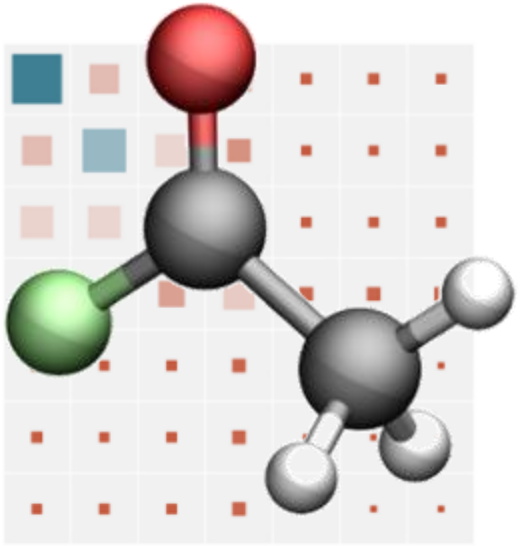
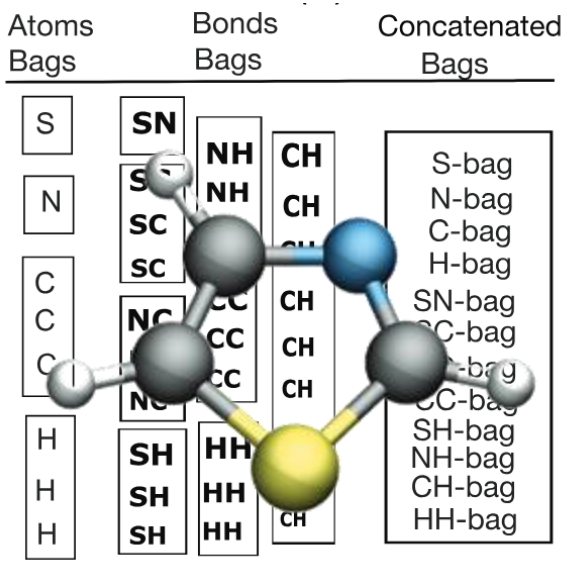
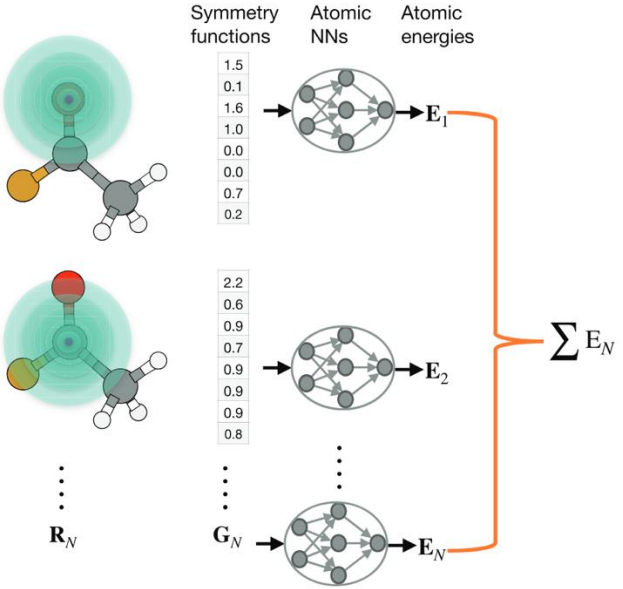
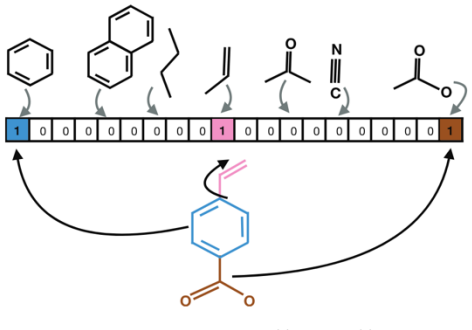








Molecular Featurization

image sources: S. Raghunathan, U. D. Priyakumar, Int. J. Quantum Chem. 2022, 122(7), e2687

Tokenization and One-Hot Encoding

One viable approach for molecules:

1. *Represent numerically via one-hot encoding over chemical vocabulary*
2. *Tokenize SMILES strings into finite set of (n-gram) characters*

One-hot Encoding – bitwise/binary description of categorical variables

e.g.,

cat

01100011011000010111010**0**

0

100

rabbit

01110010011000010110001**0**

1

010

duck

01100100011101010110001**1**

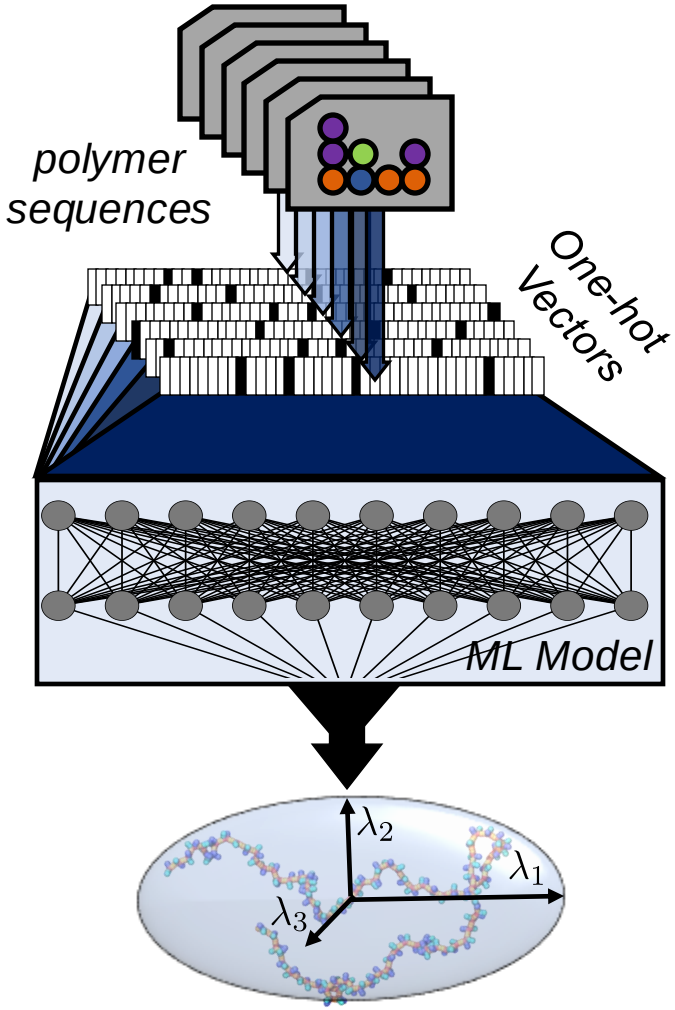
2

001

English; binary; numerical assignment; one-hot vector

Tokenization and One-Hot Encoding

Usually OHE is the first reasonable thing to try if you have countable units

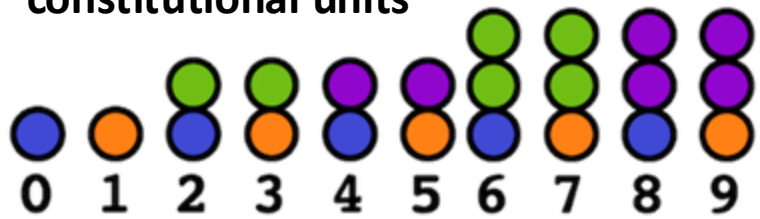


characterization

Webb et al. *Sci. Adv.* **2020**

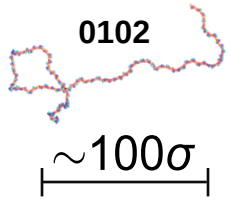
10 allowable "constitutional units"

$R_k \in \{0, \ldots, 9\}$

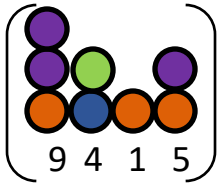


$\left(-R_1-R_2-R_3-R_4-\right)_{200}$

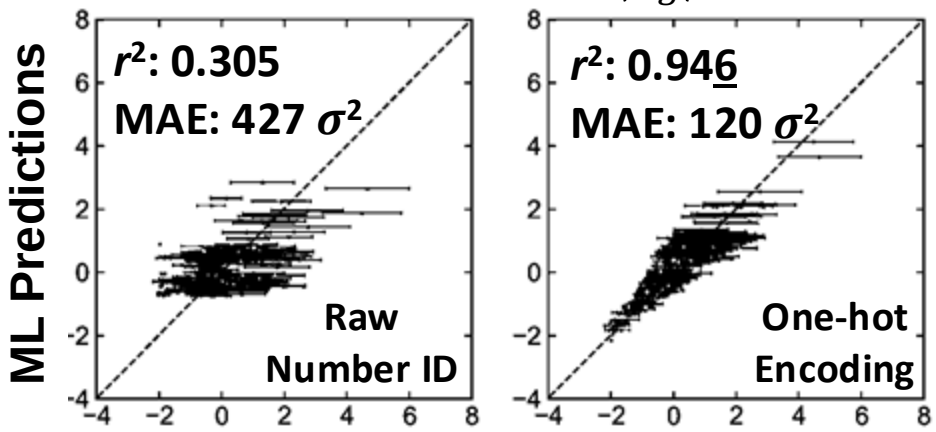
0254



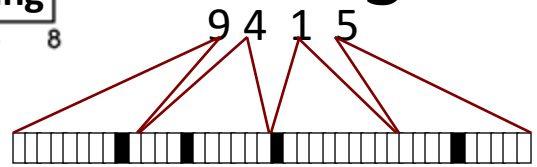
raw number ID



Standardized $\langle R_g^2 \rangle$



one-hot-encoding



Tokenization and One-Hot Encoding

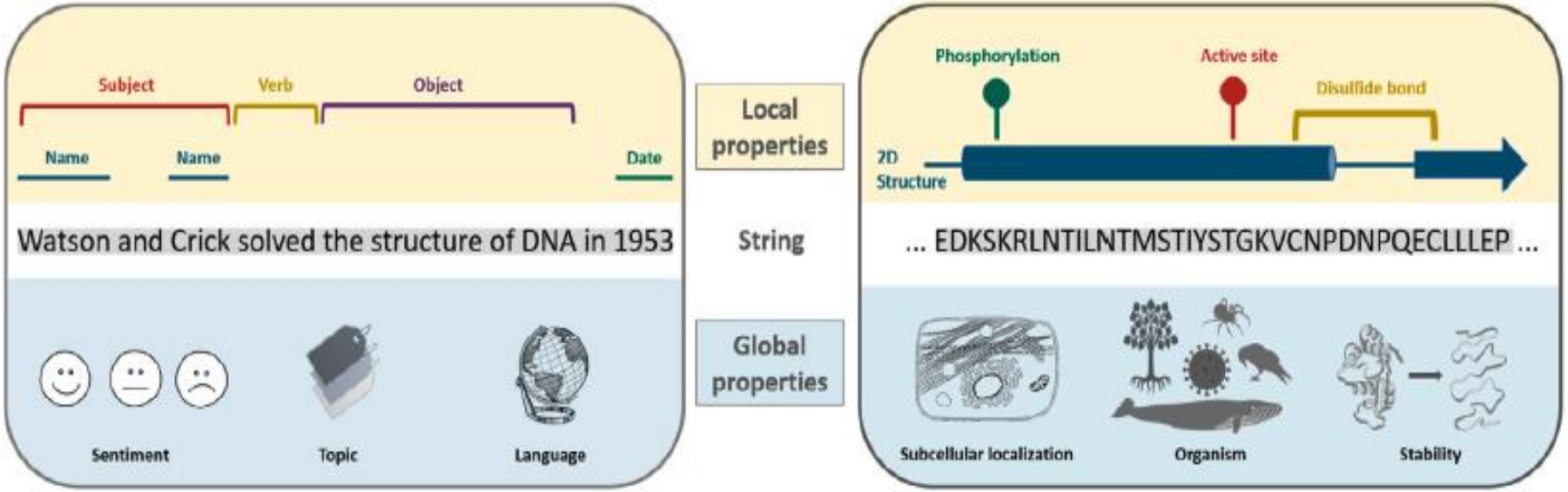
Tokenization – fundamental task in **Natural Language Processing** that yields ***tokens***

Tokens – basic building blocks of the Natural Language

words
n-grams
characters

} the collection of tokens forms a vocabulary

Techniques for NLP have been very naturally extended for bioinformatics tasks



MSTIYSTGKVCNP...

[*start*] [M] [S] [T] [I] [Y] [S] [T] [G] ...

[*start*] [MS] [TI] [YS] [TG] ...

[*start* M] [STI] [YST] [GK] [VCN] ...

A simple global description of the sequence is a ***Bag-of-Words***

$$\boldsymbol{f}_p \in \mathbb{R}^{20}$$

Ofer et al. Comput. Struct. Biotechnol. (**2021**)

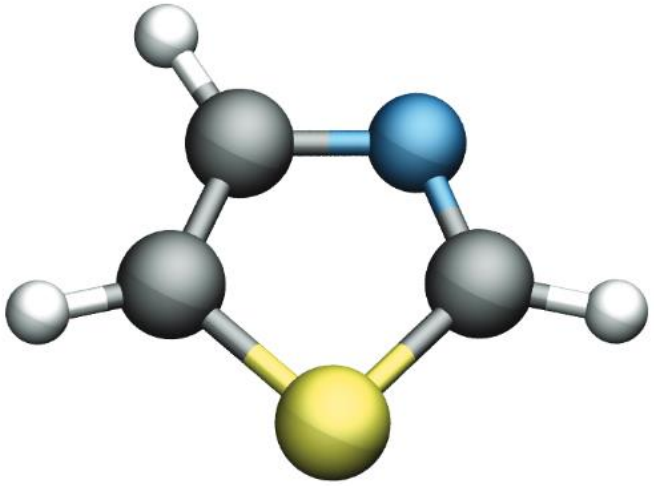
Tokenization and One-Hot Encoding

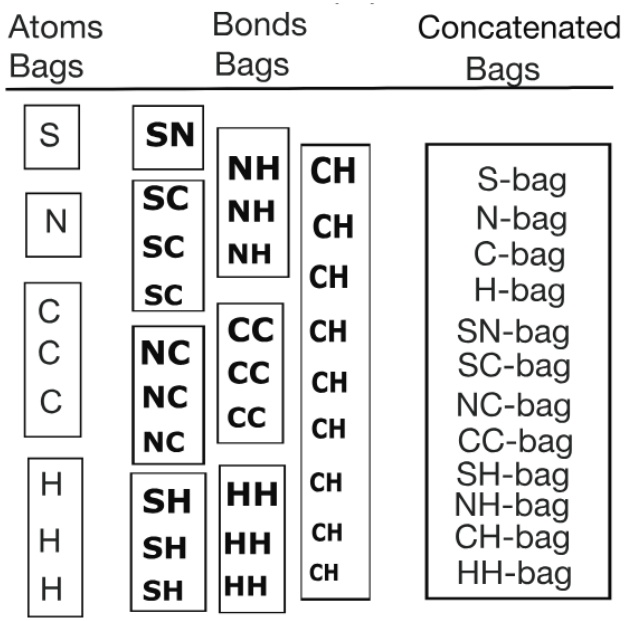
Tokenization – fundamental task in **Natural Language Processing** that yields ***tokens***

Tokens – basic building blocks of the Natural Language

words
n-grams
characters

} the collection of tokens forms a vocabulary





Bags of stuff can serve as reasonably discriminating representations of molecules;

The concatenated bag is *essentially a one-hot vector* with elements equal to the number of occurrences.

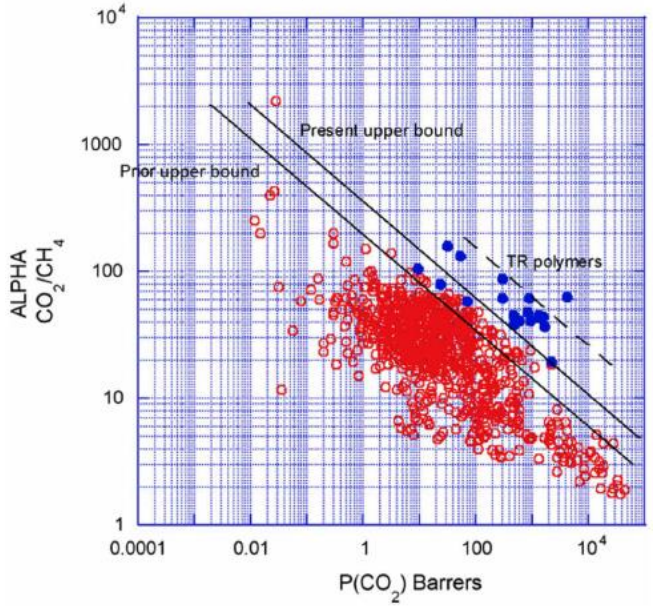
S. Raghunathan, U. D. Priyakumar, Int. J. Quantum Chem. 2022, 122(7), e2687

Tokenization and One-Hot Encoding Notebook

Tokenization and One-Hot Encoding

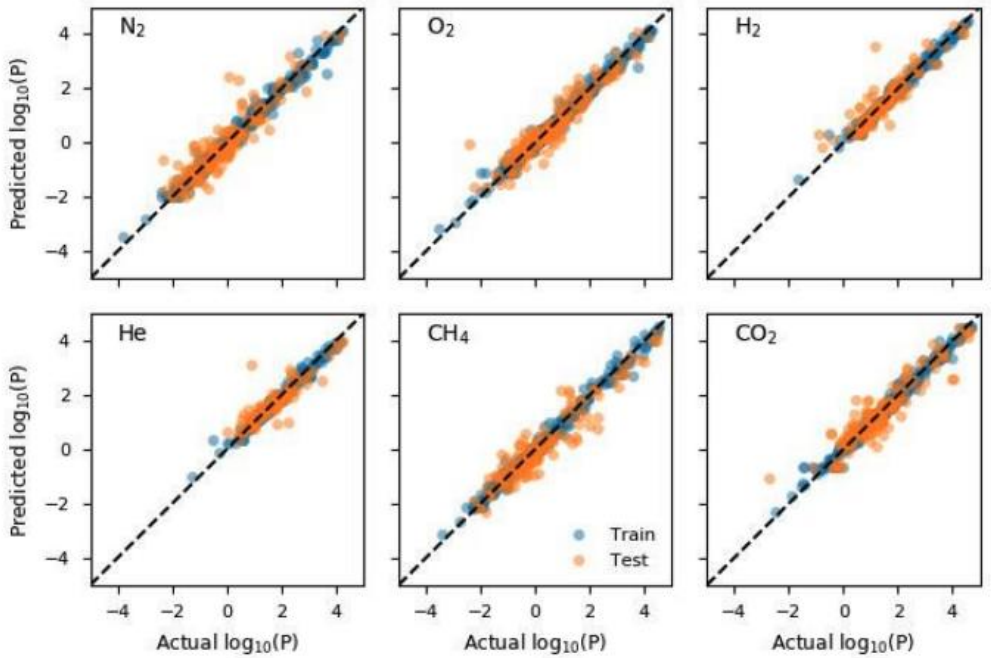
Exercise: *Develop one-hot encoded representations for these atmospheric gases from SMILES*

Gas Permeability in Polymer Membranes



Robeson J. Membr. Sci. **2008**

Separate Machine Learning Models



Barnett et al. Sci. Adv. **2020**

SMILES Conversion

N_2 → N#N
O_2 → O=O
H_2 → [HH]
He → [He]
CH_4 → C
CO_2 → C(=O)=O

Screening for tokens and generating one-hots

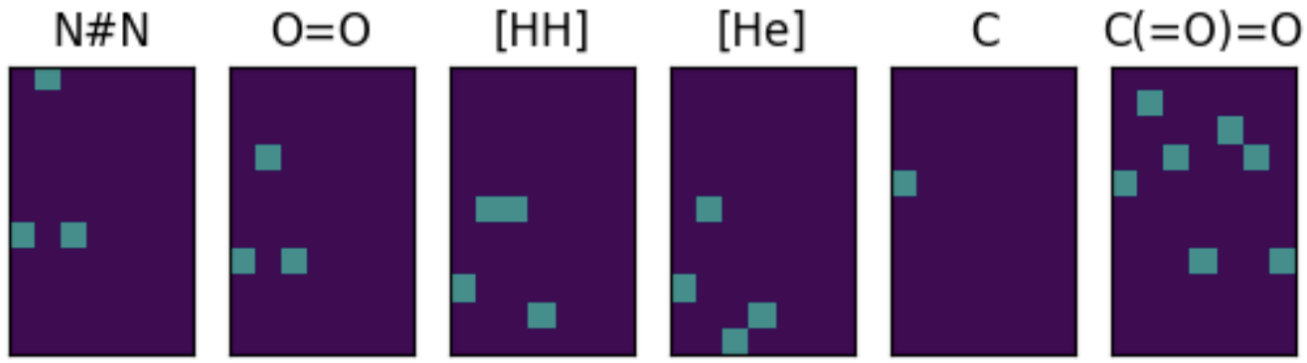
```
[1]  import numpy as np
     mols = ['N#N','O=O','[HH]','[He]','C','C(=O)=O']
```

```
[2]  # Let's determine our vocabulary over the set provided
     # this code does not work for parsing SMILES, in general
     # we just want to keep things simples
     vocab = set()

     # iterate over all molecules to extract unique characters
     M = 0 # to contain maximum number of characters in any string...
     for mol in mols:
         mol = list(mol)
         if len(mol) > M:
             M = len(mol)
         while mol:
             v = mol.pop()
             if v not in vocab:
                 vocab.add(v)
```

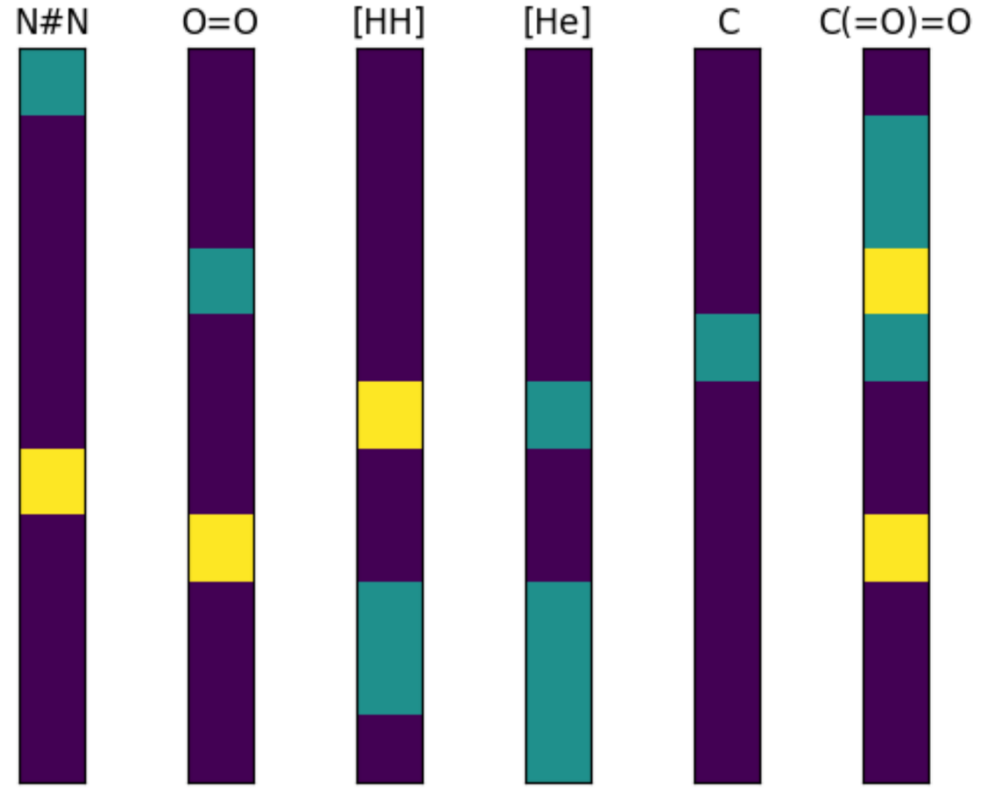
```
# we will create a mapping dictionary that
# points each character to a one-hot vector
N  = len(vocab)
IN = np.eye(N)
charMap = {}
for i,c in enumerate(sorted(list(vocab))):
    charMap[c] = IN[i,:]

''' function that takes a SMILES string
and returns a (padded) array with ohe
vectors as indicated by CMap
M - dimension to pad to if necessary
N - dimension of OHE '''
def Smi2Arr(SmiStr,CMap,M,N):
    Arr   = np.zeros([M,N])
    for i,c in enumerate(list(SmiStr)):
        Arr[i,:] = CMap[c][:]
    return Arr

# create feature arrays
featArr = []
for mol in mols:
    featArr.append(Smi2Arr(mol,charMap,M,N))
```

A new way to view molecules

```
# visualize the encodings
import matplotlib.pyplot as plt
def make_plot(ax,mol,Arr):
    ax.imshow(Arr.T,vmin=0,vmax=2)
    ax.axes.xaxis.set_visible(False)
    ax.axes.yaxis.set_visible(False)
    ax.set_title(mol)
    plt.grid(True,which='minor',color='w')

fig, axs = plt.subplots(1,len(mols))

for i, (mol,Arr) in enumerate(zip(mols,featArr)):
    make_plot(axs[i],mol,Arr)
```




```
# generate compressed feature vectors
featVec = [np.sum(Arr,axis=0) for Arr in featArr]

fig, axs = plt.subplots(1,len(mols))

for i, (mol,Vec) in enumerate(zip(mols,featVec)):
    make_plot(axs[i],mol,Vec[:,np.newaxis].T)
```



Chemically Expressive Representations

Extended Connectivity Fingerprints

A more topologically informed version of OHE are so-called

Extended Connectivity Fingerprints (ECFPs)

Rogers and Hahn. "Extended-Connectivity Fingerprints." *J. Chem. Inf. Model* **2010**

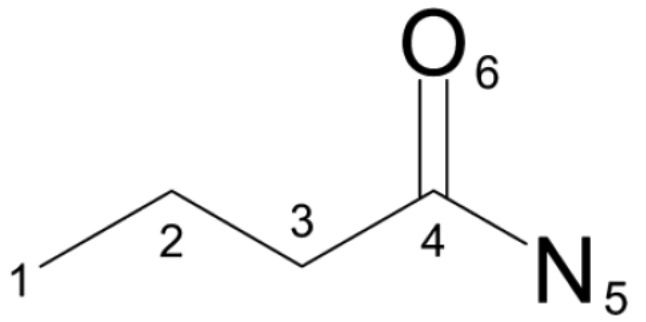
Basic Algorithm

1. **Assign each atom an identifier**
2. Iteratively update identifier based neighboring atoms
3. Remove/count duplicates
4. Fold identifiers into an *N*-bit vector

ECFPs are essentially one-hot encodings over substructures in molecular graphs

- Number of nearest-neighboring heavy atoms
- Number of bonds to heavy atoms
- Atomic Number and Mass
- Number of bonded hydrogens
- Linear or ring topology

Hash function → **integer identifier**



1: 734603939
2: 1559650422
3: 1559650422
4: -1100000244
5: 1572579716
6: -1074141656

Extended Connectivity Fingerprints

A more topologically informed version of OHE are so-called

Extended Connectivity Fingerprints (ECFPs)

Rogers and Hahn. "Extended-Connectivity Fingerprints." *J. Chem. Inf. Model* **2010**

Basic Algorithm

1. Assign each atom an identifier
2. **Iteratively update identifier based neighboring atoms**
3. Remove/count duplicates
4. Fold identifiers into an *N*-bit vector

ECFPs are essentially one-hot encodings over substructures in molecular graphs

Extended Connectivity Fingerprints

A more topologically informed version of OHE are so-called

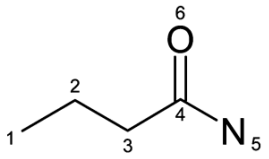
Extended Connectivity Fingerprints (ECFPs)

Rogers and Hahn. "Extended-Connectivity Fingerprints." *J. Chem. Inf. Model* **2010**

Basic Algorithm

1. Assign each atom an identifier
2. Iteratively update identifier based neighboring atoms
3. **Remove/count duplicates**
4. **Fold identifiers into an *N*-bit vector**

- *theoretically, identifiers could be treated as indices of bits in a large set (2^{32})*
- *practically, identifiers are assigned bits in a fixed-size vector*



> <ECFP_0>	> <ECFP_2>	> <ECFP_4>	> <ECFP_6>	
734603939	734603939	734603939	734603939	[A]C
1559650422	1559650422	1559650422	1559650422	[A]C[A]
-1100000244	-1100000244	-1100000244	-1100000244	[A]C(=[A])[A]
1572579716	1572579716	1572579716	1572579716	[A]N
-1074141656	-1074141656	-1074141656	-1074141656	[A]=O
	863188371	863188371	863188371	[A]CC
	-1793471910	-1793471910	-1793471910	[A]CCC
	-1789102870	-1789102870	-1789102870	[A]CCC(=[A])[A]
	-1708545601	-1708545601	-1708545601	[A]CC(=O)N
	-932108170	-932108170	-932108170	[A]C(=[A])N
	2099970318	2099970318	2099970318	[A]C(=O)[A]
		-87618679	-87618679	[A]C(=[A])CCC
		1112638790	1112638790	CCCC(=O)N
		-627599602	-627599602	[A]CCC(=O)N

extracted substructures in SMARTS

Extended Connectivity Fingerprints

A more topologically informed version of OHE are so-called

Extended Connectivity Fingerprints (ECFPs)

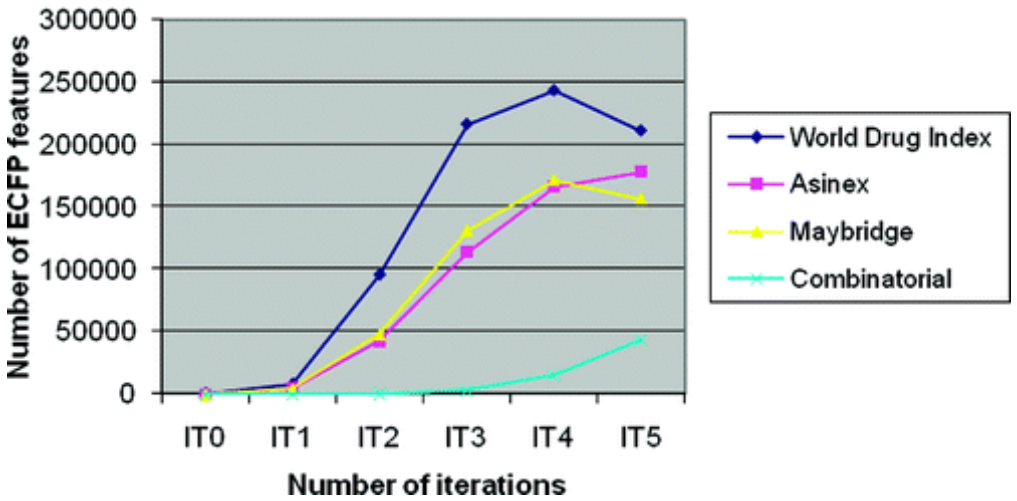
Rogers and Hahn. "Extended-Connectivity Fingerprints." *J. Chem. Inf. Model* **2010**

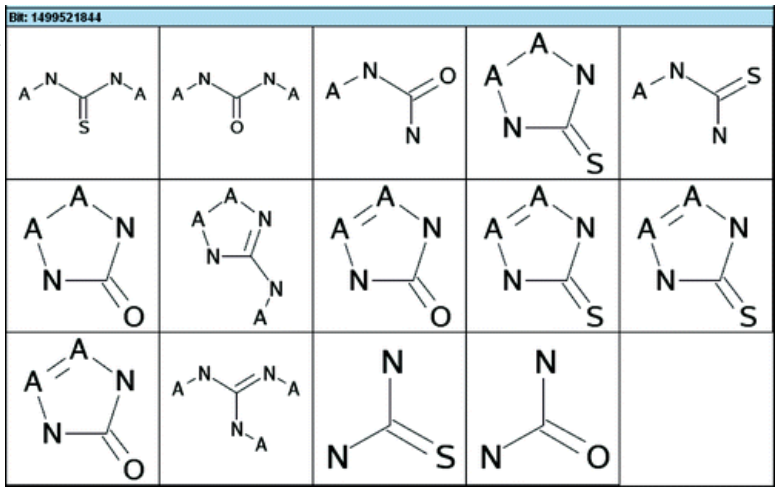
Advantages

- computationally efficient to generate and readily available
- flexible and robust
- expressive of positive/negative structure

Disadvantages

- obfuscates/does not utilize chemical properties or 3D arrangement
- sparse representation
- not scalable to variety of systems
- potentially non-unique identifiers





Descriptor (Feature) Vectors

Vectors of physiochemical properties or system characteristics may also provide good representations of inputs for ML tasks

Examples of Physiochemical Descriptors

e.g.,

- *No. of X structure,…*
- *log P, ASA, shape parameters, …*
- *dipole moment, polarizability, …*
- *electronic energy, Δh_f, IP, ε_{gap}, …*
- *simulation-derived quantities*
- *experimental measurements*

Descriptors can describe local or global characteristics. Some may be readily available or easily obtained, while others can be complicated/expensive to acquire.

Chemical Science

Chem. Sci., 2019, **10**, 6697

Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis†

Yehia Amar,[a] Artur M. Schweidtmann,[b] Paul Deutsch,[c] Liwei Cao[ad] and Alexei Lapkin*[ad]

Table 1 List of solvent molecular descriptors used in this work

	Descriptor (units)	Source
random compilation by some guy	Molecular weight (g mol^{-1})	Stenutz[45]
	Density (g mL^{-1})	Stenutz[45]
	Molar volume (mL mol^{-1})	Stenutz[45]
	Refractive index (—)	Stenutz[45]
	Molecular refractive power (mL mol^{-1})	Stenutz[45]
	Dielectric constant (—)	Stenutz[45]
	Dipole moment (D)	Stenutz[45]
	Melting point (°C)	Stenutz[45]
	Boiling point (°C)	Stenutz[45]
thermodynamic property prediction model	Viscosity (cP)	COSMO*therm*[39]
	ln $P_{octanol-water}$ partition coefficient (—)	COSMO*therm*[39]
	Vapour pressure (mbar)	COSMO*therm*[39]
	Henry's constant of H_2 in solvent (bar)	COSMO*therm*[39]
	ln(γ) activity coefficient of I in solvent (—)	COSMO*therm*[39]
	$\sigma'_1 - \sigma'_3$ profiles segmented into three (—)	COSMO*therm*[39]
	σ_1–σ_5 profiles segmented into five (—)	COSMO*therm*[39]
	t_1–t_4: principal components from PCA (—)	—

Descriptor (Feature) Vectors

Vectors of physiochemical properties or system characteristics may also provide good representations of inputs for ML tasks

Examples of Physiochemical Descriptors

e.g.,

- *No. of X structure,…*
- *log P, ASA, shape parameters, …*
- *dipole moment, polarizability, …*
- *electronic energy, Δh_f, IP, ε_{gap}, …*
- *simulation-derived quantities*
- *experimental measurements*

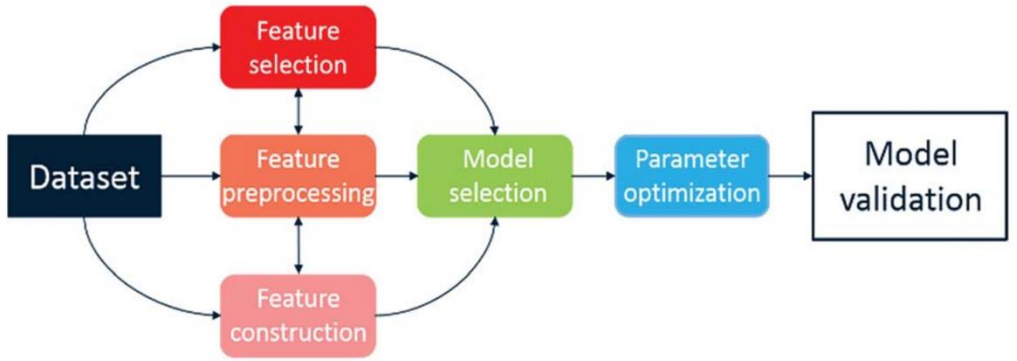
Descriptors can describe local or global characteristics. Some may be readily available or easily obtained, while others can be complicated/expensive to acquire.

Chemical Science

Chem. Sci., 2019, **10**, 6697

Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis†

Yehia Amar,[a] Artur M. Schweidtmann,[b] Paul Deutsch,[c] Liwei Cao[ad] and Alexei Lapkin*[ad]



The selection of descriptors may be part of the model-building process

Descriptor (Feature) Vectors

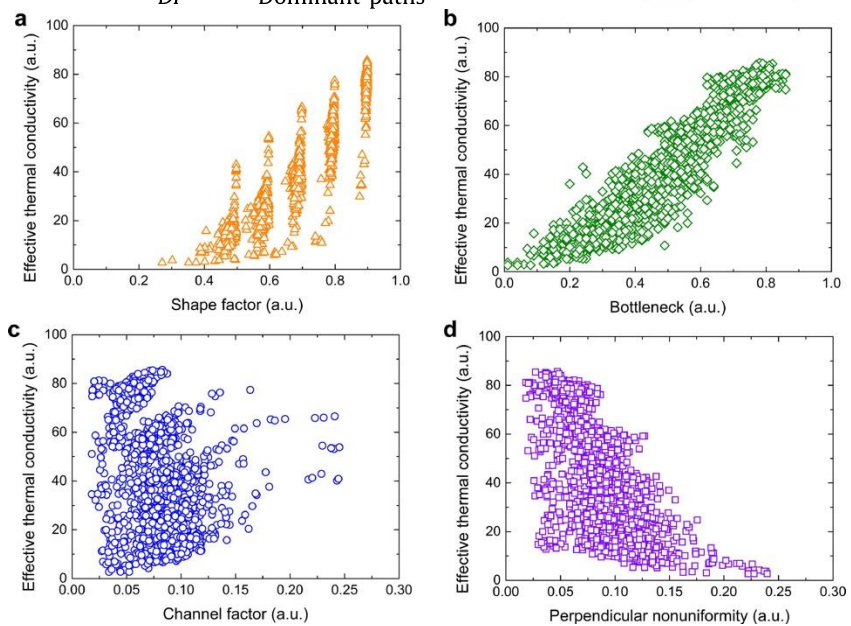
Descriptor vectors can provide suitable representations for non-molecular systems; they are likely tailored to the materials class

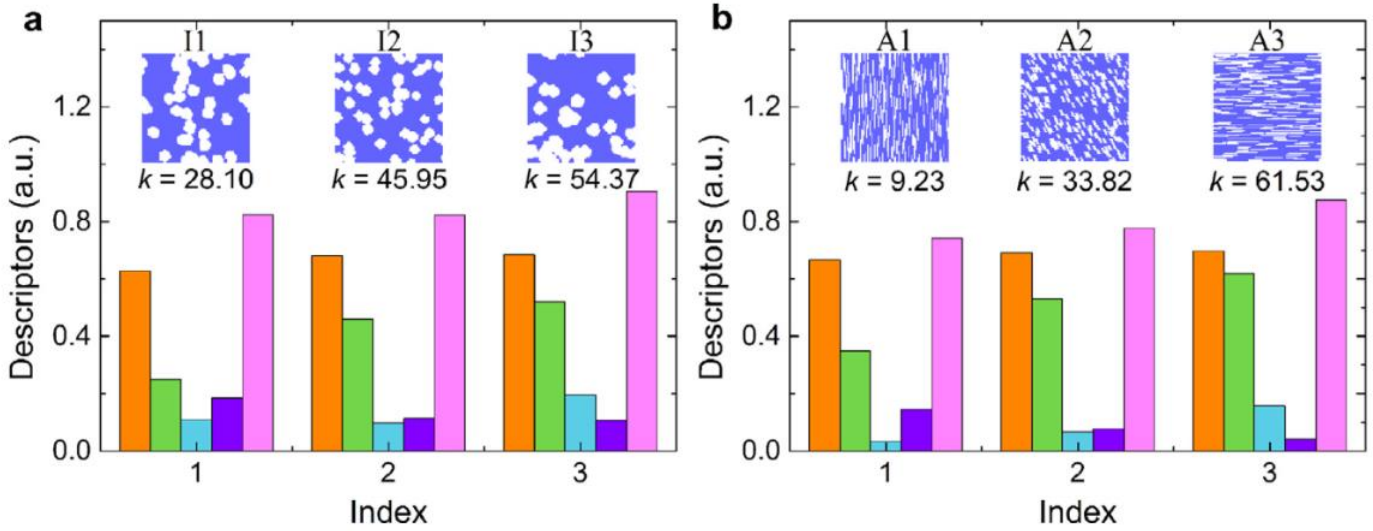
Machine learning prediction of thermal transport in porous media with physics-based descriptors

Han Wei[a], Hua Bao[a,*], Xiulin Ruan[b,*]

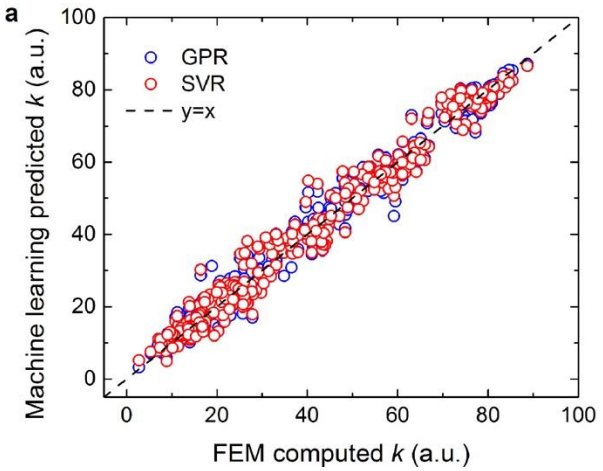
Available online 17 July 2020

ε	Porosity
c_d	Core distribution probability
g_i	Directional growth probability
A_x	matrix cross-section along x direction
A_y	matrix cross-section along y direction
SF	Shape factor
BL	Bottleneck
PN	Perpendicular nonuniformity
CF	Channel factor
DP	Dominant paths





These enable quantitative "structure" to property relationships by machine learning



Descriptor (Feature) Vectors

Descriptor vectors can provide suitable representations for non-molecular systems; they are likely tailored to the materials class

THE JOURNAL OF PHYSICAL CHEMISTRY A

Article

Cite This: J. Phys. Chem. A 2019, 123, 6080–6087

pubs.acs.org/JPCA

A Robust Machine Learning Algorithm for the Prediction of Methane Adsorption in Nanoporous Materials

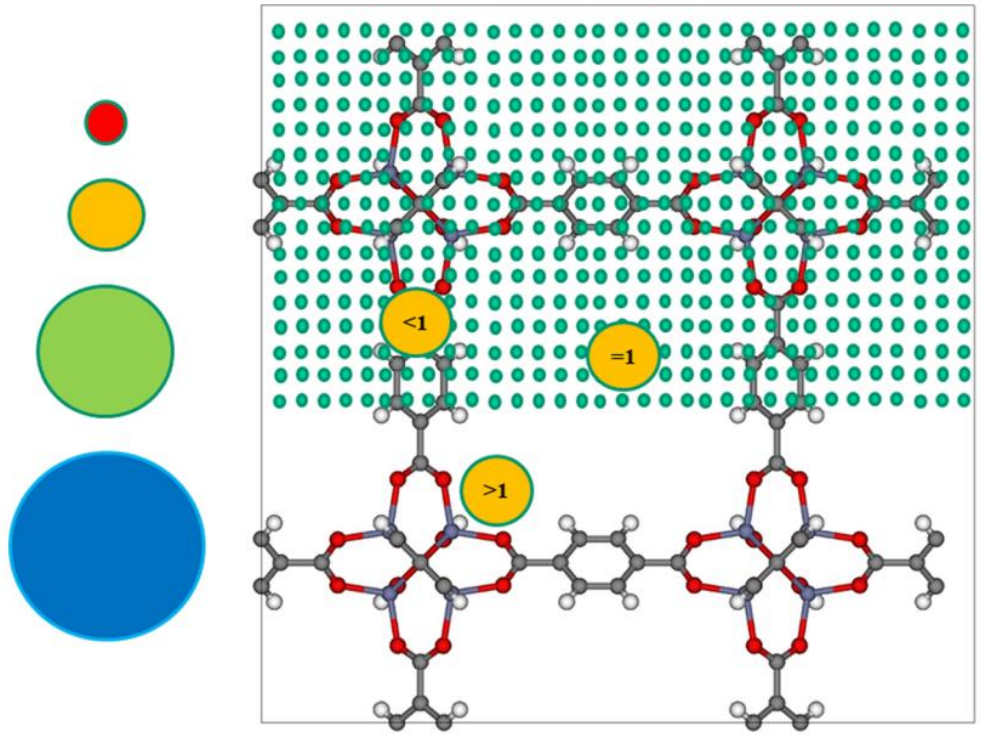
George S. Fanourgakis,*,† Konstantinos Gkagkas,‡ Emmanuel Tylianakis,¶ Emmanuel Klontzas,†,§ and George Froudakis*,†

materials specific

- readily available widely used
- creative easily computed

Table 1. Descriptors Used in the Present Study in the ML Methods[a]

descriptor	minimum	maximum	mean
Standard Structural Features			
void fraction	0.02	0.97	0.43
pore volume ($cm^3\ g^{-1}$)	0.07	7.46	0.49
density ($g\ cm^{-3}$)	0.13	5.18	1.37
grav. surface area ($m^2\ g^{-1}$)	0.0	6832.6	829.4
pore-limiting diameter (Å)	2.4	71.5	4.83
largest cavity diameter (Å)	2.74	71.64	6.79
Probe Atoms			
probe-1 (σ = 2.5 Å, ε/k_B = 50 K)	0.08	5.14	1.12
probe-2 (σ = 3.0 Å, ε/k_B = 50 K)	0.04	12.37	1.53
probe-3 (σ = 3.5 Å, ε/k_B = 50 K)	0.0	36.35	2.30
probe-4 (σ = 4.0 Å, ε/k_B = 50 K)	0.0	131.7	3.79



$$\text{probe} - (a) = \frac{1}{N}\sum_{i=1}^{N} \exp(-\beta E_i^{(a)})$$

Descriptor (Feature) Vectors

Descriptor vectors can provide suitable representations for non-molecular systems; they are likely tailored to the materials class

THE JOURNAL OF PHYSICAL CHEMISTRY A

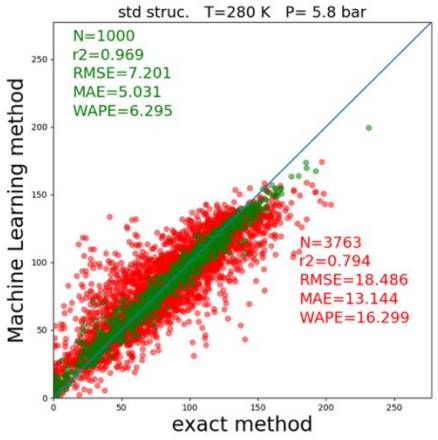
Article

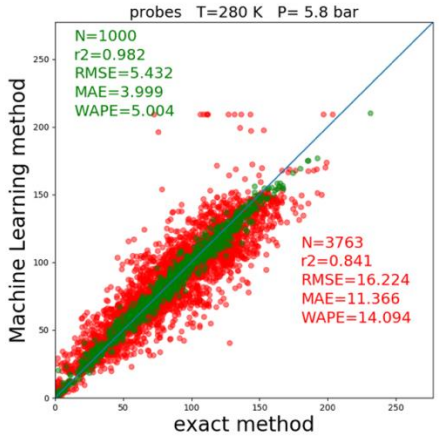
Cite This: J. Phys. Chem. A 2019, 123, 6080–6087

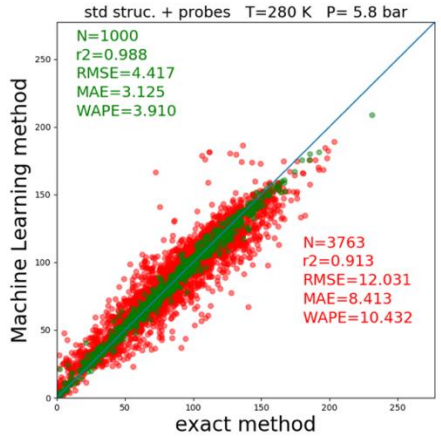
pubs.acs.org/JPCA

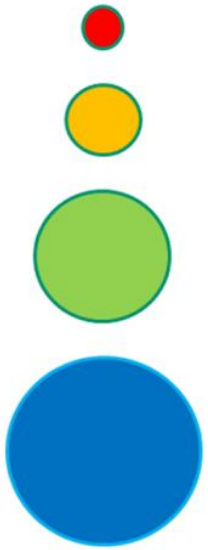
A Robust Machine Learning Algorithm for the Prediction of Methane Adsorption in Nanoporous Materials

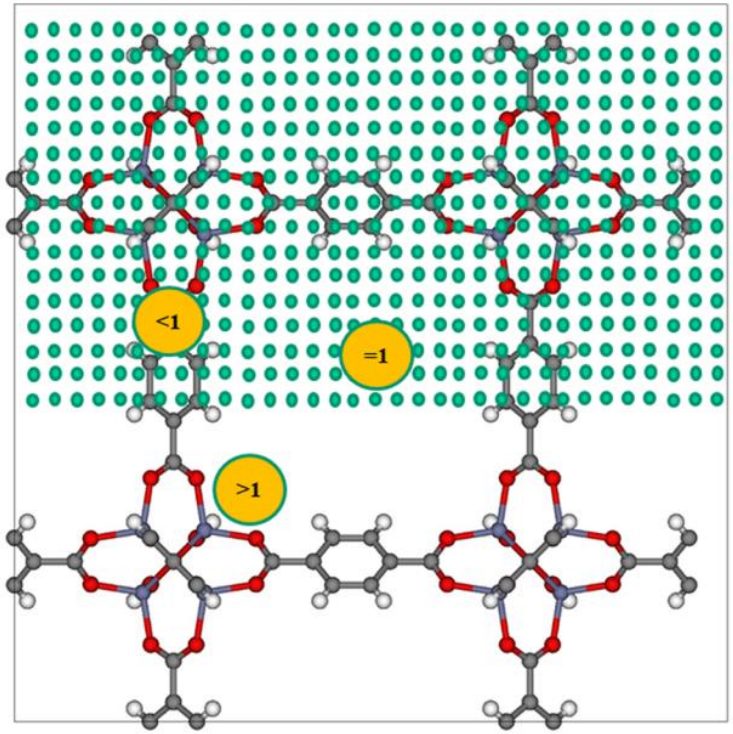
George S. Fanourgakis,*,† Konstantinos Gkagkas,‡ Emmanuel Tylianakis,¶ Emmanuel Klontzas,†,§ and George Froudakis*,†







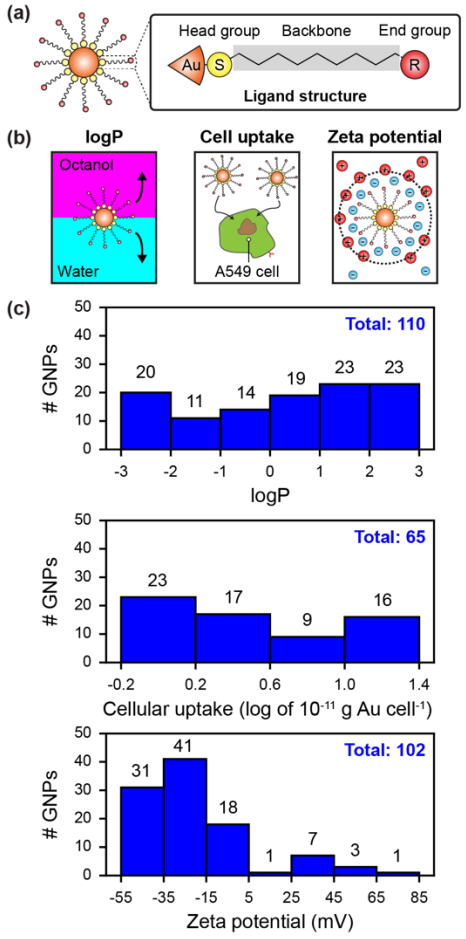




$$\text{probe} - (a) = \frac{1}{N}\sum_{i=1}^{N} \exp(-\beta E_i^{(a)})$$

Simulation-derived Descriptors

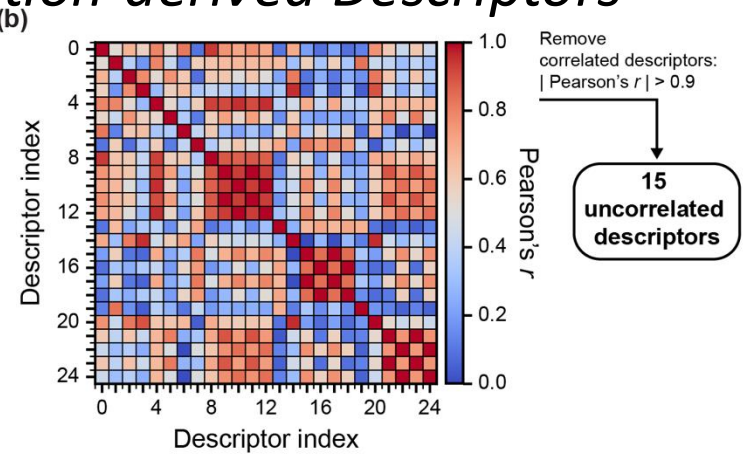
Experimental Dataset

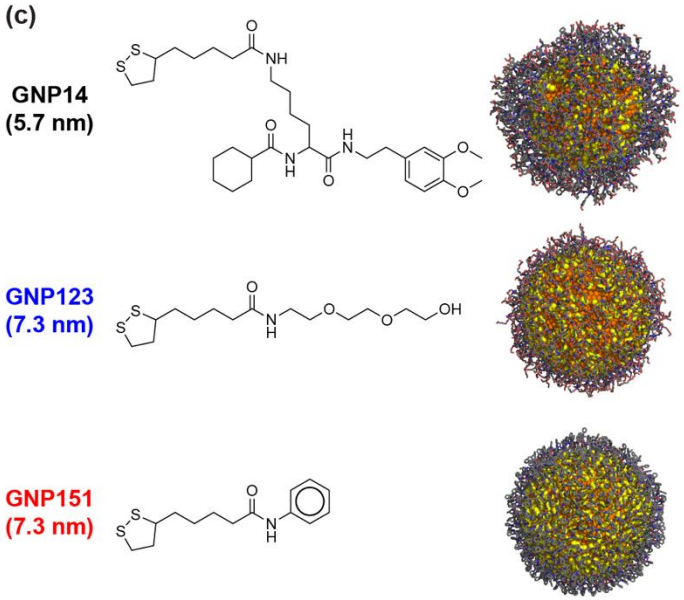


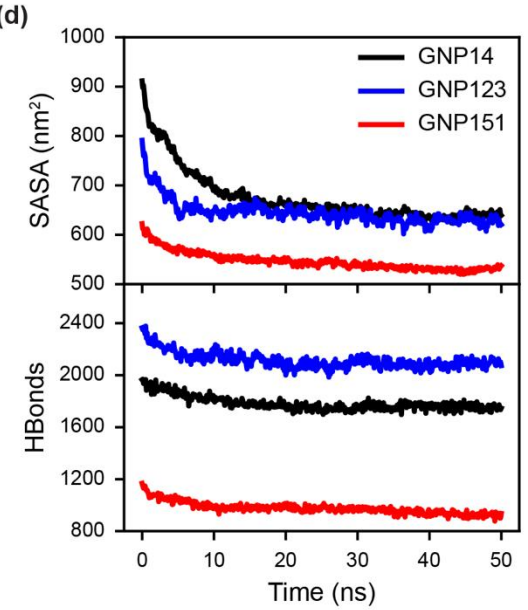
Simulation-derived Descriptors

(a)

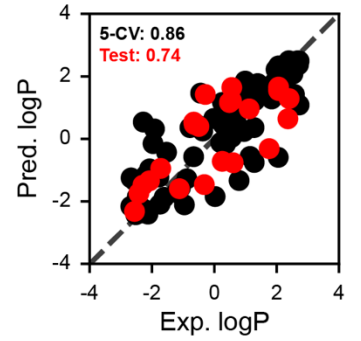
Descriptor	Description
SASA	Solvent accessible surface area
HBonds	#Lig.-water hydrogen bonds
RMSF	Lig. root-mean-squared fluctuation
E(NP-S)	GNP-solvent LJ energies
$\Delta\phi$	Radial electrostatic potential drop
e	Lig. eccentricity (*i.e.* asphericity)
RDF(Au)	Gold radius from RDFs
RDF(Lig.)	Ligand length from RDFs
RDF(H_2O)	Radius at bulk water density from RDFs
R_g	Ligand radius of gyration
⋮	**25 descriptors**

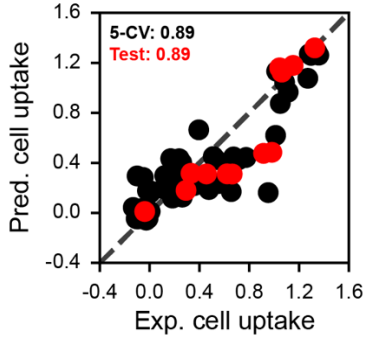


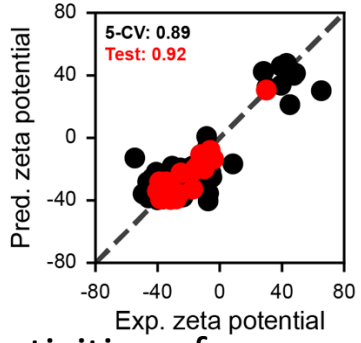




ML Models







Chew, Pedersen, Van Lehn. "Predicting the Physicochemical Properties and Biological Activities of Monolayer-Protected Gold Nanoparticles using Simulation-Derived Descriptors"

Other Structural/Geometric Input Representations

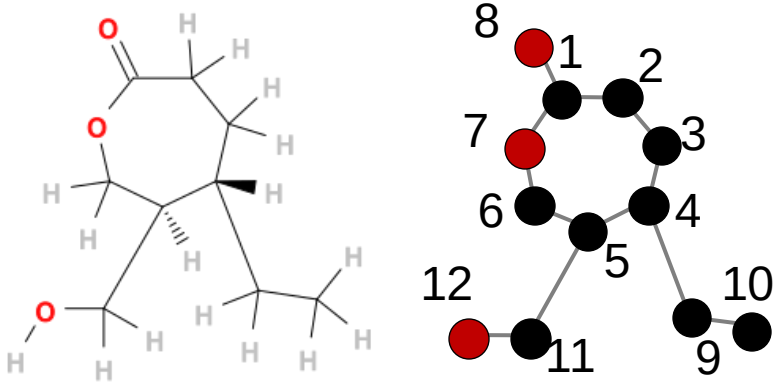
Geometric ***representations*** can provide some additional and potentially critically important information for training ML models

An important consideration for such representations is…
equivariance and ***invariance***

Noether's Theorem – *something about how local and global symmetries relate to conservation laws*

Common considerations:

Translation *Rotation* *Permutation*



$$N, \boldsymbol{r}^N = (\boldsymbol{r}_1, \cdots, \boldsymbol{r}_N)$$
$$\boldsymbol{z}^N = (z_1, \cdots, z_N)$$

Think about shifting the molecule, rotating it, or re-labeling the atoms.

What properties should change? Which shouldn't?

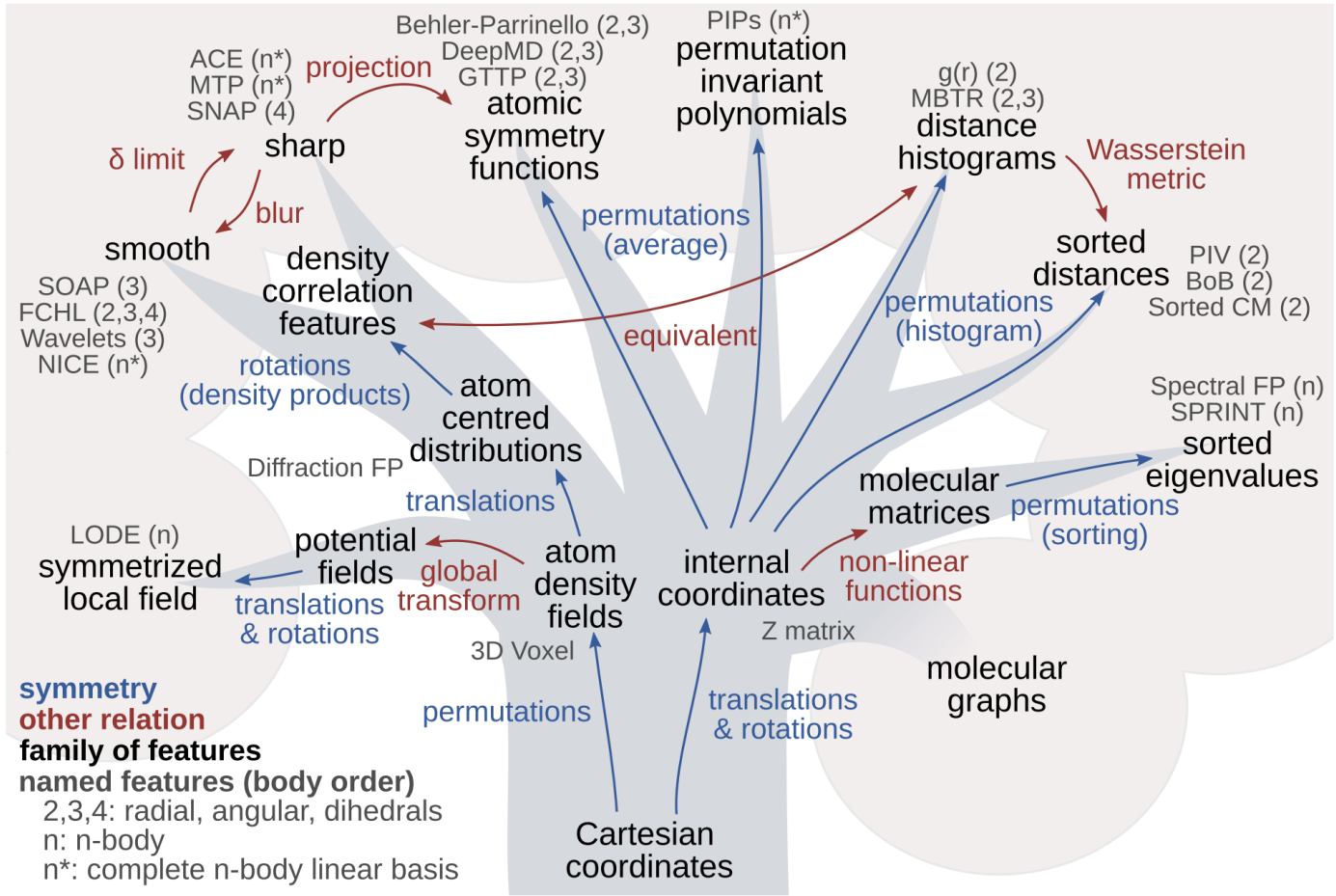
In principle, ML algorithms would be able to "learn" important symmetries; however, that may require substantial training/data. ML is likely to progress much more smoothly if important physics/symmetries are directly encoded in the representation.

Other Structural/Geometric Input Representations

Physics-Inspired Structural Representations for Molecules and Materials

Felix Musil, Andrea Grisafi, Albert P. Bartók, Christoph Ortner, Gábor Csányi, and Michele Ceriotti*

Cite This: *Chem. Rev.* 2021, 121, 9759–9815 Read Online

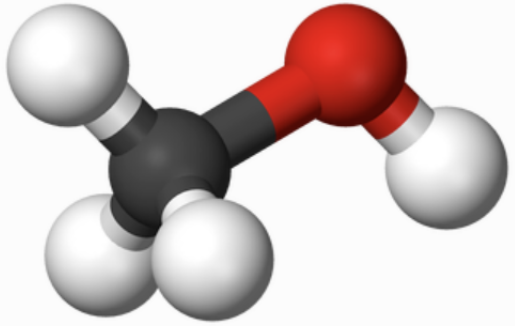


Some Examples

Coulomb Matrix

$$M_{ij}^{\text{Coul}} = \begin{cases} 0.5 z_i^{2.4} \text{ for } i = j \\ \frac{z_i z_j}{r_{ij}} \text{ for } i \neq j \end{cases}$$

- represents entire molecule/fragment



$$\begin{bmatrix} 36.9 & 33.7 & 5.5 & 3.1 & 5.5 & 5.5 \\ 33.7 & 73.5 & 4.0 & 8.2 & 3.8 & 3.8 \\ 5.5 & 4.0 & 0.5 & 0.35 & 0.56 & 0.56 \\ 3.1 & 8.2 & 0.35 & 0.5 & 0.43 & 0.43 \\ 5.5 & 3.8 & 0.56 & 0.43 & 0.5 & 0.56 \\ 5.5 & 3.8 & 0.56 & 0.43 & 0.56 & 0.5 \end{bmatrix}$$

Some Examples

Atom-centered Symmetry Functions

$$f_c(r_{ij}) = \begin{cases} \frac{1}{2}\left[\cos\left(\frac{\pi r_{ij}}{r_c}\right)+1\right] & \text{for } r_{ij} \le r_c \\ 0 & \text{for } r_{ij} > r_c \end{cases}$$

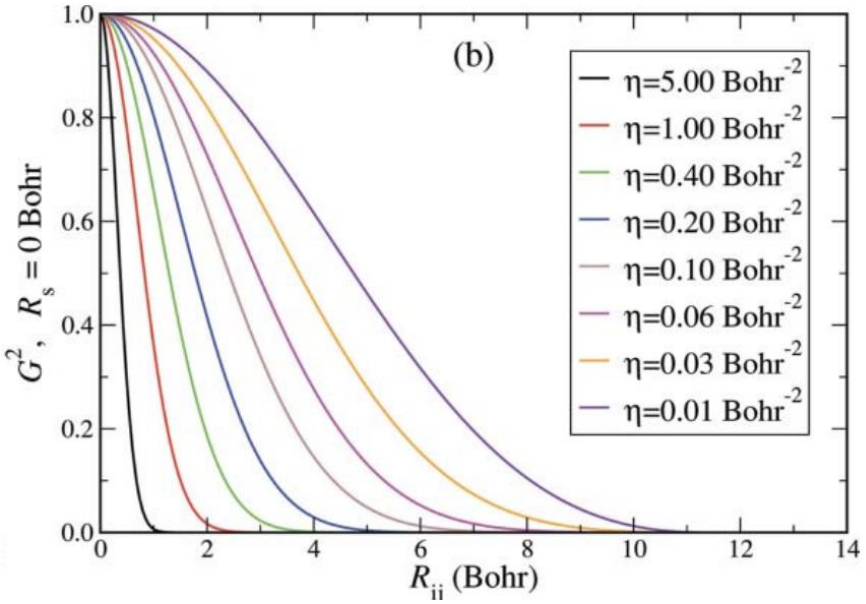
$$G_i^1 = \sum_j f_c(r_{ij})$$

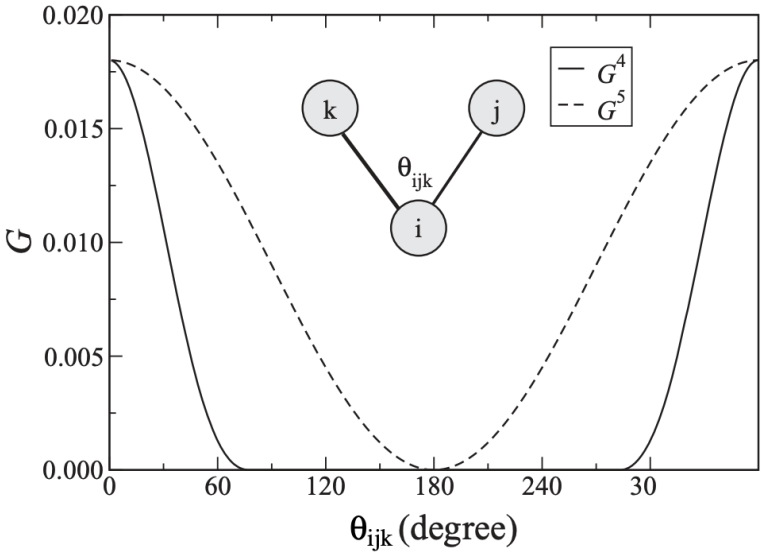
$$G_i^2 = \sum_j e^{-\eta(r_{ij}-r_s)^2} f_c(r_{ij})$$

$$G_i^3 = \sum_j \cos(\kappa r_{ij}) f_c(r_{ij})$$

$$G_i^4 = 2^{1-\zeta} \sum_{j,k\neq i} (1+\lambda\cos\theta_{ijk})^\zeta e^{-\eta(r_{ij}^2+r_{ik}^2+r_{jk})^2} f_c(r_{ij}) f_c(r_{ik}) f_c(r_{jk})$$

$$G_i^5 = 2^{1-\zeta} \sum_{j,k\neq i} (1+\lambda\cos\theta_{ijk})^\zeta e^{-\eta(r_{ij}^2+r_{ik}^2} f_c(r_{ij}) f_c(r_{ik})$$



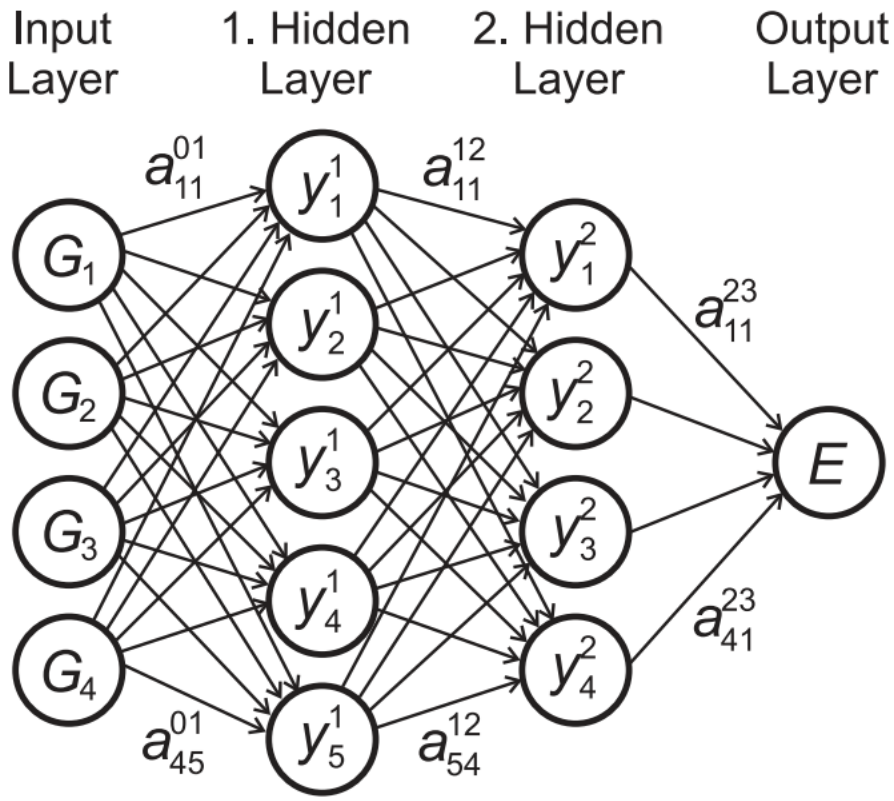


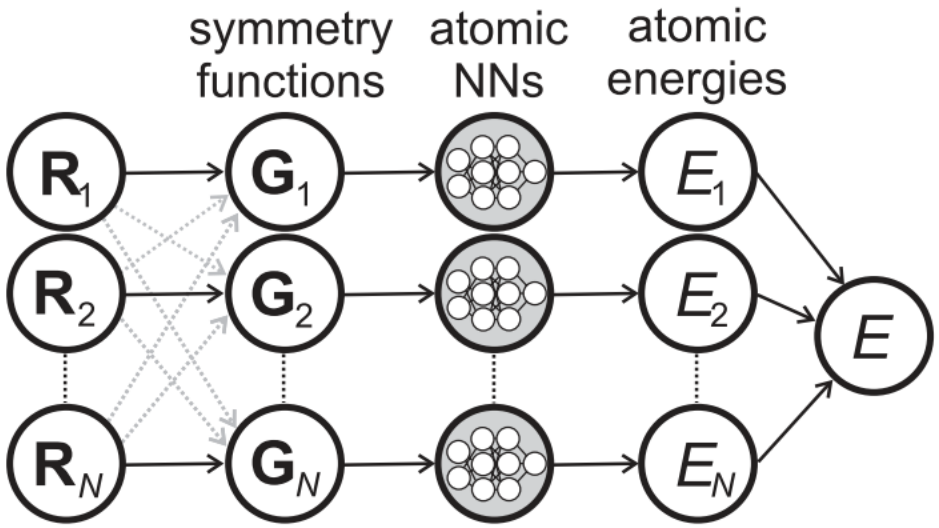
- represents local atom environment
- parameters adjust sensitivity to particular perturbations

Behler *J. Chem. Phys.* **2011**

Some Examples

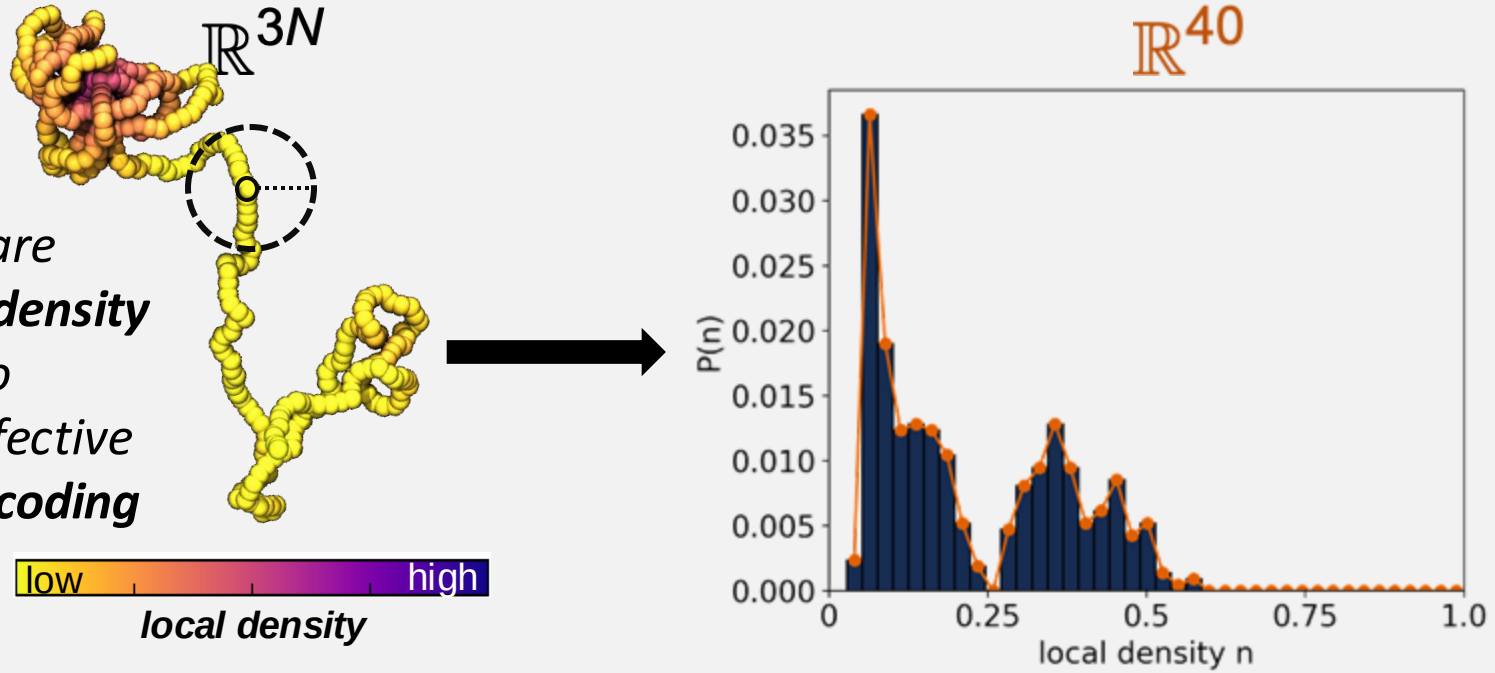
Atom-centered Symmetry Functions: **Application to "Machine Learning Potentials"**





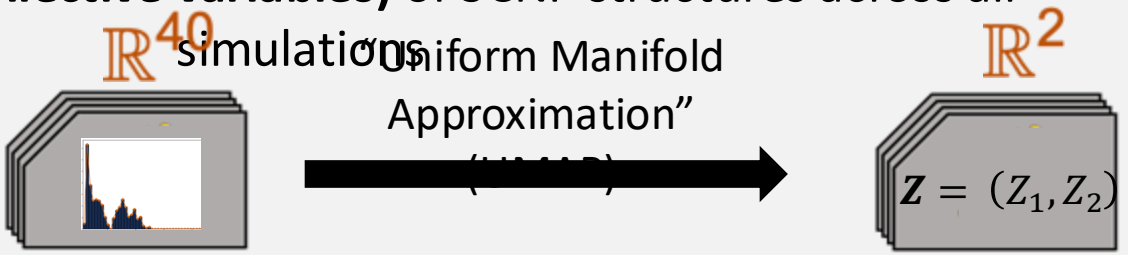
Behler *J. Chem. Phys.* **2011**

Some Examples



Coordinates** are converted to **density histograms** to provide an effective **numerical encoding of chain morphology

Unsupervised ML can then generate an **organized, low-dimensional** representations **(collective variables)** of SCNP structures across all simulations



7,680 SCNPs × 20 snapshots = **153,600 total structures**

Patel, Colmenares, Webb. *ACS Polymers Au.* ***2023***

Some Examples

Broad morphological classification of SCNP structures:

1. Compact globules
2. Tadpole and pearl-necklace
3. Deformable globules and rods