An Ontological Approach to Medicinal Chemistry

Michael Grüninger and Carmen Chui

Semantic Technologies Lab, University of Toronto

July 24, 2017

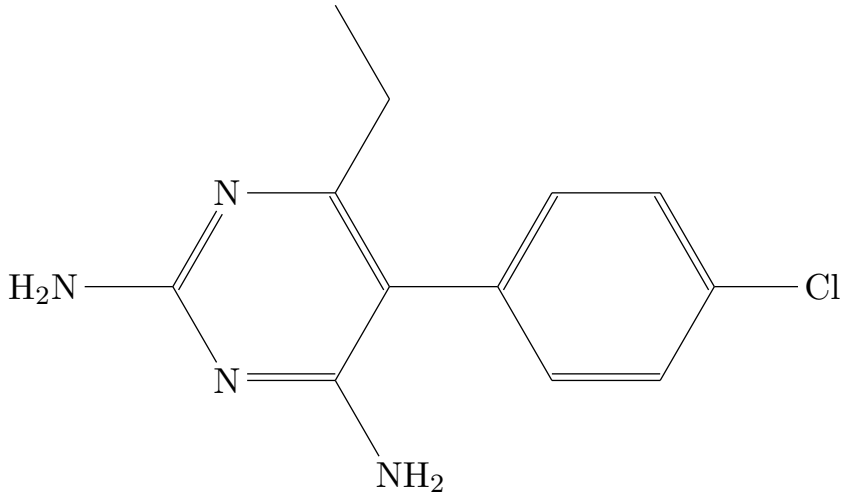
Agenda

1. Motivations
   - Ontologies
   - Competency Questions
   - Ontological Commitments as Requirements
2. The Molecular Structure Ontology (MoSt)
   - Overview
   - Examples
   - Scaffolds in Chemistry
3. Summary

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



Approach #1: Cheminformatics (Identifiers)

- Names:
  - Pyrimethamine; 58-14-0; Daraprim; Chloridine; 5-(4-chlorophenyl)-6-ethylpyrimidine-2,4-diamine; Ethylpyrimidine;
- IUPAC name:

  `5-(4-chlorophenyl)-6-ethylpyrimidine-2,4-diamine`
- InChI identifier:

  `InChI=1S/C12H13ClN4/c1-2-9-10(11(14)17-12(15)16-9)7-3-5-8(13)6-4-7/h3-6H,2H2,1H3,(H4,14,15,16,17)`

Approach #2: Cheminformatics (molfile output)

```
 17 18  0  0  0  0            999 V2000
   18.1293  -10.4004    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   17.4147  -10.8129    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   18.1293   -9.5754    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   18.8437  -10.8129    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   17.4147  -11.6379    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   16.7003  -10.4004    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   18.8437   -9.1629    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
   17.4147   -9.1629    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   19.5582  -10.4004    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
   18.8437  -11.6379    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
   16.7003  -12.0504    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   15.9858  -10.8129    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   19.5582   -9.5754    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   16.7003   -9.5754    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   15.9858  -11.6379    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   20.2727   -9.1629    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
   15.2713  -12.0504    0.0000 Cl  0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  1  3  2  0  0  0  0
  1  4  1  0  0  0  0
  2  5  2  0  0  0  0
  2  6  1  0  0  0  0
            ...
 13 16  1  0  0  0  0
 15 17  1  0  0  0  0
  9 13  1  0  0  0  0
 12 15  1  0  0  0  0
M  END
```

Approach #3: Chemical Entities of Biological Interest (ChEBI)

(Excerpts[1] extracted from ChEBI)

- pyrimethamine (CHEBI:8673) **has role** antimalarial (CHEBI:38068)
- pyrimethamine (CHEBI:8673) **has role** antiprotozoal drug (CHEBI:35820)
- pyrimethamine (CHEBI:8673) **has role** EC 1.5.1.3 (dihydrofolate reductase) inhibitor (CHEBI:50683)
- pyrimethamine (CHEBI:8673) **is a** aminopyrimidine (CHEBI:38338)
- pyrimethamine (CHEBI:8673) **is a** monochlorobenzenes (CHEBI:83403)

[1] `http://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:8673`

Approach #4: Molecular Graph Theory

- Represent a molecule as a graph
  - Vertices represent atoms
  - Edges represent bonds

How do people use these approaches?

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

How do people use these approaches?

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?

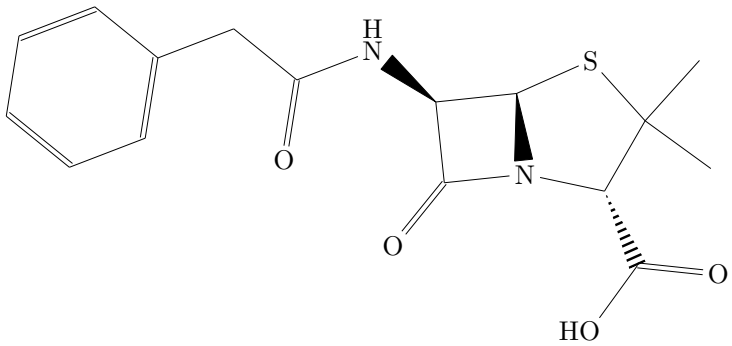


Figure: Penicillin ($C_9H_{11}N_2O_4S$)

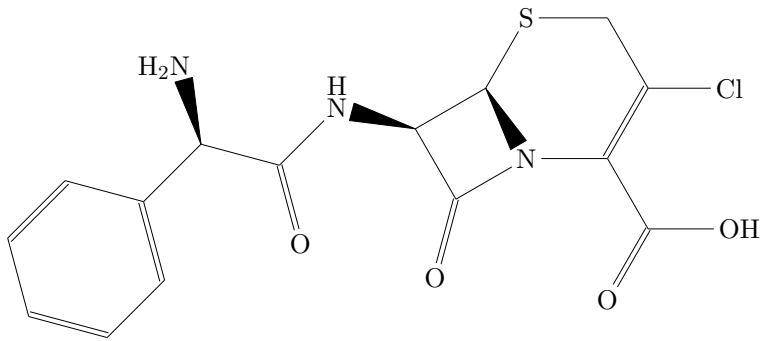


Figure: Cefaclor ($C_{15}H_{14}ClN_3O_4S$)

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?

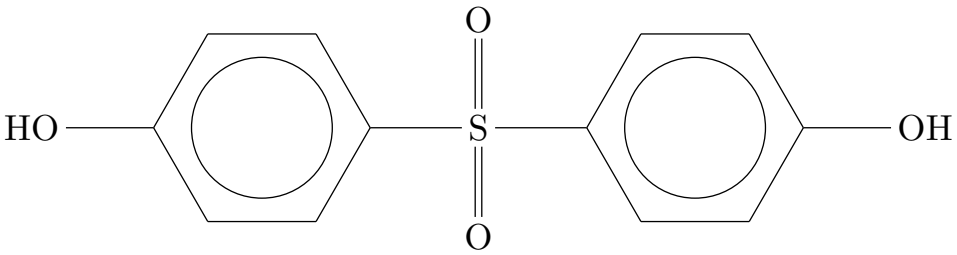


Figure: Bisphenol S (BPS)

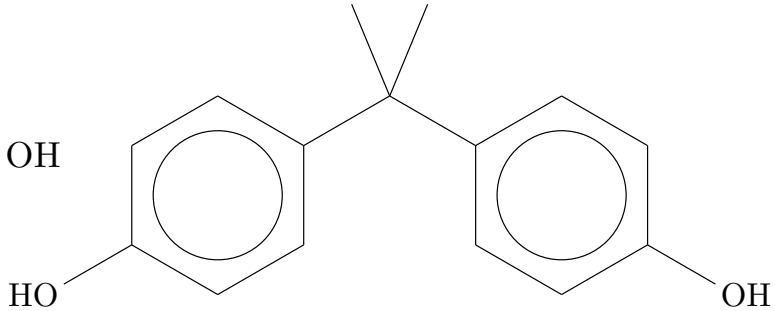


Figure: Bisphenol A (BPA)

Competency Questions

Synthesis

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

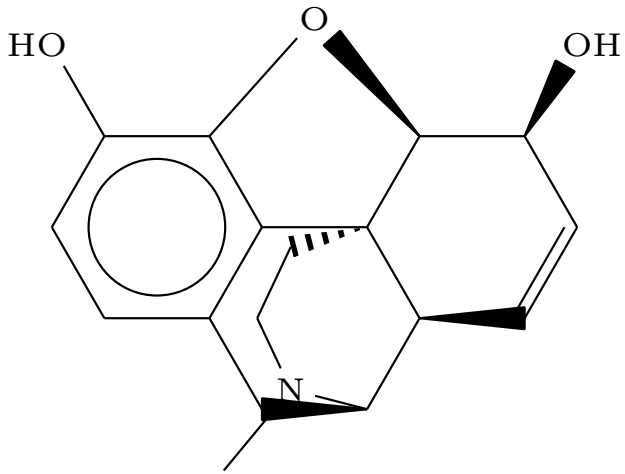


Figure: Morphine ($C_{17}H_{19}NO_3$)

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:

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 *primitives* in the ontology related by a notion of *composition*.

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

R-5 The carbon backbone of 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.

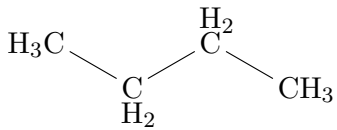


Figure: Butane

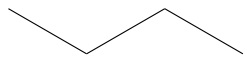


Figure: Butane (Carbon Backbone)

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 are two overlapping atoms that share a bond between the groups.

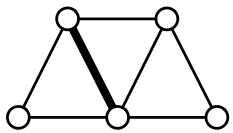


Figure: Fusion (Rings)

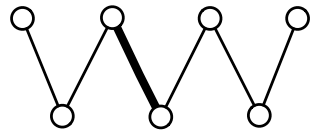


Figure: Fusion (Chain)

Ontological Commitments as Requirements (cont.)

Connections Between Functional Groups

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

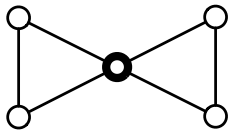


Figure: Spiro (Rings)

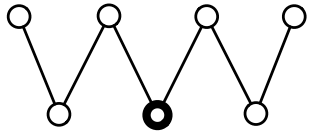


Figure: Spiro (Chain)

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.

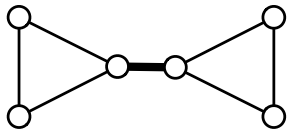


Figure: Tether (Rings)

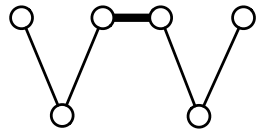


Figure: Tether (Chain)

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).

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_skeleton** satisfies R-5

most_graph

most_graph contains axioms to describe atoms and bonds and how they are connected

- Identical to existing molecular graph theory work

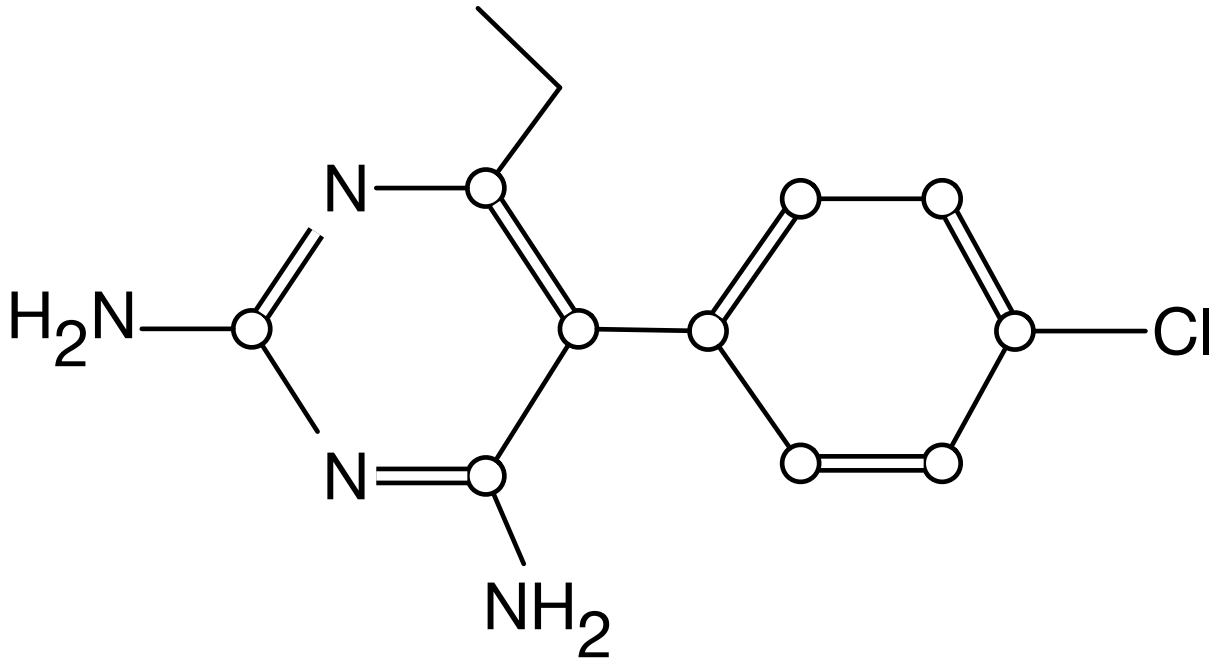


Figure: Daraprim ($C_{12}H_{13}ClN_4$)

most_graph

most_graph contains axioms to describe atoms and bonds and how they are connected

- Identical to existing molecular graph theory work

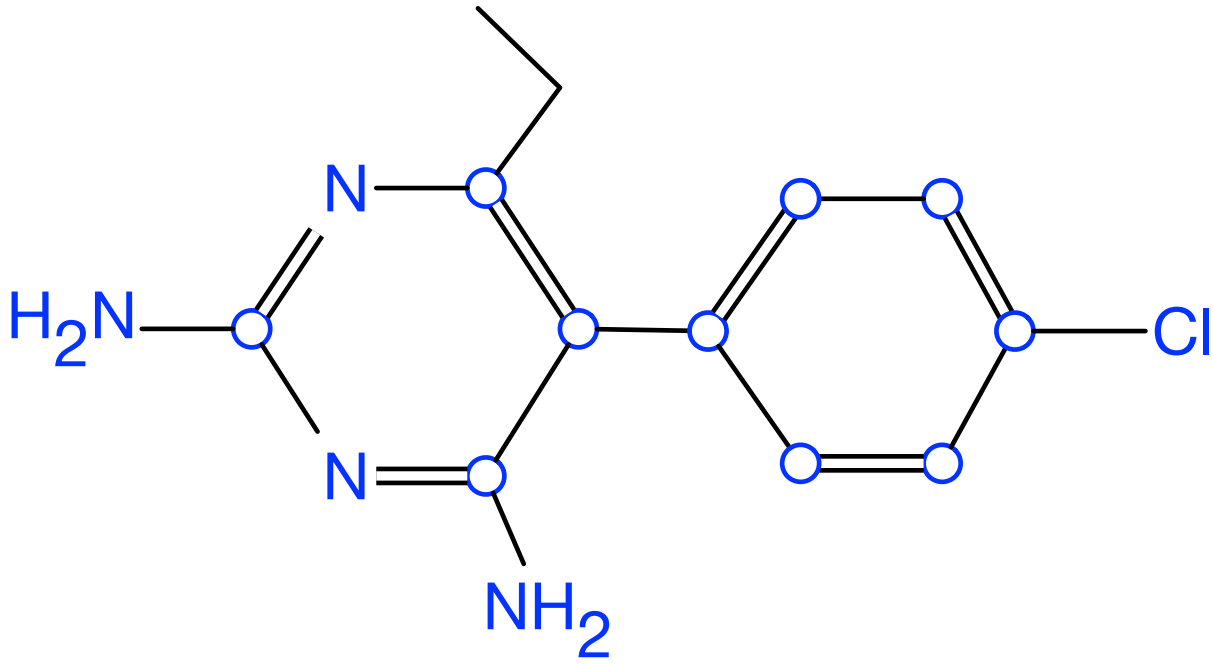


Figure: Daraprim ($C_{12}H_{13}ClN_4$)

most_graph

most_graph contains axioms to describe atoms and bonds and how they are connected

- Identical to existing molecular graph theory work

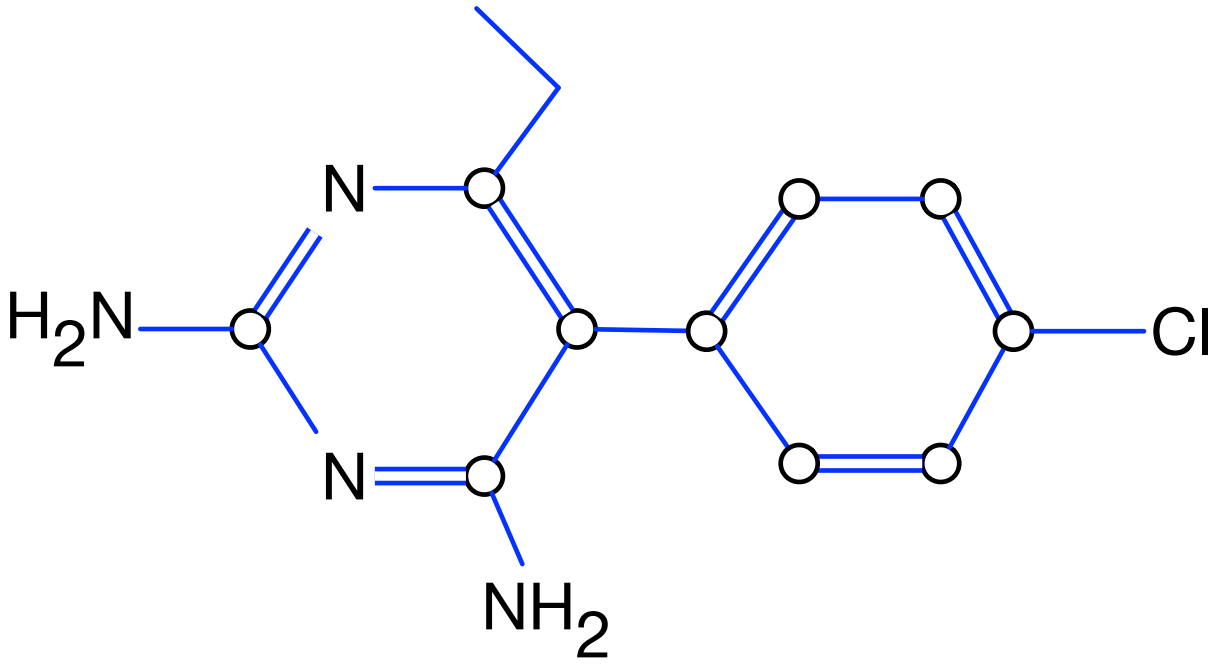


Figure: Daraprim ($C_{12}H_{13}ClN_4$)

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

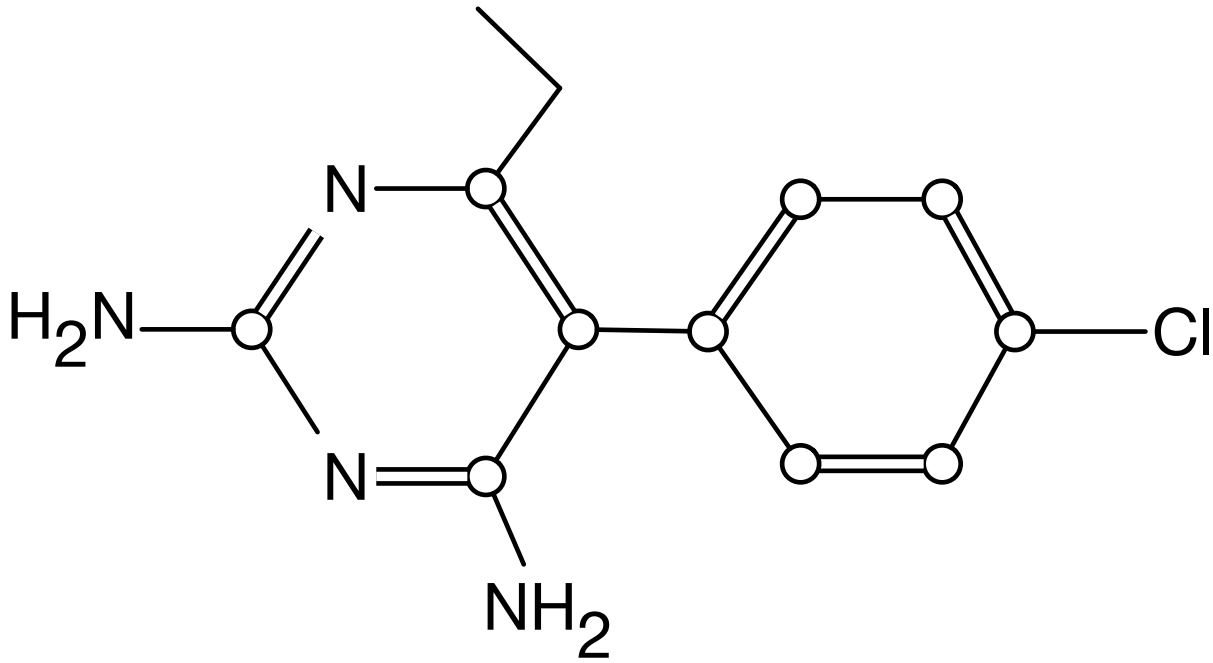


Figure: Daraprim ($C_{12}H_{13}ClN_4$)

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

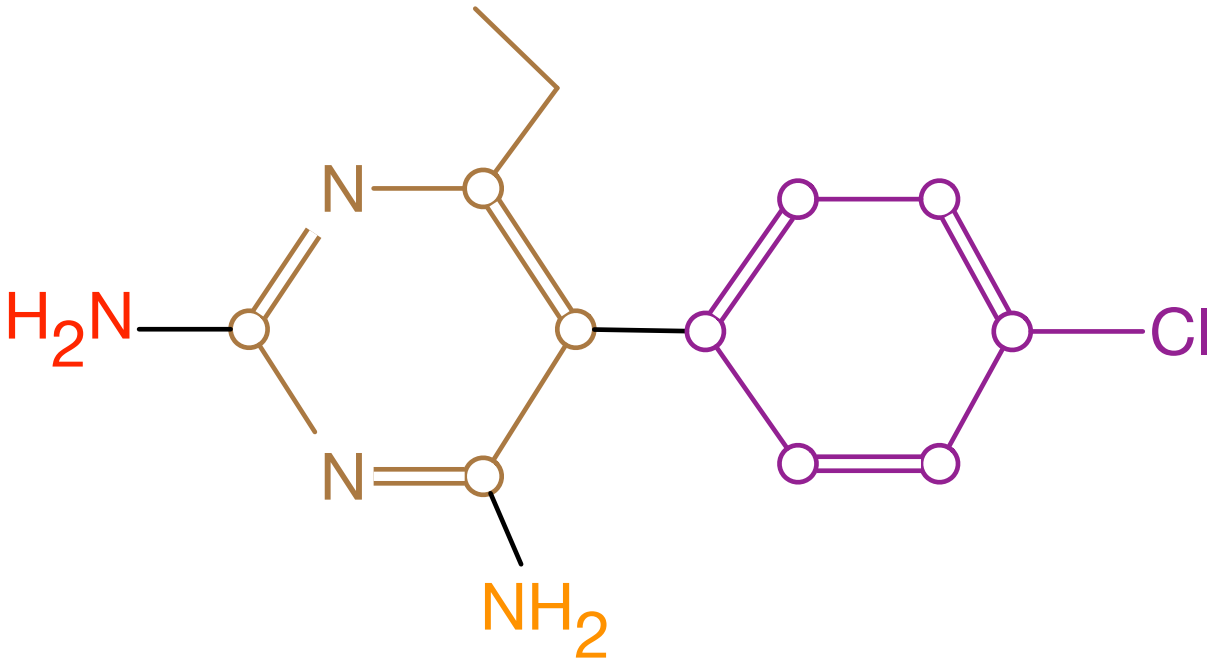


Figure: Daraprim ($C_{12}H_{13}ClN_4$)

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

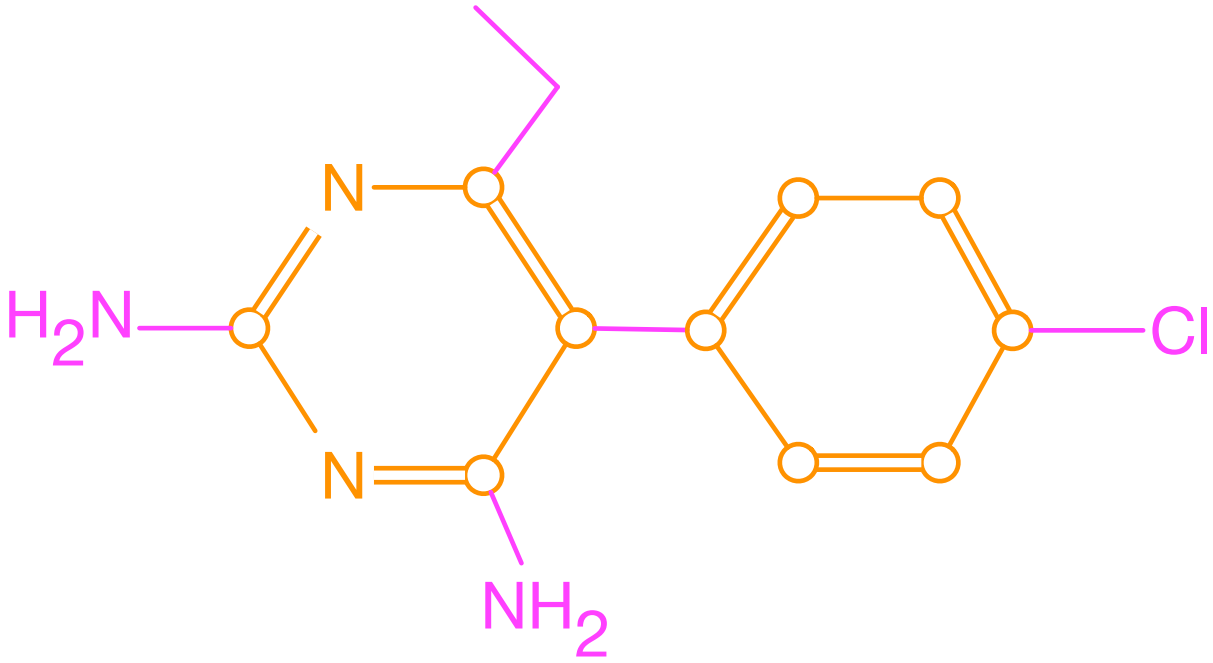
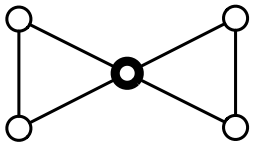


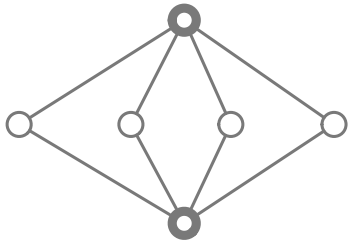
Figure: Daraprim ($C_{12}H_{13}ClN_4$)

most_attachment

Unique Spiro

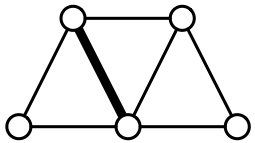


✓

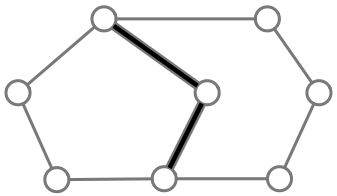


✗

Unique Fusion



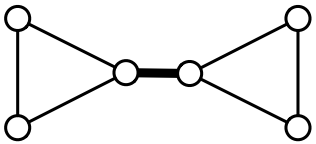
✓



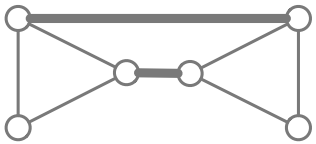
considered a bridge in the ontology

✗

Unique Tether



✓



✗

most_attachment (cont.)

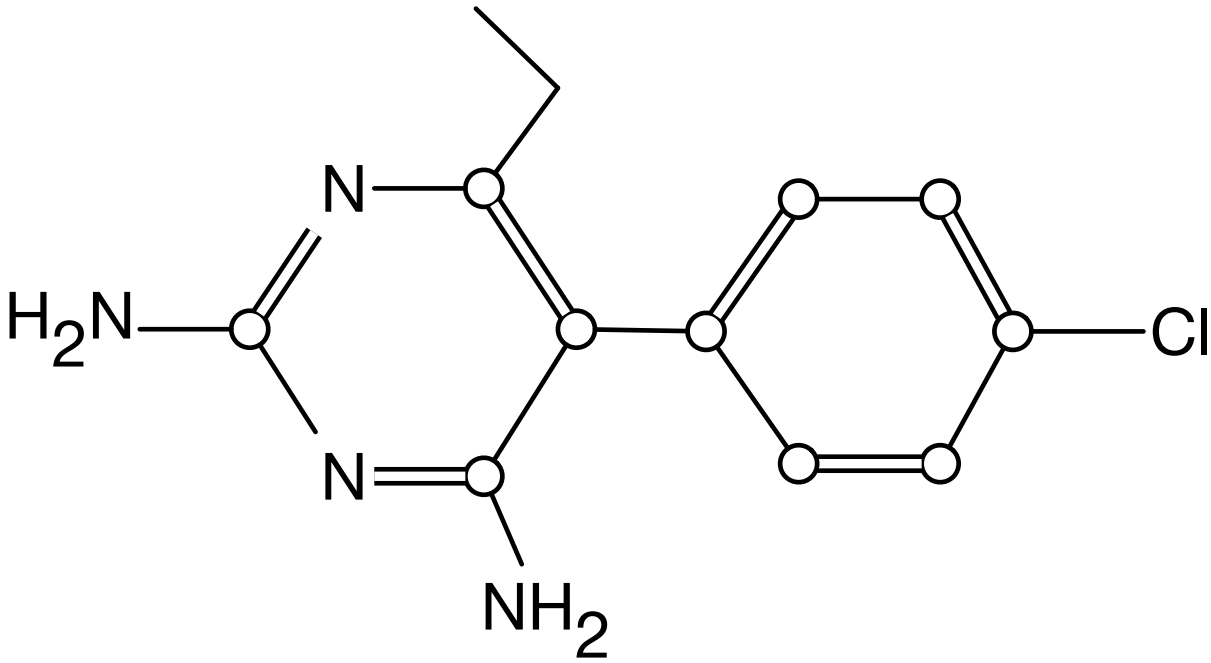


Figure: Daraprim ($C_{12}H_{13}ClN_4$)

most_attachment (cont.)

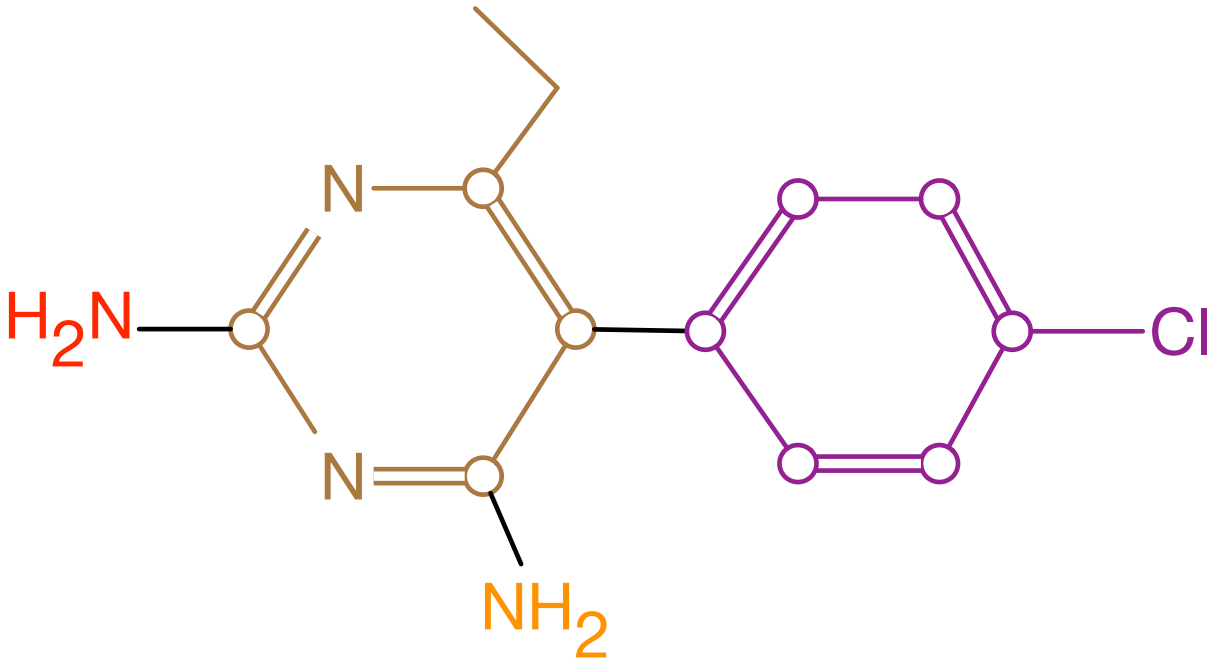


Figure: Daraprim ($C_{12}H_{13}ClN_4$)

most_attachment (cont.)

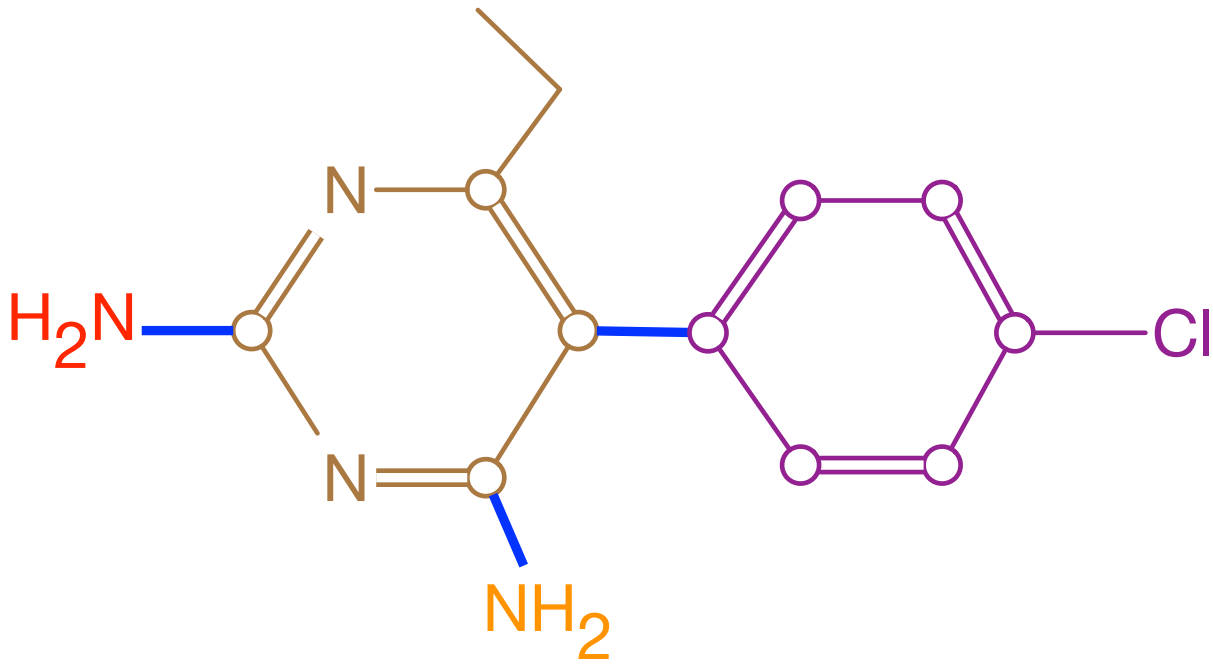


Figure: Daraprim ($C_{12}H_{13}ClN_4$)

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

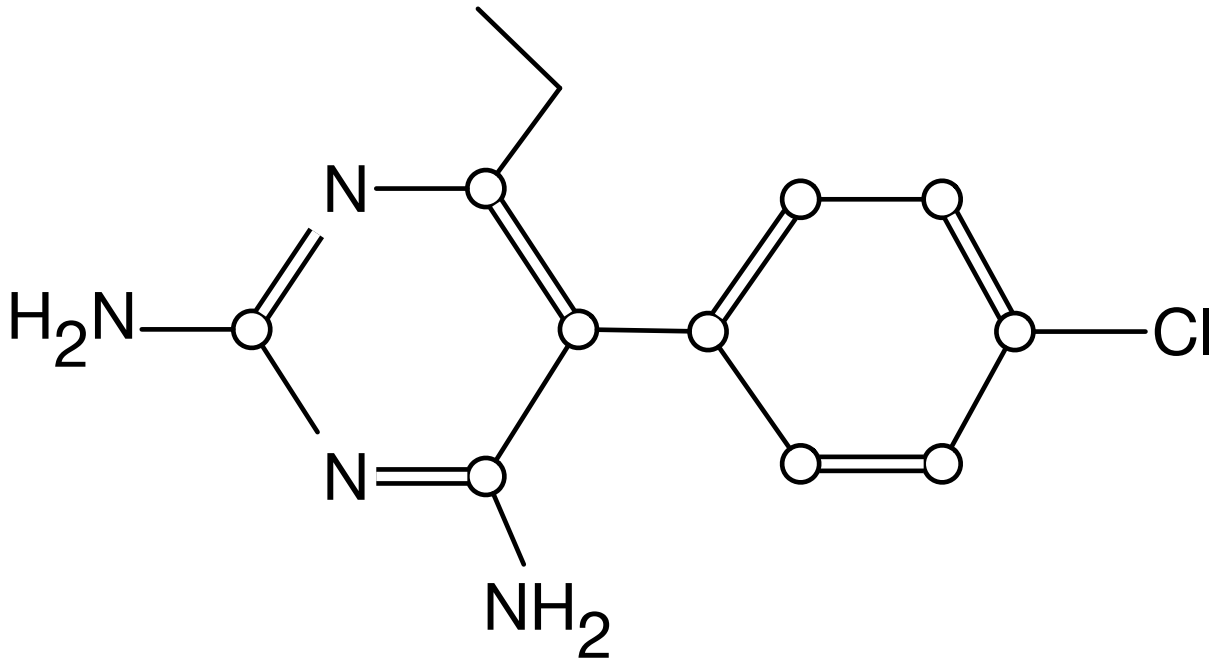


Figure: Daraprim ($C_{12}H_{13}ClN_4$)

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

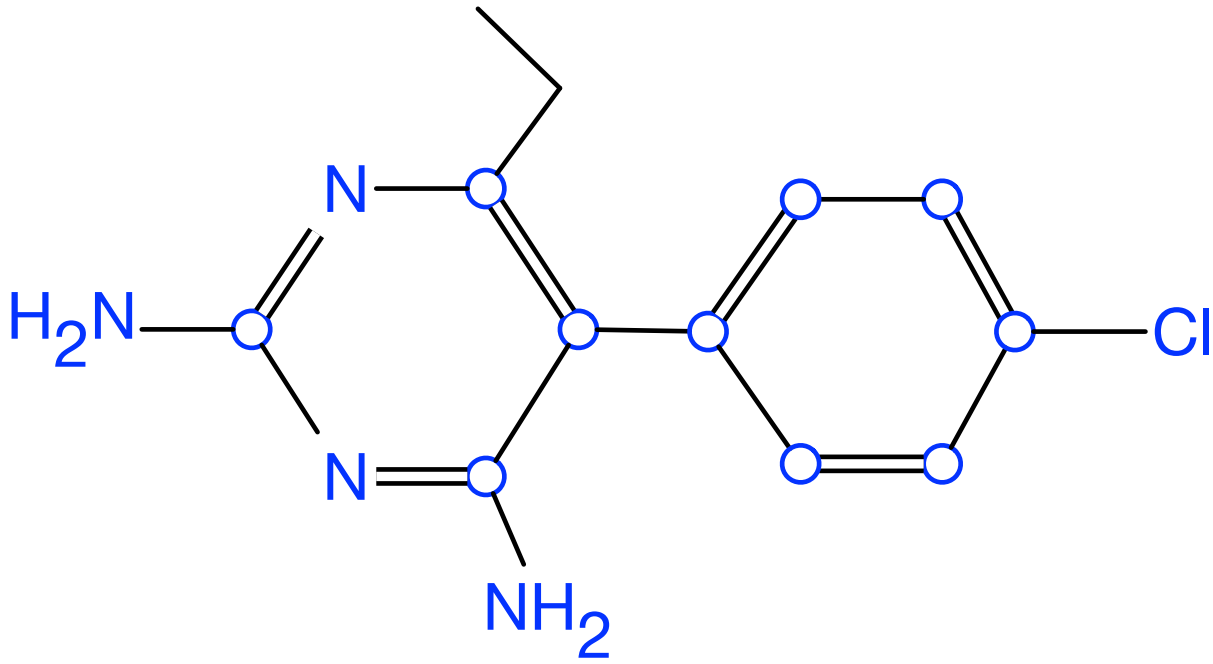


Figure: Daraprim ($C_{12}H_{13}ClN_4$)

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

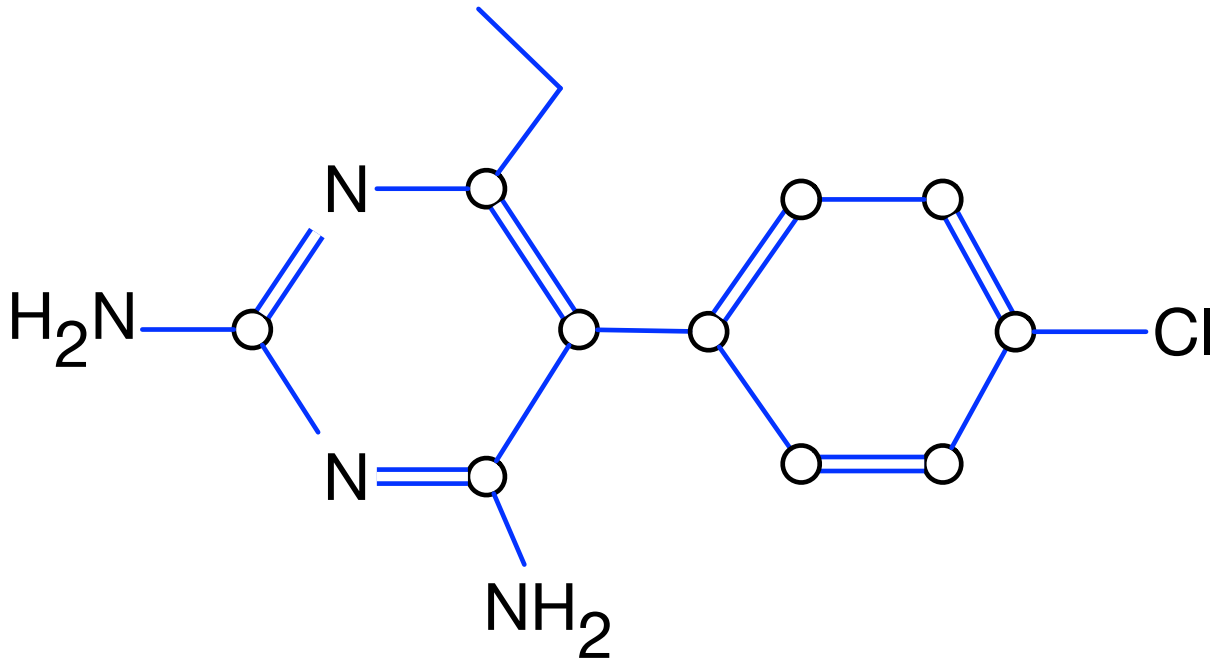


Figure: Daraprim ($C_{12}H_{13}ClN_4$)

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

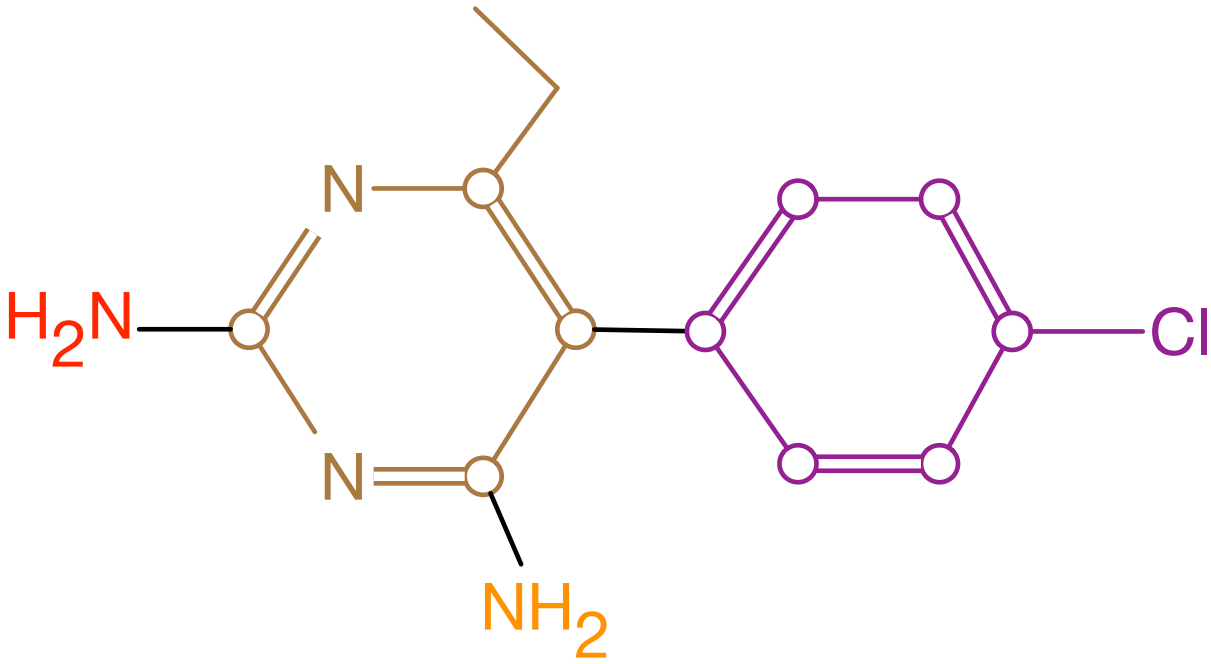


Figure: Daraprim ($C_{12}H_{13}ClN_4$)

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

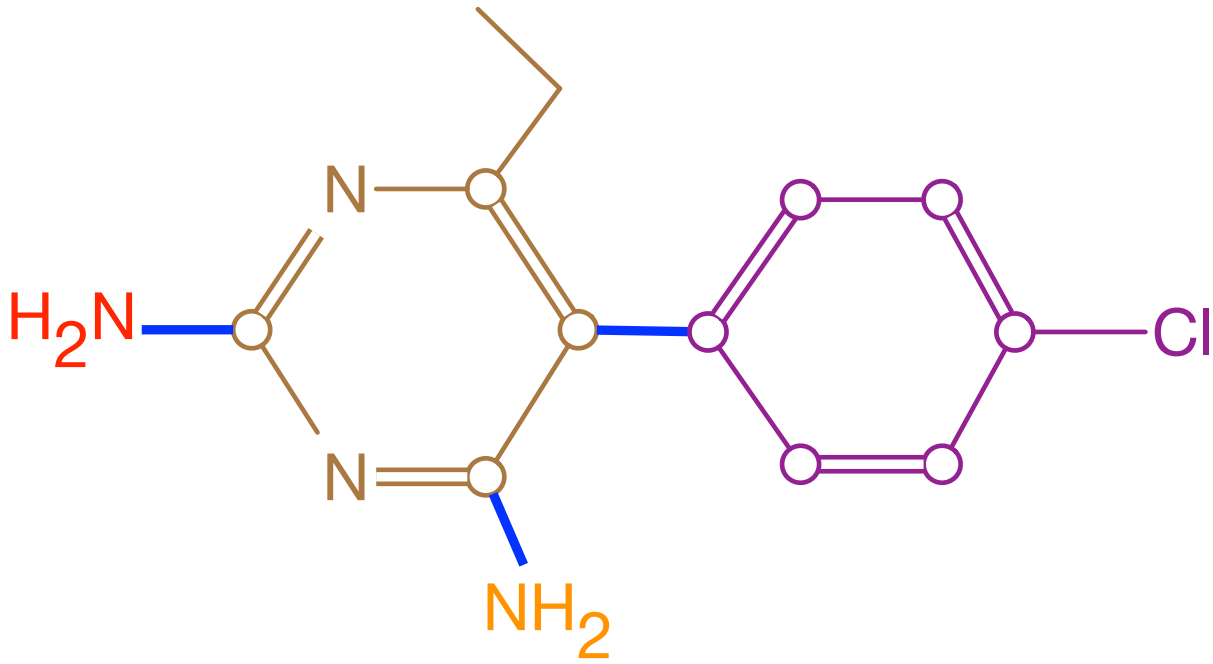


Figure: Daraprim ($C_{12}H_{13}ClN_4$)

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

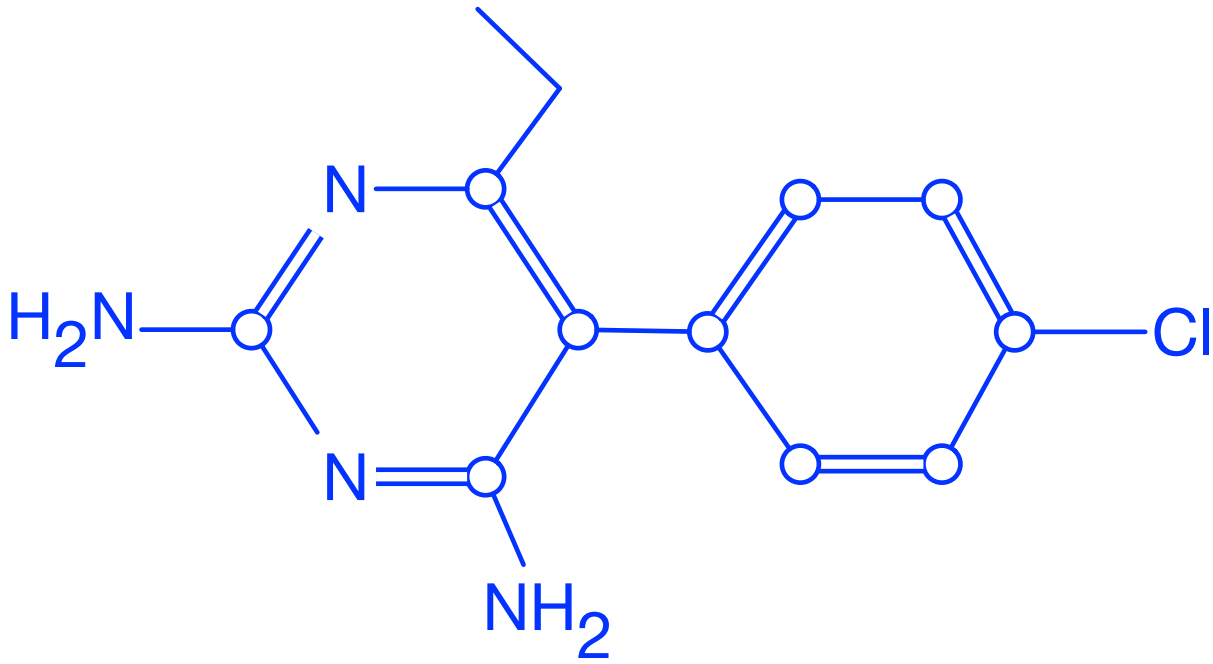


Figure: Daraprim ($C_{12}H_{13}ClN_4$)

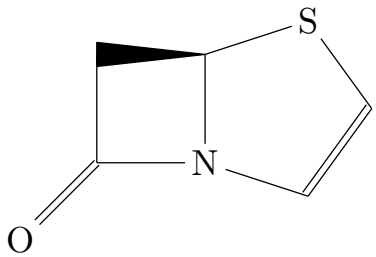
Molecular Descriptions

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)

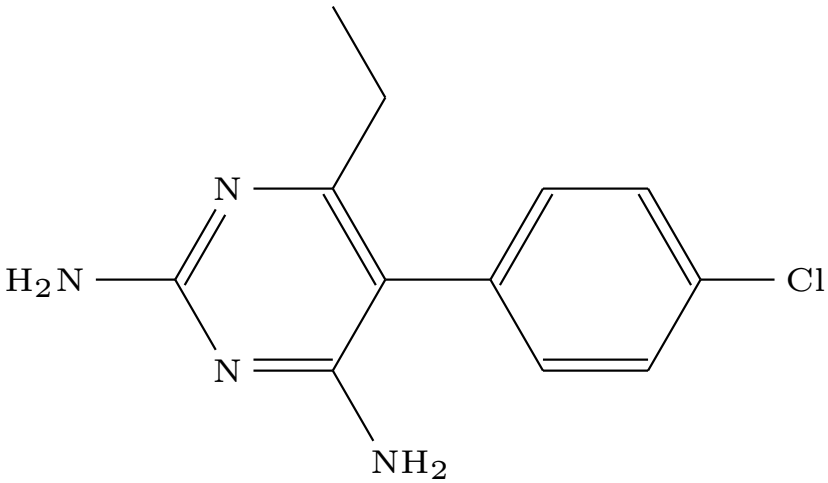
- Full name: 2,3-didehydropenam
- (5R)-4-thia-1-azabicyclo[3.2.0]hept-2-ene with a 7-keto substituent
- Core structure found in penicillin



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

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

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



$$\forall x\ pyrimethamine(x) \equiv skeleton(x) \wedge$$
$$\exists w \exists y \exists a_1 \exists a_2 \exists u \exists z \exists b_1 \exists b_2 \exists b_3\ chlorophenyl(w) \wedge$$
$$ethyl\_pyrimidine(y) \wedge$$
$$amine(a_1) \wedge amine(a_2) \wedge mol(w, x) \wedge$$
$$mol(y, x) \wedge mol(a_1, x) \wedge mol(a_2, x) \wedge$$
$$tether(w, y, b_1) \wedge$$
$$tether(y, a_1, b_2) \wedge tether(y, a_2, b_3)$$

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.)

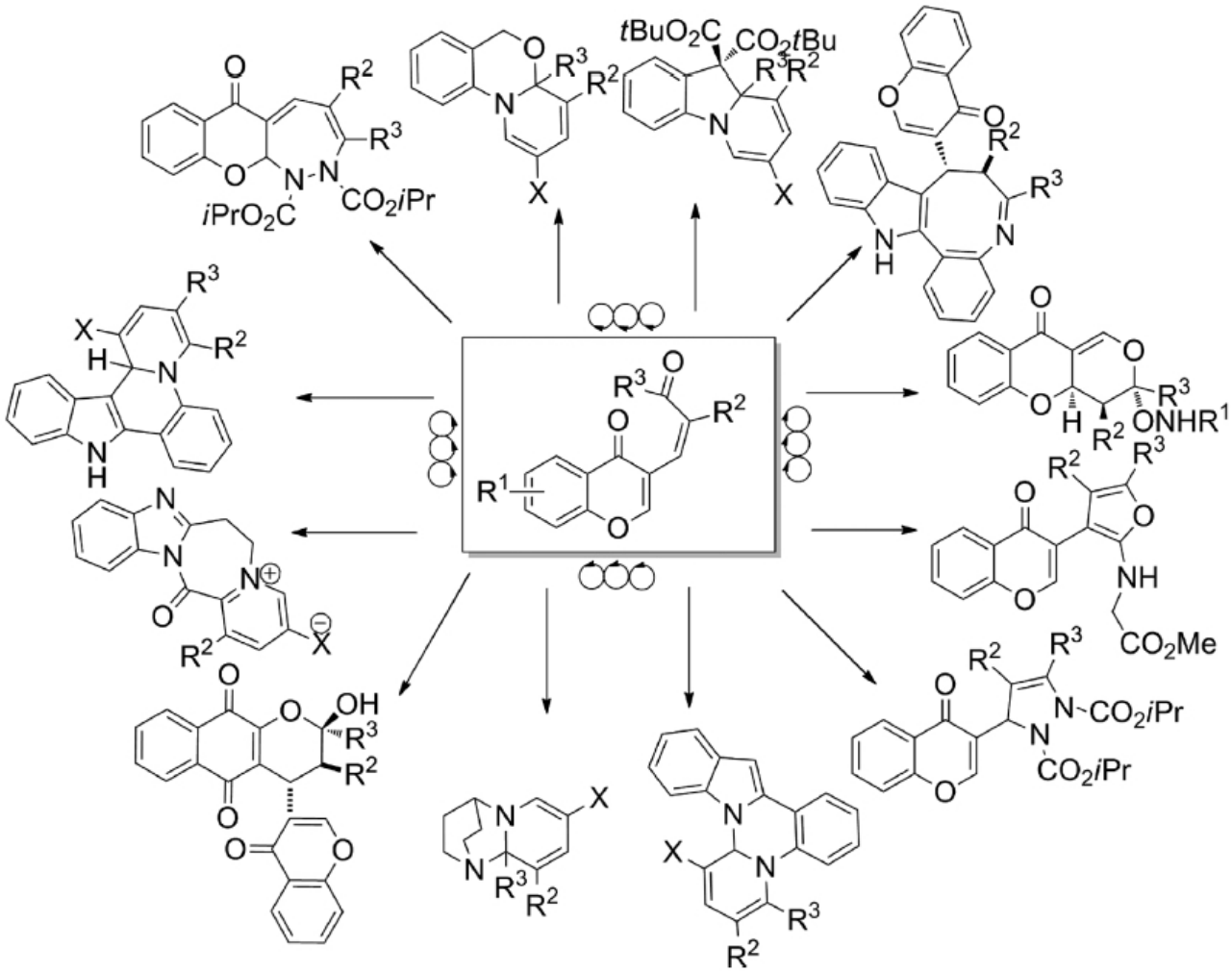


Figure: A map of skeletal diversity with scaffolds in [Lac+12].

Scaffolds in Chemistry (cont.)

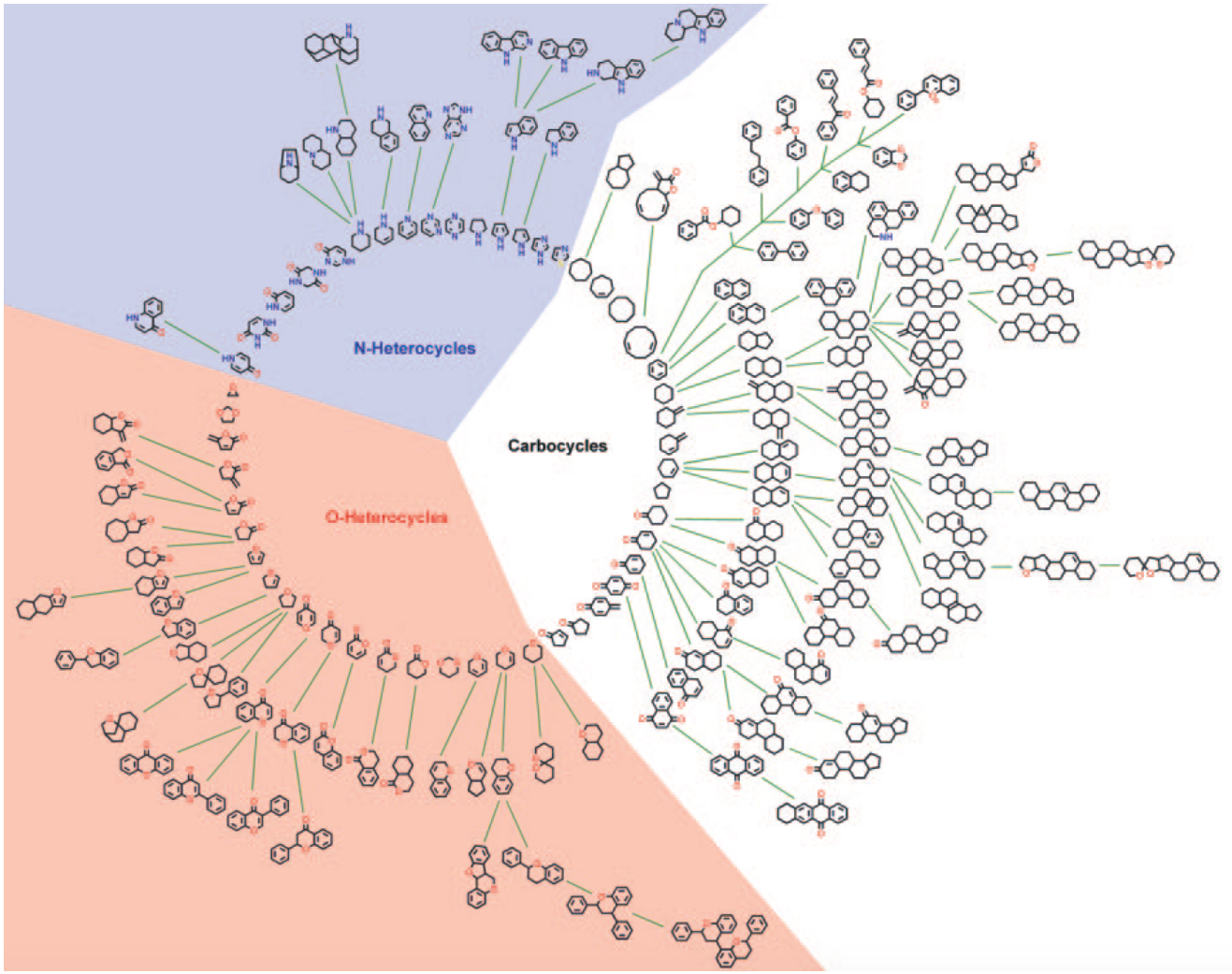


Figure: Natural product (NP) scaffold tree as presented in [Koc+05].

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.

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. DOI: 10.1002/9783527665143.ch01. 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. DOI: 10.3233/978-1-61499-660-6-285. URL: https://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. DOI: 10.1073/pnas.0503647102. eprint: http://www.pnas.org/content/102/48/17272.full.pdf. URL: http://www.pnas.org/content/102/48/17272.abstract.

References & Additional Links II

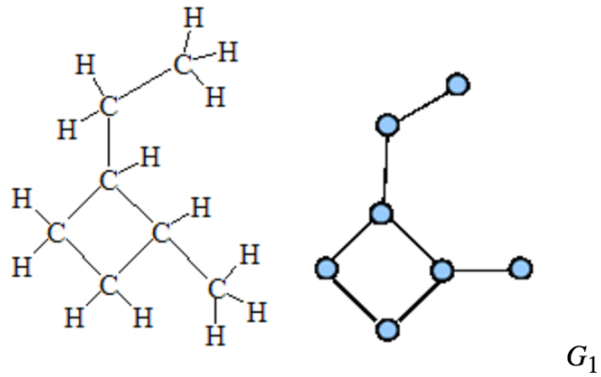
[Lac+12] Hugo Lachance et al. "Charting, Navigating, and Populating Natural Product Chemical Space for Drug Discovery". In: *Journal of Medicinal Chemistry* 55.13 (2012). PMID: 22537178, pp. 5989–6001. DOI: 10.1021/jm300288g. eprint: http://dx.doi.org/10.1021/jm300288g. URL: http://dx.doi.org/10.1021/jm300288g.

[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. DOI: 10.1021/ci600338x. eprint: http://dx.doi.org/10.1021/ci600338x. URL: http://dx.doi.org/10.1021/ci600338x.

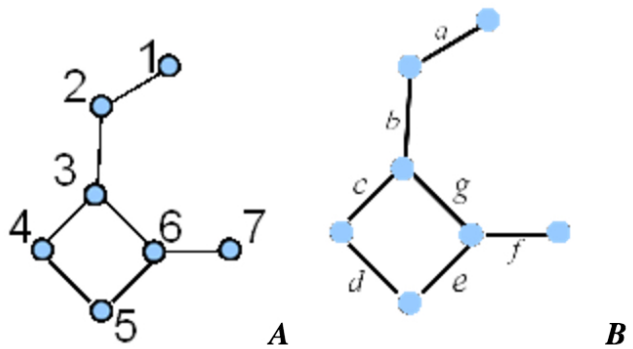
Appendix

Molecular Graphs

Vertices and edges are labelled:



(a) Hydrogen-depleted graph G_1



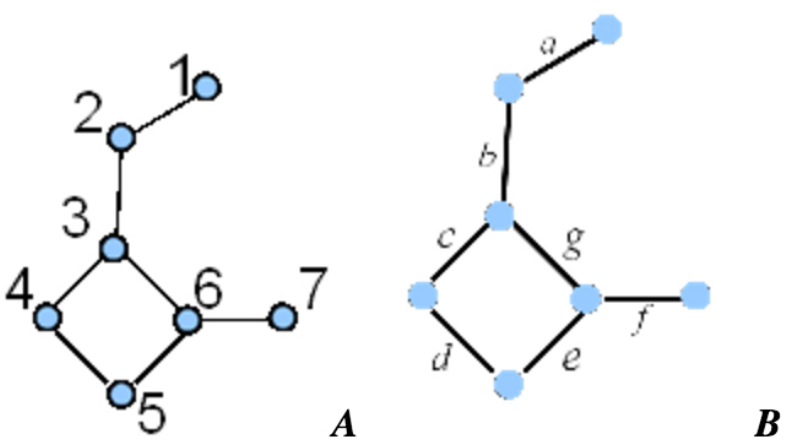
(b) Vertex-labeled (A) and edge-labeled (B) graph of G_1

Figure: 1-ethyl-2-methylcyclobutane

Molecular Graphs (cont.)

Graphs can also be represented with a **Vertex-Adjacency Matrix** $({}^{v}\mathbf{A})$[2]:

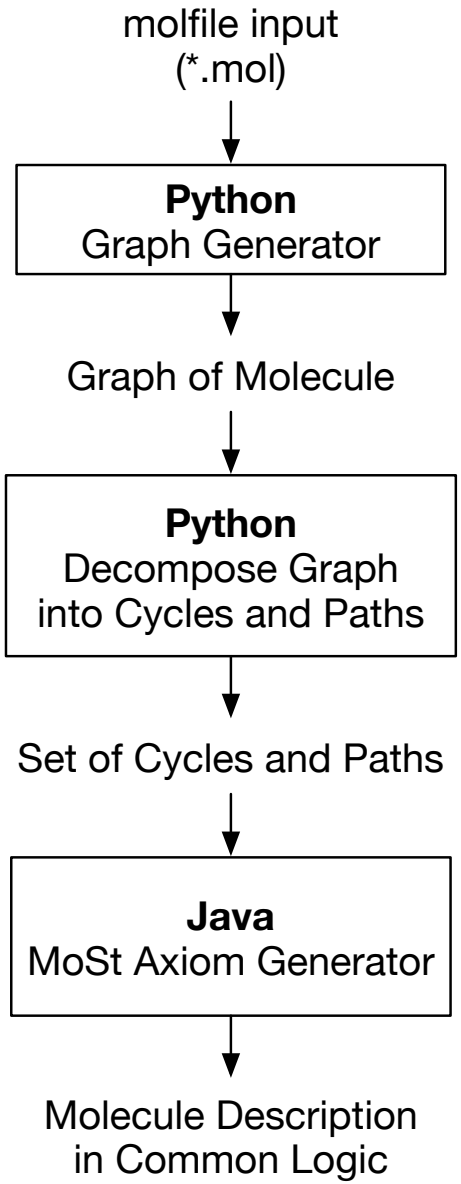
$$[{}^{v}\mathbf{A}]_{ij} = \begin{cases} 1, \text{if vertices } i \text{ and } j \text{ are adjacent} \\ 0, \text{otherwise} \end{cases} \tag{1}$$



$${}^{v}\mathbf{A}(G_1) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

[2] See `http://www.sicmm.org/~FAMNIT-knjiga/wwwANG/index1.htm`

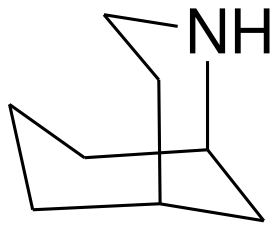
Molecule Description Generator



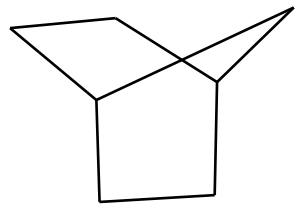
- Developed in Python with the help of an M.Eng. student to parse a `.mol` file to decompose the molecular graph into functional groups
- A module programmed in Java takes this functional group information to output a molecule description in first-order logic

most_attachment

What about bridged molecules?



(a) morphan



(b) norbornane

most_attachment

Bridges = a chain that is spiroed at one end and tethered at the other

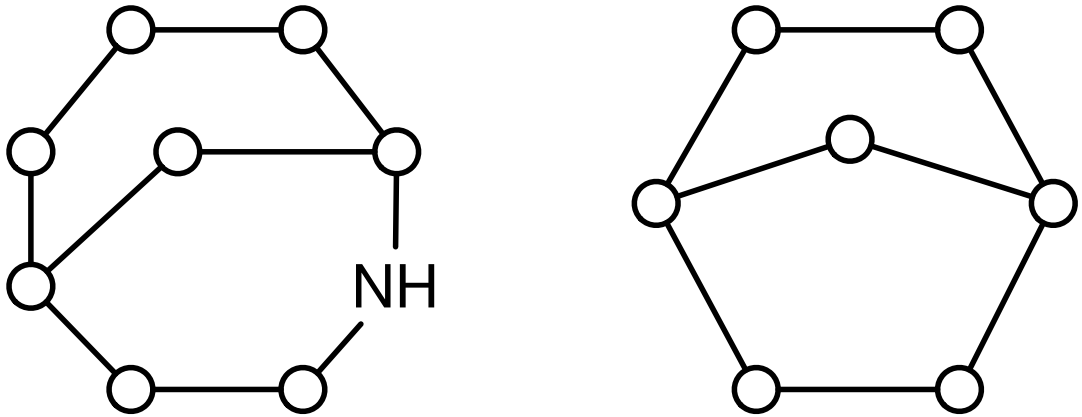
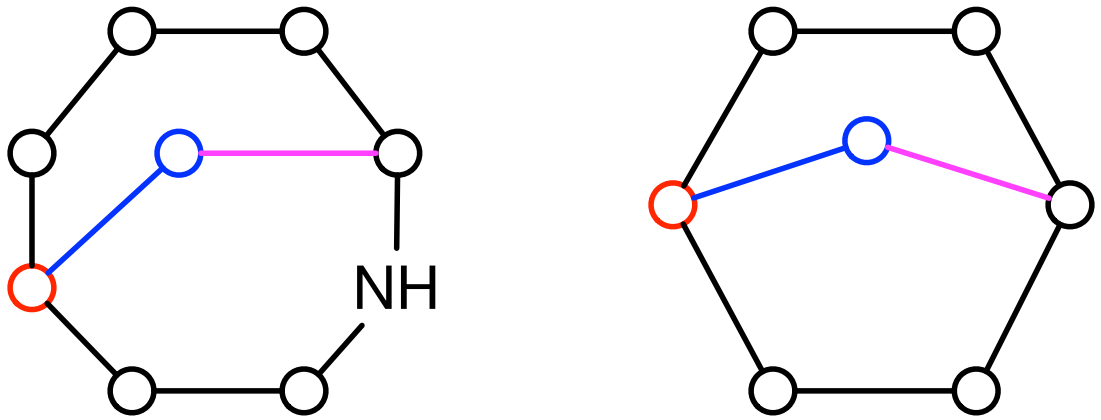


Figure: Morphan and Norbornane (redrawn topologically).

most_attachment

Bridges = a chain that is spiroed at one end and tethered at the other



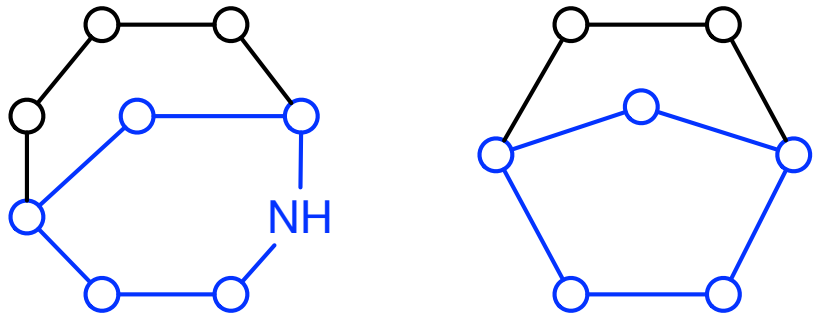
red = spiro atom blue = chain purple = tether

Figure: Morphan and Norbornane (redrawn topologically).

most_ringbond

most_ringbond

- Not all cycles in a graph correspond to a ringed functional group in chemistry
- Using an algorithmic approach, all simple cycles can be identified as rings, and the leftover paths in the graph become chains.



Reasoning with Molecular Descriptions

When we talk about **reasoning**, we are talking about queries that can be answered with a computer.

- Our first-order theorem prover of choice is **Prover9**: `https://www.cs.unm.edu/~mccune/prover9/download/`

Scaffolds - Additional Information

1	Remove three-member heterocycles
2	Retain macrocycles of greater than 11 members
3	Remove rings first by longest acyclic linker
4	Retain spiro, nonlinear, fused and bridged rings
5	Retain bridged over spiro rings
6	Remove rings of size 3, 5, and 6 first
7	Fully aromatic rings should not be removed if remaining system is not aromatic
8	Remove rings with fewest heteroatoms first
9	If (8) is equal, use precedence relationship of N > O > S
10	Remove smaller rings first
11	Retain saturated rings
12	Remove rings with a heteroatom connected to a linker
13	Tiebreaking rule based on alphabetic ordering of a canonical SMILES representation

Figure: Scaffold rules. (Table 1 in [Bro13]).

Generating Scaffolds Example

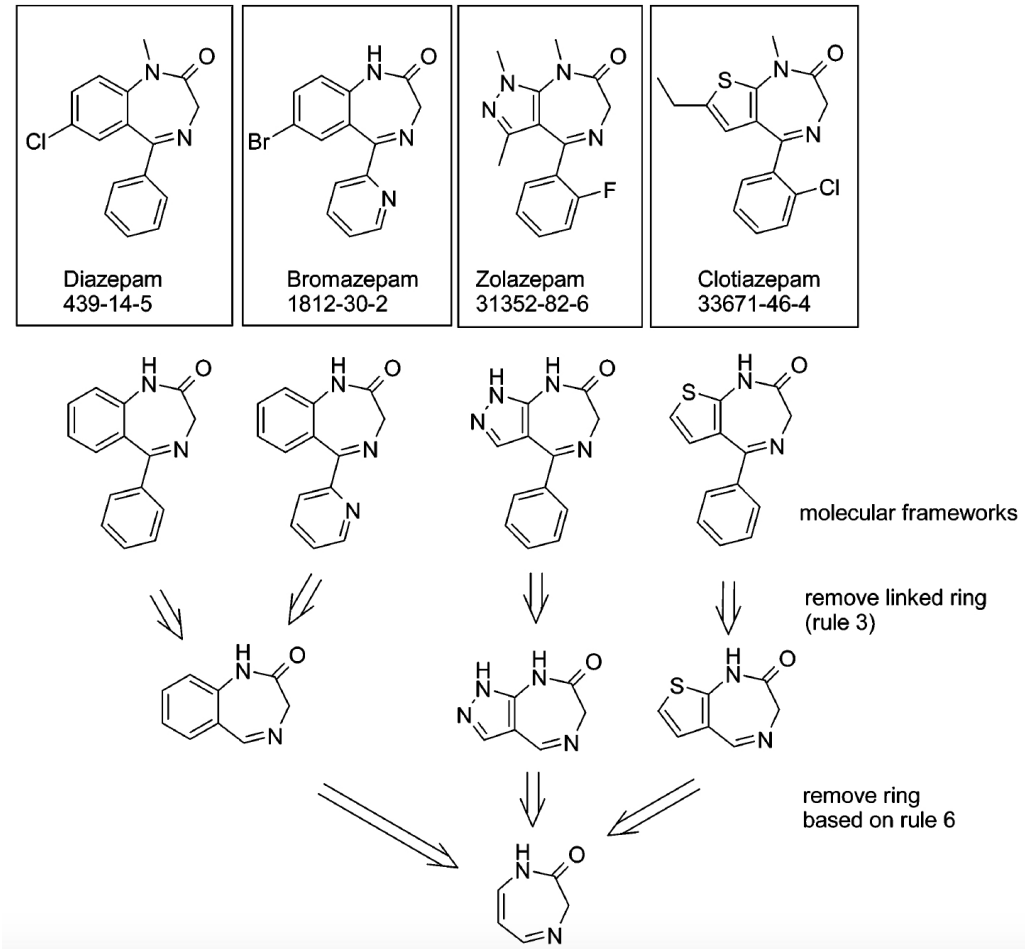


Figure: Classifying anxiolytics (diazepam, bromazepam, zolazepam, and clotiazepam) via the scaffold tree approach. (Scheme 18 in [Sch+07]).

Scaffold Tree Example

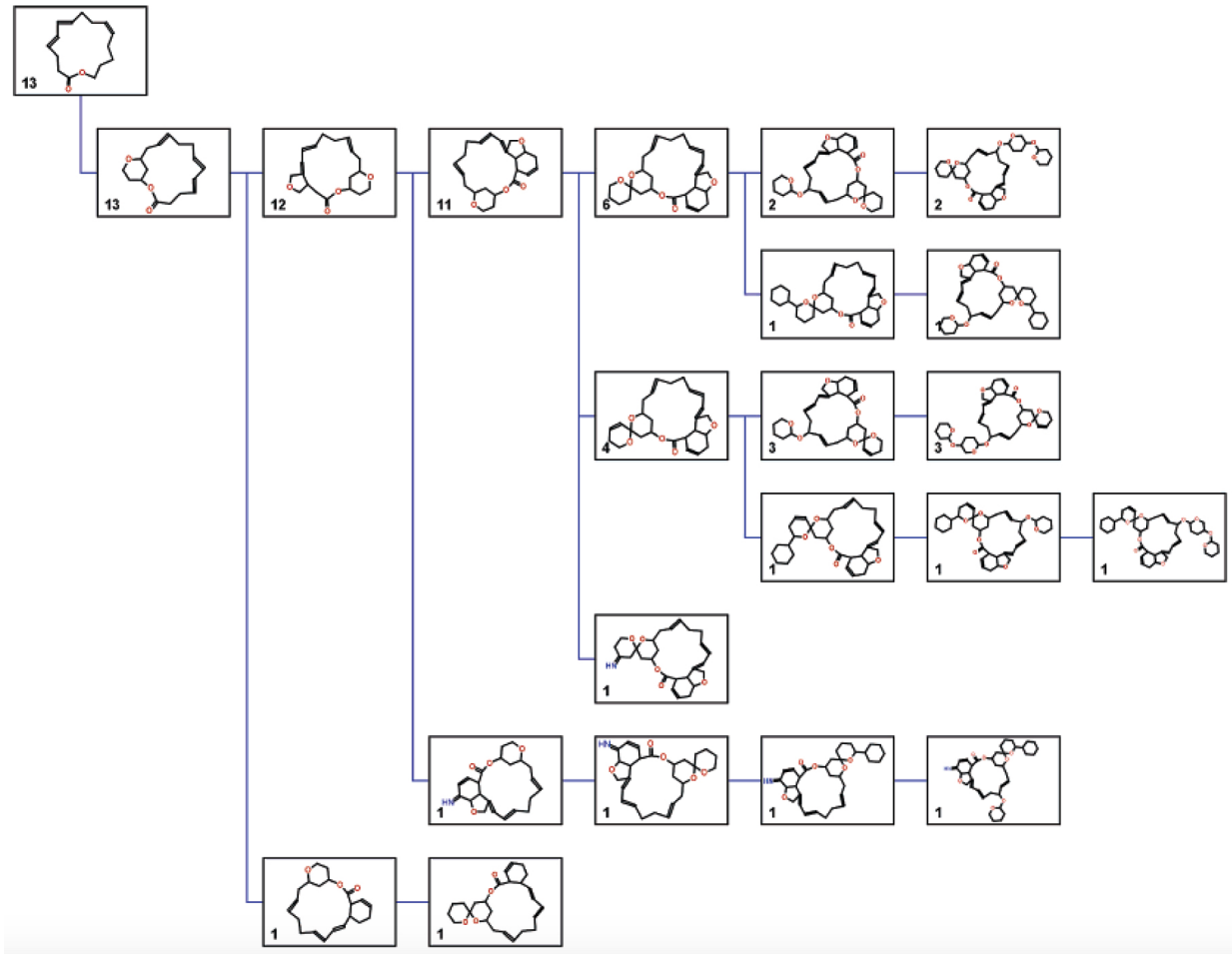


Figure: Scaffold tree for macrocyclic insecticides (Figure 4 in [Sch+07]).