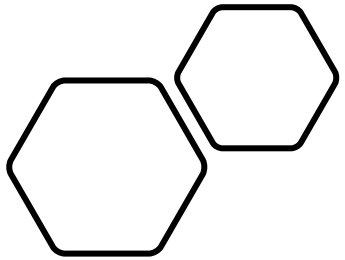
A tour through Molecular Representations in AI-driven Drug Discovery



Laurianne David

David, L., Thakkar, A., Mercado, R. *et al.* Molecular representations in AI-driven drug discovery: a review and practical guide. *J Cheminform* **12,** 56 (2020). https://doi.org/10.1186/s13321-020-00460-5

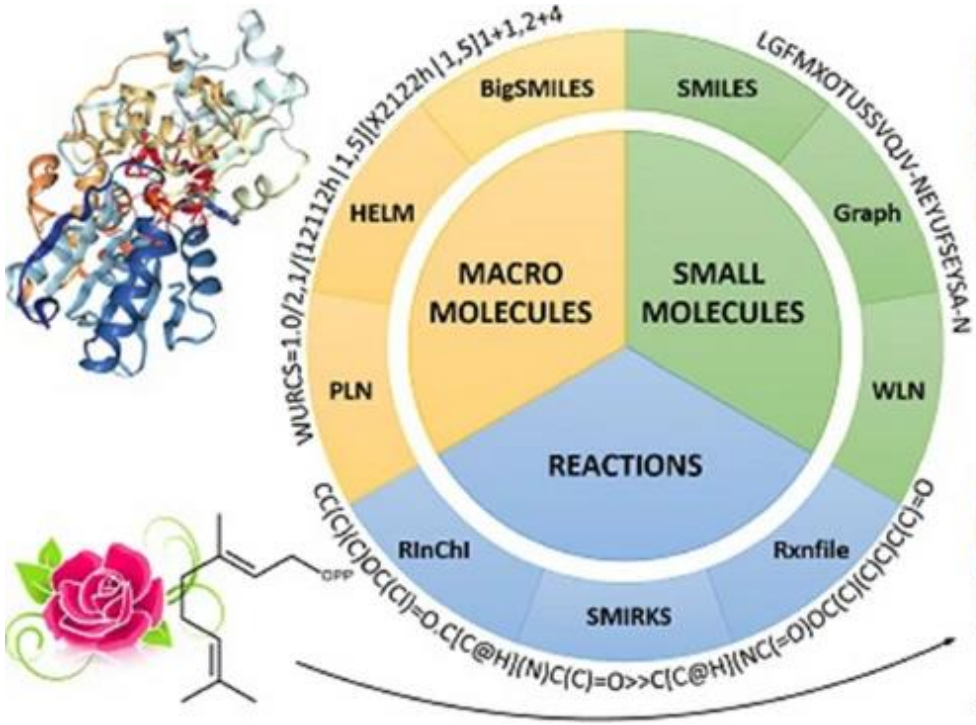
The Field of Molecular Representation

SMALL MOLECULES | REACTIONS | MACROMOLECULES | GRAPHICAL REPRESENTATIONS

- Molecular representations have been of interest since the 19th century and many challenges arose to create a functional representation
- Challenges are related to the diversity of chemical structures and their complexity

- Empirical formula of Alanine: $C_3H_7NO_2$
- This formula can also match sarcosine and lactamide.

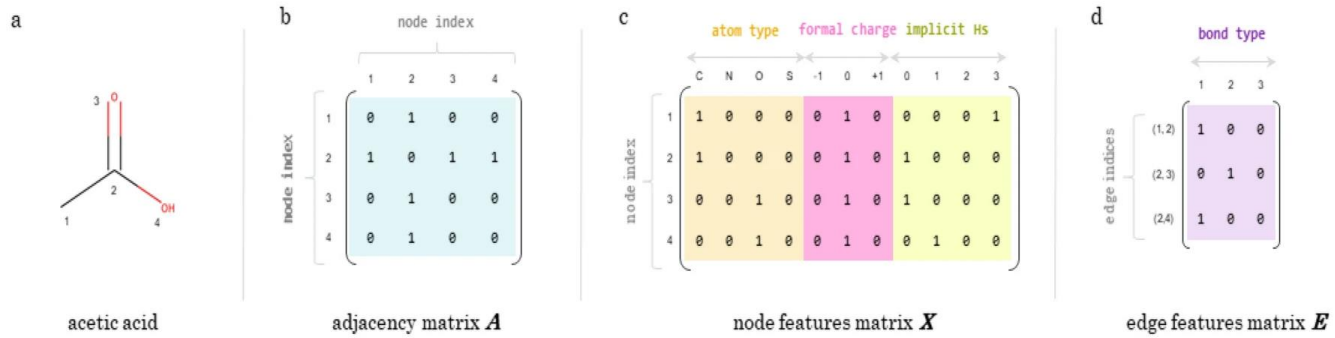
- A chemical representation is any encoding of a chemical compound
- Linear notations are referred to as notations



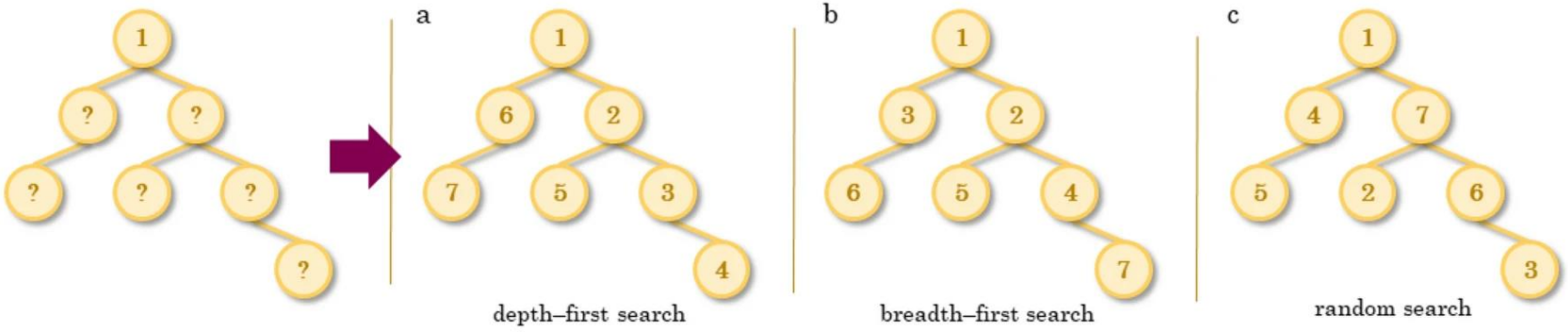
Graph Representations for Small Molecules

SMALL MOLECULES | REACTIONS | MACROMOLECULES | GRAPHICAL REPRESENTATIONS

- In a molecular graph representation, the atoms and bonds of a molecule are represented as a set of nodes and edges.
- How the atoms are connected can be represented in an adjacency matrix.
- The node order used in a matrix representation is determined by a graph traversal algorithm.



**Notion: One-hot encoding*



- Matrix representations of graphs are node-order dependent.
- If consistency is required, depth-first search and breadth-first search are applicable.
- If necessity of noisier data: random search (e.g. deep-learning)

Graph Representations for Small Molecules

SMALL MOLECULES | REACTIONS | MACROMOLECULES | GRAPHICAL REPRESENTATIONS

- **Advantages of Graph Representation**

- 3D information can be encoded into a graph
- Graphs are interpretable
  - Substrings of a SMILES string do not necessarily correspond to a valid graph

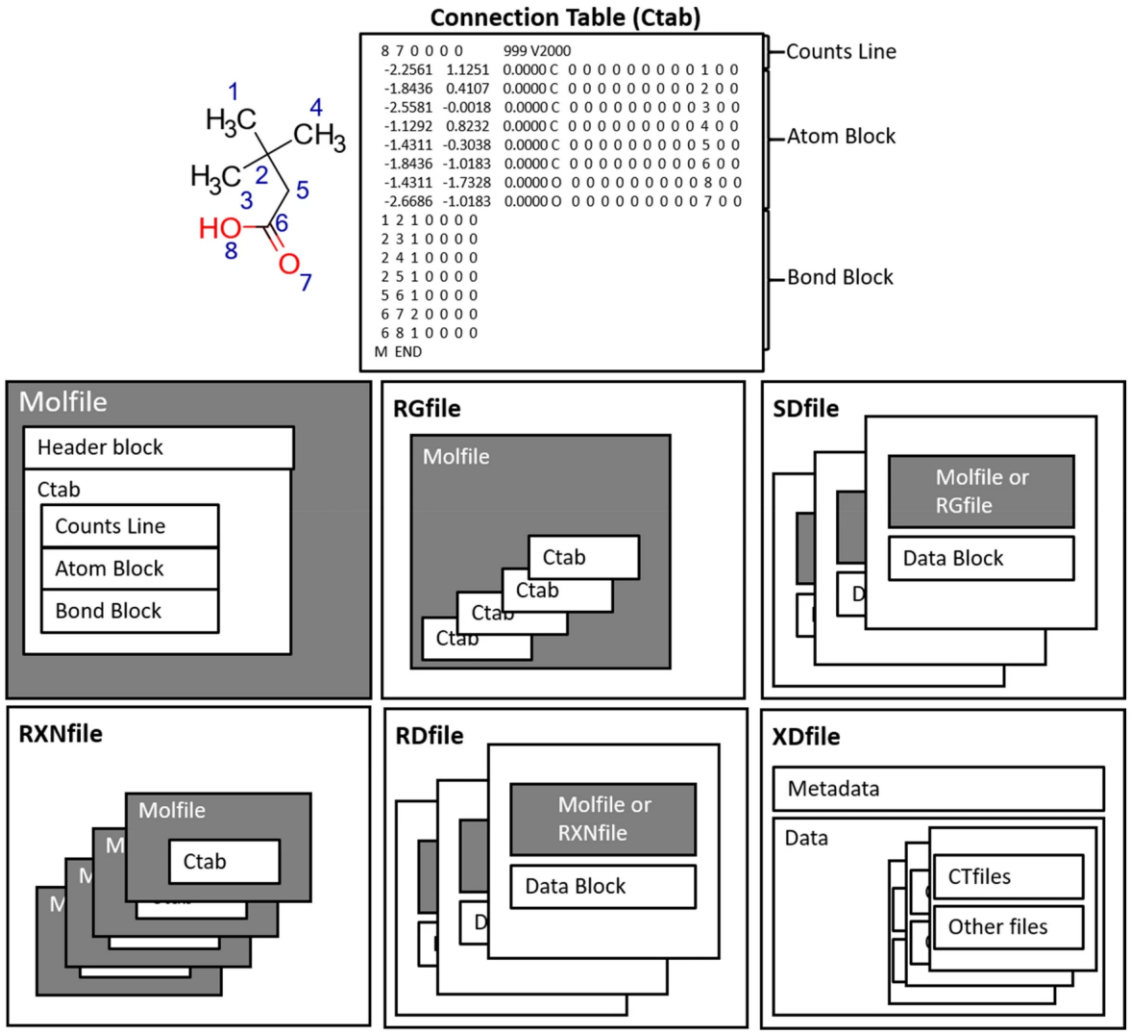
- **Limitations of Graph Representation**

- Many types of molecules cannot be described by the graph model:
  - Structure containing delocalized bonds (e.g. coordination compounds)
  - Molecule containing any of the following: polycentric bonds, ionic bonds, metal-metal bonds (e.g. organometallic)
- Graphs might not be meaningful for molecules where the arrangement of atoms is constantly changing in 3D space (e.g. tautomers)
- Molecular graphs are not compact (both memory-wise and literally)

Graph Representations: Related Formats

SMALL MOLECULES | REACTIONS | MACROMOLECULES | GRAPHICAL REPRESENTATIONS

- Connection table (Ctab): composed of 6 parts (counts line, atom block, bond block, atom list block, text block and properties block
- Ctab are the core building block around which CT files (Chemical Table files) are built
- Molfile : ensemble of Ctab
- RGfile: handling of queries
- SD File: ensemble of Molfile or Rgfile
- RXNfile: description of single reactions
- RDfile: storage of reactions as well as associated data
- XDfile: XML-based format for the transfer of structures or reactions and their associated metadata



Linear Notations of Small Molecules

SMALL MOLECULES | REACTIONS | MACROMOLECULES | GRAPHICAL REPRESENTATIONS

- Matrix representations:
  - require a large amount of disk space
  - Are not well adapted to basic cheminformatics analysis
- Molecules are often represented as strings of characters encoding the Ctab.

D-Alanine (implicit H)	
Molfile	612 bytes
SMILES	15 bytes
InChI	59 bytes

- Linear notations are compact, easy to manipulate.

- **Storyline of Linear Notations**

Alchemy: Compounds and elements named based on properties (e.g. aqua fortis = nitric acid, sweet oil of vitriol = diethyl ether)

↓

20th century : Systematic nomenclature of organic chemistry described in the IUPAC Color Books and used for literature, patents, governments legislation.

↓

1949-1961: International standard for electronic chemical notations

The IUPAC Quest for a Universal Notation

SMALL MOLECULES | REACTIONS | MACROMOLECULES | GRAPHICAL REPRESENTATIONS

- The IUPAC requested an international standard for electronic chemical notations in 1949
- **11 desirable properties ("desiderata"):**
  1. Simplicity of use
  2. Ease of printing and typewriting
  3. Conciseness
  4. Recognizability
  5. Ability to generate a unique chemical nomenclature
  6. Compatibility with accepted practices of inorganic chemical nomenclature
  7. Uniqueness
  8. Generation of an unambiguous and useful enumeration pattern
  9. Ease of manipulation by machine methods
  10. Exhibition of associations
  11. Ability to deal with partial indeterminates

The IUPAC Quest for a Universal Notation

SMALL MOLECULES | REACTIONS | MACROMOLECULES | GRAPHICAL REPRESENTATIONS

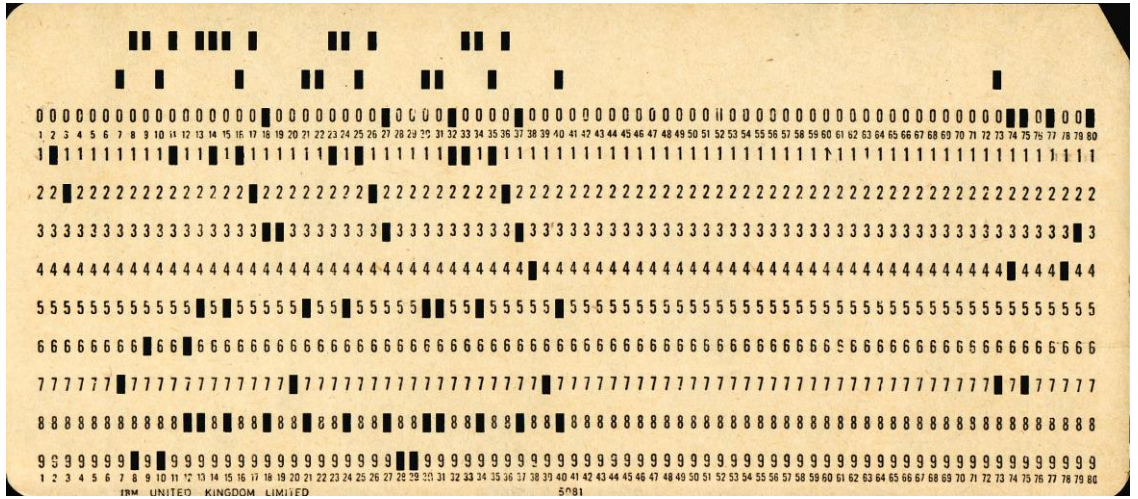
- The IUPAC requested an international standard for electronic chemical notations in 1949
- **11 desirable properties ("desiderata"):**
  1. Simplicity of use
  2. Ease of printing and typewriting
  3. Conciseness
  4. Recognizability
  5. Ability to generate a unique chemical nomenclature
  6. Compatibility with accepted practices of inorganic chemical nomenclature
  7. Uniqueness
  8. Generation of an unambiguous and useful enumeration pattern
  9. Ease of manipulation by machine methods
  10. Exhibition of associations
  11. Ability to deal with partial indeterminates

- **IUPAC Formalism (1964)**
- Notations can be classified as:
  - Unique (i.e. one notation for a given compound)
  - Non-unique (i.e. more than one notation for a given compound)
  - Ambiguous (i.e. notation will regenerate more than one compounds)
  - Non-ambiguous (i.e. notation will regenerate only the original compound)
- Chosen notation by IUPAC in 1961: Dyson cyphering
  - Drawbacks: contained many arbitrary rules / could not be handled on standard typewriters or ordinary punched-card machines

The IUPAC Quest for a Universal Notation

SMALL MOLECULES | REACTIONS | MACROMOLECULES | GRAPHICAL REPRESENTATIONS

Typewriter



Punched-card machine

The IUPAC Quest for a Universal Notation

SMALL MOLECULES | REACTIONS | MACROMOLECULES | GRAPHICAL REPRESENTATIONS

- The IUPAC requested an international standard for electronic chemical notations in 1949
- **11 desirable properties ("desiderata"):**
  1. Simplicity of use
  2. Ease of printing and typewriting
  3. Conciseness
  4. Recognizability
  5. Ability to generate a unique chemical nomenclature
  6. Compatibility with accepted practices of inorganic chemical nomenclature
  7. Uniqueness
  8. Generation of an unambiguous and useful enumeration pattern
  9. Ease of manipulation by machine methods
  10. Exhibition of associations
  11. Ability to deal with partial indeterminates

- **IUPAC Formalism (1964)**
- Notations can be classified as:
  - Unique (i.e. one notation for a given compound)
  - Non-unique (i.e. more than one notation for a given compound)
  - Ambiguous (i.e. notation will regenerate more than one compounds)
  - Non-ambiguous (i.e. notation will regenerate only the original compound)
- Chosen notation by IUPAC in 1961: Dyson cyphering
  - Drawbacks: could not be handled on standard typewriters or ordinary punched-card machines / contained many arbitrary rules
- Most used notation by the community: Wiswesser Line Notation (WLN)

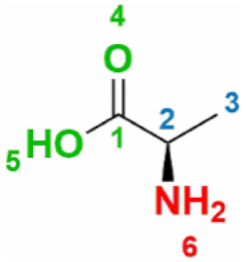
The IUPAC Quest for a Universal Notation

SMALL MOLECULES | REACTIONS | MACROMOLECULES | GRAPHICAL REPRESENTATIONS

- **Wiswesser Line Notation**

E	Bromine atom	F	Fluorine atom
G	Chlorine atom	H	Hydrogen atom
I	Iodine atom	Q	Hydroxyl group, -OH
R	Benzene ring	S	Sulfur atom
U	Double bond	UU	Triple bond
V	Carbonyl, -C(=O)-		
C	Unbranched carbon multiply bonded to non-carbon atom		
K	Nitrogen atom bonded to more than three other atoms		
L	First symbol of a carbocyclic ring notation		
M	Imino or imido -NH- group		
N	Nitrogen atom, hydrogen free, bonded to fewer than 4 atoms		
O	Oxygen atom, hydrogen-free		
T	First symbol of a heterocyclic ring notation		
W	Non-linear dioxo group, as in $-NO_2$ or $-SO_2-$		
X	Carbon attached to four atoms other than hydrogen		
Y	Carbon attached to three atoms other then hydrogen		
Z	Amino and amido NH_2 group		
<digit>	Digits "1" to "9" denote unbranched alkyl chains		
&	Sidechain terminator or, after a space, a component separator		

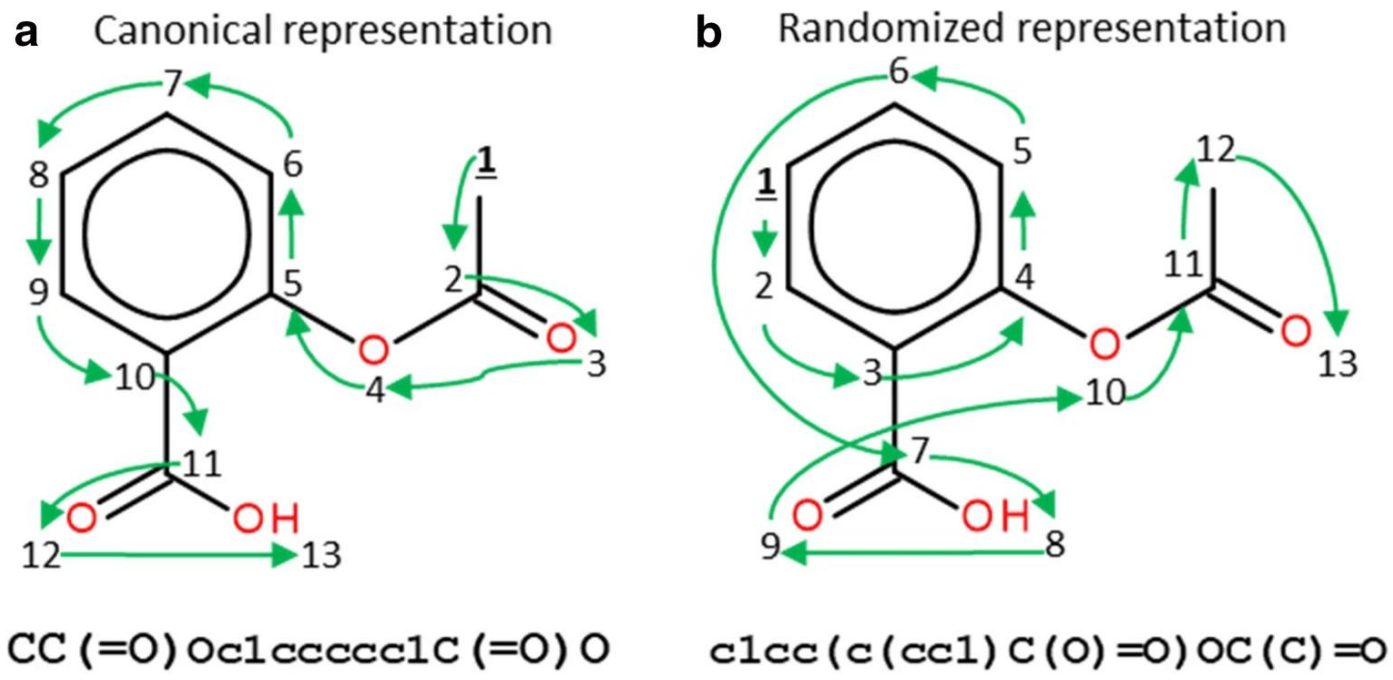
WLN for Alanine: QVYZ



Simplified Molecular Input Line Entry System (SMILES)

SMALL MOLECULES | REACTIONS | MACROMOLECULES | GRAPHICAL REPRESENTATIONS

- SMILES were developed by Weininger et al. in 1988
- SMILES are:
  - Non-unique and unambiguous
  - Obtained by assigning a number to each atom in the molecule, then traversing the molecular graph using that order
  - Rdkit version uses depth-first search
- SMILES can encode for stereochemistry
- SMILES can be generated canonically
- Canonical implementations vary between companies and research teams
- Randomized SMILES can be generated by changing the atom numbering in the structure:
  - Such SMILES can be employed for data augmentation in QSAR modelling or molecular generative modelling



Simplified Molecular Input Line Entry System (SMILES)

SMALL MOLECULES → REACTIONS → MACROMOLECULES → GRAPHICAL REPRESENTATIONS

Notations derived from SMILES

OpenSMILES

- Developed in 2007
- Provide a SMILES standard form

ChemAxon Extended SMILES (CXSMILES)

- Special features can be stored after the SMILES string

SMILES Arbitrary Target Specification (SMARTS)

- Extension of SMILES for substructure searching
- Can specify isotopes or bonds types

Reaction SMILES and SMIRKS

- In Reaction SMILES, molecules in reactants, agents and products can be represented by a SMILES string
- SMIRKS patterns define generic reaction transformations.

BigSMILES

- Encodes various types of polymers

Molecular Descriptors

SMALL MOLECULES | REACTIONS | MACROMOLECULES | GRAPHICAL REPRESENTATIONS

- Molecular descriptors such as structural keys and hashed fingerprints are notations encoding physicochemical, structural, topological, and/or electronical properties of a compound
- Descriptors are unique and ambiguous
- Structural keys are bit strings, encoding for the absence (0) and presence (1) of a specific chemical group (e.g. MACCS Keys)
- Chemical fingerprints are vectors containing indexed elements encoding for physicochemical or structural properties (e.g. Extended Connectivity Fingerprints ECFP)

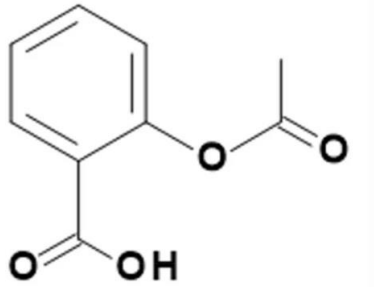
```
smartsPatts = {
  1: ('?', 0),  # ISOTOPE
  #2:('[#104,#105,#106,#107,#106,#109,#110,#111,#112]',0),  # atomic num >103 Not complete
  2: ('[#104]', 0),  # limit the above def'n since the RDKit only accepts up to #104
  3: ('[#32,#33,#34,#50,#51,#52,#82,#83,#84]', 0),  # Group IVa,Va,VIa Rows 4-6
  4: ('[Ac,Th,Pa,U,Np,Pu,Am,Cm,Bk,Cf,Es,Fm,Md,No,Lr]', 0),  # actinide
  5: ('[Sc,Ti,Y,Zr,Hf]', 0),  # Group IIIB,IVB (Sc...)
  6: ('[La,Ce,Pr,Nd,Pm,Sm,Eu,Gd,Tb,Dy,Ho,Er,Tm,Yb,Lu]', 0),  # Lanthanide
  7: ('[V,Cr,Mn,Nb,Mo,Tc,Ta,W,Re]', 0),  # Group VB,VIB,VIIB
  8: ('[!#6;!#1]1~*~*~*~1', 0),  # QAAA@1
  9: ('[Fe,Co,Ni,Ru,Rh,Pd,Os,Ir,Pt]', 0),  # Group VIII (Fe...)
  10: ('[Be,Mg,Ca,Sr,Ba,Ra]', 0),  # Group IIa (Alkaline earth)
  11: ('*1~*~*~*~1', 0),  # 4M Ring
  12: ('[Cu,Zn,Ag,Cd,Au,Hg]', 0),  # Group IB,IIB (Cu..)
  13: ('[#8]~[#7](~[#6])~[#6]', 0),  # ON(C)C
  14: ('[#16]-[#16]', 0),  # S-S
```

Some keys from MACCS Keys (rdkit/MACCSkeys.py at master · rdkit/rdkit · GitHub)

International Chemistry Identifier (InChI)

SMALL MOLECULES | REACTIONS | MACROMOLECULES | GRAPHICAL REPRESENTATIONS

- InChI is an open-source canonical notation introduced in 2006 by NIST
- InChI are composed of multiple layers such as the *Main, Charge, Stereochemical* and *Isotopic layers* (non exhaustive lit)
- InChI might not be decodable back to the molecular graph of origin
- A hashed version of InChI, InChIKey is used for library searching
- There exists a universal SMILES system which relies on the InChI representation



InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)

Chemical formula | Atom connections | Hydrogen atoms

Representation for Chemical Reactions

SMALL MOLECULES | REACTIONS | MACROMOLECULES | GRAPHICAL REPRESENTATIONS

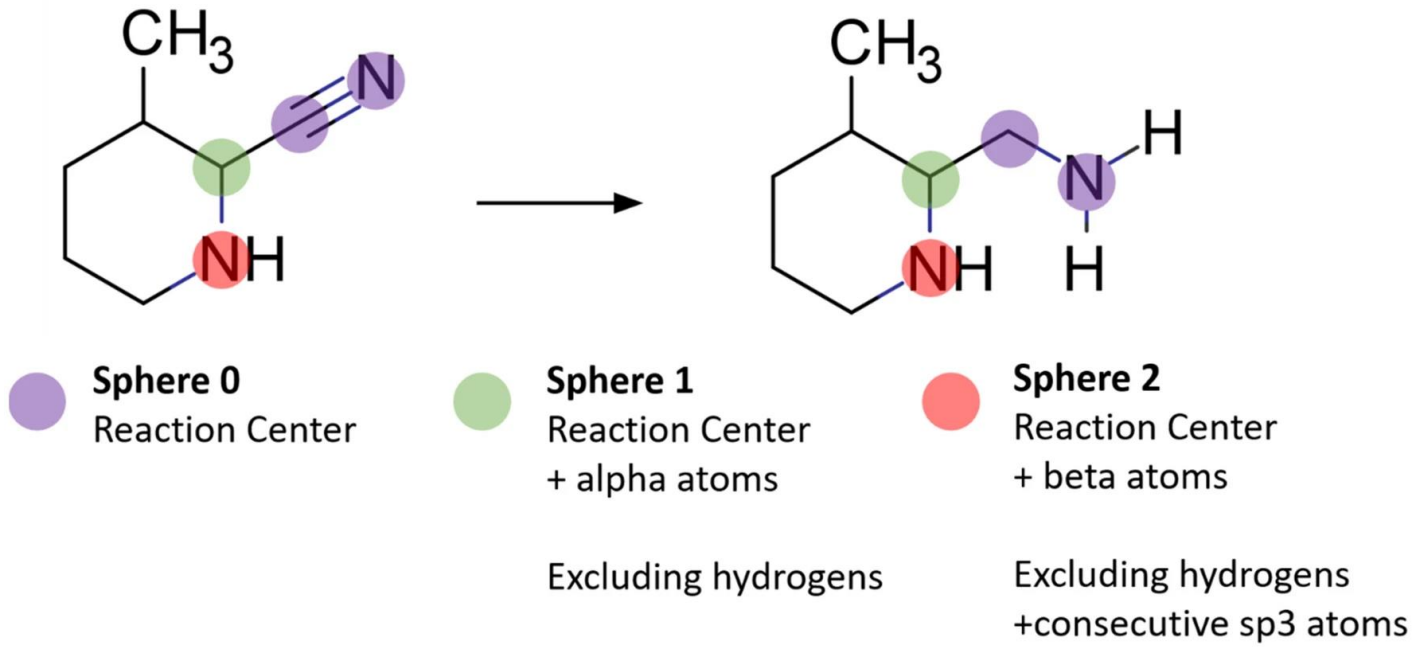
- Approximately 141 million reactions have been recorded since 1840
- Chemical reactions represent the interconversion of one set of molecules into another related set, under a set of specific conditions
- Graphical representations of reactions are not easily machine-readable
- Several notations and representations are available to solve this problem

Notations (non-exhaustive)	Details
Reaction SMILES	Encode reactants, agent and product Storage of reactions conditions or reaction center not supported
SMIRKS	Define generic reactions transformations Can describe reaction centre, enumerate libraries
RInChI	Direction of reaction can be described Enables identification of duplicate reactions ProcAuxInfo extension stores metadata (yield, temperature, …)

Representation for Chemical Reactions

SMALL MOLECULES | REACTIONS | MACROMOLECULES | GRAPHICAL REPRESENTATIONS

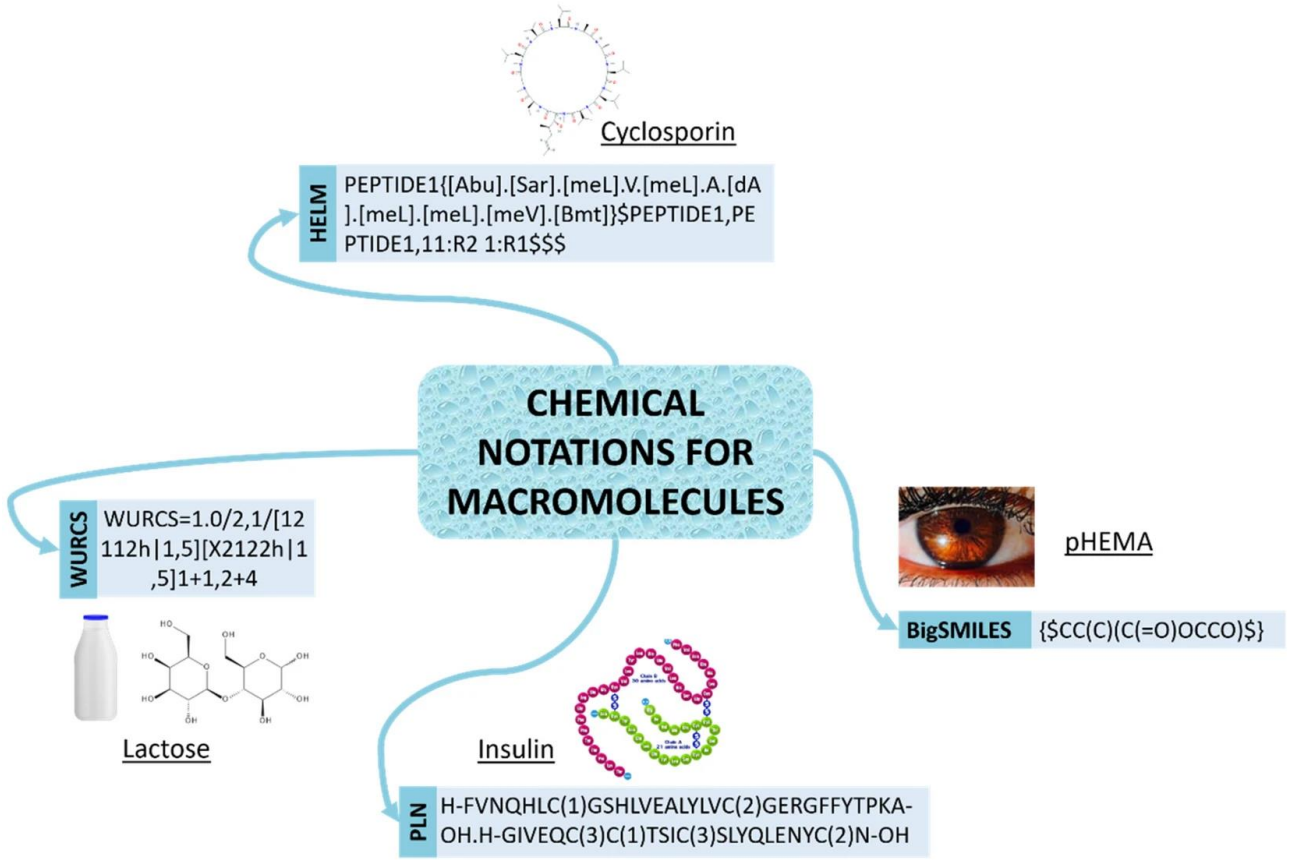
- A method to represent reactions was developed by Saller et. al. (**InfoChem CLASSIFY**)
- The steps are the following:
  1. Identify and extract the reaction centre (i.e. set of atoms that have changed their number of implicit hydrogens, valency, number of electrons, etc …)
  2. Atom hash codes are calculated for all atoms belonging to the reaction centre. A unique representation of the reaction centre is obtained



Representation for Macromolecules

SMALL MOLECULES | REACTIONS | MACROMOLECULES | GRAPHICAL REPRESENTATIONS

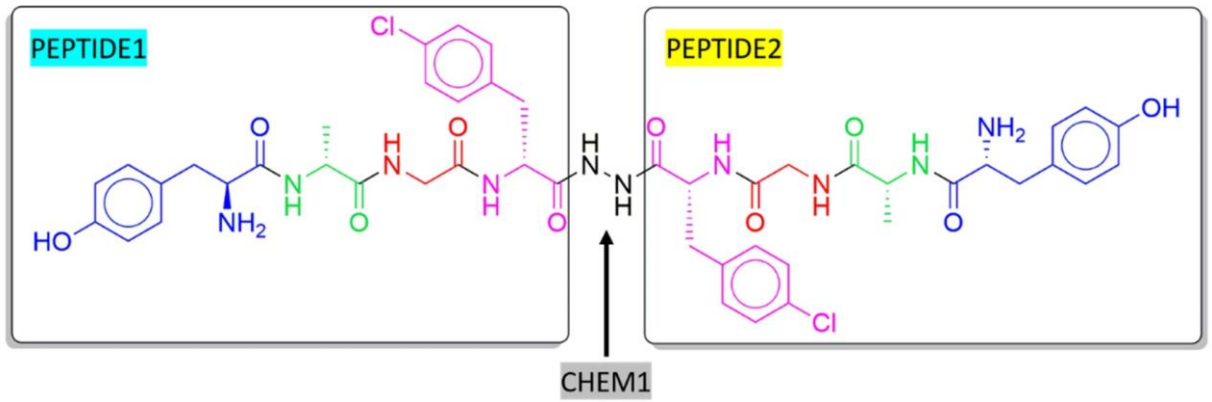
- Macromolecules are peptides, proteins, glycans, polymeric drugs.
- Peptides are chains of 2 to 40 amino-acids, linked by peptide bonds. Therapeutic peptides are valuable in terms of selectivity and affinity.
- Glycans are ensemble of monosaccharides linked glycosidically.
- Polymers are commonly used as drug deliverers. Nonetheless, some polymers have been used as active ingredients (e.g. Renagel®, Kayexalate®).
- It can be beneficial to treat these molecules on a different basis than the atomistic basis.



Amino-acid based structures

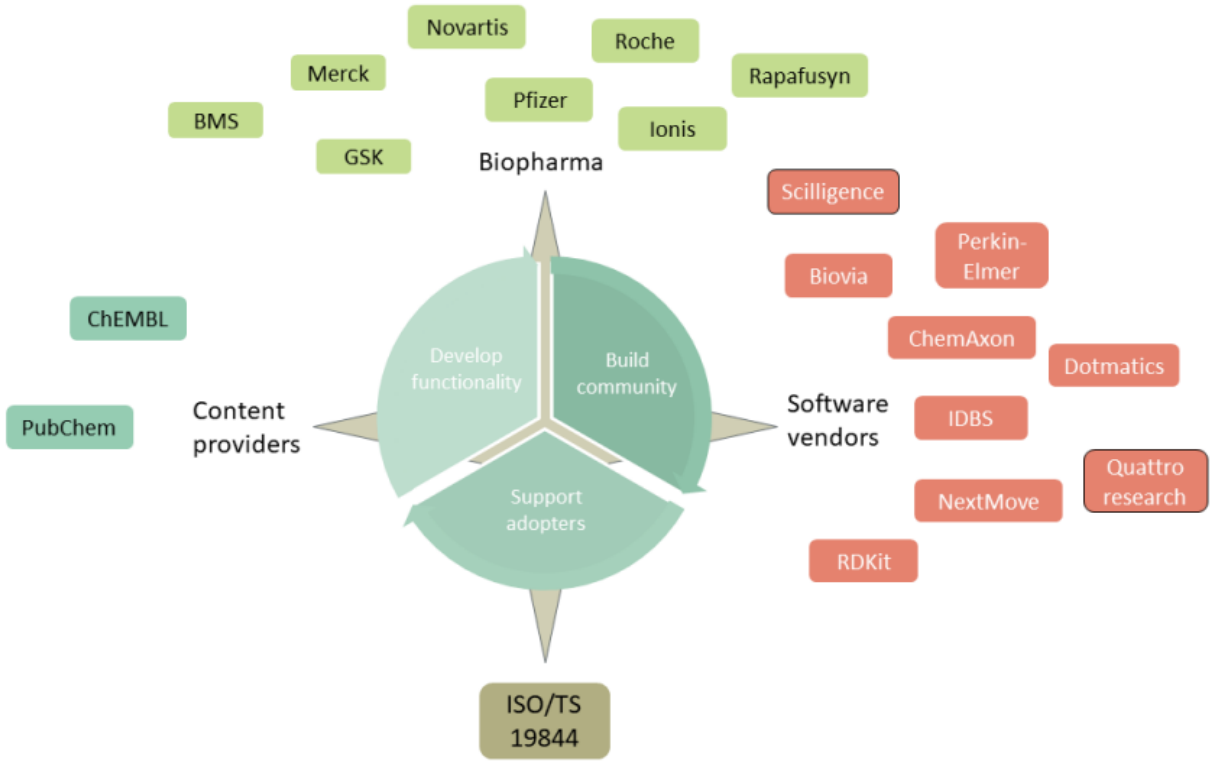
SMALL MOLECULES | REACTIONS | MACROMOLECULES | GRAPHICAL REPRESENTATIONS

- The Hierarchical Editing Language for Macromolecules (HELM) was developed by Pfizer under the auspices of the Pistoia Alliance.
- HELM can be used to describe peptides, antibodies, polymers, nucleotides, …
- HELM is widely adopted by the community



HELM Notation

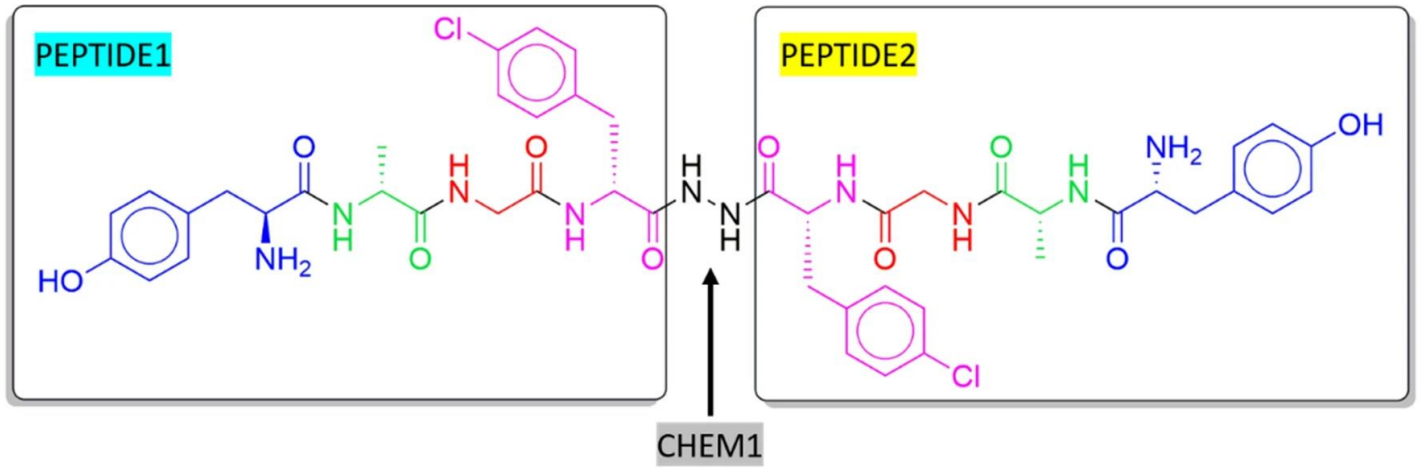
PEPTIDE1{Y.[dA].G.F}|PEPTIDE2{Y.[dA].G.F}|CHEM1{[R1]NN[R2]}$
PEPTIDE1,CHEM1,4:R2-1:R1|PEPTIDE2,CHEM1,4:R2:1R1$$$V2.0



HELM Project - Pistoia Alliance

Amino-acid based structures

SMALL MOLECULES | REACTIONS | MACROMOLECULES | GRAPHICAL REPRESENTATIONS



Polymer sequence defined first

Natural AA = single letter code

Non-natural AA = multi-letter code enclosed in square brackets

Branches = round brackets

PEPTIDE1{Y.[dA].G.F}

|PEPTIDE2{Y.[dA].G.F}|CHEM1{[R1]NN[R2]}

$PEPTIDE1,CHEM1,4:R2-1:R1

|PEPTIDE2,CHEM1,4:R2:1R1

$$$V2.0

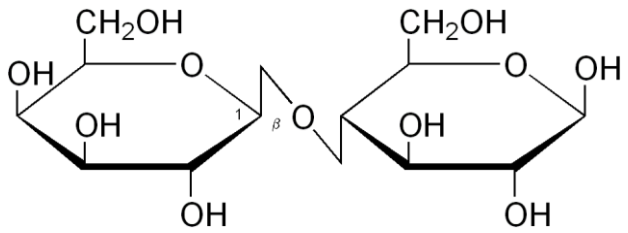
$ followed by connection between different types of polymer

Peptide 1 attached to CHEM1

Key Macromolecules

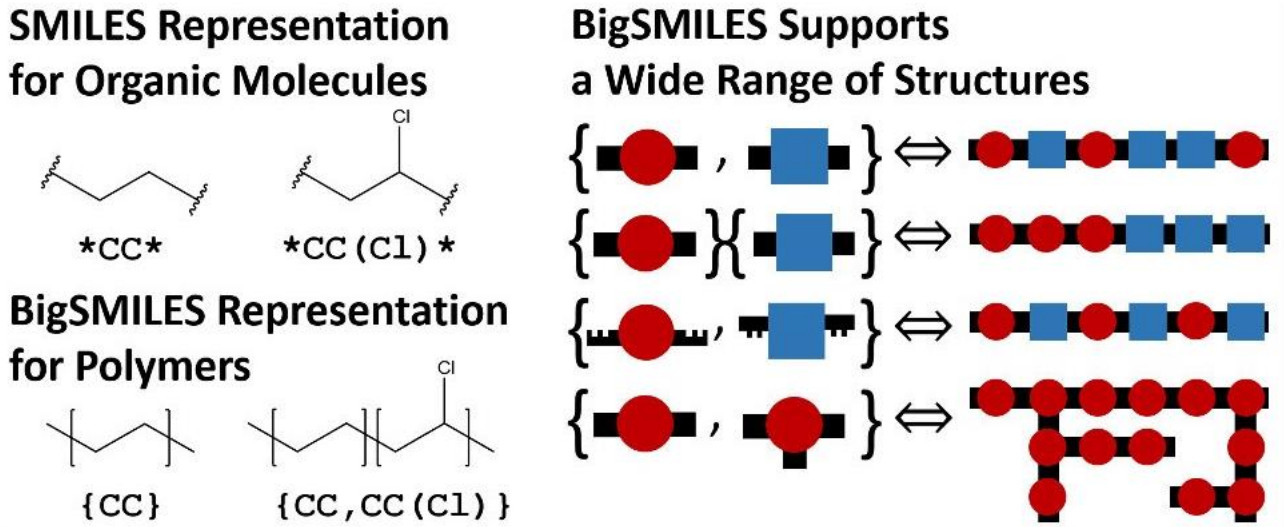
SMALL MOLECULES | REACTIONS | **MACROMOLECULES** | GRAPHICAL REPRESENTATIONS

- **Glycans**
- A main representation for glycans is the Web3 Unique Representation of Carbohydrate Structures (WURCS)
- It is used in the International Glycan Structure Repository (GlyTouCan)



WURCS=2.0/2,2,1/[a2122h-1b_1-5][a2112h-1b_1-5]/1-2/a4-b1

- **Polymeric drugs**
- The BigSMILES can encode for different types of polymers (e.g. homopolymers, branched polymers,…)

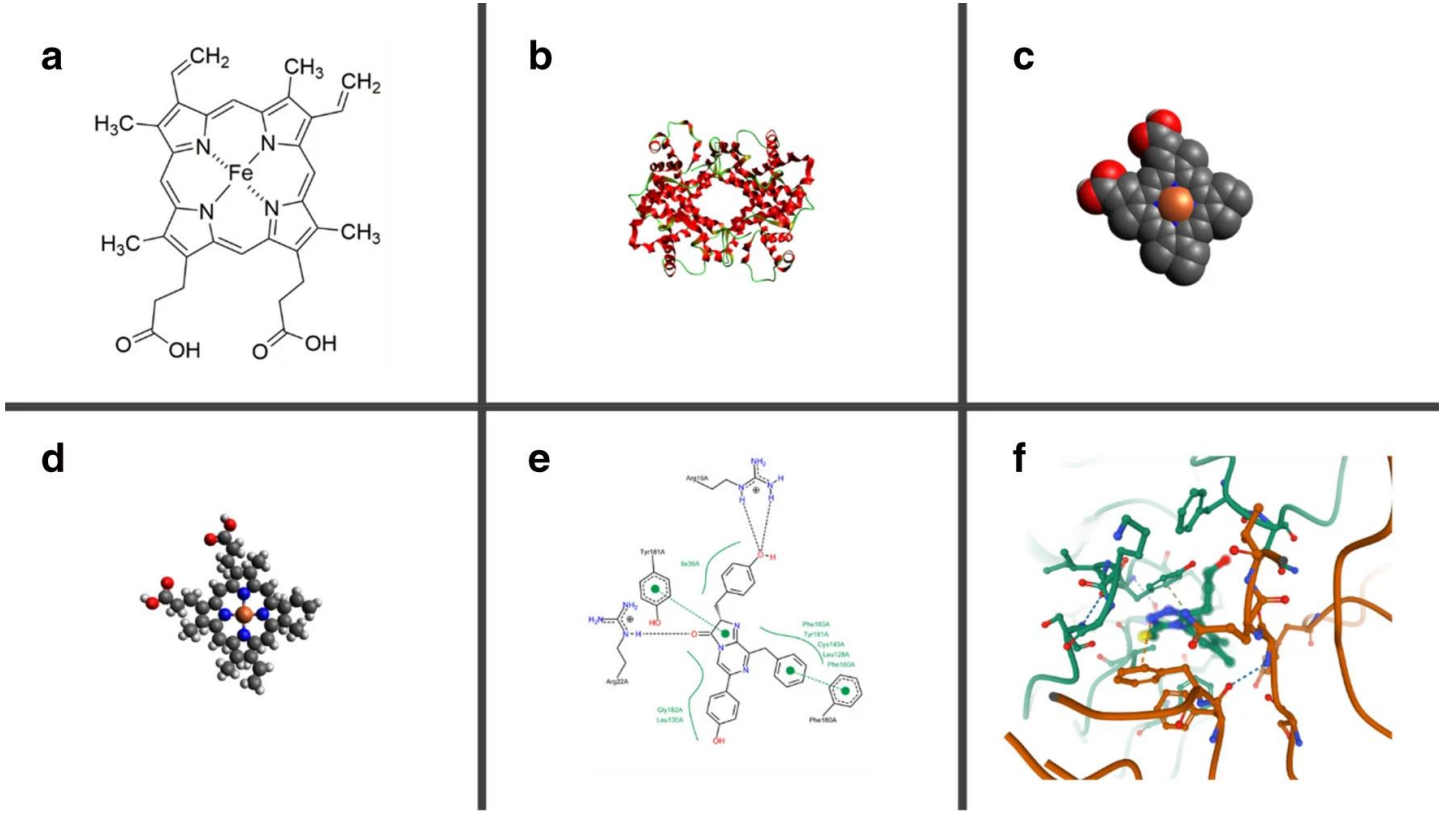


[BigSMILES – Olsen Research Group (mit.edu)]

Graphical Representations

SMALL MOLECULES | REACTIONS | MACROMOLECULES | GRAPHICAL REPRESENTATIONS

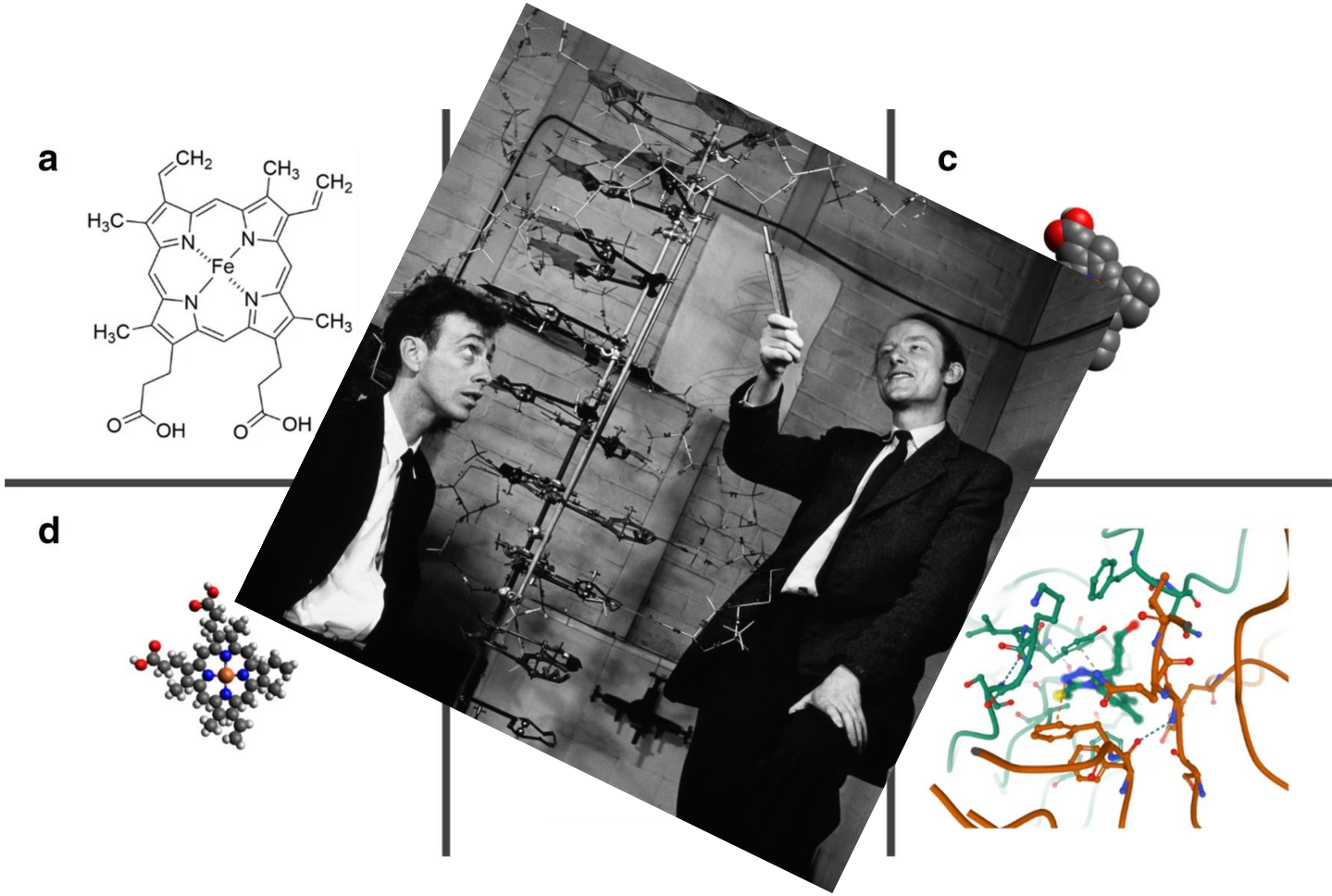
- Apart from notations and graph representations, molecules can be depicted – in 2D or 3D
- 2D depictions are used for structure-activity investigations
- 3D depictions are useful in docking and mechanistic studies
- *Side note*
  - Published data often have molecules given as 2D depictions, which makes it complicated to mine the data
  - Optical Chemical Recognition (OCR) systems are required to translate a depiction into representation or notation.



Graphical Representations

SMALL MOLECULES | REACTIONS | MACROMOLECULES | GRAPHICAL REPRESENTATIONS

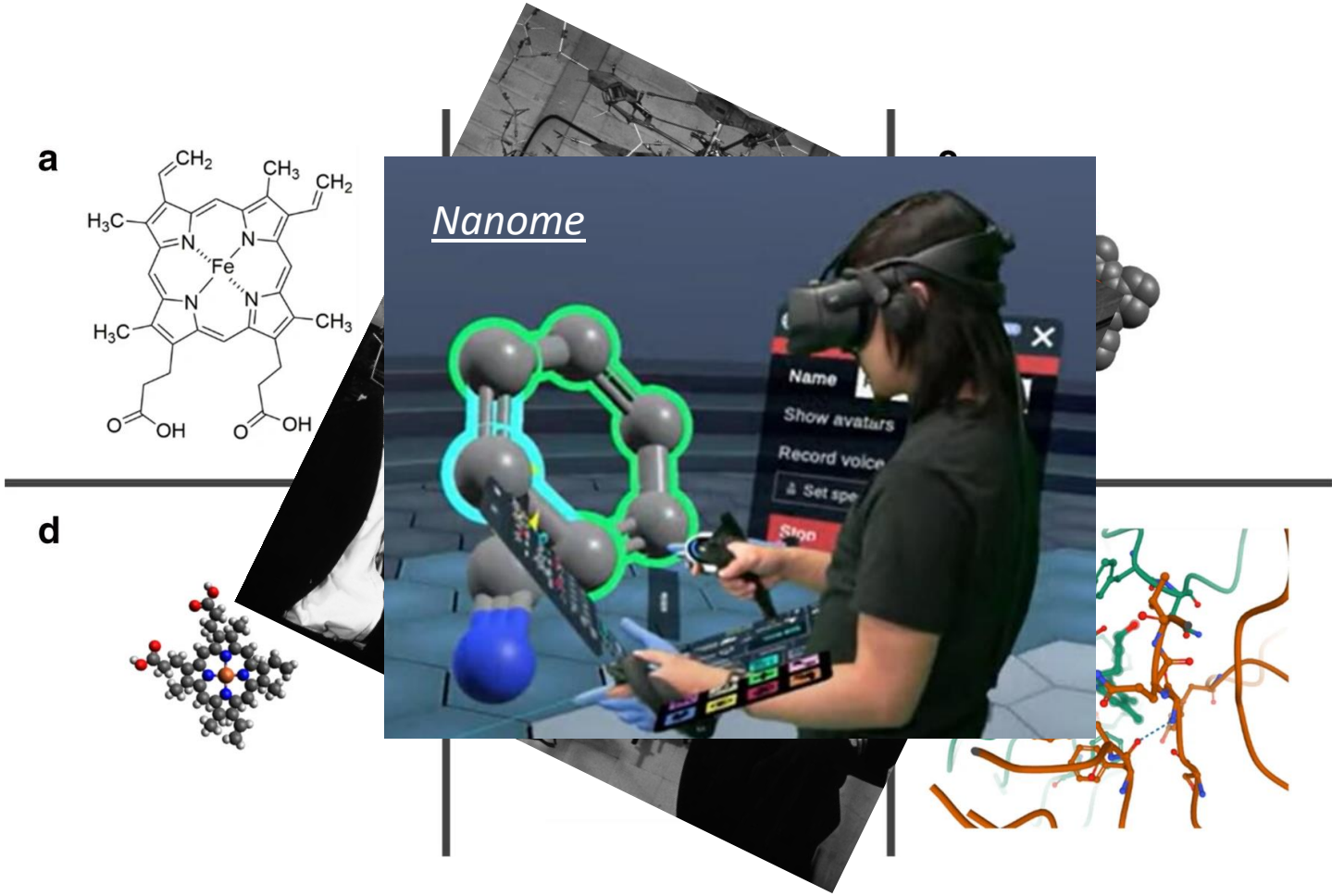
- Apart from notations and graph representations, molecules can be depicted – in 2D or 3D
- 2D depictions are used for structure-activity investigations
- 3D depictions are useful in docking and mechanistic studies
- *Side note*
  - Published data often have molecules given as 2D depictions, which makes it complicated to mine the data
  - Optical Chemical Recognition (OCR) systems are required to translate a depiction into representation or notation.



Graphical Representations

SMALL MOLECULES | REACTIONS | MACROMOLECULES | GRAPHICAL REPRESENTATIONS

- Apart from notations and graph representations, molecules can be depicted – in 2D or 3D
- 2D depictions are used for structure-activity investigations
- 3D depictions are useful in docking and mechanistic studies
- *Side note*
  - Published data often have molecules given as 2D depictions, which makes it complicated to mine the data
  - Optical Chemical Recognition (OCR) systems are required to translate a depiction into representation or notation.



Conclusions

SMALL MOLECULES → REACTIONS → MACROMOLECULES → GRAPHICAL REPRESENTATIONS

- Molecular representations must describe:
  - different types of structures (small molecules, peptides, polymers, …)
  - with different properties (stereochemistry, valence, …)
- Precise representations for specific structures are needed to optimize the process of AI-driven discovery
- Our aim was to provide a general overview of these representations, some which are well known in our field of AI-driven drug discovery, and some which are specialized

Notation system	Representation
Generic name	D-Alanine
IUPAC name	(2R)-2-aminopropanoic acid (English) acide (2R)-2-aminopropanoïque (French) (2R)-2-аминопропановая (Russian)
WLN[a]	QVYZ [46]
SMILES	C[C@H](C(=O)O)N
InChI	InChI = 1S/C3H7NO2/c1-2(4)3(5)6/h2H,4H2,1H3,(H,5,6) InChI = 1S/C3H7NO2/c1-2(4)3(5)6/h2H,4H2,1H3,(H,5,6)/t2-/m1/s1
InChI Key	QNAYBMKLOCPYGJ-UWTATZPHSA-N
HELM	PEPTIDE1{[dA]}
Three-letter symbol	D-Ala
Protein Line Notation (PLN)	H-{d}A-OH

Acknowledgments

- Rocío Mercado (rociomer@mit.edu)
- Amol Thakkar (tha@zurich.ibm.com)
- Ola Engkvist (AZ)

CO-AUTHORS

https://jcheminf.biomedcentral.com/articles/10.1186/s13321-020-00460-5

- Hongming Chen
- Noé Sturm
- Thierry Kogel
- Lionel Colliandre

- All the BigChem Fellows and PIs

The project leading to this presentation received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 676434, "Big Data in Chemistry" ("BIGCHEM", http://bigchem.eu). This presentation reflects only the authors' view, and neither the European Commission nor the Research Executive Agency are responsible for any use that may be made of the information it contains.

MARIE CURIE ACTIONS

Bibliography

- **Molecular representations in AI-driven drug discovery: a review and practical guide | Journal of Cheminformatics | Full Text (biomedcentral.com)**
- rdkit/MACCSkeys.py at master · rdkit/rdkit · GitHub
- Polymers as drugs—Advances in therapeutic applications of polymer binding agents - Connor - 2017 - Journal of Polymer Science Part A: Polymer Chemistry - Wiley Online Library
- HELM Project - Pistoia Alliance
- PepSeA: Peptide Sequence Alignment and Visualization Tools to Enable Lead Optimization | Journal of Chemical Information and Modeling (acs.org)
- BigSMILES – Olsen Research Group (mit.edu)
- Nanome