UNIVERSITÉ DE GENÈVE

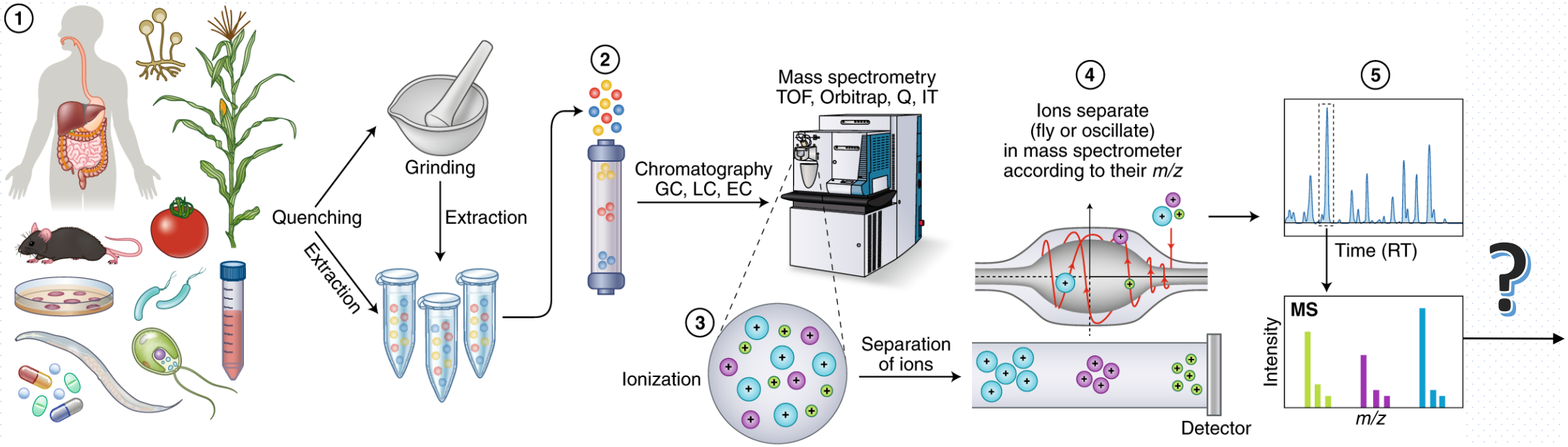
LSMS

COMPOUND IDENTIFICATION FROM MASS SPECTROMETRY AND *IN SILICO* MOLECULAR FRAGMENTATION

Stepan Stepanovic

Postdoc at Life Sciences Mass Spectrometry group
Department of Inorganic and Analytical Chemistry, University of Geneva
24 Quai Ernest Ansermet, CH-1211 Geneva 4, Switzerland

Identification of analytes in a sample



Sample preparation and extraction

- Avoid environmental perturbation during harvesting
- Control environment: harvesting at the same time and under the same conditions
- Snap-freezing in liquid nitrogen
- Enzyme quenching: completely terminate all enzyme activities
- Standards spiked into the quenching solvent
- Grinding, isolation of cells, fast-filtration or aspiration

Sample replication and randomization

- At least four biological replicates, preferably more
- Technical and analytic replicates are worthy of consideration
- Randomization of samples throughout workflows is essential
- In large-scale studies, quality-control samples and batch correction are essential

Chromatography–mass spectrometry

- Separation methods, composition of the mobile phase, column properties and injection volume
- Metabolites are within their range of detection
- Avoid ion suppression: dilution of extracts, sonication, filtration or centrifugation, recovery test
- Choosing ionization source and type of detection mode, MS method, scan number and speed, MS/MS and energy for fragmentation

Nat. Methods, **2021**, *18*, 747–756

Identification of analytes in a sample

OPEN-ACCESS SPECTRAL LIBRARIES:

MassBank

MONA
MEDICAL OPEN NETWORK FOR AI

m/z
CLOUD

Access issues

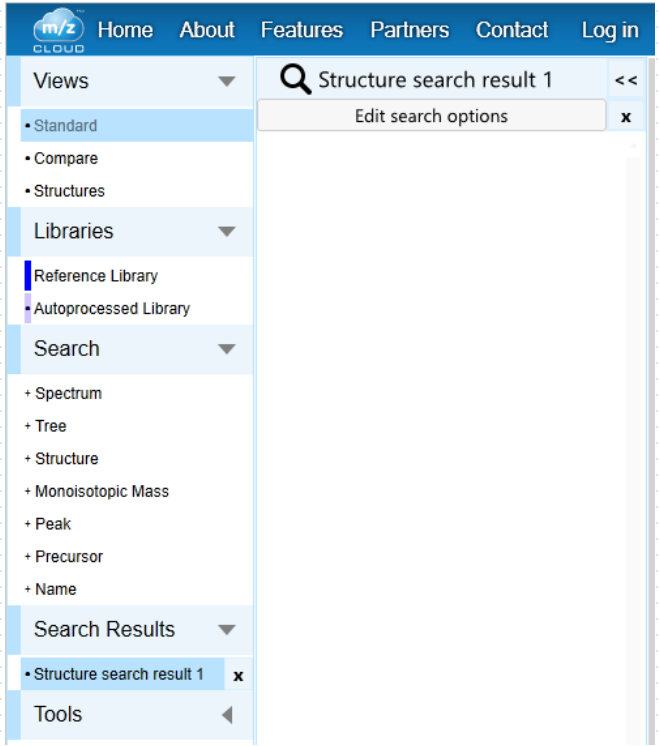
m/z CLOUD™

Microsoft® Silverlight™

Supported Operating Systems and Browsers

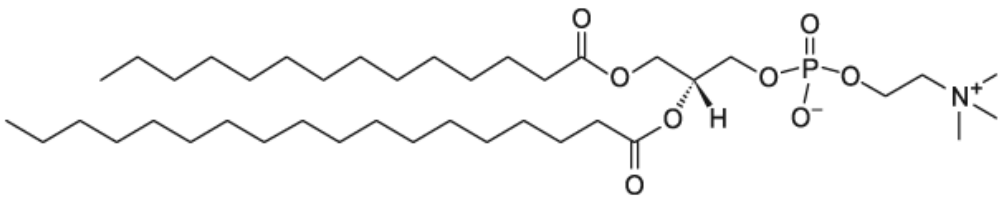
		Internet Explorer	Edge	Mozilla Firefox	Chrome	Safari	Desktop Installation
Microsoft Windows	XP	✗	N/A	✗	✗	✗	✗
	Vista	✗	N/A	✗	✗	✗	✗
	7	✓	N/A	✗	✗	✗	✓
	8	✓	N/A	✗	✗	✗	✓
	10	✓	✗	✗	✗	✗	✓
Mac OS X	10.7.5 (Lion)	✗	N/A	✗	✗	✓	✓
	10.9.2 (Mavericks)	✗	N/A	✗	✗	✓	✓
Linux	All	✗	N/A	✗	✗	✗	✗

Precursor Ion Fingerprinting (PIF) technique



Comparison of spectra

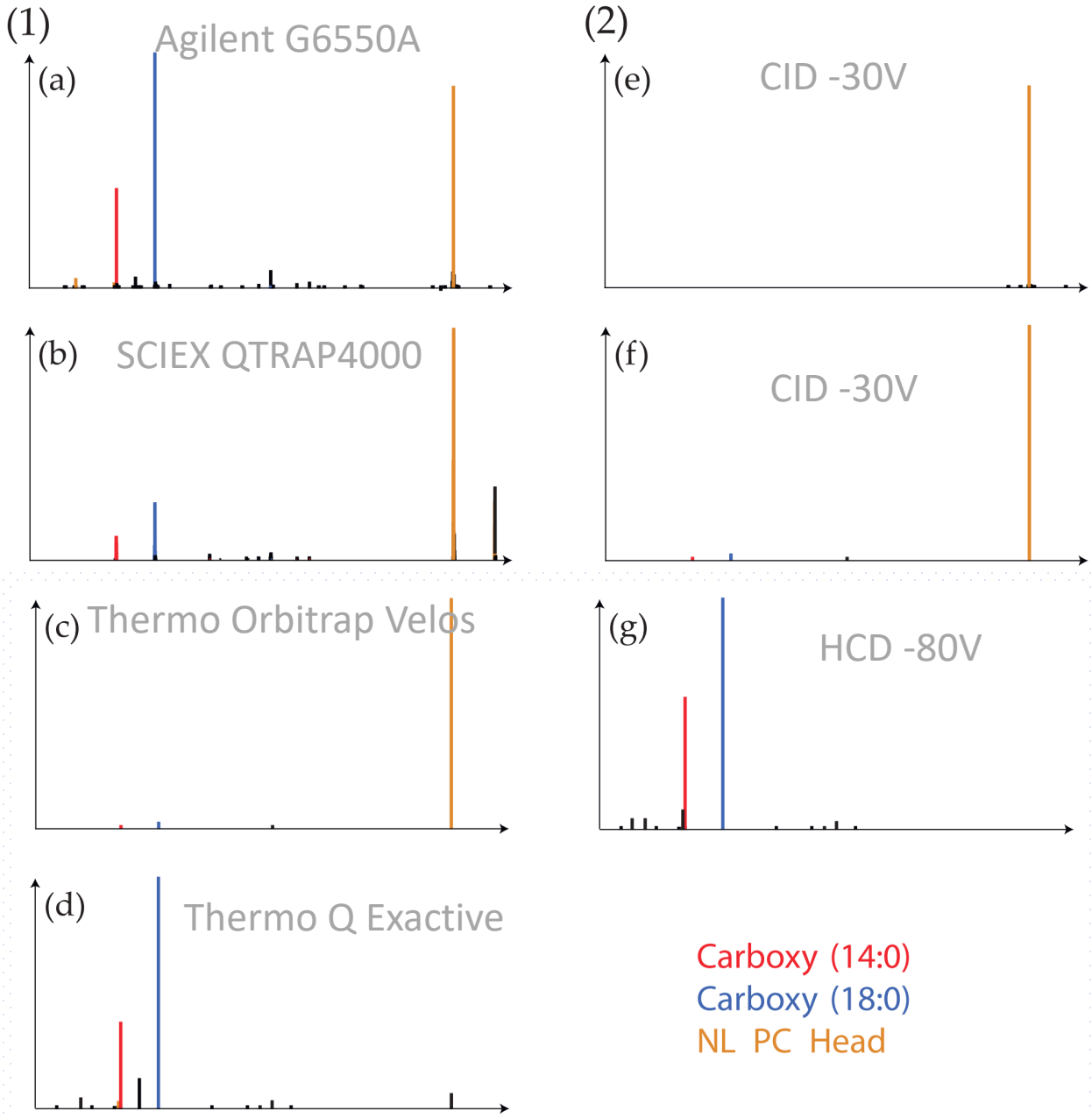
Molecular and fragment peak intensity highly depends on :
(1) Equipment used
(2) Ionization method



14:0-18:0 PC

1-myristoyl-2-stearoyl-sn-glycero-3-phosphocholine

extensive shorthand notation for mass spectrometry-derived lipid identifications



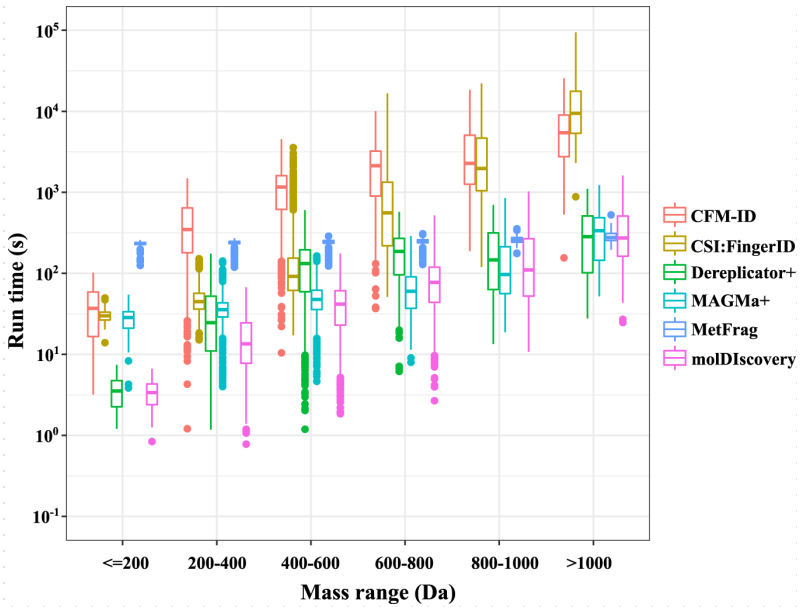
Briefings in Bioinformatics, **2021**, 22, *6*, bbab073

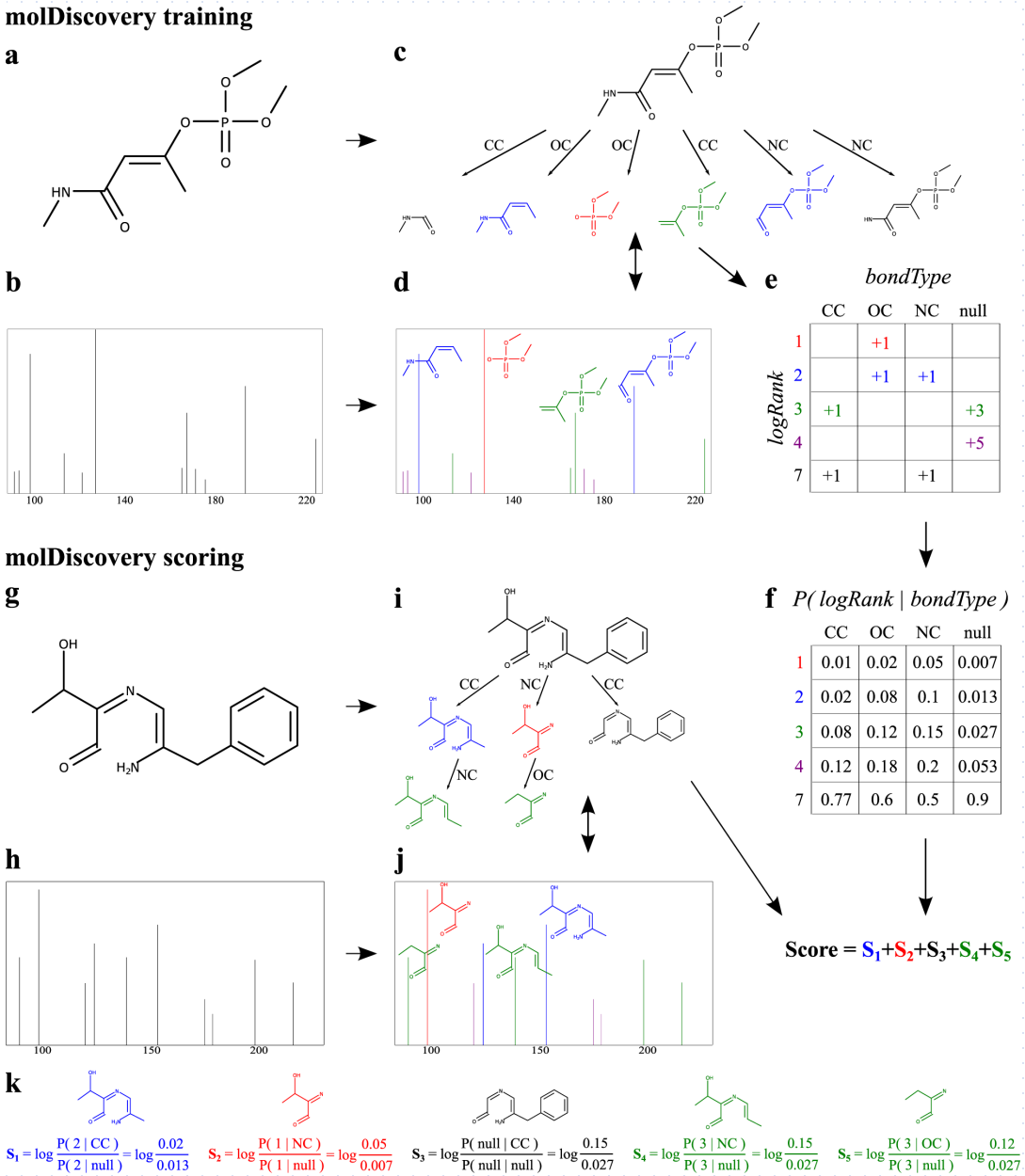
MolDiscovery: learning mass spectrometry fragmentation of small molecules

Liu Cao [1,4], Mustafa Guler [1,4], Azat Tagirdzhanov [2,3], Yi-Yuan Lee [1], Alexey Gurevich [2] & Hosein Mohimani [1✉]

Nature Comm., **2021**, 12:3718

molDiscovery is a mass spectral **database search method** that improves both efficiency and accuracy of small molecule identification by learning a probabilistic model to match small molecules with their mass spectra





Challenges in similarity search

Manual interpretation:

- × Tedious
- × Time-consuming
- × Error-prone

Bioinformatic tools:

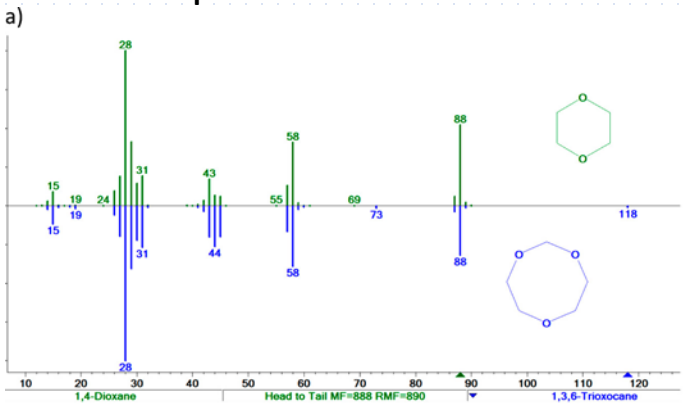
- ✓ Much faster
- ✓ Automatized
- × Still error-prone (FALSE POSITIVES)

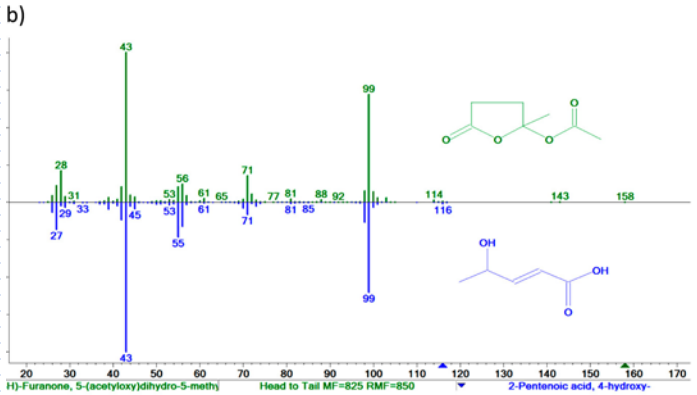
Types of false positives occurring in similarity searches:

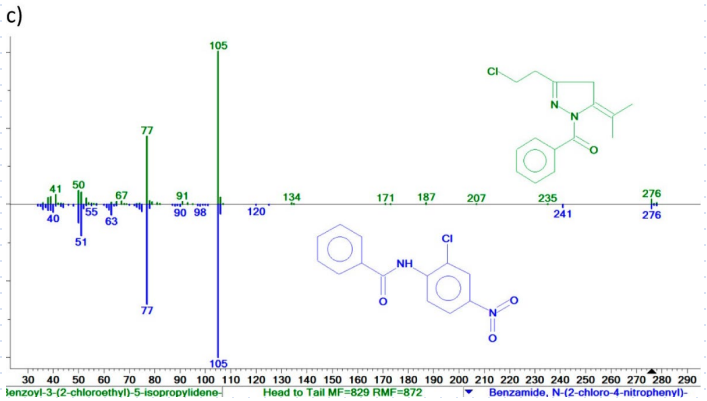
- lost structural information upon fragmentation
- accidental match
- identical by MS same connectivity
- isomers with minimal differences in fragmentation
- uncertain structure in library

Anal. Chem. **2012**, 84, 17, 7274–7282

Examples:







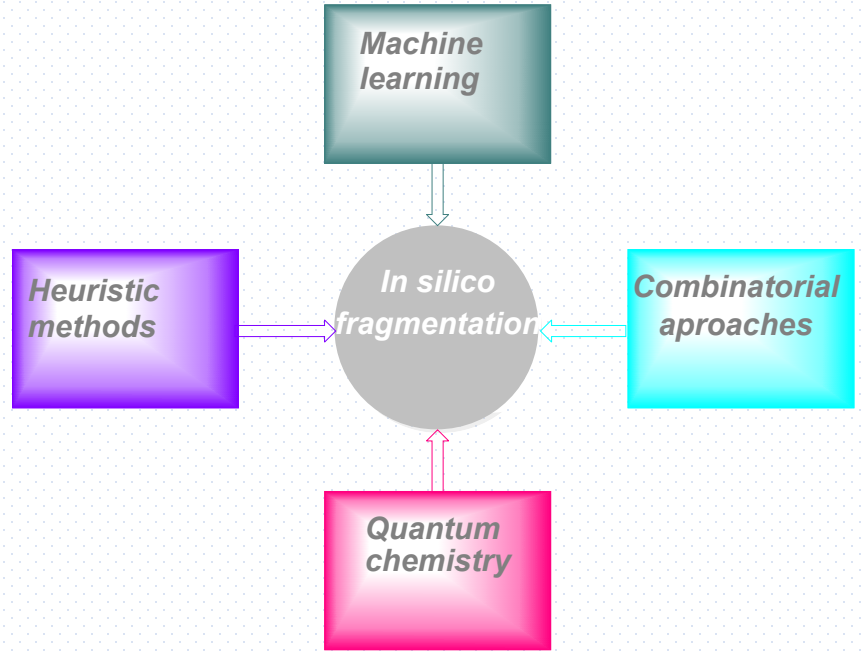
Untargeted analysis of metabolites

Human Metabolome Database (HMDB 5.0) currently contains MS^2 spectra for only ~4000 compounds, although it contains a total of ~250000 human metabolites

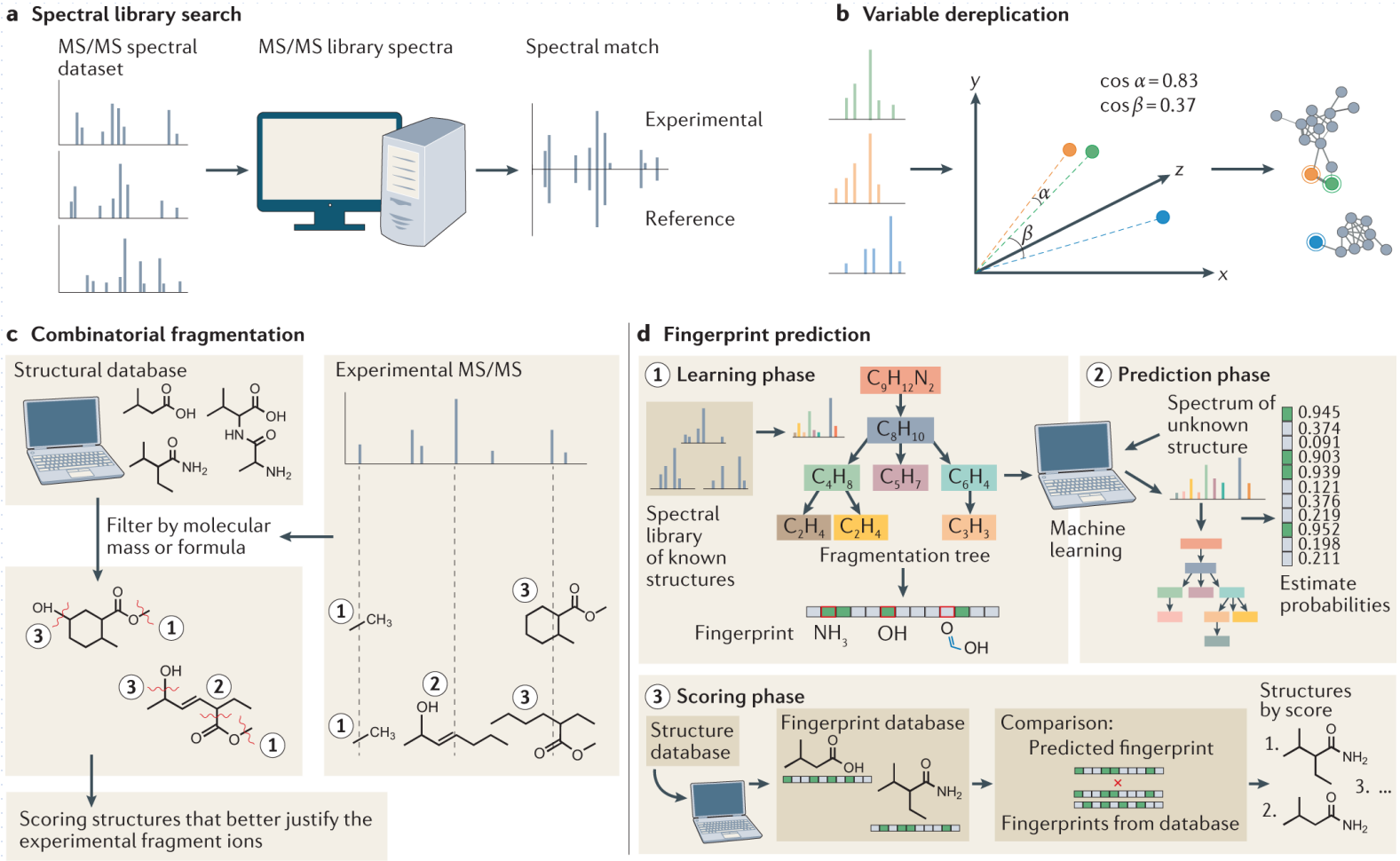
Only ~2% of spectra in an untargeted metabolomics experiment can be annotated

Computational methods

in silico fragmentation of analytes and spectral classifiers:



Computational tools for metabolite annotation, substructure assessment and chemical classification



Bauermeister, A., Mannochio-Russo, H., Costa-Lotufo, L.V. *et al.* *Nat Rev Microbiol* **20**, 143–160 (2022). https://doi.org/10.1038/s41579-021-00621-9

Selected milestones towards *in silico* fragmentation modelling

1965 — DENDRAL

Pioneering AI research at Stanford with the major goal of automatic structural elucidation of compounds by mass spectral data; led to general models of fragmentation, as well as some class-specific MS fragmentation rules.

1992 — MASSIMO

A knowledge-based approach combining chemical expertise with experimental MS data; utilizes logistic regression in order to predict fragmentation probabilities for various types of reactions happening during electron ionization (EI) MS.

2011 — LipidData Analyzer

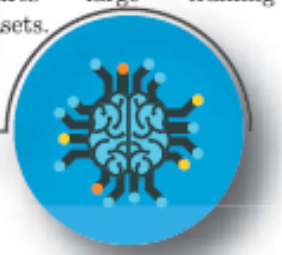
A heuristic approach utilizing a novel 3D algorithm to allow identification and quantification of lipids in LC-MS data through expert-curated fragmentation rules.

2013 — QCEIMS

Introduces a quantum chemistry framework for the prediction of complete mass spectra; utilizes Born–Oppenheimer *ab initio* molecular dynamics to generate *in silico* EI mass spectra from input compounds without the need for additional data.

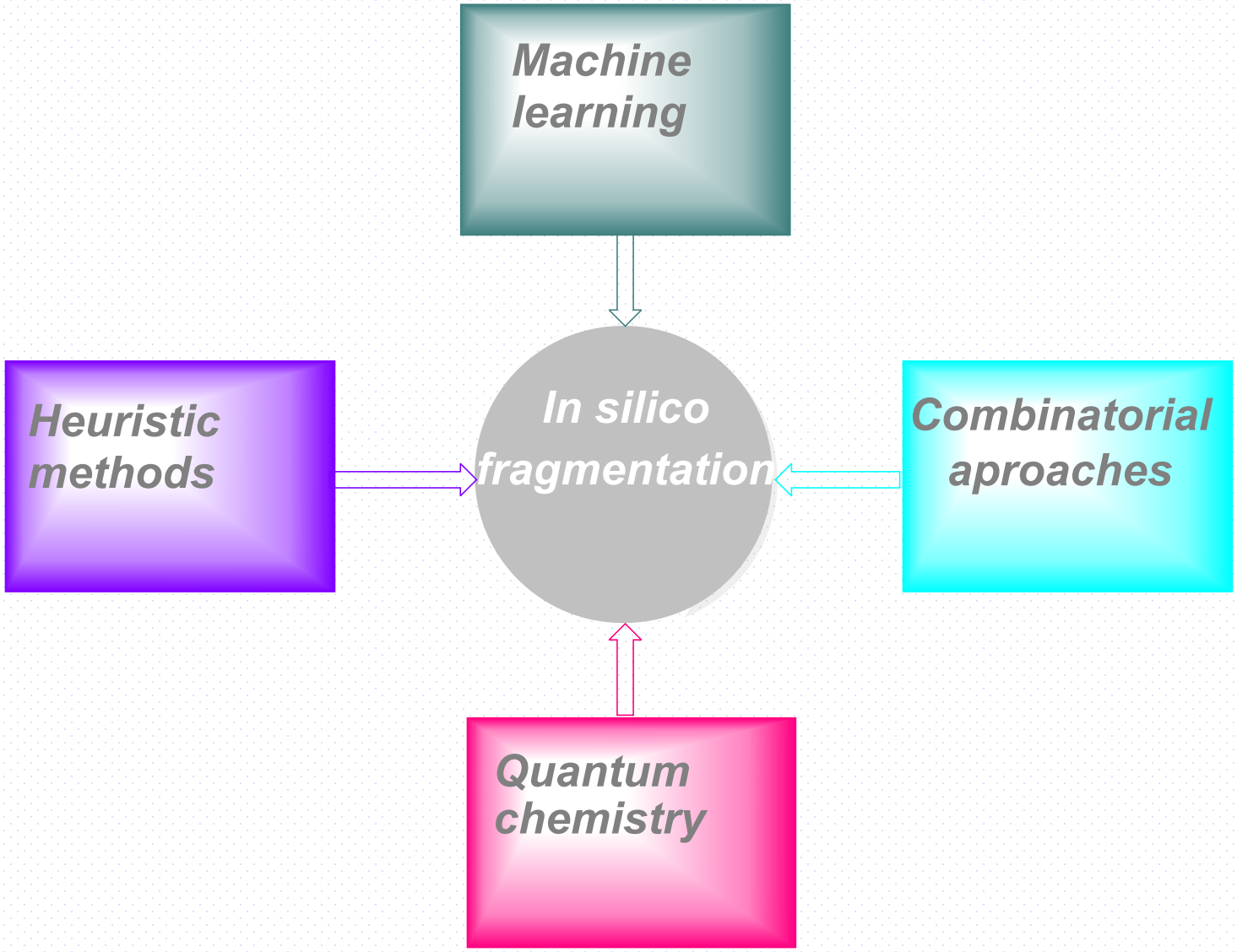
2014 — CFM-ID

Introduces a novel probabilistic generative method to model the MS fragmentation process; model parameters are learned directly from mass spectrometry data; requires large training datasets.

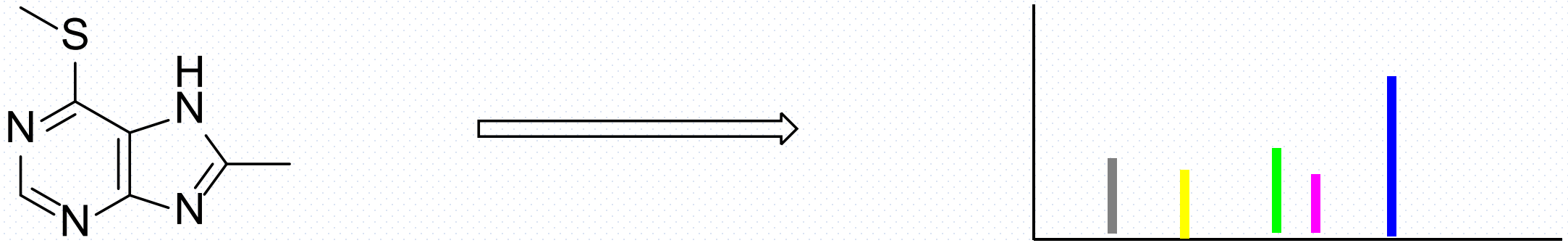


2020 — SIRIUS+ CSI:FingerID

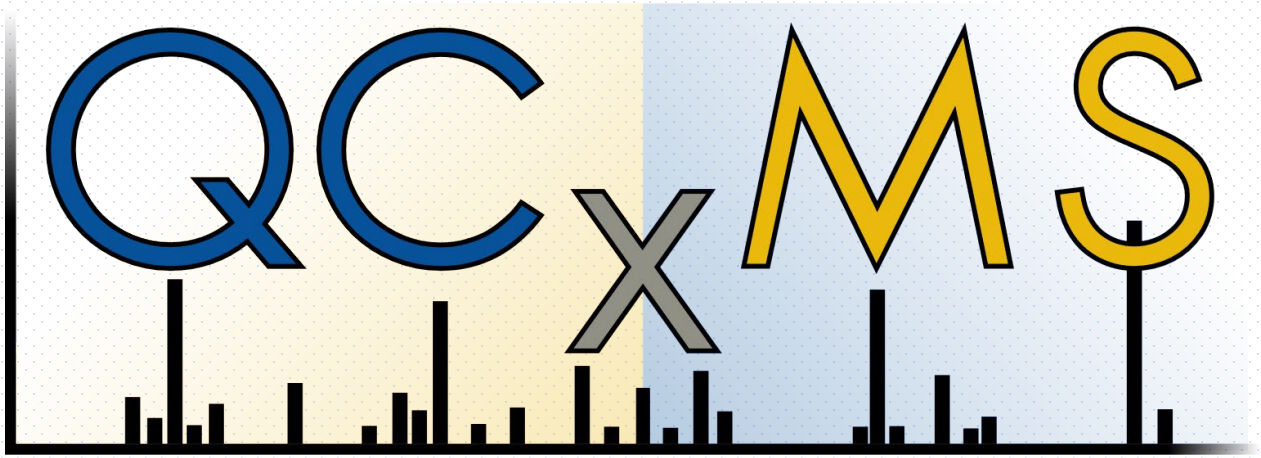
ML method the computes a structural features from MS spectrum.



Krettler, C. A., Thallinger G. G. *Briefings in bioinformatics* vol. 22,5 (2021): bbab073. DOI: 10.1093/bib/bbab073


S
H
N
N
N
N

on-the-fly computed potential energy surfaces with Born-Oppenheimer *ab initio* molecular dynamics



quantum chemical (QC) based program that enables users to calculate mass spectra (MS) using Born-Oppenheimer Molecular Dynamics (MD)

Model: semi empirical, DFT

MS methods: **EI, CID**, simulâtes full spectra

S. Grimme . *Angew. Chem. Int. Ed.*, **2013**, 52, 6306-631. DOI: 10.1002/anie.201300158
Koopman, J.; Grimme, S. *From J. Am. Soc. Mass Spectrom*, **2021**. DOI: 10.1021/jasms.1c00098

QCMS²

The method is automated and has been implemented into the BRABO and STOCK software packages

Conformational search with AM1, fragmentation pathways with B3LYP/6311+G*, TS energies using QST3

Quantum Chemistry + Heuristics *Platform (utilizes* straightforward calculations of bond orders and energies of possible fragments).

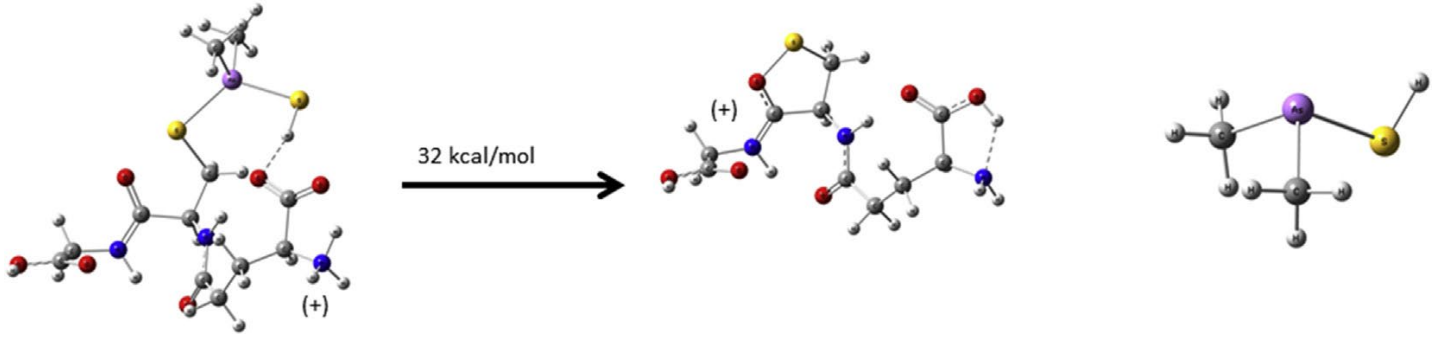
Model: DFT

MS methods: **EI, CID**, generates insights into the fragmentation process

Cautereels, J.; Van Hee, N.; Chatterjee, S.; Van Alsenoy, C.; Lemière, F.; Blockhuys, F. *J. Mass Spectrom.* **2020**, *55*, e4446, 10.1002/jms.4446

Quantum Chemical Fragment Precursor Tests - QC-FPT

Perl sctipts + gaussian DFT(B3LYP/6311+G*)



Quantum Chemistry + Combinatorial Optimization
(cleaves bonds of candidate precursors to identify the fragment precursor).

Users have to generate candidate precursors complete with 3D structures as input, which makes it infeasible

Fragments (focuses on ones that are already known) are generated by systematic bond cleavages of precursors, only their thermodynamic feasibility is investigated

Janesko BG, Li L, Mensing R. *Anal Chim Acta* 2017;**995**:52–64. DOI: 10.1016/j.aca.2017.09.034

ChemFrag

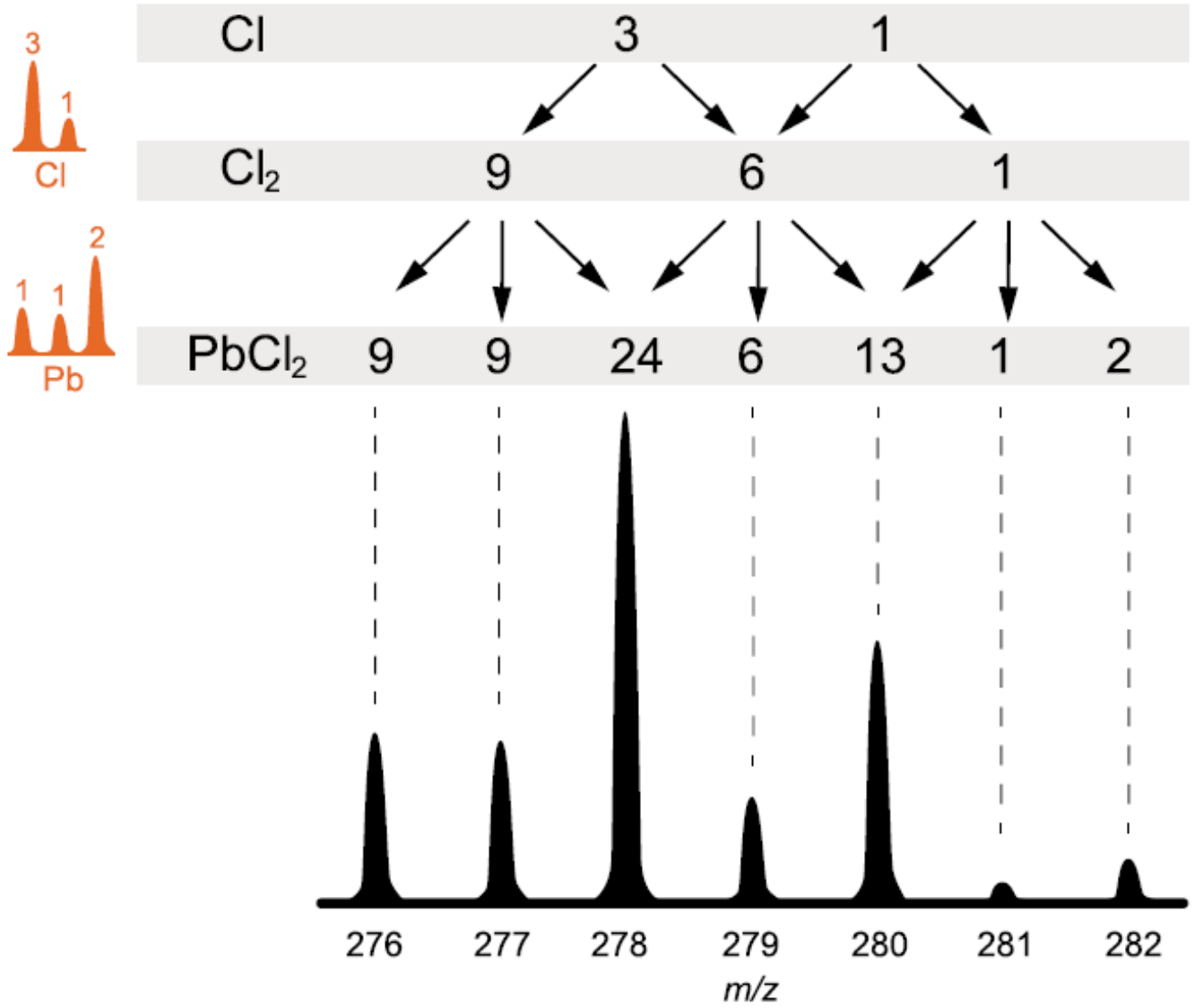
Quantum Chemistry + Heuristics *Platform (*the most unstable bonds (PM7) are cleaved and empirically derived rules are applied to model rearrangements).

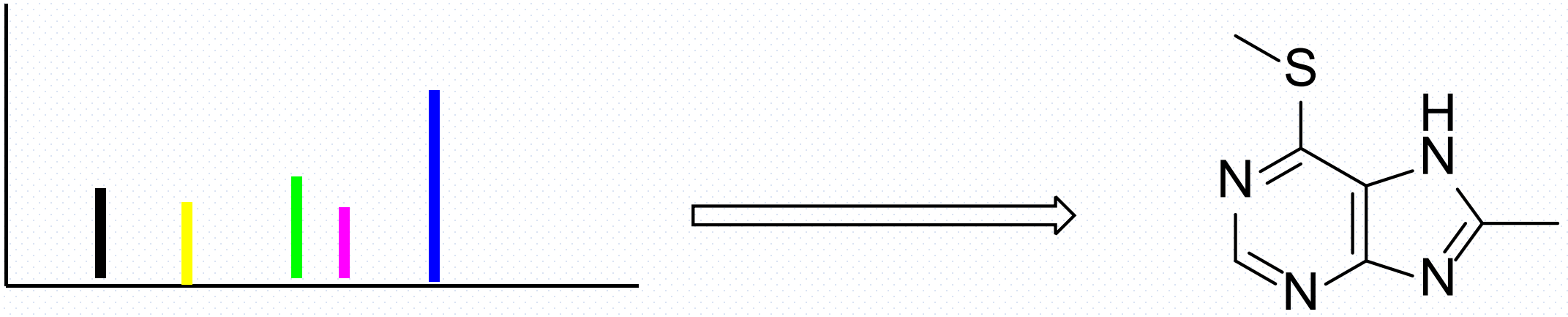
The QC approach is combined with a rule-based Approach (do not cover all possible processes) for fast modelling of rearrangement reactions (up to minuts). Not available.

MS methods: **CID**, barcode spectra

Schüler JA, Neumann S,Müller-Hannemann M, *et al. J Mass Spectrom* 2018;**53**(11):1104–15. **DOI: 10.1002/jms.4278**

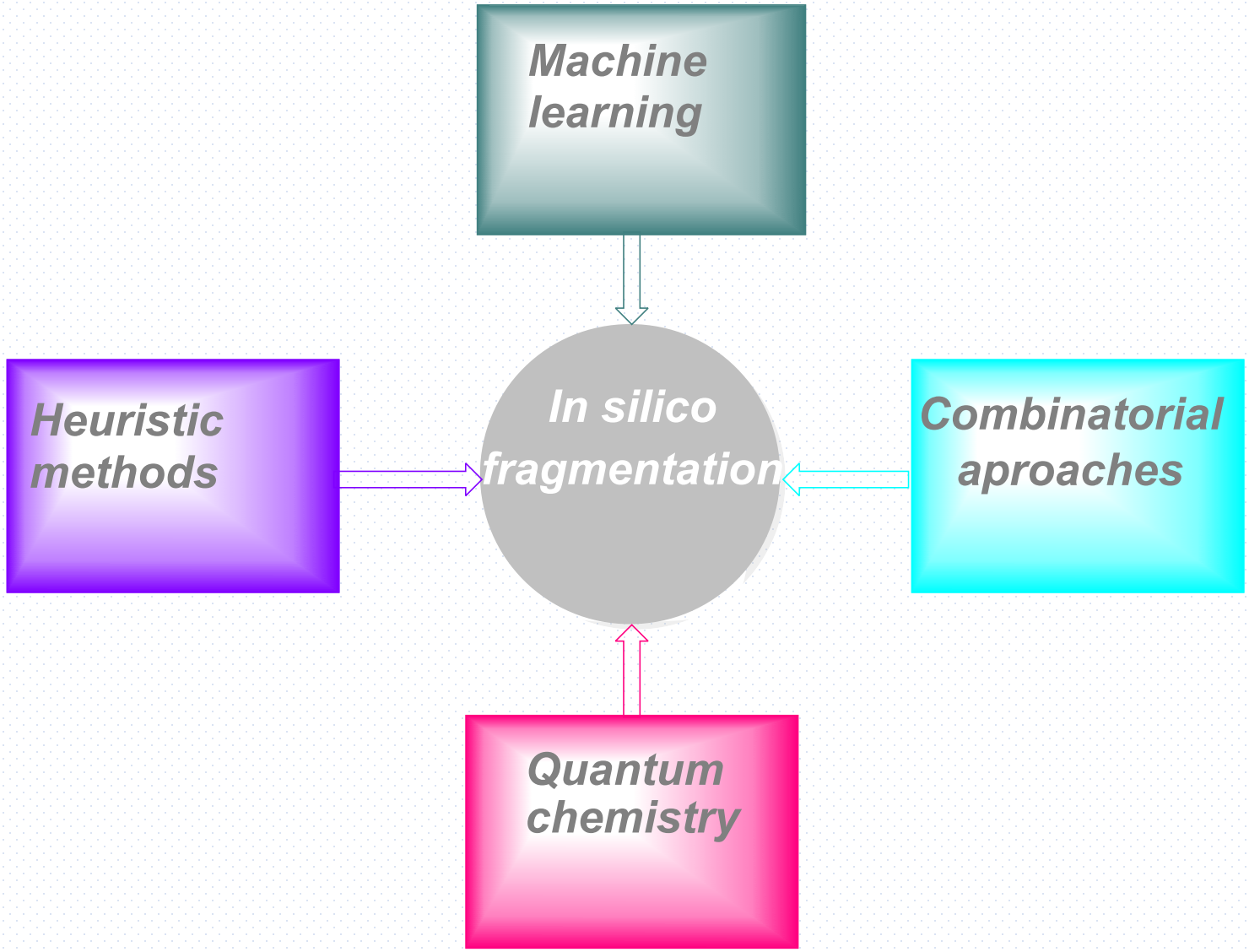
ISOTOPE PATTERN

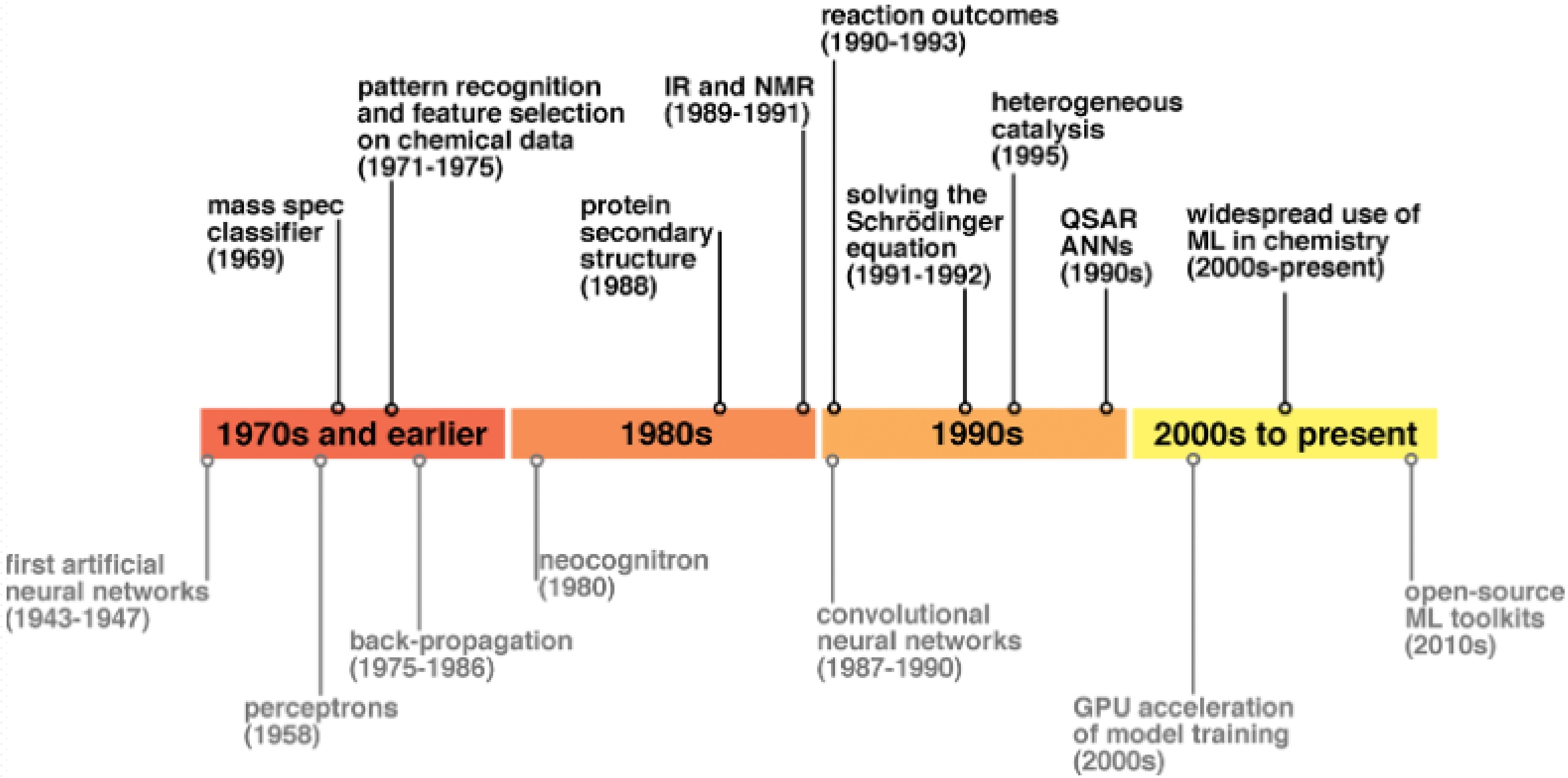




The ‘inverse problem’

THIS PAGE INTENTIONALLY LEFT BLANK


Machine learning
Heuristic methods
In silico fragmentation
Combinatorial aproaches
Quantum chemistry


reaction outcomes
(1990-1993)
pattern recognition
and feature selection
on chemical data
(1971-1975)
IR and NMR
(1989-1991)
heterogeneous
catalysis
(1995)
mass spec
classifier
(1969)
protein
secondary
structure
(1988)
solving the
Schrödinger
equation
(1991-1992)
QSAR
ANNs
(1990s)
widespread use of
ML in chemistry
(2000s-present)
1970s and earlier
1980s
1990s
2000s to present
first artificial
neural networks
(1943-1947)
perceptrons
(1958)
back-propagation
(1975-1986)
neocognitron
(1980)
convolutional
neural networks
(1987-1990)
GPU acceleration
of model training
(2000s)
open-source
ML toolkits
(2010s)

The New York Times

A.I. Predicts the Shape of Nearly Every Protein Known to Science

DeepMind has expanded its database of microscopic biological mechanisms, hoping to accelerate research into all living things.

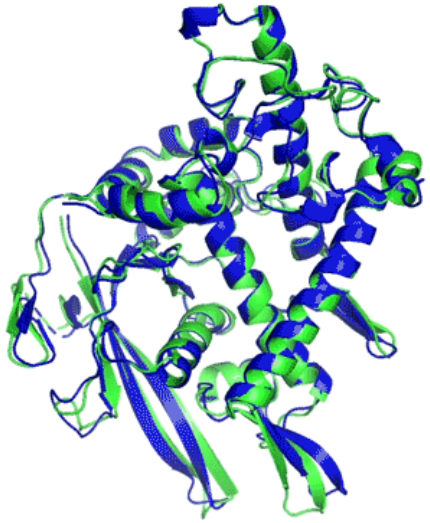
Forbes

AI

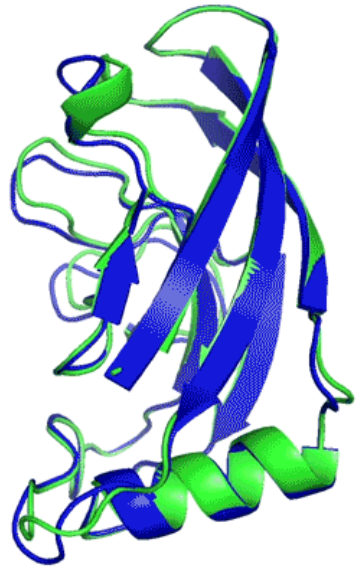
AlphaFold Is The Most Important Achievement In AI—Ever

The Guardian

DeepMind uncovers structure of 200m proteins in scientific leap forward



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)



T1049 / 6y4f
93.3 GDT
(adhesin tip)

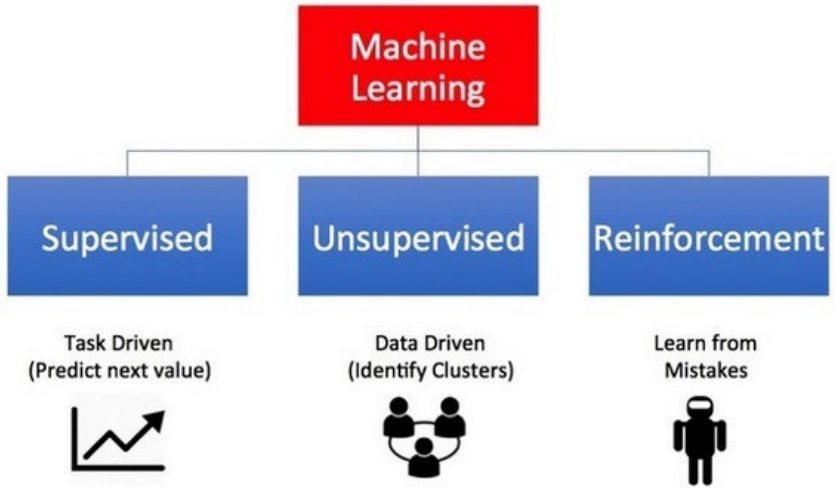
Experimental result

Computational prediction

Languages

Python 92.4% Jupyter Notebook 5.1%
Shell 2.1% Dockerfile 0.4%

Jumper, J., Evans, R., *et al.* *Nature* **596**, 583–589 (2021). DOI:10.1038/s41586-021-03819-2



Supervised Learning

Data: (x, y)
x is data, y is label

Goal: Learn function to map
$x \rightarrow y$

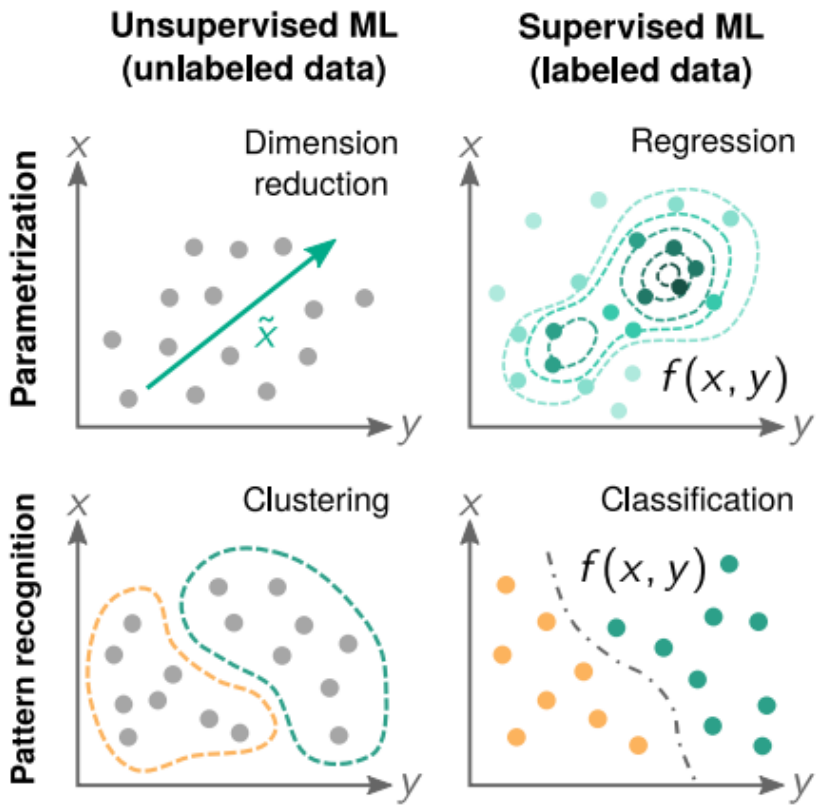
Examples: Classification, regression, object detection, semantic segmentation, etc.

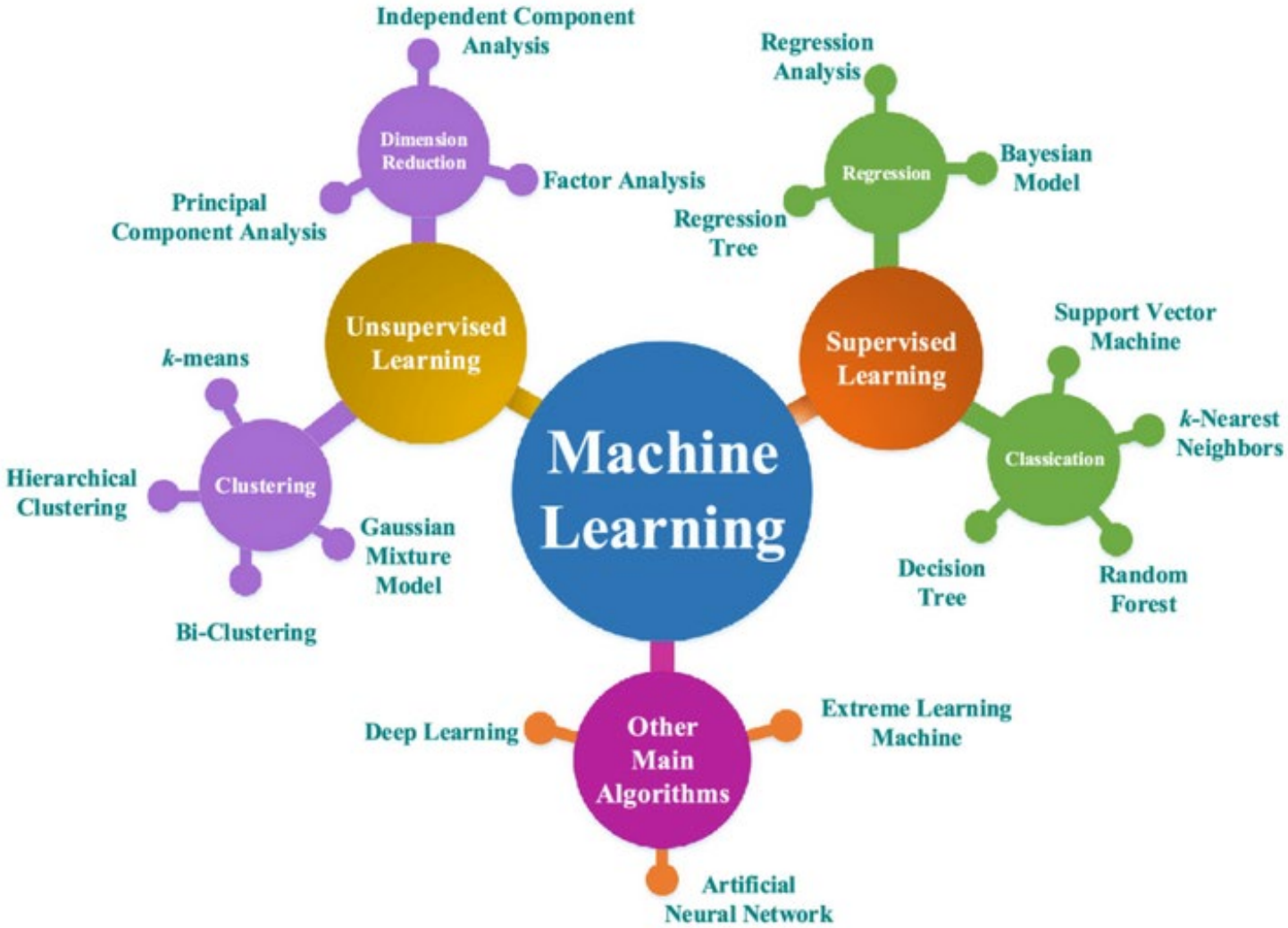
Unsupervised Learning

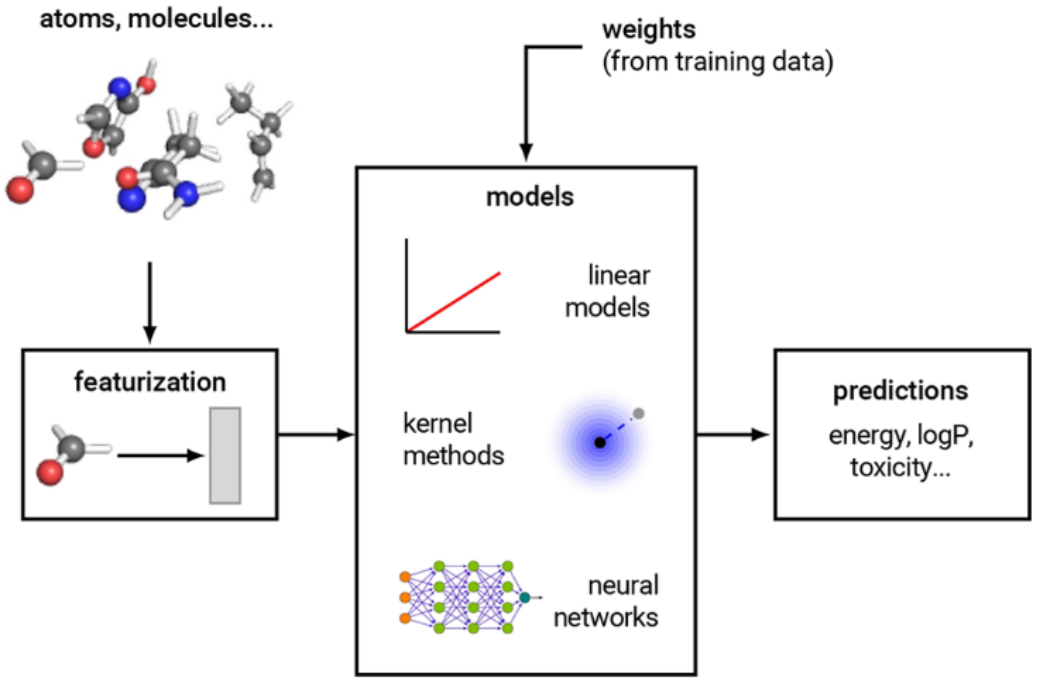
Data: x
x is data, no labels!

Goal: Learn some *hidden* or *underlying structure* of the data

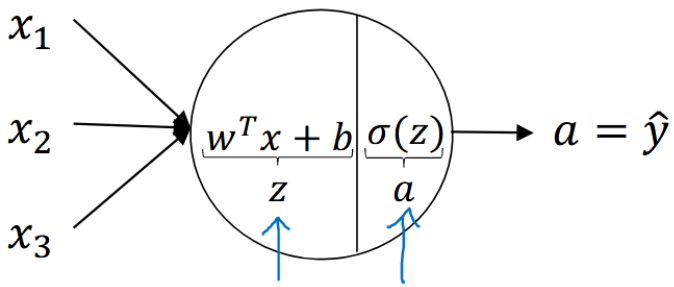
Examples: Clustering, feature or dimensionality reduction, etc.

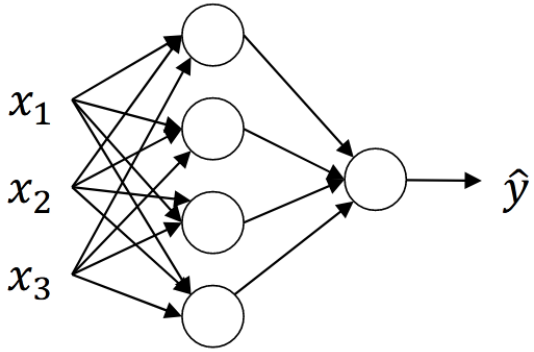



Machine Learning
Unsupervised Learning
Supervised Learning
Other Main Algorithms
Dimension Reduction
Clustering
Regression
Classication
Independent Component Analysis
Principal Component Analysis
Factor Analysis
k-means
Hierarchical Clustering
Gaussian Mixture Model
Bi-Clustering
Regression Analysis
Regression Tree
Bayesian Model
Support Vector Machine
k-Nearest Neighbors
Decision Tree
Random Forest
Deep Learning
Extreme Learning Machine
Artificial Neural Network


atoms, molecules...
weights
(from training data)
featurization
models
linear models
kernel methods
neural networks
predictions
energy, logP, toxicity...

Neural Network Representation





$z = w^T x + b$

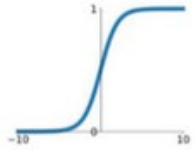
$a = \sigma(z)$

*The **loss** of our network measures the cost incurred from incorrect predictions*

Activation Functions

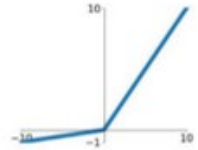
Sigmoid

$\sigma(x) = \frac{1}{1+e^{-x}}$



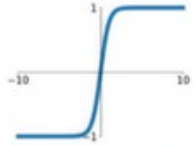
Leaky ReLU

$\max(0.1x, x)$



tanh

$\tanh(x)$

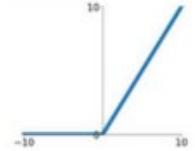


Maxout

$\max(w_1^T x + b_1, w_2^T x + b_2)$

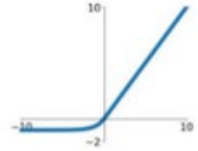
ReLU

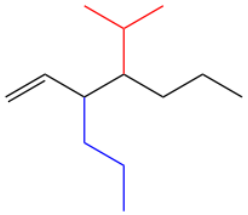
$\max(0, x)$



ELU

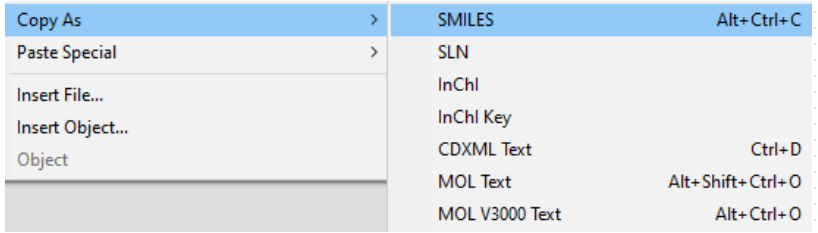
$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$

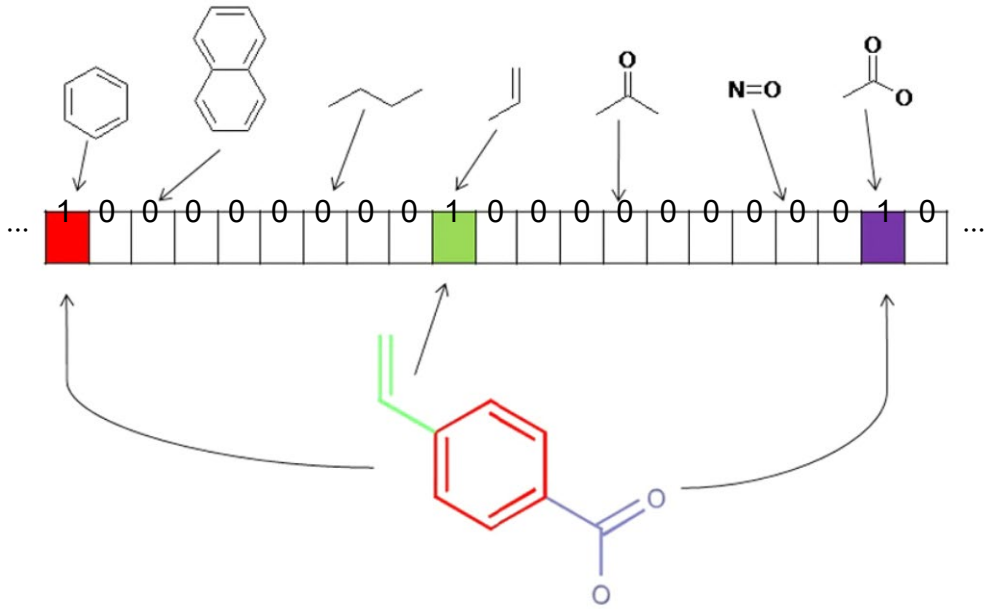


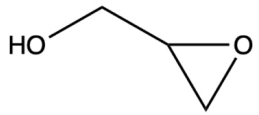


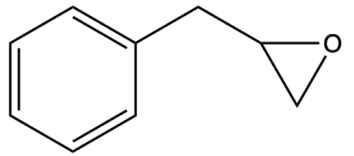
→ C=CC(CCC)C(C(C)C)CCC

How easiest to obtain SMILES?
Draw a structure in ChemDraw:







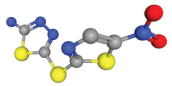


SMILES	OCC1CO1	c1ccc(CC2CO2)cc1
SMARTS	[#6]1-[#6](-[#8]-1)-[#6]-[#8]	[#6]1:[#6]:[#6]:[#6](-[#6]-[#6H]2-[#6]-[#8]-2):[#6]:[#6]:1
DeepSMILES	OCCCO3	ccccCCCO3))))cc6
SELFIES	[O][C][C][C][O][Ring1][Ring1]	[C][=C][C][=C][Branch1][#Branch1][C][C][C][O][Ring1][Ring1][C][=C][Ring1][#Branch2]

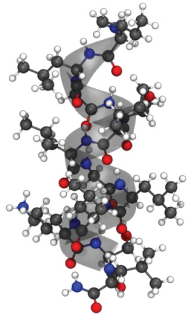
ML in Chem

Experimental information → **Representation formats** → **Machine learning models** → **Predictions and generations**

Structural data



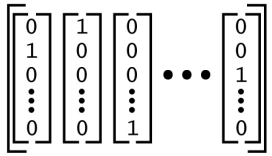
Dynamical data



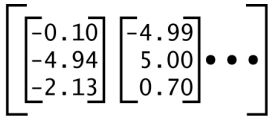
Physicochemical properties

- Charge
- Hydrophobicity
- Polarizability
- Atomic properties

One-hot encoding



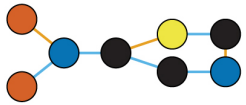
Feature vector



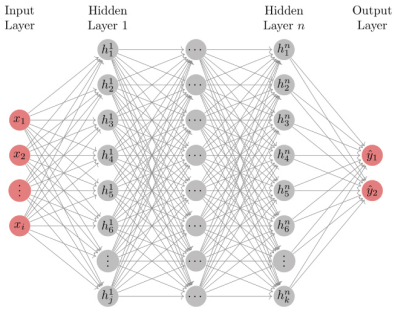
SMILES string

C1=C(SC(=N1)SC2....

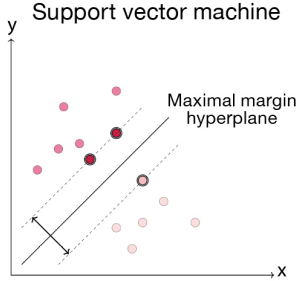
Graph



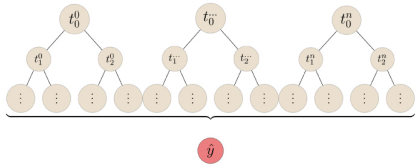
Neural networks



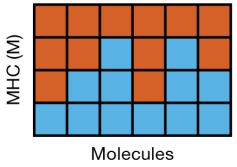
Support vector machine



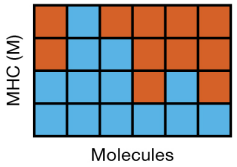
Random forest



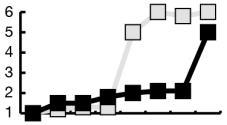
Antimicrobial activity



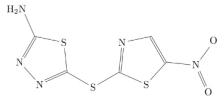
Hemolytic activity



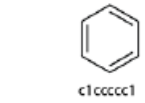
Development of resistance



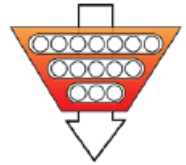
Novel molecules



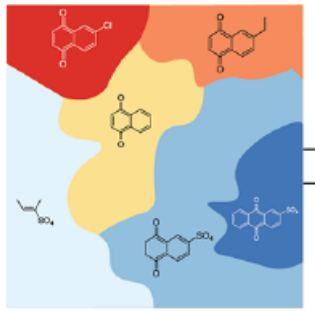
SMILES input

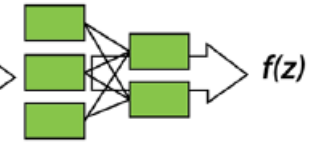


ENCODER Neural Network



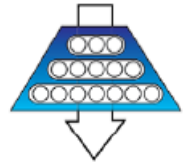
CONTINUOUS MOLECULAR REPRESENTATION (Latent Space)



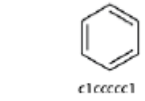


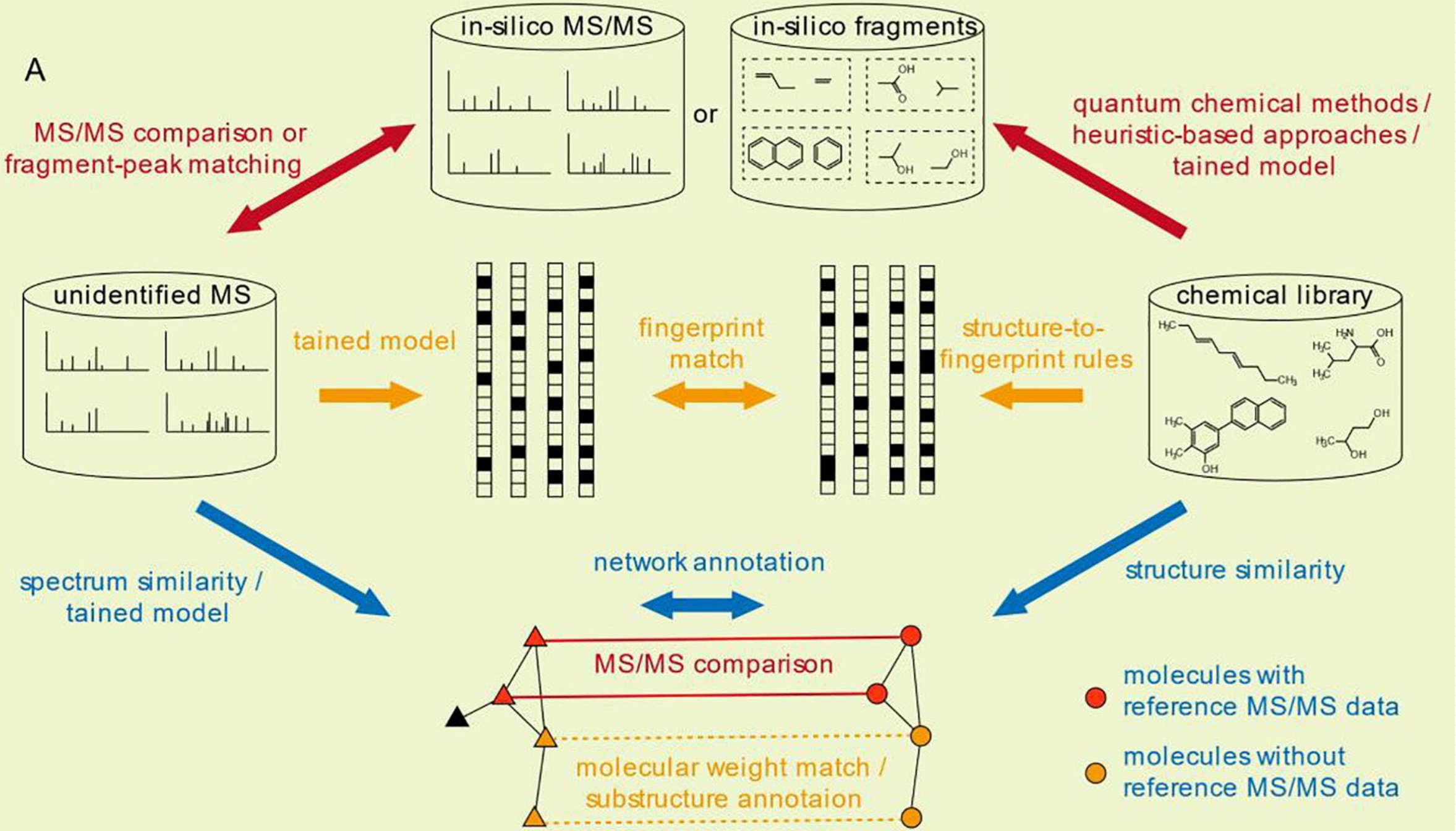
PROPERTY PREDICTION

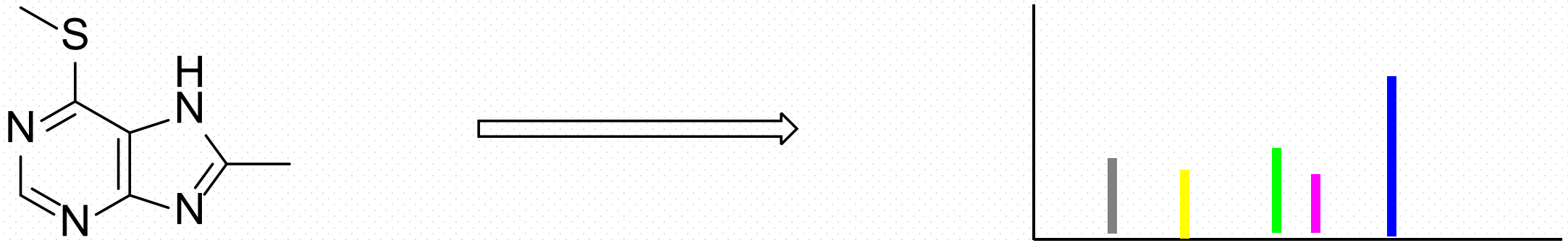
DECODER Neural Network



SMILES output




A
in-silico MS/MS
or
in-silico fragments
MS/MS comparison or
fragment-peak matching
quantum chemical methods /
heuristic-based approaches /
tained model
unidentified MS
tained model
fingerprint
match
structure-to-
fingerprint rules
chemical library
spectrum similarity /
tained model
network annotation
structure similarity
MS/MS comparison
molecular weight match /
substructure annotaion
molecules with
reference MS/MS data
molecules without
reference MS/MS data


S
H
N
N
N
N

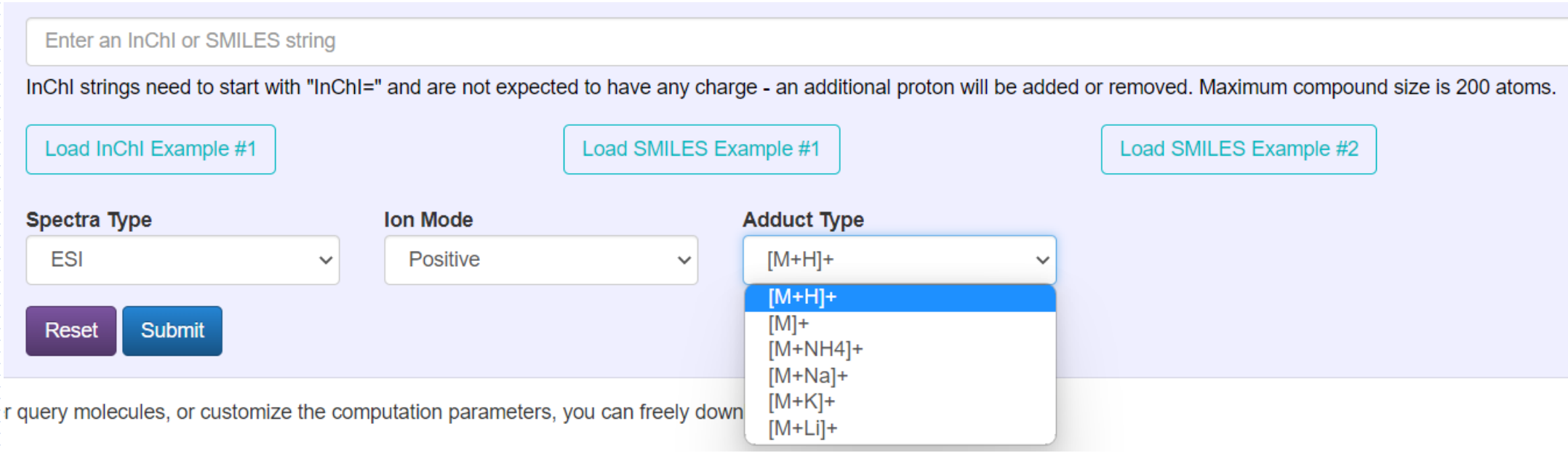
CFM-ID

Competitive Fragmentation Modeling for Metabolite Identification

Machine Learning + Heuristics (SID, CID)

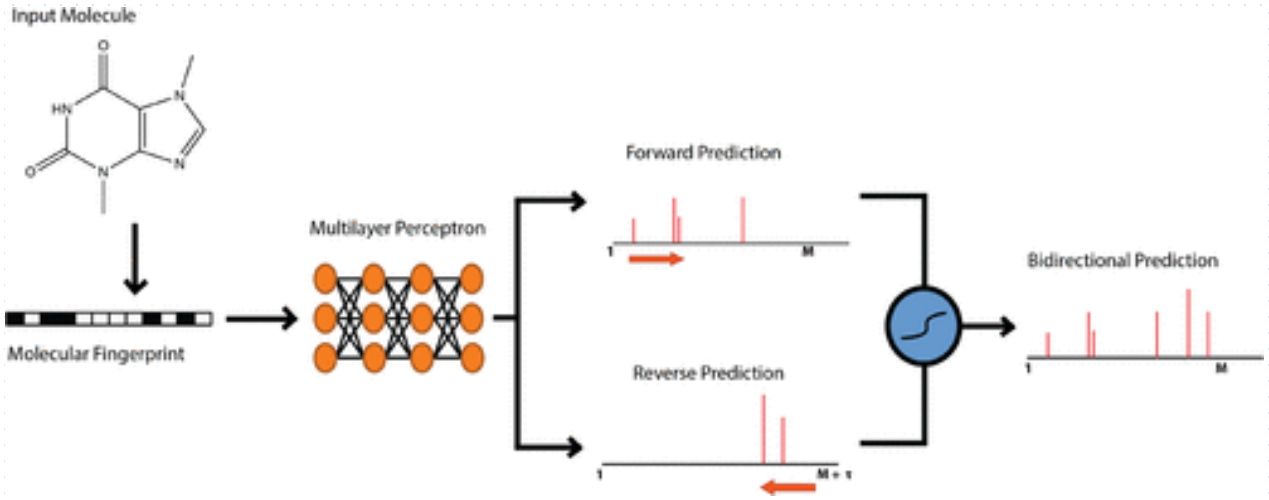
Break atom pairs, ion and neutral loss root path, Gasteiger charges, H movement, ring features; model predicts break tendencies which are mapped to probabilities using a softmax function

Expert curated rules for 21 lipid classes in version 3.0.



Fei Wang, Dana Allen, Siyang Tian, Eponine Oler, Vasuk Gautam, Russell Greiner, Thomas O Metz, David S Wishart
Nucleic Acids Research, Volume 50, Issue W1, 5 July 2022, Pages W165–W174, DOI:10.1093/nar/gkac383

Neural Electron–Ionization Mass Spectrometry (NEIMS)



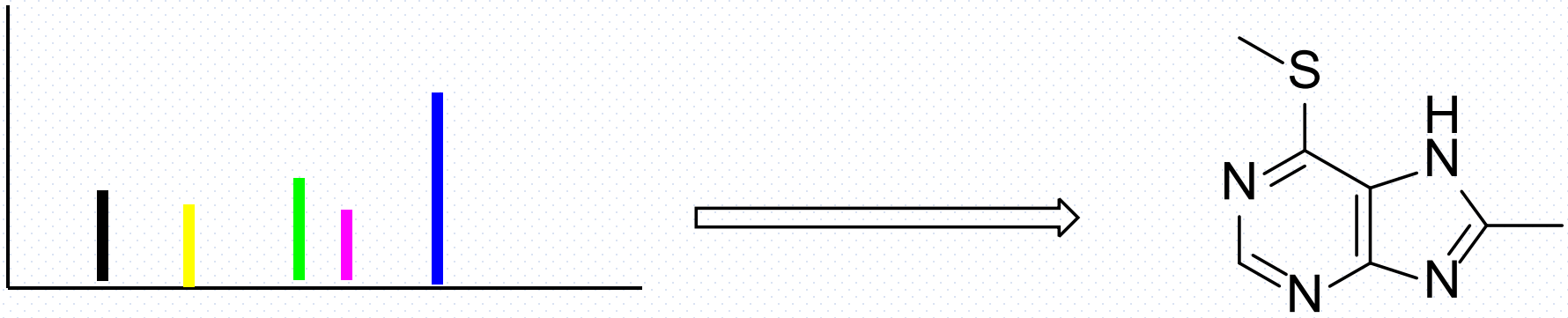
A bidirectional neural network for EI MS

forward model for 'local fragmentation'

reverse model for neutral losses

directly predicts spectra, very fast processing

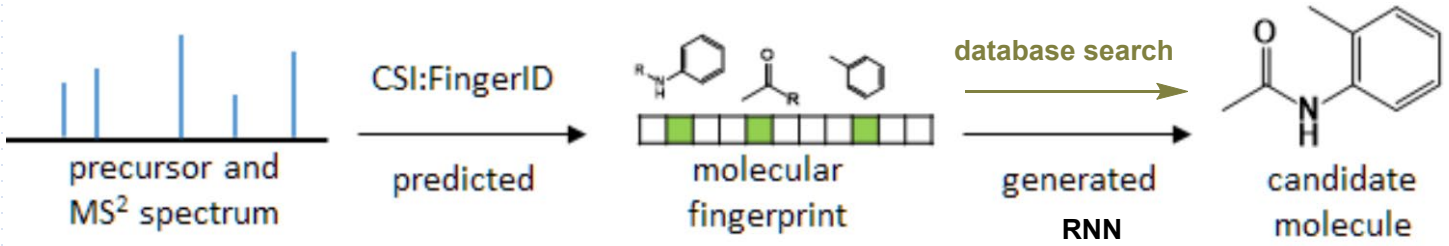
Jennifer N. Wei, David Belanger, Ryan P. Adams, and D. Sculley ACS Central Science 2019 5 (4), 700-708
DOI: 10.1021/acscentsci.9b00085

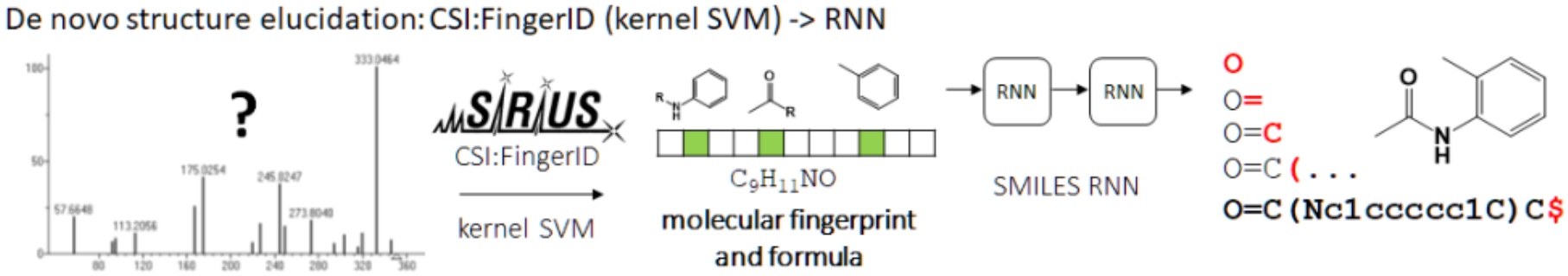

S
H
N
N
N
N

CSI:FingerID

Method the computes a fragmentation tree that best explains the unknown molecule fragmentation spectrum.

Fragmentation tree $\overset{ML}{\Rightarrow}$ molecular fingerprint





Training on a large dataset. Supports various adducts and provides an intuitive GUI (inside the SIRIUS suite).

Only CID. Reliability of fragmentation trees decreases rapidly for larger compounds with a M>500 Dalton

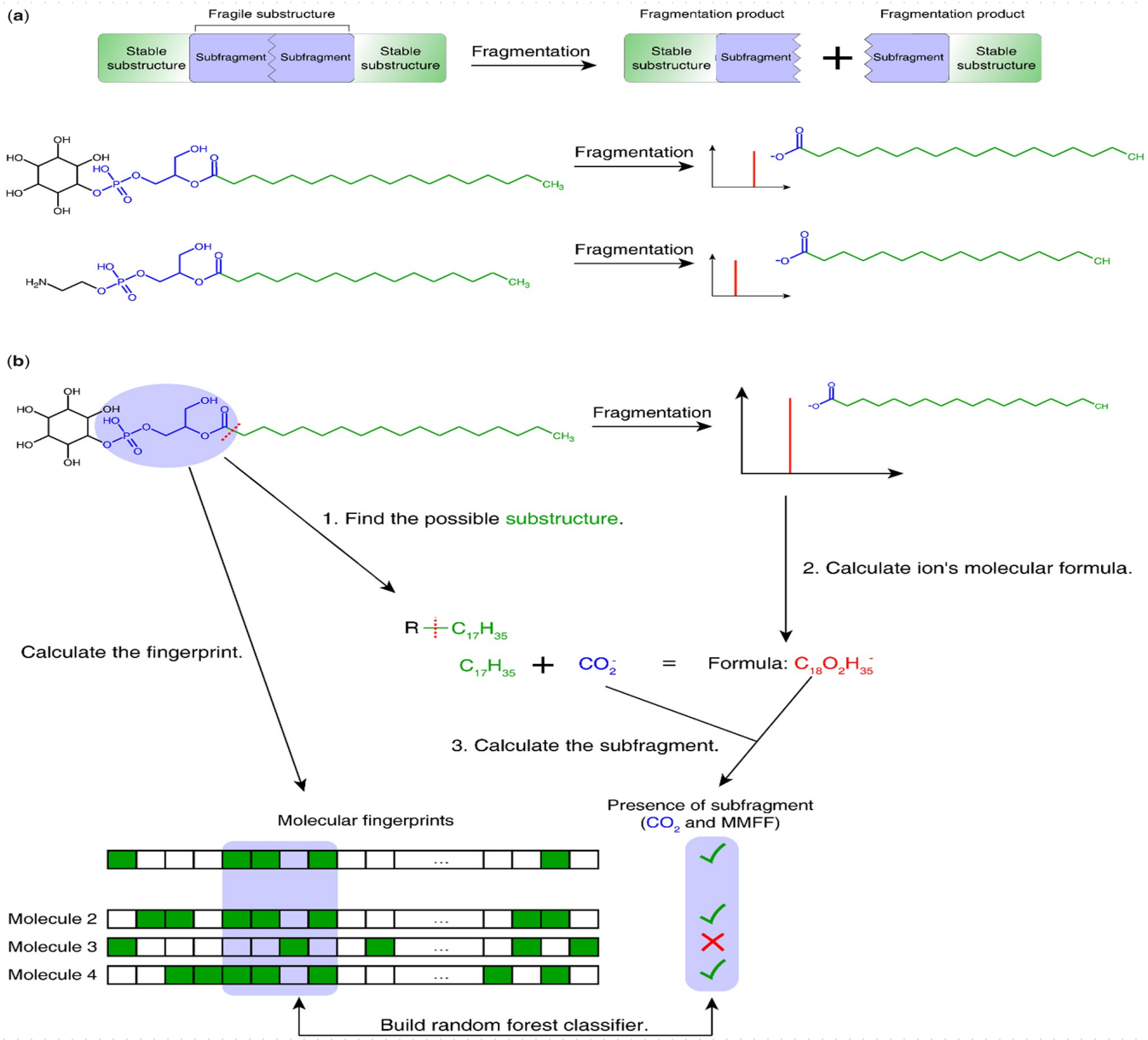
Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Proc. Natl Acad. Sci. USA 112, 12580–12585 (2015) DOI:10.1073/pnas.150978811

Stravs, M.A., Dührkop, K., Böcker, S. *et al.* *Nat Methods* **19**, 865–870 (2022). DOI:10.1038/s41592-022-01486-3

SF-Matching

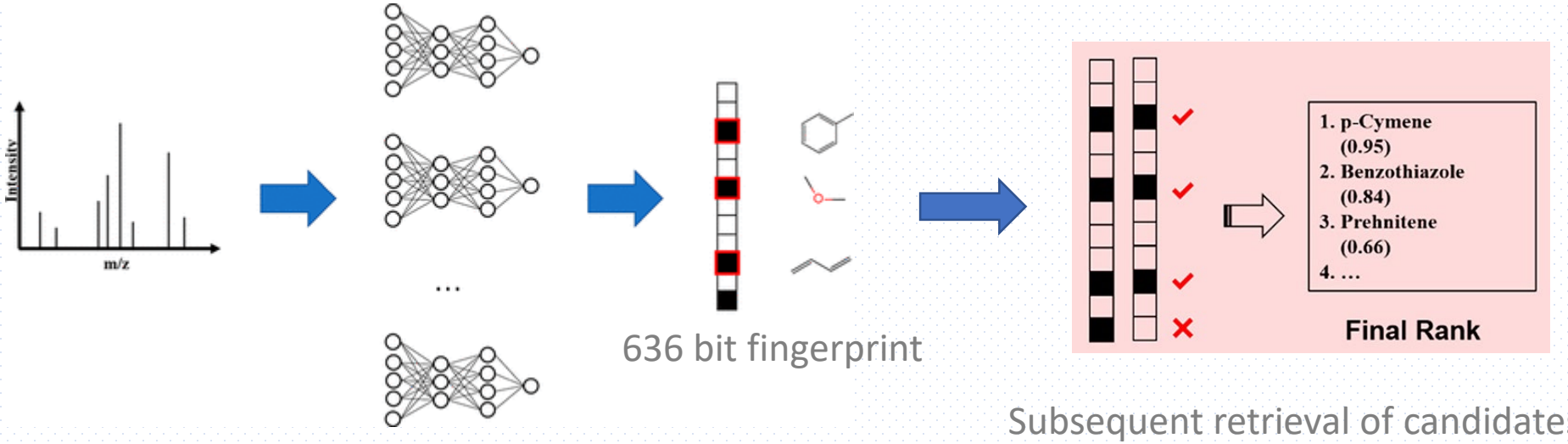
Idea: molecules with similar structural features will exhibit similar fragmentation patterns

Only CID. No adducts.
Python script, no GUI.



Li Y, Kuhn M, Gavin AC, *et al*. *Bioinformatics* 2020;**36**(4):1213–8, DOI:10.1093/bioinformatics/btz736

DeepEI



Subsequent retrieval of candidates from various databases ranked by Jaccard similarity scores

Lower prediction performance for NIST compounds

Ji H, Deng H, Lu H, *et al. Anal Chem* 2020;**92**(13):8649–53, DOI:10.1021/acs.analchem.0c01450

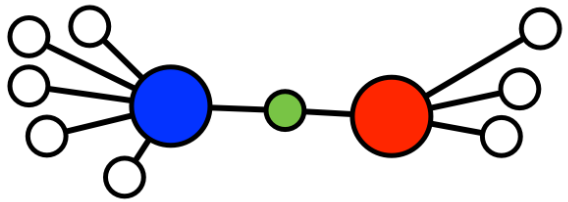
Unsupervised Machine Learning

molecular substructures from

MESSAR

MS2 substructure patterns

Full data analysis pipeline including data pre-processing, extracting of relevant substructures, nteractive exploration of the results and automatic annotations of fragments through MAGMa



MS2LDA

Unsupervised Substructure Discovery

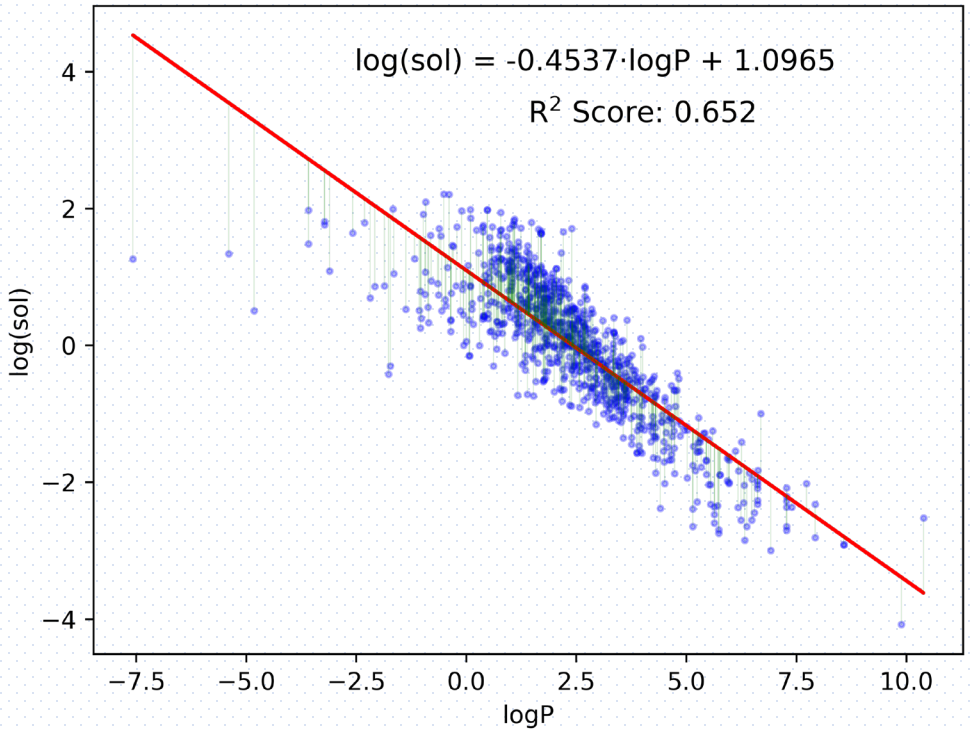
co-occurrence of mass fragments and neutral losses

Also available as web service. It can be used as a good structural hypothesis generator and to identify functional classes of unknown spectra that share substructures. Provides orthogonal information to MS2LDA

Additional applications of ML in HPLC/MS

RDKit
Open-Source Cheminformatic and Machine Learning

ESOL (water solubility) dataset, 1128 molecules



- **Compound Identification and Characterization**
- **Prediction of Retention Times**
- **Optimization of Mobile Phase**
- **Ion Mobility and DMS Modifier Analysis**
- **Ionization Type and Kinetic Energy Optimization**
- **Adduct Formation Analysis**
- **Prediction of Physicochemical Properties**
- **Quality Control and Batch Analysis**
- **Metabolomics and Biomarker Discovery**
- **Environmental Monitoring**
- **Food Safety Analysis**
- **Drug Discovery and Development**