

AI and ML for Chemists

Topic 4.1: Representations

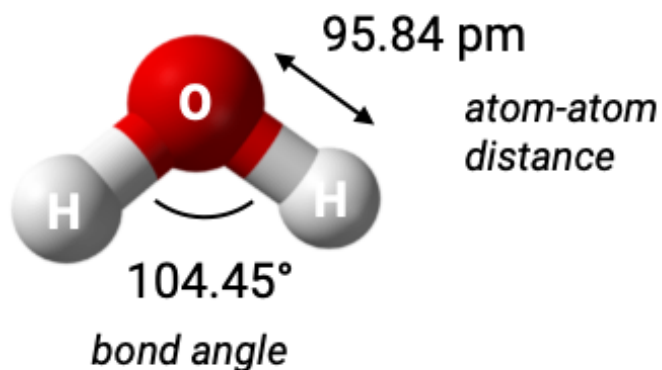
Pre-lecture Explore

Dr. Ganna (Anya) Gryn'ova

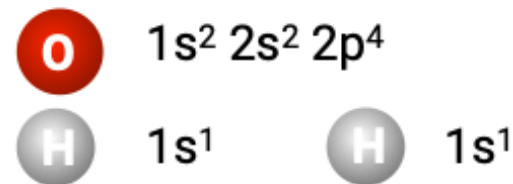
Representing chemistry

$$y = f(x)$$

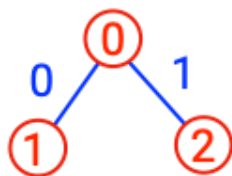
output prediction (label) input function (feature)



8 + 1 + 1 = 10 electrons

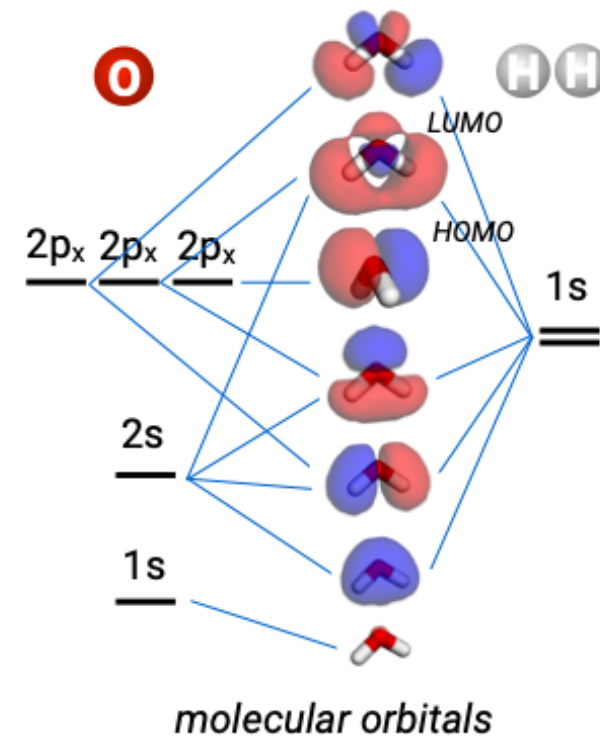
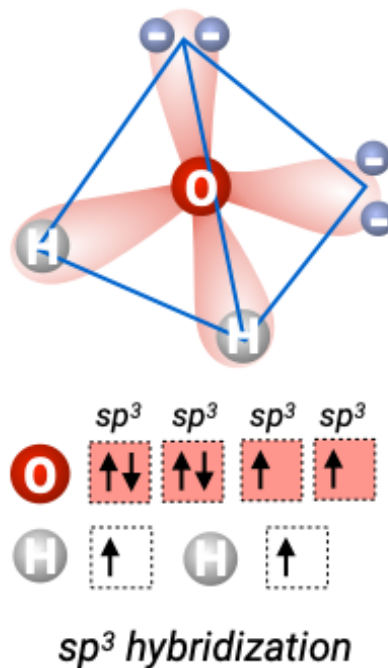
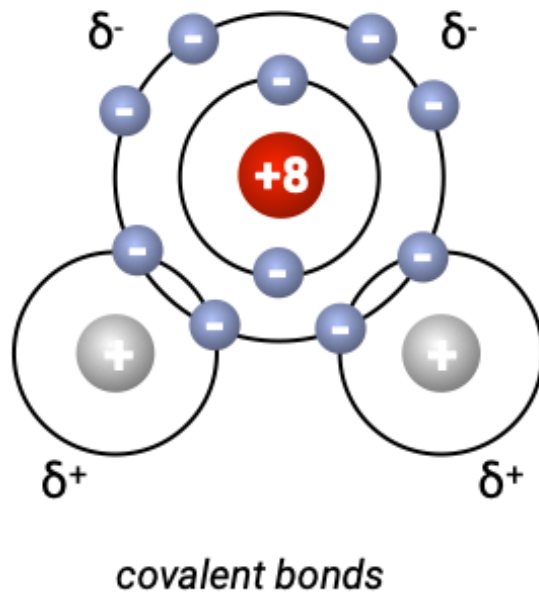


A molecular graph (RDKit)



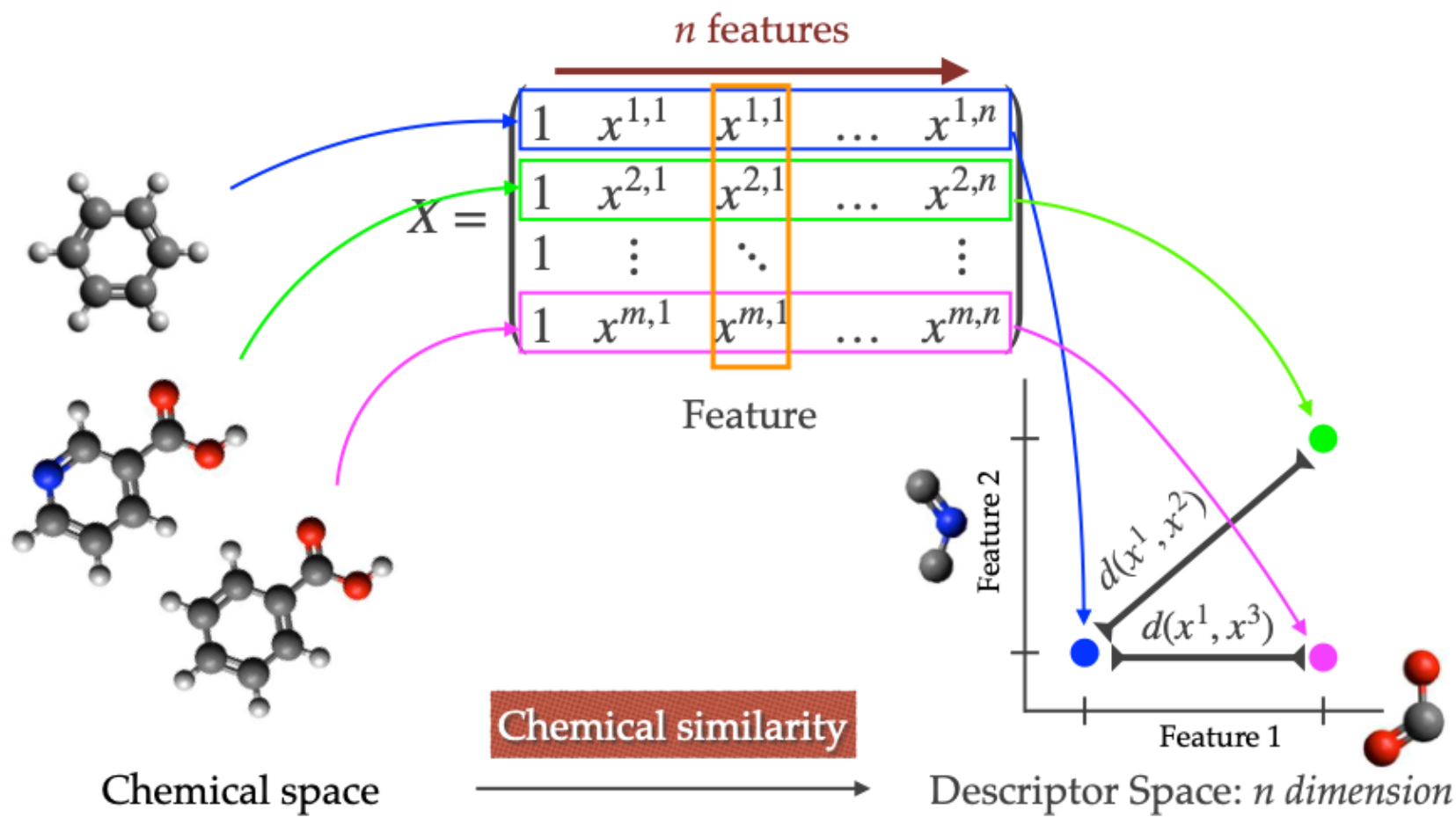
	0	1	2	
AtomicNum	8	1	1	Atom Invariants
TotalDegree	2	1	1	
TotalNumHs	0	0	0	
FormalCharge	0	0	0	
deltaMass	0	0	0	
IsInRing	0	0	0	

	0	1	
BondType	0	0	Bond Invariants
Stereo	0	0	



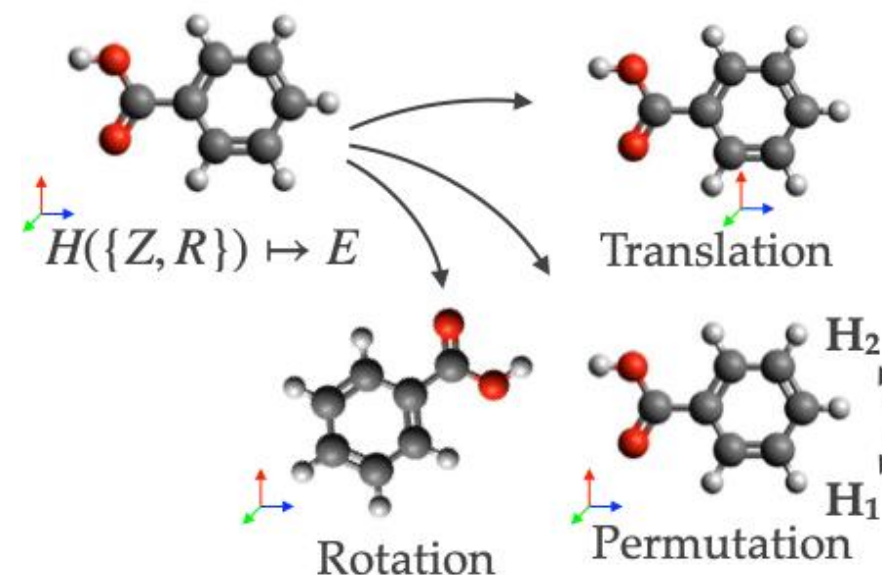
Features and descriptors

To learn, a machine requires a numerical **representation** of a molecule encoding the information used in subsequent ML.



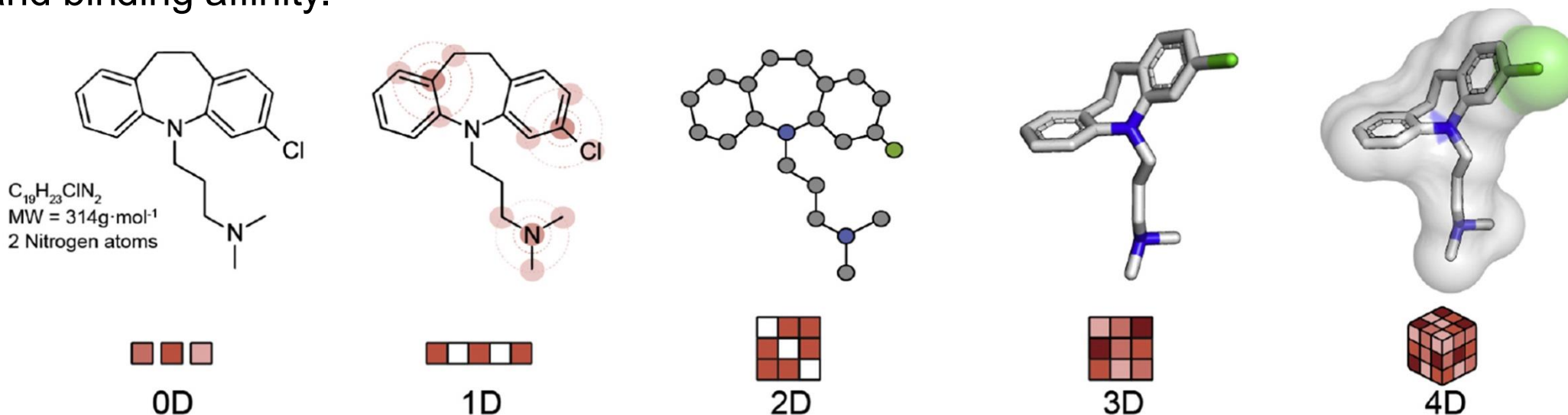
A good descriptor (representation)...

- is correlated with the properties of the molecules that are to be predicted.
- obeys physics.
- generates distinct values for structurally different molecules, even if the structural differences are small (non-degenerate, unique).
- is adapted to the size of the molecules that are used in the machine learning algorithm.
- is cheap to compute / store.
- changes smoothly with gradual changes in the molecular structure.
- is not trivially / linearly related to other descriptors.
- [is transferrable across elements.]
- [discriminates between isomers.]
- [is invariant with respect to atom labels and symmetries.]
- [is decodable / reversible.]



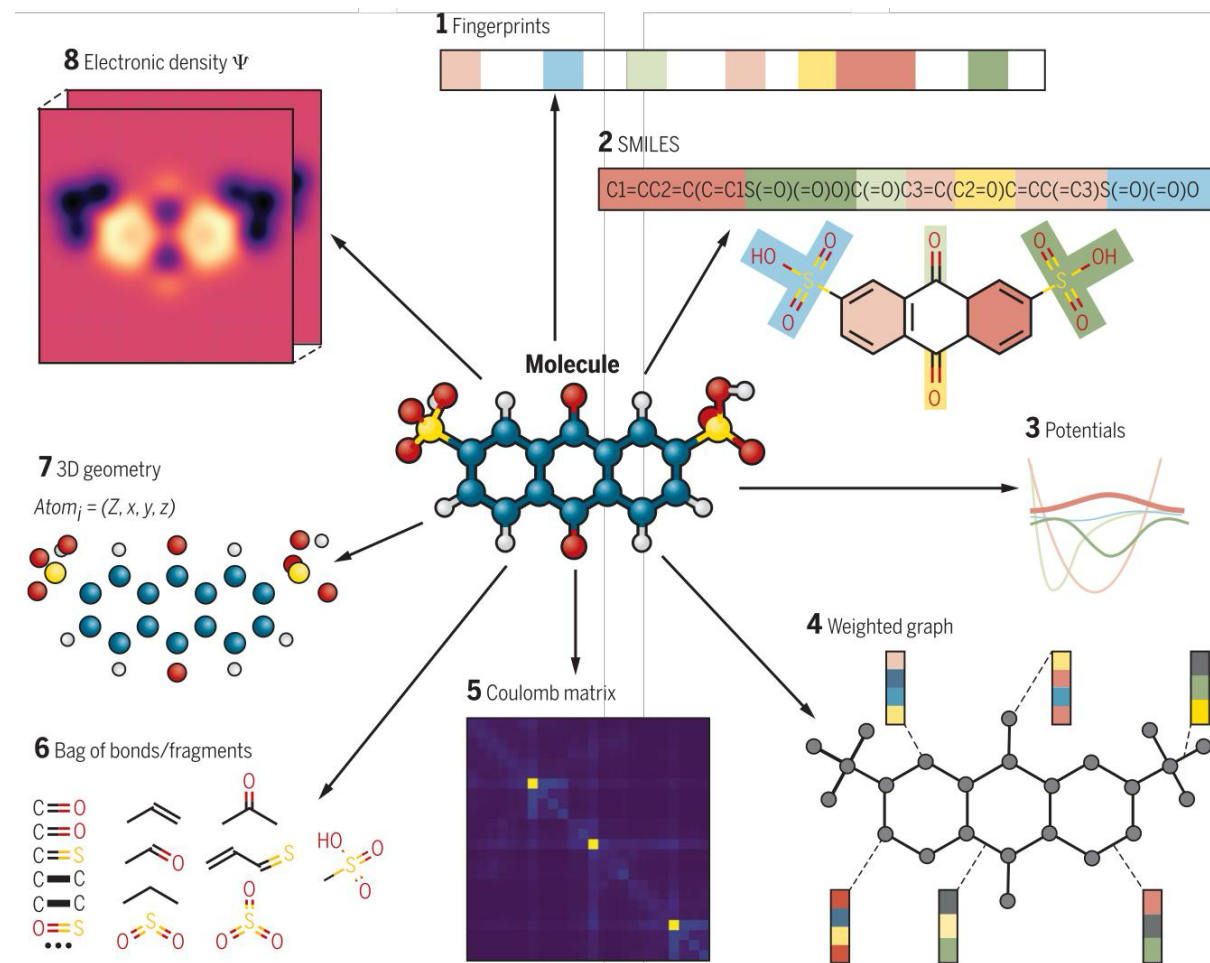
Types of descriptors

- **0D**: atom and bond counts, molecular weights; **do not provide any information about the molecular structure or connectivity of atoms.**
- **1D**: lists of substructures / fragments such as functional groups, e.g., fingerprints.
- **2D**: provide information on molecular topology often based on the graph representation of the molecules.
- **3D**: provide information about the spatial coordinates of atoms of a molecule, encoding molecular geometry.
- **4D**: "grid-based descriptors that introduce a fourth dimension to a 3D descriptor, e.g., protein-ligand binding affinity.



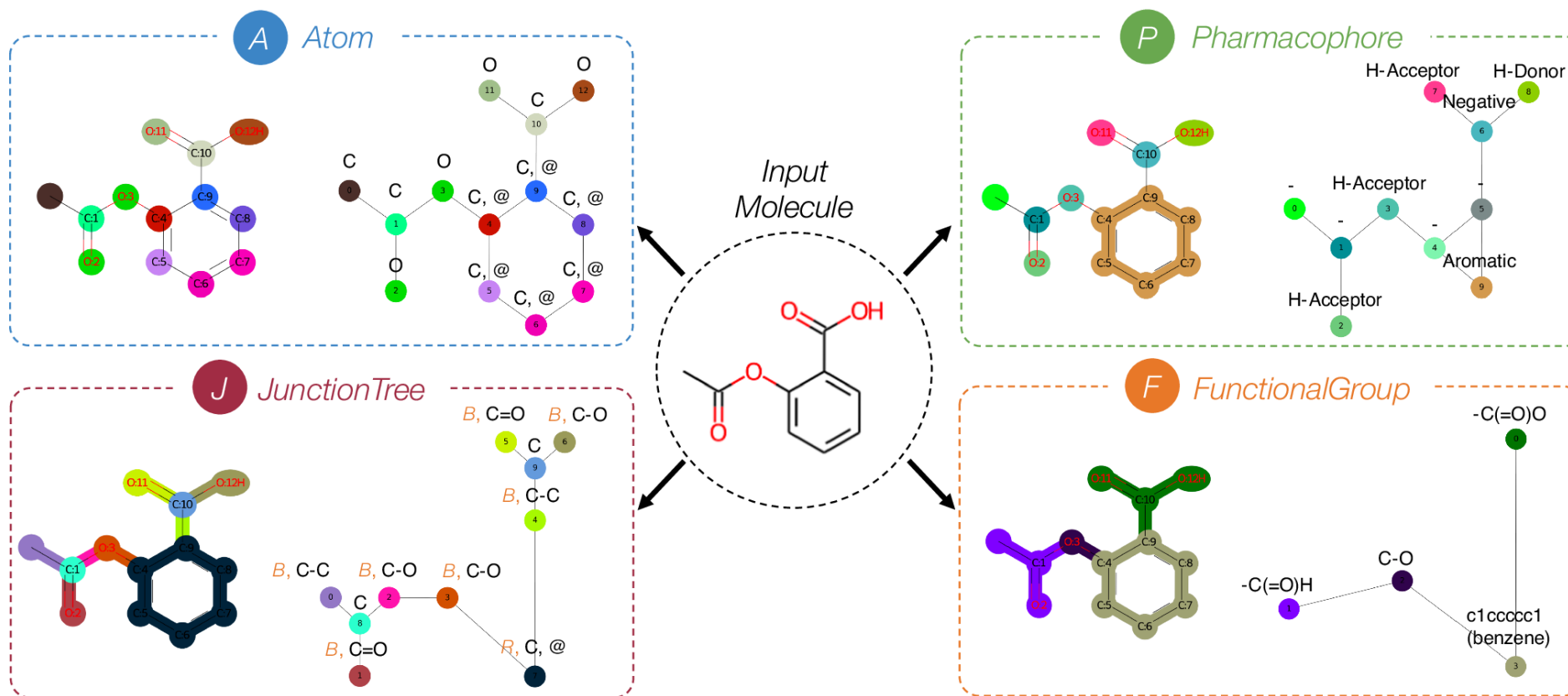
Molecular representations

1. A fingerprint vector that quantifies presence or absence of molecular environments;
2. SMILES strings that use simplified text encodings to describe the structure of a chemical species;
3. Potential energy functions that could model interactions or symmetries;
4. A graph with atom and bond weights;
5. Coulomb matrix;
6. Bag of bonds (BoB) and bag of fragments;
7. 3D geometry with associated atomic charges;
8. The electronic density.



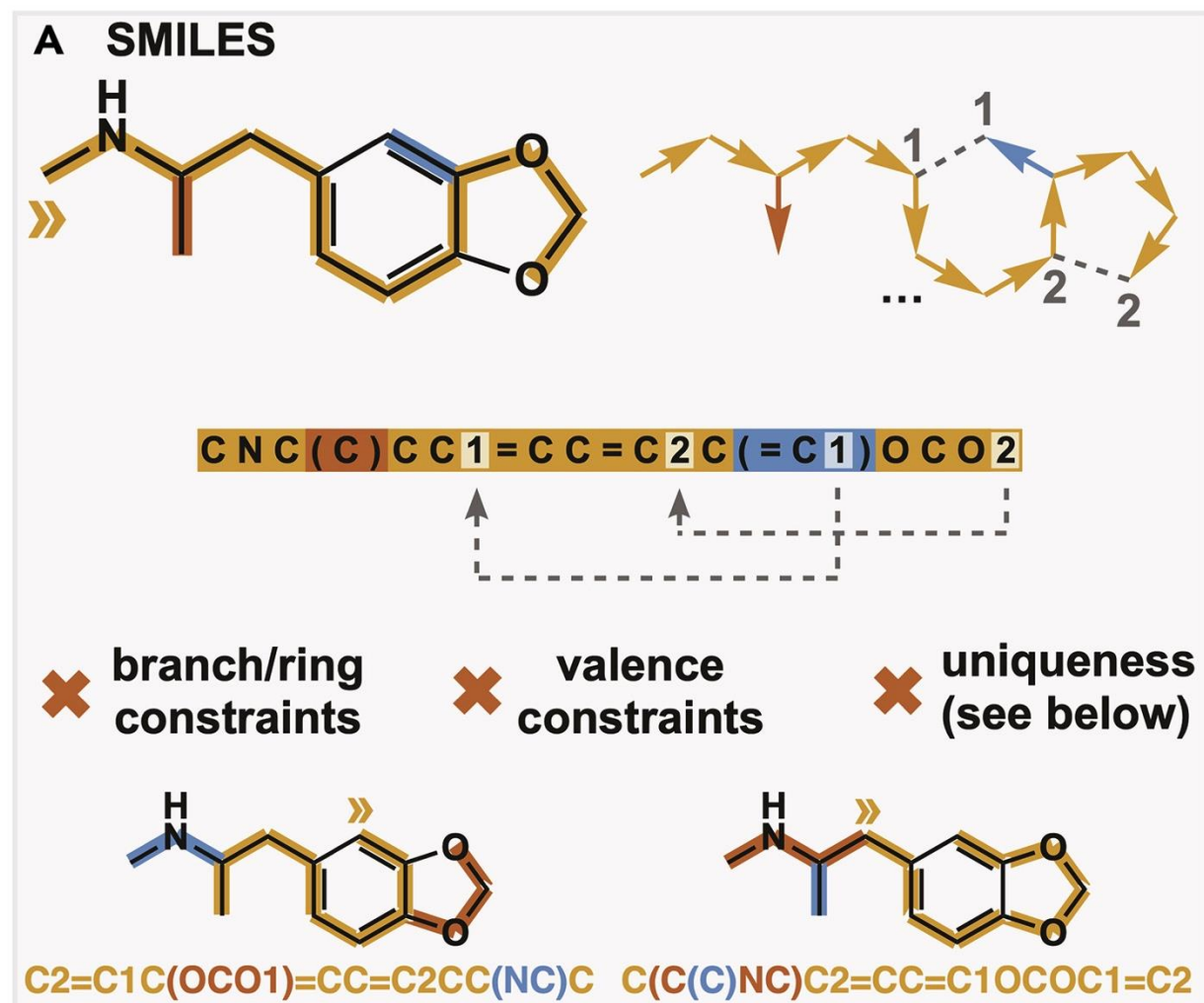
Molecular graph

is a connected, undirected graph which admits a one-to-one correspondence with the structural formula of a chemical compound in which the vertices of the graph correspond to atoms of the molecule and edges of the graph correspond to chemical bonds between these atoms.



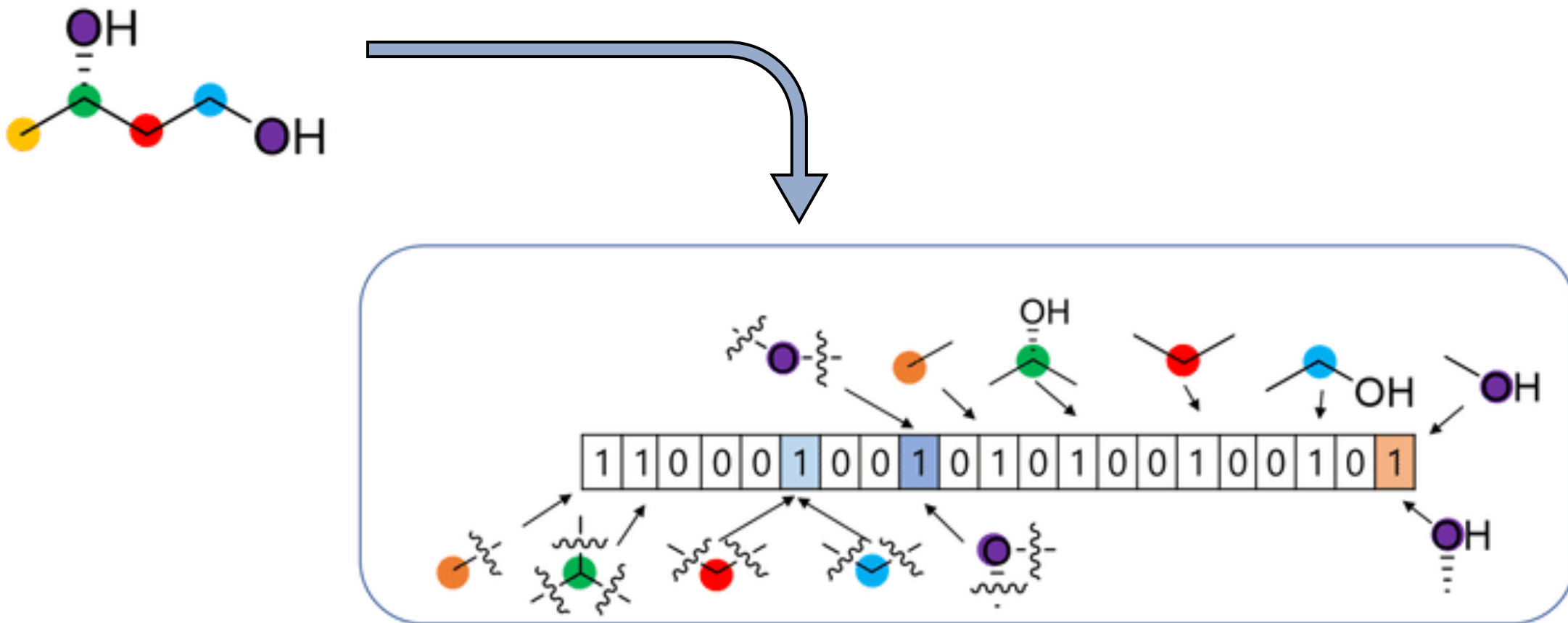
SMILES

Simplified Molecular Input Line Entry System is a specification in the form of a **line notation** for describing the structure of chemical species using short ASCII strings.



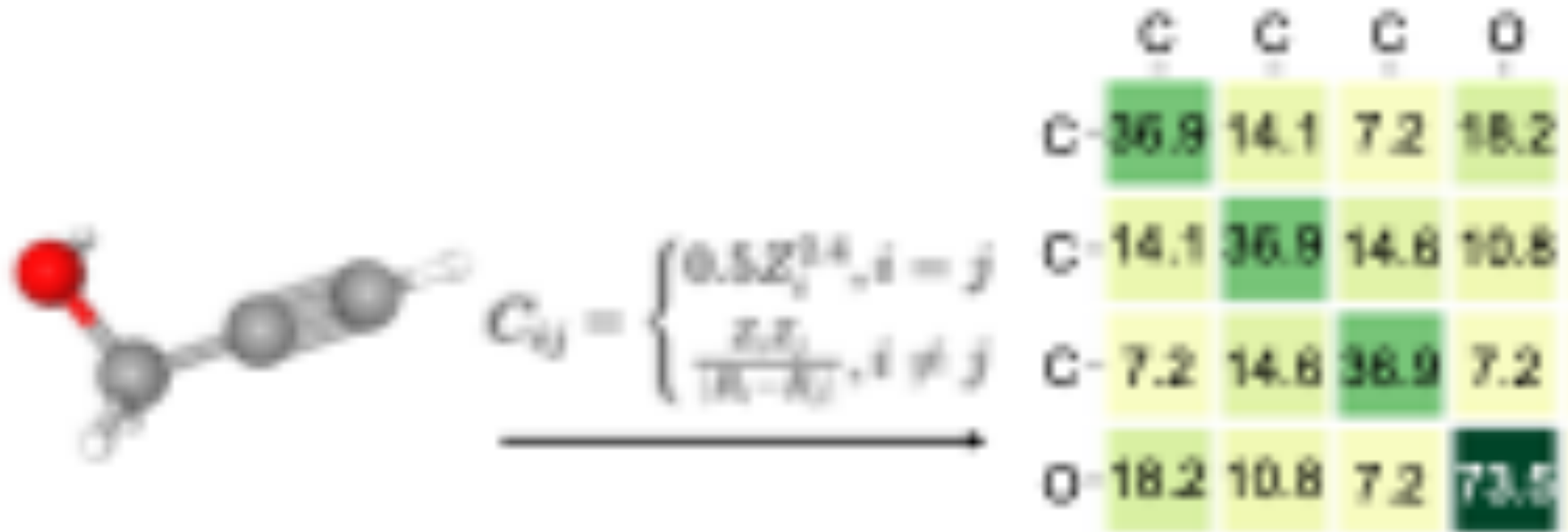
Morgan fingerprint

– a representation of a molecule that identifies the presence of **substructures / fragments** within it.



Coulomb matrix

is a simple global descriptor which mimics the **electrostatic interaction** between nuclei.

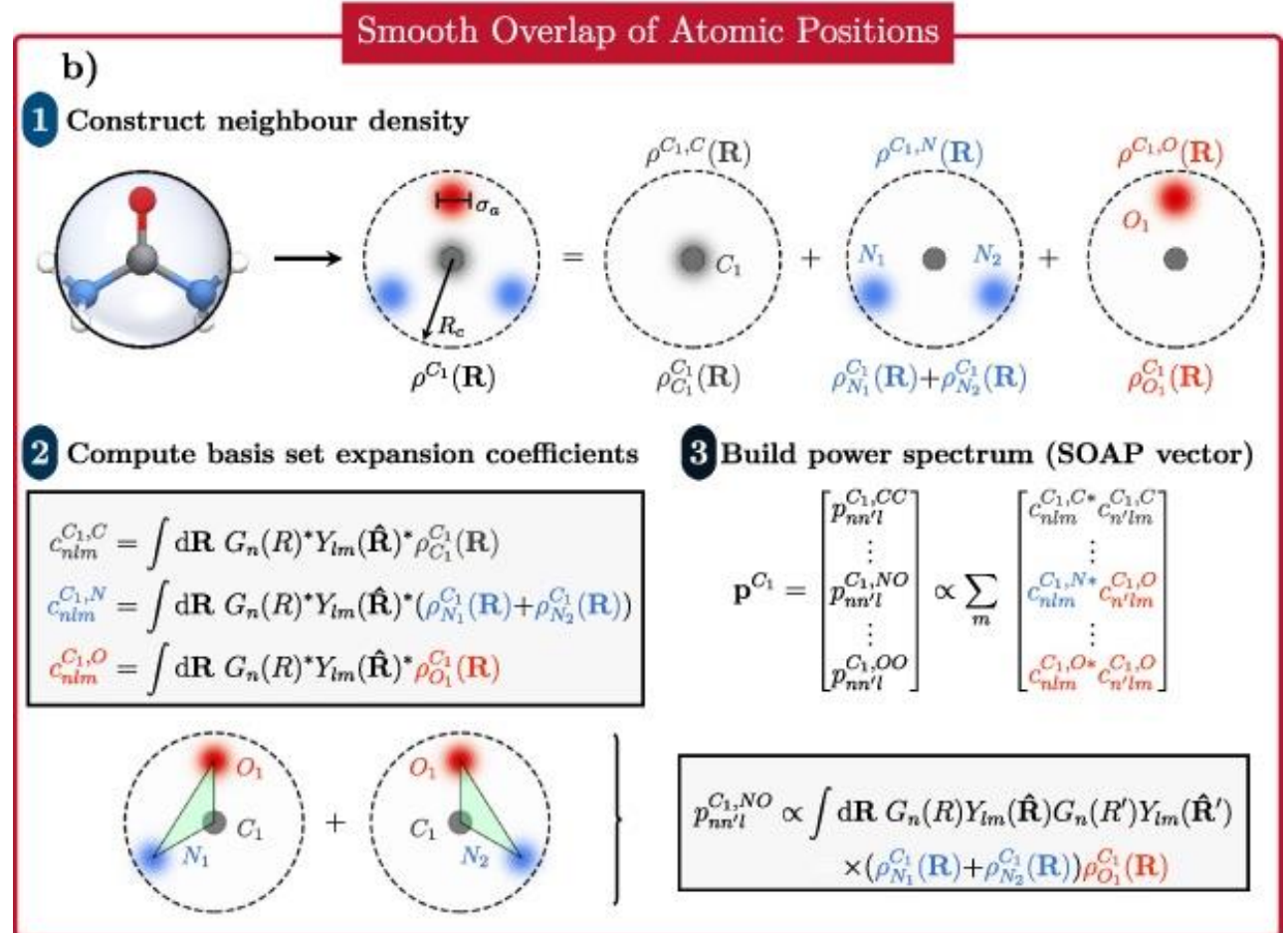


The diagonal elements can be seen as the interaction of an atom with itself and are essentially a polynomial fit of the atomic energies to the nuclear charge Z_i . The off-diagonal elements represent the Coulomb repulsion between nuclei i and j .

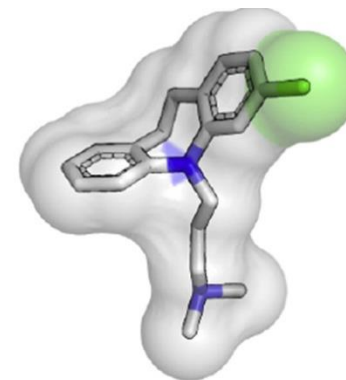
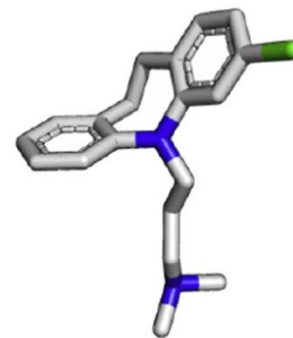
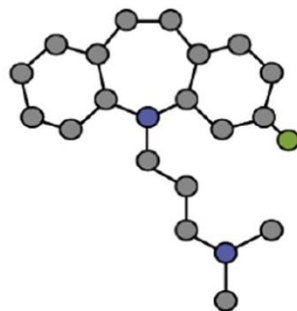
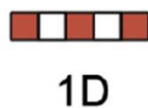
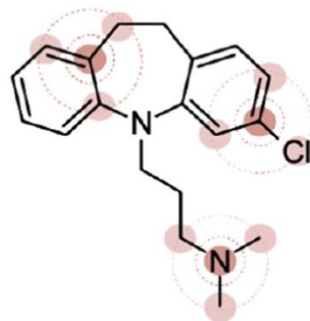
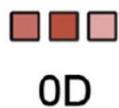
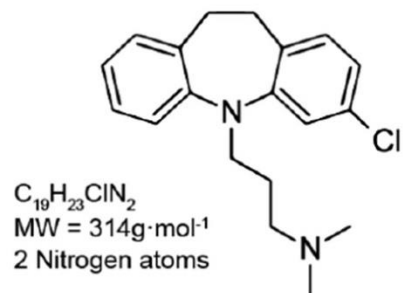
Smooth overlap of atomic positions

(SOAP) is a descriptor that encodes regions of atomic geometries by using a local expansion of a Gaussian smeared atomic density with orthonormal functions based on spherical harmonics and radial basis functions.

Schematic of the smooth overlap of atomic positions descriptor constructed for a non-planar urea molecule. (1) First, atomic positions are transformed into the neighbour density ρ , which is permutationally invariant. (2) Next, ρ is expanded in a local basis of radial functions and spherical harmonics, Y_{lm} . (3) Finally, summing up the square modulus of the expansion coefficients c_{nlm} over the index m provides the rotational invariance power spectrum p .

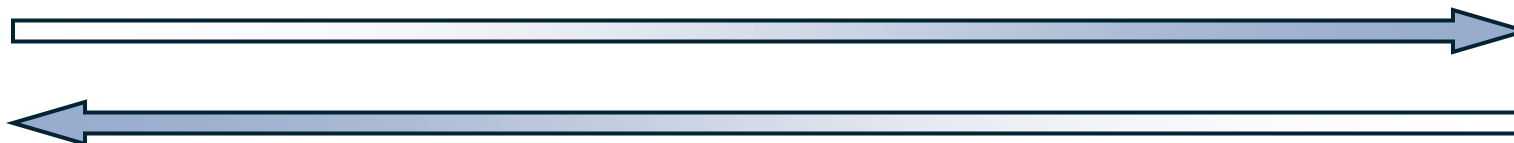


Types of descriptors



SMILES,
graphs

Richness of information



Physics-based,
quantum-inspired

Cost and complexity