Molecular Feature Vectors

(Representing molecules in a way that is good for ML)

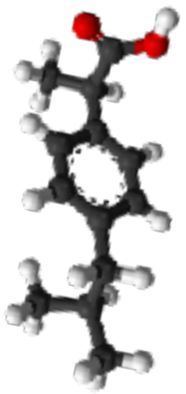
HUBDATA

IIIT, HYDERABAD

CCNSB

@deva_priyakumar

Recap



Z_i, R_i

x

Software
(eg. Gaussian16)

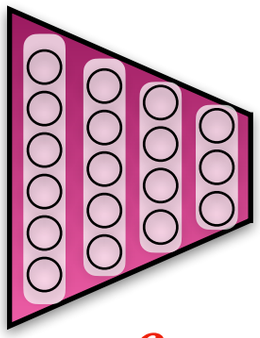
f

Properties
(eg. Electronic Energy)

$y = f(x)$

$\boldsymbol{x}$ and $\boldsymbol{f}$ are known;
we compute $\boldsymbol{y}$

x'



f

Properties
(eg. Solvation Free Energy)

$y = f(x')$

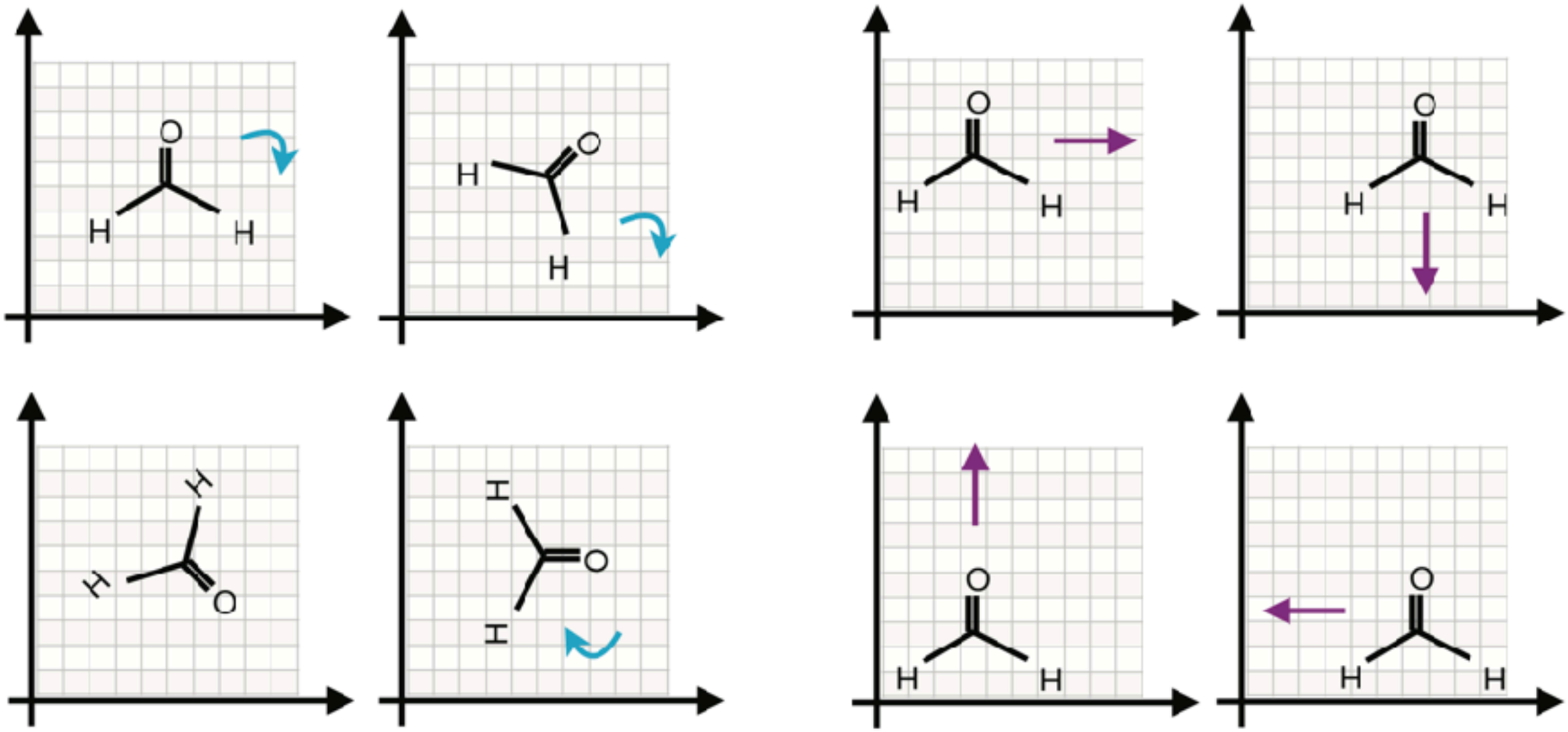
$\boldsymbol{x}$ and $\boldsymbol{y}$ are known;
we model $\boldsymbol{f}$

Machine Learning

Blank

In a Computational Tool, How do we Represent Molecules?

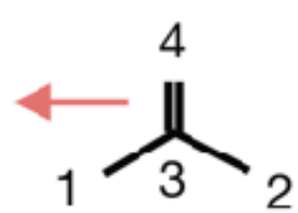
- Cartesian coordinates
- Internal coordinates
- Why unsuitable?
  - Not invariant to rotation
  - Not invariant to translation
  - Not invariant to change of atom ordering
  - Not a constant size feature vector

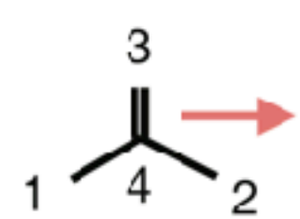


Rotations

Translations

$$\begin{bmatrix} H & 0 & 1 & 0 \\ 0 & H & 1 & 0 \\ 1 & 1 & C & 2 \\ 0 & 0 & 2 & O \end{bmatrix}$$



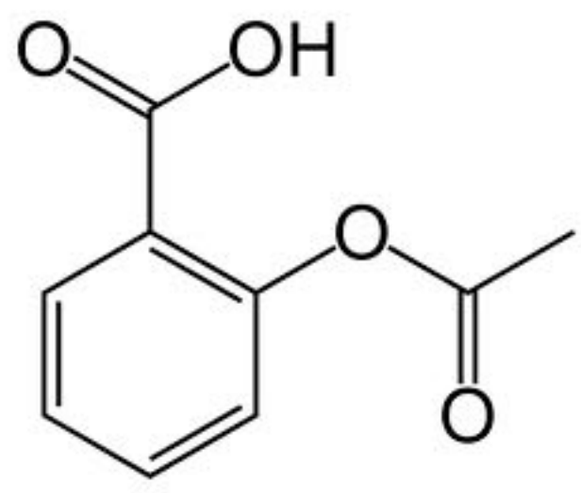


$$\begin{bmatrix} H & 0 & 0 & 1 \\ 0 & H & 0 & 1 \\ 0 & 0 & O & 2 \\ 1 & 1 & 2 & C \end{bmatrix}$$

Permutations

Molecular Representation

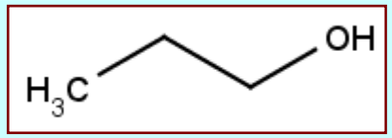
- Name of the Molecule - Aspirin or 2-acetoxybenzoic acid
- Molecular Formula - $C_9O_4H_8$
- Line diagram -



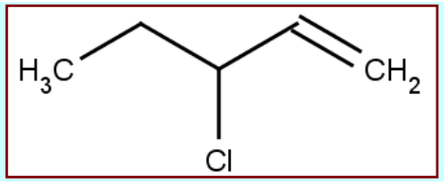
- Machine learning friendly featurizations
  - capture crucial information about the molecule
  - invariant to rotation
  - invariant to translation
  - invariant to the ordering of atoms
  - preferably a constant dimension for molecules of any size

(1) SMILES

- Atoms given as atomic symbols
- No hydrogens (implicit)
- Double bonds represented as “=” and triple bonds by “#”
- Branches by parenthesis
- Rings represented by allocating digits to the connecting atoms



SMILES representation : **CCCO**

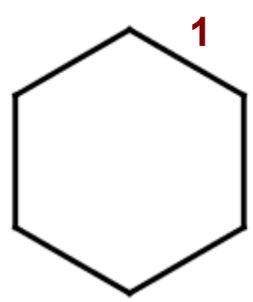


SMILES: **CCC(Cl)C=C**

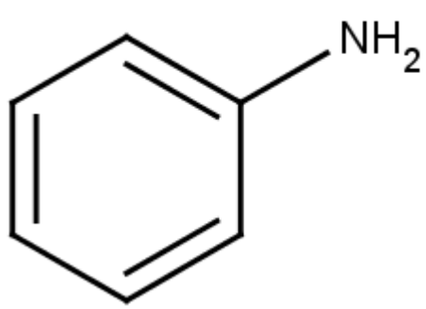
http://infochim.u-strasbg.fr/FC/docs/Intro/FC_DB.pdf

- Rings represented by allocating digits to the connecting atoms

- Aromatic rings represented by lowercase



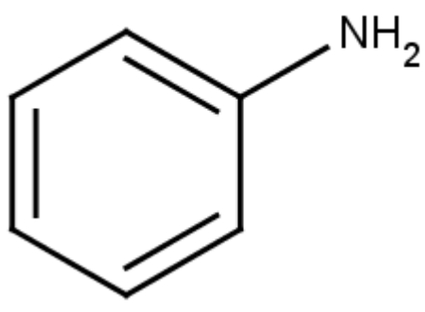
SMILES: **C1CCCCC1**



SMILES: **Nc1ccccc1**

- Unambiguous - a given SMILES string represents a unique structure.

- But, how about the reverse?



SMILES: **Nc1ccccc1**

but also **c1ccccc1N**

or **c1cc(N)ccc1**

- Canonical representation of SMILES (**FIND OUT**)

(2) Molecular Fingerprints

- Bit string - set the bits depending on the presence or absence of a given list of features

- Number of bits determined by the number of structural keys in the whole dataset.

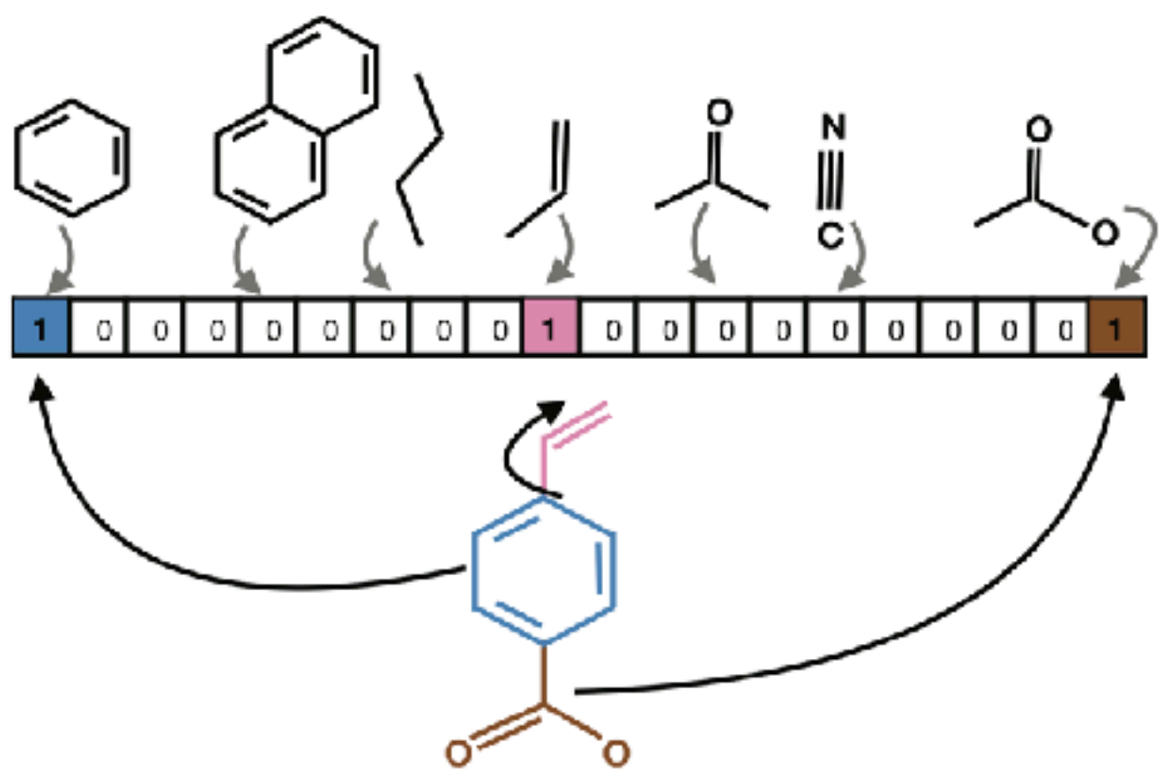
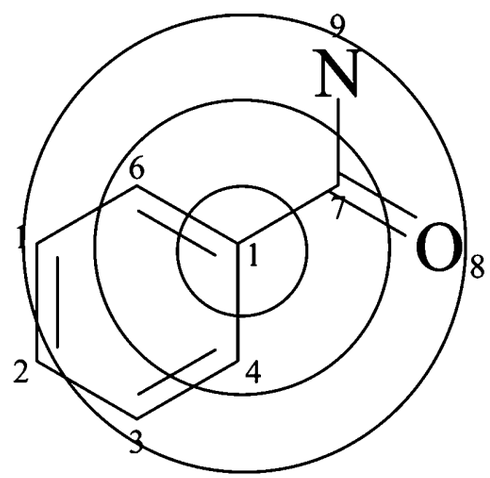


FIGURE 3 Representation of a molecular fingerprint encoded the presence (1) or absence (0) of certain substructures in a compound. This molecule is represented by a vector of length 20 consisting of binary numbers

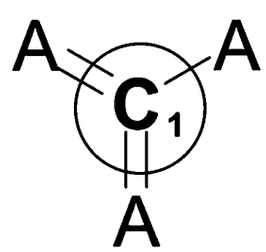
Extended Connectivity Fingerprints (ECFP)

- Each molecule decomposed into submodules originating from each heavy atom
- Each of these assigned with a unique identifier
- This is extended through bonds to generate larger substructures (and corresponding identifiers)
- Hash all the substrutures into a fixed length binary fingerprint representation.

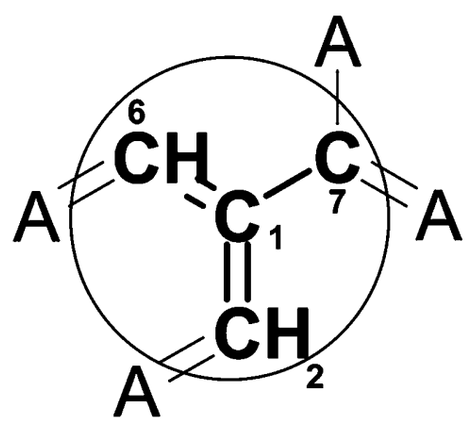
J. Chem. Inf. Model. **2010,** *50,* 742–754



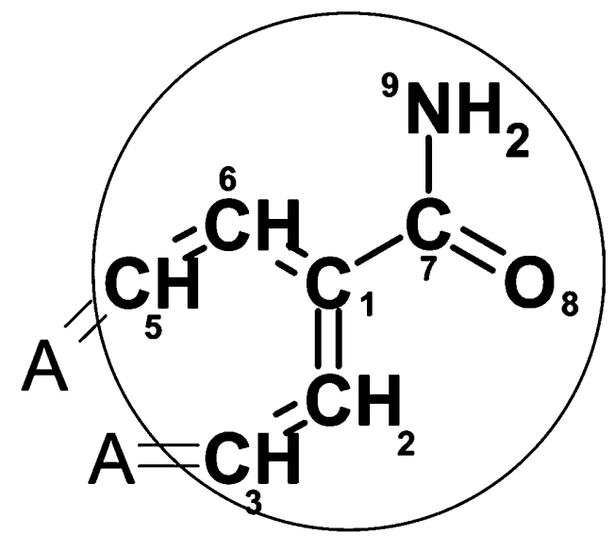
Considering atom 1 in benzoic acid amide



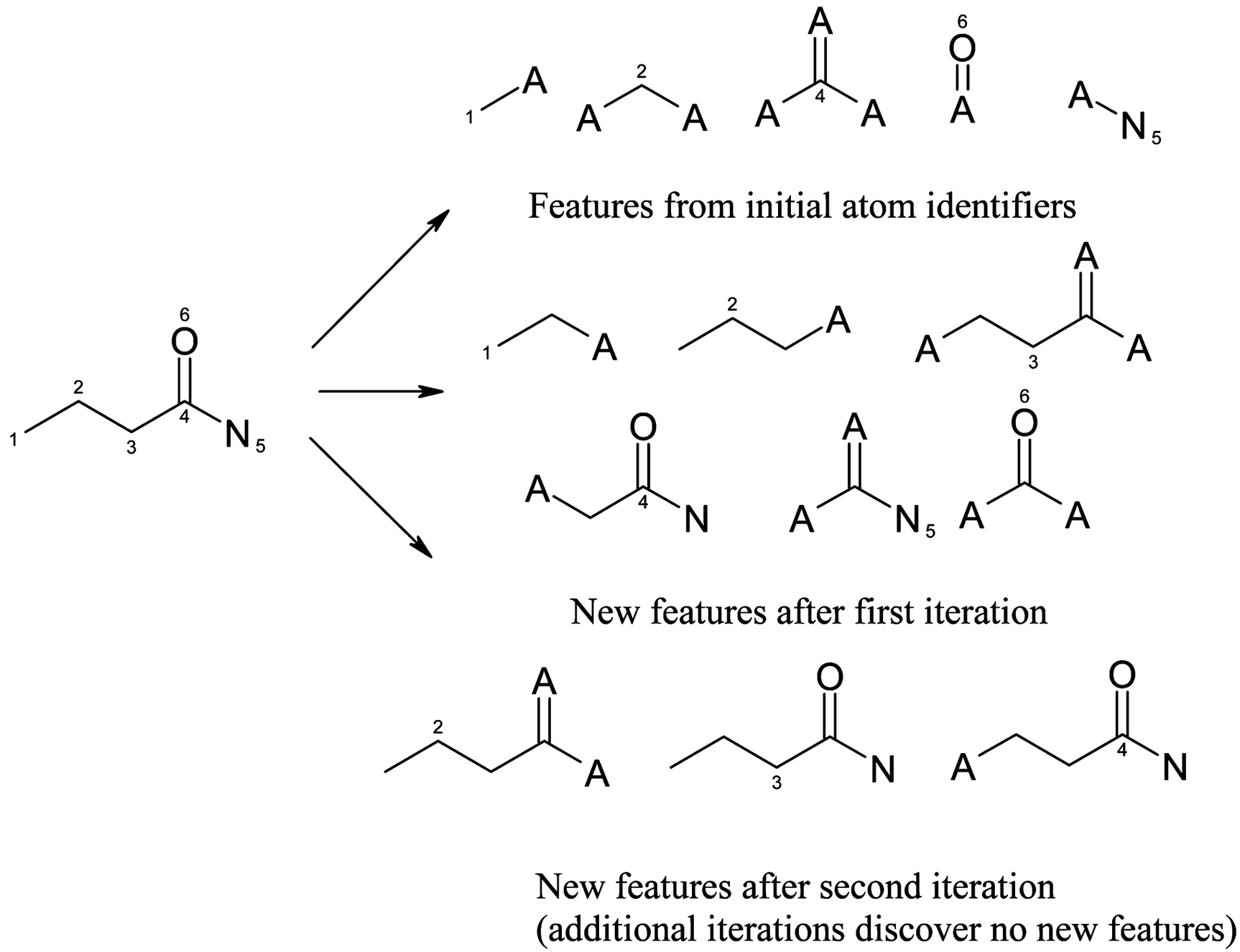
Iteration 0

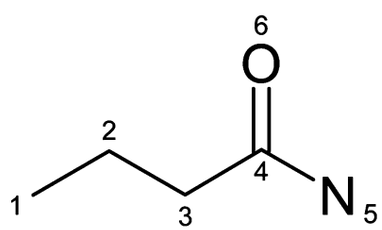


Iteration 1



Iteration 2


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



> <ECFP_0>	> <ECFP_2>	> <ECFP_4>	> <ECFP_6>
734603939	734603939	734603939	734603939
1559650422	1559650422	1559650422	1559650422
-1100000244	-1100000244	-1100000244	-1100000244
1572579716	1572579716	1572579716	1572579716
-1074141656	-1074141656	-1074141656	-1074141656
	863188371	863188371	863188371
	-1793471910	-1793471910	-1793471910
	-1789102870	-1789102870	-1789102870
	-1708545601	-1708545601	-1708545601
	-932108170	-932108170	-932108170
	2099970318	2099970318	2099970318
		-87618679	-87618679
		1112638790	1112638790
		-627599602	-627599602

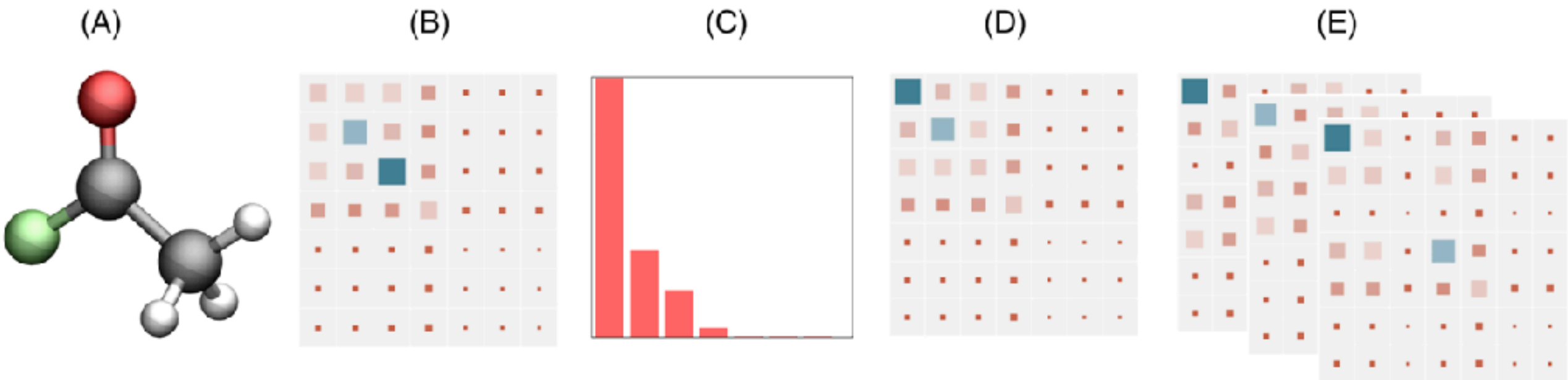
- Unique
- Rotational/translational invariant
- But, no information about the three-dimensional structure

Why do we need to represent the 3D structure?

(3) Coulomb Matrix

$$M_{IJ} = \begin{cases} 0.5 Z_I^{2.4} & \text{for } I = J \\ \frac{Z_I Z_J}{|\boldsymbol{R}_I - \boldsymbol{R}_J|} & \text{for } I \neq J \end{cases}$$

- Off-diagonal elements => Coulomb repulsion between atoms I and J
- Diagonal elements => Polynomial fit of atomic self-energy


(A)
(B)
(C)
(D)
(E)

- Not invariant to swapping of atomic indices
- Tricks:
  - Use randomly permuted sets of matrices
  - Use eigen values of the Coulomb matrix…
- Not invariant to molecular size - pad with zeroes

(4) Bag of Bonds

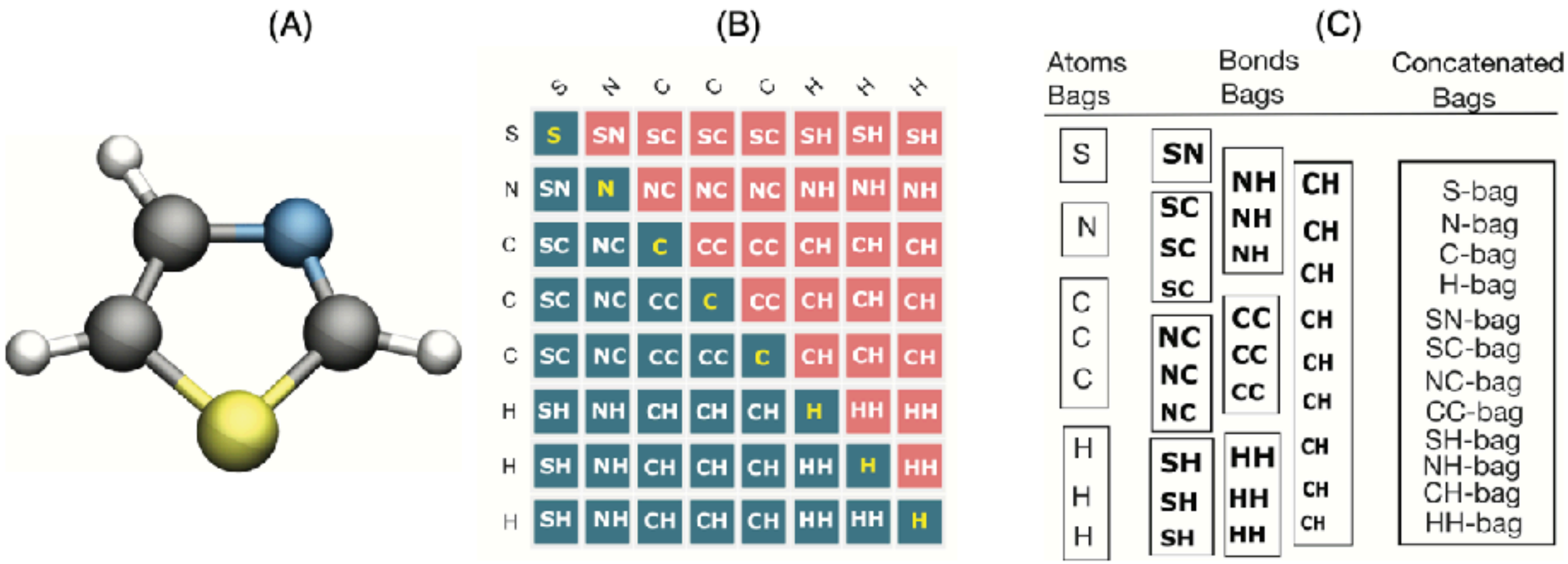


FIGURE 5 Schematic view of the bag-of-bonds (BoB) representation: (A) ball and stick representation of the thiazole (C_3H_3NS) molecule, (B) Coulomb matrix elements, (C) different Coulomb matrix entries (off-diagonal) are sorted into different bags. Here, concatenated bags are not padded with zeros for clarity

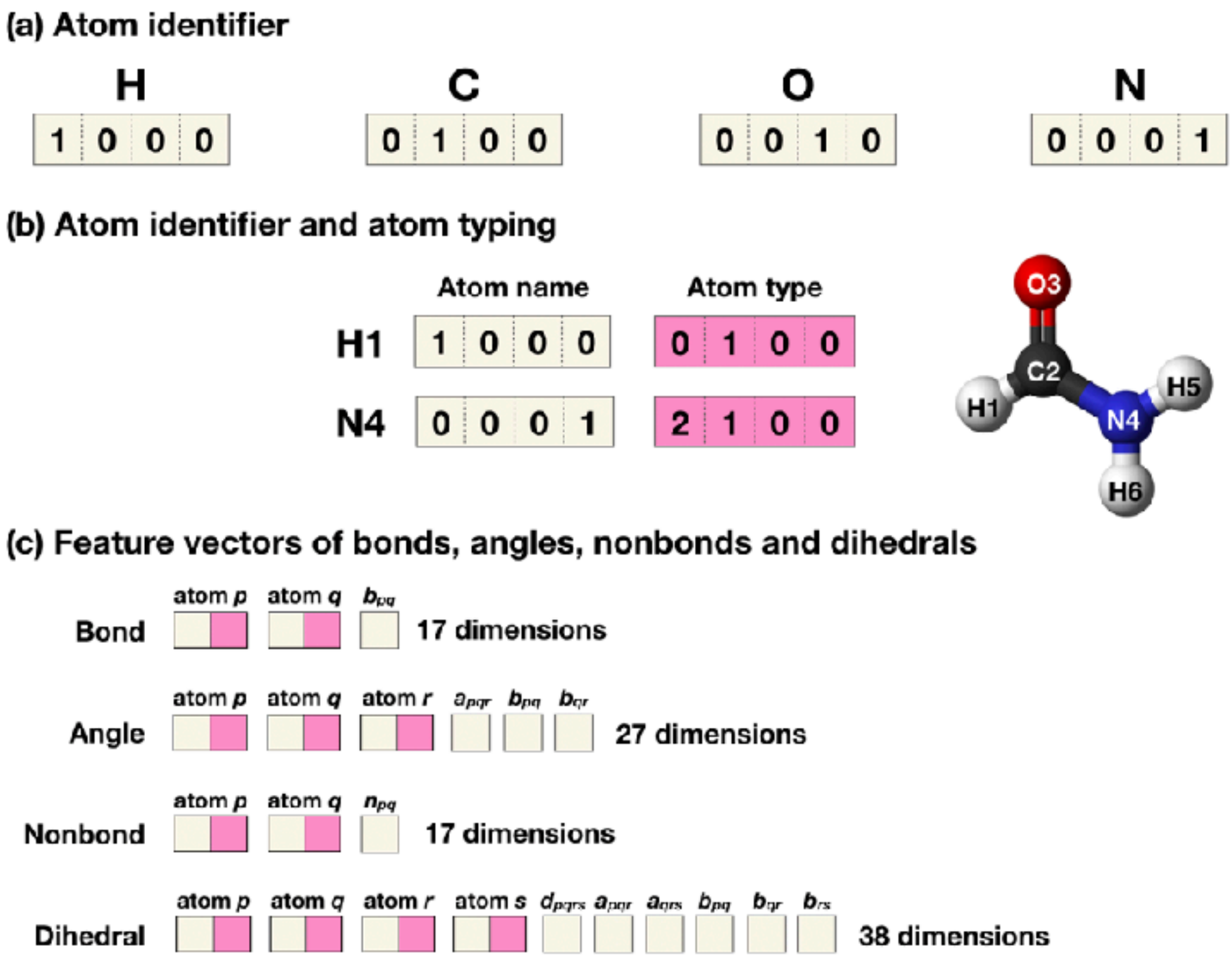
Each entry in the bag computed as

$$Z_i Z_j = \frac{Z_i Z_j}{|R_i - R_j|}$$

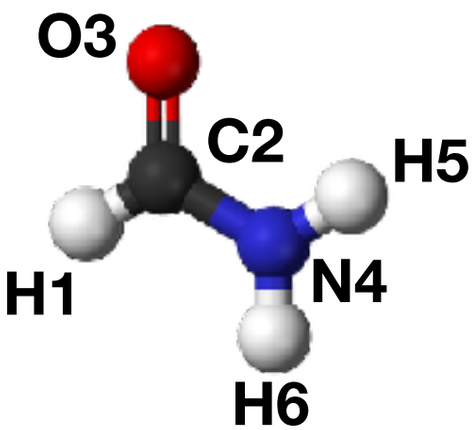
J. Phys. Chem. Lett. 2015, 6, 2326–2331

- Each entry in the bag is sorted
- Then sorted in a specific order
- Padded with zeros to account for molecular size invariance

(5) BAND



https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.26128



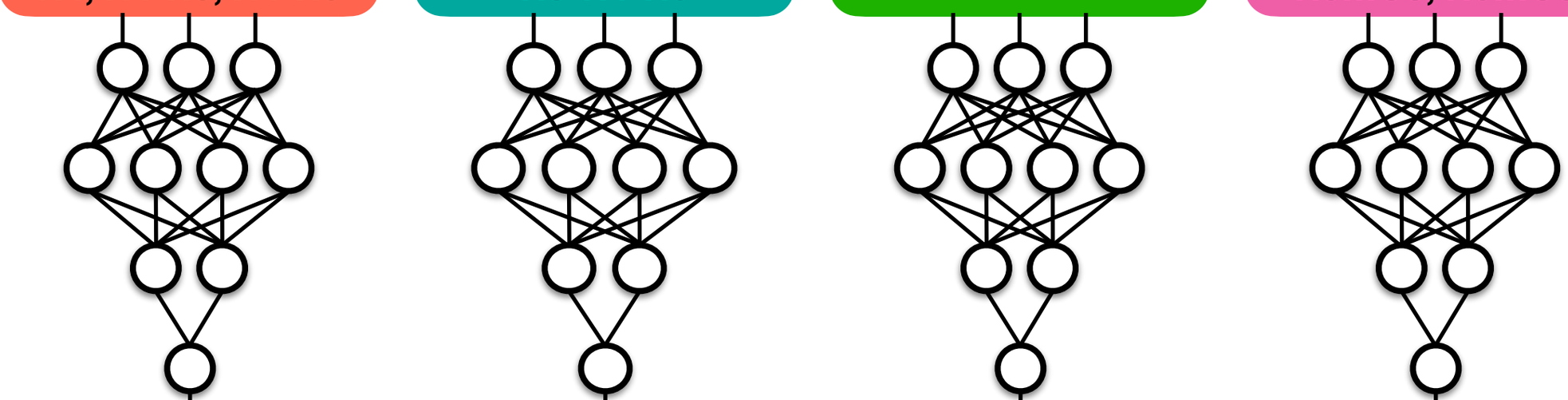
Cartesian coords.
$r_1, r_2, \ldots r_N; Z_1, Z_2, \ldots Z_N$

Bonds
H1-O3, C2-O3, C2-N4, N4-H5, N4-H6

Angles
H1-C2-O3, H1-C2-N4, N4-C2-O3, H5-N4-C2, N6-N4-C2, H5-N4-H6

Dihedrals
H1-C2-N4-H5, H1-C2-N4-H6, H5-N4-C2-O3, H6-N4-C2-O3

Nonbonds
H1…O3, H1…N4, O3…N4, C2…H6, C2…H5, H5…H6 H1…H6, H1…H5, H5…O3, H6…O3



E_b E_a E_d E_n

$$E = E_b + E_a + E_d + E_n$$

TO Learn:

SELFIES

ANI - Originally Symmetry Fn. by Behler & Parinello

SchNet

Slide added after the lecture

- https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870
- https://jcheminf.biomedcentral.com/articles/10.1186/s13321-020-00460-5