# Machine learning on molecules and molecular fingerprints

Jakub Adamczyk, Piotr Ludynia
Faculty of Computer Science, AGH

# General plan

- **Part I: intro to drug discovery**
  - drug development & design process
  - RDKit
- **Part II: molecular property prediction**
  - molecular fingerprints & ML
  - scikit-fingerprints
- **Part III: virtual screening**
  - screening & searching
  - molecular filters

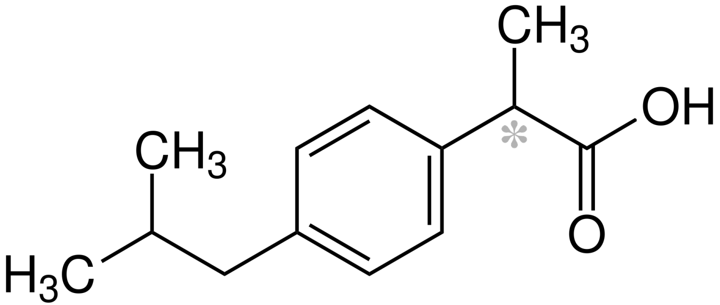
# Chemoinformatics

- **interdisciplinary science** between chemistry and informatics, with heavy influence of AI/ML:
    - chemical databases
    - molecular similarity searching
    - predicting properties of molecules
    - 3D simulations, generative models
- very similar to computational chemistry, often hard to tell the difference

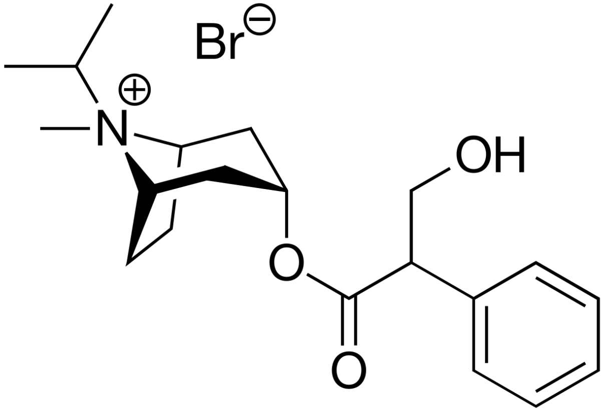
# Intro to drug discovery

# Drug

- typically a **small molecule**, i.e. <1000 daltons, <50 atoms (roughly)
- drug (a.k.a. **ligand**) typically **binds** to protein (often cell receptors), regulating its function
- **antagonists** dampen the effect, e.g. non-steroidal anti-inflammatory drugs
- **agonists** increase the effect, e.g. dilation of muscles in asthma treatment



Ibuprofen

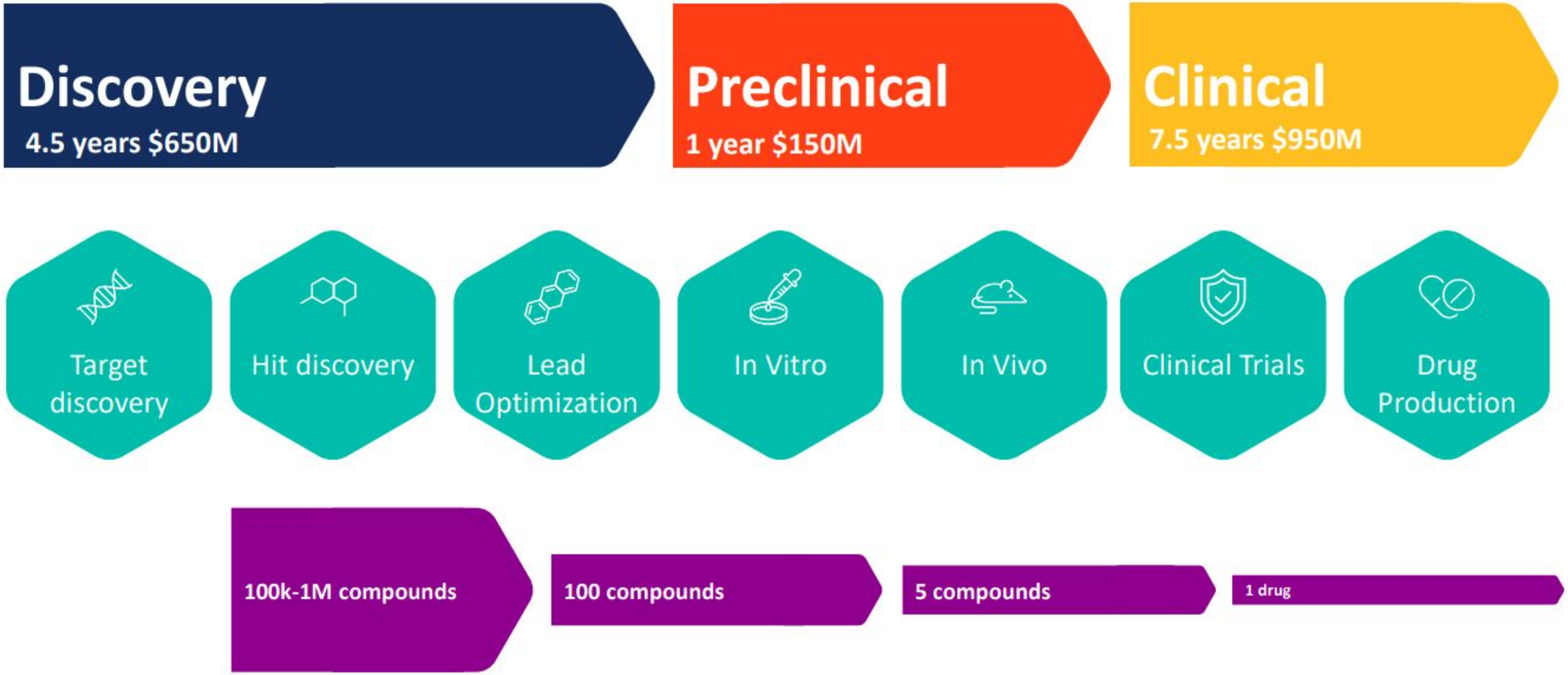


Ipratropium bromide

Sources: Wikipedia (Ibuprofen), Wikipedia (Ipratropium bromide)

# Drug discovery

- **multi-stage process** of discovery, optimization, and testing new drug



Discovery
4.5 years $650M

Preclinical
1 year $150M

Clinical
7.5 years $950M

Target discovery

Hit discovery

Lead Optimization

In Vitro

In Vivo

Clinical Trials

Drug Production

100k-1M compounds

100 compounds

5 compounds

1 drug

Source: "Enhancing Drug Discovery with Machine Learning: ADMET Property Modeling" M. Kowiel, GHOST Day 2024

# Drug discovery is expensive

- **on average:**
  - total cost >1 billion $
  - 1 out of 5000 drugs from preclinical get approved
  - ~10% of clinically tested drugs get approved
- **goal of ML** in drug discovery is to make it:
  - faster
  - cheaper
  - safer
- applied generally during **discovery** phase, to optimize what is tested on animals & humans

# Target discovery

- **goal:** identify specific molecular target to apply drugs to
- **questions:**
  - what biological mechanism causes the disease?
  - how can we regulate it?
- **input:** medical & biological knowledge
- **result:** DNA fragment, protein etc. and desired effect (e.g. we need an antagonist)
- **AI/ML tools:**
  - protein folding, e.g. AlphaFold
  - protein function prediction, e.g. DeepFRI
  - protein-protein interaction (PPI) prediction

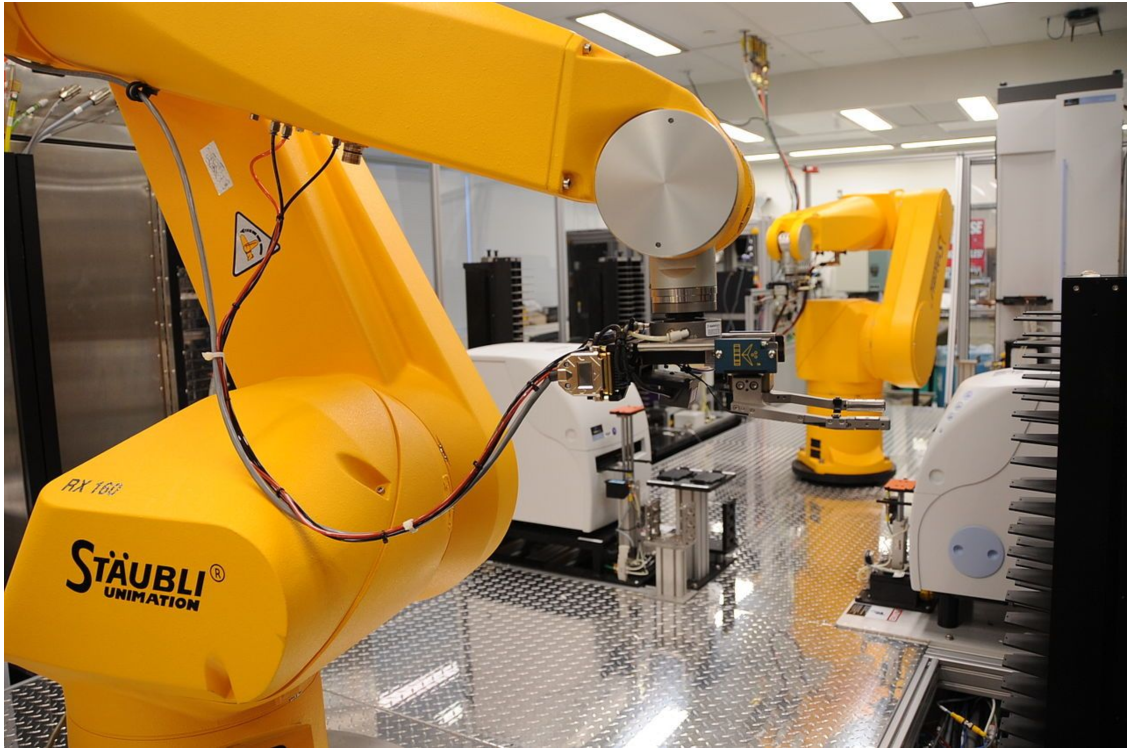
# Hit discovery (lead generation)

Hit discovery

- **goal:** identify promising compounds for more detailed testing
- **questions:**
    - which groups of molecules we should consider?
    - which can we discard?
- **input:** huge libraries (~100k-billions) of potential compounds for a given target
- **result:** small subset (~1-10k) of promising molecules, called **hits** or **leads**
- **AI/ML tools:**
    - virtual screening, e.g. molecular filters, protein-ligand docking
    - similarity searching

# High Throughput Screening (HTS)

- robotic wet labs, highly automated
- testing of compounds' properties on massive scale
- **pros:** fast, quite cheap
- **cons:** not very accurate



Source: Wikipedia (High-throughput screening)

# Lead optimization

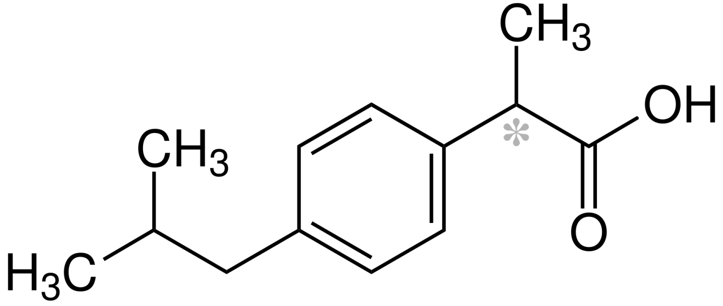
Lead Optimization

- **goal:** detect, refine and select actual drug candidates
- **questions:**
  - what are properties of molecules, e.g. toxicity, solubility?
  - how can we change their structure to have better properties?
- **input:** small subset (~1-10k) of promising molecules (hits/leads)
- **result:** ~tens-hundreds of molecules for actual wet lab testing
- **AI/ML tools:**
  - molecular property prediction, e.g. ADMET models
  - molecular generative models, e.g. genetic algorithms, diffusion models

# Chemistry recap
# & molecule processing

# Representing molecules

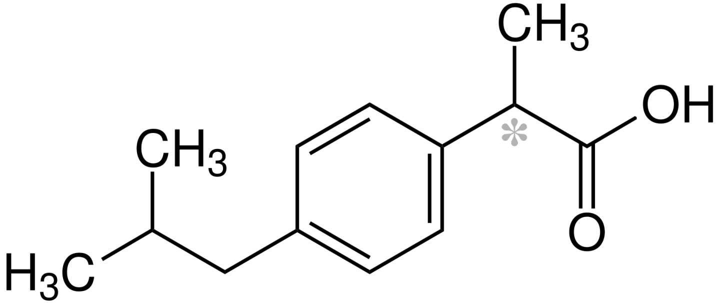
- **molecular graph** - typically processed in ML, attributed graph
- **SMILES (Simplified Molecular Input Line Entry System):**
  - string, can be transformed into graph
  - used for storage and sending data
  - not unique, loses some information, only 2D structure - but quite convenient
- other formats: SMARTS (molecular regex), SELFIES (generative models), SMIRKS (reactions)

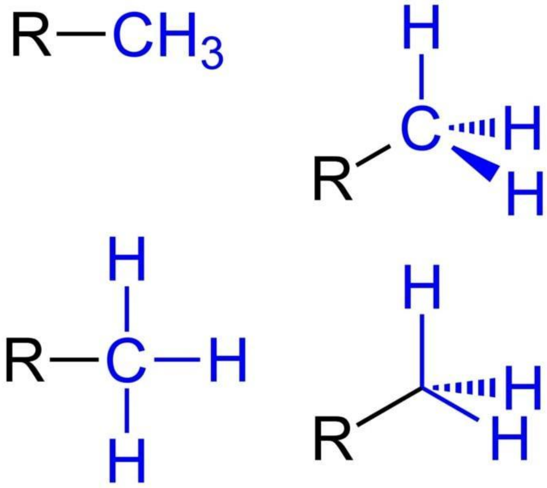


```
CC(C)Cc1ccc([C@@H](C)C(=O)O)cc1
```

# Chemistry recap

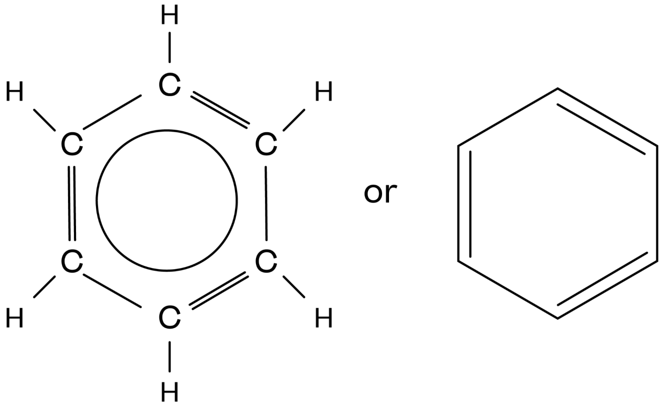
- drawings like those are called **skeletal formulas**
- **elements:**
  - carbon - "empty" atom
  - hydrogens typically **implicit**
  - other elements drawn explicitly
  - common **functional groups** often written as text
- **bonds:**
  - single, double, triple
  - solid wedge - up, towards the reader
  - dashed wedge - down, from the reader





# Chemistry recap

- **rings:**
    - cycles of atoms
    - often made of carbons
- **aromaticity:**
    - some rings are **aromatic** - surprisingly strong and stable
    - special atom/bond type
    - multiple atoms in a ring share electrons together, binding them strongly
    - also called **Kekule structures**, and computing this is kekulization



Benzene

# Chemoinformatics software

- **shockingly** for computer scientists, in chemistry a lot of software is:
  - proprietary, closed-source
  - horribly expensive
  - GUI-only, or with custom micro "programming languages"
  - quite old, e.g. Pipeline Pilot 1999, MOE 1994, ChemAxon Marvin 1999
- a de facto open source standard is **RDKit:**
  - created by Rational Discovery, open sourced when it shut down
  - creator & main maintainer: Gregory Landrum, ETH Zurich
  - written in C++, has Python wrapper (and others)

# Coding time 1

- https://github.com/j-adamczyk/molecular_ml_workshops
- notebook 1, RDKit intro

# Further reading

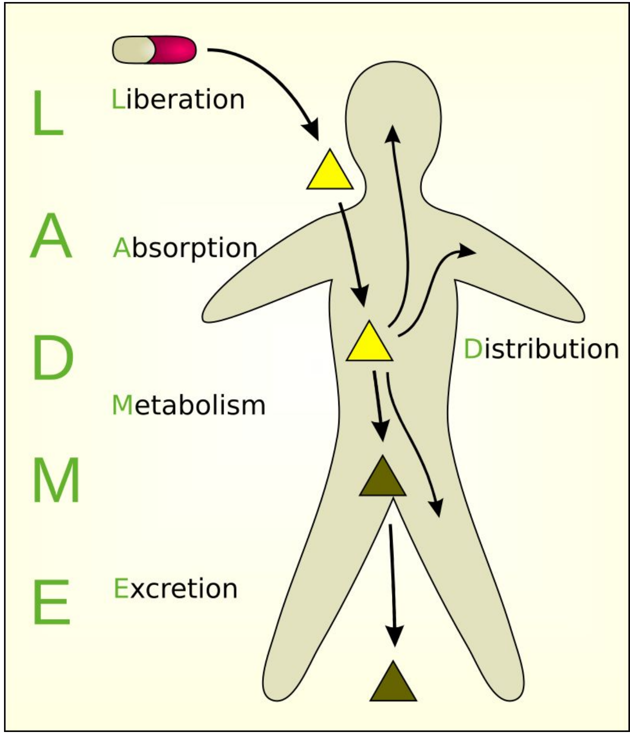
- "Enhancing Drug Discovery with Machine Learning: ADMET Property Modeling" M. Kowiel
- "Drug discovery and development: introduction to the general public and patient groups" N. Singh et al.
- "Artificial intelligence in drug discovery and development" D. Paul et al.
- "Artificial intelligence in drug discovery: applications and techniques" J. Deng et al.

- "Scikit-fingerprints: easy and efficient computation of molecular fingerprints in Python" J. Adamczyk, P. Ludynia

# Molecular property prediction

# Molecular property prediction

- **goal:** predict certain molecular properties
- **examples:**
  - ADMET: absorption, distribution, metabolism, excretion
  - toxicity
  - physicochemical properties, e.g. solubility, boiling point
  - biological activity, e.g. active/non-active against HIV
- **ML perspective:**
  - typical input: attributed graph
  - classification or regression
  - often: small data, imbalanced, multioutput (multitask)



By Scivit - Own work, CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=89573357

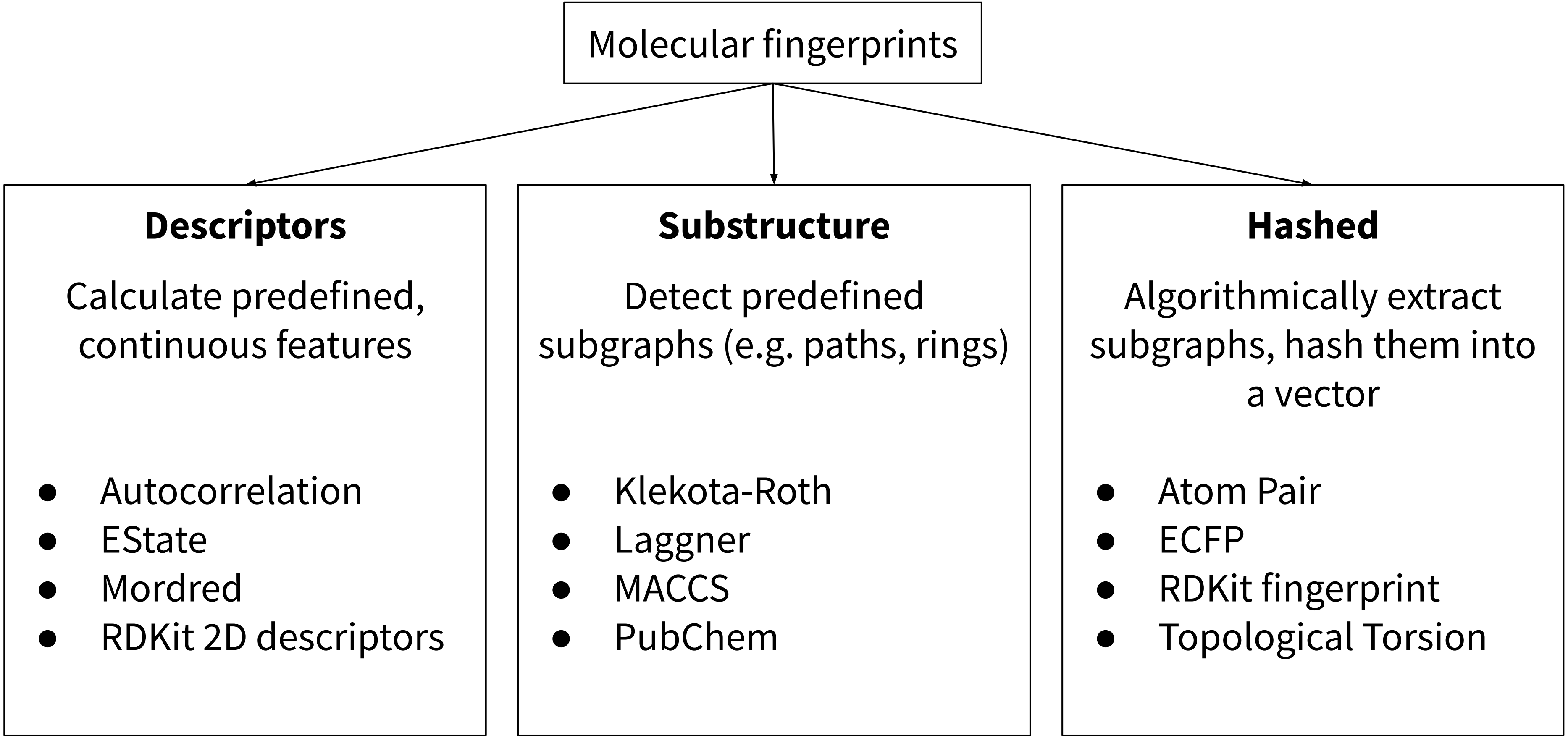
# Molecules as graphs

- molecules are **graphs:**
  - sets of unordered vertices (atoms) and edges (bonds)
  - **topology** (structure), i.e. what is connected
  - **functional** information, e.g. atom element types, bond orders
- they are **non-Euclidean** structures:
  - unordered, permutation-invariant (in contrast to e.g. text, images)
  - no natural distance or similarity between molecules
- so we need to extract features, turning graphs into **feature vectors**
- this turns graph classification into tabular classification

# Molecular fingerprints

- **molecular fingerprints:**
  - algorithms for automated feature extraction from molecules
  - permutation-invariant, i.e. we could reorder atoms and still get the same features
  - typically high-dimensional (e.g. 1024, 2048), sparse vectors
  - often binary, i.e. feature exists or not in a molecule
- not reversible and potentially surjective
- alternative to e.g. **graph neural networks (GNNs)** and **SMILES transformers**
  - they are interesting and powerful, but hard to train
  - fingerprints are much cheaper, faster, and often give better results

# Types of fingerprints



Molecular fingerprints

**Descriptors**

Calculate predefined, continuous features

- Autocorrelation
- EState
- Mordred
- RDKit 2D descriptors

**Substructure**

Detect predefined subgraphs (e.g. paths, rings)

- Klekota-Roth
- Laggner
- MACCS
- PubChem

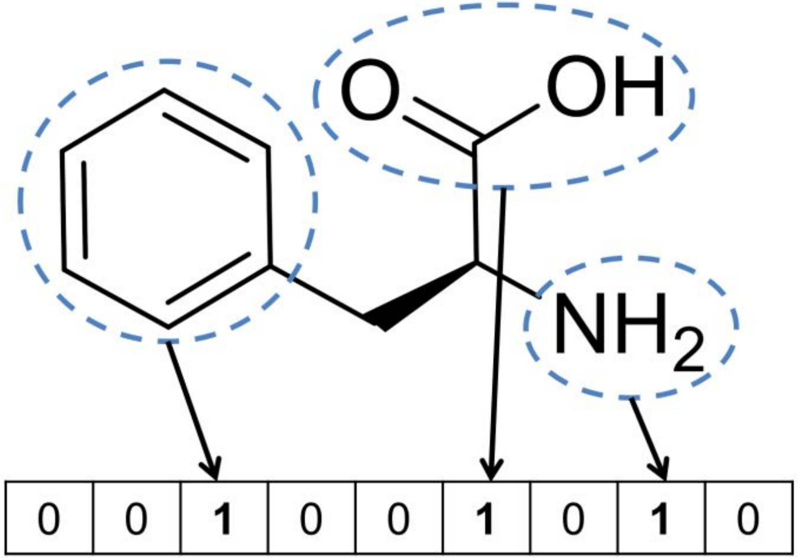
**Hashed**

Algorithmically extract subgraphs, hash them into a vector

- Atom Pair
- ECFP
- RDKit fingerprint
- Topological Torsion

# Substructure fingerprints

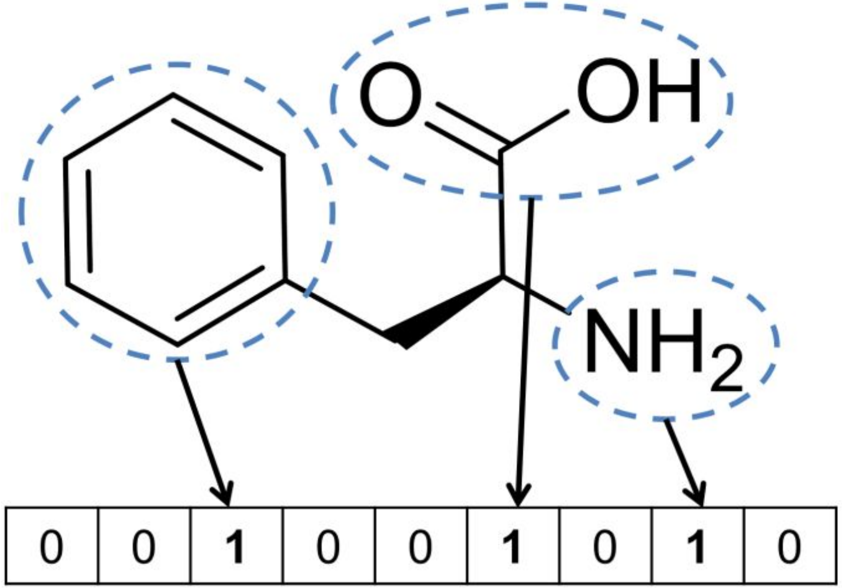
- **explicitly** define features (typically with SMARTS) to extract from molecule
- rings, atoms of given type, functional groups etc. - **substructures**
- typically **binary (bit)**, check if/else condition
- examples:
    - MACCS: 166 features
    - Klekota-Roth: 4860 features
    - PubChem: 881 features
- **pros:** great for similarity searching, interpretability
- **cons:** can be slow, bad for new domains



Source: "Chemical Data Formats, Fingerprints, and Other Molecular Descriptions for Database Analysis and Searching" D. Bajusz et al.

# MACCS fingerprint

- **substructure** fingerprint
- 166 bits, features defined by expert chemists
- examples:
  - fewer than 3 oxygen atoms
  - -S-S- bond
  - a ring of size 4

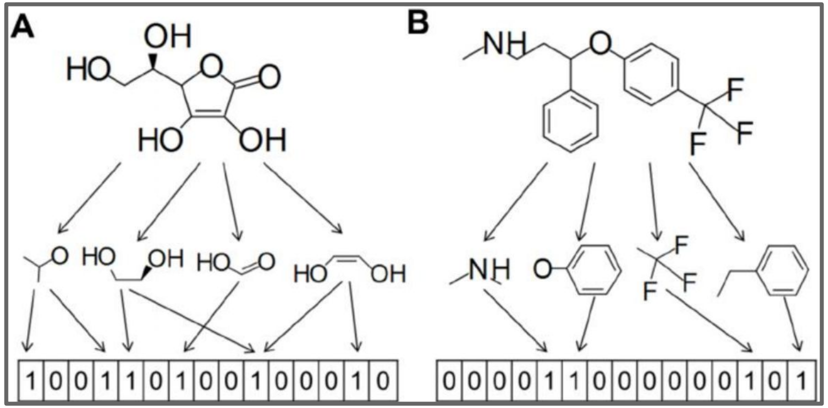


| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |

Source: "Chemical Data Formats, Fingerprints, and Other Molecular Descriptions for Database Analysis and Searching" D. Bajusz et al.

# Hashed fingerprints

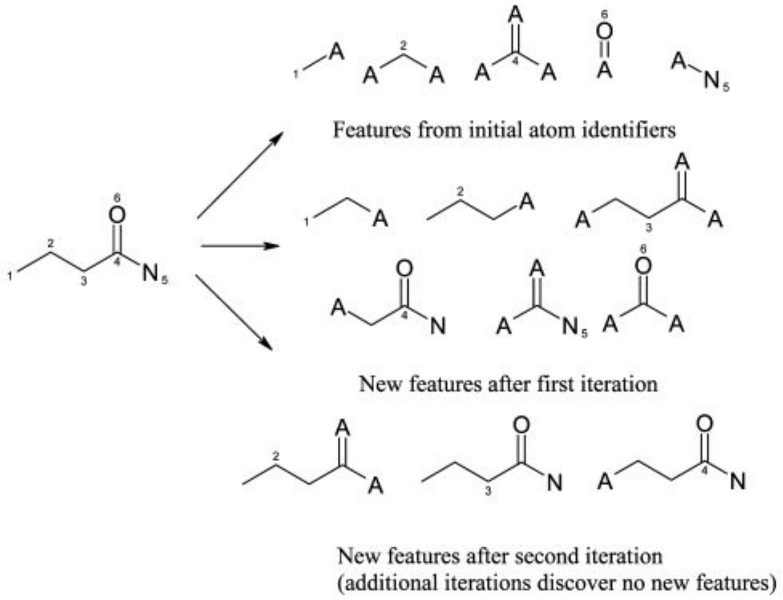
- **idea:** create a dictionary detecting subgraphs with a given shape
- all follow the same **steps:**
  - extract **subgraphs**, e.g. atom neighborhoods of given radius
  - each subgraph gets a unique ID
  - create an all-zero vector, e.g. 2048 bits
  - **hash** IDs onto this vector, e.g. with modulo operation
- outputs are **not interpretable** - they are just some subgraphs



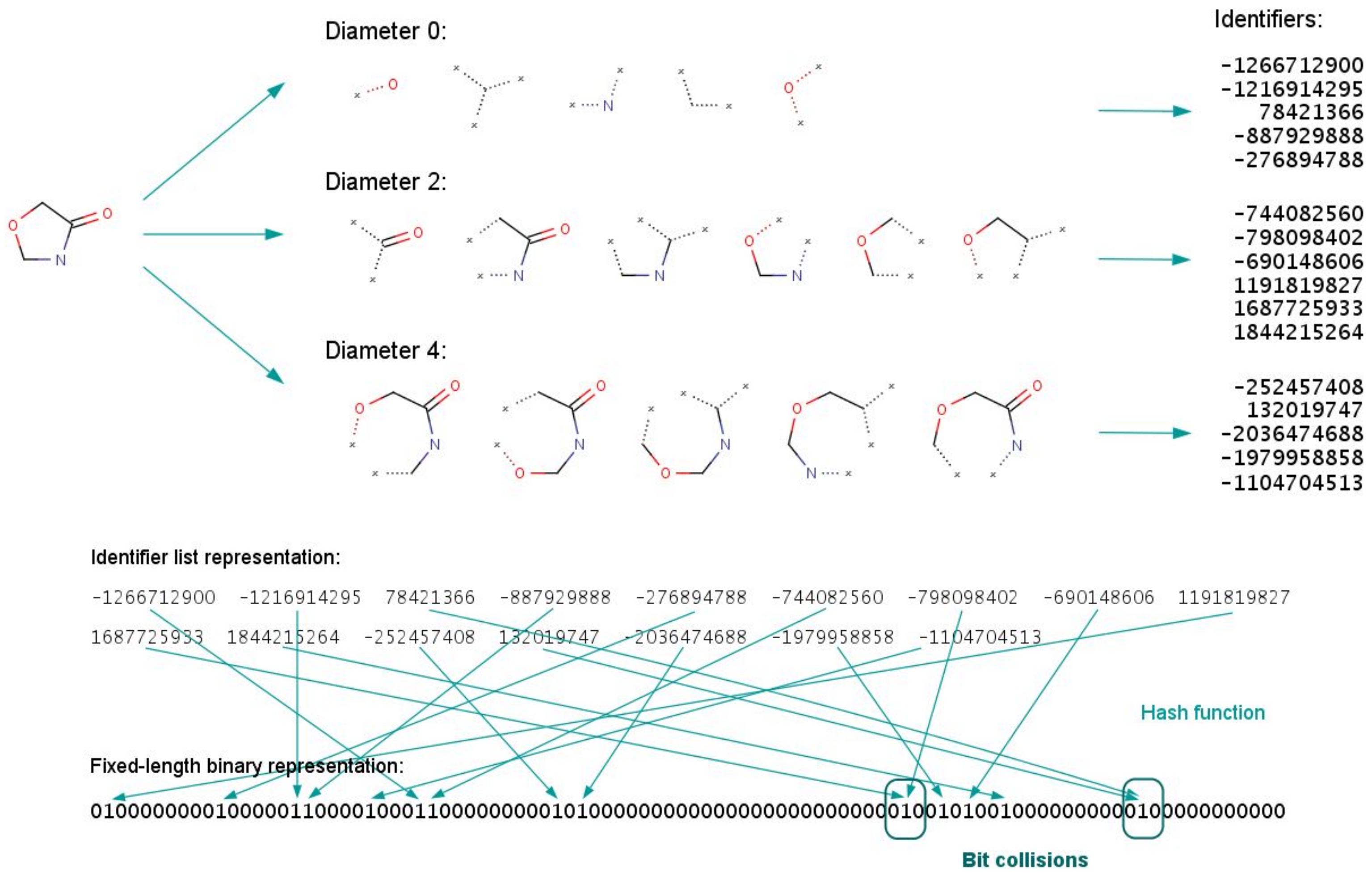
Source: "Average Information Content Maximization-A New Approach for Fingerprint Hybridization and Reduction" M. Śmieja, D. Warszycki

# Extended Connectivity Fingerprint (ECFP)

- ECFP is **the most popular** hashed fingerprint
- it extracts **circular** subgraphs with given radius:
    - atoms
    - atom + neighbors
    - atom + neighbors + their neighbors
    - …
- works **great** on average
- graph neural networks (GNNs) work in similar manner


Features from initial atom identifiers
New features after first iteration
New features after second iteration
(additional iterations discover no new features)

Source: "Extended-Connectivity Fingerprints" D. Rogers, M. Hahn



Diameter 0:

Diameter 2:

Diameter 4:

Identifiers:

-1266712900
-1216914295
78421366
-887929888
-276894788

-744082560
-798098402
-690148606
1191819827
1687725933
1844215264

-252457408
132019747
-2036474688
-1979958858
-1104704513

Identifier list representation:

-1266712900 -1216914295 78421366 -887929888 -276894788 -744082560 -798098402 -690148606 1191819827
1687725933 1844215264 -252457408 132019747 -2036474688 -1979958858 -1104704513

Hash function

Fixed-length binary representation:

0100000000100000110000100011000000000101000000000000000000000000000100101001000000000001000000000000

Bit collisions

Source: [https://docs.chemaxon.com/display/docs/fingerprints_extended-connectivity-fingerprint-ecfp.md](https://docs.chemaxon.com/display/docs/fingerprints_extended-connectivity-fingerprint-ecfp.md)

# Binary vs count fingerprints

- instead of **detecting** features, we can also **count** them
- most fingerprints have both variants
- **binary/bit:**
  - detect if a given feature or hash index appears at all
  - better for **molecular similarity search**
  - efficient boolean vectors processing
- **count:**
  - store the number of feature occurrences
  - preserve more information
  - often better for molecular property prediction

# Coding time 2

- https://github.com/j-adamczyk/molecular_ml_workshops
- notebook 2, molecular property prediction

# Further reading

- "Extended-Connectivity Fingerprints" D. Rogers, M. Hahn
- "Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models" D. Jiang et al.
- "Comparative analysis of molecular fingerprints in prediction of drug combination effects" B. Zagidullin et al.
- "A Python library for efficient computation of molecular fingerprints" M. Szafarczyk et al.
- "Scikit-fingerprints: easy and efficient computation of molecular fingerprints in Python" J. Adamczyk, P. Ludynia

# Virtual screening

# Screening

- **screening:**
  - identify potentially active compounds to synthesize and test in a lab
  - from a huge initial data, e.g. hundreds of thousands
- **high-throughput screening (HTS)** typically provides initial, imprecise labels
- we can also use literature and previous publicly available results
- **virtual screening (VS)** is a computational approach, where we use ML to select candidates, often from vast amounts of data
- molecules are often **filtered** first to remove "bad" molecules, e.g. toxic, reactive, typical false positives

# Virtual screening

- **main VS types:**
  - ligand-based, where we use 2D molecule graph and activity information
  - structure-based, where we use 3D ligand-protein docking simulations
- **ligand-based** is large-scale, cheap, and is basically binary classification (active/non-active)
- binary classification active/non-active
- **extremely imbalanced** problem, ~0.01%-1% actives (positive class)
- testing is expensive, so we select only top ranked molecules
- model must quickly find the few good candidates - **early enrichment**

# Molecular filters

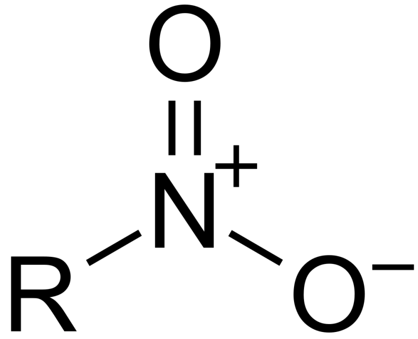
- **idea:** remove clearly unwanted molecules, e.g.:
  - too high mass = probably won't go through membranes = reject
  - has nitro group = is probably mutagenic / genotoxic = reject
- **molecular filters** have two main types:
  - property filters, which check if properties (e.g. mass, logP) are in reasonable range
  - substructure filters, which remove compounds with problematic substructures
- designed for **given purpose**, e.g. druglike, leadlike, orally available, pesticides
- however, they **limit the chemical space** that we consider, so we can e.g. allow 1 violation of filter rules

# Property filter - Lipinski Rule of 5

- **Lipinski Rule of 5 (Ro5)** - oldest molecular filter, made by Pfizer
- Lipinski observed that most of their approved drugs had basic properties in reasonable range
- **orally available** drug can have at most 1 violation of:
    - mass < 500 daltons
    - number of hydrogen bond donors (HBD) < 5
    - number of hydrogen bond acceptors (HBA) < 10
    - logP < 5
- this guarantees small, reasonably lipophilic molecules

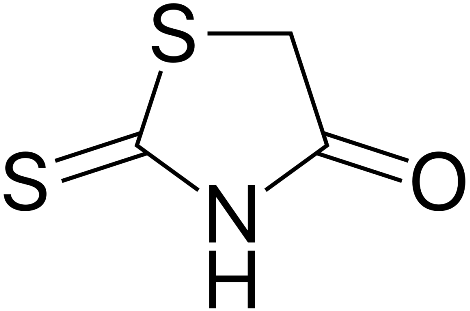
# Substructure filter - PAINS

- **Pan Assay Interference Compounds (PAINS)** - most commonly used set of problematic structures, defined with SMARTS patterns
- designed to filter out false positives - highly reactive, often toxic molecules, which show very commonly in HTS
- 3 different sets of substructures: A, B, C
- example substructures:



Nitro group:

toxicity



Rhodanine group:

poor selectivity

# Evaluating virtual screening

- few molecules can be lab-tested, so models must select the few potential actives accurately
- we have *n* actives among *N* molecules (*n* << *N*, e.g. *n*=30, *N*=15000)
- metrics should measure **early enrichment** - highest ranked molecules are most important
- **common metrics:** AUROC, EF, RIE, BEDROC
- **enrichment factor EF(*X*):**
    - defined for fraction *X* of dataset that we select, e.g. 1%, 5%
    - number of actives found, divided by random picking performance:

      EF(X) = (num actives in top *N* * *X*) / (*n* * *X*)

      N - dataset size, n - total actives
    - min value: 0, max value: 1/*X* if *X* >= *n*/*N*, and *N*/*n* otherwise

# Coding time 3

- https://github.com/j-adamczyk/molecular_ml_workshops
- notebook 3, virtual screening

# Further reading

- "The Light and Dark Sides of Virtual Screening: What Is There to Know?" A. Gimeno et al.
- "Molecular fingerprint similarity search in virtual screening" A. Cereto-Massagué et al.
- "An overview of molecular fingerprint similarity search in virtual screening" I. Muegge, P. Mukherjee
- "Performance of machine-learning scoring functions in structure-based virtual screening" M. Wójcikowski et al.
- "Open-source platform to benchmark fingerprints for ligand-based virtual screening" S. Riniker, G. Landrum
- "Heterogeneous Classifier Fusion for Ligand-Based Virtual Screening: Or, How Decision Making by Committee Can Be a Good Thing" S. Riniker et al.