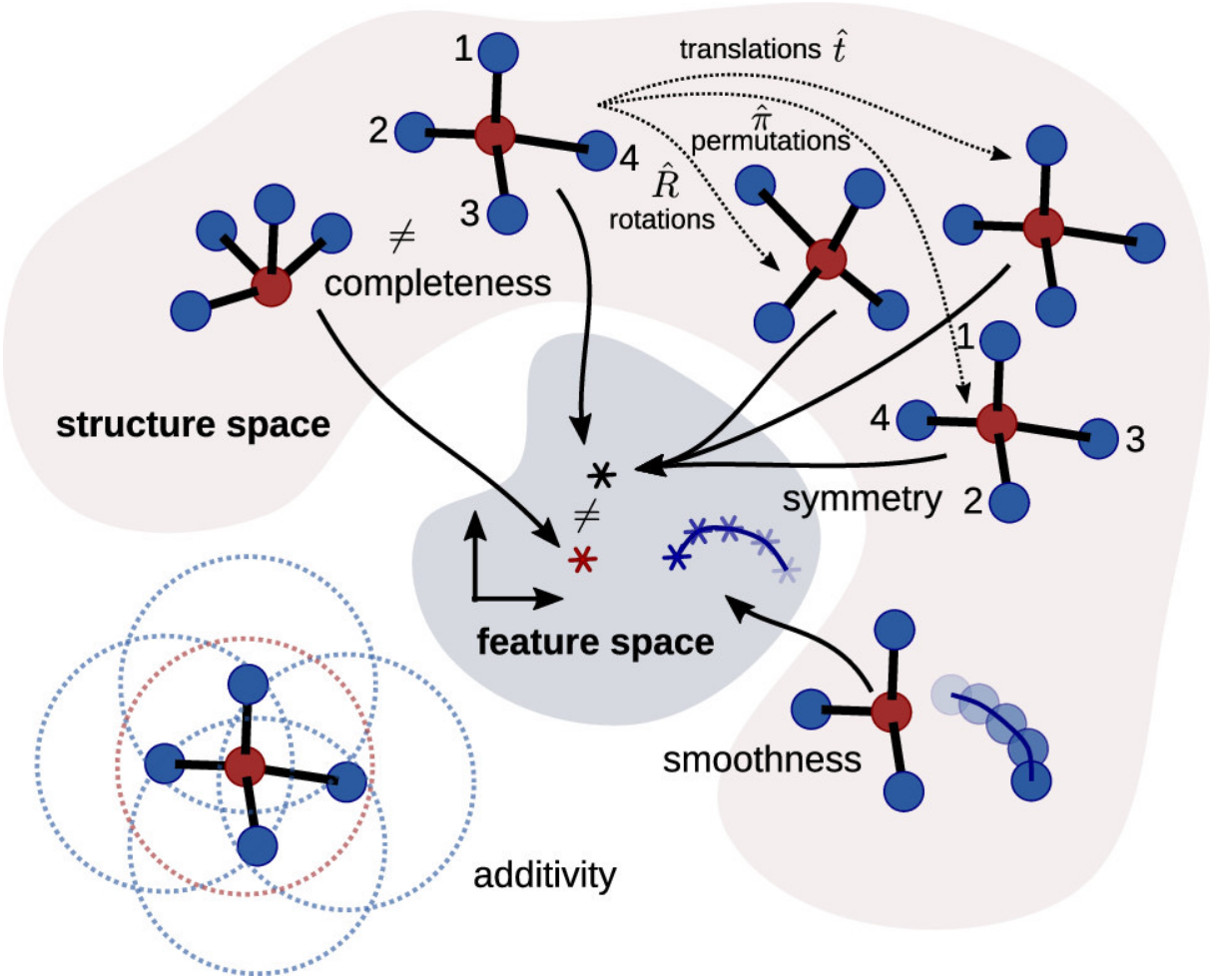
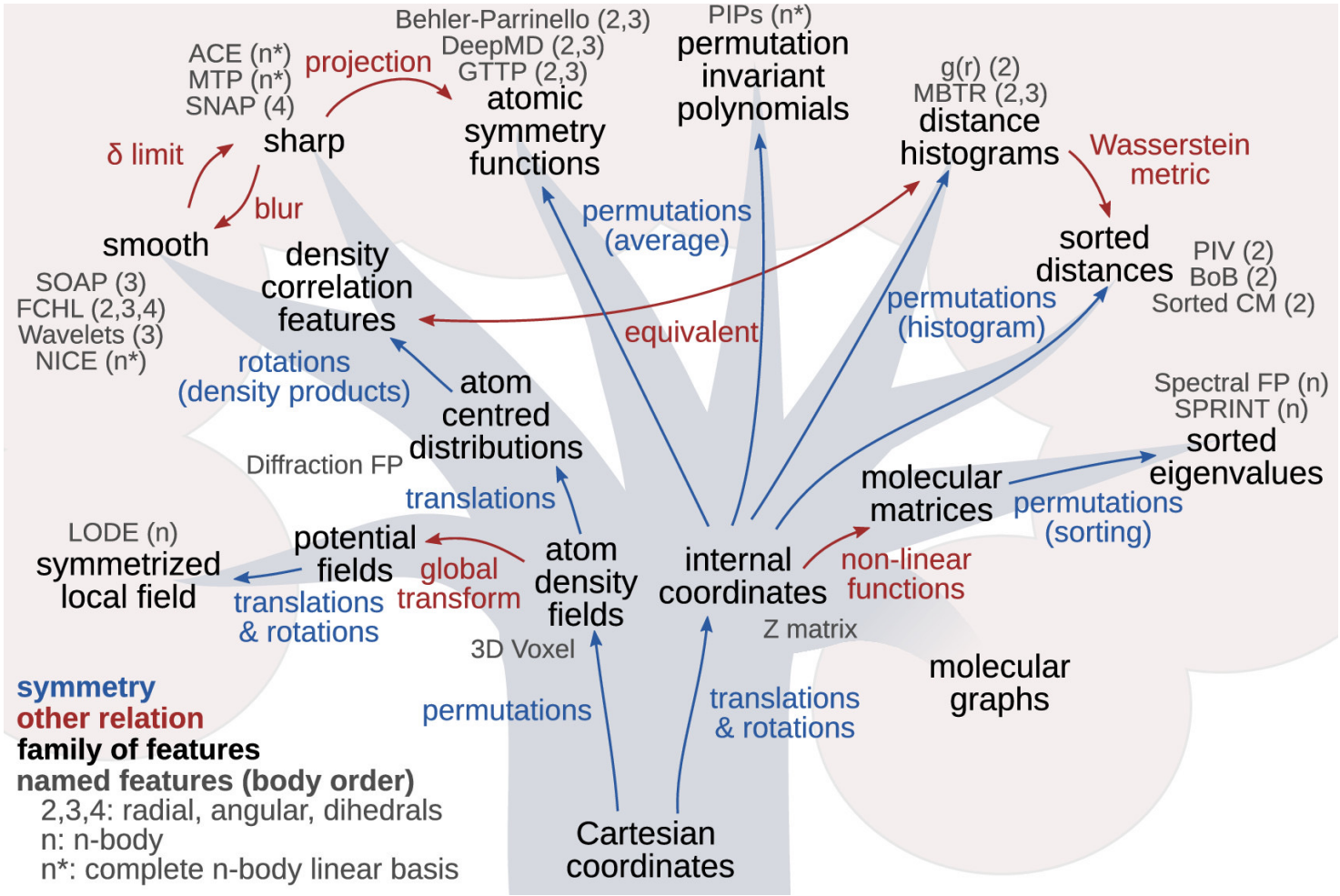
Representations

Requirements



Musil et al., *Chem Rev*, 2021.

Overview



Musil et al., *Chem Rev*, 2021.

Compact Representations Useful

Expected error given sufficient data

$$\exp(-c/h_{X,\Omega})$$

Positive constant c

Function of the „fill distance“

$$h_{X,\Omega} \equiv \sup_{y \in \Omega} \min_{x_j \in X} \|y - x_j\|_2$$

Training data X
Domain Ω

Shorter fill distance: steeper learning curve, i.e. more data efficient model

Madych, Nelson, *J Approx Theory*, 1992.

Categorial: One-Hot

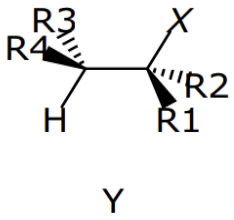
Categorial data vs regression

Solution: „binary“, „dummy“, „one-hot“ encoding

Encode *n* categories as vector of length *n-1* with one category (arbitrary) being the null vector.

Example:
- A: (0, 0)
- B: (1, 0)
- C: (0, 1)

Chemistry example: (ABBA|CD) -> (0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0|0,1,0,0,1)



	A	B	C	D	E
R*k*	H	NO_2	CN	CH_3	NH_2
X	F	Cl	Br		
Y	H	F	Cl	Br	

GFvR, SN Heinen, M Bragato, Anatole von Lilienfeld, *Mach. Learn.: Sci. Technol.* 2020.

Graph-based

Often: molecules = well-defined bonds

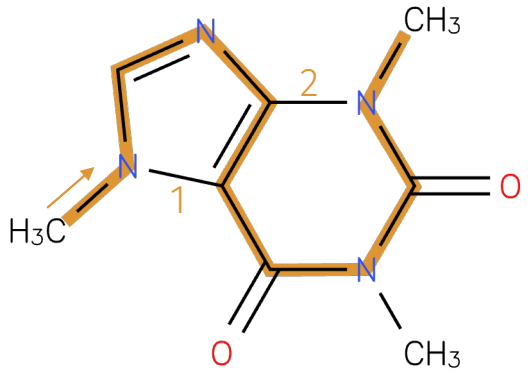
Adjacency matrix
- 1 if atoms i and j are bonded
- 0 otherwise

$$\begin{matrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{matrix}$$

Bond order matrix
- Bond order if atoms i and j are bonded
- 0 otherwise

$$\begin{matrix} 0 & 2 & 0 \\ 2 & 0 & 2 \\ 0 & 2 & 0 \end{matrix}$$

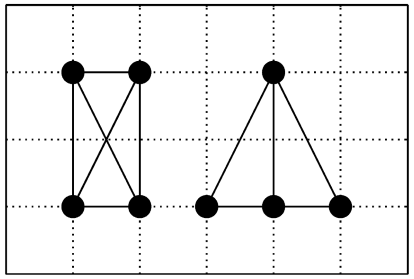
SMILES
- Bonds: Nothing (1), = (2), # (3), $ (4) O=C=O
- Partial charges [Na+]
- Fragments: . [Na+].[Cl-]
- Rings: labels C1CCCCC1
- Branches: parentheses



CN1C=NC2=C1C(=O)N(C)C(=O)N2C

Distance-based

Many-body descriptions, e.g. all pairwise distances



	s	l	m
s		m	l
l	m		s
m	l	s	

	l	l	m
l		m	s
l	m		s
m	s	s	

Even three- and four-body interactions not unique [1].

Chemistry: **Coulomb Matrix [2]**

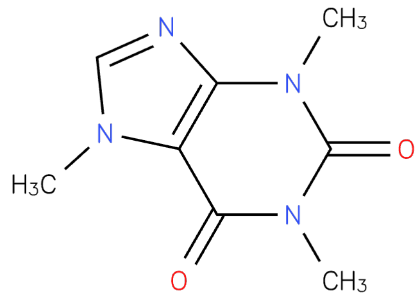
- Diagonal: $0.5 Z_i^{2.4}$
- Off-diagonal: $Z_i Z_j / \|\mathbf{R}_i - \mathbf{R}_j\|$
- Problem: sorting, uniqueness

[1] Pozdnyakov et al., *Phys Rev Lett* 2020. [2] M. Rupp et al., *Phys Rev Lett* 2012.

Fingerprints

Collect relevant features of a molecule into a vector.

- Fixed list of functional groups (e.g. Joback method)
- Hash of generated circular atom environments (e.g. ECFP)
- Kinds of generated local environments (e.g. Morgan)
  - Element, # heavy neighbors, # protons, charge, part of ring…
- Fixed checklist of features (e.g. MACCS keys)
  - Subgraphs



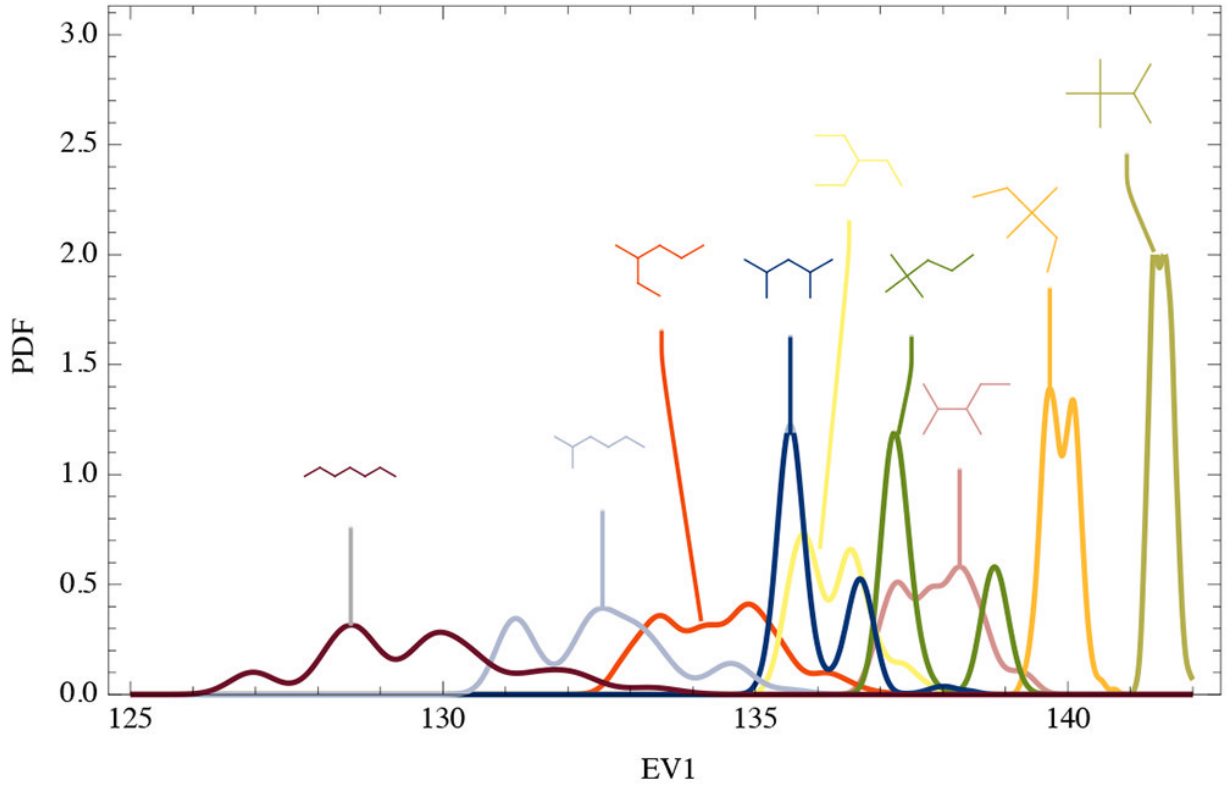
Morgan (feature: count)
10565946: 2, 348155210: 1, 476388586: 1, 540046244: 1, 553412256: 1, 864942730: 2, 909857231: 1, 1100037548: 1, 1333761024: 1, 1512818157: 1, 1981181107: 1, 2030573601: 1, 2041434490: 1, 2092489639: 3, 2246728737: 3, 2370996728: 1, 2877515035: 1, 2971716993: 1, 2975126068: 2, 3140581776: 1, 3217380708: 4, 3218693969: 1, 3462333187: 1, 3657471097: 3, 3796970912: 1

MACCS keys:
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0]

Spectra

Condense a matrix into its eigenvalues

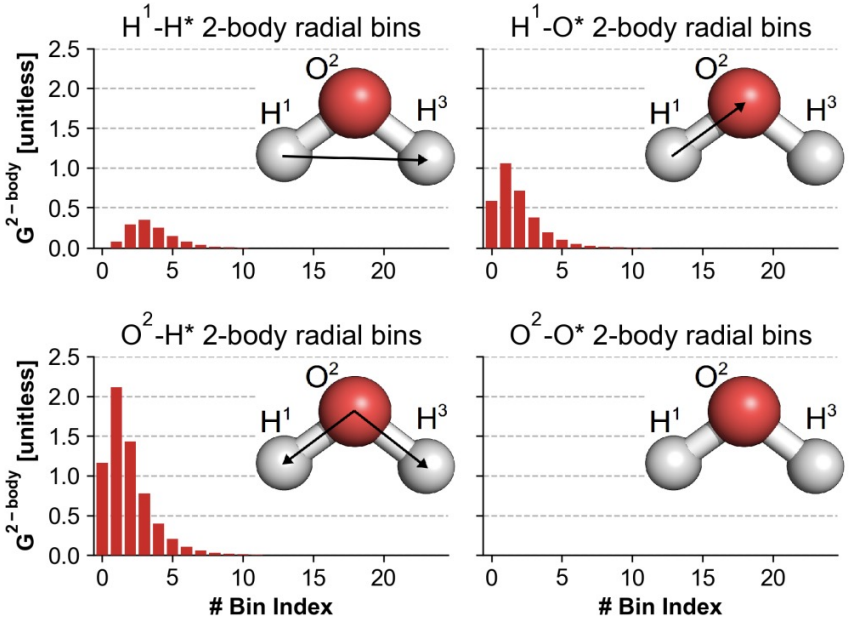
- No sorting issue, as permutationally invariant
- Smaller: N instead of N^2
- Lossy

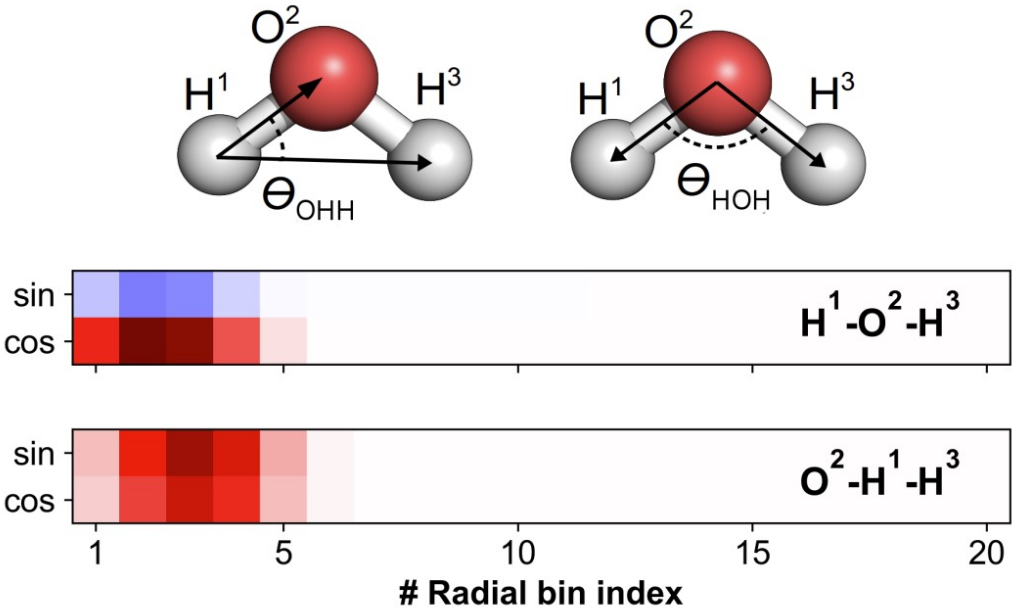


J. Schrier, *J Chem Inf Model*, 2020.

Densities

Smear positions into densities (FCHL / SOAP)





Christensen et al, *J Chem Phys*, **2020**.

Summary Representations

- Helpful properties of a representation
  - Unique
  - Smooth in representation space
  - Changes in geometry smooth in representation
  - Complete
  - Invariant under transformations (translation, rotation, permutation)
  - Compact
  - Additive
  - Invertible
  - Fast
  - Simple
  - Containing many-body effects