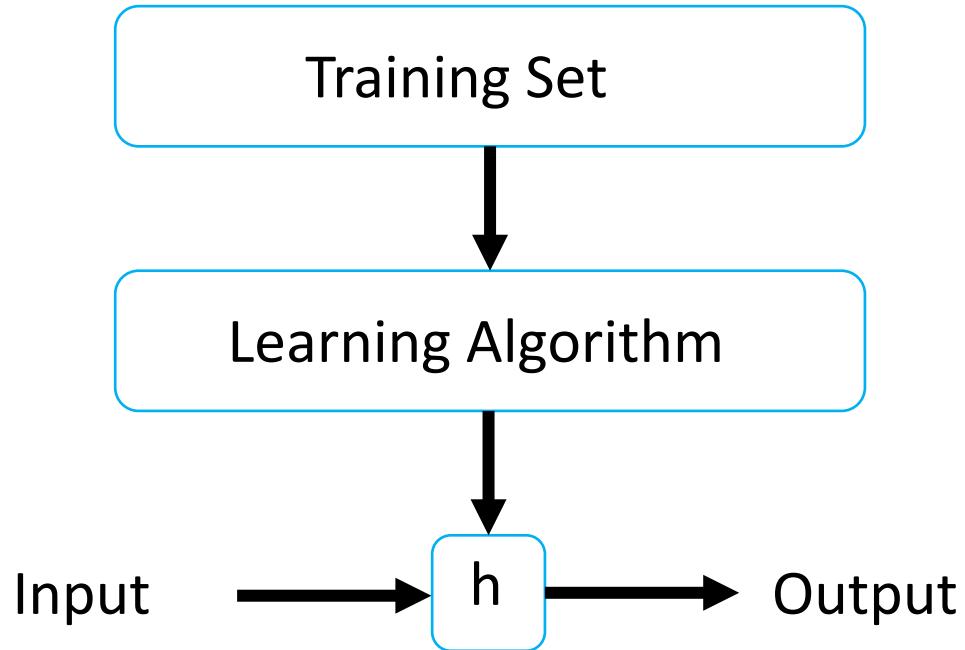


Molecular Featurization and Chemical Descriptors

Review

- MLR



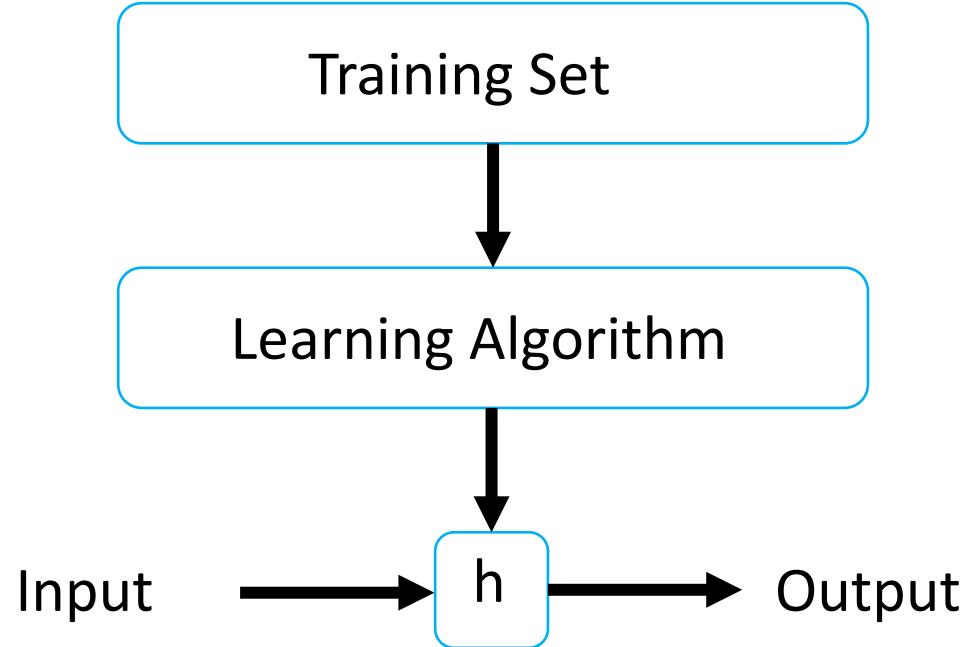
$$\theta^T X = y$$

$$\theta = (X^T X)^{-1} X^T y$$

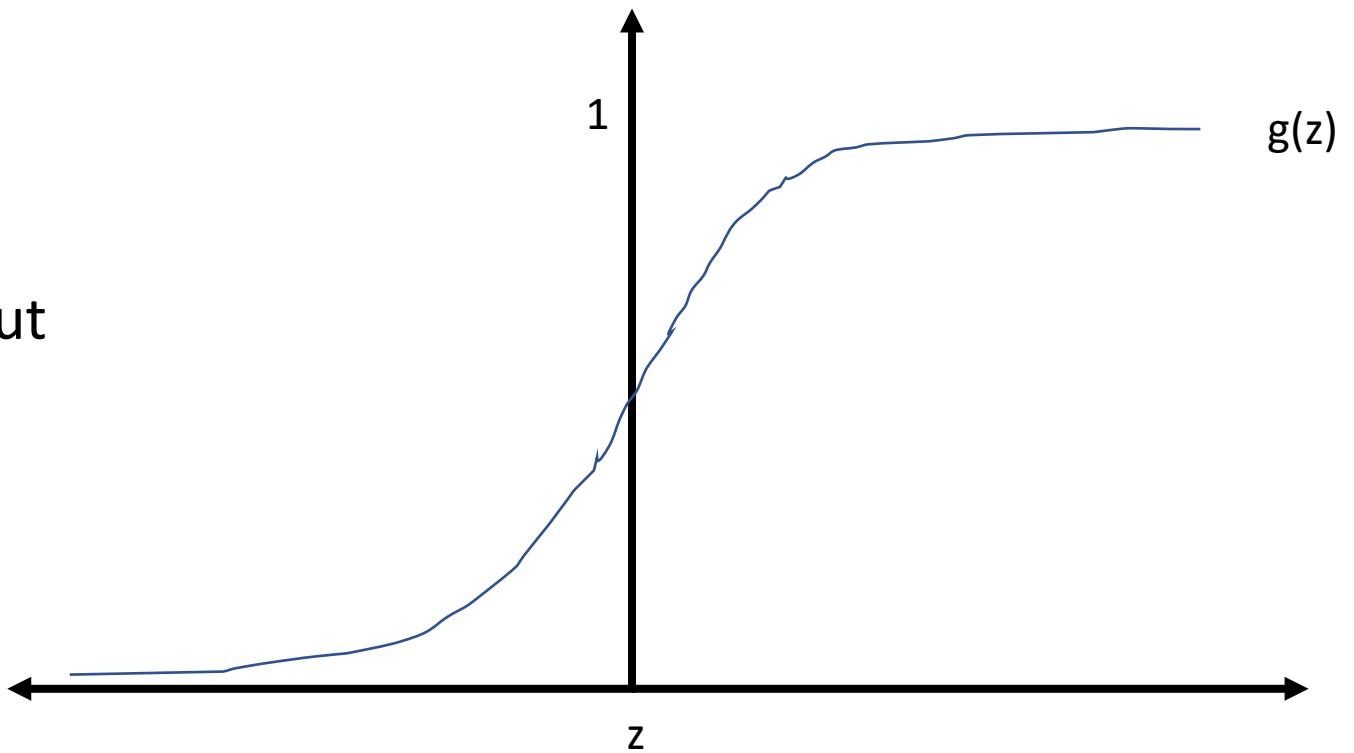
$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2$$

Review

- Classification vs. Regression

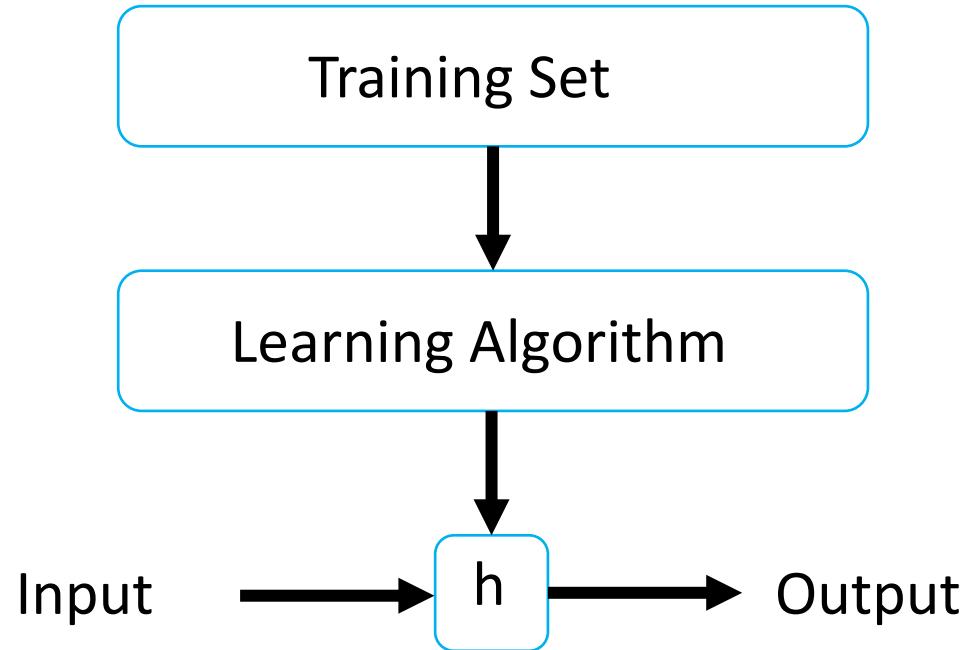


$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

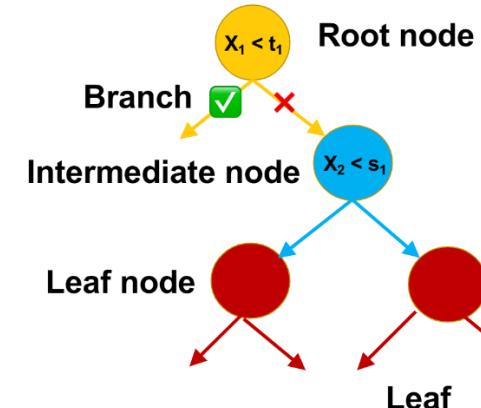


Review

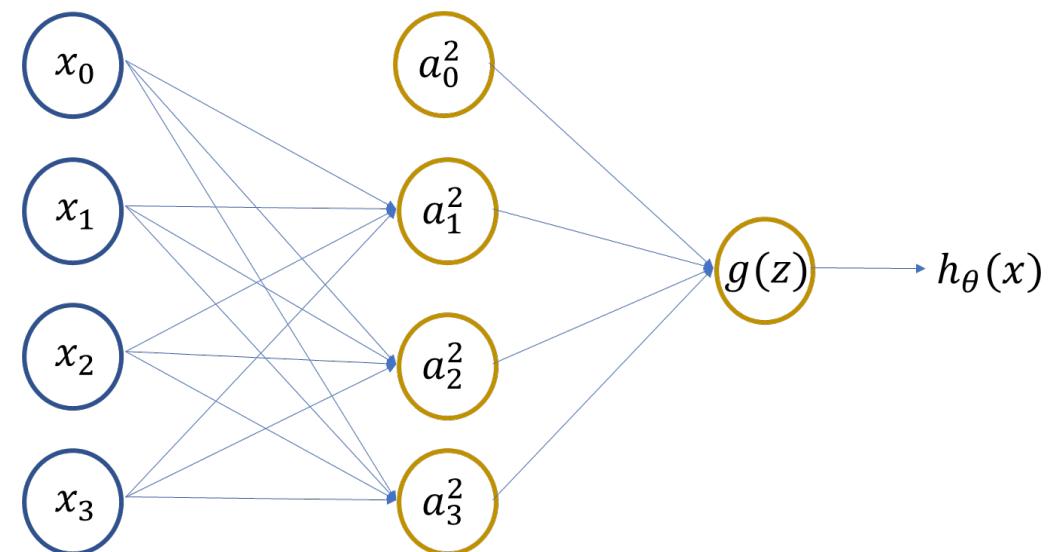
- More methods



Decision Tree



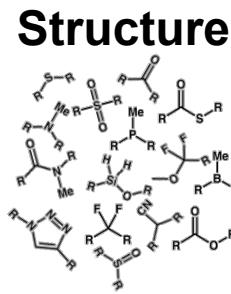
Depth of the tree
= maximum
number of
branches to
reach a leaf



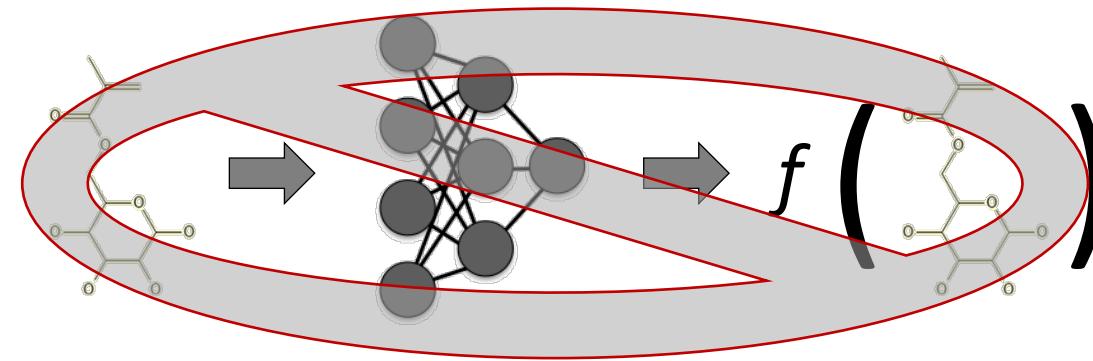
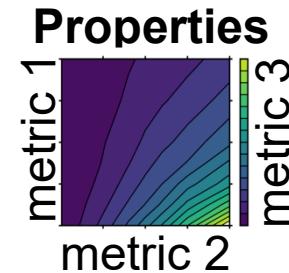
Machine Learning Meets Molecules

A critical task for utilizing machine learning algorithms is data representation

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



Prediction
Design



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

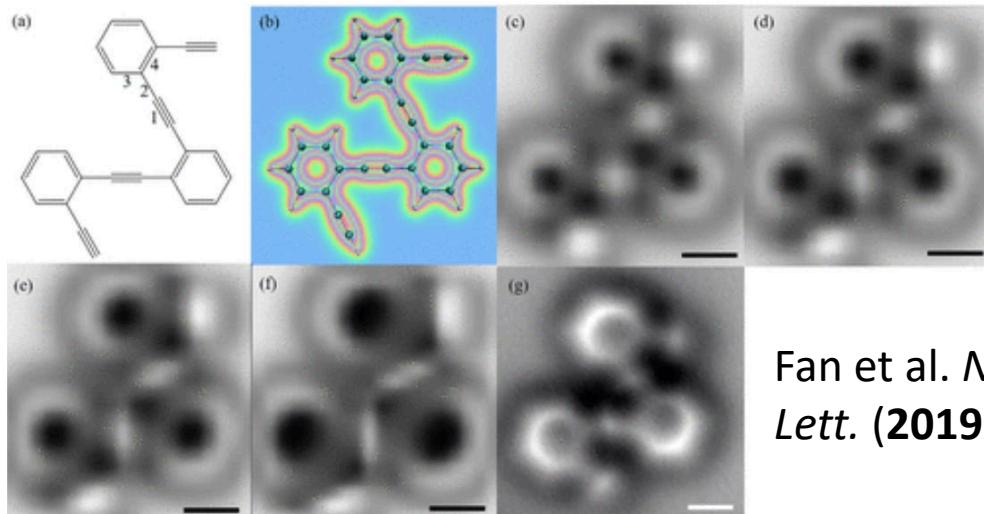
*This conversion process is referred to as **molecular featurization***

What is a molecule?

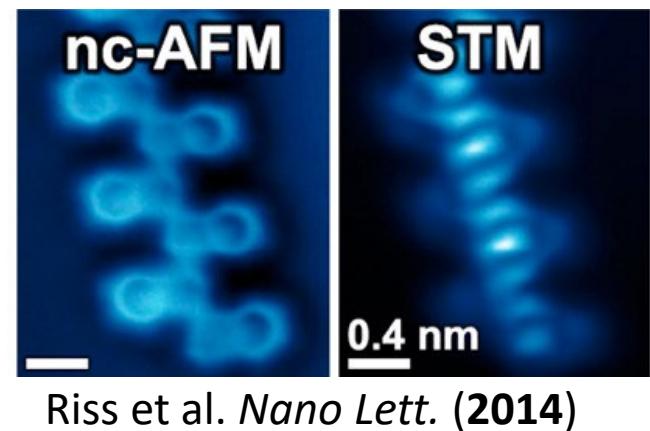
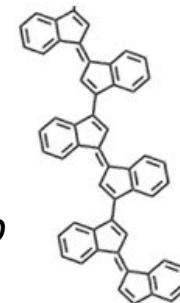
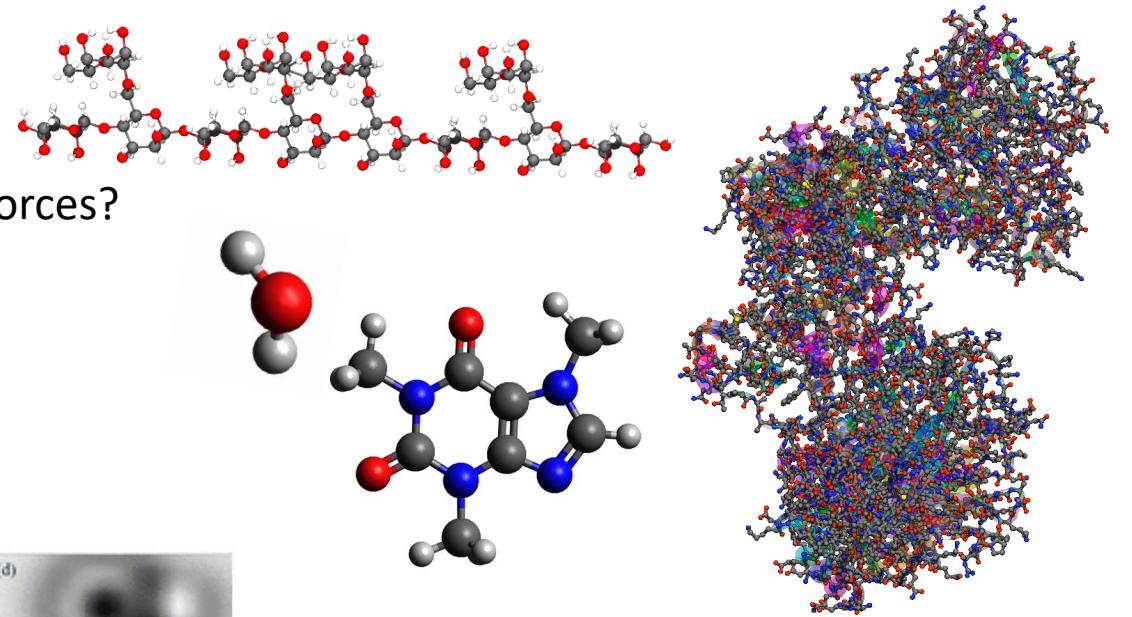
Before we specifically address how to featurize a molecule, we should settle on what a molecule *is* since that may dictate the approach

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?



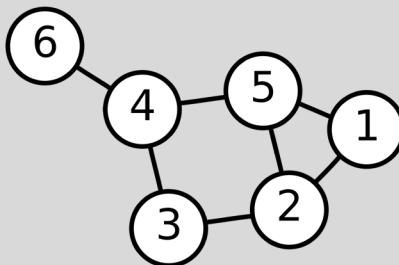
Fan et al. *Nano Lett.* (2019)



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

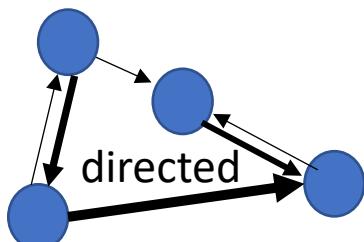
\mathcal{V}

Vertices/nodes – indicates objects

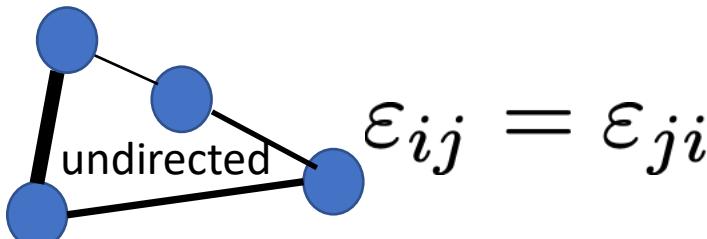
\mathcal{E}

Edges – indicates pairwise relationship amongst objects

graphs may be *directed* or *undirected*



vs.



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

Graphs can be conveniently represented as matrices

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

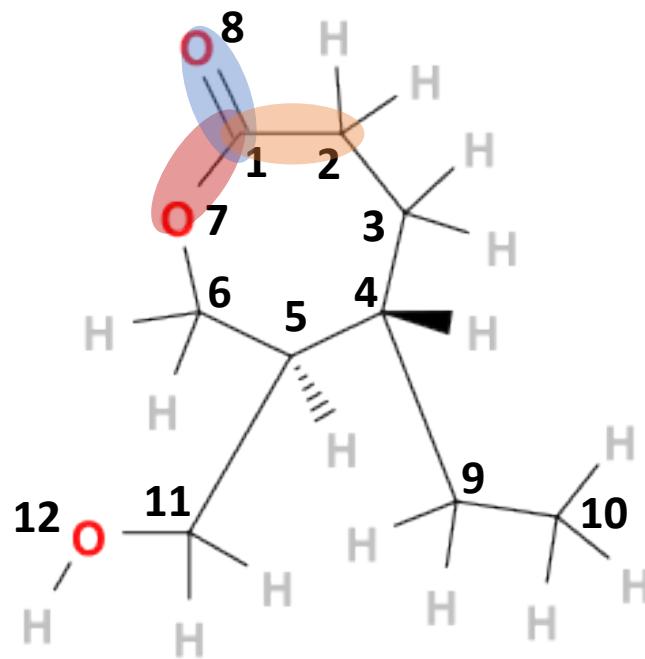
e.g.,

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

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

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
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

*ignoring hydrogens

Here, entries denote presence of a bond (or not) → often referred to as an Adjacency matrix, but the premise can easily be 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: Simplified Molecular-Input Line Entry System (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*

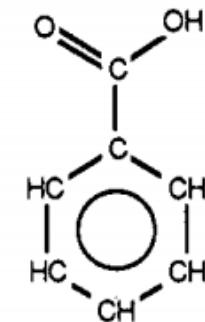
Daylight Chemical Information Systems, Irvine, California 92714

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)*

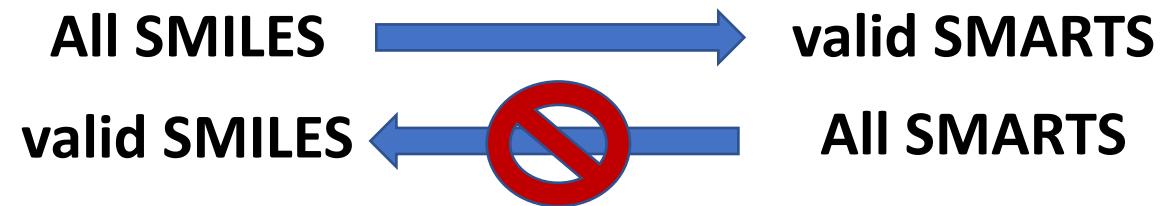
c1ccccc1C(=O)O

Text-based Representations

Common starting points for describing molecular graphs are text strings

An extension: SMILES Arbitrary Target Specification (SMARTS)

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



→ 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	

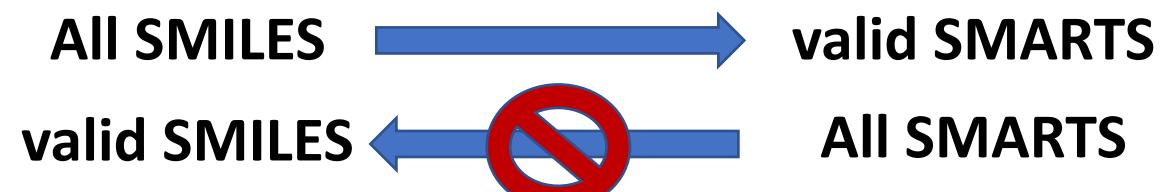
SMARTS Logical Operators		
Symbol	Expression	Meaning
exclamation	!e1	not e1
ampersand	e1&e2	a1 and e2 (high precedence)
comma	e1,e2	e1 or e2
semicolon	e1;e2	a1 and e2 (low precedence)

Text-based Representations

Common starting points for describing molecular graphs are text strings

An extension: SMILES Arbitrary Target Specification (SMARTS)

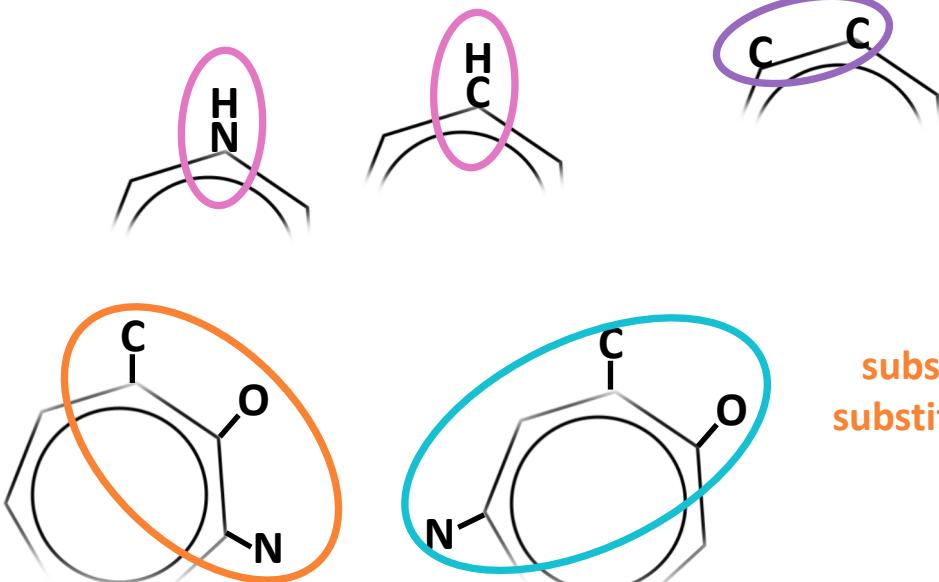
SMARTS is not for representing molecular structures but *chemical patterns*



→ 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

either aromatic carbon or nitrogen and exactly one hydrogen

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

same as above but O and N likely para

- “cc”
- [c,n;H1]
- “Caa(O)aN”
- “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
C1CCCCCCCC1	CCCCCC%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/CCCCC/1	C=C/CCCCC/8
C\1=C/CCCCC1	C=C/CCCCC/8
B(c1ccccc1)(O)O	Bcccccc6))))O)O
Cn1cccc-2nccc12	Cnccccnccc9-5
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

Other string-based representations

Wiswesser line notation

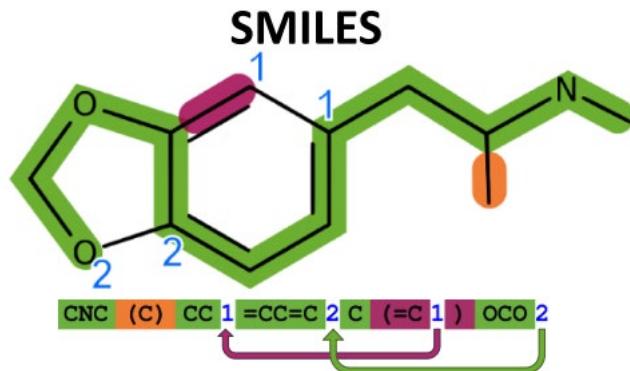
SYBYL Line Notation

IUPAC International Chemical Identifier

Text-based Representations

Common starting points for describing molecular graphs are text strings

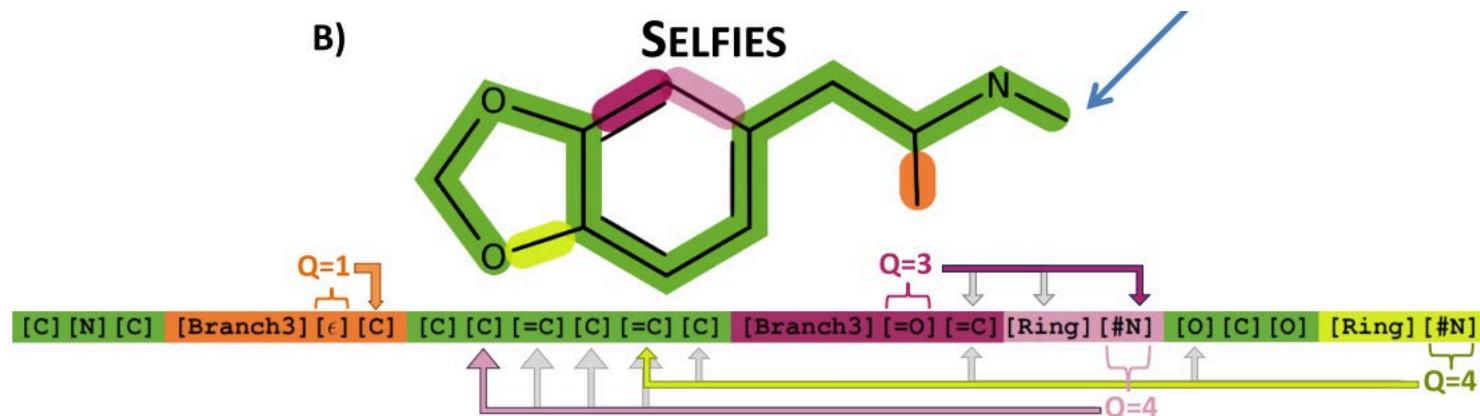
Self-referencing Embedded Strings (SELFIES)



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

Mario Krenn^{1,2,3} , Florian Häse^{1,2,3,4}, Akshat Kumar Nigam², 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 Embedded Strings (SELFIES)

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

Mario Krenn^{1,2,3} , Florian Häse^{1,2,3,4}, Akshat Kumar Nigam², Pascal Friederich^{2,5} and Alan Aspuru-Guzik^{1,2,3,6}

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

Formal Grammar Rules

State of Derivation	Rule Vectors													
	[ε]	[F]	[=O]	[#N]	[O]	[N]	[=N]	[C]	[=C]	[#C]	[Branch1]	[Branch2]	[Branch3]	[Ring]
X ₀	X ₀	F X ₁	0 X ₂	N X ₃	0 X ₂	N X ₃	N X ₃	C X ₄	C X ₄	C X ₄	ign X ₀	ign X ₀	ign X ₀	ign X ₀
X ₁ → ε	F	0	N	0 X ₁	N X ₂	N X ₂	C X ₃	C X ₃	C X ₃	ign X ₁	ign X ₁	ign X ₁	R(Q)	
X ₂ → ε	F	=0	=N	0 X ₁	N X ₂	=N X ₁	C X ₃	=C X ₂	=C X ₂	B(Q,X ₅)X ₁	B(Q,X ₅)X ₁	B(Q,X ₅)X ₁	R(Q) X ₁	
X ₃ → ε	F	=0	#N	0 X ₁	N X ₂	=N X ₁	C X ₃	=C X ₂	#C X ₁	B(Q,X ₅)X ₂	B(Q,X ₆)X ₁	B(Q,X ₅)X ₂	R(Q) X ₂	
X ₄ → ε	F	=0	#N	0 X ₁	N X ₂	=N X ₁	C X ₃	=C X ₂	#C X ₁	B(Q,X ₅)X ₃	B(Q,X ₇)X ₁	B(Q,X ₆)X ₂	R(Q) X ₃	
X ₅ → C	F	0	N	0 X ₁	N X ₂	N X ₂	C X ₃	C X ₃	C X ₃	X ₅	X ₅	X ₅		
X ₆ → C	F	=0	=N	0 X ₁	N X ₂	=N X ₁	C X ₃	=C X ₂	=C X ₂	X ₆	X ₆	X ₆		
X ₇ → C	F	=0	#N	0 X ₁	N X ₂	=N X ₁	C X ₃	=C X ₂	#C X ₁	X ₇	X ₇	X ₇		
Q → 1	1	2	3	4	5	6	7	8	9	10	11	12	13	14

conversion to molecular graph

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

1. Start in X₀ → F X₁

2. F X₁ → F C X₃

3. F C X₃ → F C = C X₂

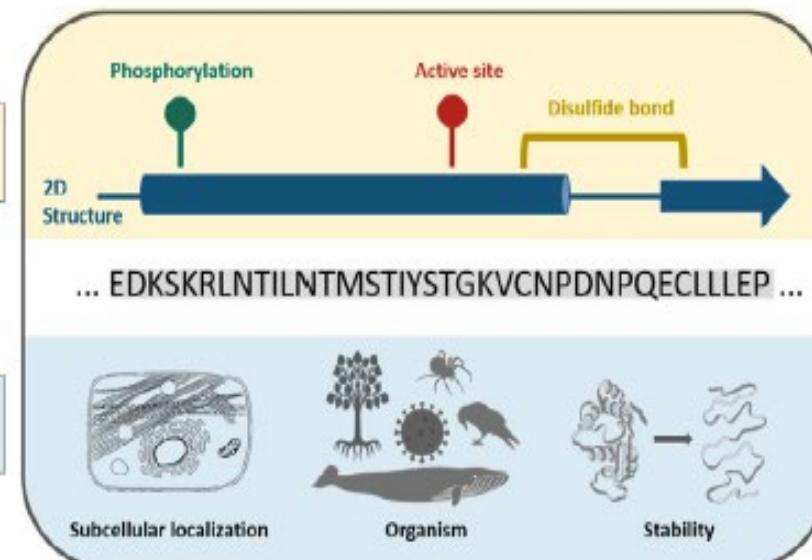
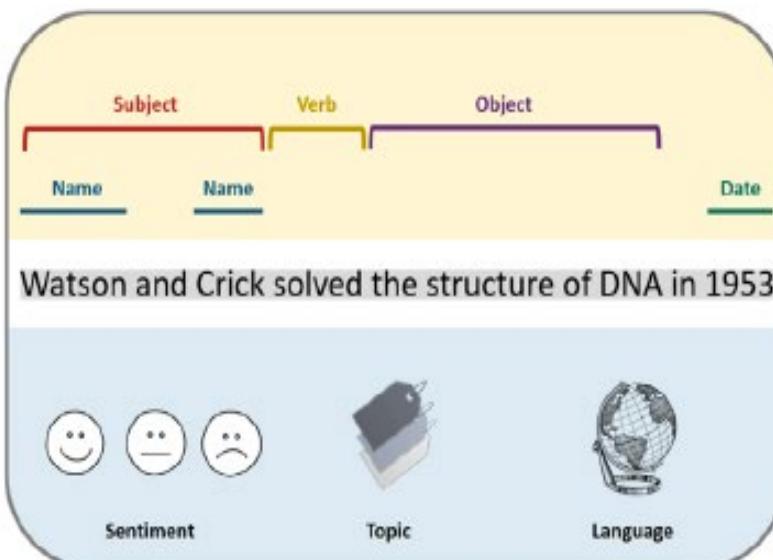
4. F C = C X₂ → F C = C = N

Tokenization and One-Hot Encoding

Tokenization – fundamental task in **Natural Language Processing** that yields ***t*okens**

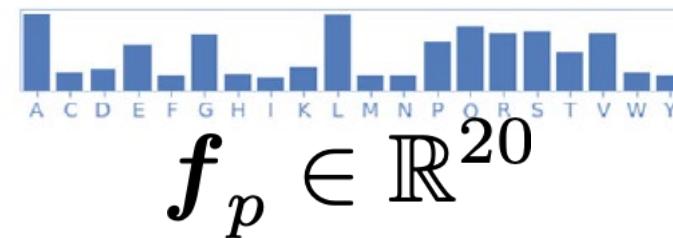
Tokens – basic building blocks of the Natural Language words n-grams characters } the collection of tokens forms a vocabulary

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***



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

011000110110000101110100

0

100



rabbit

011100100110000101100010

1

010



duck

011001000111010101100011

2

001

English; binary; numerical assignment; one-hot vector

Tokenization and One-Hot Encoding

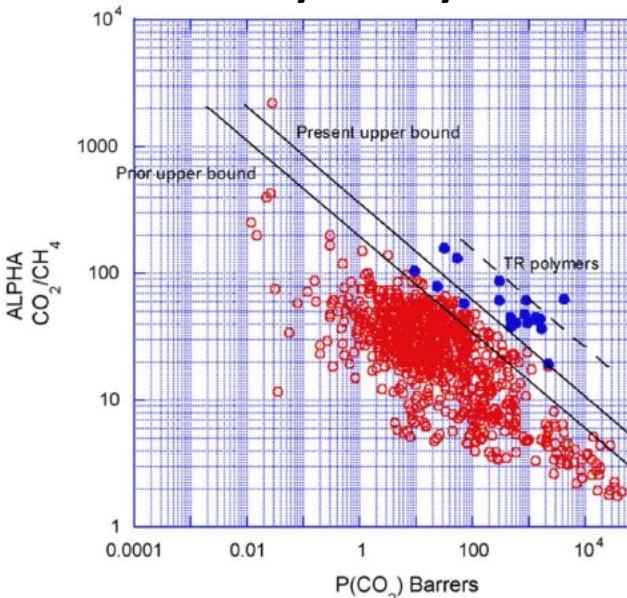
One viable approach
for molecules:

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

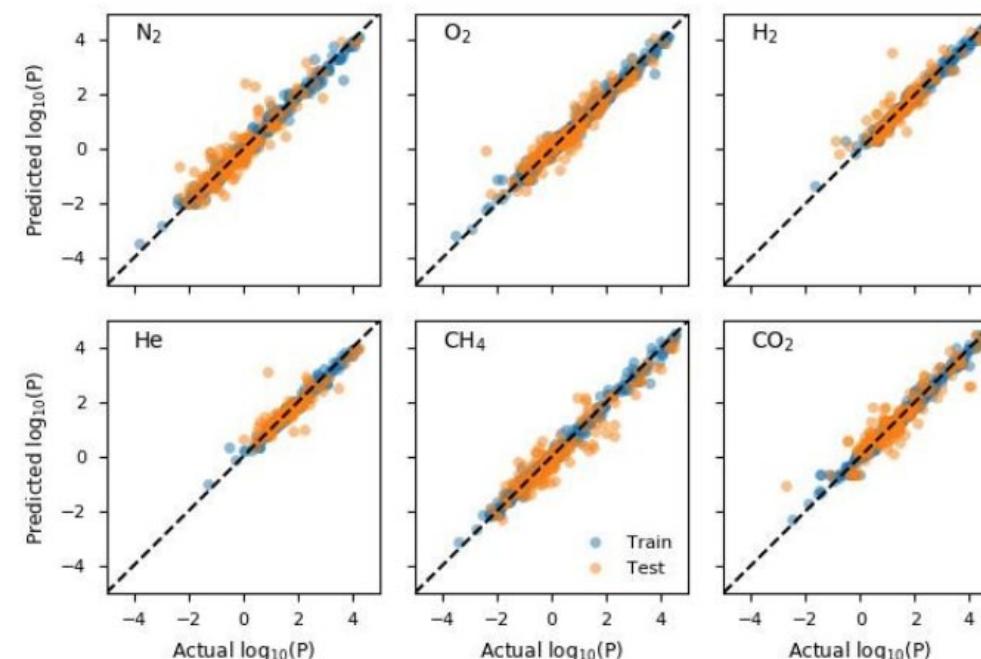
Exercise: Develop one-hot encoded representations for these atmospheric gases from SMILES

Gas Permeability in Polymer Membranes



Robeson J. Membr. Sci. 2008

Separate Machine Learning Models



Barnett et al. Sci. Adv. 2020

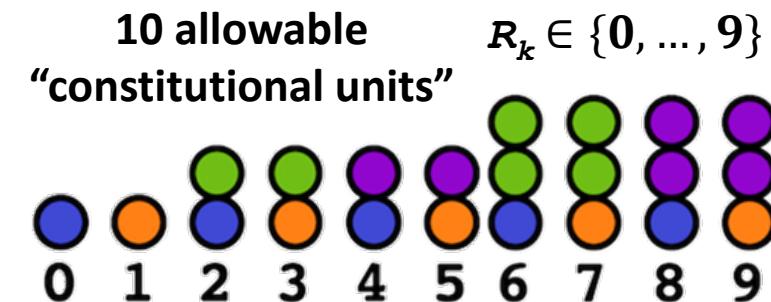
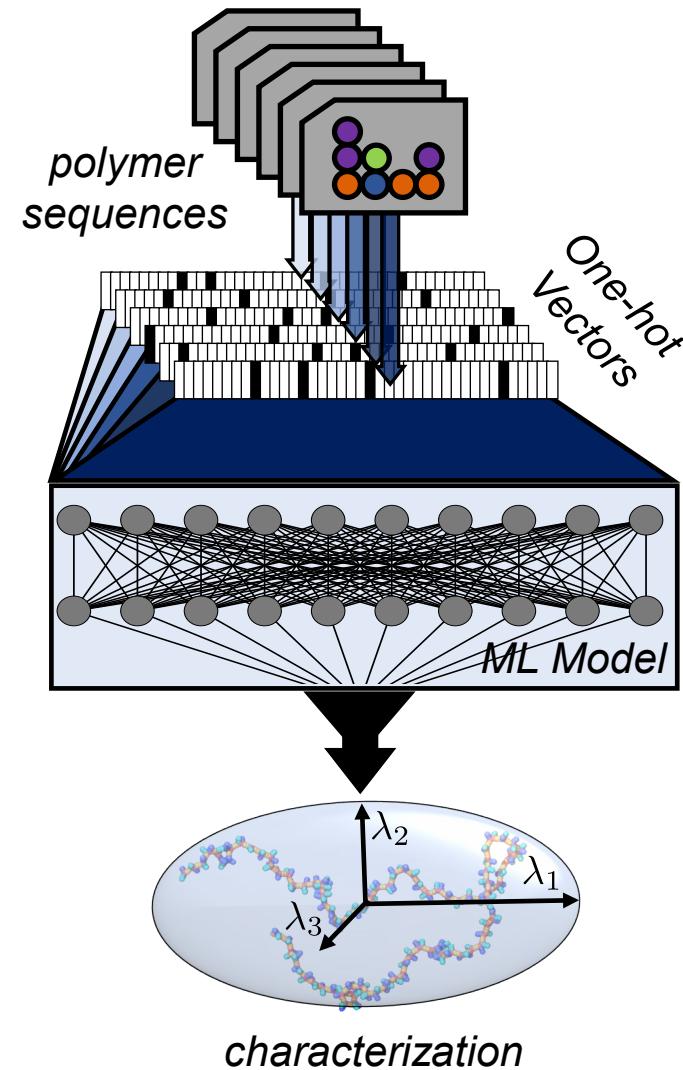
SMILES Conversion

$\text{N}_2 \rightarrow \text{N}\#\text{N}$
 $\text{O}_2 \rightarrow \text{O}=\text{O}$
 $\text{H}_2 \rightarrow [\text{HH}]$
 $\text{He} \rightarrow [\text{He}]$
 $\text{CH}_4 \rightarrow \text{C}$
 $\text{CO}_2 \rightarrow \text{C}(=\text{O})=\text{O}$

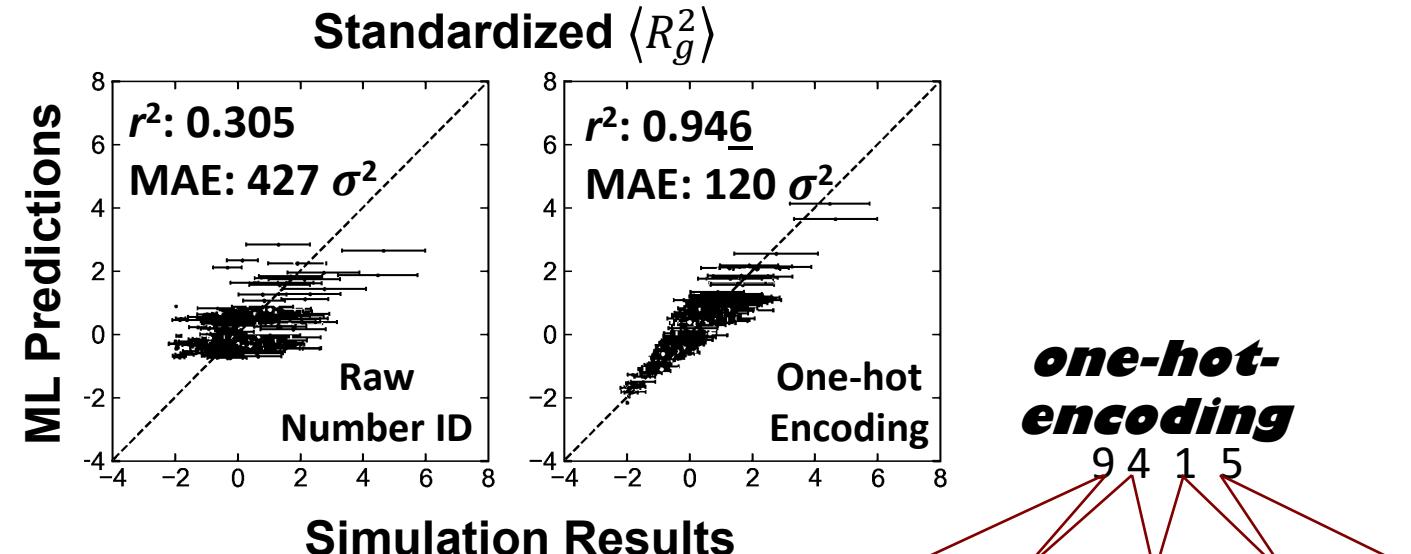
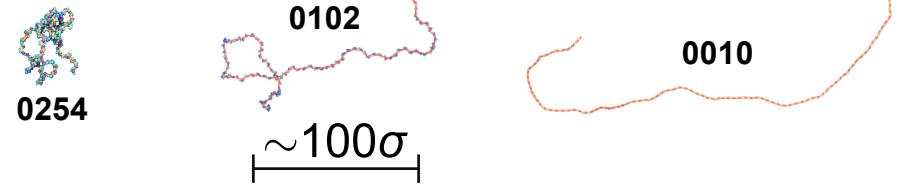
[go to notebook](#)

Tokenization and One-Hot Encoding

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



$$\left[-R_1 - R_2 - R_3 - R_4 - \right]_{200}$$



Extended Connectivity Fingerprints

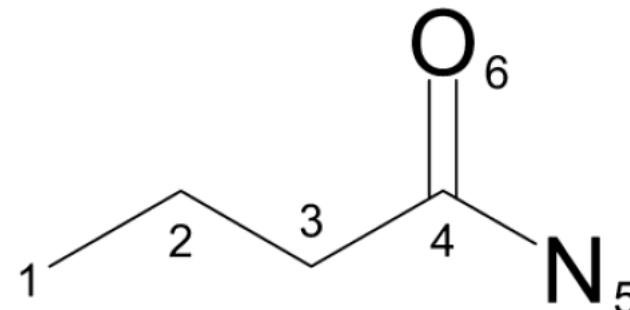
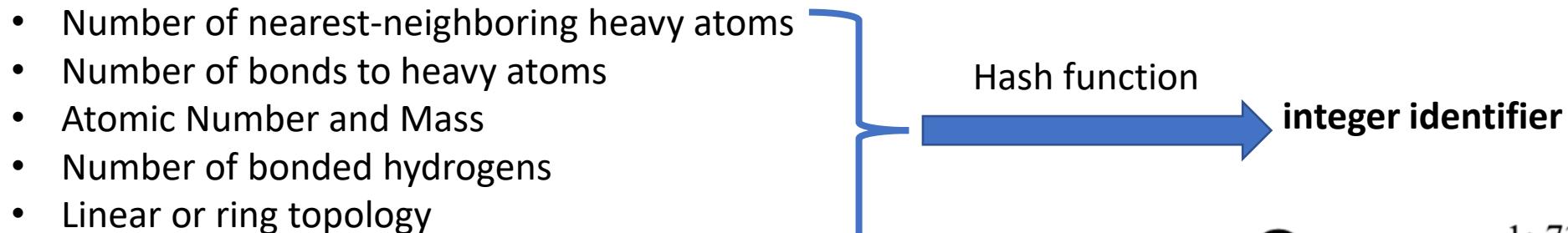
A more topologically informed version of OHE are so-called
Extended Connectivity Fingerprints (ECFPs)

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



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



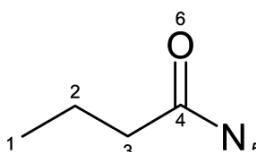
Extended Connectivity Fingerprints

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



> <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

- *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*

extracted
substructures
in SMARTS

Extended Connectivity Fingerprints

A more topologically informed version of OHE are so-called
Extended Connectivity Fingerprints (ECFPs)

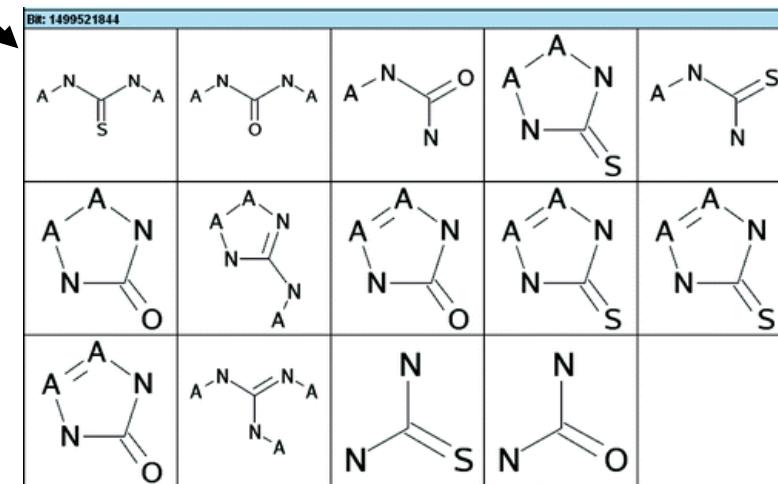
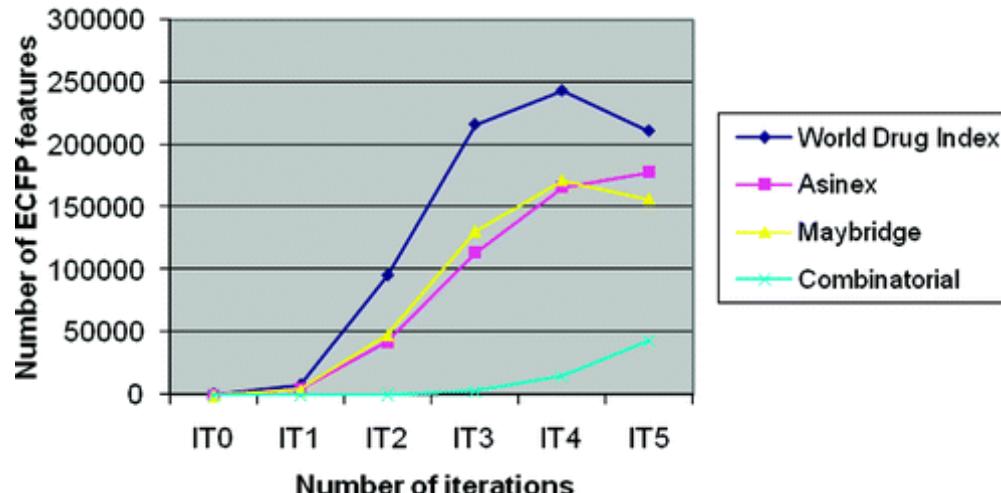
Rogers and Hahn. "Extended-Connectivity Fingerprints." *J. Chem. Inf. Model* 2010

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 -

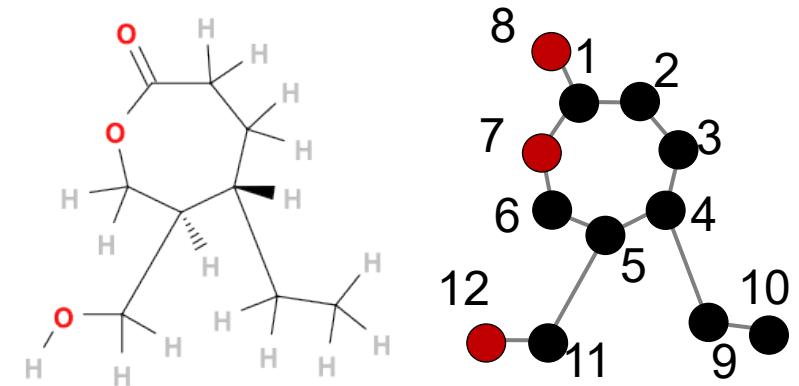
Common considerations:

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.



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

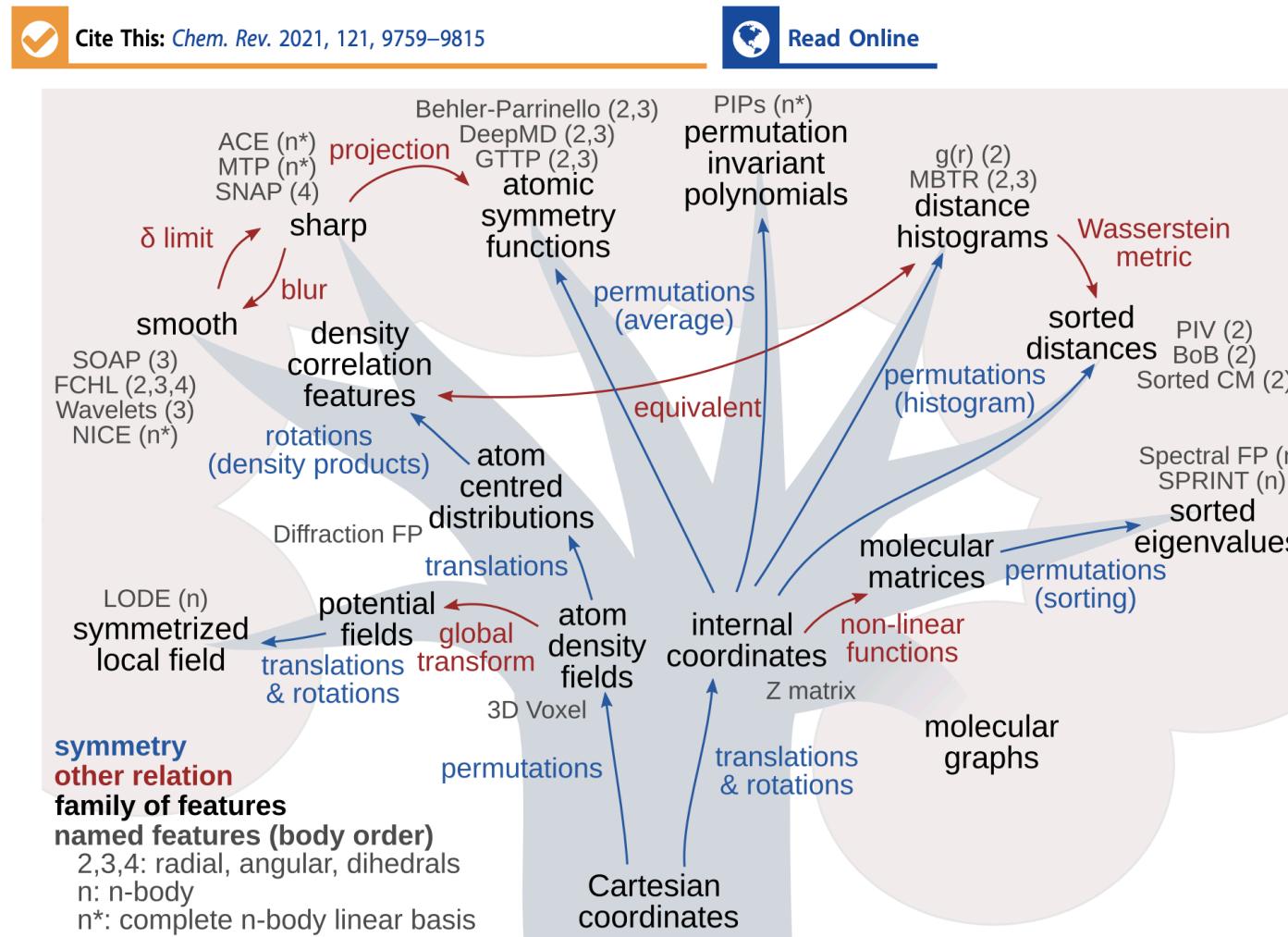
Think about shifting the molecule, rotating it, or re-labeling the atoms.

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*

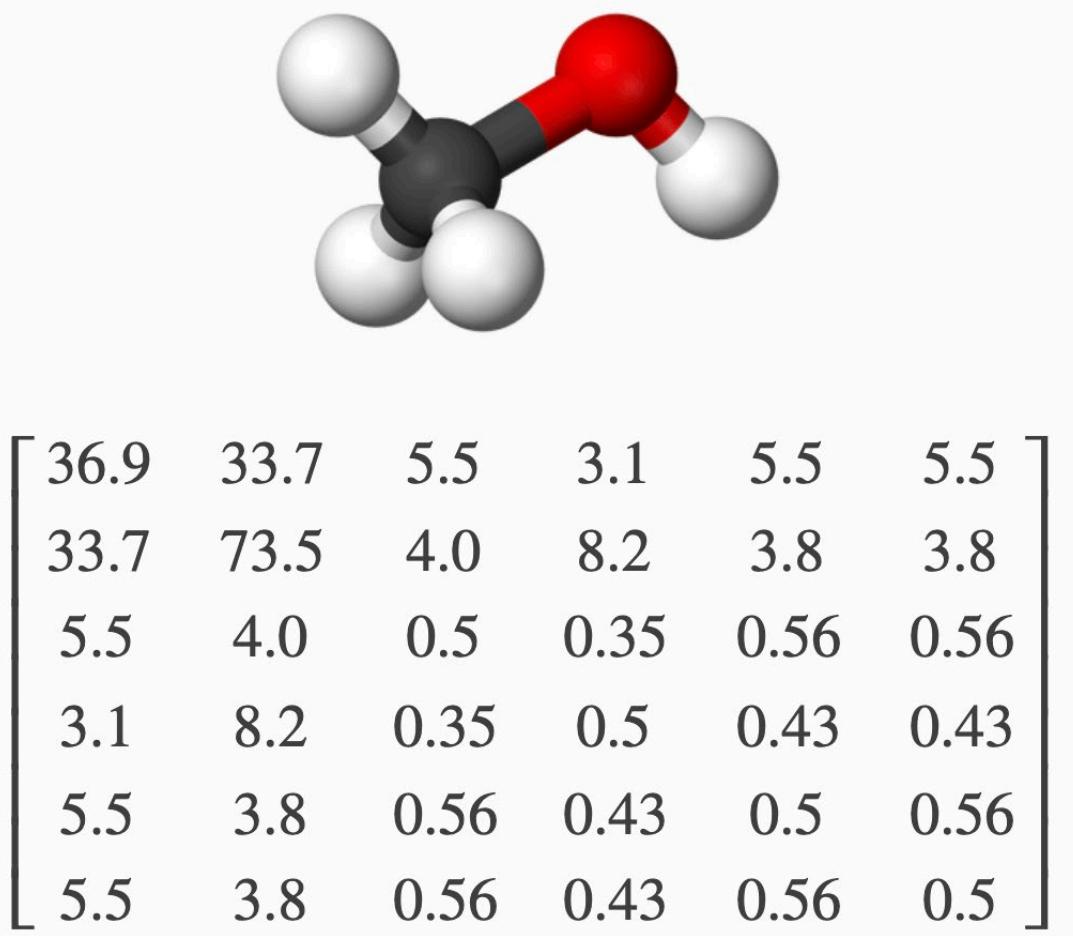


Some Examples

Coulomb Matrix

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

- represents entire molecule/fragment



Some Examples

Atom-centered Symmetry Functions

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

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

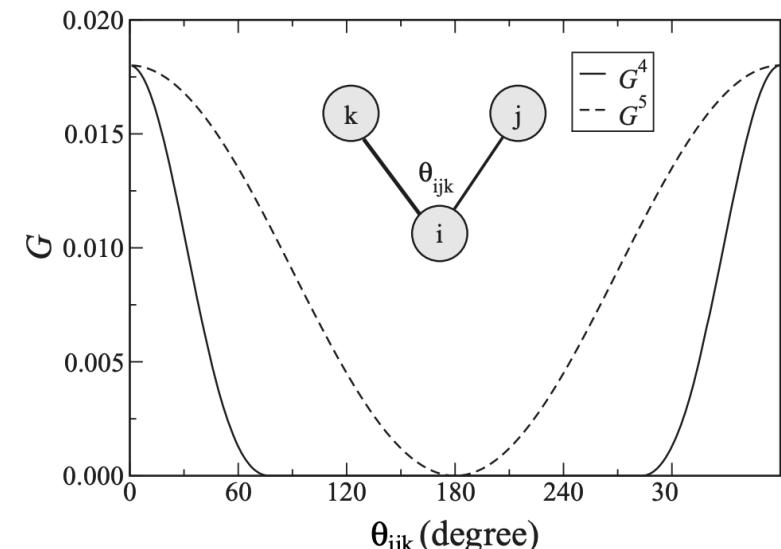
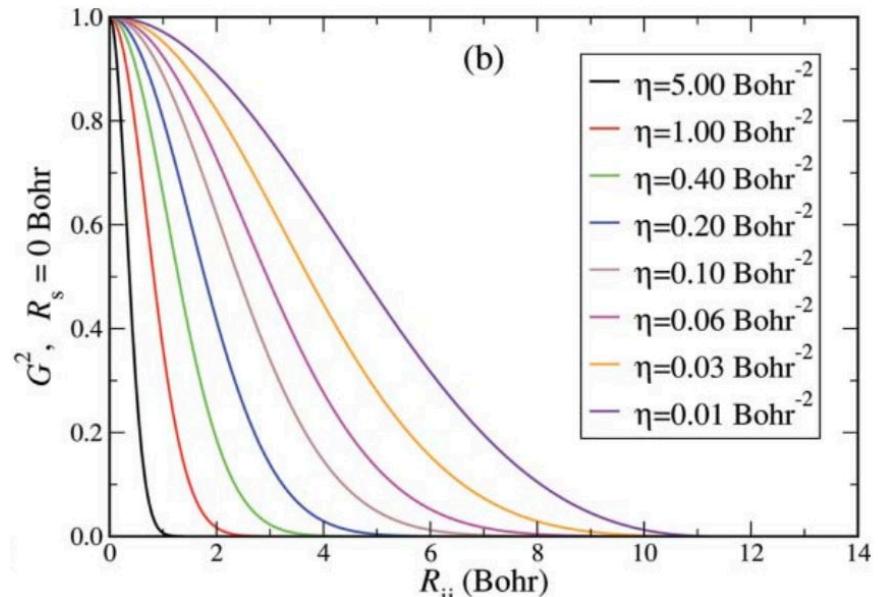
$$G_i^2 = \sum_j e^{-\eta(r_{ij}-r_s)^2} f_c(r_{ij})$$

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

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

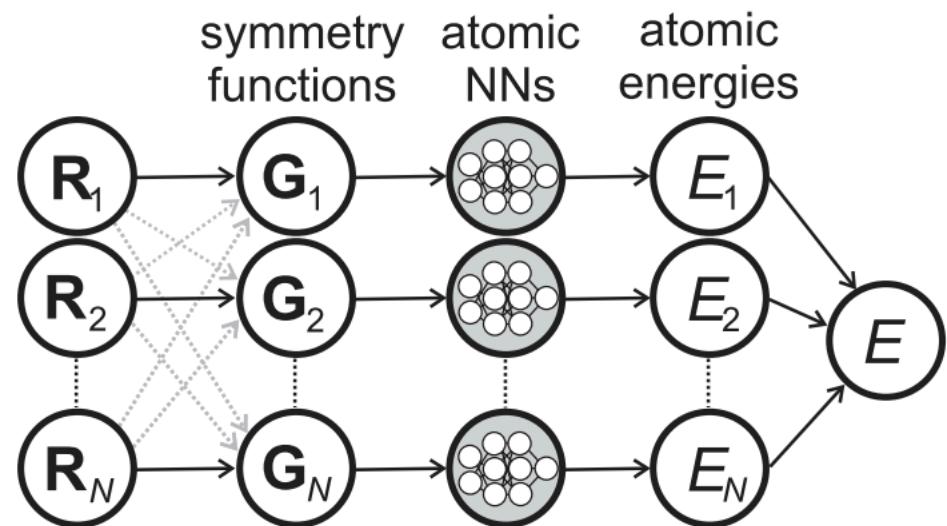
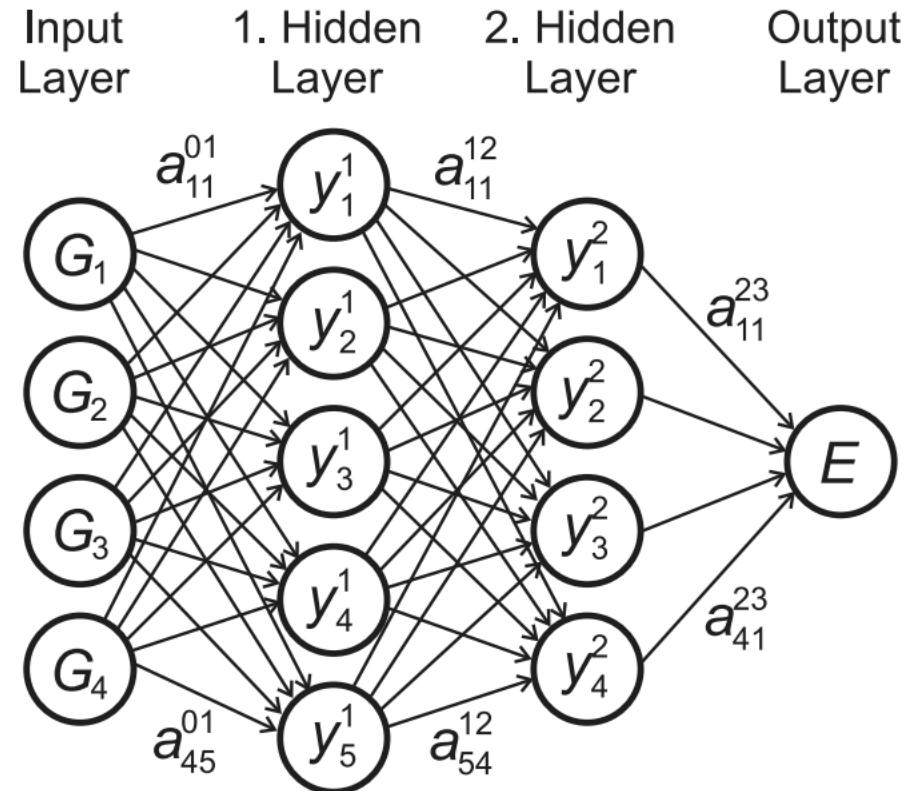
$$G_i^5 = 2^{1-\zeta} \sum_{j,k \neq i} (1 + \lambda \cos \theta_{ijk})^\zeta e^{-\eta(r_{ij}^2 + r_{ik}^2)} f_c(r_{ij}) f_c(r_{ik})$$

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



Some Examples

Atom-centered Symmetry Functions: Application to “Machine Learning Potentials”



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

random compilation
thermodynamic property prediction model

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
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 ($^{\circ}\text{C}$)	Stenutz ⁴⁵
Boiling point ($^{\circ}\text{C}$)	Stenutz ⁴⁵
Viscosity (cP)	COSMOtherm ³⁹
$\ln P_{\text{octanol-water}}$ partition coefficient (—)	COSMOtherm ³⁹
Vapour pressure (mbar)	COSMOtherm ³⁹
Henry's constant of H_2 in solvent (bar)	COSMOtherm ³⁹
$\ln(\gamma)$ activity coefficient of I in solvent (—)	COSMOtherm ³⁹
$\sigma'_1 - \sigma'_3$ profiles segmented into three (—)	COSMOtherm ³⁹
$\sigma_1 - \sigma_5$ profiles segmented into five (—)	COSMOtherm ³⁹
$t_1 - t_4$: principal components from PCA (—)	COSMOtherm ³⁹
—	—

Descriptor (Feature) Vectors

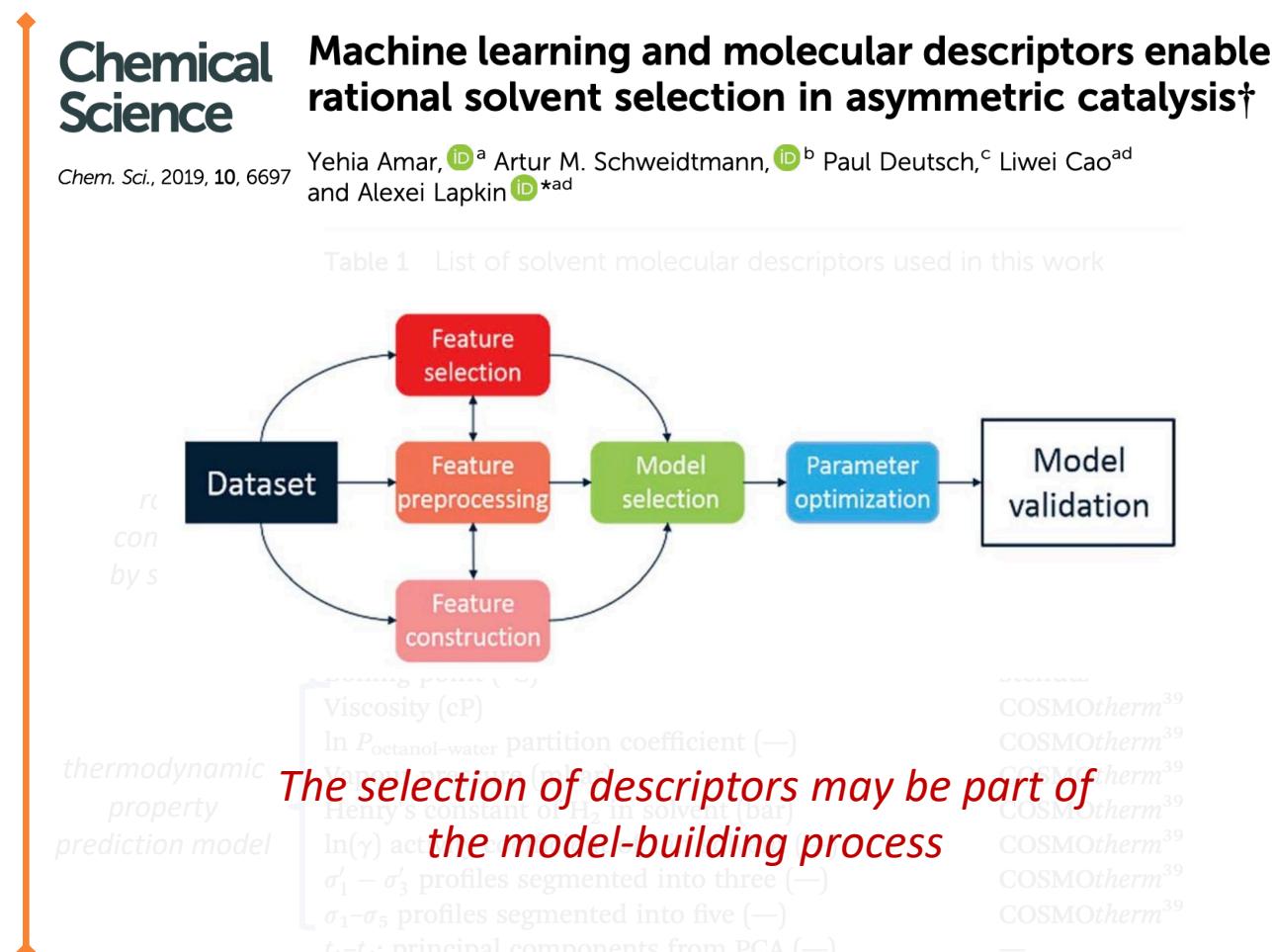
Vectors of physiochemical properties or system characteristics may also provide good representations of inputs for ML tasks

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.



Descriptor (Feature) Vectors

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

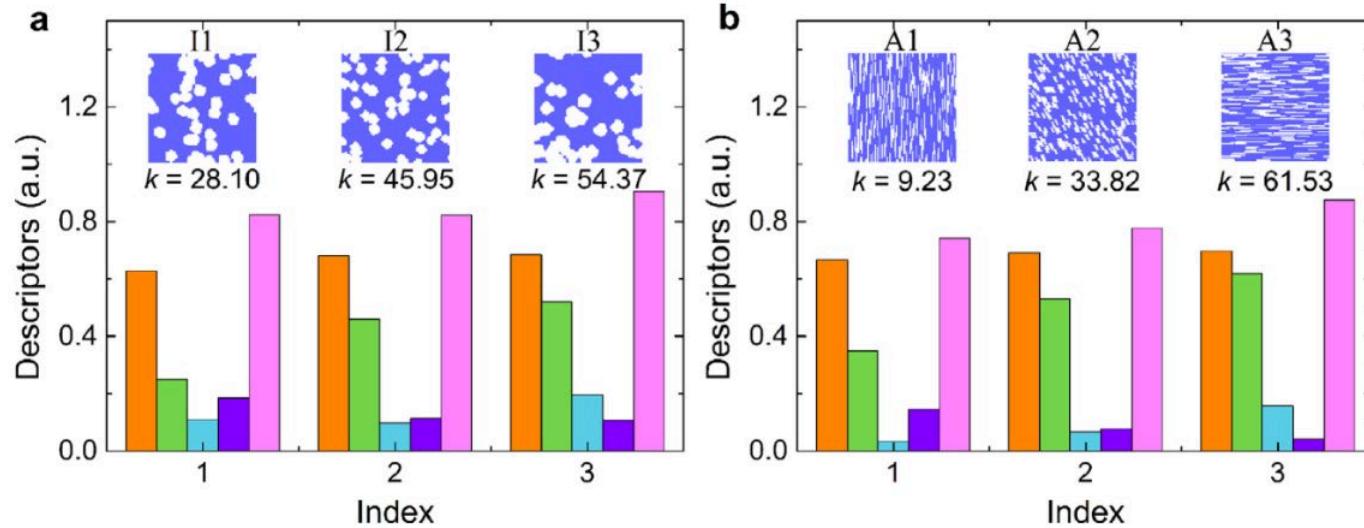
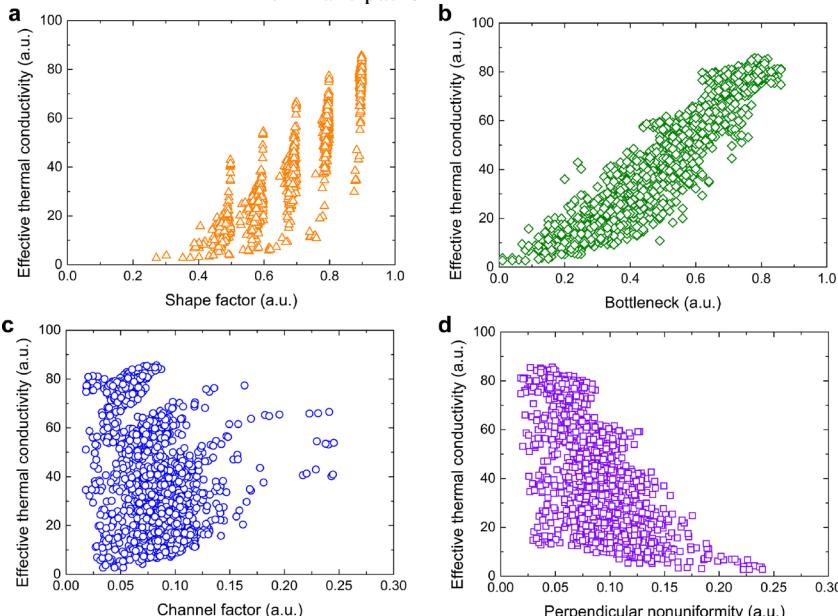


Machine learning prediction of thermal transport in porous media with physics-based descriptors

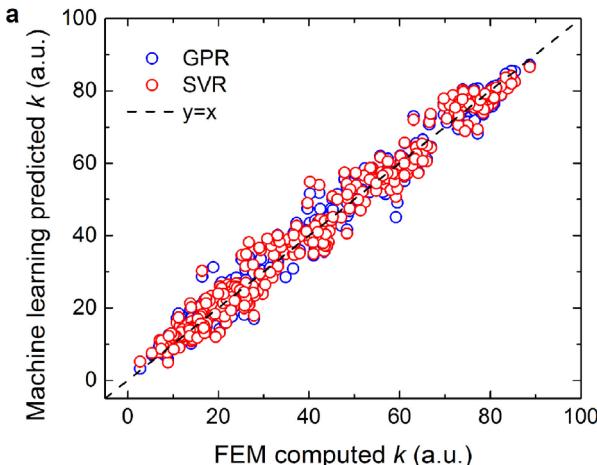
Han Wei^a, Hua Bao^{a,*}, Xiulin Ruan^{b,*}

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

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^{*,†,||}

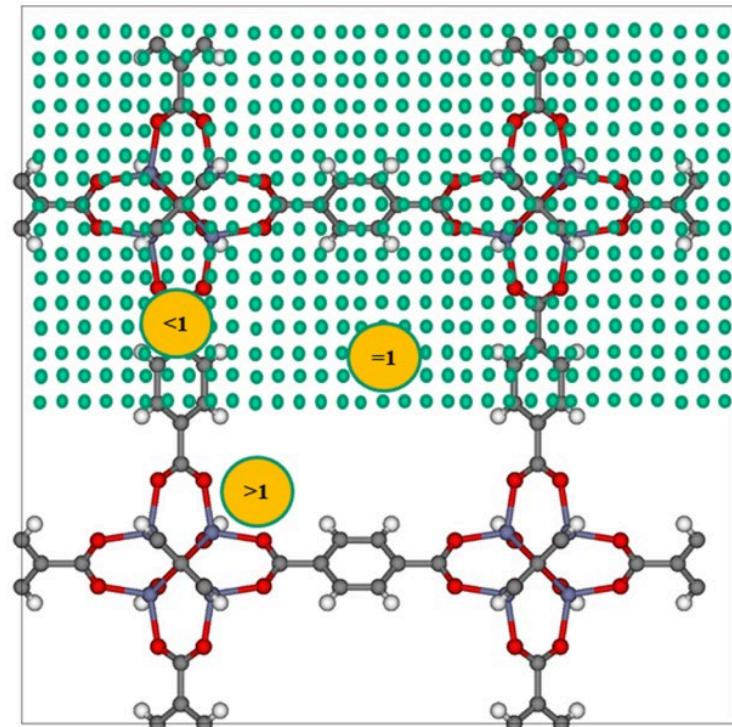
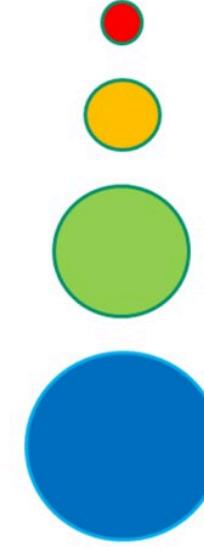
Table 1. Descriptors Used in the Present Study in the ML Methods^a

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

materials specific

readily available widely used

creative easily computed



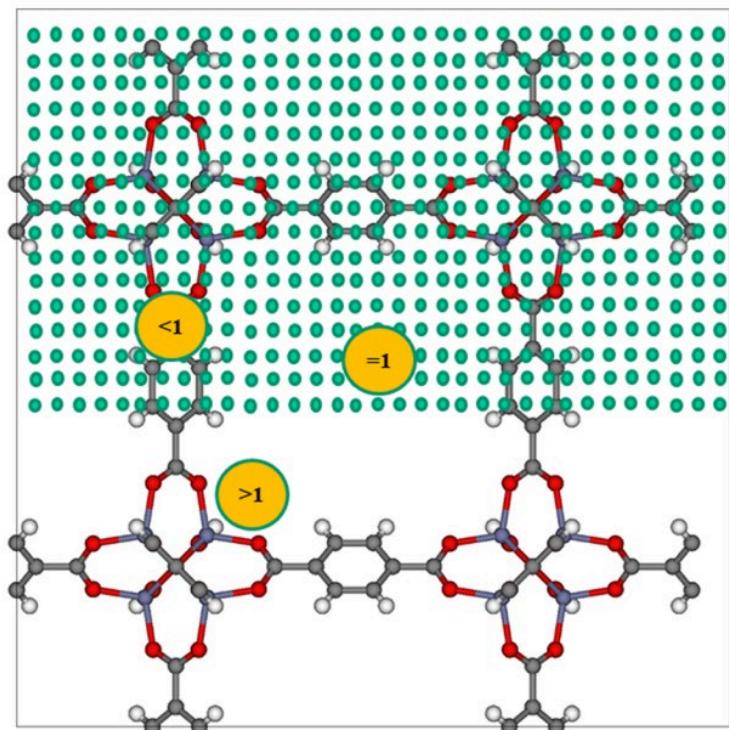
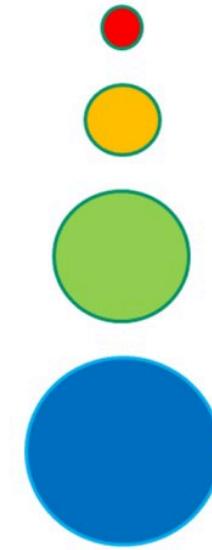
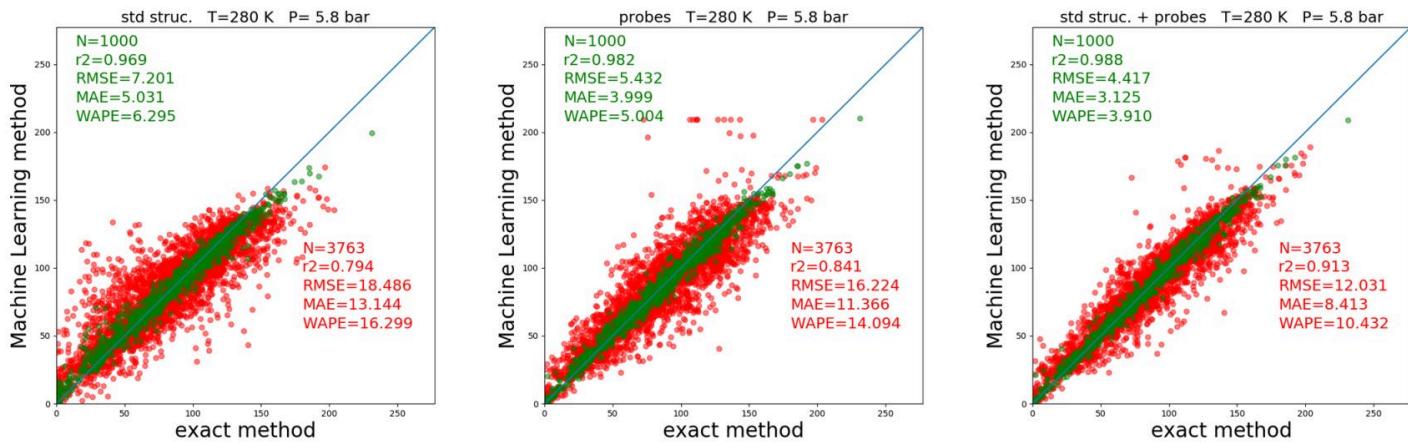
$$\text{probe - } (a) = \frac{1}{N} \sum_{i=1}^N \exp(-\beta E_i^{(a)})$$

Descriptor (Feature) Vectors

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

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

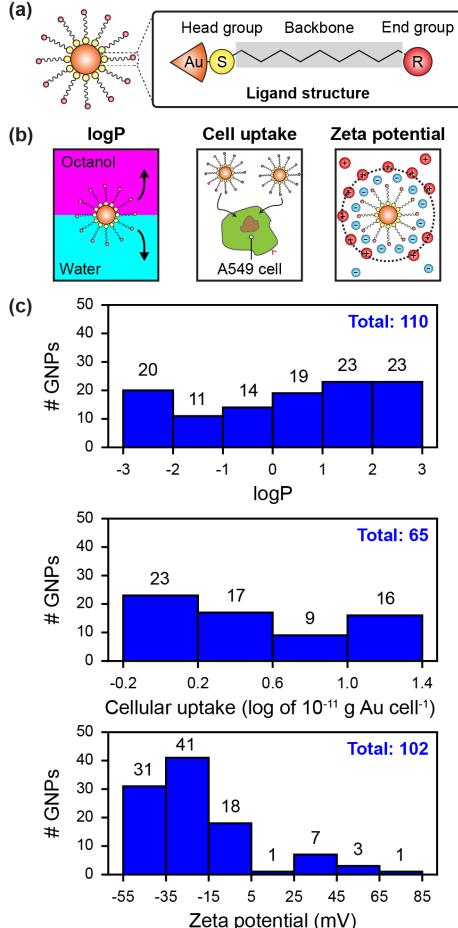
George S. Fanourgakis,^{*,†} Konstantinos Gkagkas,^{‡,§} Emmanuel Tylianakis,[¶] Emmanuel Klontzas,^{†,§} and George Froudakis^{*,†,||}



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

Simulation-derived Descriptors

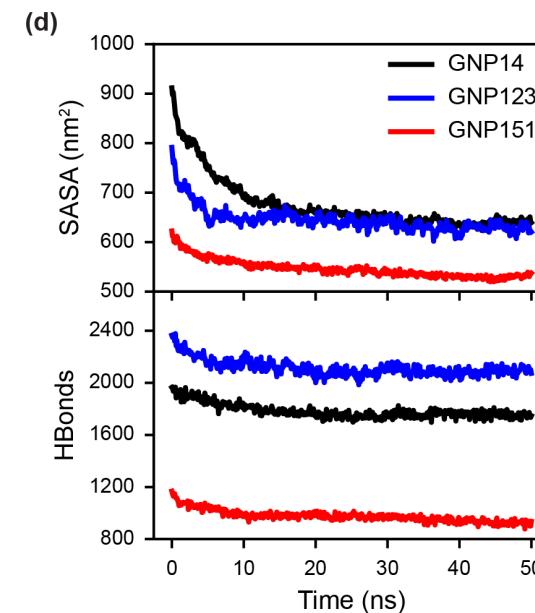
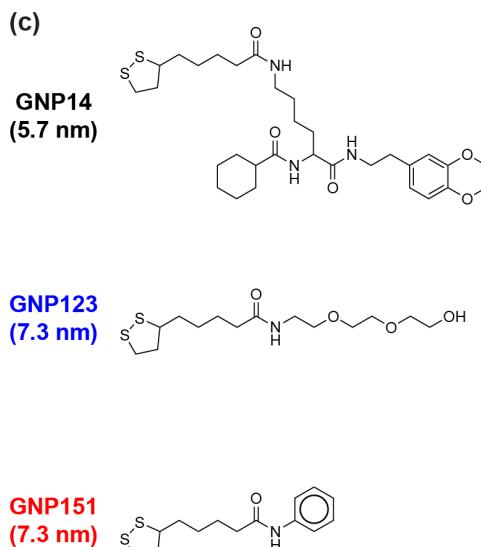
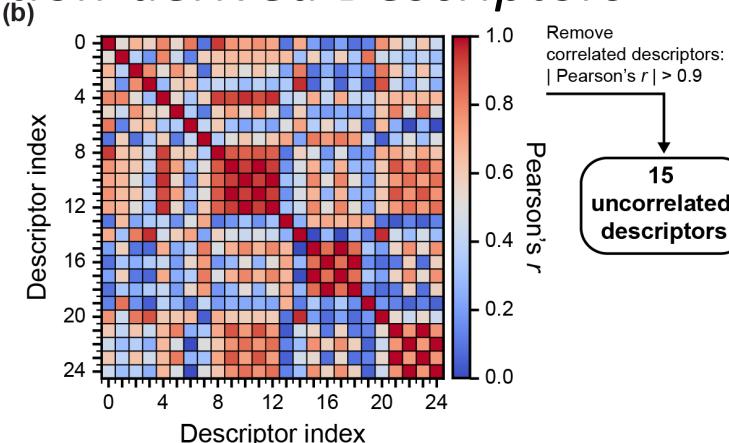
Experimental Dataset



(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>i.e.</i> 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

Simulation-derived Descriptors



ML Models

