Molecular Representation

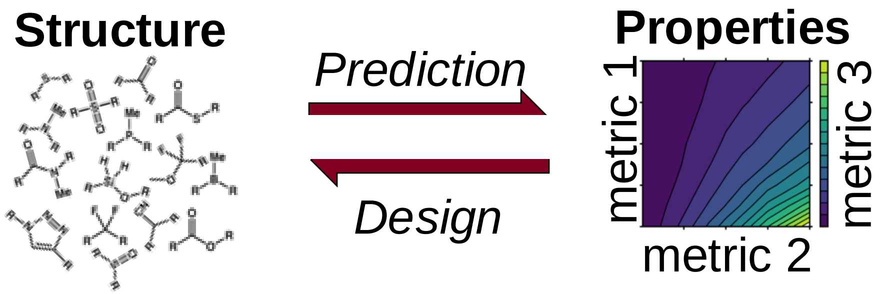
From Lecture 11 (basic deep learning with Keras)

- ***Keras*** provides a convenient programming interface for deep learning that allows us to have efficient/verified implementations of many standard procedures/algorithms relevant to deep neural networks
- For simple feed-forward, densely connected neural networks, one may use the **Sequential API**. For more complicated architectures, the **Functional API** may be needed.
- The basic procedure is *build* → *compile* → *fit* → *predict*
- In terms of optimizers, **NAdam** is usually a solid choice, but there are many options available
- In terms of loss functions, there are again many available, but it is easy to define your own should any of those be inadequate

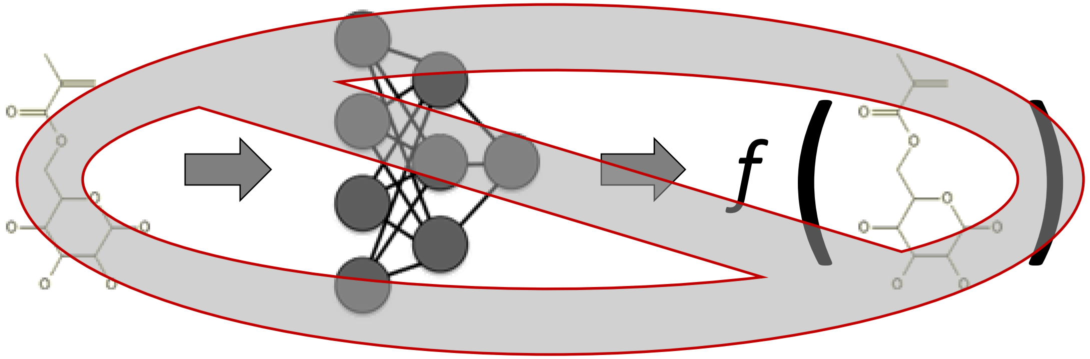
Machine Learning Meets Molecules

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 this process/goal, and we are interested in using ML to facilitate it



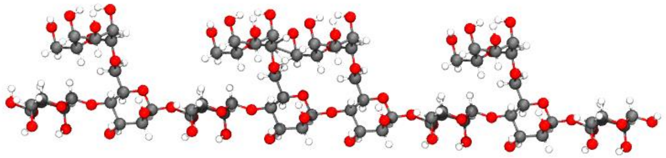
However, stick drawing or chemical name is a bit of problem for a neural network or other ML algorithm

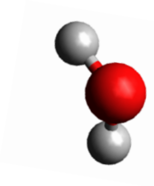
- Need technical methods to convert molecular structures into machine-readable formats (e.g., numerical vectors) that can be processed as inputs to ML algorithms
- The transformation should carry information that is useful to the prediction task in distinguishing amongst chemical moieties

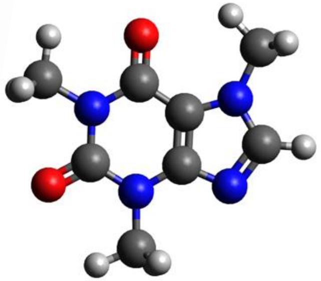
This conversion process is referred to as ***molecular featurization***

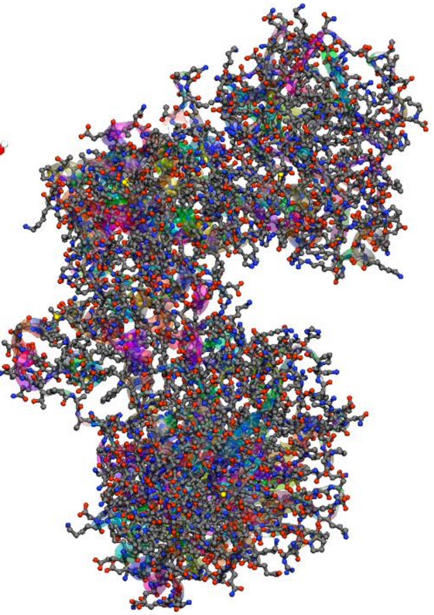
What is a molecule?

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



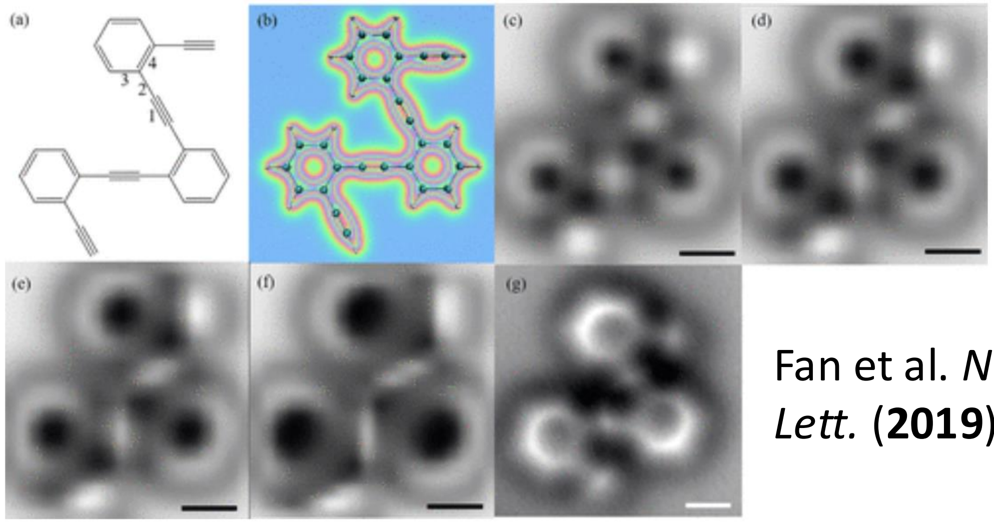




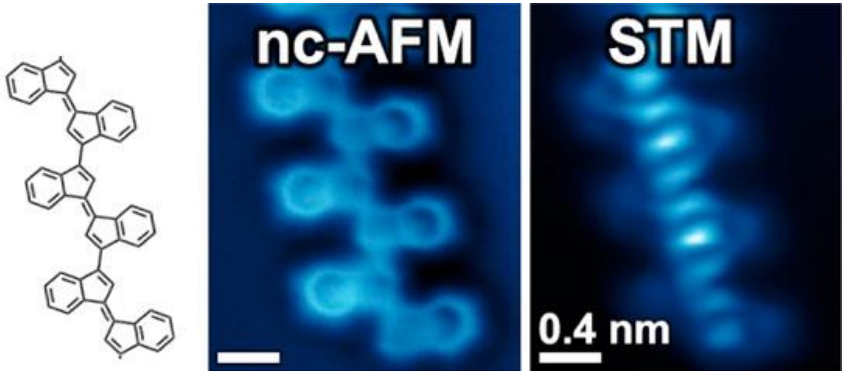


Some possible characteristics of molecules

- 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?

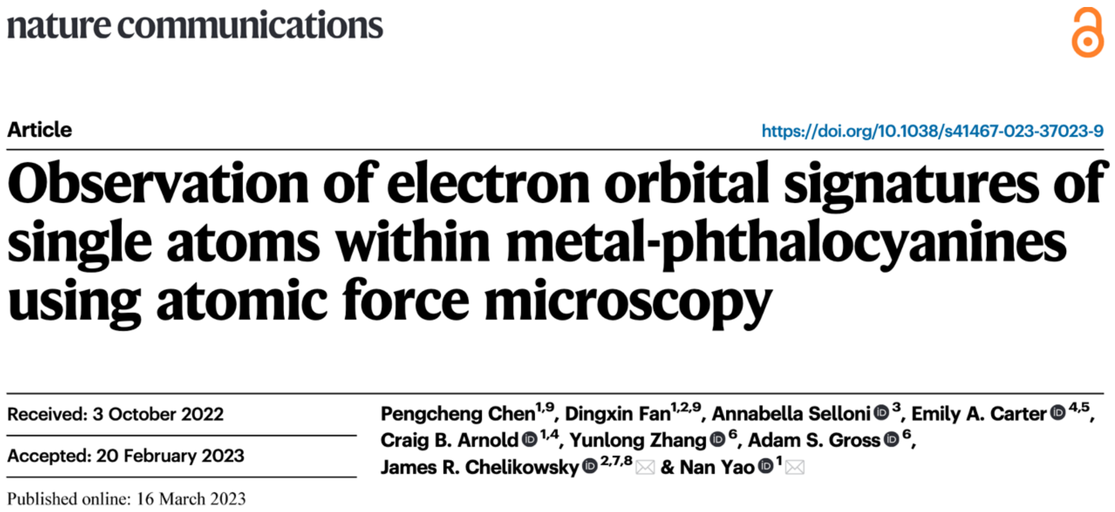


Fan et al. *Nano Lett.* (**2019**)



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

What is a molecule?



nature communications

Article

https://doi.org/10.1038/s41467-023-37023-9

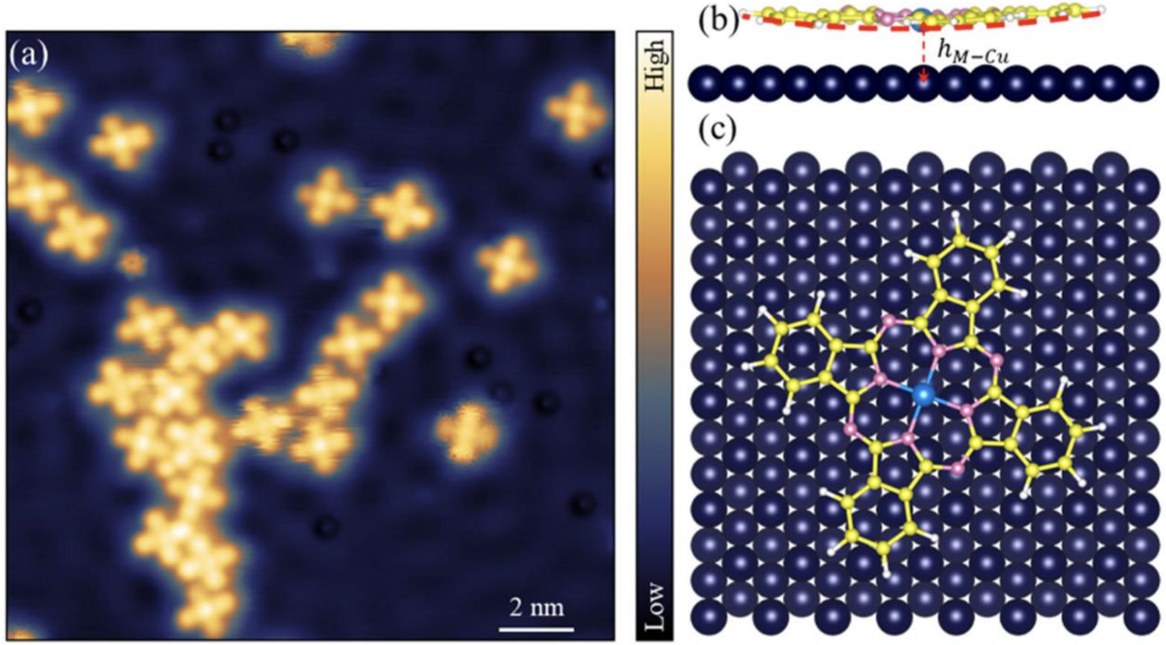
Observation of electron orbital signatures of single atoms within metal-phthalocyanines using atomic force microscopy

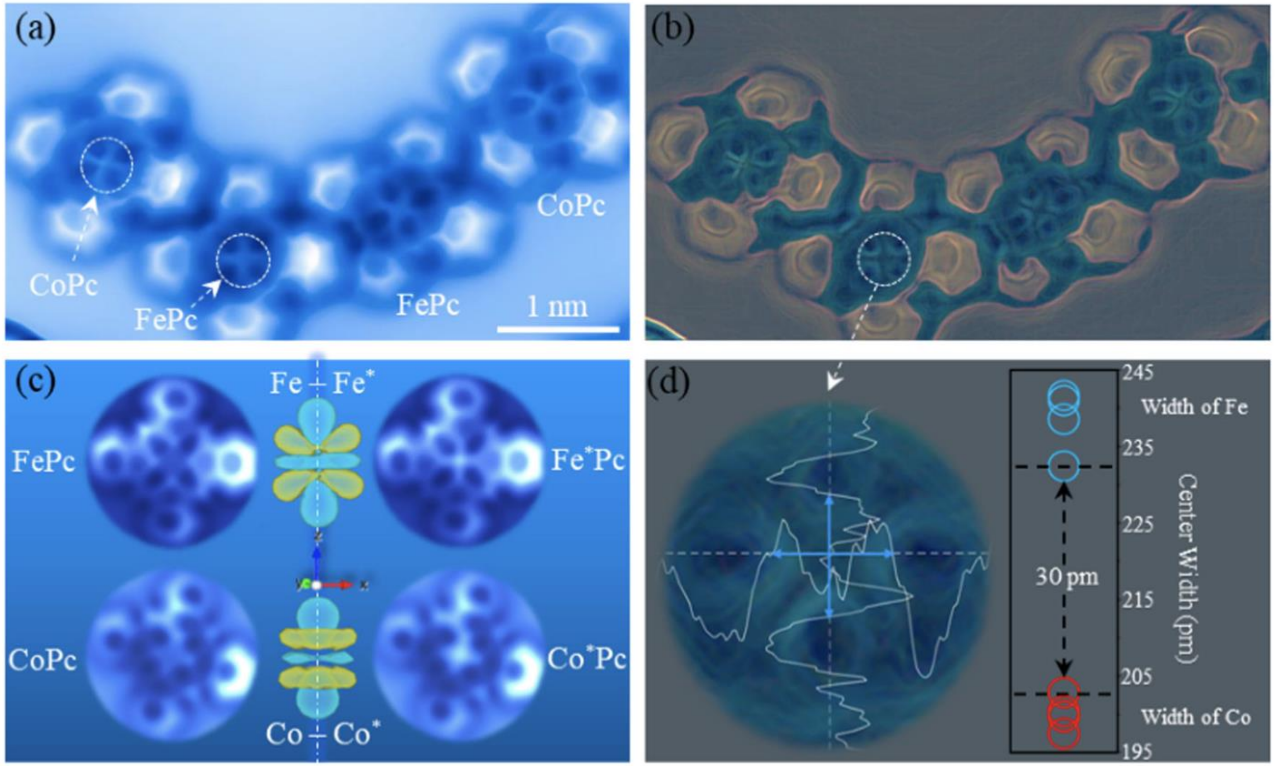
Received: 3 October 2022

Accepted: 20 February 2023

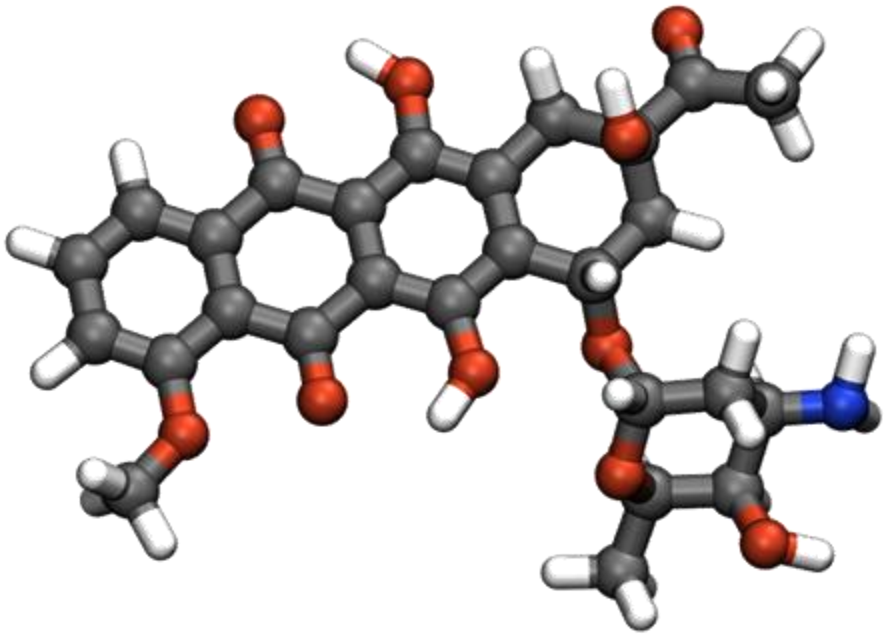
Published online: 16 March 2023

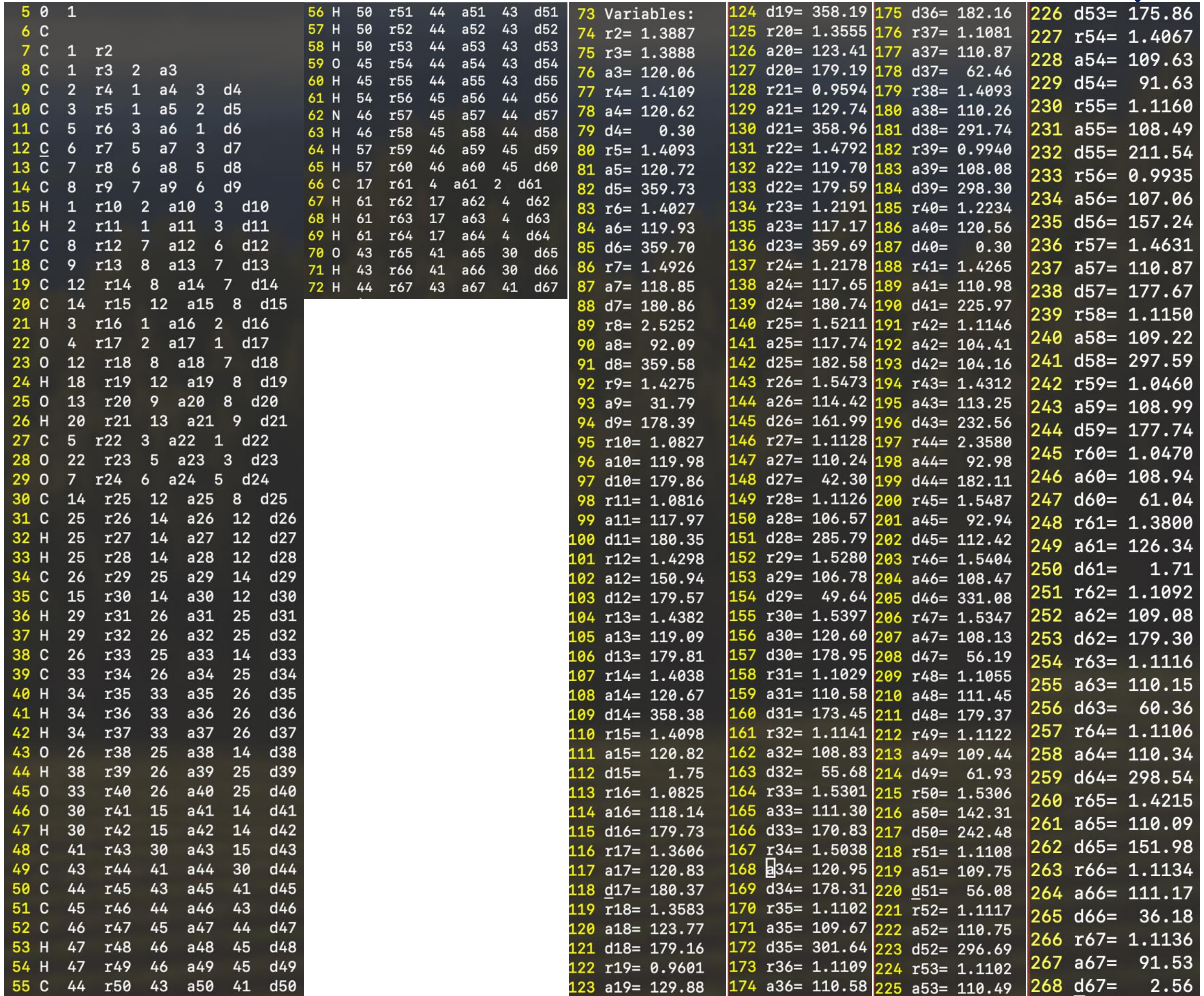
Pengcheng Chen[1,9], Dingxin Fan[1,2,9], Annabella Selloni[3], Emily A. Carter[4,5], Craig B. Arnold[1,4], Yunlong Zhang[6], Adam S. Gross[6], James R. Chelikowsky[2,7,8] & Nan Yao[1]





Molecules as particles… or waves





```
0  1
C
C  1  r2
C  1  r3  2  a3
C  2  r4  1  a4  3  d4
C  3  r5  1  a5  2  d5
C  5  r6  3  a6  1  d6
C  6  r7  5  a7  3  d7
C  7  r8  6  a8  5  d8
C  8  r9  7  a9  6  d9
H  1  r10  2  a10  3  d10
H  2  r11  1  a11  3  d11
C  8  r12  7  a12  6  d12
C  9  r13  8  a13  7  d13
C  12  r14  8  a14  7  d14
C  14  r15  12  a15  8  d15
H  3  r16  1  a16  2  d16
O  4  r17  2  a17  1  d17
O  12  r18  8  a18  7  d18
H  18  r19  12  a19  8  d19
O  13  r20  9  a20  8  d20
H  20  r21  13  a21  9  d21
C  5  r22  3  a22  1  d22
O  22  r23  5  a23  3  d23
O  7  r24  6  a24  5  d24
C  14  r25  12  a25  8  d25
C  25  r26  14  a26  12  d26
H  25  r27  14  a27  12  d27
H  25  r28  14  a28  12  d28
C  26  r29  25  a29  14  d29
C  15  r30  14  a30  12  d30
H  29  r31  26  a31  25  d31
H  29  r32  26  a32  25  d32
C  26  r33  25  a33  14  d33
C  33  r34  26  a34  25  d34
H  34  r35  33  a35  26  d35
H  34  r36  33  a36  26  d36
H  34  r37  33  a37  26  d37
O  26  r38  25  a38  14  d38
H  38  r39  26  a39  25  d39
O  33  r40  26  a40  25  d40
O  30  r41  15  a41  14  d41
H  30  r42  15  a42  14  d42
C  41  r43  30  a43  15  d43
C  43  r44  41  a44  30  d44
C  44  r45  43  a45  41  d45
C  45  r46  44  a46  43  d46
C  46  r47  45  a47  44  d47
H  47  r48  46  a48  45  d48
H  47  r49  46  a49  45  d49
C  44  r50  43  a50  41  d50
```

```
H  50  r51  44  a51  43  d51
H  50  r52  44  a52  43  d52
H  50  r53  44  a53  43  d53
O  45  r54  44  a54  43  d54
H  45  r55  44  a55  43  d55
H  54  r56  45  a56  44  d56
N  46  r57  45  a57  44  d57
H  46  r58  45  a58  44  d58
H  57  r59  46  a59  45  d59
H  57  r60  46  a60  45  d60
C  17  r61  4  a61  2  d61
H  61  r62  17  a62  4  d62
H  61  r63  17  a63  4  d63
H  61  r64  17  a64  4  d64
O  43  r65  41  a65  30  d65
H  43  r66  41  a66  30  d66
H  44  r67  43  a67  41  d67
```

```
Variables:
r2= 1.3887
r3= 1.3888
a3= 120.06
r4= 1.4109
a4= 120.62
d4=   0.30
r5= 1.4093
a5= 120.72
d5= 359.73
r6= 1.4027
a6= 119.93
d6= 359.70
r7= 1.4926
a7= 118.85
d7= 180.86
r8= 2.5252
a8=  92.09
d8= 359.58
r9= 1.4275
a9=  31.79
d9= 178.39
r10= 1.0827
a10= 119.98
d10= 179.86
r11= 1.0816
a11= 117.97
d11= 180.35
r12= 1.4298
a12= 150.94
d12= 179.57
r13= 1.4382
a13= 119.09
d13= 179.81
r14= 1.4038
a14= 120.67
d14= 358.38
r15= 1.4098
a15= 120.82
d15=   1.75
r16= 1.0825
a16= 118.14
d16= 179.73
r17= 1.3606
a17= 120.83
d17= 180.37
r18= 1.3583
a18= 123.77
d18= 179.16
r19= 0.9601
a19= 129.88
```

```
d19= 358.19
r20= 1.3555
a20= 123.41
d20= 179.19
r21= 0.9594
a21= 129.74
d21= 358.96
r22= 1.4792
a22= 119.70
d22= 179.59
r23= 1.2191
a23= 117.17
d23= 359.69
r24= 1.2178
a24= 117.65
d24= 180.74
r25= 1.5211
a25= 117.74
d25= 182.58
r26= 1.5473
a26= 114.42
d26= 161.99
r27= 1.1128
a27= 110.24
d27=  42.30
r28= 1.1126
a28= 106.57
d28= 285.79
r29= 1.5280
a29= 106.78
d29=  49.64
r30= 1.5397
a30= 120.60
d30= 178.95
r31= 1.1029
a31= 110.58
d31= 173.45
r32= 1.1141
a32= 108.83
d32=  55.68
r33= 1.5301
a33= 111.30
d33= 170.83
r34= 1.5038
a34= 120.95
d34= 178.31
r35= 1.1102
a35= 109.67
d35= 301.64
r36= 1.1109
a36= 110.58
```

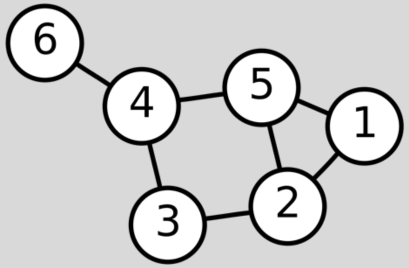
```
d36= 182.16
r37= 1.1081
a37= 110.87
d37=  62.46
r38= 1.4093
a38= 110.26
d38= 291.74
r39= 0.9940
a39= 108.08
d39= 298.30
r40= 1.2234
a40= 120.56
d40=   0.30
r41= 1.4265
a41= 110.98
d41= 225.97
r42= 1.1146
a42= 104.41
d42= 104.16
r43= 1.4312
a43= 113.25
d43= 232.56
r44= 2.3580
a44=  92.98
d44= 182.11
r45= 1.5487
a45=  92.94
d45= 112.42
r46= 1.5404
a46= 108.47
d46= 331.08
r47= 1.5347
a47= 108.13
d47=  56.19
r48= 1.1055
a48= 111.45
d48= 179.37
r49= 1.1122
a49= 109.44
d49=  61.93
r50= 1.5306
a50= 142.31
d50= 242.48
r51= 1.1108
a51= 109.75
d51=  56.08
r52= 1.1117
a52= 110.75
d52= 296.69
r53= 1.1102
a53= 110.49
```

```
d53= 175.86
r54= 1.4067
a54= 109.63
d54=  91.63
r55= 1.1160
a55= 108.49
d55= 211.54
r56= 0.9935
a56= 107.06
d56= 157.24
r57= 1.4631
a57= 110.87
d57= 177.67
r58= 1.1150
a58= 109.22
d58= 297.59
r59= 1.0460
a59= 108.99
d59= 177.74
r60= 1.0470
a60= 108.94
d60=  61.04
r61= 1.3800
a61= 126.34
d61=   1.71
r62= 1.1092
a62= 109.08
d62= 179.30
r63= 1.1116
a63= 110.15
d63=  60.36
r64= 1.1106
a64= 110.34
d64= 298.54
r65= 1.4215
a65= 110.09
d65= 151.98
r66= 1.1134
a66= 111.17
d66=  36.18
r67= 1.1136
a67=  91.53
d67=   2.56
```

Molecular Graphs

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

Mathematical graph



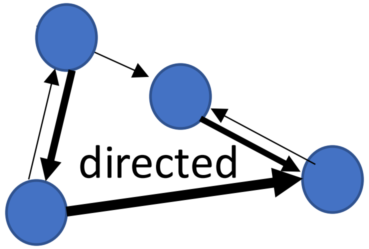
$$\mathcal{G} = (\mathcal{V}, \mathcal{E})$$

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

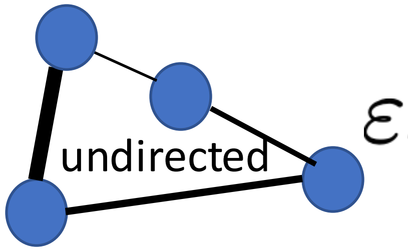
$\mathcal{V}$ ***Vertices/nodes*** – indicates objects

$\mathcal{E}$ ***Edges*** – indicates pairwise relationship amongst objects

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



vs.



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

Graphs can be conveniently represented as matrices

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

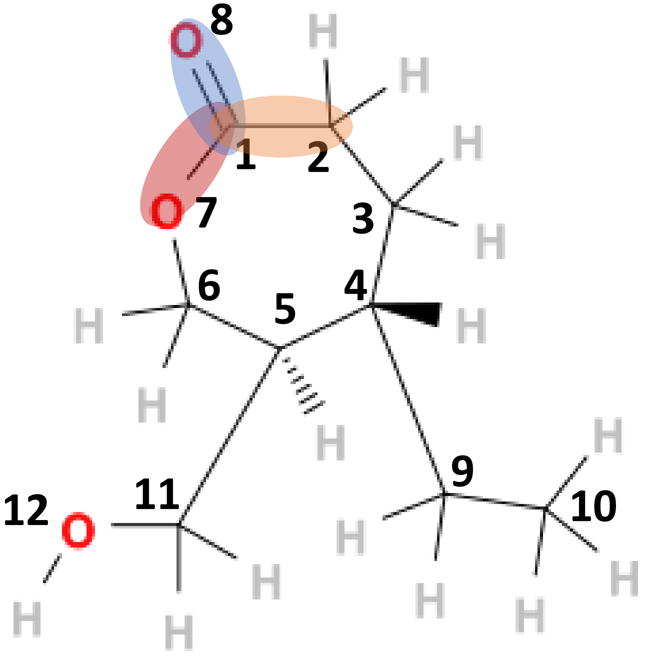
<u>e.g.,</u>

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

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 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.

This representation is typically deficient in the description of 3D structure/conformation, chirality, …

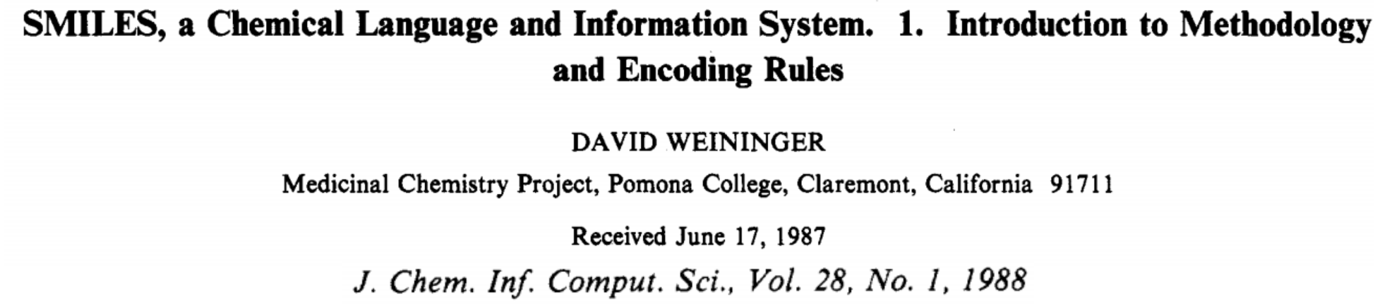
Text-based Representations

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)



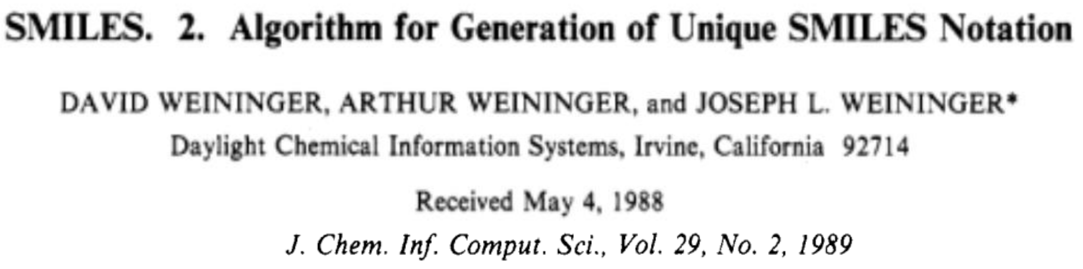
SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules

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*

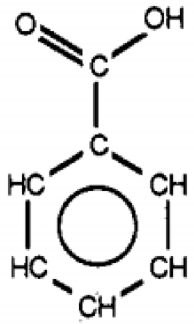
Daylight Chemical Information Systems, Irvine, California 92714

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*



c1ccccc1C(=0)O

not natively canonical (additional algorithms do this)

Text-based Representations

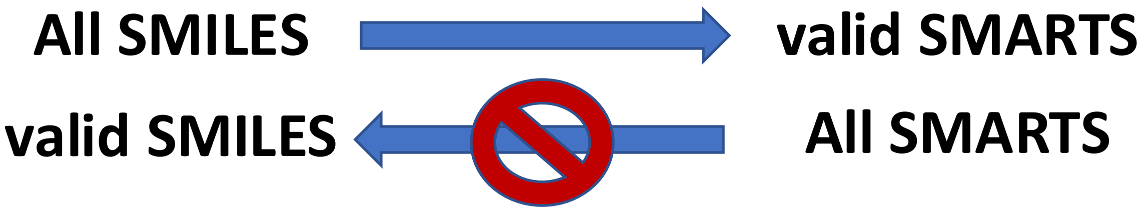
Common starting points for describing molecular graphs are text strings

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

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

→ database queries

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



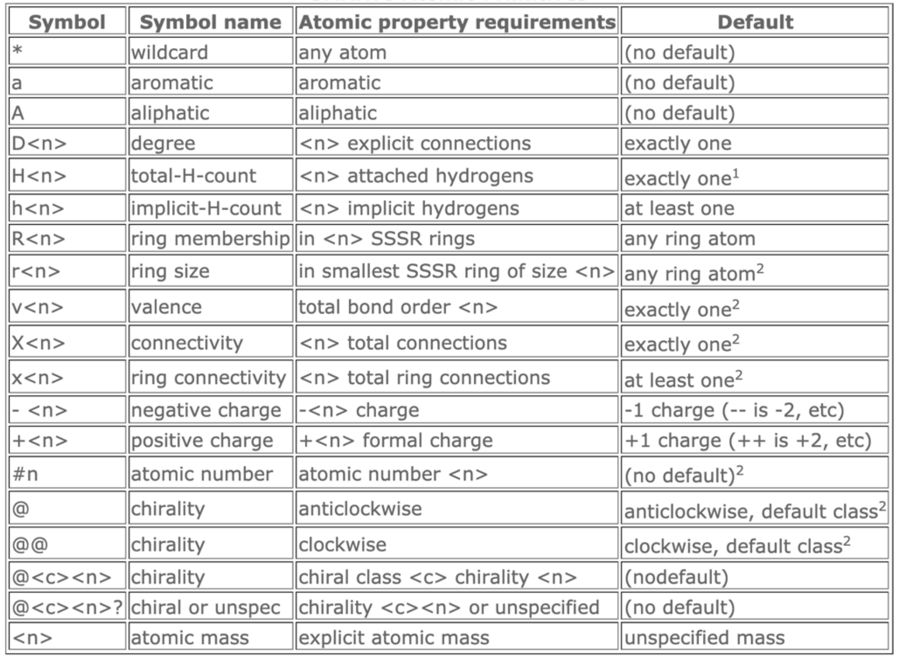
All SMILES

valid SMARTS

valid SMILES

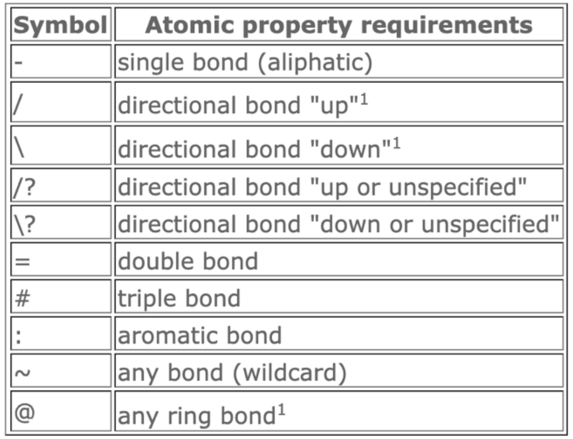
All SMARTS

SMARTS Atomic Primitives



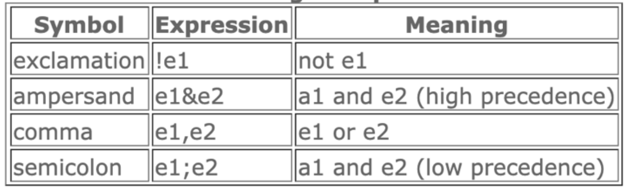
Symbol	Symbol name	Atomic property requirements	Default
*	wildcard	any atom	(no default)
a	aromatic	aromatic	(no default)
A	aliphatic	aliphatic	(no default)
D<n>	degree	<n> explicit connections	exactly one
H<n>	total-H-count	<n> attached hydrogens	exactly one[1]
h<n>	implicit-H-count	<n> implicit hydrogens	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)

Decoding Exercise

draw out/describe the substructures from SMARTS

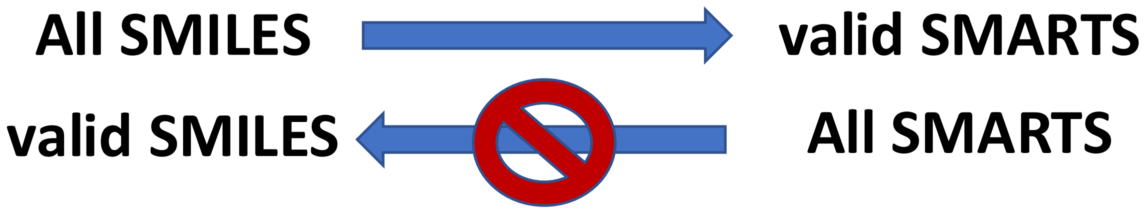
- “cc”
- [c,n;H1]
- “Caa(O)aN”
- “Ca(aO)aaN”

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

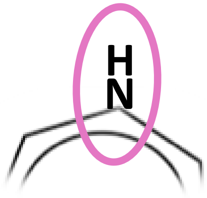
valid SMILES ← All SMARTS

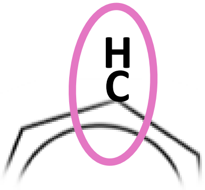
→ database queries

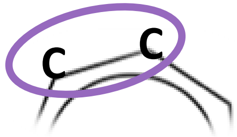
→ substructure searches *(finding a subgraph of the molecular graph)*

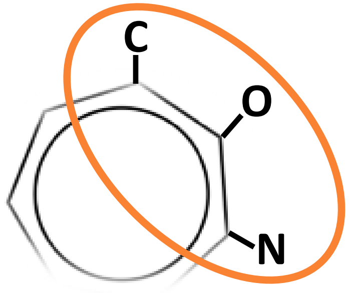
Decoding Exercise

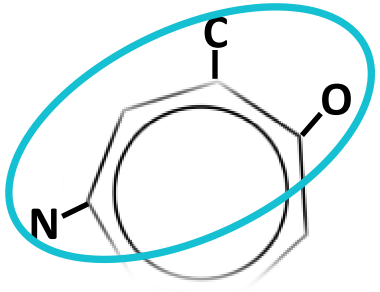
draw out/describe the substructures from SMARTS











any pair of bonded aromatic carbons

- “cc”
- [c,n;H1]

either aromatic carbon or nitrogen and exactly one hydrogen

- “Caa(O)aN”

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

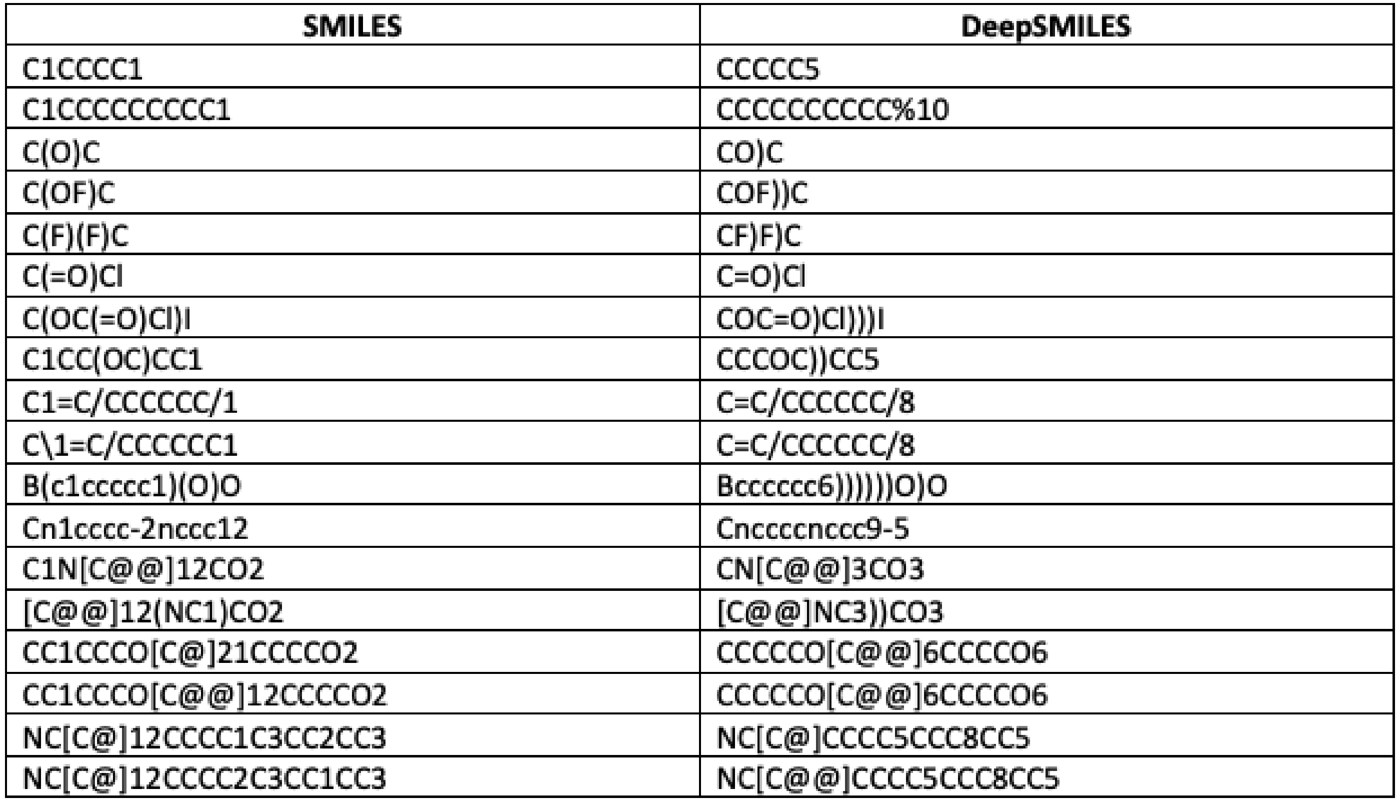
- “Ca(aO)aaN”

same as above but O and N likely para

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	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

10.26434/chemrxiv.7097960.v1

<u>Other string-based representations</u>

Wiswesser line notation

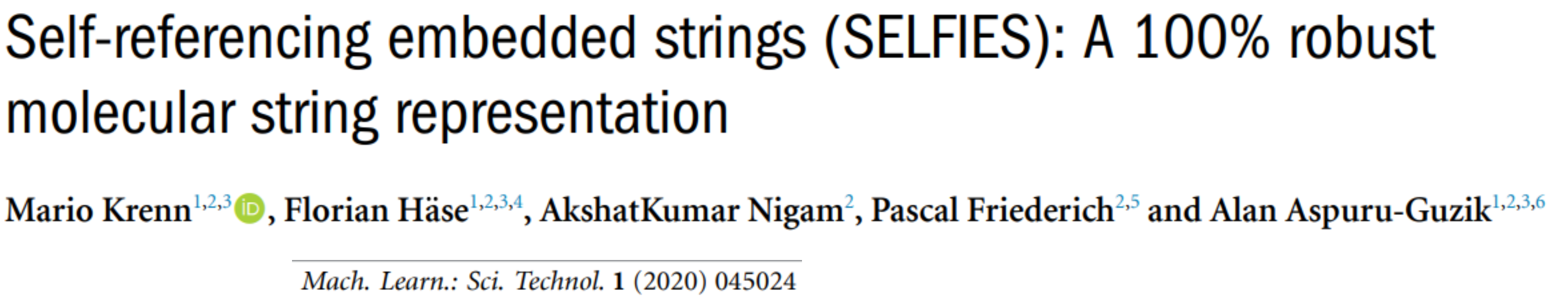
SYBYL Line Notation

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

Text-based Representations

Common starting points for describing molecular graphs are text strings

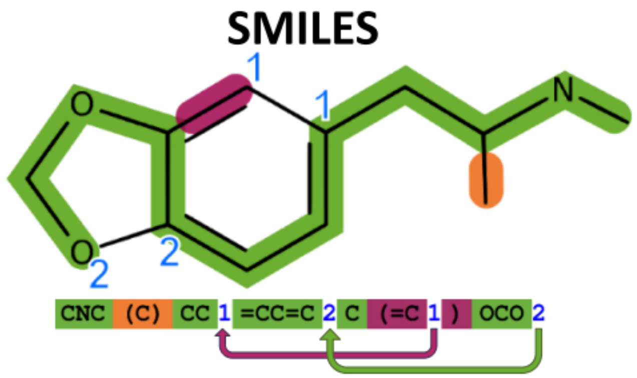
Self-referencing **E**mbedded **S**trings (SELFIES)

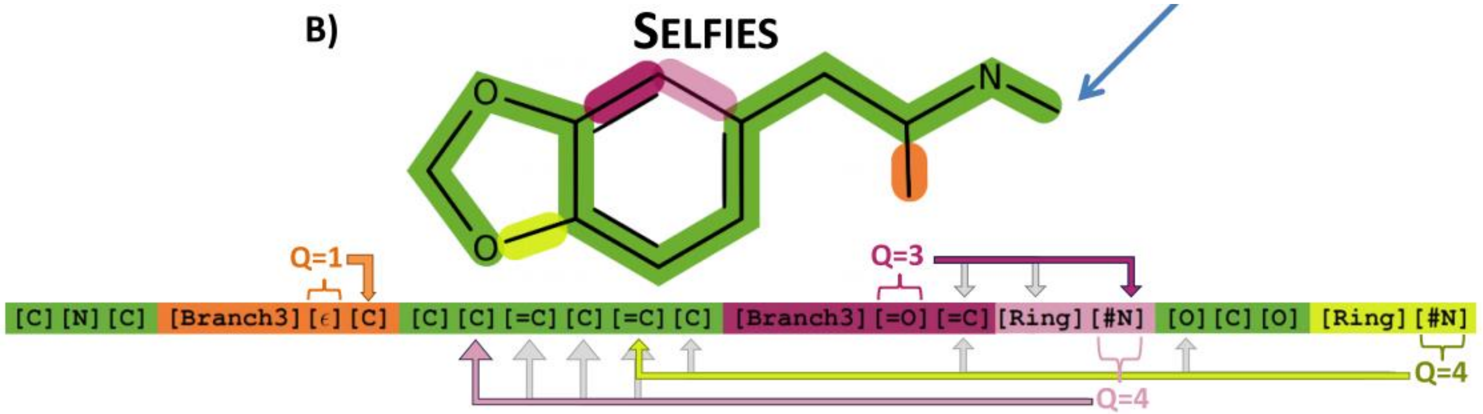


Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation

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

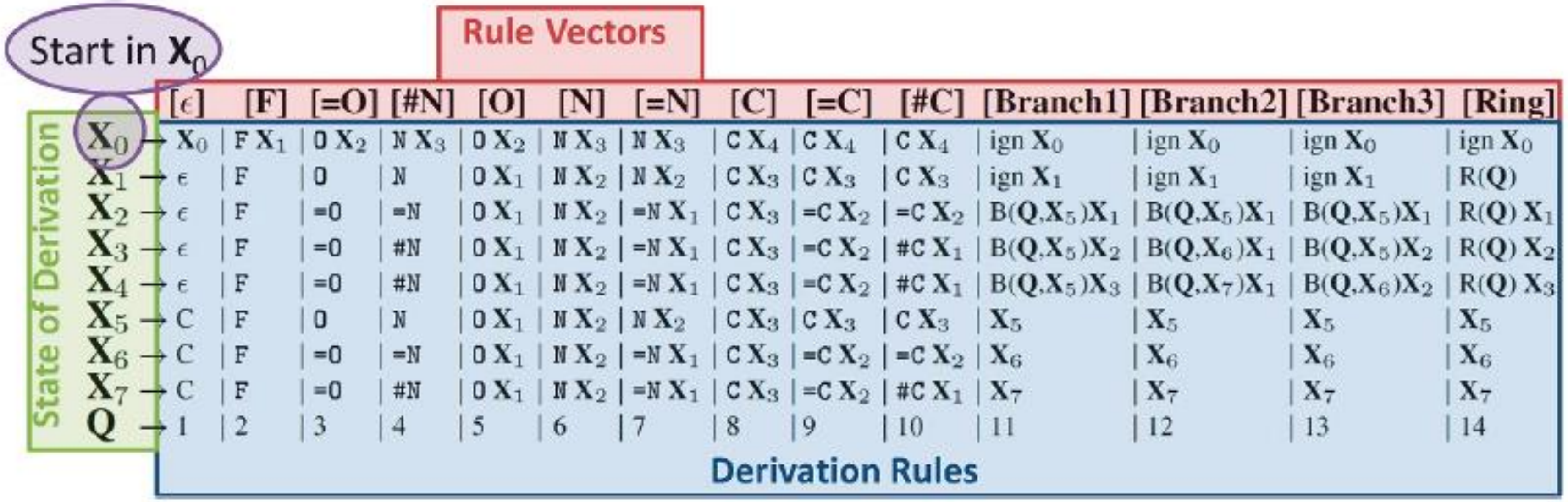
Self-referencing **E**mbedded **S**trings (SELFIES)

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 X_0
Rule Vectors
State of Derivation
Derivation Rules

	$[\epsilon]$	[F]	[=O]	[#N]	[O]	[N]	[=N]	[C]	[=C]	[#C]	[Branch1]	[Branch2]	[Branch3]	[Ring]
$X_0 \rightarrow$	X_0	F X_1	O X_2	N X_3	O 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
$X_1 \rightarrow$	ϵ	F	O	N	O 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)
$X_2 \rightarrow$	ϵ	F	=O	=N	O 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
$X_3 \rightarrow$	ϵ	F	=O	#N	O 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
$X_4 \rightarrow$	ϵ	F	=O	#N	O 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
$X_5 \rightarrow$	C	F	O	N	O X_1	N X_2	N X_2	C X_3	C X_3	C X_3	X_5	X_5	X_5	X_5
$X_6 \rightarrow$	C	F	=O	=N	O X_1	N X_2	=N X_1	C X_3	=C X_2	=C X_2	X_6	X_6	X_6	X_6
$X_7 \rightarrow$	C	F	=O	#N	O X_1	N X_2	=N X_1	C X_3	=C X_2	#C X_1	X_7	X_7	X_7	X_7
$Q \rightarrow$	1	2	3	4	5	6	7	8	9	10	11	12	13	14

<u>conversion to molecular graph</u>

[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