Mechanical & Industrial Engineering
UNIVERSITY OF TORONTO

A Molecular Structure Ontology for Medicinal Chemistry

Carmen Chui, Michael Grüninger

June 16, 2020

Research Objective

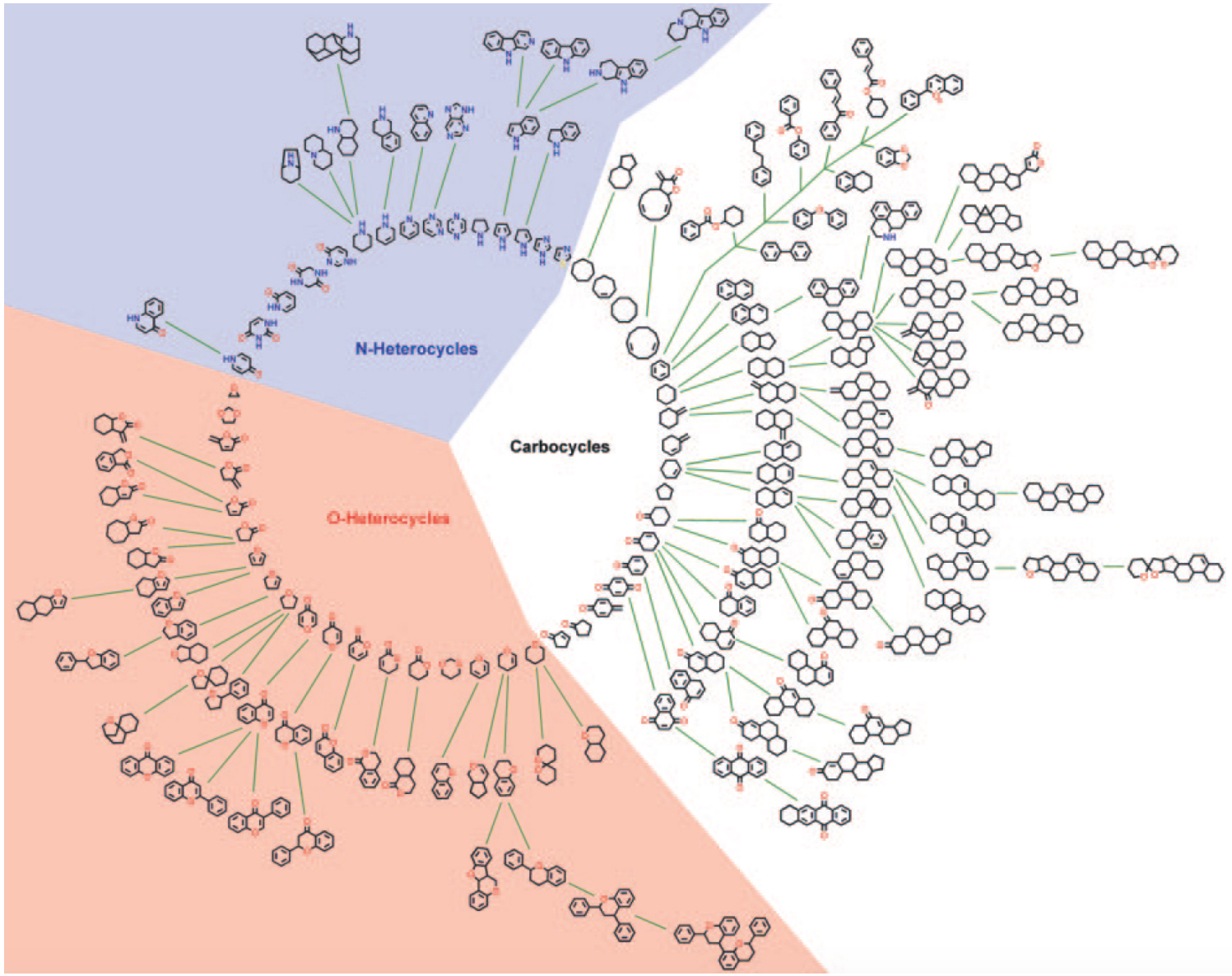
Objective

Existing work in cheminformatics discusses the notion of 'chemical space' to describe all possible organic molecules to be considered when searching for new drugs [RA12].
We want to provide **ontological foundations for chemical space**, where the central idea is that chemical space is characterized by the *shape* and *structure* of molecules.

Topological Structure

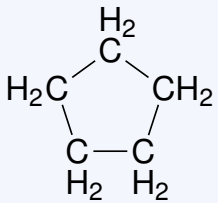
With this work, we focus on **topological structure** of molecules in a **static** context.
Chemical spaces and the scaffold tree approach do not talk about the full geometry (such as stereochemistry pertaining to isomers, bond angles, etc.) since only the 'core' is examined. Instead, we wanted the ontology to be geared toward supporting the scaffold tree approach.

Chemical Space: Scaffold Tree in [Koc+05]

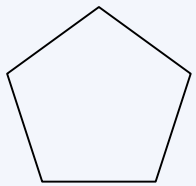


Motivations: What do we mean by shape?

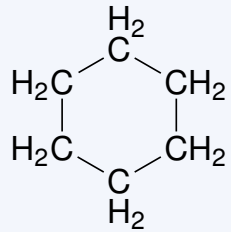
2D Shapes & Polygons



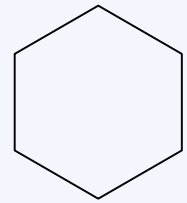
Cyclopentane



Pentagon

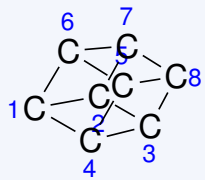


Cyclohexane

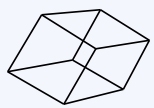


Hexagon

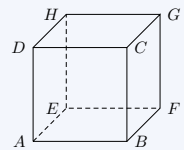
3D Shapes (Polyhedra) & Molecules



Cubane

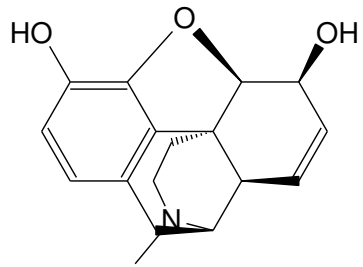


Skeletal Formula



Cube

Current Approaches to Represent Molecular Shape



Chemical name [Nat15]:

Morphine; Morphinum; Morphia; Morphium; Morphin

IUPAC name:

(4R,4aR,7S,7aR,12bS)-3-methyl-2,4,4a,7,7a,13-hexahydro-1H-4,12-methanobenzofuro[3,2-e]isoquinoline-7,9-diol

SMILES:

CN1CC[C@]23C4=C5C=CC(O)=C4O[C@H]2[C@@H](O)C=C[C@H]3[C@H]1C5

InChI identifier:

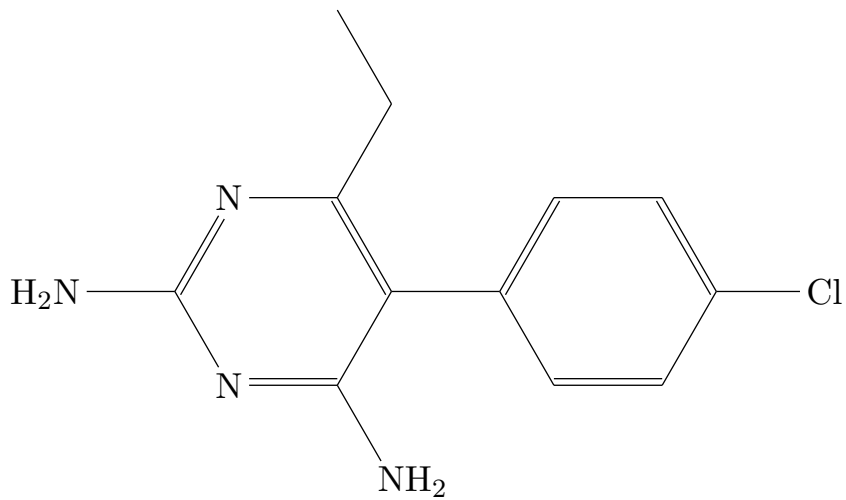
InChI=1S/C17H19NO3/c1-18-7-6-17-10-3-5-13(20)16(17)21-15-12(19)4-2-9(14(15)17)8-11(10)18/h2-5,10-11,13,16,19-20H,6-8H2,1H3/t10-,11+,13-,16-,17-/m0/s1

InChI key: BQJCRHHNABKAKU-KBQPJGBKSA-N

How would you describe and represent this molecule?

Pyrimethamine/Daraprim ($C_{12}H_{13}ClN_4$)

- Used to treat malaria and *Pneumocystis jirovecii* pneumonia (PCP)
- Price hike controversy in 2015: Turing Pharmaceuticals raised the price from 13.50 USD to 750 USD per tablet



How do people use these approaches?

- Cheminformatics uses *special-purpose* tailor-made algorithms
- Focus is more on classification with an algorithmic approach

The Benefit of Automated Reasoning

Another approach is to use **automated deduction**.

- Given the information about the structure of a molecule, we can use **software** to deduce consequences.
- This requires the **representation of knowledge** of molecular structure.

Ontologies

In **artificial intelligence**, an *ontology* is a formal specification of knowledge in some domain.

- Ontologies make the semantics of the domain terminology **explicit**.
- For example, chemical terminology includes:
  - Rings
  - Chains
  - Functional groups
  - Skeletons
- An ontology for molecular structure would need to **capture the semantics of these terms**.

What's the Problem?

Cheminformatics Approaches

- No *semantics* for symbols used to represent molecules
- No *reasoning* capabilities about shape

Ontological Approaches

- Not enough *semantics* for shape representation
- No *reasoning* capabilities about shape

Competency Questions

Similarity

- Which molecules have common substructures with a given molecule?
- Which antibiotics contain a β-lactam ring?
- What are molecules that contain two fused rings?
- Which molecules contain a given functional group?

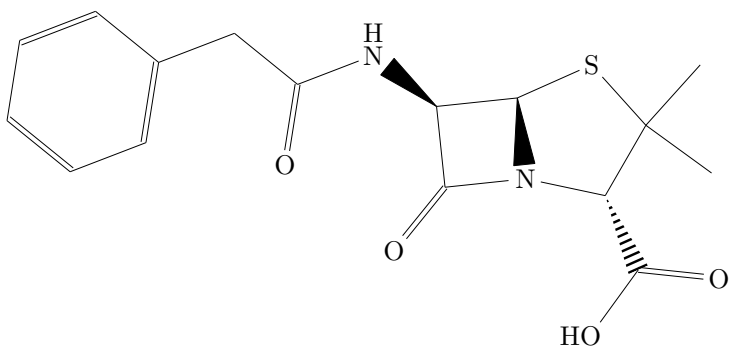


Figure: Penicillin ($C_9H_{11}N_2O_4S$)

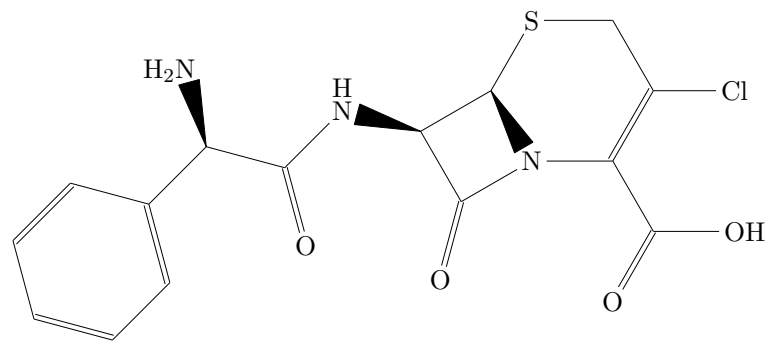


Figure: Cefaclor ($C_{15}H_{14}ClN_3O_4S$)

Competency Questions

Substitution

- What molecules are equivalent to molecule x after we substitute substructure y with substructure z?
  - e.g., dimethylmethylene group $C(CH_3)_2$ in Bisphenol A (BPA) can be replaced with a sulfone group SO_2 in Bisphenol S (BPS).
- What molecules have the same shape if you substitute one functional group with another?

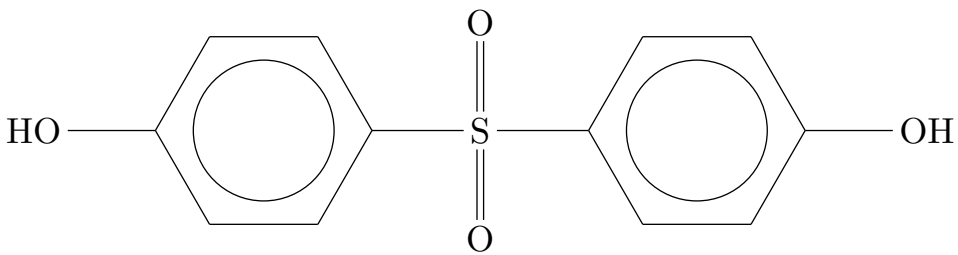


Figure: Bisphenol S (BPS)

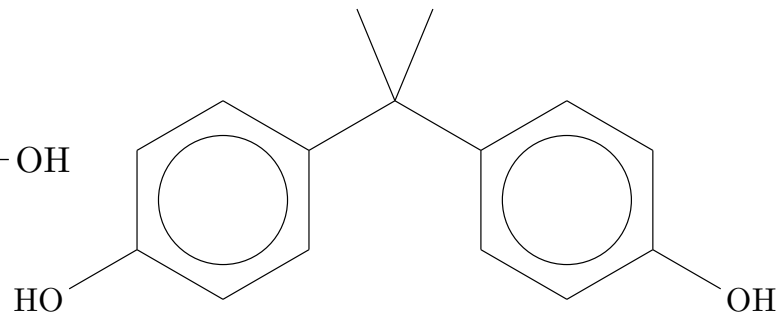


Figure: Bisphenol A (BPA)

Competency Questions

Synthesis

- What molecules contain the combination of elements/atoms x, y, z?
- What molecules contain functional groups x and y?

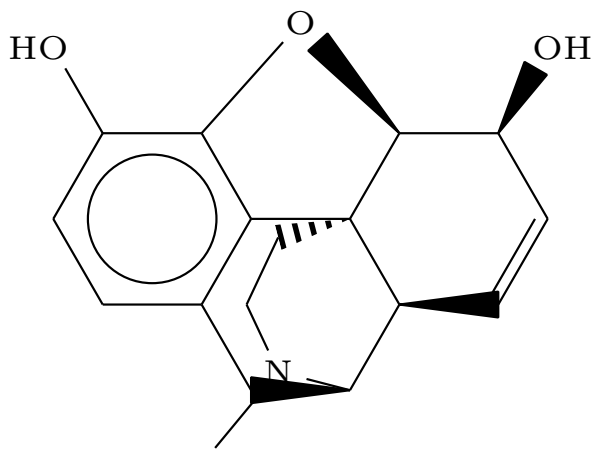


Figure: Morphine ($C_{17}H_{19}NO_3$)

Requirements for the Ontology & Its Models

Competency questions guided the overall design of the MOlecular Structure ontology (MoSt)

Requirements & Semantic Conditions for Representing Shape

- Molecules must be represented as graphs
- Components of molecules must be elements of the domain
- Attachments between functional groups (spiro, tether, fusion) must also be represented

Requirements for the Models of the Ontology

- 1-to-1 correspondence of models of MoSt with molecules
- Intended models of the ontology are molecules
- Unintended models of the ontology are not molecules

Ontological Commitments as Requirements

We can extract requirements that drive the design of the ontology to ensure we capture all of the information required to represent molecular structure:

Ontology Requirements

R-1 The ontology must represent the properties of elements, functional groups, connections between functional groups and components of molecules, along with a classification of molecules with respect to their structures.

R-2 The ontology must represent molecules as graphs, such that molecules can be decomposed into their primitive functional groups.

Ontological Commitments as Requirements (cont.)

Atoms, Bonds, Functional Groups

R-3 Atoms, bonds, and functional groups are elements of the domain (*things* in the ontology).

R-4 Primitive functional groups are *rings* and *chains*, which correspond to induced cyclic and path subgraphs, respectively.

R-5 The carbon backbone of organic molecules forms the essential structure of a given compound, which we call a '*skeleton*'. The skeleton consists of the various combinations of rings, chains, and atoms that are not in any functional groups.

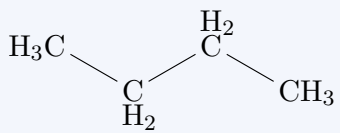


Figure: Butane

Figure: Butane (Carbon Backbone)

Ontological Commitments as Requirements (cont.)

Connections Between Functional Groups

Functional groups can be categorized according to how groups can share atoms or are bonded together; we categorize the various types of connections as requirements for the ontology:

R-6 The ontology must represent the *fusion* connection: there is an overlapping bond between the groups.

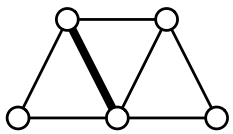


Figure: Fusion (Rings)

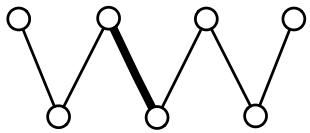


Figure: Fusion (Chain)

Ontological Commitments as Requirements (cont.)

Connections Between Functional Groups

R-7 The ontology must represent the *spiro* connection: rings or chains share an atom.

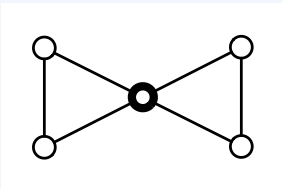


Figure: Spiro (Rings)

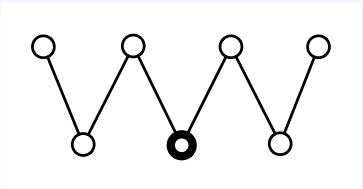


Figure: Spiro (Chain)

Ontological Commitments as Requirements (cont.)

Connections Between Functional Groups

R-8 The ontology must represent the *tether* connection: no atoms between the groups are shared, but any two groups are bonded together by a bond between each group.

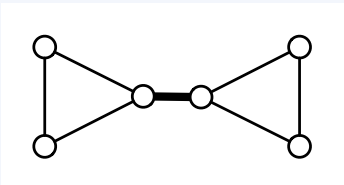


Figure: Tether (Rings)

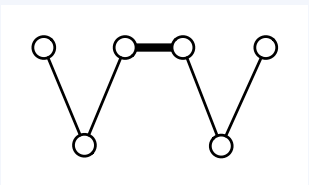


Figure: Tether (Chain)

Molecular Structure Ontology (MoSt)

We propose an ontology that combines both the cheminformatics and ontological approaches to describing molecules, and adheres to the requirements.

With this ontology, we can represent:

- Rings
- Chains
- Functional groups
- Skeletons
- Classes of bonds
- Classes of elements
- Connections between functional groups
- Connections between skeletons

What does this mean?

Knowing that the ontology is able to represent various molecular structure concepts, what does this mean?

- MoSt consists of statements that can be written in both **English** and in a **formal logic** (first-order logic; FOL).
- In this work, we have axiomatized these sentences using the Common Logic (CL) syntax (ISO/IEC 24707) and in the Prover9 syntax.
  - Common Logic: `https://www.iso.org/standard/39175.html`
  - Prover9: `https://www.cs.unm.edu/~mccune/prover9/manual/2009-11A/`
- For readability in print, we write out the axioms using FOL.

What does this mean? (cont.)

Example: Ring Definition

In English:

> *"A **ring** is a group that contains atoms that are not ends and not forks."*

In logic:

$$\forall x\, ring(x) \equiv group(x) \wedge (\forall y\, (atom(y) \wedge mol(y, x) \supset (\neg end(y, x) \wedge \neg fork(y))))$$

x is a ring *if and only if* x is a group and every atom in the group is not an end and not a fork.

What does this mean? (cont.)

Example: Groups & Skeletons

In English:

*"Every **group** is contained in a unique **skeleton**."*

In logic:

$$\forall x\ group(x) \supset \exists y\ skeleton(y) \wedge mol(x, y) \wedge (\neg \exists w\ (skeleton(w) \wedge mol(x, w) \wedge w \neq y))$$

If x is a group, there exists a skeleton y where x is in y, and there does not exist a skeleton w that contains x.

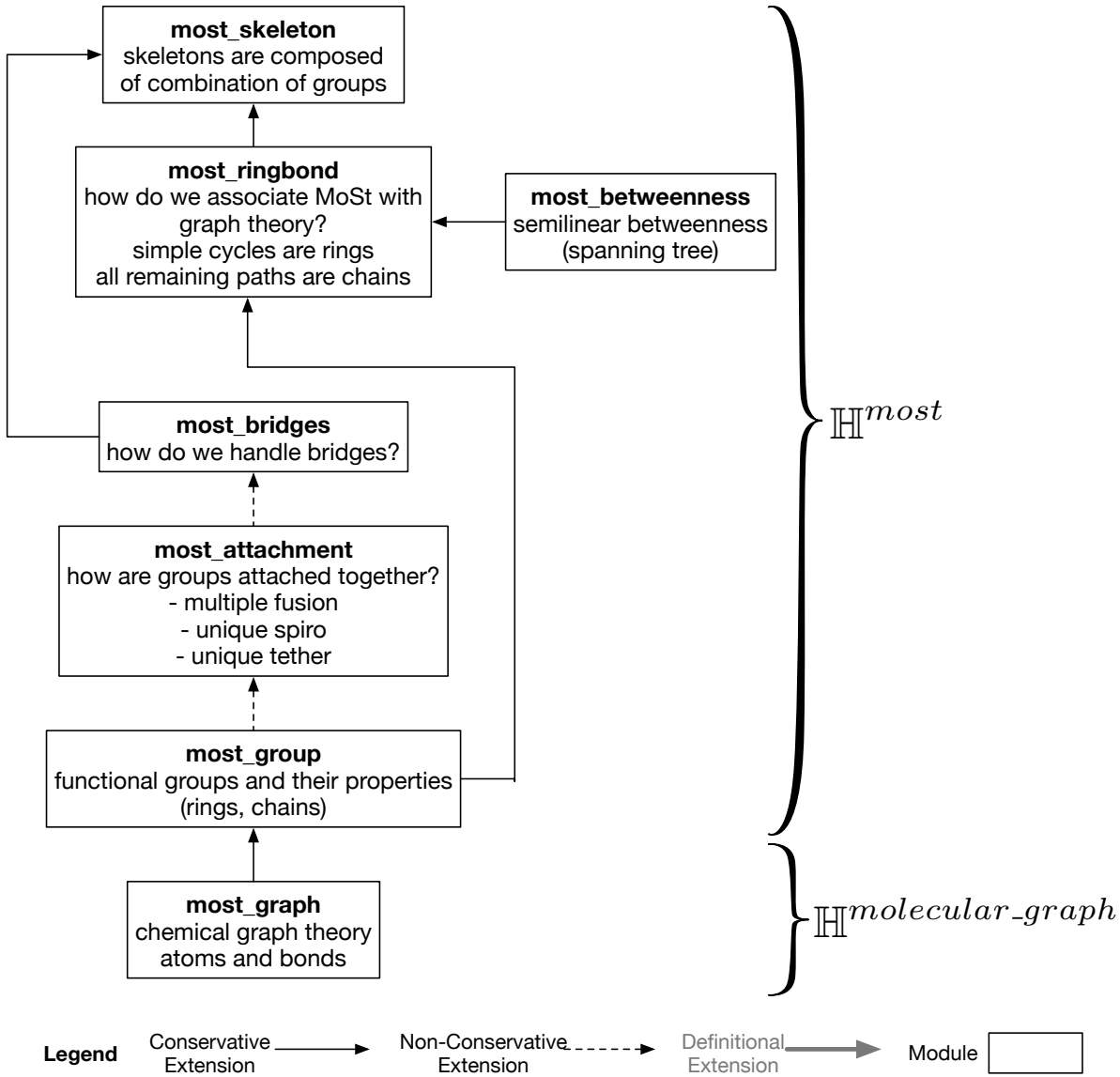
Modules of MoSt

The ontology satisfies each of the requirements and consists of various modules:

- **most_graph** satisfies R-1, R-2, R-3
- **most_root** satisfies R-1, R-2, R-3, R-4
- **most_attachment** satisfies R-6, R-7, R-8
- **most_bridges** satisfies R-5
- **most_skeleton** satisfies R-5

Modules of MoSt (cont.)

Graphically, the modules of the ontology are organized as such:



most_graph

$most\_graph$ contains axioms to describe atoms and bonds and how they are connected

- Identical to existing molecular graph theory work

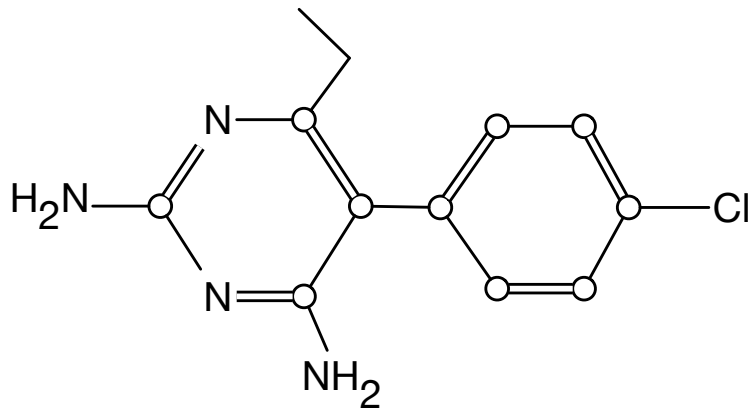


Figure: Daraprim ($C_{12}H_{13}CIN_4$)

most_root

$most\_root$ adds an additional layer of semantics:

- Functional groups are elements of the domain
- Properties of functional groups
- Rings are cycles in the graph
- Chains are paths in the graph

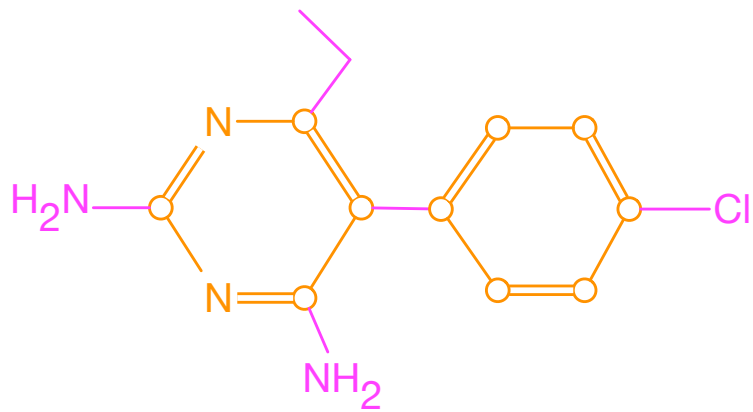


Figure: Daraprim ($C_{12}H_{13}CIN_4$)

most_attachment

Permissible Attachment for Rings

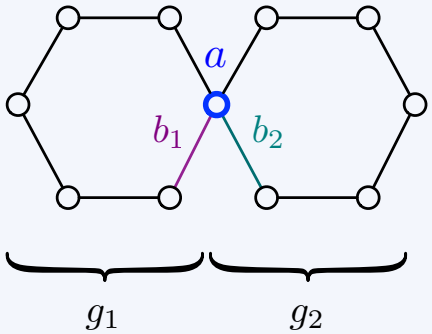


Figure: Unique spiro

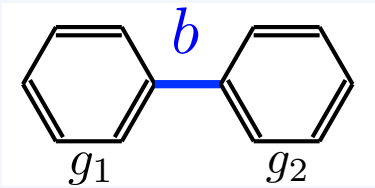


Figure: Unique Tether

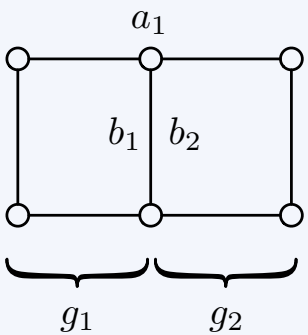


Figure: Fused rings at one bond

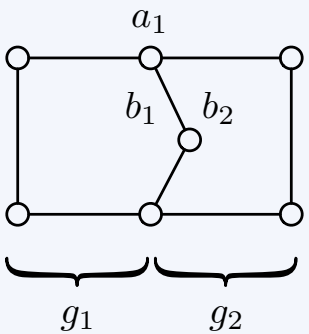
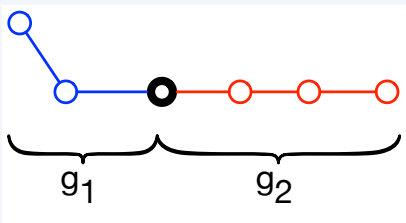


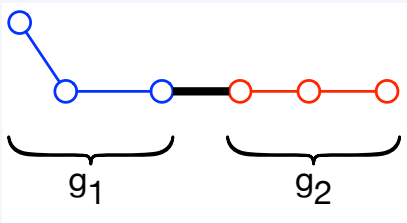
Figure: Multiple fusion for bridged molecules

most_attachment (cont.)

Permissible Attachment for Chains

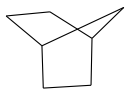
- Chains cannot be fused
- Chains can be connected via spiro or tether attachments





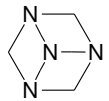
most_bridges

$most\_bridges$ allows us to describe bridged molecules as molecules that are fused at *multiple* bonds



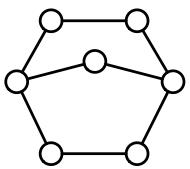
Norbornane (C_7H_{12})

Adamantane ($C_{10}H_{16}$)

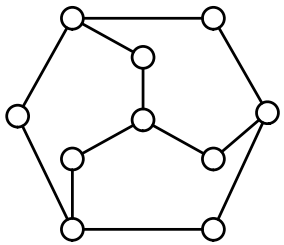


Hexamine ($C_6H_{12}N_4$)

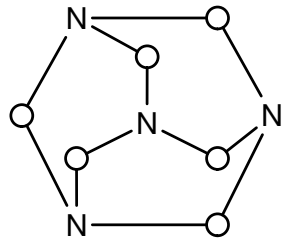
We can redraw them *topologically*:



Norbornane (C_7H_{12})



Adamantane ($C_{10}H_{16}$)



Hexamine ($C_6H_{12}N_4$)

most_skeleton

$most\_skeleton$ combines everything together

- Skeleton = composed of various functional groups connected together
- Skeletons are also elements of the domain in the ontology

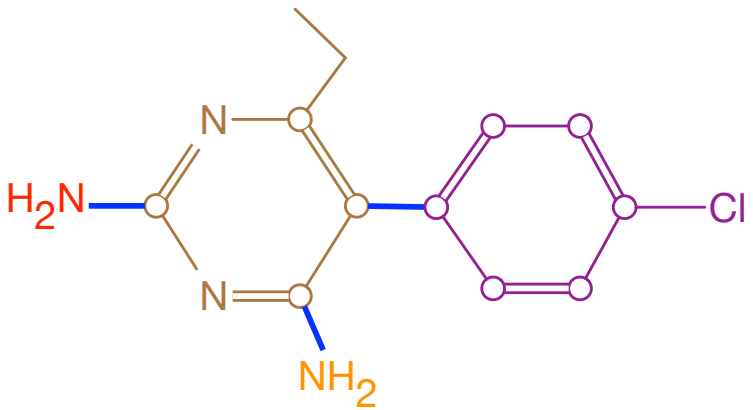
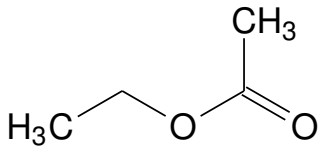


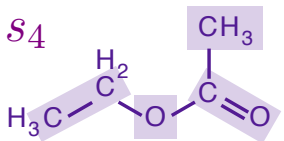
Figure: Daraprim ($C_{12}H_{13}CIN_4$)

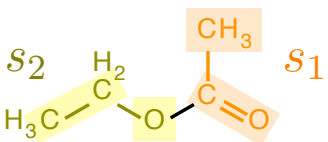
Breaking Down Skeletons

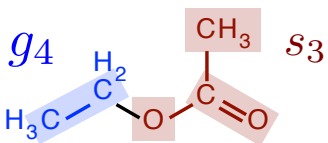
Consider Ethyl Acetate ($C_4H_8O_2$):

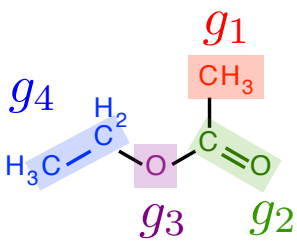


Corresponding Skeletal Diagrams

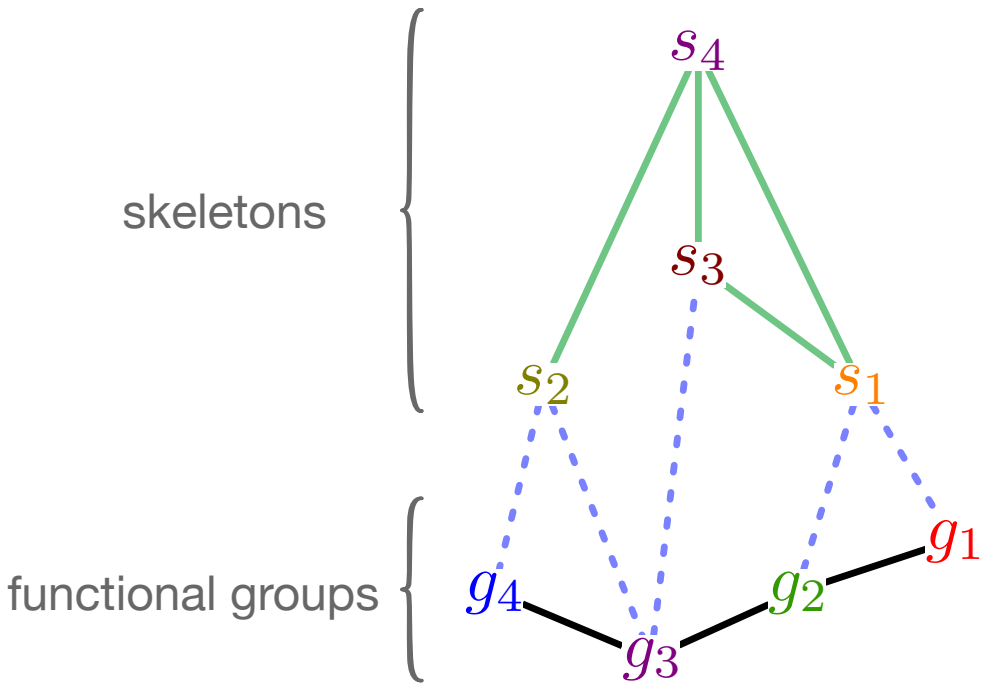






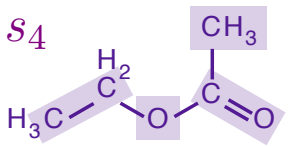


Breakdown of Skeletons & Groups

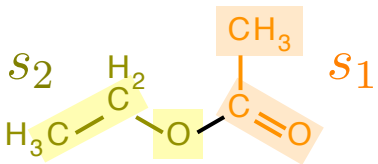


Legend connectedness ——— incidence - - - - - mereology ———

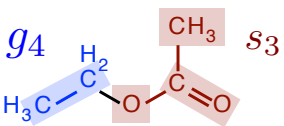
Breaking Down Skeletons (cont.)



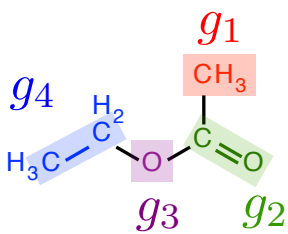
$$\forall x\ ethyl\_acetate(x) \supset\ skeleton(x)$$



$$\begin{aligned}\forall x\ ethyl\_acetate(x) \equiv\ & \exists s_1 \exists s_2 \exists b_1\ skeleton(x) \wedge \\ & acetic\_acid(s_1) \wedge ethanol(s_2) \wedge \\ & mol(s_1, x) \wedge mol(s_2, x) \wedge \\ & tether(s_1, s_2, b_1)\end{aligned}$$



$$\begin{aligned}\forall x\ ethyl\_acetate(x) \equiv\ & \exists g_4 \exists s_3 \exists b_1\ skeleton(x) \wedge \\ & acetyl\_oxy(s_3) \wedge ethane(g_4) \wedge \\ & mol(s_3, x) \wedge mol(g_4, x) \wedge \\ & tether(s_1, s_2, b_1)\end{aligned}$$



$$\begin{aligned}\forall x\ ethyl\_acetate(x) \equiv\ & \exists g_1 \exists g_2 \exists g_3 \exists g_4 \exists b_1 \exists b_2 \exists b_3\ skeleton(x) \wedge \\ & methyl(g_1) \wedge carbonyl(g_2) \wedge ether(g_3) \wedge \\ & ethane(g_4) \wedge mol(g_1, x) \wedge mol(g_2, x) \wedge \\ & mol(g_3, x) \wedge mol(g_4, x) \wedge tether(g_1, g_2, b_1) \wedge \\ & tether(g_2, g_3, b_2) \wedge tether(g_3, g_4, b_3)\end{aligned}$$

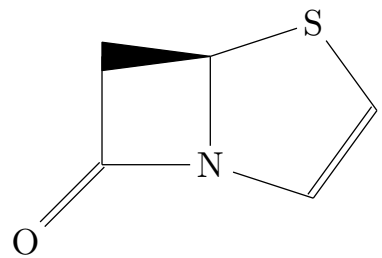
Molecular Descriptions

Now that we've summarized the components of MoSt, what does this mean?

- We can describe molecules as logical axioms in **first-order logic**, which are compact enough to be read by humans.
- With a **theorem prover**, we can use the axioms to reason about molecular shape, and answer the similarity/substitution/synthesis queries.

Example: Penem (C_5H_5NOS)

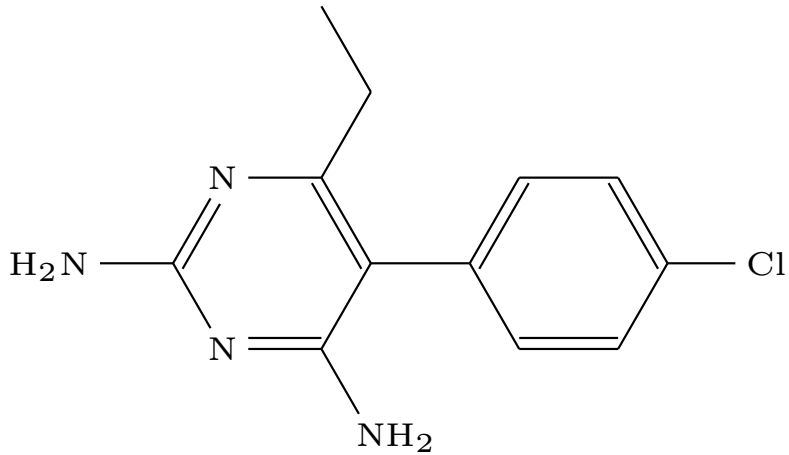
- Full name: 2,3-didehydropenam
- (5R)-4-thia-1-azabicyclo[3.2.0]hept-2-ene with a 7-keto substituent
- Core structure found in penicillin



$$\forall x\, penem(x) \equiv \exists g_1 \exists g_2\, skeleton(x) \wedge thiazolidine(g_1) \wedge beta\_lactam(g_2) \wedge mol(g_1, x) \wedge mol(g_2, x) \wedge fused(g_1, g_2)$$

Example: Daraprim ($C_{12}H_{13}ClN_4$)

- Full name: 5-(4-chlorophenyl)-6-ethylpyrimidine-2,4-diamine



$$\begin{aligned}\forall x\, pyrimethamine(x) \equiv\ & skeleton(x) \wedge \exists w \exists y \exists z \\ & \exists a_1 \exists a_2 \exists b_1 \exists b_2\ chlorophenyl(w) \wedge \\ & 2,4\_Diaminopyrimidine(y) \wedge \\ & ethane(z) \wedge mol(w,x) \wedge \\ & mol(y,x) \wedge mol(z,x) \wedge \\ & tether(w,y,b_1) \wedge \\ & tether(y,z,b_2)\end{aligned}$$

Scaffolds in Chemistry

In addition to representing molecules, we also want to use the ontology to **design new molecules**. This provides us with the opportunity to aid the scaffold work done in the medicinal chemistry community:

- **Scaffolds** are a fixed part of a molecule where functional groups can be substituted or exchanged.
- Structures sharing a scaffold can often be assumed to share a common synthetic pathway [Sch+07].
- Used to define classes of chemical compounds and reduce the chemical search space in drug design [Bro13].

Scaffolds in Chemistry (cont.)

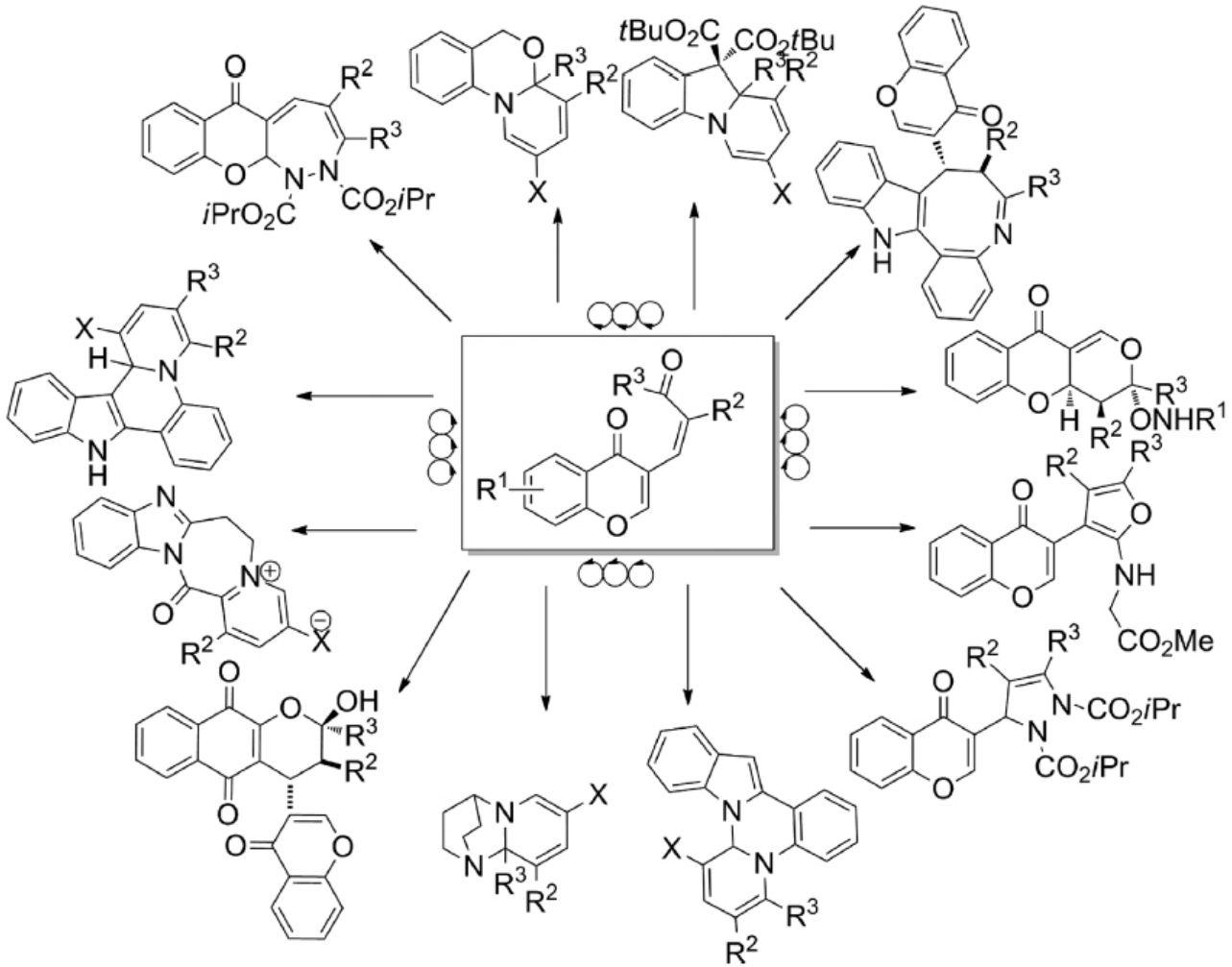


Figure: A map of skeletal diversity with scaffolds in [Lac+12].

Summary

- We have proposed a set of competency questions and requirements for an ontology of molecular structure.
- We have developed an axiomatization for such an ontology.
  - See our FOIS 2016 conference paper [CG16] for more details.
  - MoSt axioms in Common Logic: `http://colore.oor.net/most/`

Thank You!

Any Questions?

Feedback

If you have any questions or comments, feel free to contact us:

- Professor Michael Grüninger: gruninger@mie.utoronto.ca
- Carmen Chui: cchui@mie.utoronto.ca

Visitor Address:

Semantic Technologies Lab
Room 8120 (BA 8120)
Bahen Centre for Information Technology
40 St George St
University of Toronto
http://stl.mie.utoronto.ca

References & Additional Links I

[Bro13] Nathan Brown. “Identifying and Representing Scaffolds”. In: *Scaffold Hopping in Medicinal Chemistry*. Wiley-VCH Verlag GmbH & Co. KGaA, 2013, pp. 1–14. ISBN: 9783527665143. URL: http://dx.doi.org/10.1002/9783527665143.ch01.

[CG16] Carmen Chui and Michael Grüninger. “A Molecular Structure Ontology for Medicinal Chemistry”. In: *Formal Ontology in Information Systems - Proceedings of the 9th International Conference, FOIS 2016, Annecy, France, July 6-9, 2016*. 2016, pp. 285–298. URL: http://dx.doi.org/10.3233/978-1-61499-660-6-285.

[Koc+05] Marcus A. Koch et al. “Charting biologically relevant chemical space: A structural classification of natural products (SCONP)”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.48 (2005), pp. 17272–17277. URL: http://www.pnas.org/content/102/48/17272.abstract.

[Lac+12] Hugo Lachance et al. “Charting, Navigating, and Populating Natural Product Chemical Space for Drug Discovery”. In: *Journal of Medicinal Chemistry* 55.13 (2012). PMID: 22537178, pp. 5989–6001. URL: http://dx.doi.org/10.1021/jm300288g.

References & Additional Links II

[Nat15] National Center for Biotechnology Information. *PubChem Compound Database - Morphine (CID=5288826)*. 2015. URL: https://pubchem.ncbi.nlm.nih.gov/compound/5288826.

[RA12] Jean-Louis Reymond and Mahendra Awale. “Exploring Chemical Space for Drug Discovery Using the Chemical Universe Database”. In: *ACS Chemical Neuroscience* 3.9 (2012). PMID: 23019491, pp. 649–657. URL: https://doi.org/10.1021/cn3000422%20https://doi.org/10.1021/cn3000422.

[Sch+07] Ansgar Schuffenhauer et al. “The Scaffold Tree - Visualization of the Scaffold Universe by Hierarchical Scaffold Classification”. In: *Journal of Chemical Information and Modeling* 47.1 (2007). PMID: 17238248, pp. 47–58. URL: https://doi.org/10.1021/ci600338x.

Links to Useful Information

- First-Order Logic (FOL): https://milnepublishing.geneseo.edu/concise-introduction-to-logic/chapter/16-summary-of-first-order-logic/
- First-Order Resolution: https://en.wikipedia.org/wiki/Resolution_(logic)#Resolution_in_first_order_logic
- TPTP Syntax/Grammar: http://www.tptp.org/TPTP/SyntaxBNF.html
- The CADE ATP System Competition: http://www.tptp.org/CASC/
- Vampire Theorem Prover: https://vprover.github.io/