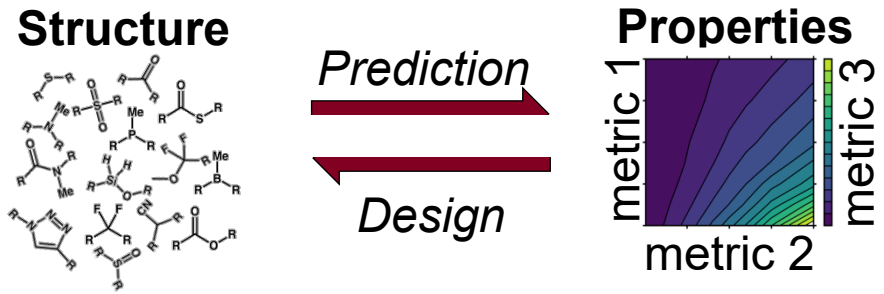
# Molecular Featurization and Chemical Descriptors

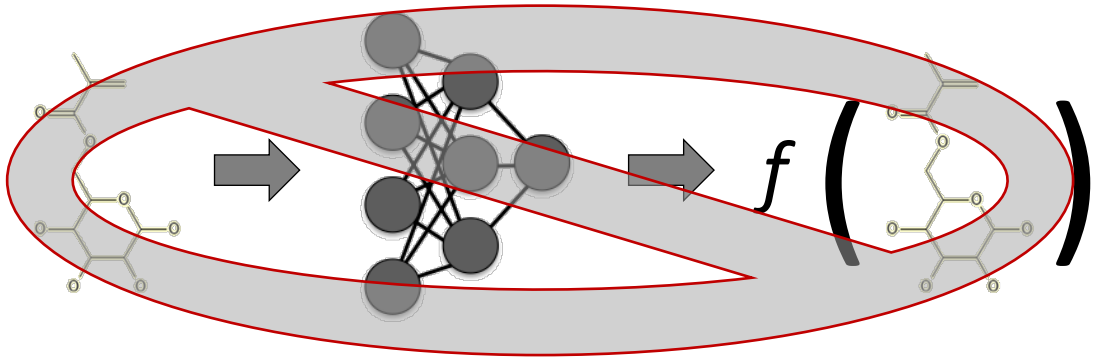
# Machine Learning Meets Molecules

A critical task for utilizing machine learning algorithms is *data representation*

*Consider the goal of developing Quantitative (Chemical) Structure Property Relationships….*




*we (as humans) know how to do this process/goal, and we are interested in using ML to facilitate it*




*However, stick drawing or chemical name is a bit of problem for a neural network or other ML algorithm*

- Need technical methods to convert molecular structures into machine-readable formats (e.g., numerical vectors) that can be processed as inputs to ML algorithms
- The transformation should carry information that is useful to the prediction task in distinguishing amongst chemical moieties

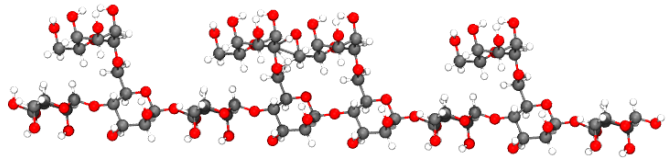
*This conversion process is referred to as **molecular featurization***

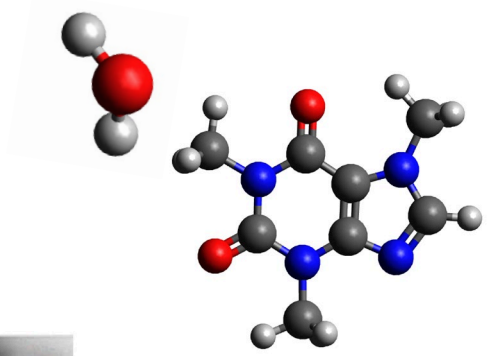
# What is a molecule?

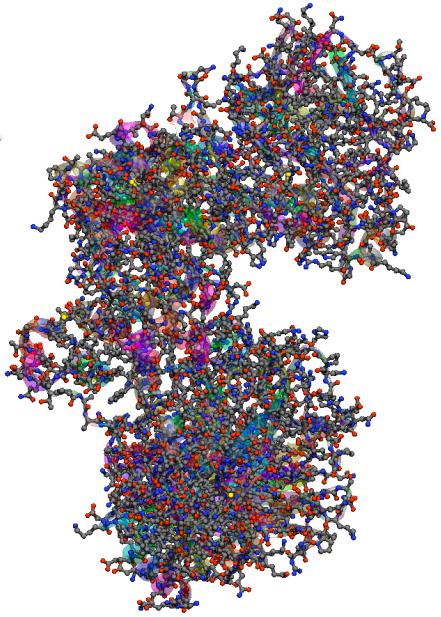
Before we specifically address how to featurize a molecule, we should settle on what a molecule *is* since that may dictate the approach

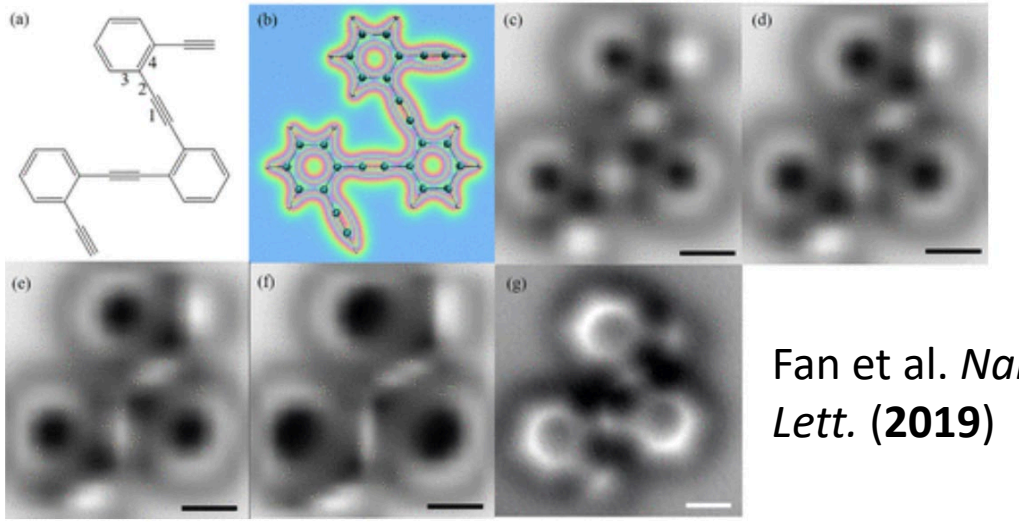
*Some possible characteristics of molecules*

- Groups of atoms that interact/are joined by physical forces?
- Notion of spatially localized electrons → bonds?
- Fundamental units of chemical reactions?
- Typified by specific chemical or physical properties?

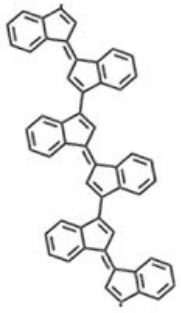


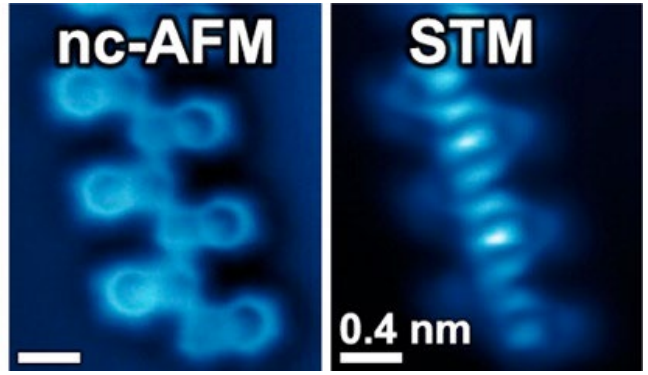







Fan et al. *Nano Lett.* (**2019**)



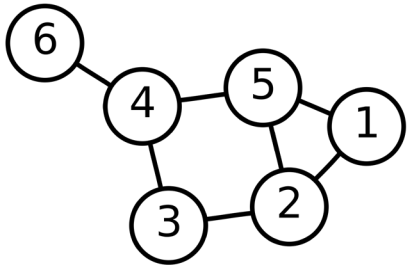



Riss et al. *Nano Lett.* (**2014**)

# Molecular Graphs

The idea of molecules being groups of connected atoms lends itself to representation as a ***graph***

***Mathematical graph***

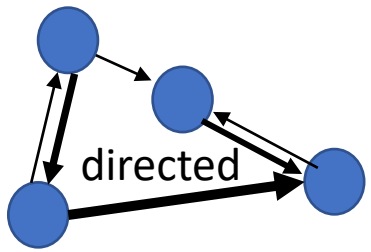



$$\mathcal{G} = (\mathcal{V}, \mathcal{E})$$

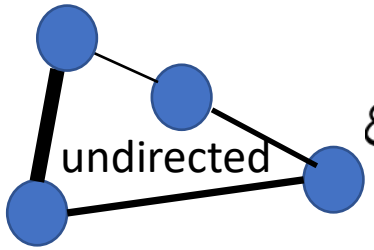
*Structure in discrete mathematics that usually demonstrates how some set of objects are related to one another*

$\mathcal{V}$ ***Vertices/nodes*** – indicates objects
$\mathcal{E}$ ***Edges*** – indicates pairwise relationship amongst objects

graphs may be ***directed*** or ***undirected***




*vs.*




$$\varepsilon_{ij} = \varepsilon_{ji}$$

Graphs can be conveniently represented as matrices

$$\boldsymbol{G}, G_{ij} = \varepsilon_{ij}$$

**e.g.,**

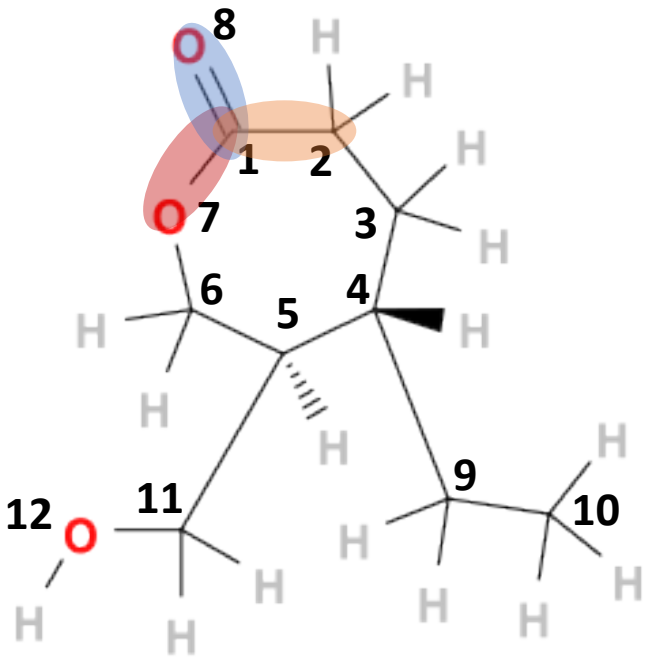
$$\boldsymbol{G} = \begin{bmatrix} 0 & \varepsilon_{12} & 0 & 0 & \varepsilon_{15} & 0 \\ \varepsilon_{21} & 0 & \varepsilon_{23} & 0 & \varepsilon_{25} & 0 \\ 0 & \varepsilon_{32} & 0 & \varepsilon_{34} & 0 & 0 \\ 0 & 0 & \varepsilon_{43} & 0 & \varepsilon_{45} & \varepsilon_{46} \\ \varepsilon_{51} & \varepsilon_{52} & 0 & \varepsilon_{54} & 0 & 0 \\ 0 & 0 & 0 & \varepsilon_{64} & 0 & 0 \end{bmatrix}$$

# Molecular Graphs

The idea of molecules being groups of connected atoms lends itself to representation as a ***graph***

In a molecular graph

Vertices → Atoms or particles
Edges → bonds or interactions




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

***ignoring hydrogens***

Here, entries denote presence of a bond (or not) → often referred to as an Adjacency matrix, but the premise can easily encode other bits of information

Nodes can report properties of each atom (element, charge, hybridization state)

Edges can indicate bond (order), distances, electronic properties, etc.

Typically deficient in the description of 3D structure/conformation, chirality, …

# Text-based Representations

Common starting points for describing molecular graphs are text strings

**Such text strings should…**

- *be human-readable (not necessarily be intuitive)*
- *have well-defined rules to facilitate disambiguation*
- *Ideally possess canonicalization procedures*

***Most popular, pervasive:*** **S**implified **M**olecular-**I**nput **L**ine **E**ntry **S**ystem (SMILES)

**SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules**

DAVID WEININGER
Medicinal Chemistry Project, Pomona College, Claremont, California 91711

Received June 17, 1987
*J. Chem. Inf. Comput. Sci., Vol. 28, No. 1, 1988*

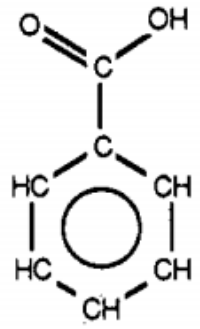
**SMILES. 2. Algorithm for Generation of Unique SMILES Notation**

DAVID WEININGER, ARTHUR WEININGER, and JOSEPH L. WEININGER*
Daylight Chemical Information Systems, Irvine, California 92714

Received May 4, 1988
*J. Chem. Inf. Comput. Sci., Vol. 29, No. 2, 1989*

**Basic Rules**

1. *Atoms indicated by atomic symbols (aromatic rings → lower case)*
2. *Inorganic elements are enclosed by brackets (as are formal charges)*
3. Bonds represented by -,=,#, and : *(single, double, triple, and aromatic); single and aromatic bonds are conventionally omitted*
4. *Branches are specified by enclosures in parentheses*
5. *Cyclic structures are indicated by breaking one bond in each ring and designating the point of opening/closure with a digit*




*not natively canonical (additional algorithms do this)*

c1ccccc1C(=0)O

# Text-based Representations

Common starting points for describing molecular graphs are text strings

***An extension:*** **SM**ILES **Ar**bitrary **T**arget **S**pecification (SMARTS)

**SMARTS** is not for representing molecular structures but ***chemical patterns***

**All SMILES** → **valid SMARTS**

**valid SMILES** ⊘ **All SMARTS**

→ database queries

→ substructure searches *(finding a subgraph of the molecular graph)*

**Decoding Exercise**

*draw out/describe the substructures from SMARTS*

- "cc"
- [c,n;H1]
- "Caa(O)aN"
- "Ca(aO)aaN"

**SMARTS Atomic Primitives**

| Symbol | Symbol name | Atomic property requirements | Default |
|---|---|---|---|
| * | wildcard | any atom | (no default) |
| a | aromatic | aromatic | (no default) |
| A | aliphatic | aliphatic | (no default) |
| D<n> | degree | <n> explicit connections | exactly one |
| H<n> | total-H-count | <n> attached hydrogens | exactly one[1] |
| h<n> | implicit-H-count | <n> implicit hydrogens | at least one |
| R<n> | ring membership | in <n> SSSR rings | any ring atom |
| r<n> | ring size | in smallest SSSR ring of size <n> | any ring atom[2] |
| v<n> | valence | total bond order <n> | exactly one[2] |
| X<n> | connectivity | <n> total connections | exactly one[2] |
| x<n> | ring connectivity | <n> total ring connections | at least one[2] |
| - <n> | negative charge | -<n> charge | -1 charge (-- is -2, etc) |
| +<n> | positive charge | +<n> formal charge | +1 charge (++ is +2, etc) |
| #n | atomic number | atomic number <n> | (no default)[2] |
| @ | chirality | anticlockwise | anticlockwise, default class[2] |
| @@ | chirality | clockwise | clockwise, default class[2] |
| @<c><n> | chirality | chiral class <c> chirality <n> | (nodefault) |
| @<c><n>? | chiral or unspec | chirality <c><n> or unspecified | (no default) |
| <n> | atomic mass | explicit atomic mass | unspecified mass |

**SMARTS Bond Primitives**

| Symbol | Atomic property requirements |
|---|---|
| - | single bond (aliphatic) |
| / | directional bond "up"[1] |
| \ | directional bond "down"[1] |
| /? | directional bond "up or unspecified" |
| \? | directional bond "down or unspecified" |
| = | double bond |
| # | triple bond |
| : | aromatic bond |
| ~ | any bond (wildcard) |
| @ | any ring bond[1] |

**SMARTS Logical Operators**

| Symbol | Expression | Meaning |
|---|---|---|
| exclamation | !e1 | not e1 |
| ampersand | e1&e2 | a1 and e2 (high precedence) |
| comma | e1,e2 | e1 or e2 |
| semicolon | e1;e2 | a1 and e2 (low precedence) |

# Text-based Representations

Common starting points for describing molecular graphs are text strings

***An extension:*** **SM**ILES **Ar**bitrary **T**arget **S**pecification (SMARTS)

**SMARTS** is not for representing molecular structures but ***chemical patterns***

**All SMILES** → **valid SMARTS**

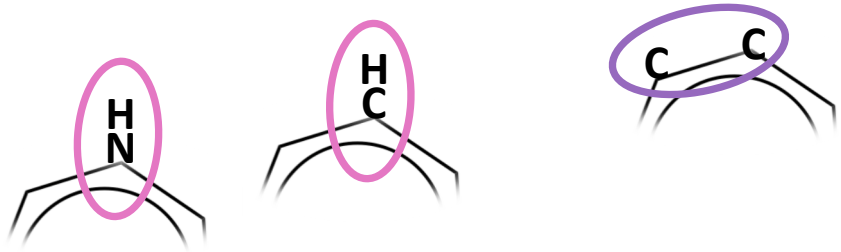
**valid SMILES** ← **All SMARTS**

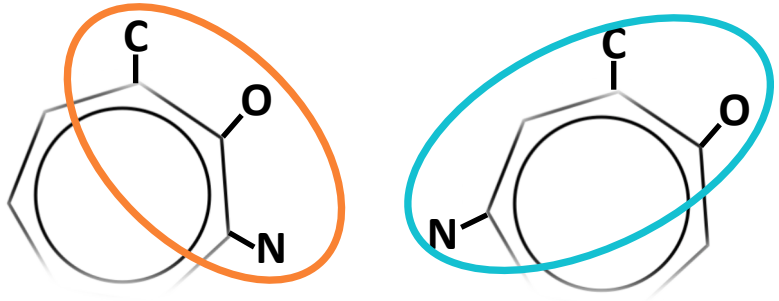
→ database queries

→ substructure searches *(finding a subgraph of the molecular graph)*

**Decoding Exercise**

*draw out/describe the substructures from SMARTS*




any pair of bonded aromatic carbons

- "CC"

either aromatic carbon or nitrogen and exactly one hydrogen

- [c,n;H1]

substituent carbon of aromatic ring that is ortho to substituent oxygen and meta to substituent nitrogen

- "Caa(O)aN"

same as above but O and N likely para

- "Ca(aO)aaN"

# Text-based Representations

Common starting points for describing molecular graphs are text strings

***DeepSMILES*** was developed to address some syntactic issues in using SMILES for "generative" models. The gist of the problem is that many perturbations to SMILES strings do not result in valid molecules.

| SMILES | DeepSMILES |
|---|---|
| C1CCCC1 | CCCCC5 |
| C1CCCCCCCCC1 | CCCCCCCCCC%10 |
| C(O)C | CO)C |
| C(OF)C | COF))C |
| C(F)(F)C | CF)F)C |
| C(=O)Cl | C=O)Cl |
| C(OC(=O)Cl)I | COC=O)Cl)))I |
| C1CC(OC)CC1 | CCCOC))CC5 |
| C1=C/CCCCCC/1 | C=C/CCCCCC/8 |
| C\1=C/CCCCCC1 | C=C/CCCCCC/8 |
| B(c1ccccc1)(O)O | Bcccccc6)))))O)O |
| Cn1cccc-2nccc12 | Cncccccnccc9-5 |
| C1N[C@@]12CO2 | CN[C@@]3CO3 |
| [C@@]12(NC1)CO2 | [C@@]NC3))CO3 |
| CC1CCCO[C@]21CCCCO2 | CCCCCO[C@@]6CCCCO6 |
| CC1CCCO[C@@]12CCCCO2 | CCCCCO[C@@]6CCCCO6 |
| NC[C@]12CCCC1C3CC2CC3 | NC[C@]CCCC5CCC8CC5 |
| NC[C@]12CCCC2C3CC1CC3 | NC[C@@]CCCC5CCC8CC5 |

10.26434/chemrxiv.7097960.v1

**Other string-based representations**

*Wiswesser line notation*

*SYBYL Line Notation*

*IUPAC **In**ternational **Ch**emical **I**dentifier*

# Text-based Representations

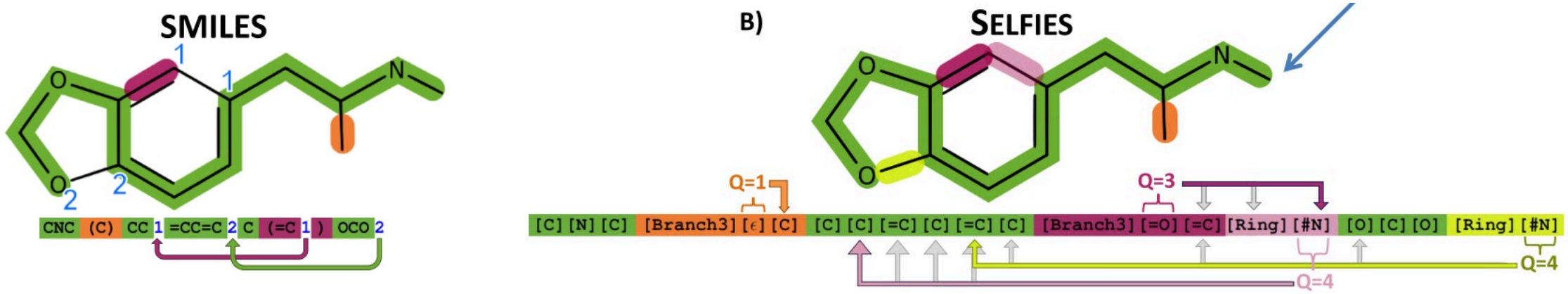
Common starting points for describing molecular graphs are text strings

**Self**-referencing **E**mbedded **S**trings (SELFIES)

Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation

Mario Krenn[1,2,3], Florian Häse[1,2,3,4], AkshatKumar Nigam[2], Pascal Friederich[2,5] and Alan Aspuru-Guzik[1,2,3,6]

*Mach. Learn.: Sci. Technol.* **1** (2020) 045024




- New kid on the block with growing utility
- developed as a “100% robust” alternative to SMILES:
  - *every SELFIES string is a valid molecule*
  - *every molecule has a SELFIES*

# Text-based Representations

Common starting points for describing molecular graphs are text strings

**Self**-referencing **E**mbedded **S**trings (SELFIES)

Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation

Mario Krenn[1,2,3], Florian Häse[1,2,3,4], AkshatKumar Nigam[2], Pascal Friederich[2,5] and Alan Aspuru-Guzik[1,2,3,6]

*Mach. Learn.: Sci. Technol.* **1** (2020) 045024

## Formal Grammar Rules

Start in $\mathbf{X}_0$

Rule Vectors

State of Derivation

| | [ϵ] | [F] | [=O] | [#N] | [O] | [N] | [=N] | [C] | [=C] | [#C] | [Branch1] | [Branch2] | [Branch3] | [Ring] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mathbf{X}_0$ → | $\mathbf{X}_0$ | F $\mathbf{X}_1$ | O $\mathbf{X}_2$ | N $\mathbf{X}_3$ | O $\mathbf{X}_2$ | N $\mathbf{X}_3$ | N $\mathbf{X}_3$ | C $\mathbf{X}_4$ | C $\mathbf{X}_4$ | C $\mathbf{X}_4$ | ign $\mathbf{X}_0$ | ign $\mathbf{X}_0$ | ign $\mathbf{X}_0$ | ign $\mathbf{X}_0$ |
| $\mathbf{X}_1$ → | ϵ | F | O | N | O $\mathbf{X}_1$ | N $\mathbf{X}_2$ | N $\mathbf{X}_2$ | C $\mathbf{X}_3$ | C $\mathbf{X}_3$ | C $\mathbf{X}_3$ | ign $\mathbf{X}_1$ | ign $\mathbf{X}_1$ | ign $\mathbf{X}_1$ | R(Q) |
| $\mathbf{X}_2$ → | ϵ | F | =O | =N | O $\mathbf{X}_1$ | N $\mathbf{X}_2$ | =N $\mathbf{X}_1$ | C $\mathbf{X}_3$ | =C $\mathbf{X}_2$ | =C $\mathbf{X}_2$ | B(Q,$\mathbf{X}_5$)$\mathbf{X}_1$ | B(Q,$\mathbf{X}_5$)$\mathbf{X}_1$ | B(Q,$\mathbf{X}_5$)$\mathbf{X}_1$ | R(Q) $\mathbf{X}_1$ |
| $\mathbf{X}_3$ → | ϵ | F | =O | #N | O $\mathbf{X}_1$ | N $\mathbf{X}_2$ | =N $\mathbf{X}_1$ | C $\mathbf{X}_3$ | =C $\mathbf{X}_2$ | #C $\mathbf{X}_1$ | B(Q,$\mathbf{X}_5$)$\mathbf{X}_2$ | B(Q,$\mathbf{X}_6$)$\mathbf{X}_1$ | B(Q,$\mathbf{X}_5$)$\mathbf{X}_2$ | R(Q) $\mathbf{X}_2$ |
| $\mathbf{X}_4$ → | ϵ | F | =O | #N | O $\mathbf{X}_1$ | N $\mathbf{X}_2$ | =N $\mathbf{X}_1$ | C $\mathbf{X}_3$ | =C $\mathbf{X}_2$ | #C $\mathbf{X}_1$ | B(Q,$\mathbf{X}_5$)$\mathbf{X}_3$ | B(Q,$\mathbf{X}_7$)$\mathbf{X}_1$ | B(Q,$\mathbf{X}_6$)$\mathbf{X}_2$ | R(Q) $\mathbf{X}_3$ |
| $\mathbf{X}_5$ → | C | F | O | N | O $\mathbf{X}_1$ | N $\mathbf{X}_2$ | N $\mathbf{X}_2$ | C $\mathbf{X}_3$ | C $\mathbf{X}_3$ | C $\mathbf{X}_3$ | $\mathbf{X}_5$ | $\mathbf{X}_5$ | $\mathbf{X}_5$ | $\mathbf{X}_5$ |
| $\mathbf{X}_6$ → | C | F | =O | =N | O $\mathbf{X}_1$ | N $\mathbf{X}_2$ | =N $\mathbf{X}_1$ | C $\mathbf{X}_3$ | =C $\mathbf{X}_2$ | =C $\mathbf{X}_2$ | $\mathbf{X}_6$ | $\mathbf{X}_6$ | $\mathbf{X}_6$ | $\mathbf{X}_6$ |
| $\mathbf{X}_7$ → | C | F | =O | #N | O $\mathbf{X}_1$ | N $\mathbf{X}_2$ | =N $\mathbf{X}_1$ | C $\mathbf{X}_3$ | =C $\mathbf{X}_2$ | #C $\mathbf{X}_1$ | $\mathbf{X}_7$ | $\mathbf{X}_7$ | $\mathbf{X}_7$ | $\mathbf{X}_7$ |
| Q → | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |

Derivation Rules

*conversion to molecular graph*

[F][=C][=C][#N]

1. Start in $\mathbf{X}_0$ → F $\mathbf{X}_1$
2. F $\mathbf{X}_1$ → F C $\mathbf{X}_3$
3. F C $\mathbf{X}_3$ → F C = C $\mathbf{X}_2$
4. F C = C $\mathbf{X}_2$ → F C = C = N

# Tokenization and One-Hot Encoding

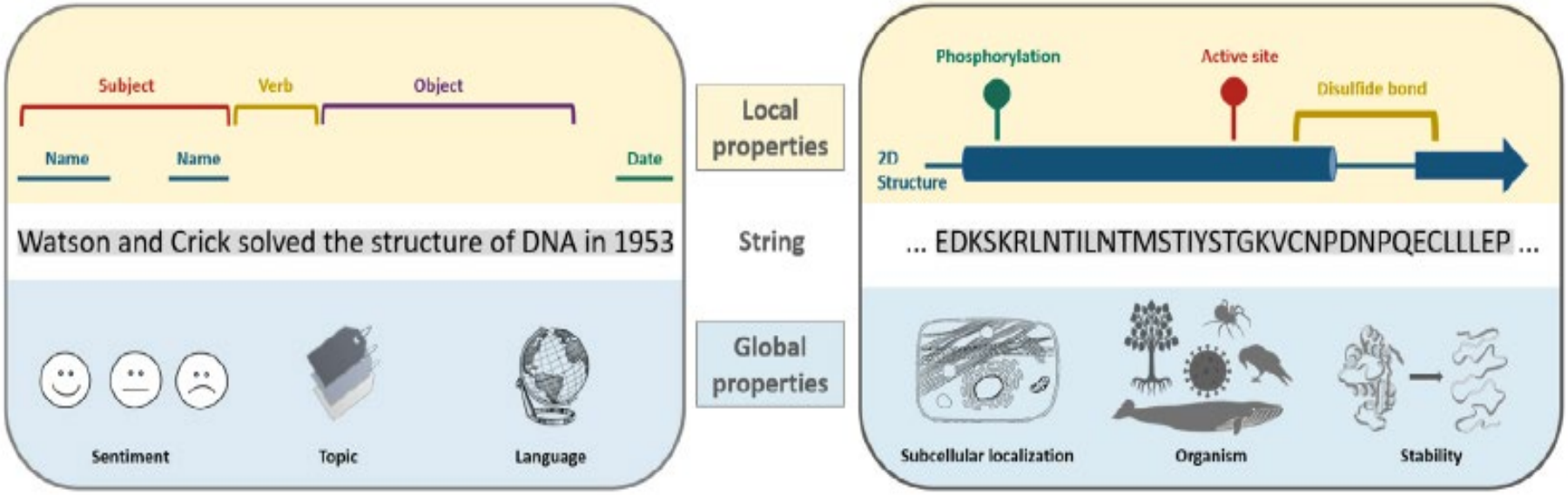
***Tokenization –*** fundamental task in **Natural Language Processing** that yields ***tokens***

***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




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

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***


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

# Tokenization and One-Hot Encoding

**One viable approach for molecules:**

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

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

e.g.,

**cat**

01100011011000010111010**0**

**100**

**rab**bit

01110010011000010110001**0**

**010**

**duc**k

01100100011101010110001**1**

**001**

**English; binary; numerical assignment; one-hot vector**

# Tokenization and One-Hot Encoding

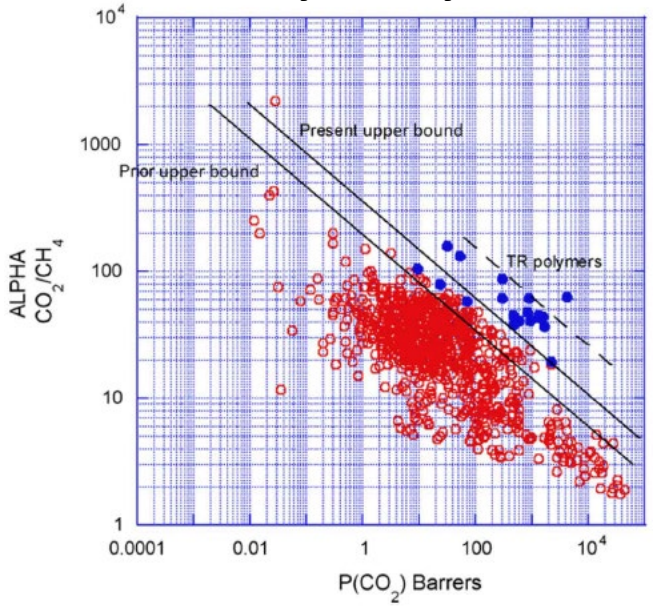
**One viable approach for molecules:**

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

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

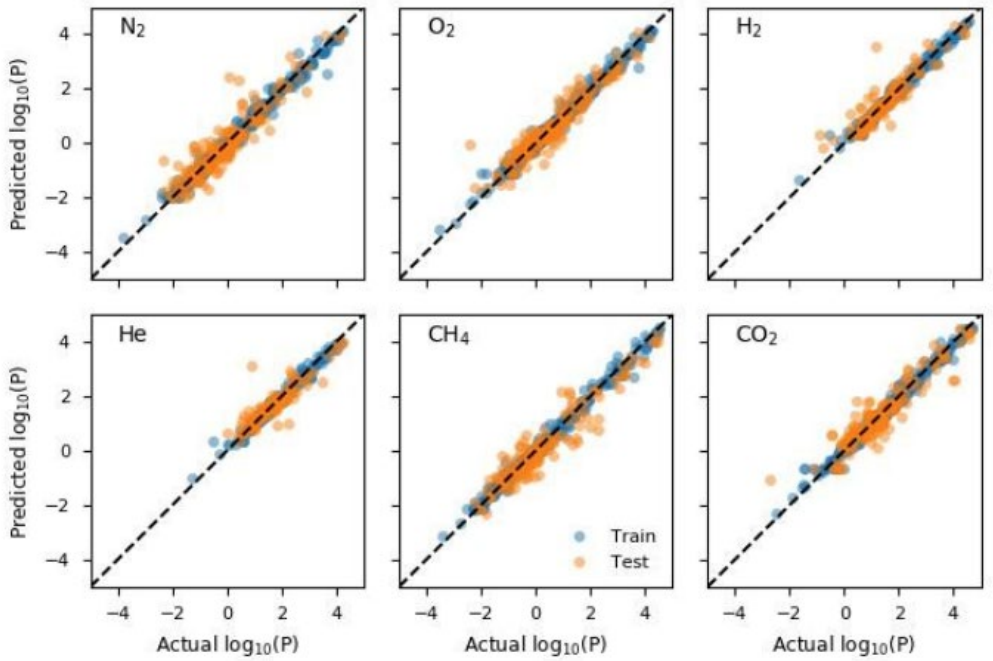
Exercise: *Develop one-hot encoded representations for these atmospheric gases from SMILES*

**Gas Permeability in Polymer Membranes**




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

**Separate Machine Learning Models**




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

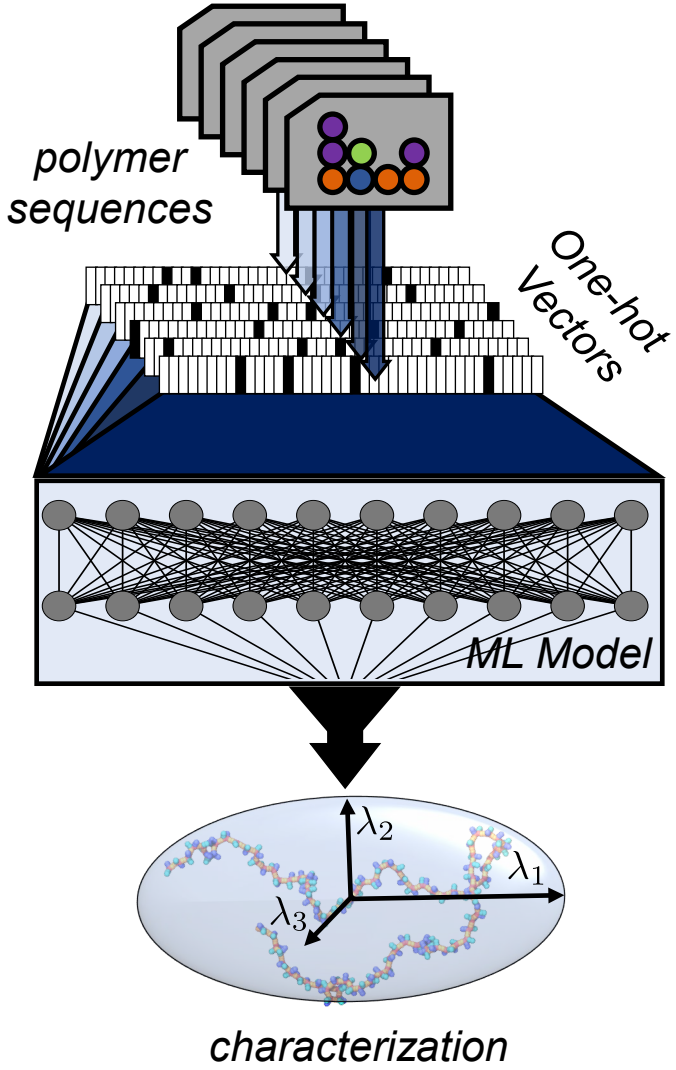
## SMILES Conversion

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

*go to notebook*

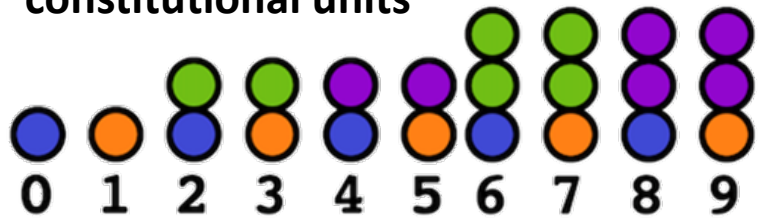
# Tokenization and One-Hot Encoding

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

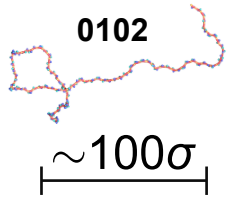


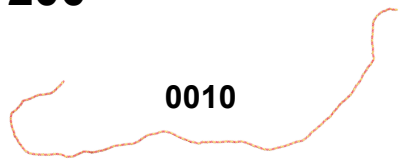

Webb et al. *Sci. Adv.* **2020**

**10 allowable "constitutional units"** $R_k \in \{0, \ldots, 9\}$

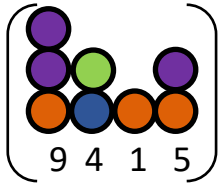



$\left(-R_1-R_2-R_3-R_4-\right)_{200}$

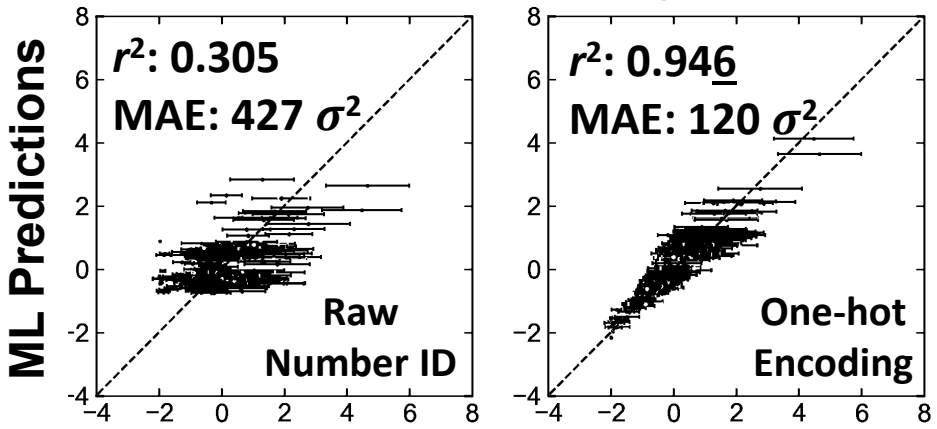



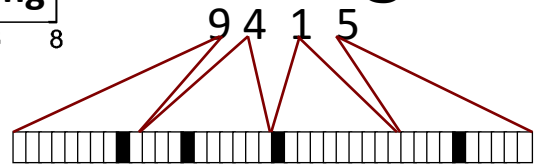

***raw number ID***




**Standardized** $\langle R_g^2 \rangle$




***one-hot-encoding***



# 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**

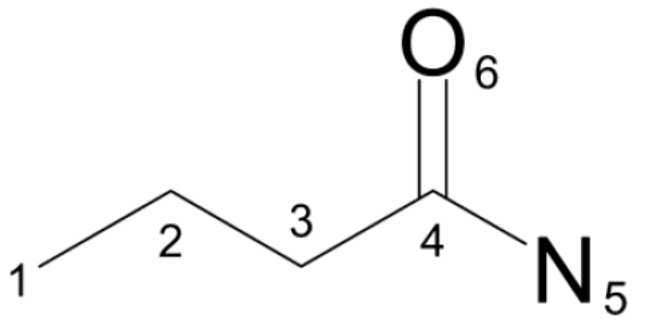
**Basic Algorithm**

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

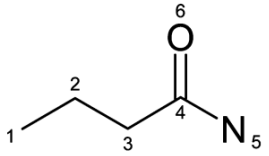
***Extended Connectivity Fingerprints (ECFPs)***

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)***

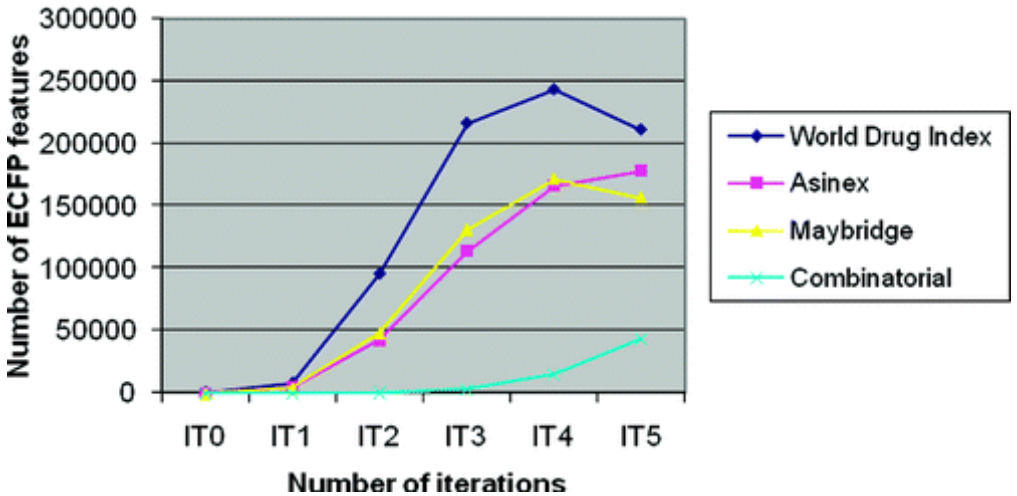
Rogers and Hahn. "Extended-Connectivity Fingerprints." *J. Chem. Inf. Model* **2010**

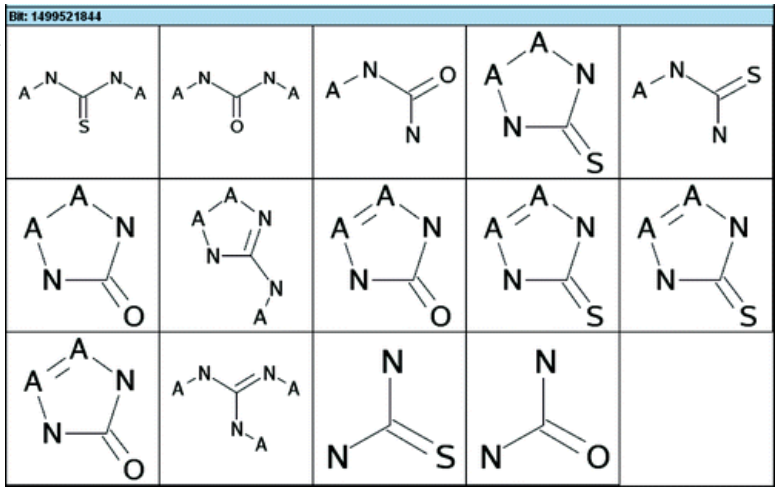
## Advantages

- 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



# Other Structural/Geometric Input Representations

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

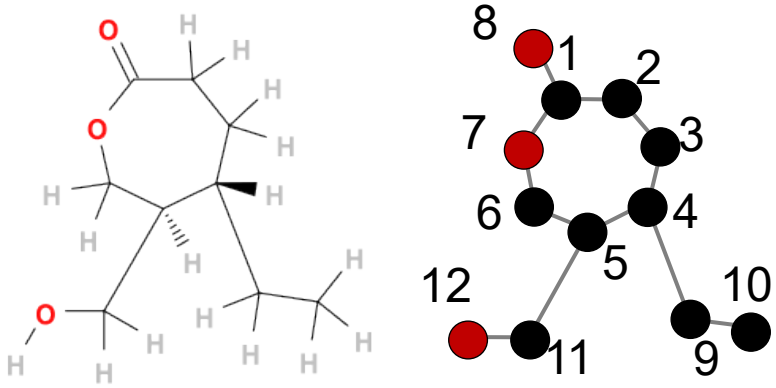
An important consideration for such representations is…
***equivariance*** and ***invariance***

**Noether Theorem -**

Common considerations:

*Translation* *Rotation* *Permutation*

*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.*




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

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

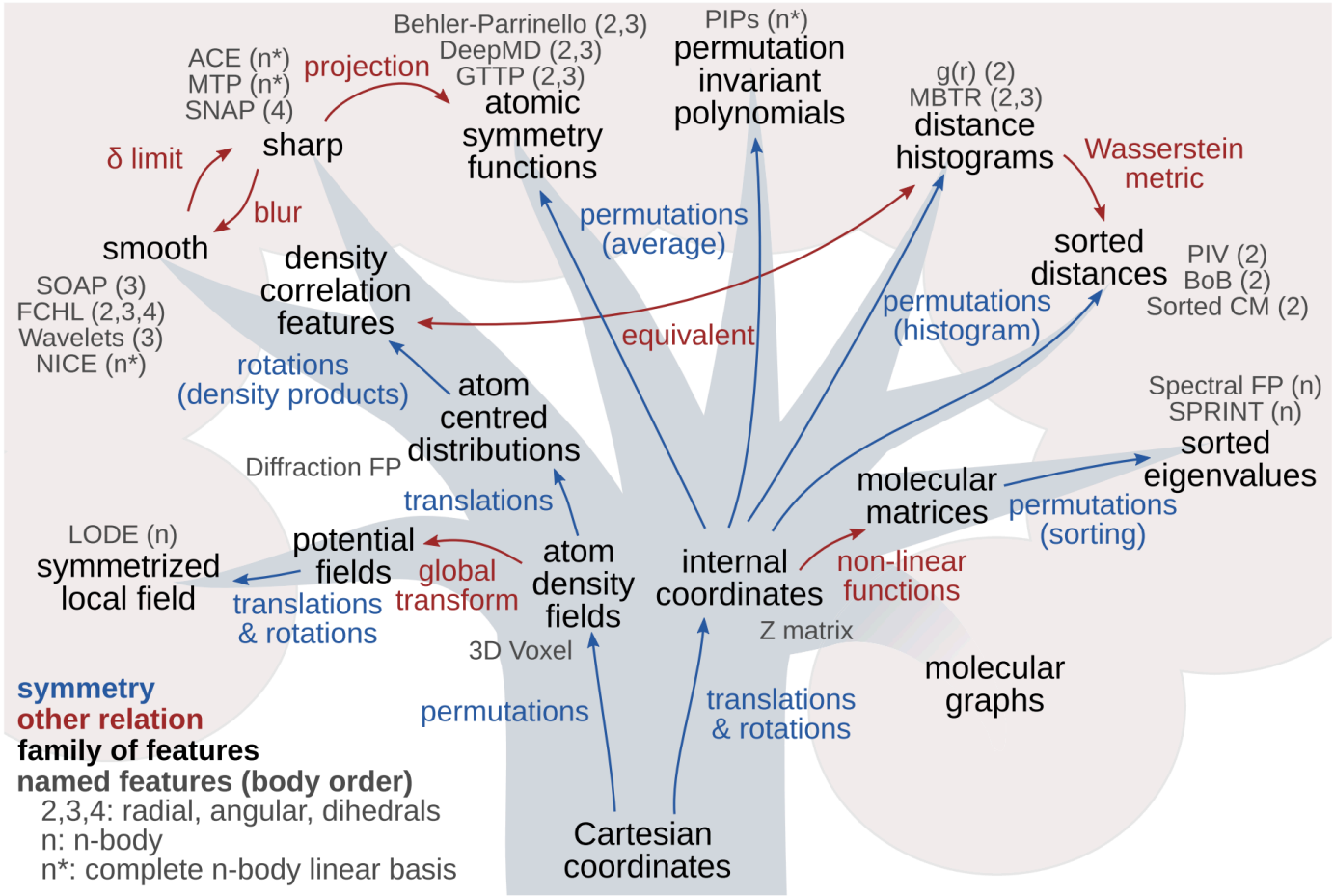
*What properties should change? Which shouldn't?*

# 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

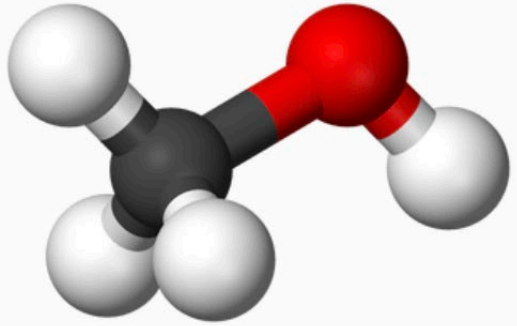


# Some Examples

## *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***

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

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

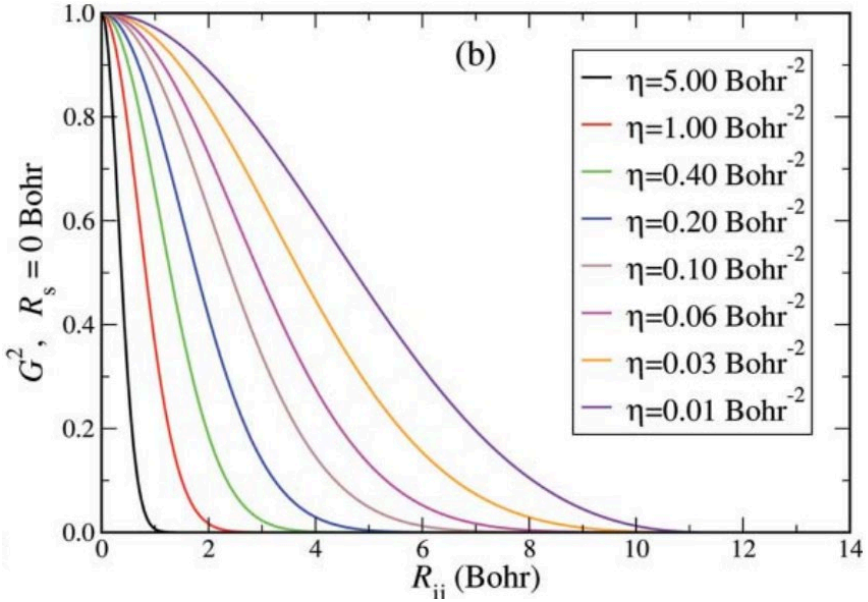
$$G_i^2 = \sum_j e^{-\eta(r_{ij}-r_s)^2} f_c(r_{ij})$$

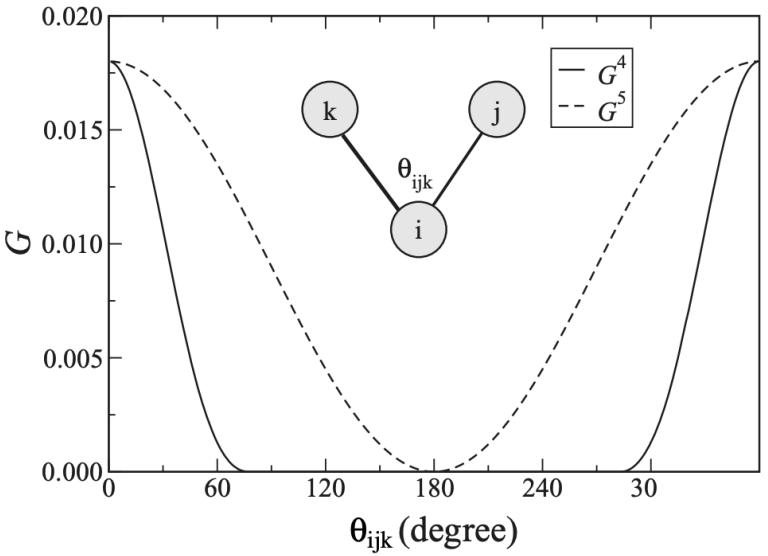
$$G_i^3 = \sum_j \cos(\kappa r_{ij}) f_c(r_{ij})$$

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

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

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

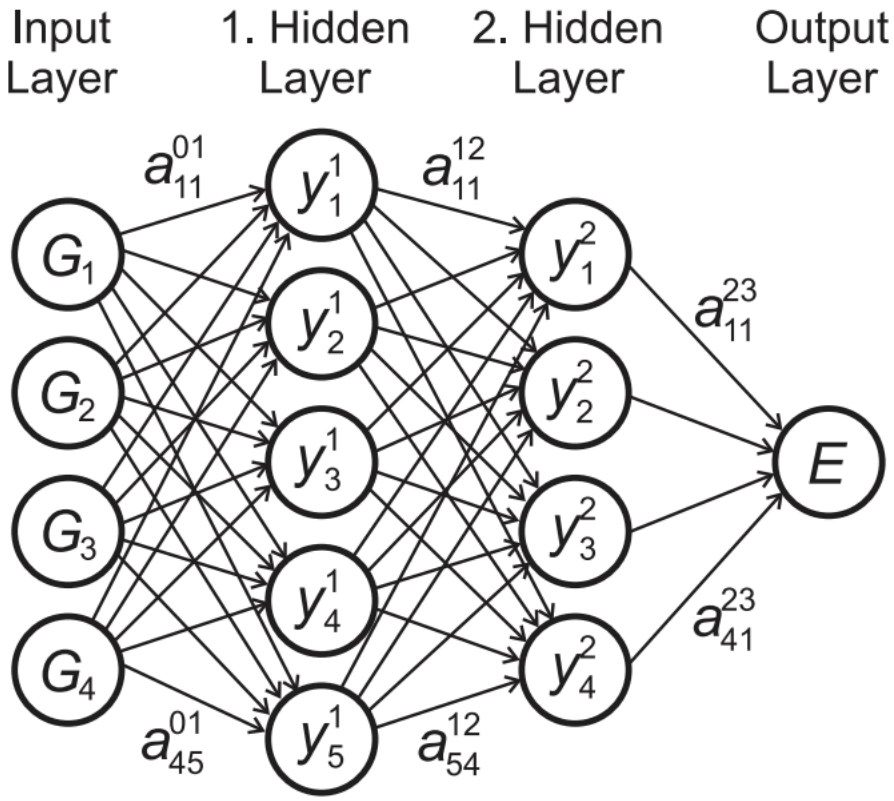


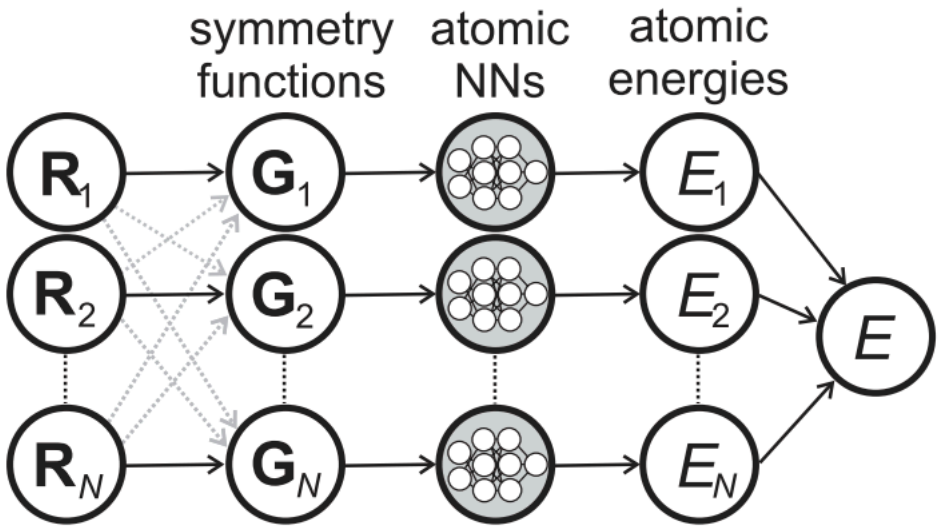




Behler *J. Chem. Phys.* **2011**

# Some Examples

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




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

# 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, $\Delta h_f$, IP, $\varepsilon_{gap}$, …*
- *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* | 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($\gamma$) activity coefficient of **I** in solvent (—) | COSMO*therm*[39] |
| | $\sigma'_1 - \sigma'_3$ profiles segmented into three (—) | COSMO*therm*[39] |
| | $\sigma_1$–$\sigma_5$ 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, $\Delta h_f$, IP, $\varepsilon_{gap}$, …*
- *simulation-derived quantities*
- *experimental measurements*

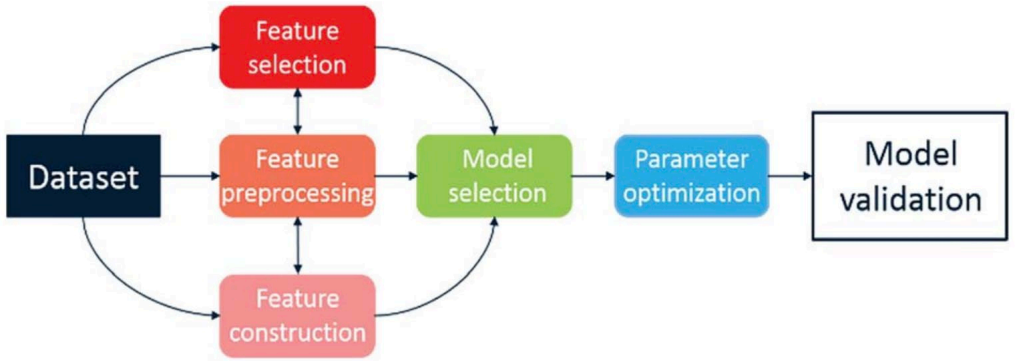
*Descriptors can describe local or global characteristics. Some may be readily available or easily obtained, while others can be complicated/expensive to acquire.*

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

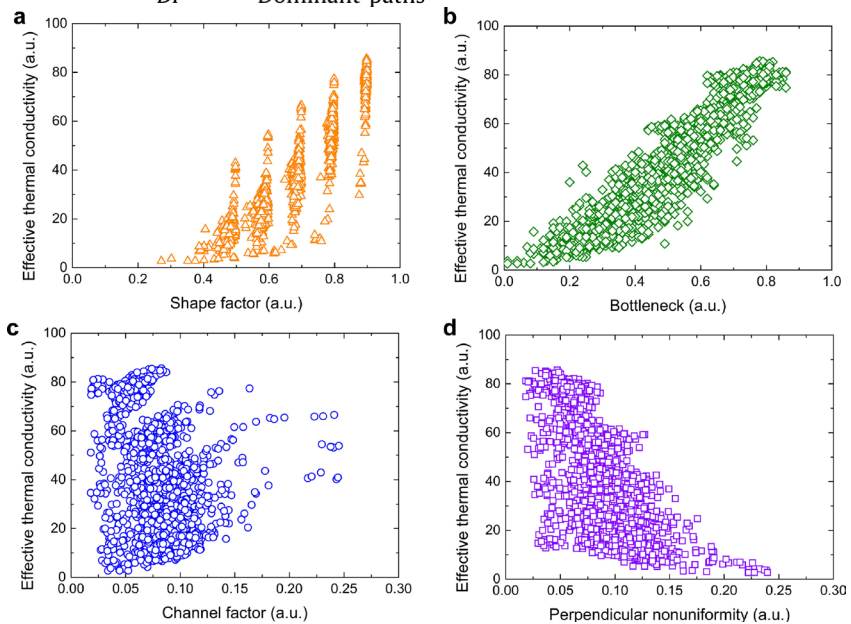
Descriptor vectors can provide suitable representations for non-molecular systems; they are likely tailored to the materials class

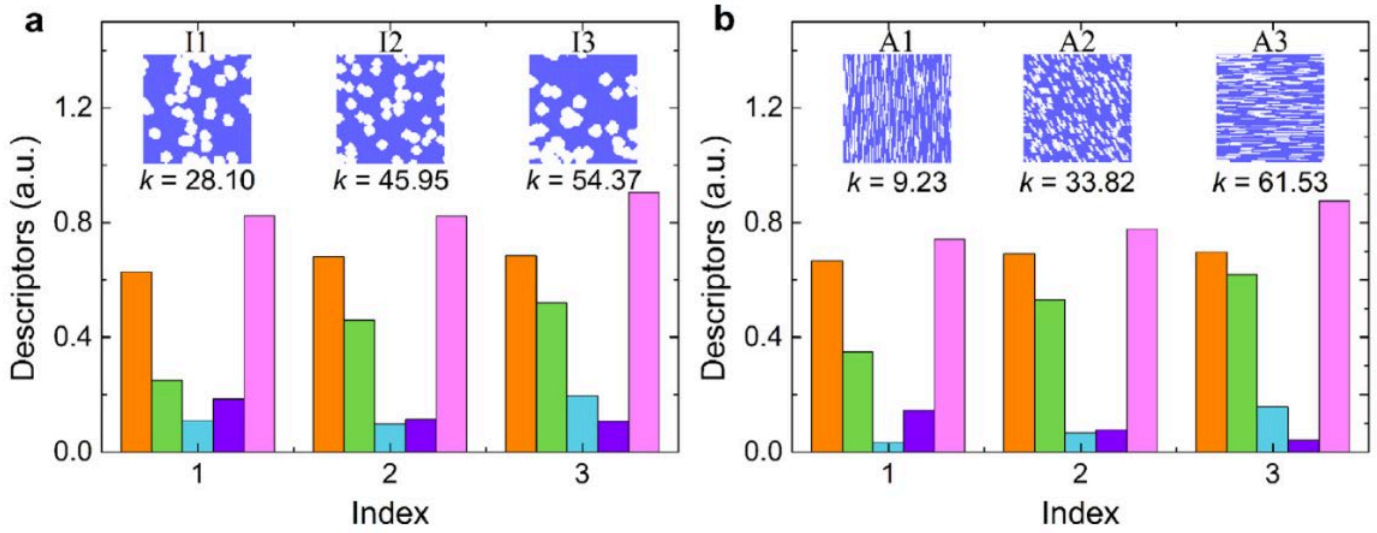

Machine learning prediction of thermal transport in porous media with physics-based descriptors

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

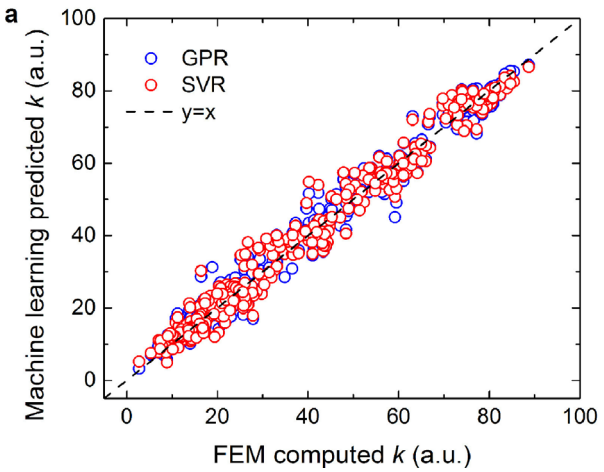
Available online 17 July 2020

| | |
|---|---|
| $\varepsilon$ | 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

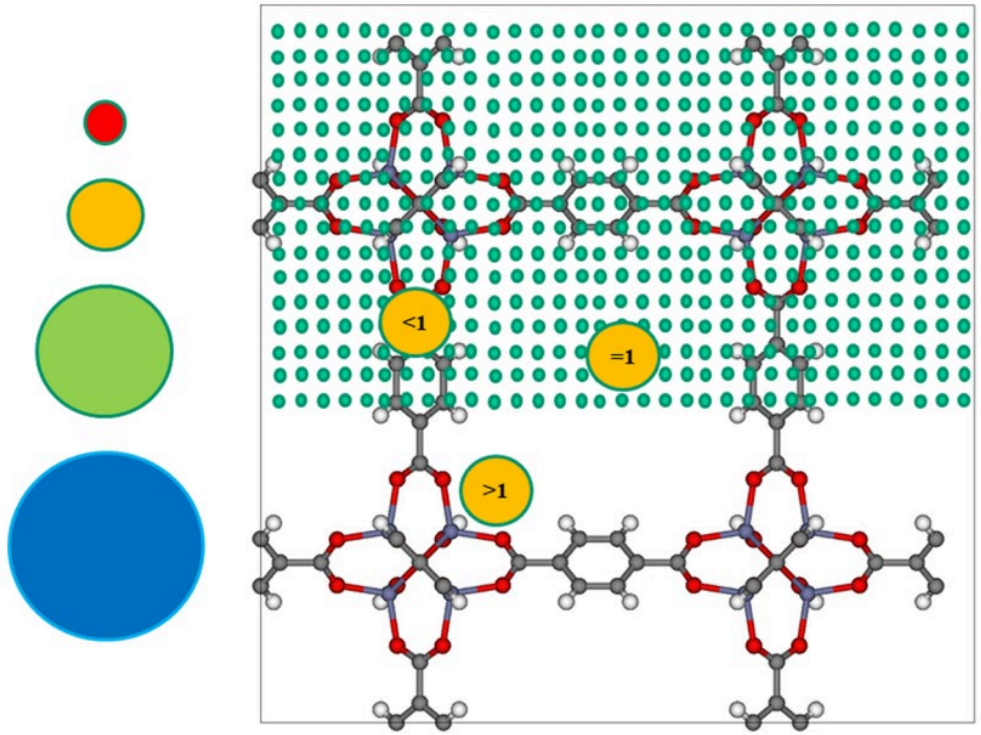
George S. Fanourgakis,*,† Konstantinos Gkagkas,‡ Emmanuel Tylianakis,¶ Emmanuel Klontzas,†,§ and George Froudakis*,†

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 ($\sigma$ = 2.5 Å, $\varepsilon/k_B$ = 50 K) | 0.08 | 5.14 | 1.12 |
| probe-2 ($\sigma$ = 3.0 Å, $\varepsilon/k_B$ = 50 K) | 0.04 | 12.37 | 1.53 |
| probe-3 ($\sigma$ = 3.5 Å, $\varepsilon/k_B$ = 50 K) | 0.0 | 36.35 | 2.30 |
| probe-4 ($\sigma$ = 4.0 Å, $\varepsilon/k_B$ = 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

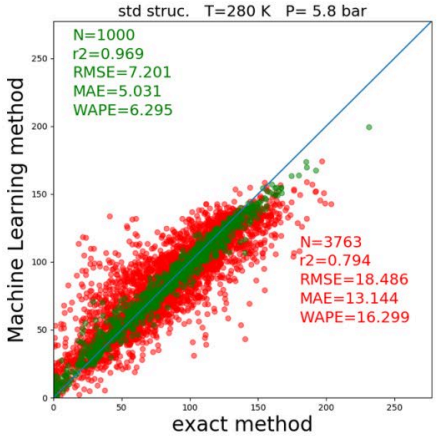
Article

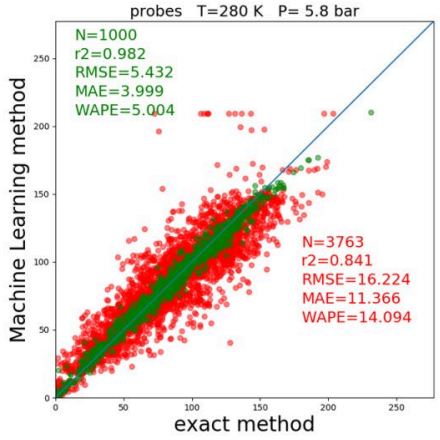
Cite This: J. Phys. Chem. A 2019, 123, 6080−6087

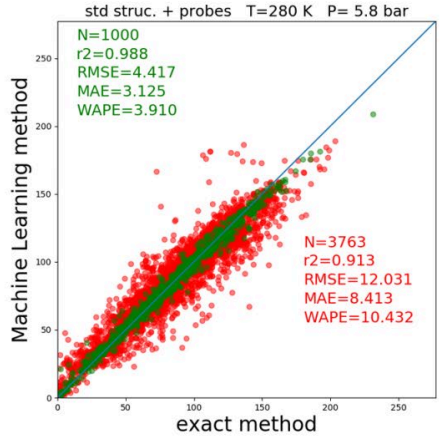
pubs.acs.org/JPCA

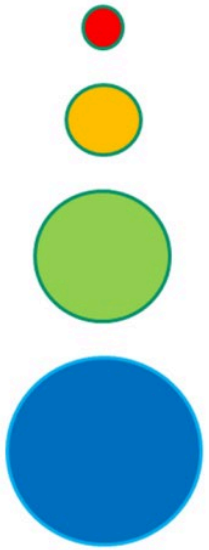
**A Robust Machine Learning Algorithm for the Prediction of Methane Adsorption in Nanoporous Materials**

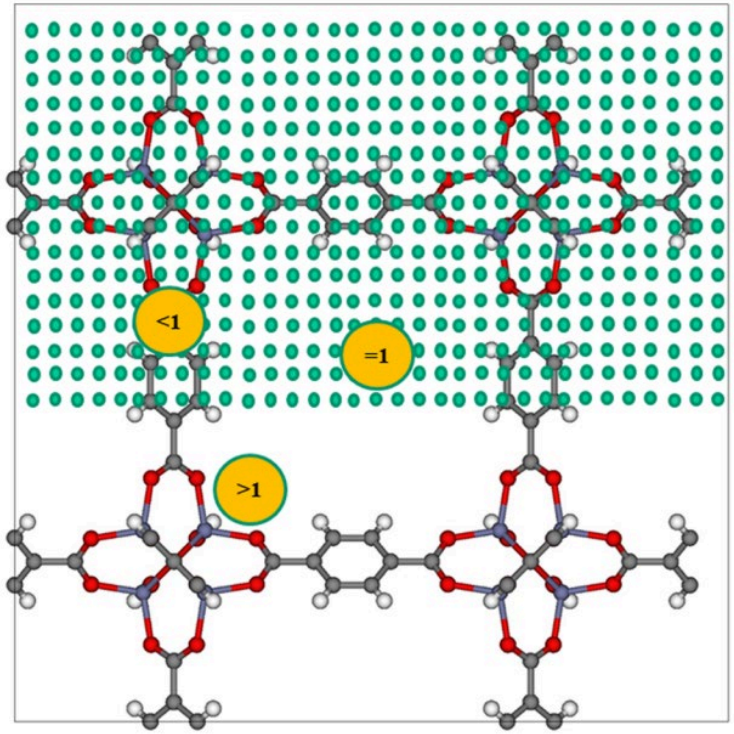
George S. Fanourgakis,*,† Konstantinos Gkagkas,‡ Emmanuel Tylianakis,¶ Emmanuel Klontzas,†,§ and George Froudakis*,†



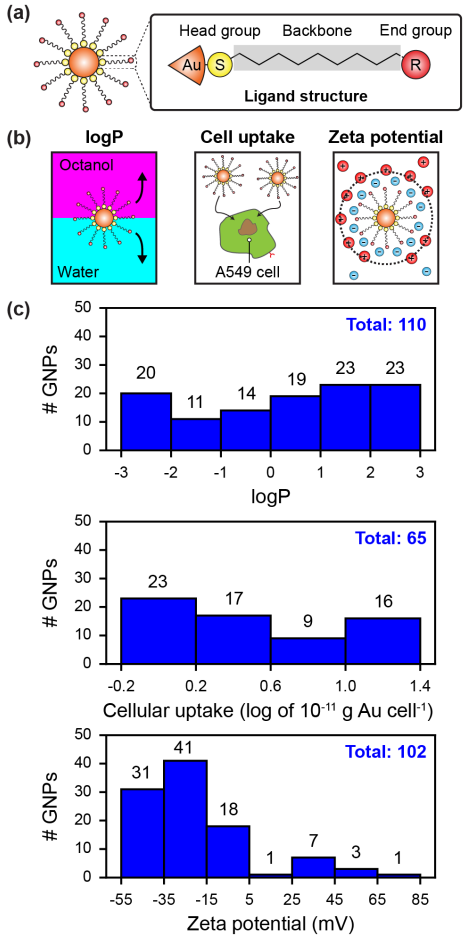






$$\text{probe} - (a) = \frac{1}{N}\sum_{i=1}^{N} \exp(-\beta E_i^{(a)})$$

# Simulation-derived Descriptors

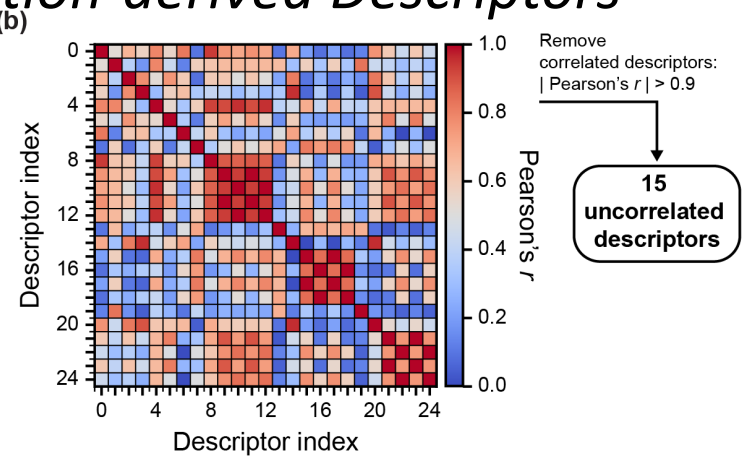
## *Experimental Dataset*

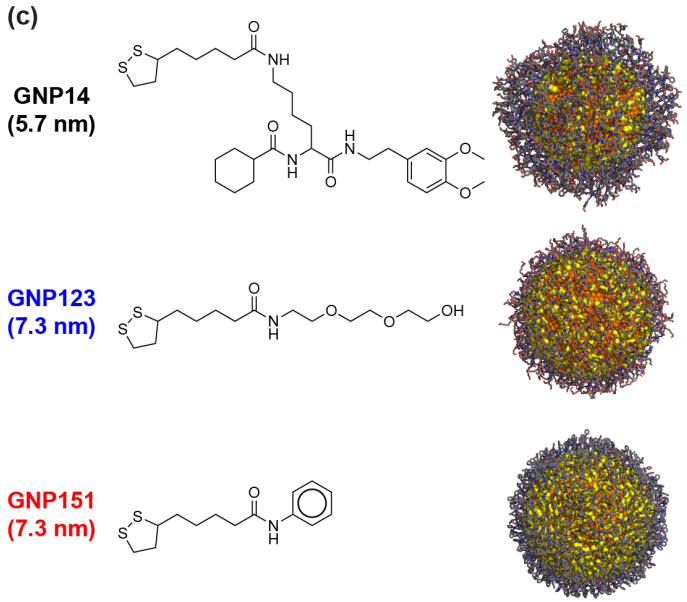


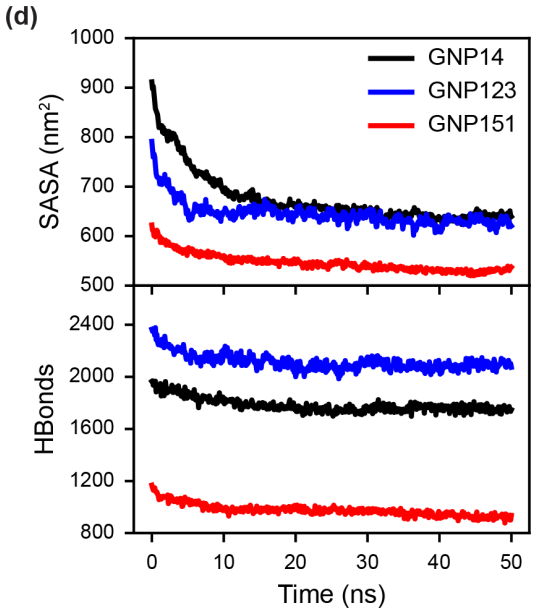

## *Simulation-derived Descriptors*

(a)

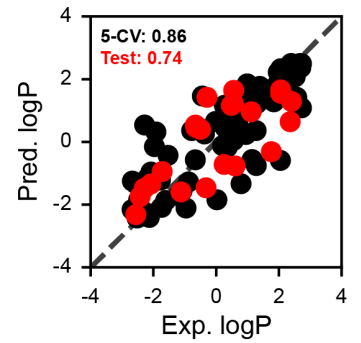
| Descriptor | Description |
|---|---|
| SASA | Solvent accessible surface area |
| HBonds | #Lig.-water hydrogen bonds |
| RMSF | Lig. root-mean-squared fluctuation |
| *E*(NP-S) | GNP-solvent LJ energies |
| $\Delta\phi$ | Radial electrostatic potential drop |
| *e* | Lig. eccentricity (*i.e.* asphericity) |
| *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** |

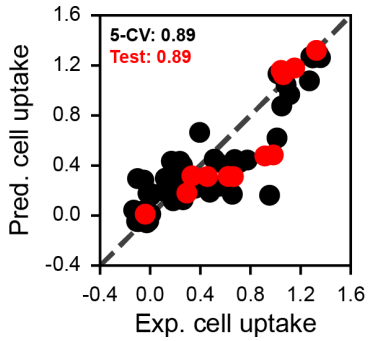


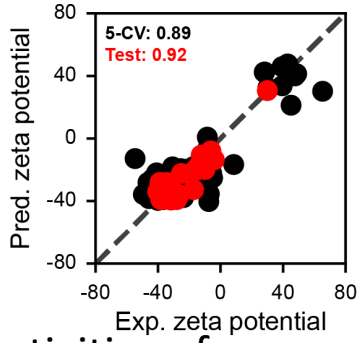




## *ML Models*




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