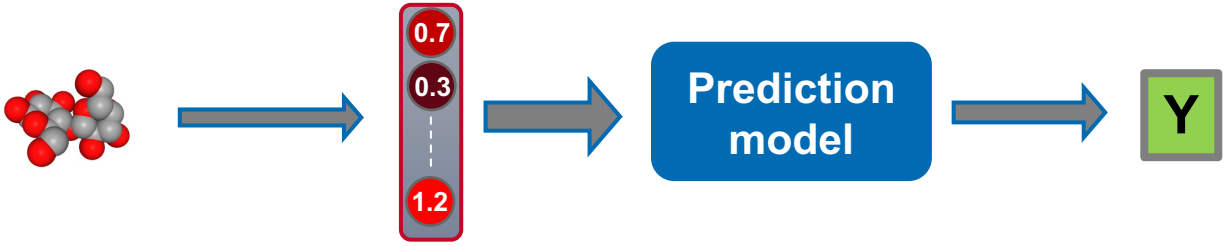
Heinrich Heine
Universität Düsseldorf

Traditional numerical representations for small molecules

Motivation



Molecular Descriptors

- Molecular Descriptors
  - Examples for entries of molecular descriptors:
    - Number of atoms, number of bonds, molecular weight, number of rings, number of bonds, …
    - Hydrophobicity (e.g. LogP), geometric descriptors (surface area, shade descriptors, ...)

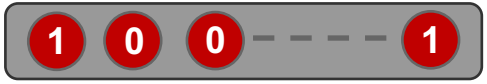
- Many features can be calculated using the Python package RDKit
  - RDKit is an open-source cheminformatics software library
  - Has a $Descriptors$ module with over 200 molecular descriptors

Landrum, Greg. "RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling."

Molecular (expert-designed) fingerprints

Heinrich Heine Universität Düsseldorf

- Molecular (expert-designed) fingerprints:
  - Binary vectors representing the presence or absence of certain substructures in the molecule
  - Examples:

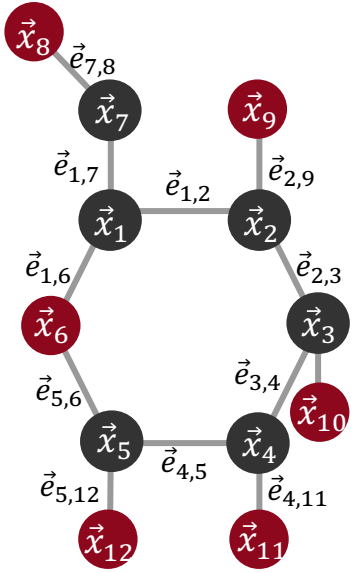


    - Graph-based:
      - Morgan Fingerprints (ECFPs)
      - RDKit Fingerprints
    - MACCS keys:
      - Each entry corresponds to a specific structural feature or characteristic of a molecule
        - Does the molecule contain a ring of size 4?
        - Are there fewer than 3 oxygen atoms present in the molecule?
      - 166-dimensional binary fingerprint
  - Can be calculated using the Python package RDKit

Durant JL, Leland BA, Henry DR, Nourse JG: Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci.*

Morgan Fingerprints (ECFPs)

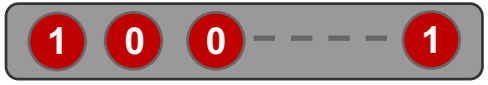
$$\vec{x}_i = \begin{pmatrix} atom\ type \\ mass \\ valence \\ charge \\ \dots \end{pmatrix} \qquad \vec{e}_{i,j} = \begin{pmatrix} bond\ type \\ part\ of\ ring \\ aromaticity \\ \dots \\ \dots \end{pmatrix}$$



- Molecules are represented as graphs:
  - Atoms are interpreted as the nodes of a graph
  - Bonds are interpreted as the edges of a graph
- Calculating atom and bond feature vectors
- Updating each vector by using information about neighboring atoms and bonds
- Mapping all atom and bond representations from all iterations to a binary vector using hash functions

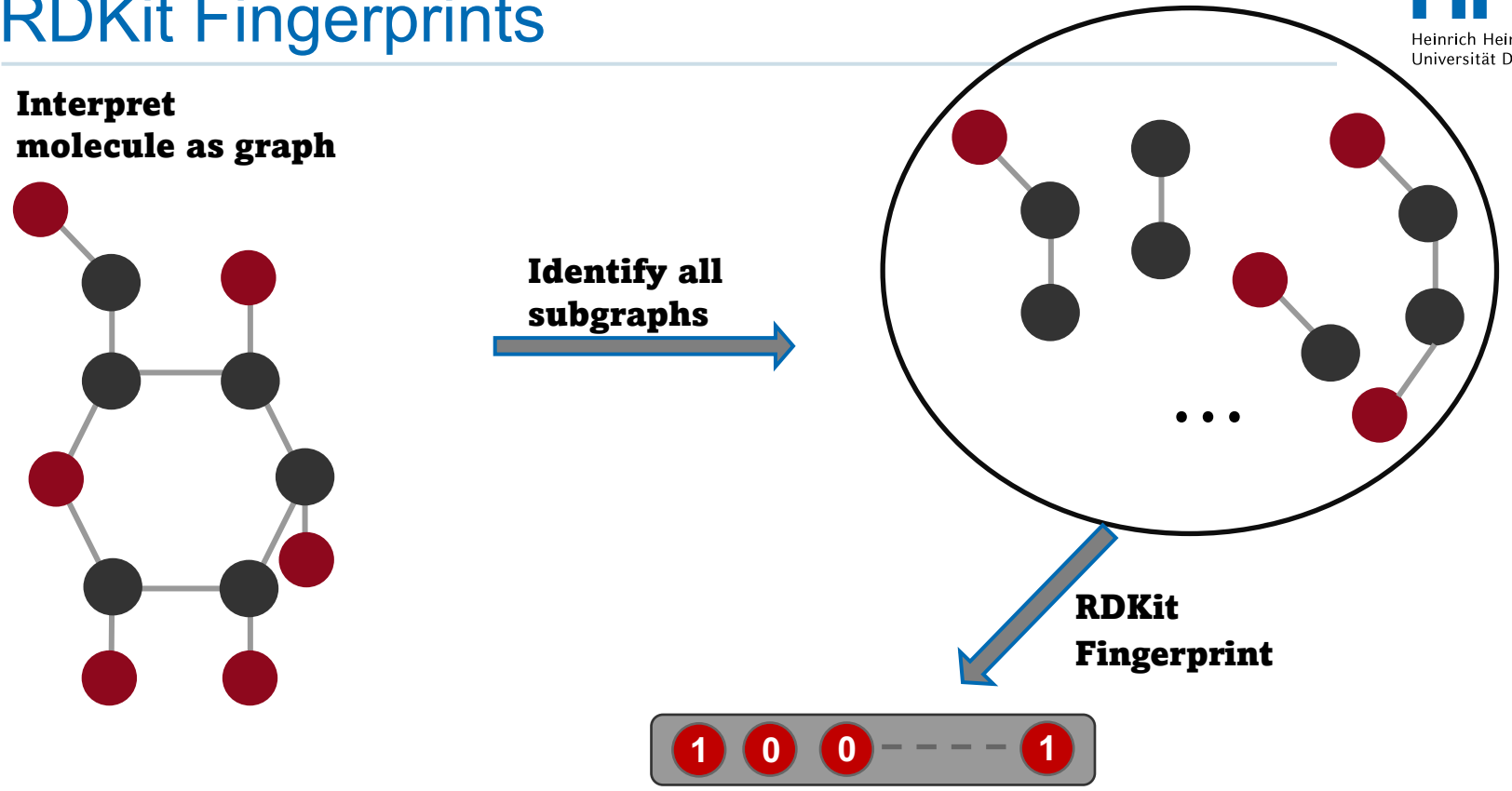
Morgan

Fingerprints



Rogers, David, and Mathew Hahn. "Extended-connectivity fingerprints." *Journal of chemical information and modeling*

RDKit Fingerprints



Landrum, Greg. "RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling."