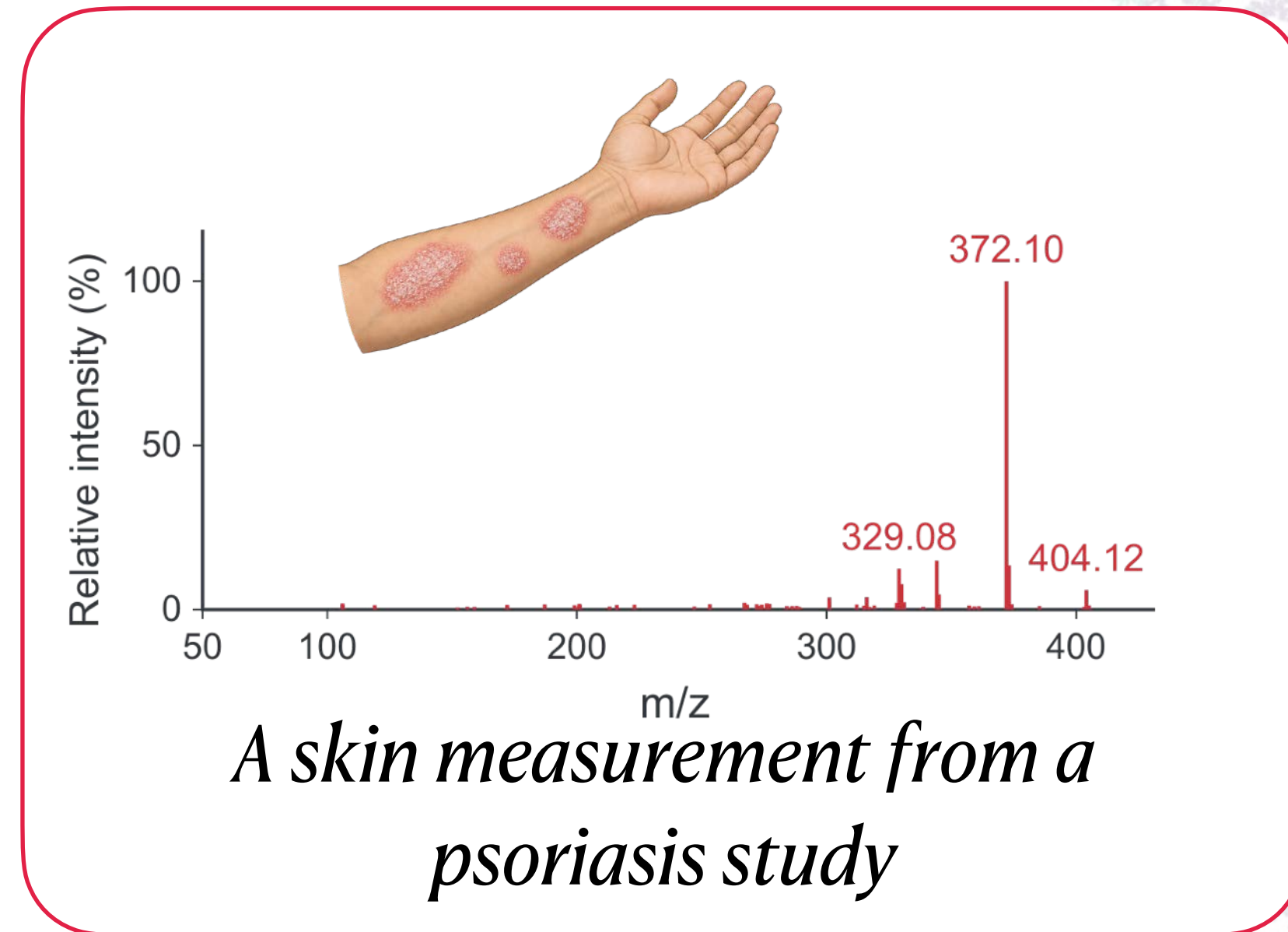
DreaMS Atlas

**Molecular network
of 200 million
mass spectra from
MassIVE**

**“The Internet of
mass
spectrometry”**

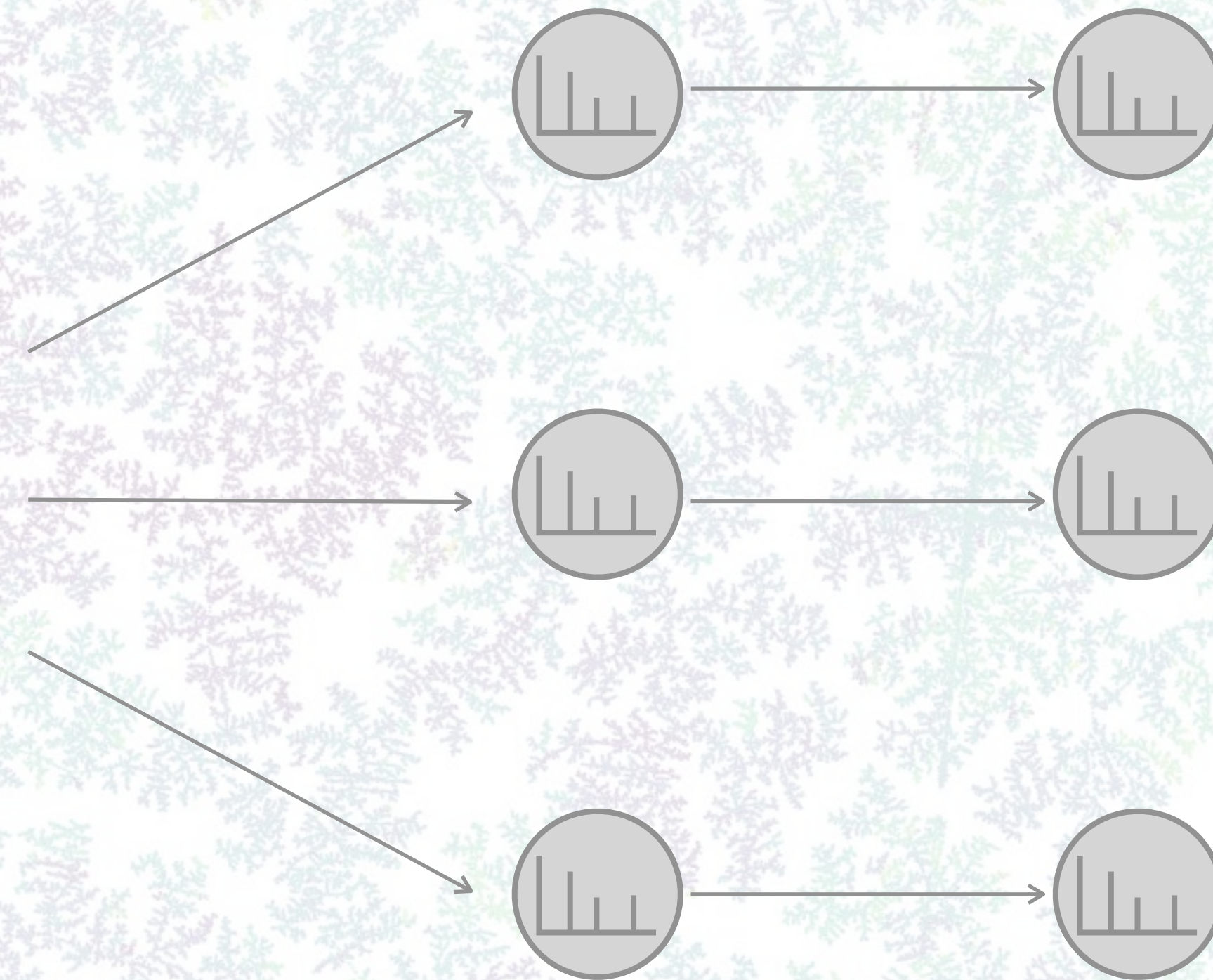


What could this skin signal be?

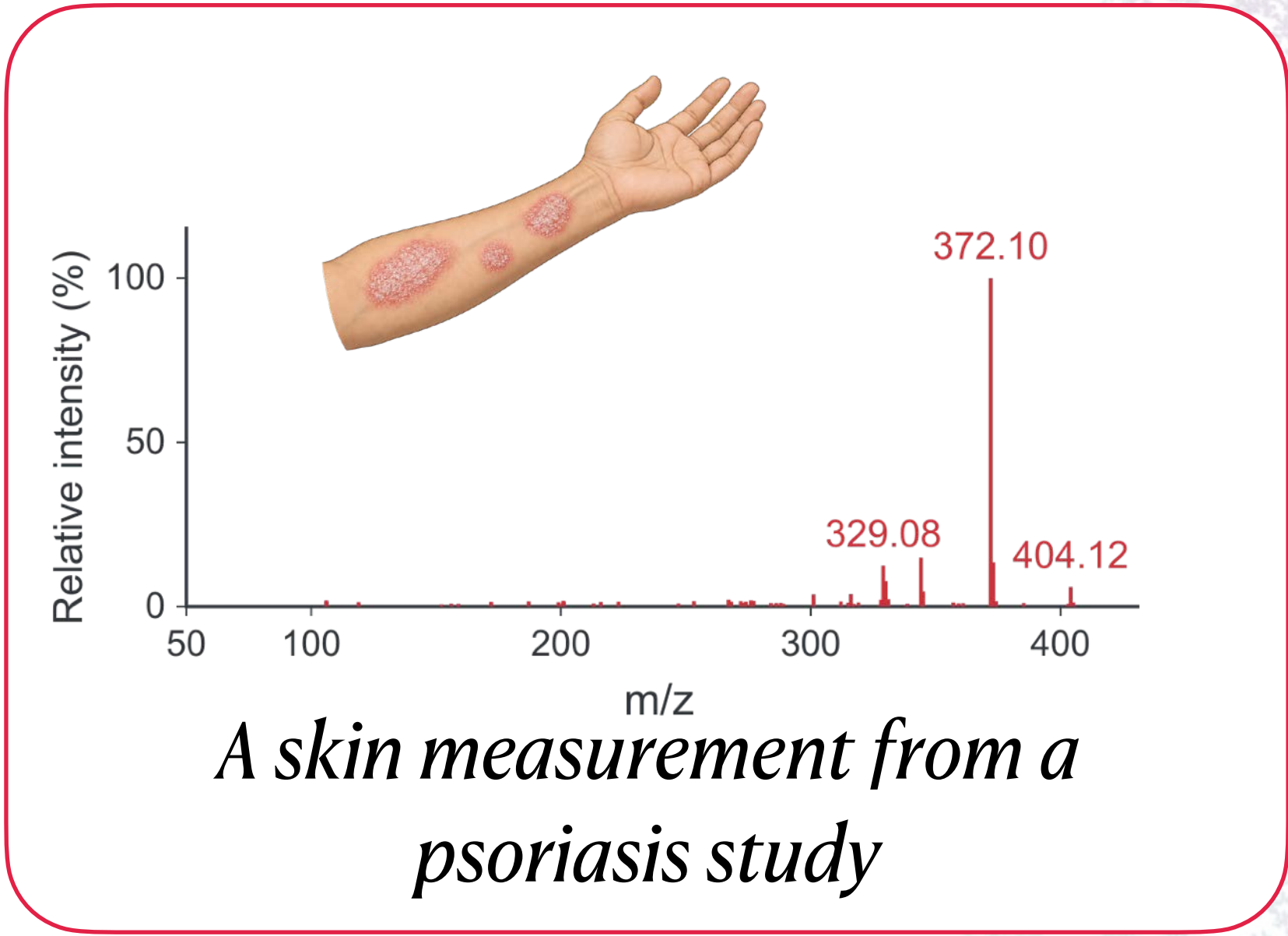


Your measurement

Unknown signal still can have
neighbourhood

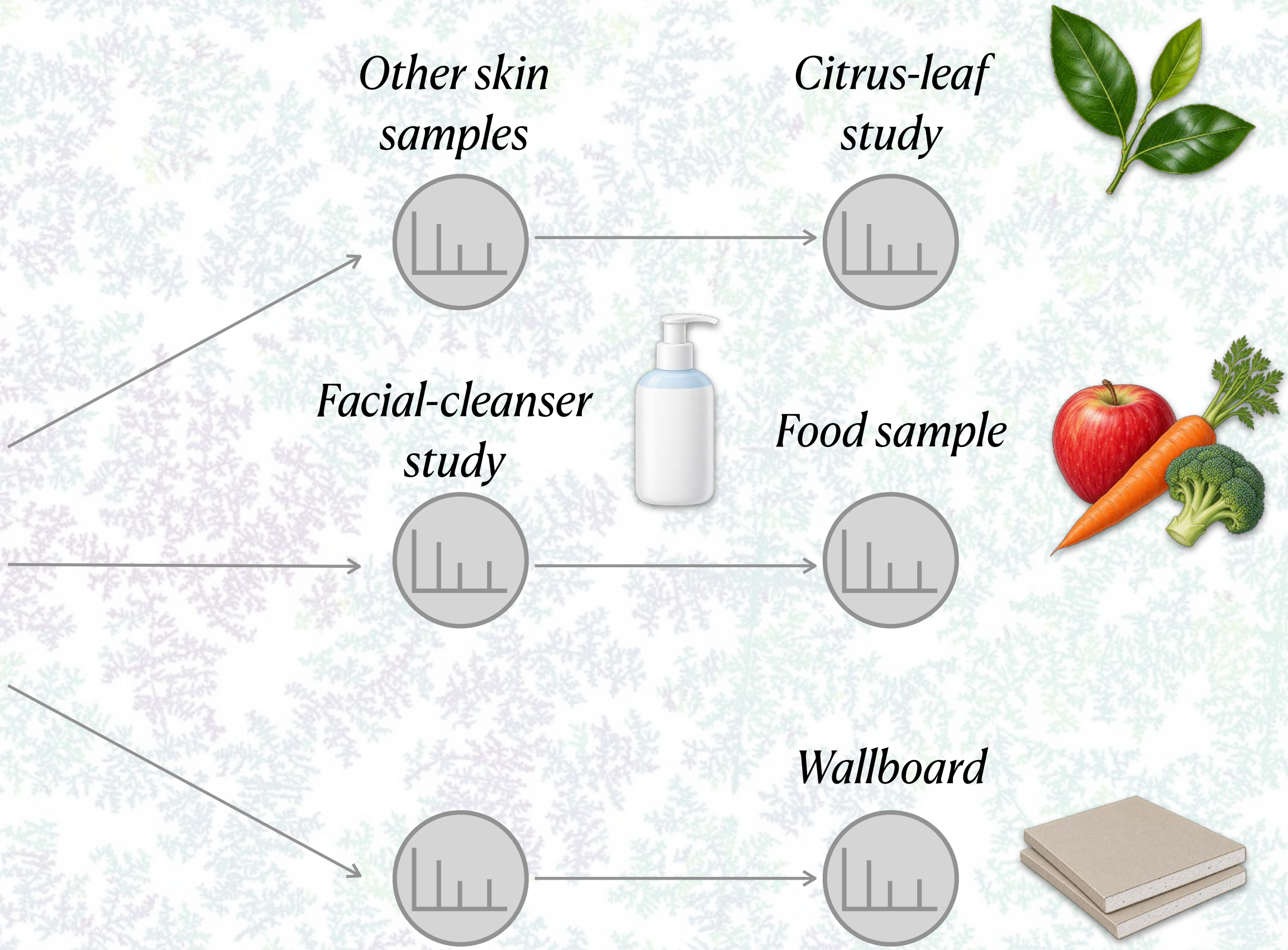


Its neighbours come from studies asking different questions

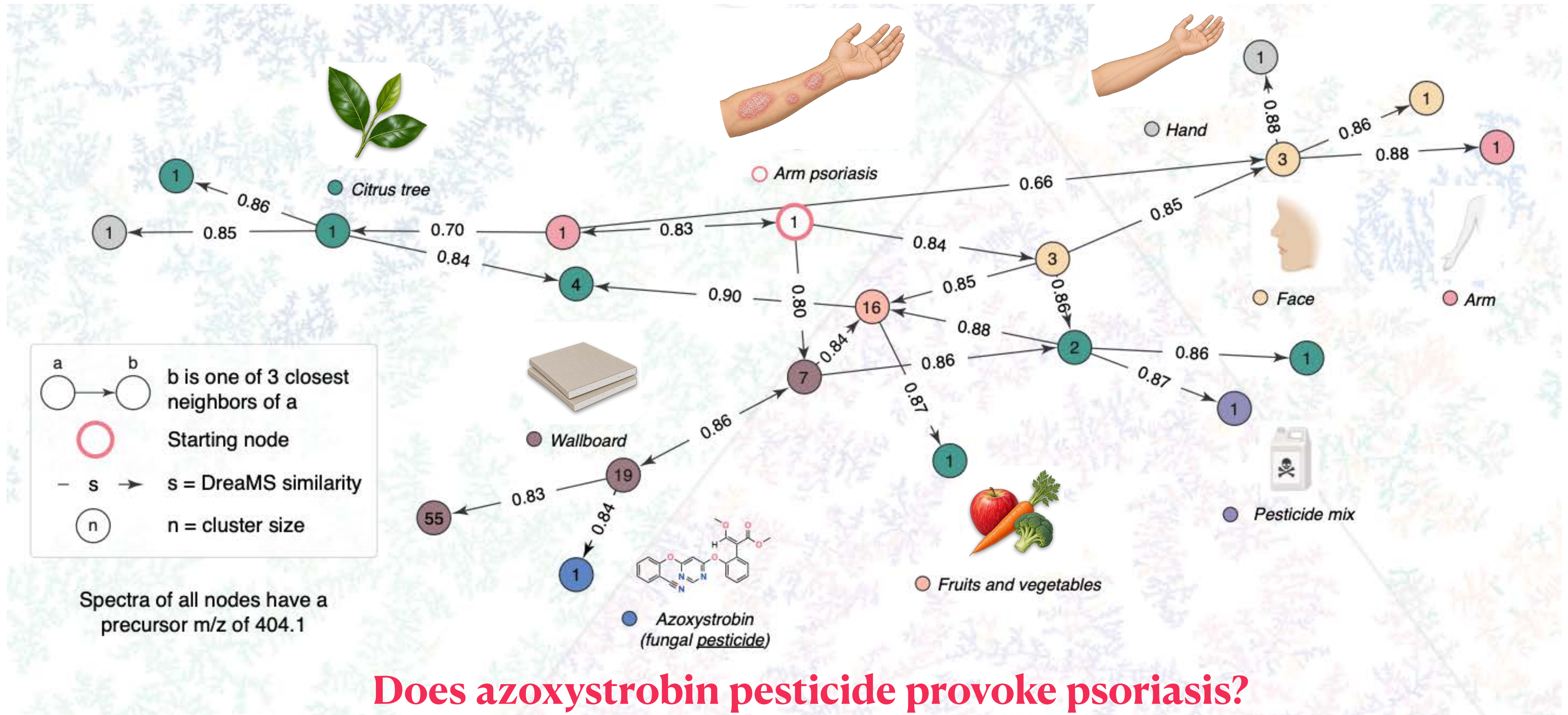


Your measurement

What might connect these observations?



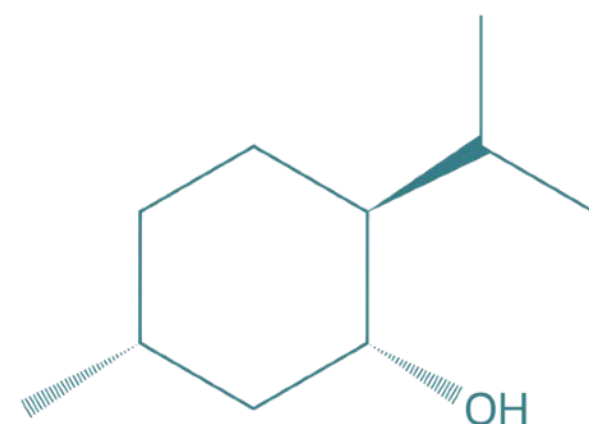
Navigating the “internet of mass spectrometry”



What if you start with a chemical family?

Menthane monoterpenoids

(–)-Menthol

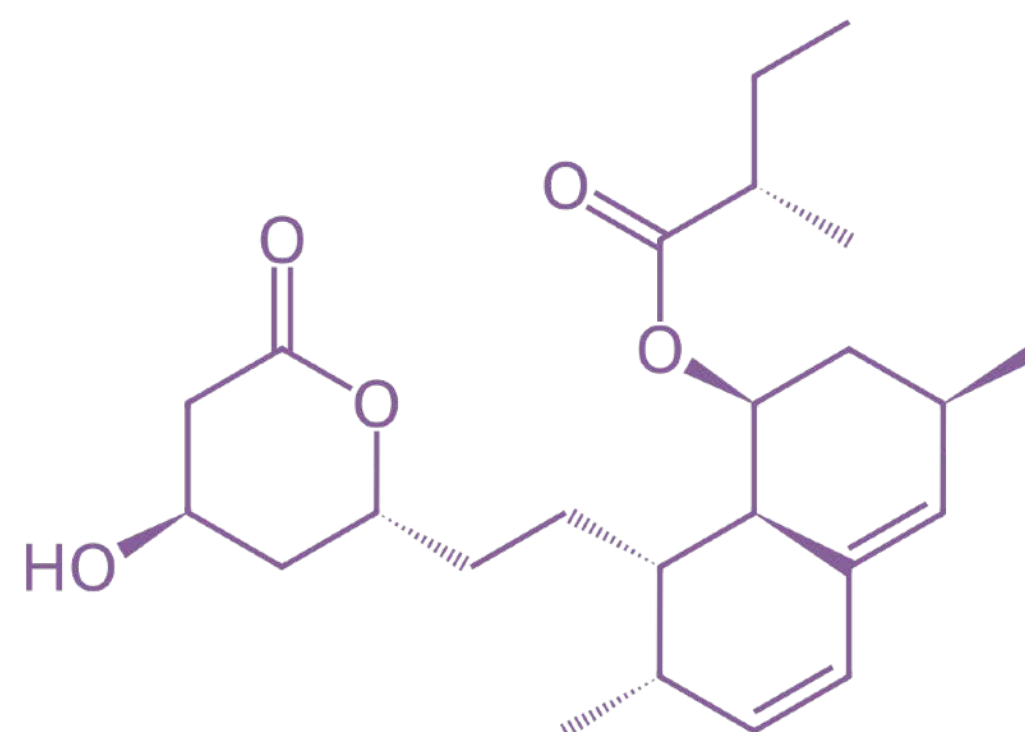


Flavouring and cooling

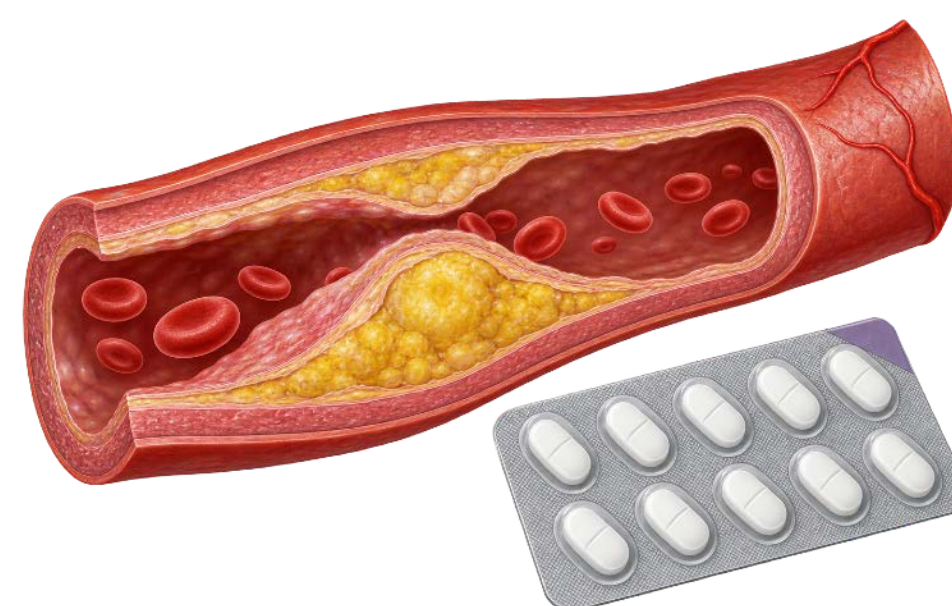


Statins

Lovastatin

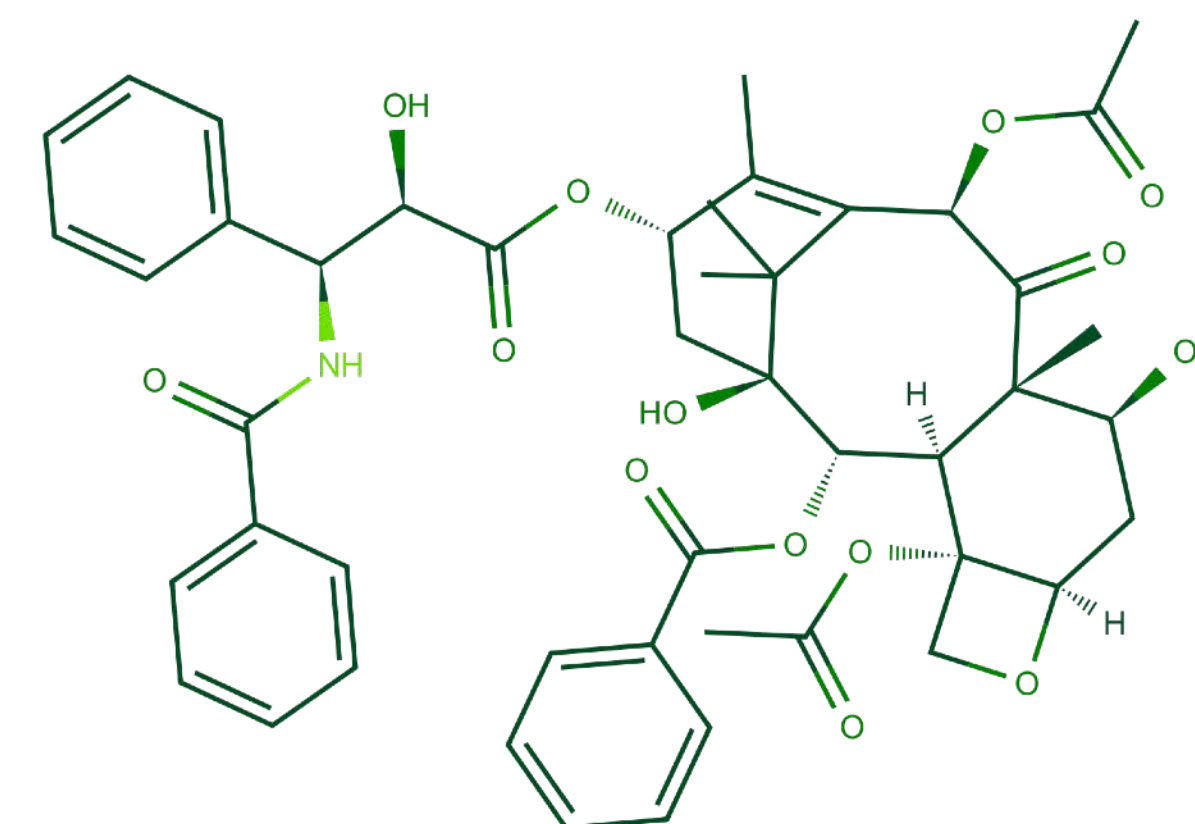


Lowering cholesterol



Taxanes

Paclitaxel (Taxol)



Chemotherapy drug



NPKB

NATURAL PRODUCTS KNOWLEDGEBASE

Search compound

Search

Compounds

Classes

Features

Organisms

Data

Find by measurement

Proposals

Deposit

The chemistry of living things, with the measurements attached

Natural small molecules linked to the spectra that identify them, the organisms that make them and the pathways that build them.

Compounds

Isochlorogenic acid A, C6H12O6, Triterpenoids, Homalomena occulta

Search

Examples

Isochlorogenic acid A

C6H12O6

Triterpenoids

Homalomena occulta

Still filling up: 2 of the datasets NPKB ships with are not in yet, including 2 untargeted runs. A run carries the features and its source organism, so those stay at zero until it lands. What is still to come.

8,912

COMPOUNDS

7

UNRESOLVED FEATURES

31

ORGANISMS

66,518

SPECTRA

12

DEPOSITIONS

0

COUNTRIES SAMPLED

HOW MUCH OF THIS KNOWLEDGEBASE IS MEASURED

Compounds in NPKB

8,912

100.0%

...with a mass spectrum

8,912

100.0%

...reviewed

8,912

100.0%

...from an authentic standard

8,727

97.9%

Counted from the entries actually held here, not estimated against a catalogue of known chemistry.

WHERE THE MATERIAL COMES FROM

No collection coordinates recorded yet.

0 collection site(s) with coordinates, across 12 deposition(s)

RECENTLY ADDED

5-OXO-D-PROLINE

NPK-0008988 · C5H7NO3

D-GLUCOSE-6-PHOSPHATE

NPK-0008987 · C6H13O9P

DEOXYCORTICOSTERONE

NPK-0008986 · C23H32O4

NPKB

Natural Products Knowledgebase

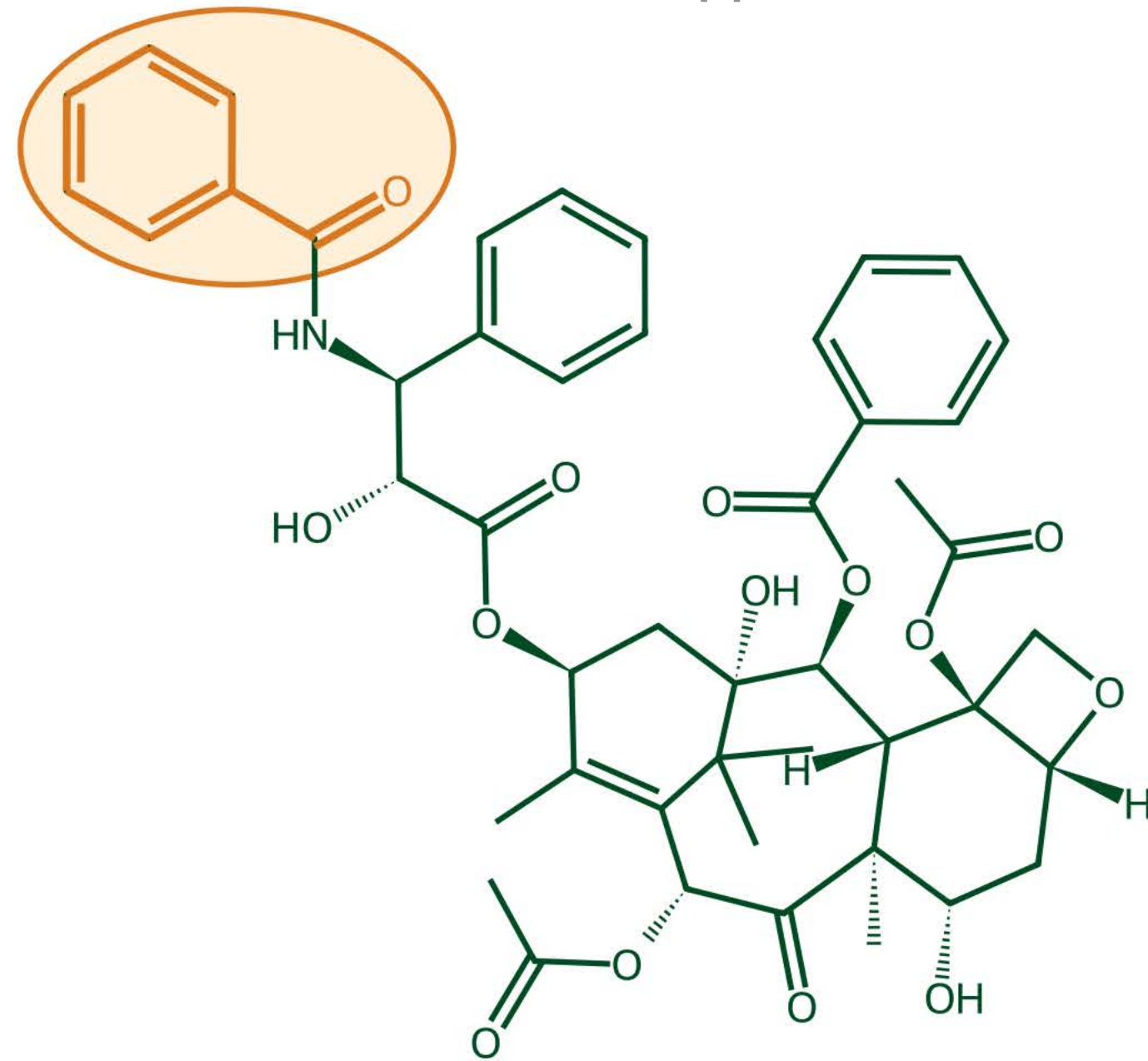
UniProt

For small molecules

Taxol is one member of a larger chemical family

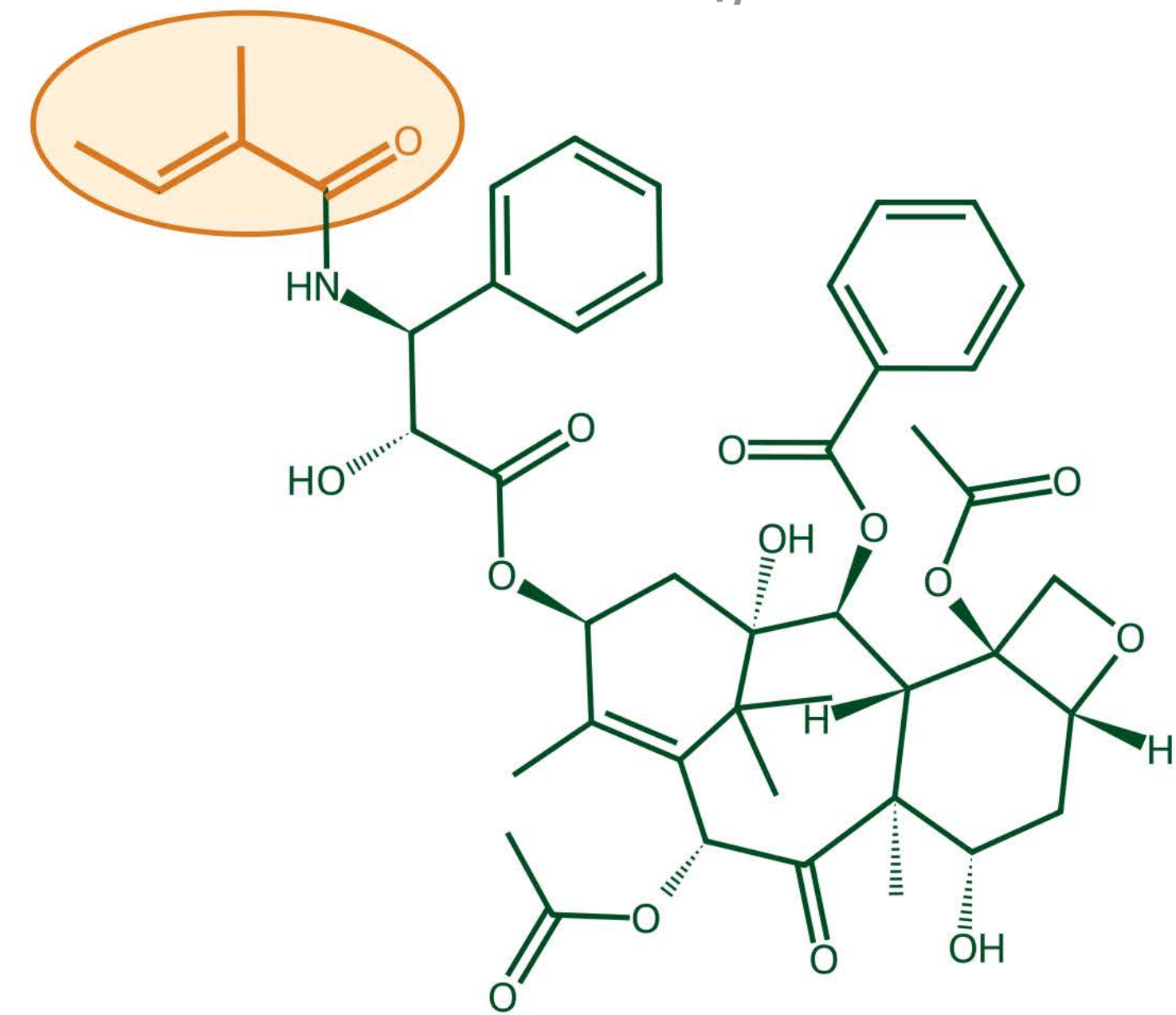
Paclitaxel (Taxol)

NPK-0002244



Cephalomannine

NPK-0004701

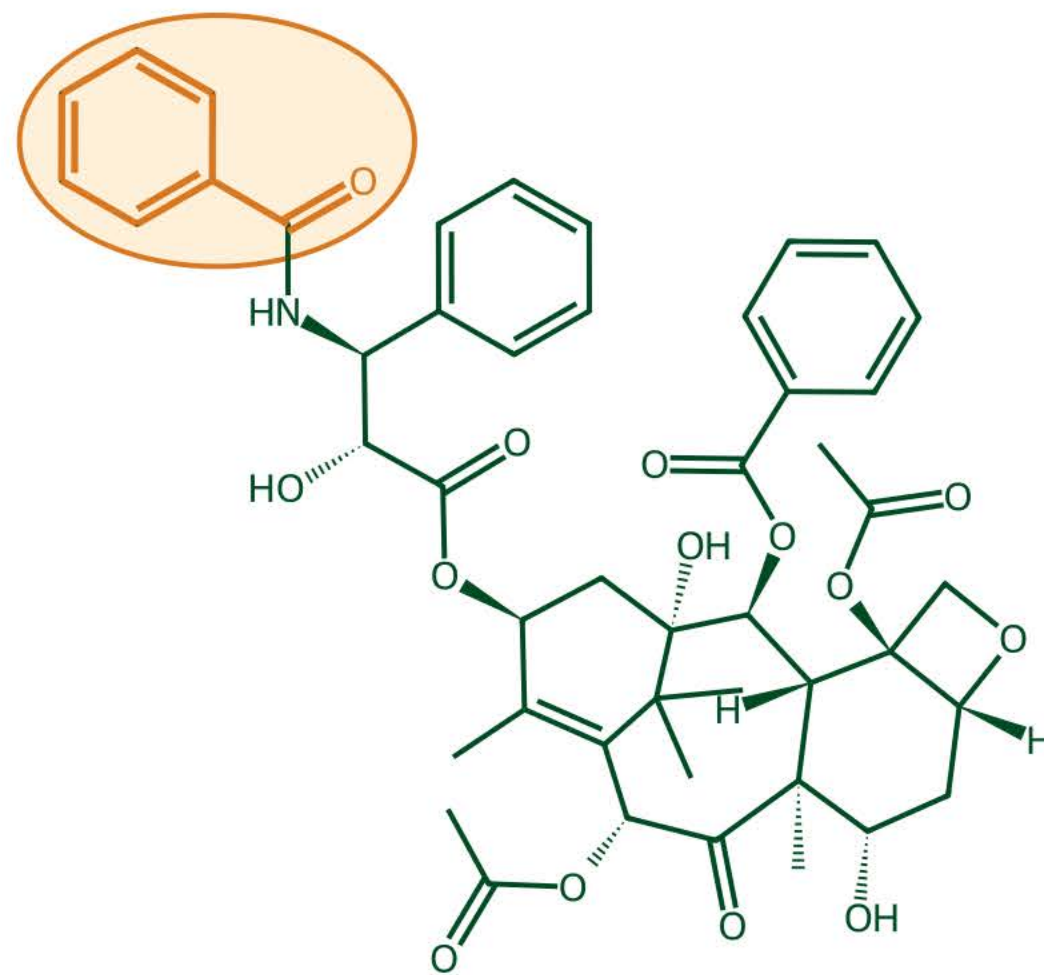


The highlighted side groups are different.

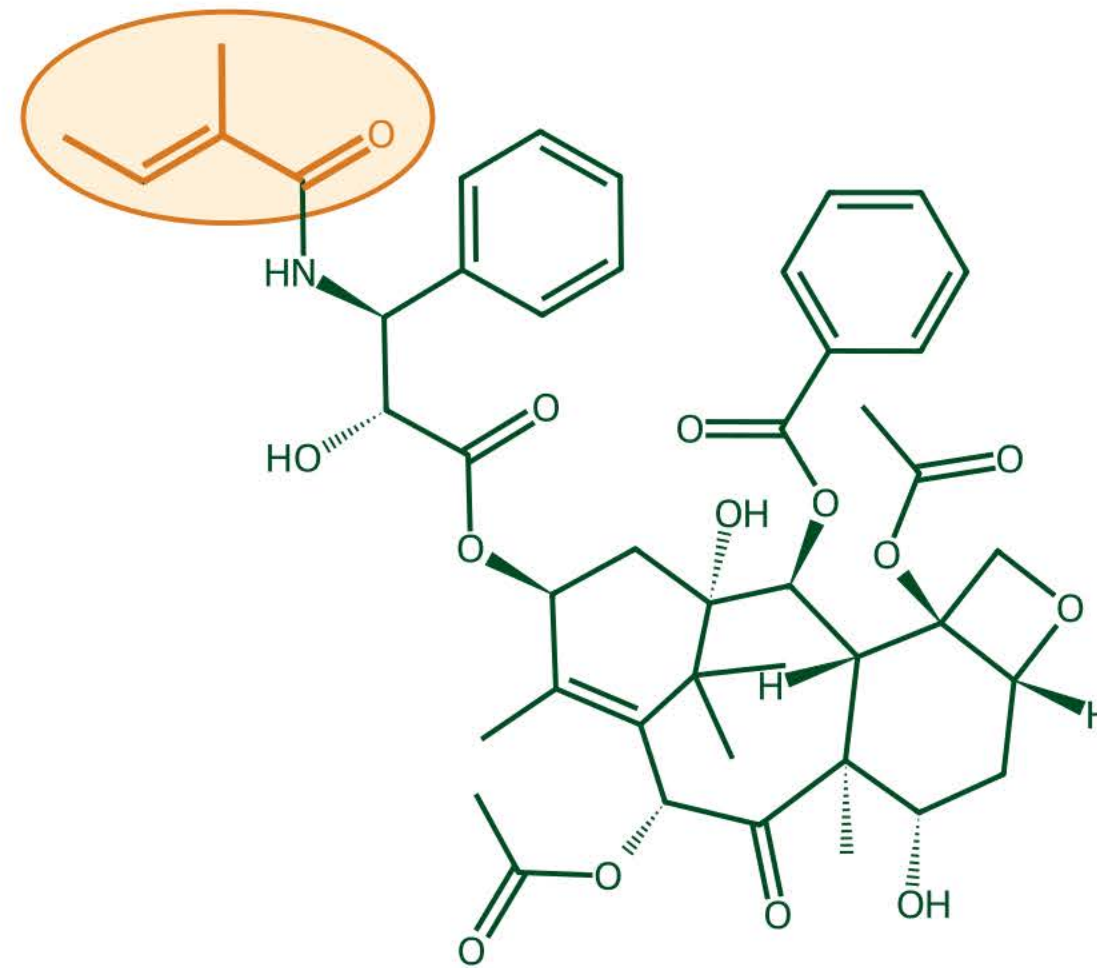
Do these related molecules behave the same way?

A small structural change can alter metabolism

Paclitaxel



Cephalomannine



Where did enzymes add an OH group?

Paclitaxel



Cephalomannine



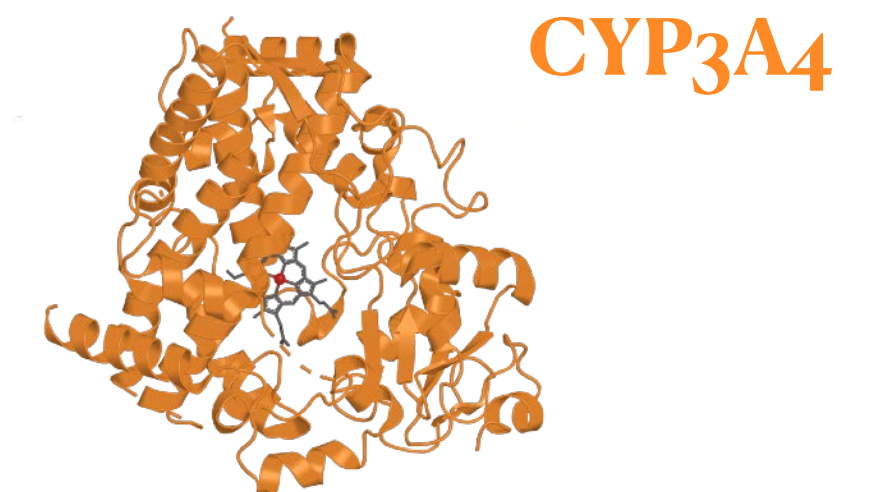
Side chain

Core (C6α)



Similar structures can be processed differently.

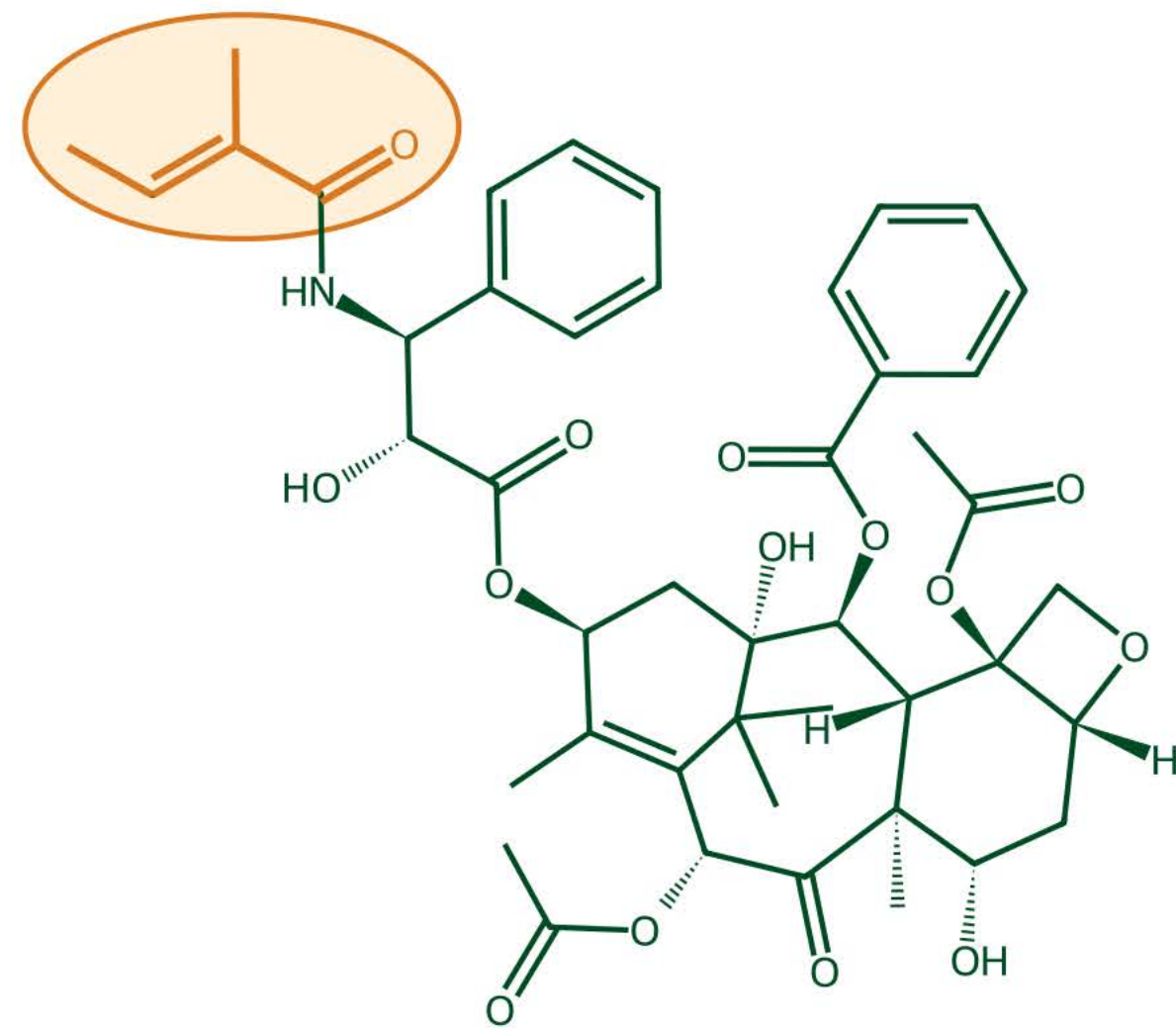
How the body processes a medicine can affect how well it works and what side effects it causes.



NPK connects a molecule to its family and the measurements behind it

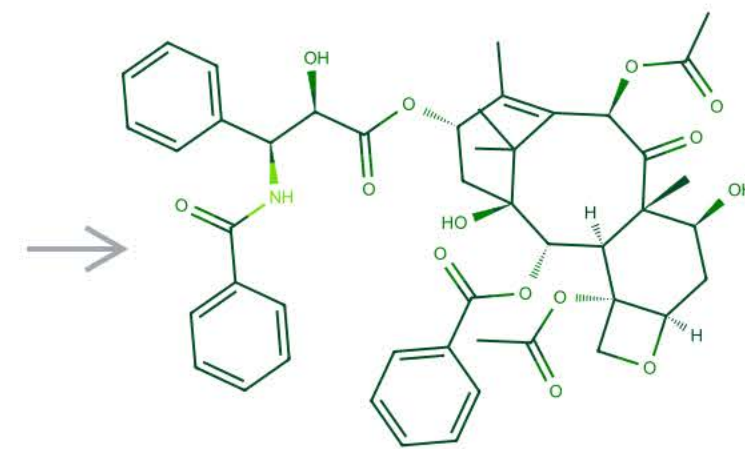
Cephalomannine

Taxane diterpenoid · $C_{45}H_{53}NO_{14}$



NPK-0004701
Structure and identifiers

Related molecule

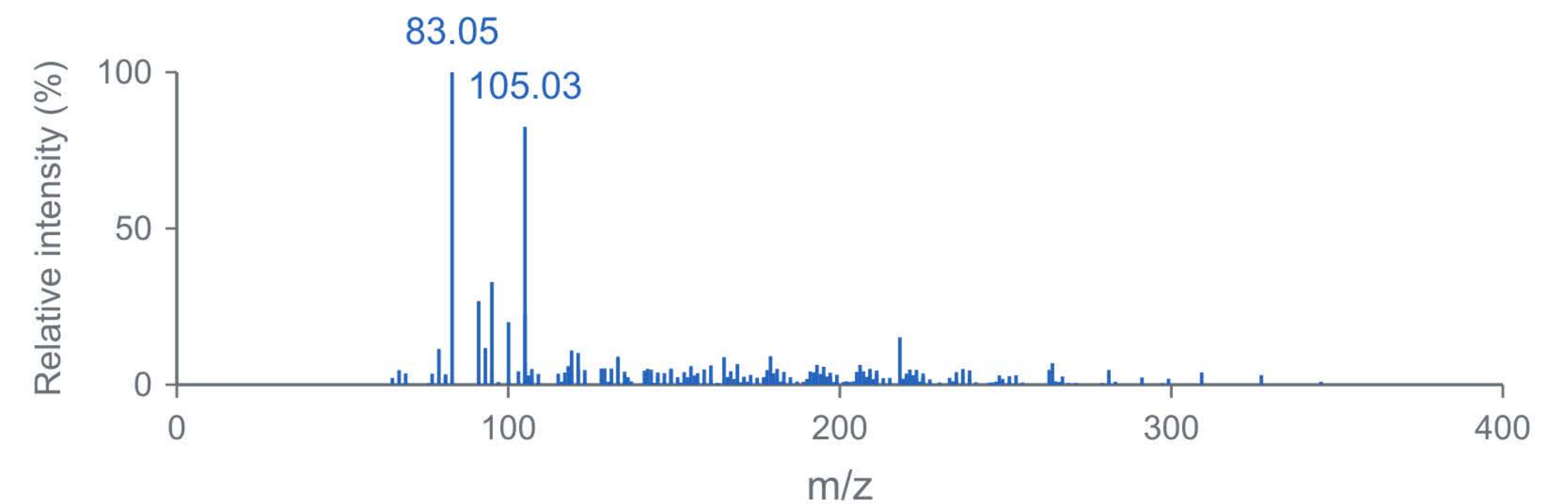


Paclitaxel (Taxol)

NPK-0002244

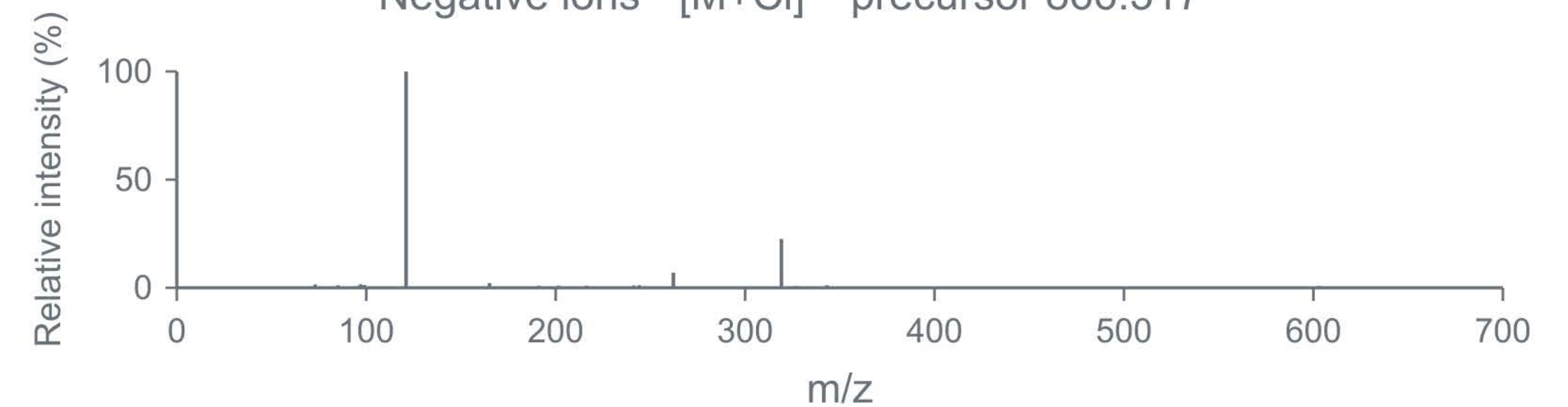
Two deposited reference spectra

Positive ions · $[M+H]^+$ · precursor 832.354



MoNA Alkaloids000154 · 202 peaks

Negative ions · $[M+Cl]^-$ · precursor 866.317



MoNA QE_Alk_neg000040 · 17 peaks

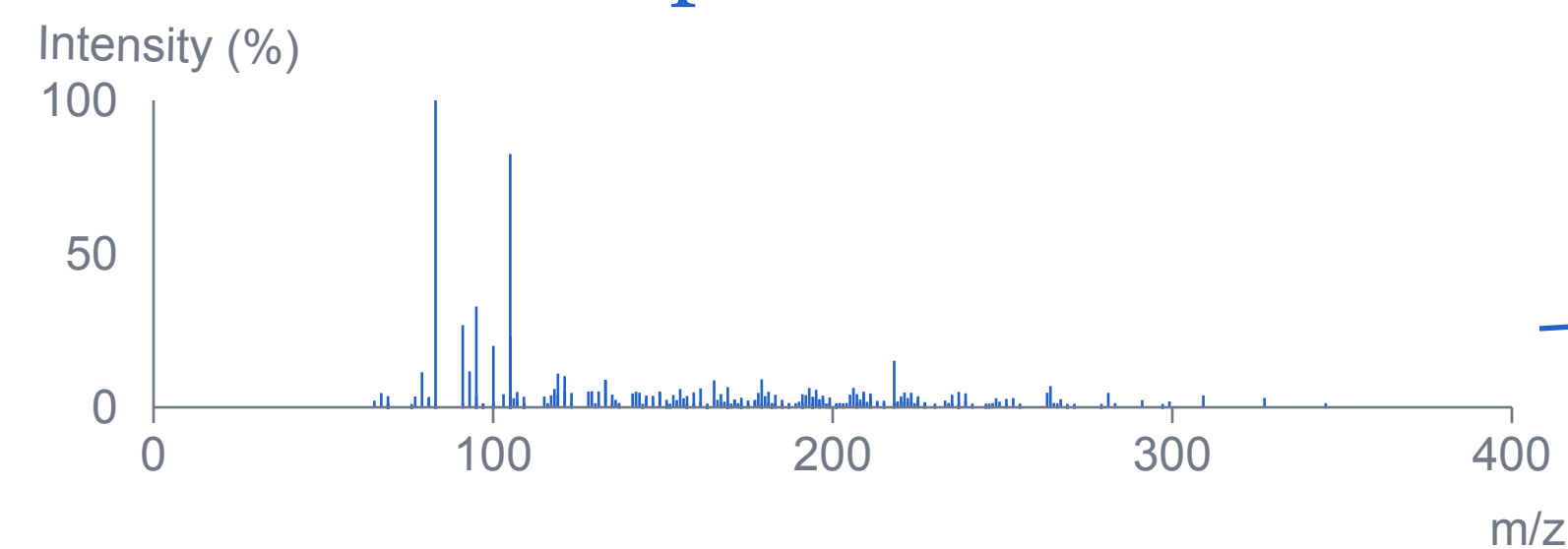
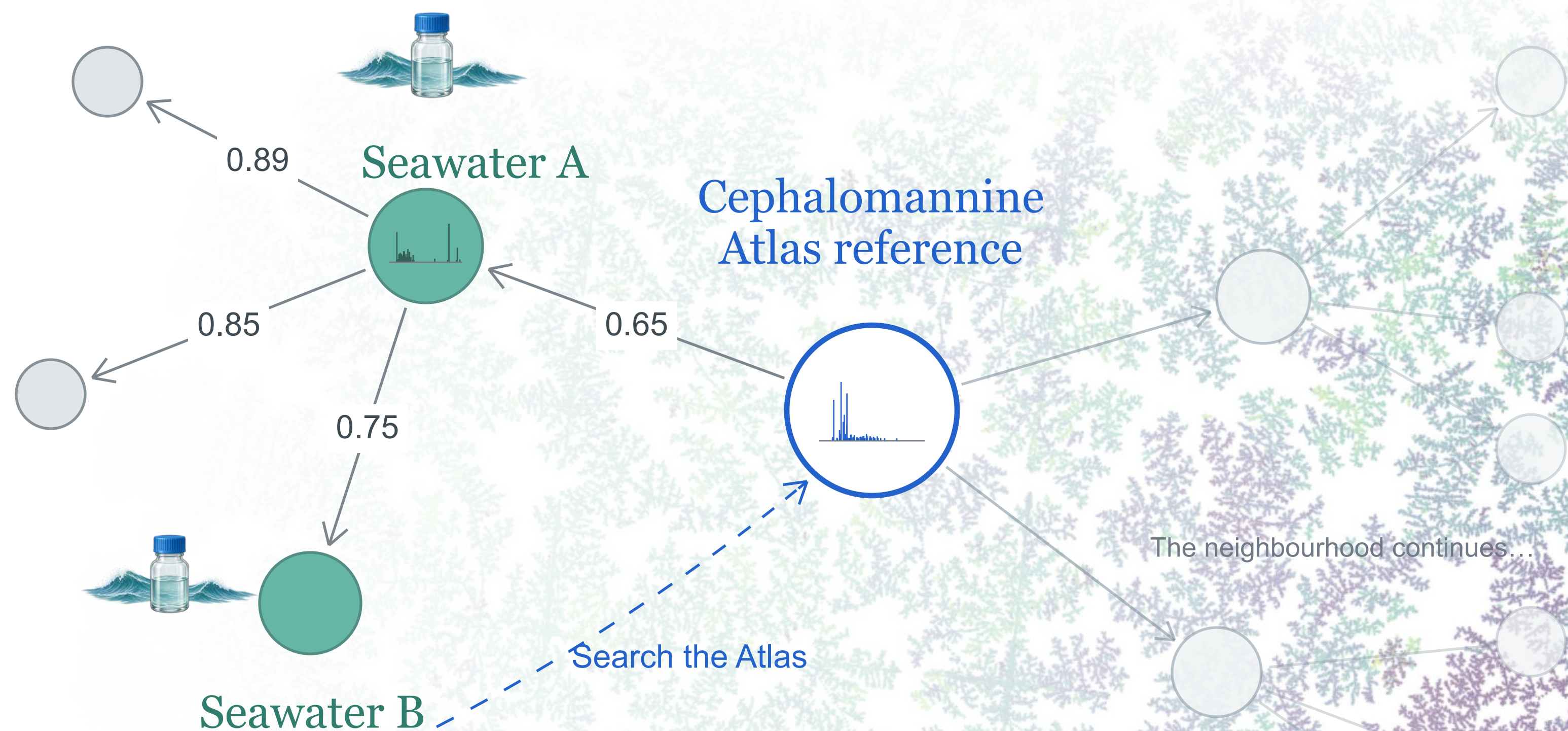
Choose a spectrum. Follow it into the Atlas.

NPKB
Natural Products Knowledgebase

Cephalomannine

The image shows a complex chemical structure, likely a diterpene or a related natural product. A specific group is highlighted with an orange oval: an isopentenyl group, which consists of a prenyl chain (3-methylbut-3-en-1-yl) attached to a carbonyl group. The rest of the molecule is a complex polycyclic system with various functional groups, including hydroxyl groups, ester groups, and a benzene ring. The structure is drawn in a skeletal format with green lines for the main framework and orange lines for the highlighted group.

NPK reference spectrum

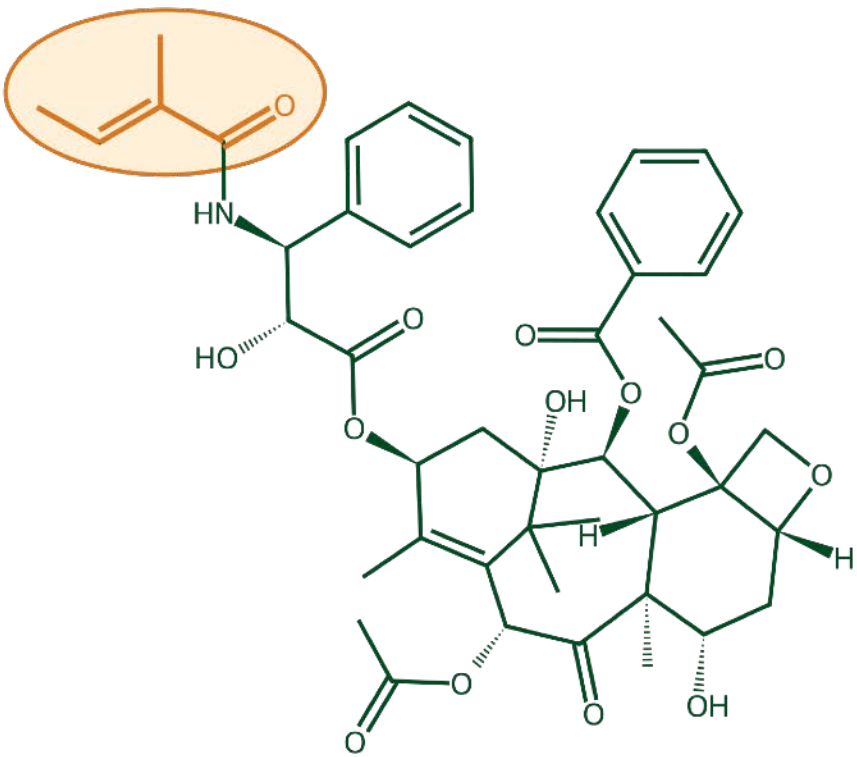
MoNA Alkaloids000154 · [M+H]⁺ · m/z 832.35

Could marine organisms make related molecules?

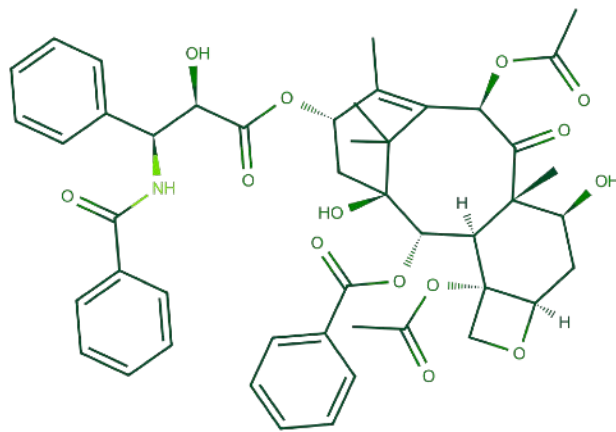
So many connections. Where do we look next?

Cephalomannine

NPK-0004701 · Taxane diterpenoid



Related NPK molecule

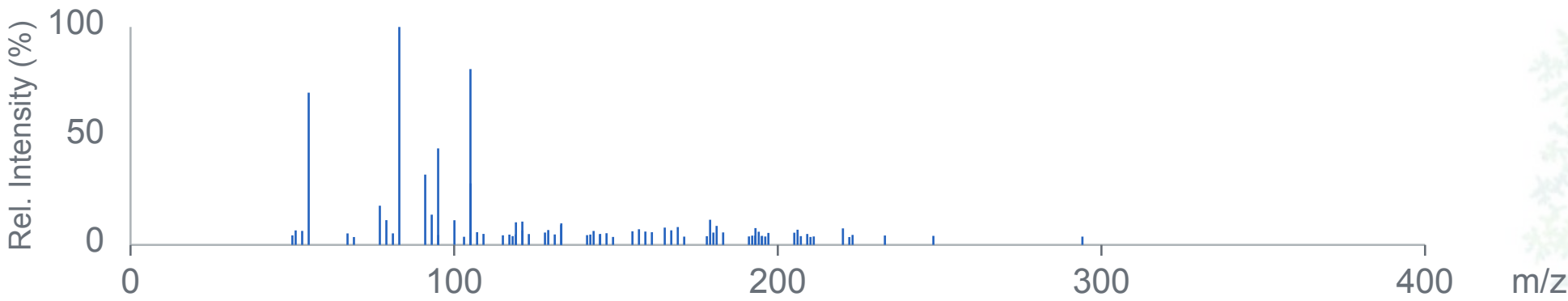


Paclitaxel (Taxol)

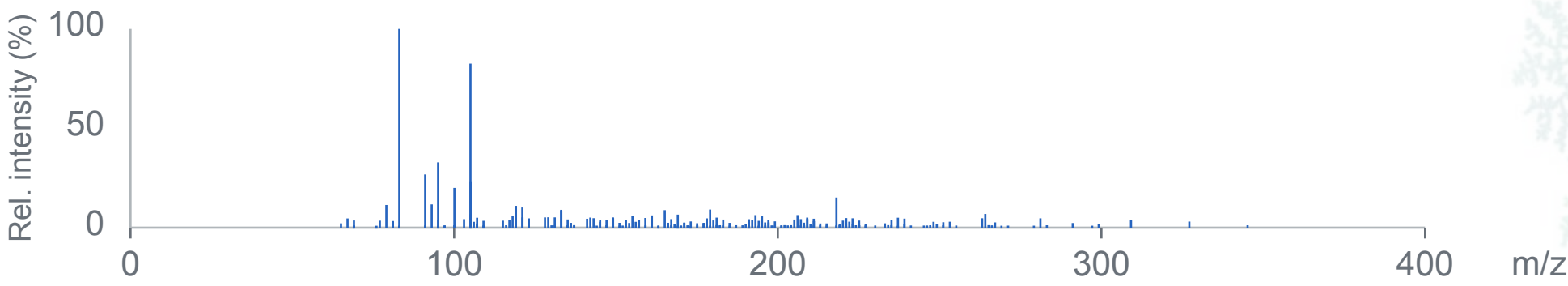
NPK-0002244

MassSpecGymID0232216

Atlas reference spectrum



NPK reference spectrum



DreaMS similarity

0.785

[M+H]⁺ · precursor m/z 832.35

Researcher + AI agent

What could explain these connections?



Examples from the reference neighbourhood

Seawater

MSV000085480

Atlas node 528664



Human-sample study

MSV000083559

Atlas node 17797298

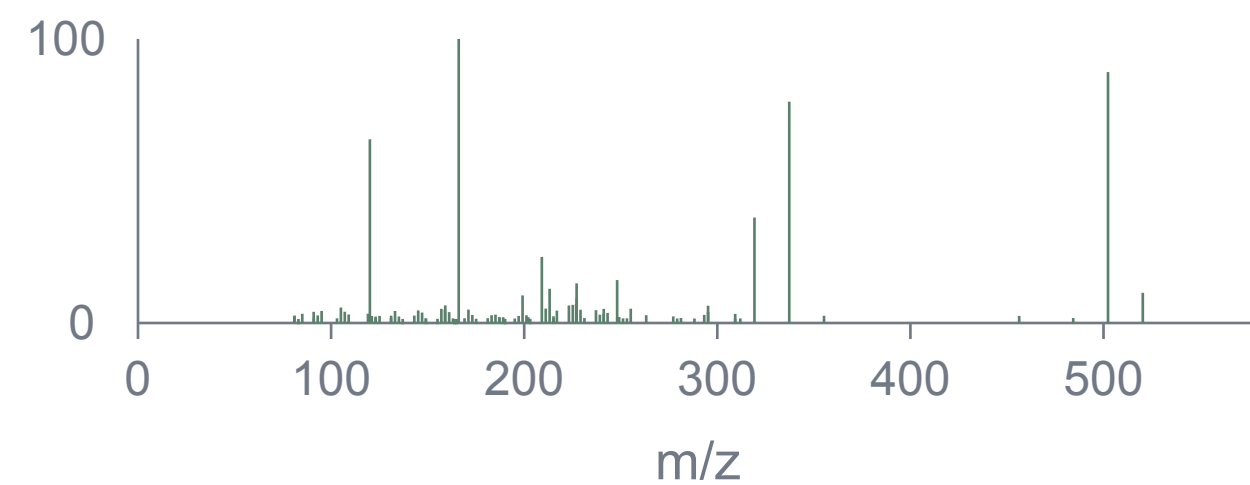


Choose what to test next

Measurements from different studies can help us ask a new biological question

Mouse microbiome study

How do microbes change the body's chemistry?



Phe-chol reference MS/MS

Earlier mouse liver-cancer study

How does inflammation promote liver cancer?



Samples with and without antibiotics

Connect evidence across studies

Human diet study

How does diet relate to our gut microbiome?



American Gut

Could gut bacteria be making these molecules?



Proposed agent support

Published discovery: Quinn and colleagues, *Nature* (2020)

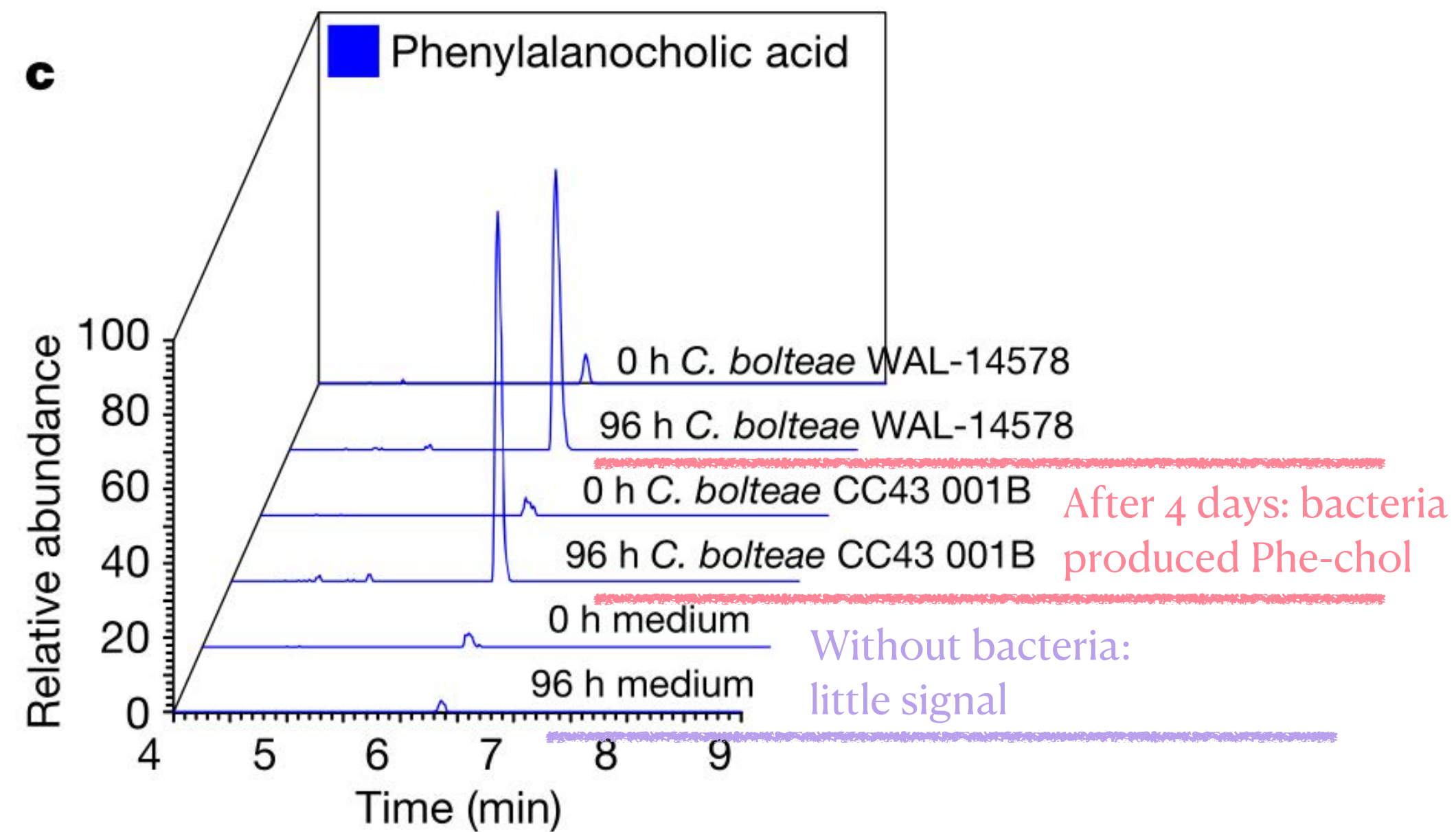


Could gut bacteria be making these molecules?

Yes. Two bacterial strains made new forms of bile acids.



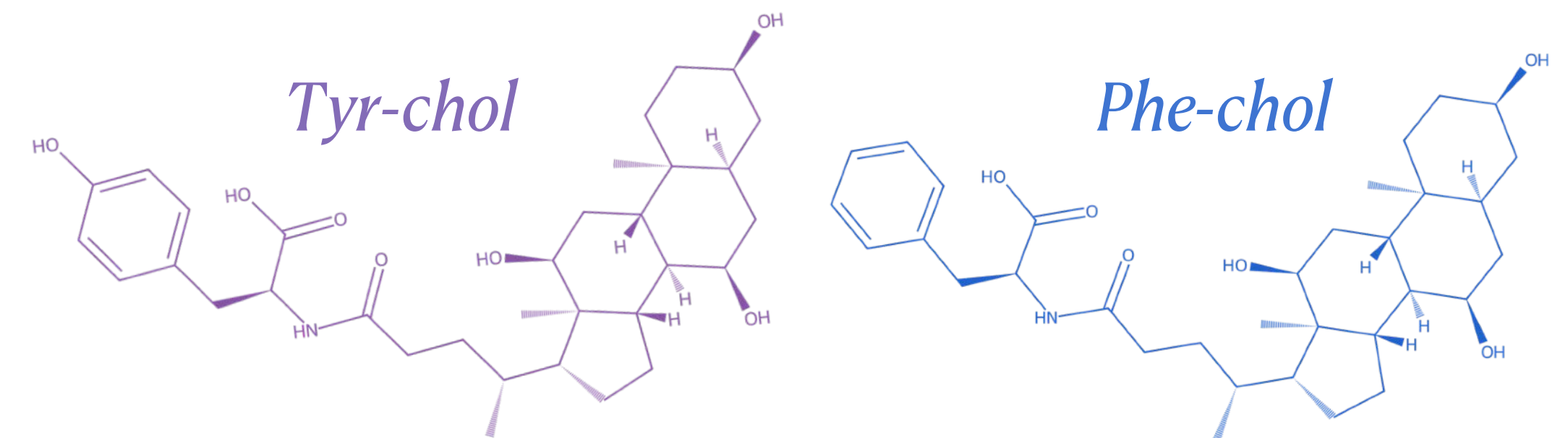
Bacteria produced the new molecules in culture



Published experiment: Phe-chol

Why does this matter?

Gut bacteria can make previously unknown versions of our own bile acids, and these molecules can affect how our cells receive signals.



Our ambition

Turn overlooked connections into testable hypotheses

DreaMS Search

Find similar spectra



Molecules and supporting evidence

DreaMS Atlas

Explore related measurements



Researcher + AI agent
Investigate together

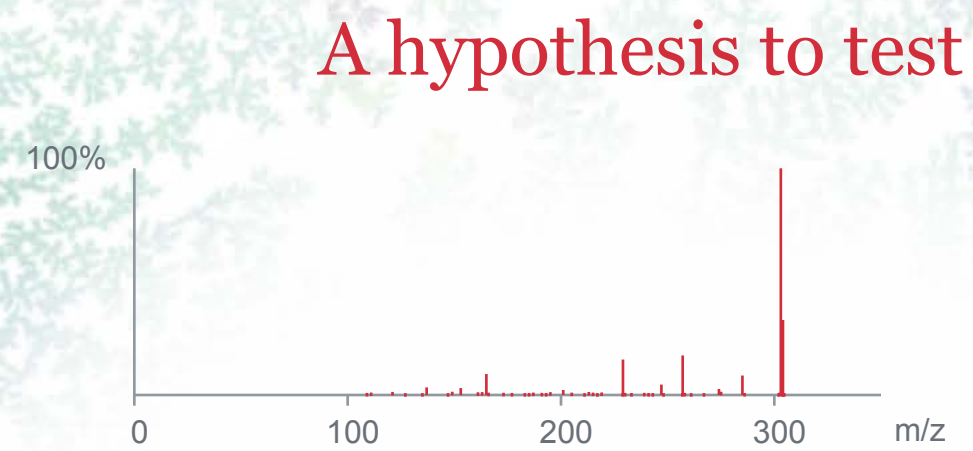
Gather evidence | Assess explanations | Propose an experiment to consider

Our next step

Test hypotheses in real experiments

DreaMS Search

Find similar spectra



What could explain this signal?



Molecules and supporting evidence

Wet-lab experiments

Does the explanation hold up?



Sample + control

New measurements → shared evidence

DreaMS Atlas

Explore related measurements



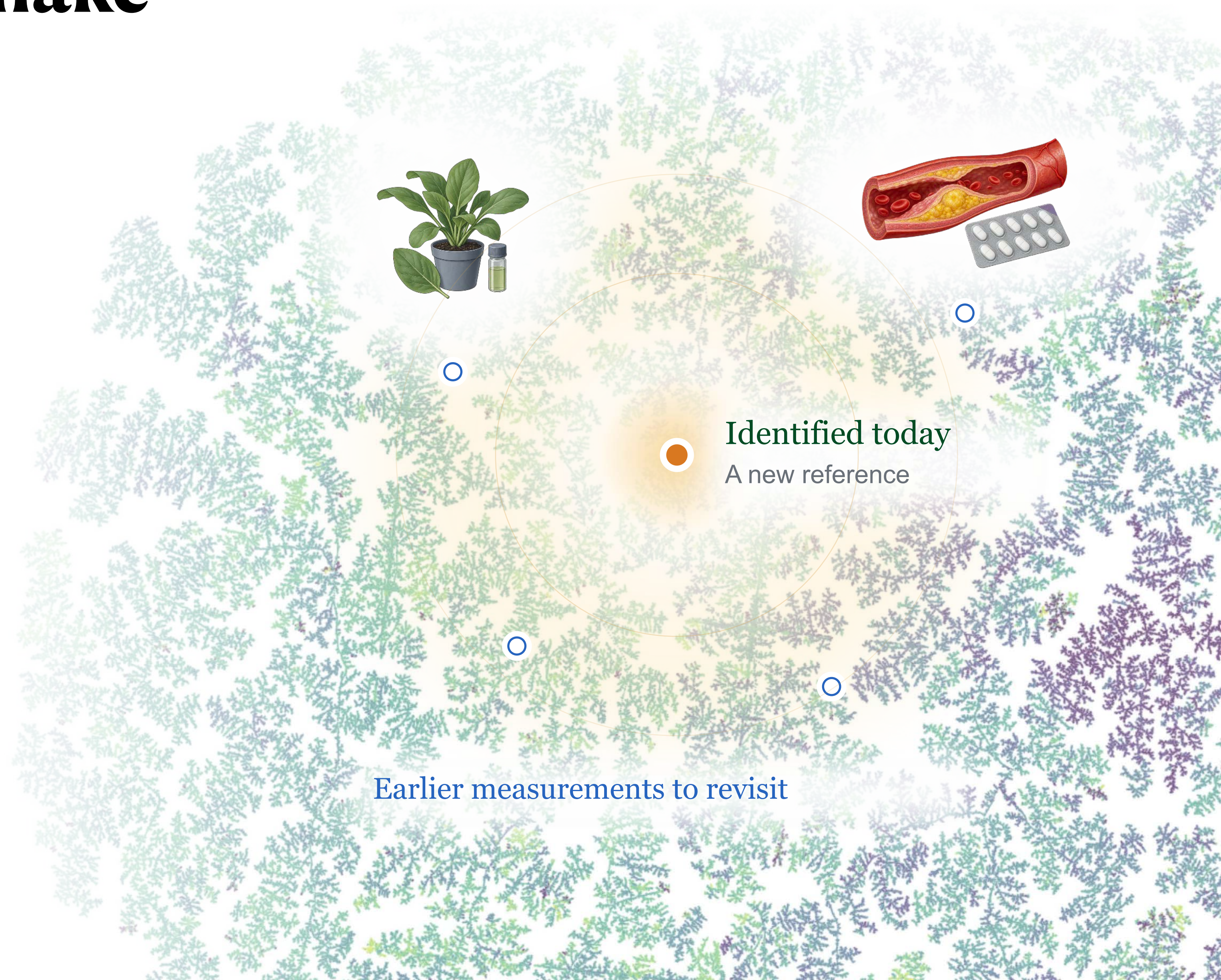
Researcher + AI agent

Plan and interpret experiments

Every discovery can make old data more useful

A new reference can help us revisit unexplained spectra from years ago.

The measurements are already there.
Our knowledge keeps growing.



DreaMS Search *Development Team*



Filip Jozefov, David Procházka, Emma
Sommerová, Boris Lukačovič



Andrej Tekel



Tito Damiani



Vlastislav Dohnal



Iuliia Nikolskaia



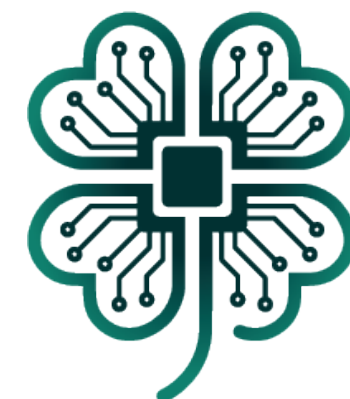
Téo Hebra



Aleš Křenek

MUNI

cesnet
....



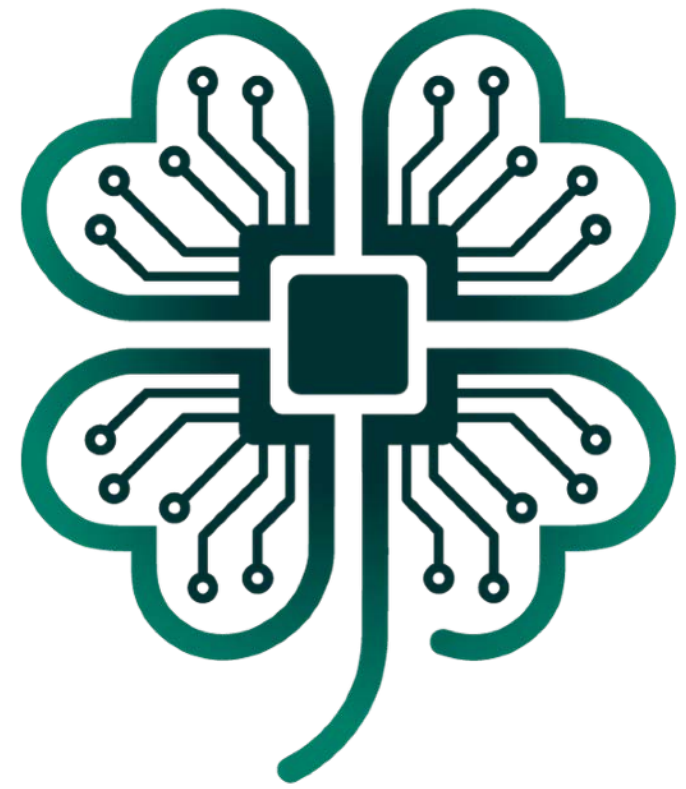
Pluskal Group
IOCB Prague

DreaMS Development Team



Josef Sivic

Tomáš Pluskal, Filip Jozefov, Roman Bushuiev, Anton Bushuiev



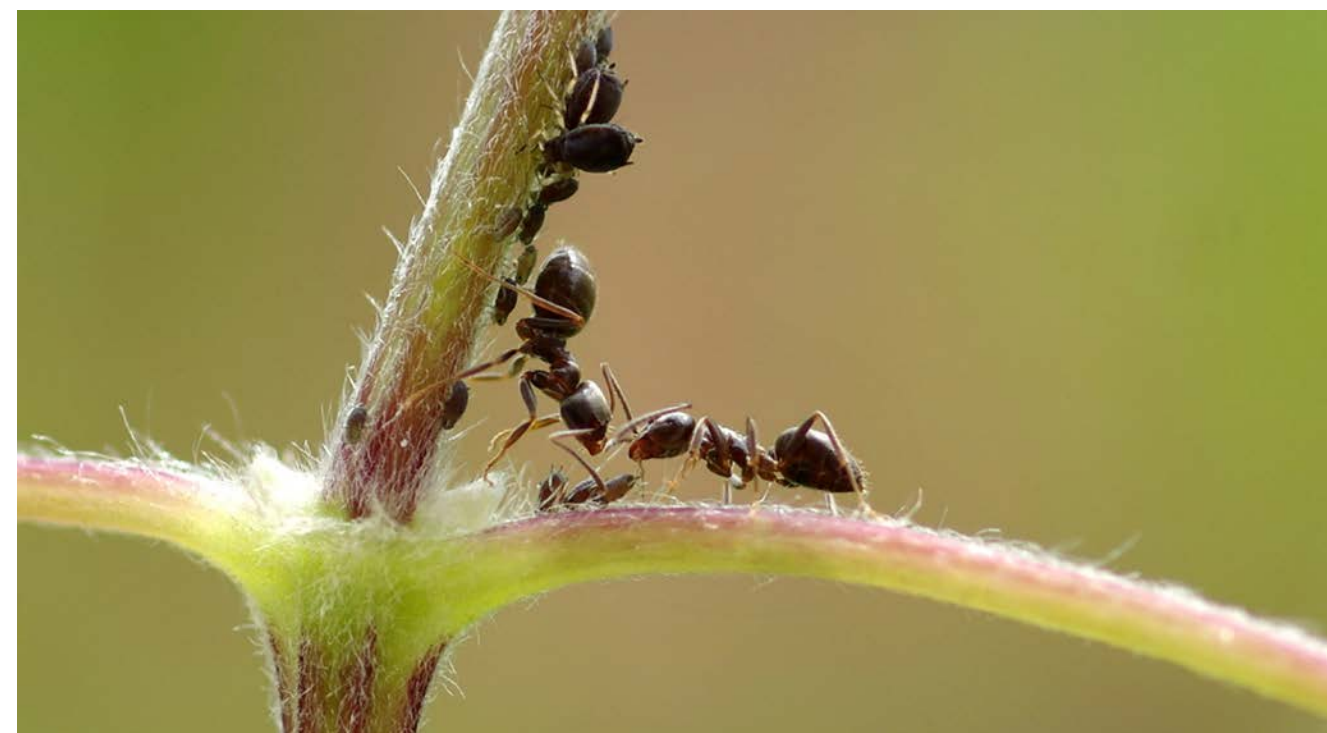
Pluskal group

Wet lab



Seoul National
University

Plant 🌱



Insects 🐛



JOO, YOUNGSUNG

Acknowledgments

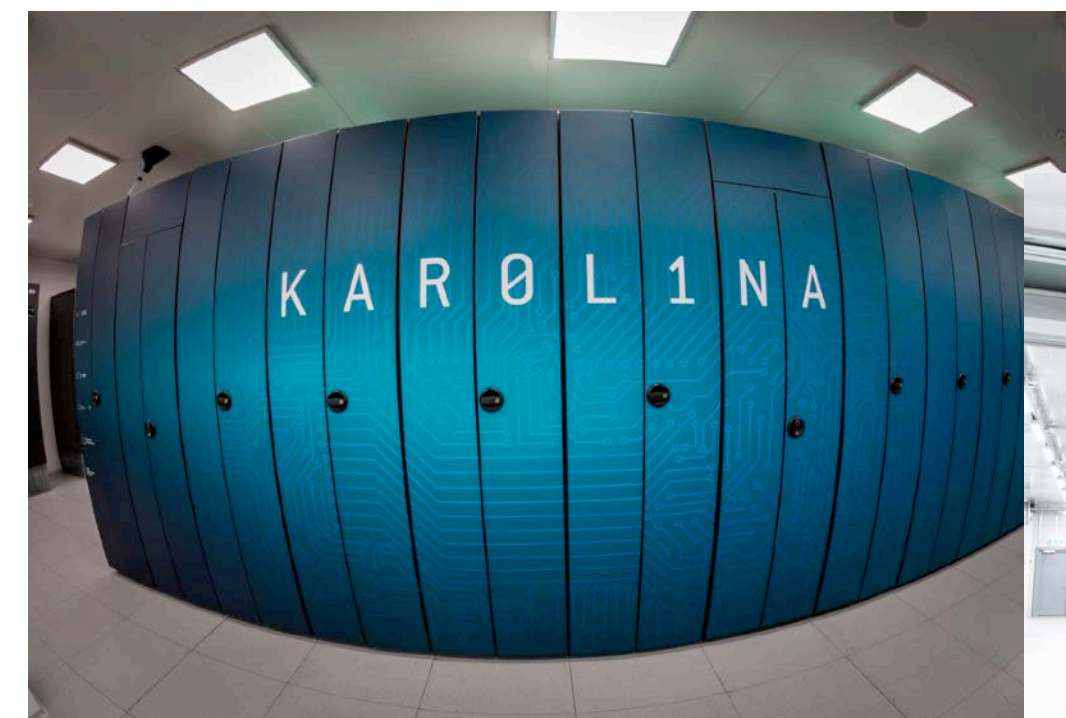


Tomáš Pluskal



Josef Sivic

Supercomputers



Industrial partners



European Research Council
Established by the European Commission



Funded by
the European Union

with support from
Google.org



Fin