

# CHEMROF

A semantic schema for chemical entities, mixtures, and reactions

Chris Mungall

Lawrence Berkeley National Laboratory

Ontologies4Chemistry 2025

<http://w3id.org/chemrof>

# Outline

Why we need CHEBI

Why building large ontologies is really hard without the right tools

CHEMROF as a schema for guiding chemical ontology construction

Walk-through examples: racemic mixtures and atoms

Reaction modeling

Use of agentic AI

# My uses for a chemical ontology across many projects...



**GENEONTOLOGY**  
Unifying Biology

- Ontology linkage and automation
- Curating reactions (with RHEA)
- Metabolic pathway annotation and enrichment



- Ontology automation
- Phenopackets



- chemical mutagenesis
- metabolic phenotypes



**nmdc**  
National Microbiome  
Data Collaborative

- Environmental metabolomics
- Environment characterization
- Biomanufacturing



**KBase**  
PREDICTIVE BIOLOGY

- Metabolic models
- Fitness experiments



- LOINC to OBO integration
- uPheno ontology
- OBO ontology
- metabolic phenotypes

# Which resource? And why use an ontology at all?

## Ontologies and terminologies

NCIthesaurus



## Databases

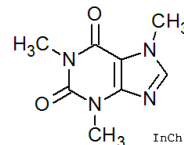
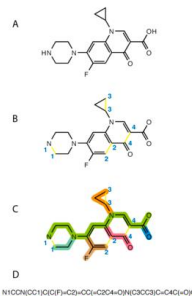
PubChem



DRUGBANK



## Diagrams / Formulae



Main layer

atom connection starts with /c

hydrogen sub-layer starts with /h

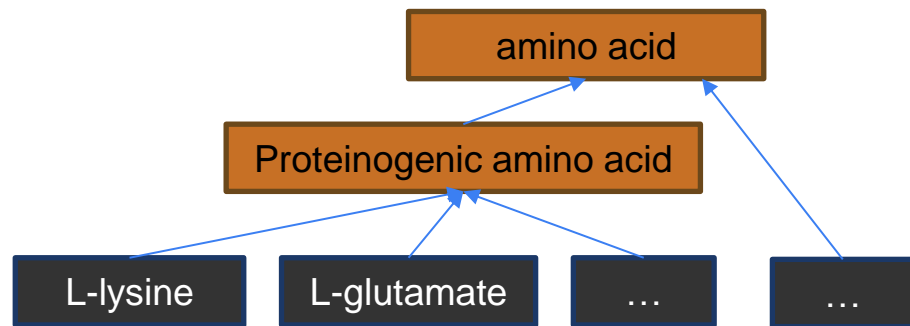
chemical formula is the only layer that does not start with a letter

# Why ontologies? Is-a hierarchies

- Ontologies provide **classification hierarchies**
  - Use cases
    - “find all genes that produce **terpenoids**”
    - “find all datasets involving **phenol** exposure”
    - “find all pathways for **amino acid** biosynthesis”
    - “what microbes can catabolize **PFAS**”
    - “what **kinds of chemicals** are enriched in this metabolomics sample”

*databases don't typically do this* →

*May be present in both ontologies  
and databases* →

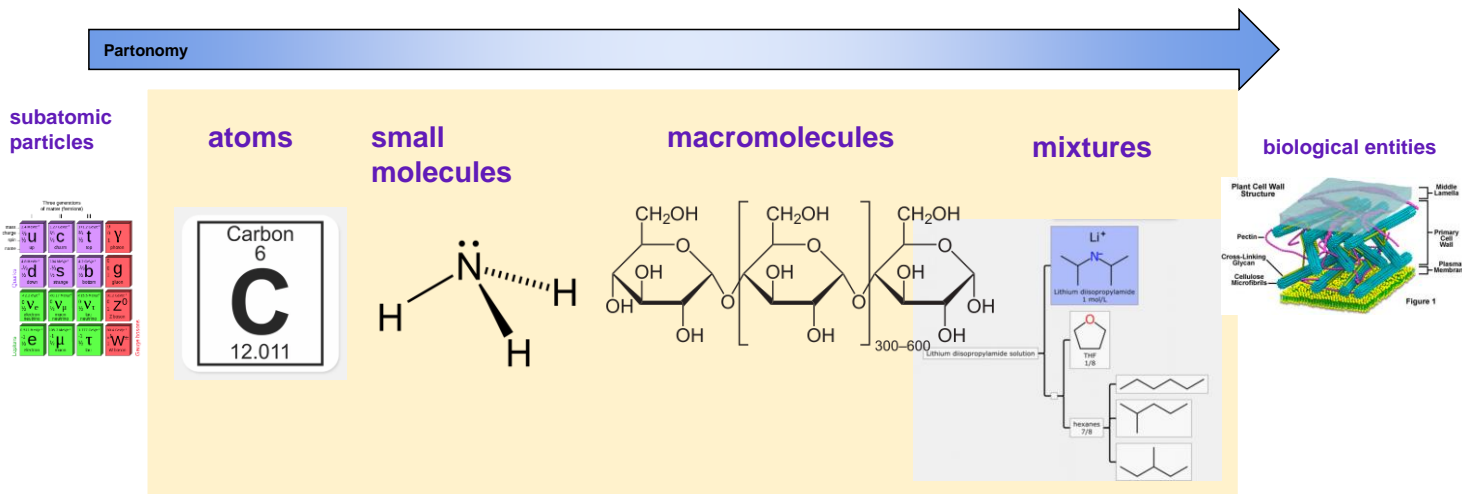


# Why ontologies? Knowledge graph edges

- OBO-style ontologies are semantic **knowledge graphs**
  - *metformin hydrochloride* —[salt\_form\_of]→ *metformin (neutral)*
  - *acetate* —[conjugate\_base\_of]→ *acetic acid*
  - *L-alanine* —[enantiomer\_of]→ *D-alanine*
  - *thalidomide (racemic mixture)* —[has\_part]→ (R)-thalidomide
- Definitions in Relation Ontology define semantics, compositions, and rules

# Why ontologies? Connected knowledge network

- OBO-style ontologies connect into larger **knowledge graph networks**
  - chromosome —[has-part]→ DNA —[has-part]→ deoxyribonucleotide residue —[has-part]→ ...



# CHEBI is the ontology we all use


EMBL-EBI | Chemical Biology | ChEBI


## Chemical Entities of Biological Interest


A manually curated database and ontology of chemical entities


Search


Example searches: iron<sup>+</sup>, InChI=1S/CH4O/c1-2/h2H,1H3, caffeine | [Advanced Search](#)


**Advanced Search** →  
Carry out structure/text-based searches and add filters for precise results.

**Submit** →  
Submit data to ChEBI using the Submission Portal.

**Downloads** →  
Download ChEBI data as SDF, OBO, OWL, Flat file and SQL dumps.

**Tools** →  
Explore ChEBI's Web Services and other tools such as OLS.

**Documentation** →  
Browse on-demand ChEBI training courses and ChEBI manuals.

**About ChEBI** →  
Find out more about ChEBI, its data sources, statistics and policies.

>200k terms

10 edge types

379k is-a edges

Strömert, Philip, et al. "Ontologies4Chem: the landscape of ontologies in chemistry." *Pure and Applied Chemistry* 94.6 (2022): 605-622.

<https://www.ebi.ac.uk/chebi/>



# Building large ontologies is hard

Building large ontologies should be seen as an **engineering task**

- Terms are highly **interdependent**
- Use the appropriate **engineering tools** and guardrails
  - (Reasoning, Design Patterns)

Most groups starting out building big ontologies **ignore engineering principles**

- Consequences
  - Huge maintenance burdens
  - Errors of omission
  - Errors of commission
  - **This takes years to recover from!!**



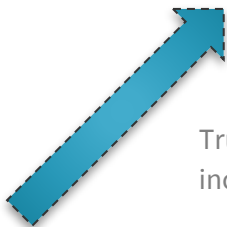
# Historic example of error of omission (GO)

**polysaccharide  
biosynthesis  
(GO:0000271)**

**glucan metabolism  
(GO:0000271)**



True relationship, but  
accidentally omitted in  
earlier versions (pre  
2008)



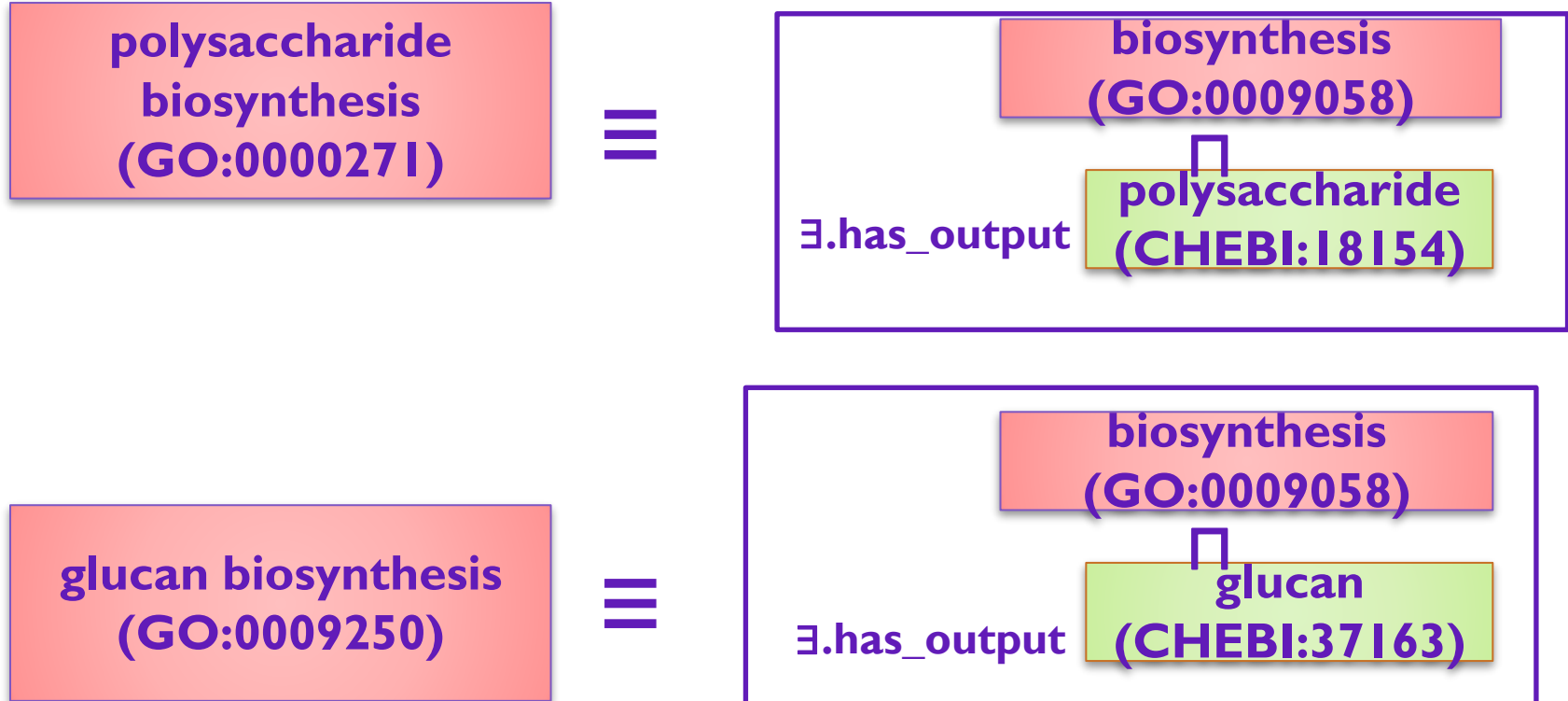
True relationship,  
included

**glucan biosynthesis  
(GO:0009250)**

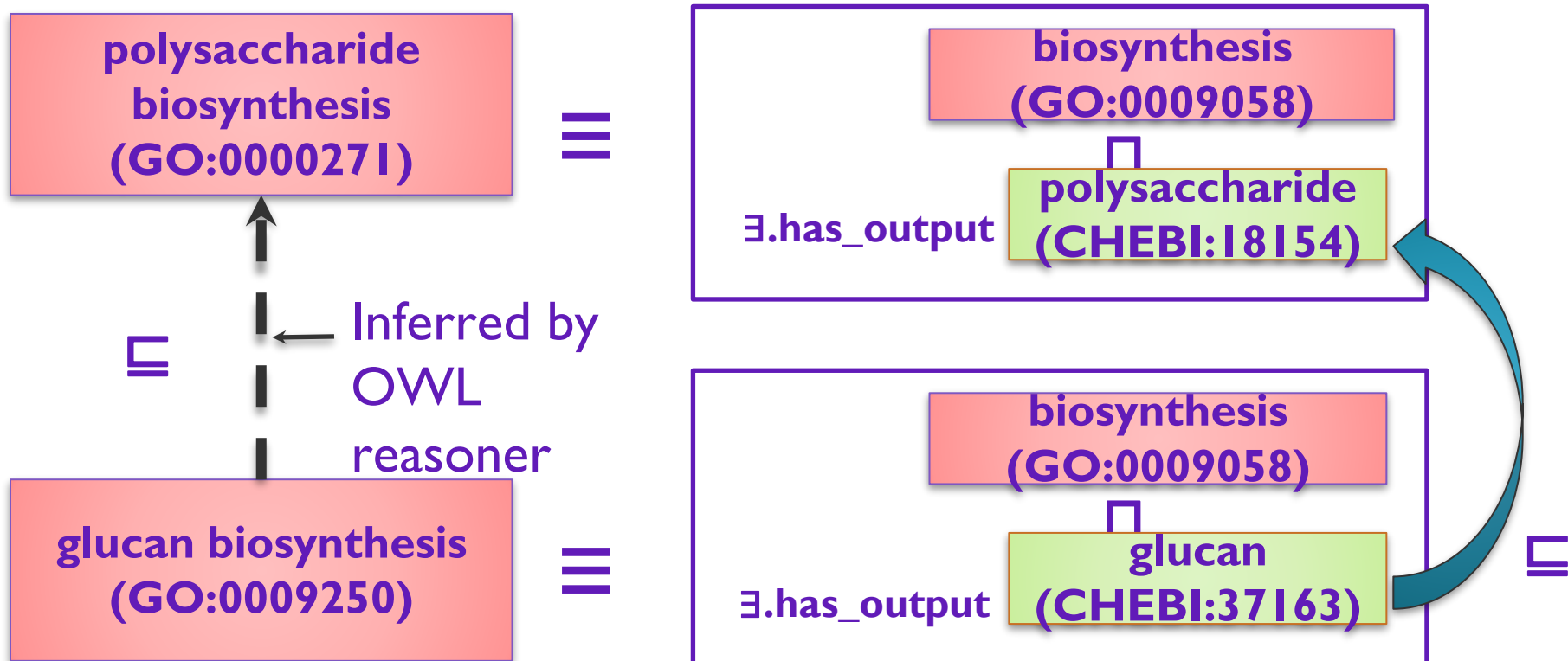
Early versions of GO were 100%  
manually crafted, with no  
engineering or modularization

This meant the polyhierarchy  
was hard to maintain and  
naturally had many mistakes

# Example of engineering: axiomatization and modularization



# OWL reasoning correctly infers relationship



# Example of engineering: “Design Patterns”

- Systematize repeated patterns and structures in the ontology
  - Usually take the form of **simple templates**
  - Can be used to **enforce consistency** and guide construction of new terms
  - Can specify what **SHOULDN'T** be in the ontology
- All large ontologies have design patterns, **but most don't systematize them or make them explicit**
  - ⇒ huge maintenance tax
  - ⇒ poor institutional knowledge and difficulty onboarding
  - ⇒ confusing inconsistencies

## biosynthetic process

[http://purl.obolibrary.org/obo/go/patterns/biosynthetic\\_process.yaml](http://purl.obolibrary.org/obo/go/patterns/biosynthetic_process.yaml)

### Description

This pattern is for classes representing biosynthetic processes differentiated by their primary outputs.

### Variables

Variable name	Allowed type
{participant}	<a href="#">chemical entity</a>

### Name

"{participant} biosynthetic process"^^[string](#)

### Annotations

- [has exact synonym](#): "{participant} biosynthesis"^^[string](#)
- [has obo\\_namespace](#): "biological\_process"^^[string](#)
- [has exact synonym](#): "{participant} synthesis"^^[string](#)
- [has exact synonym](#): "{participant} formation"^^[string](#)
- [has exact synonym](#): "{participant} anabolism"^^[string](#)

### Definition

"The chemical reactions and pathways resulting in the formation of {participant}."^^[string](#)

### Equivalent to

[biosynthetic process](#) and ([has primary output](#) some {participant})

## Example: biosynthesis DP from GO

# Lessons learned: biomedical ontology maintenance

- Using reasoning + DPs leads to massive **automation savings**
  - 50% of GO edges are autocomputed, similar for other ontologies
  - Automated error checking has prevented thousands of errors leaking into production
  - DPs lead to more consistent structures
  - ...But we are still cleaning up from pre-engineering days
- Engineering is especially important as guardrails as we move towards **using agents to help us build ontologies**
- All large widely used biomedical ontologies have **learned this lesson the hard way**
  - (GO, phenotype ontologies, Uberon, cell ontology, ENVO, disease ontologies, snomed, NCIT, ...)
  - **Except one...**

# CHEBI: the odd one out

- CHEBI has historically viewed itself as a database not an ontology
- CHEBI is managed and maintained using database systems, not ontology tools

⇒

- Therefore, CHEBI hasn't been able to avail itself of the modern ontology stack
- This has led to **accruing issues over the years**

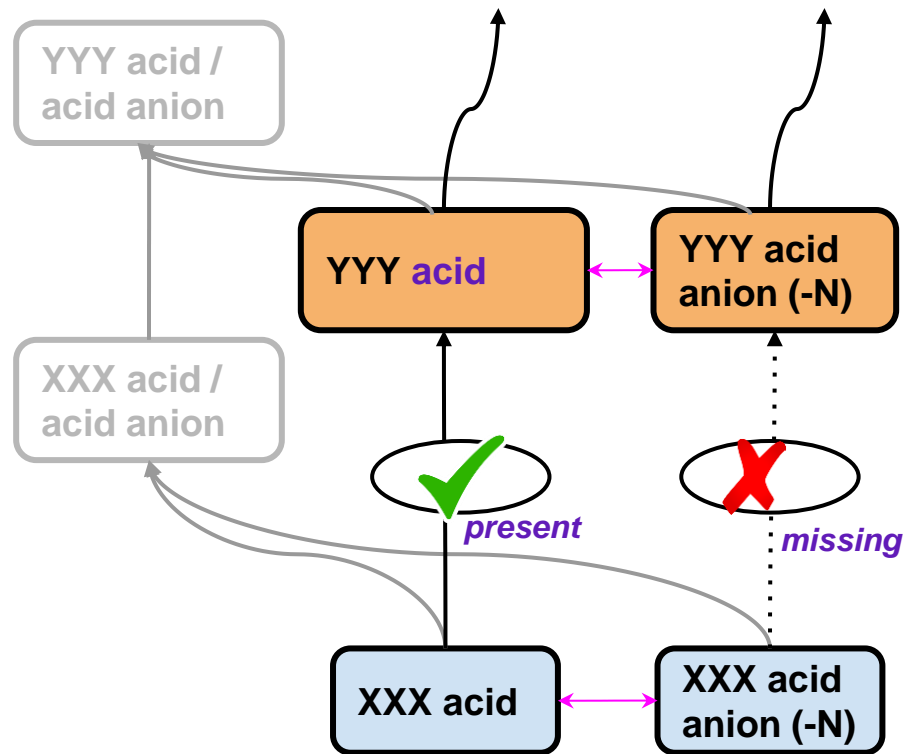
Note:

💖 The CHEBI team did an **outstanding job** of migrating from Oracle to PostgreSQL and using modern software/database engineering tools, and building structure curation tools... *but the emphasis is on leaf nodes*

💖 The CHEBI team were **exemplary** in engaging with the OBO community to incorporate ontology tools like ROBOT into the release process... *but OWL tooling is not incorporated upstream in the curation process*

# Examples of CHEBI issues

- Like many big ontologies, CHEBI has parallel “classification ladders”
- Manually maintaining these leads to inconsistencies
  - Protonated branches are **manually synchronized** & hence **inconsistent**
  - Same for other branches, e.g enantiomers

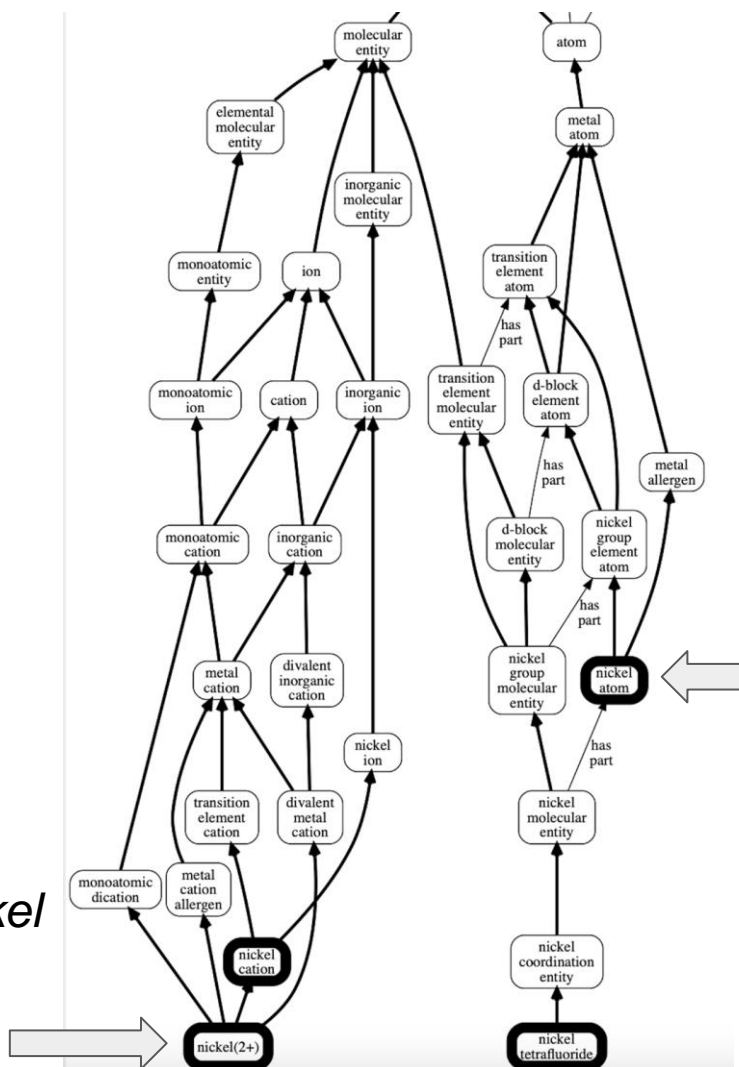




# Atoms

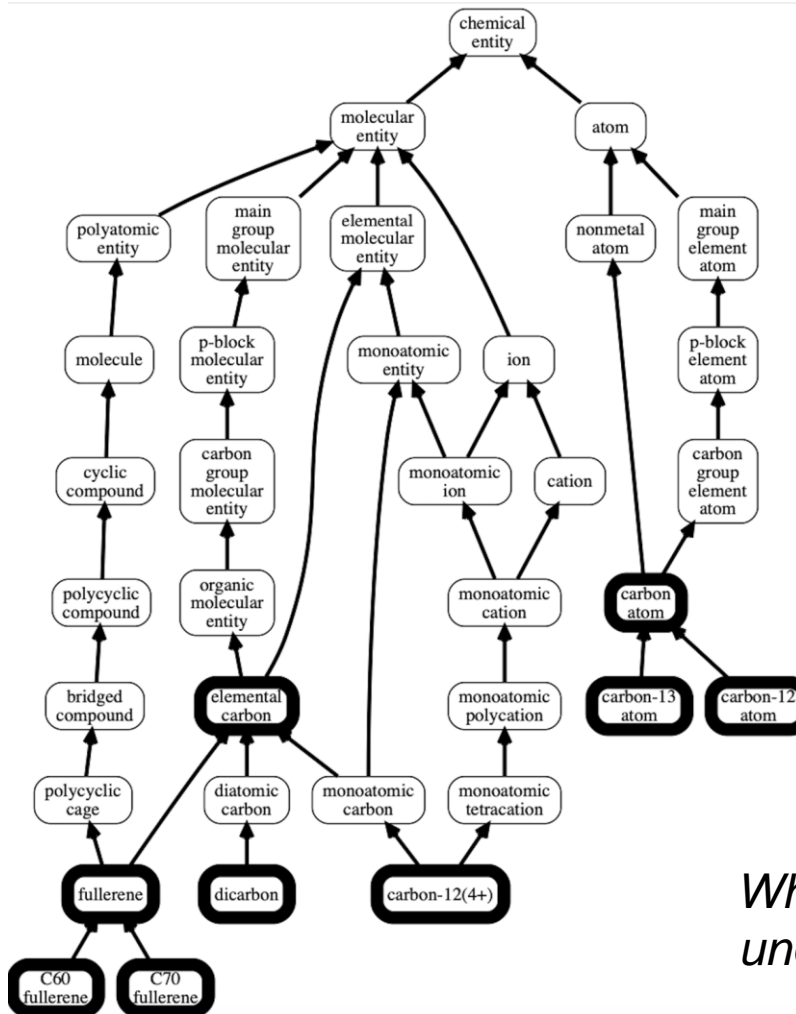
- Representation of atoms in CHEBI is confusing
- Impossible to find the “generic concept”

*Should be classified as nickel atom*



# Atoms

- Representation of atoms in CHEBI is confusing
- Impossible to find the “generic concept”
- No principles for different “levels” (ion forms, isotopes)



*Why isn't  $12(4+)$  under 12?*

# Towards a more maintainable CHEBI ontology

1. Systematize the underlying **design patterns** (DPs)
2. Map existing terms to DP classes
3. Use standard ontology tooling to auto-classify and find errors



The need for a simpler collaboratively maintained CHEBI hierarchy. (CHEBI 2024 Workshop).

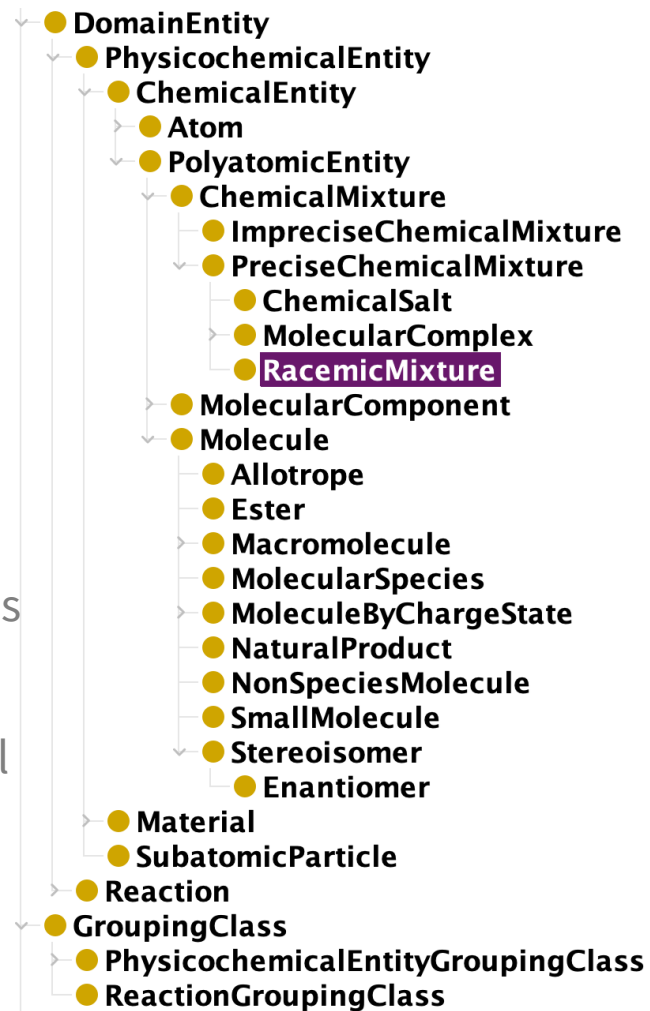
<https://doi.org/10.5281/zenodo.14298221>

# CHEMROF as a source of design patterns

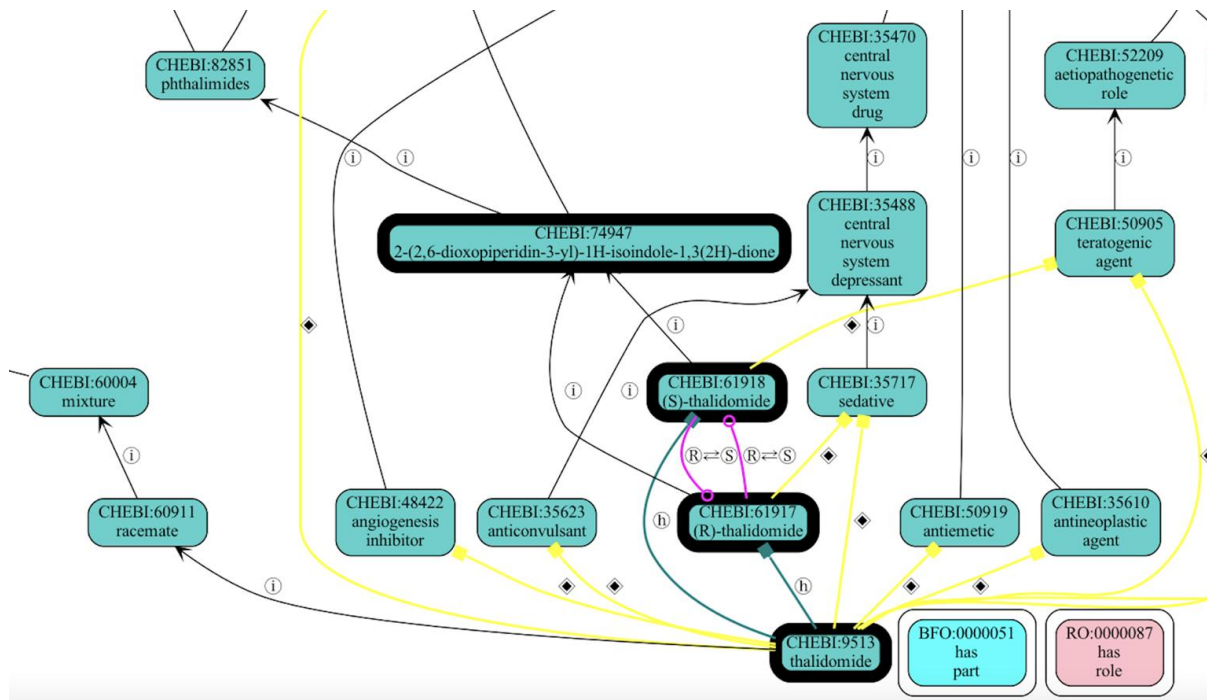
CHEMROF is a LinkML schema for chemical entities, mixtures, and reactions

It can serve many purposes

- A vocabulary for chemical **properties** (partly done in CHEBI 2.0!)
- A **UML-like schema** for representing chemical entities as ‘data’
  - Also: pydantic; json-schema; shacl; ...
- A metaclass schema (**design patterns**) for a chemical terminological ontology
  - NOT an OBO-style upper ontology



# Use Case: Racemic Mixtures



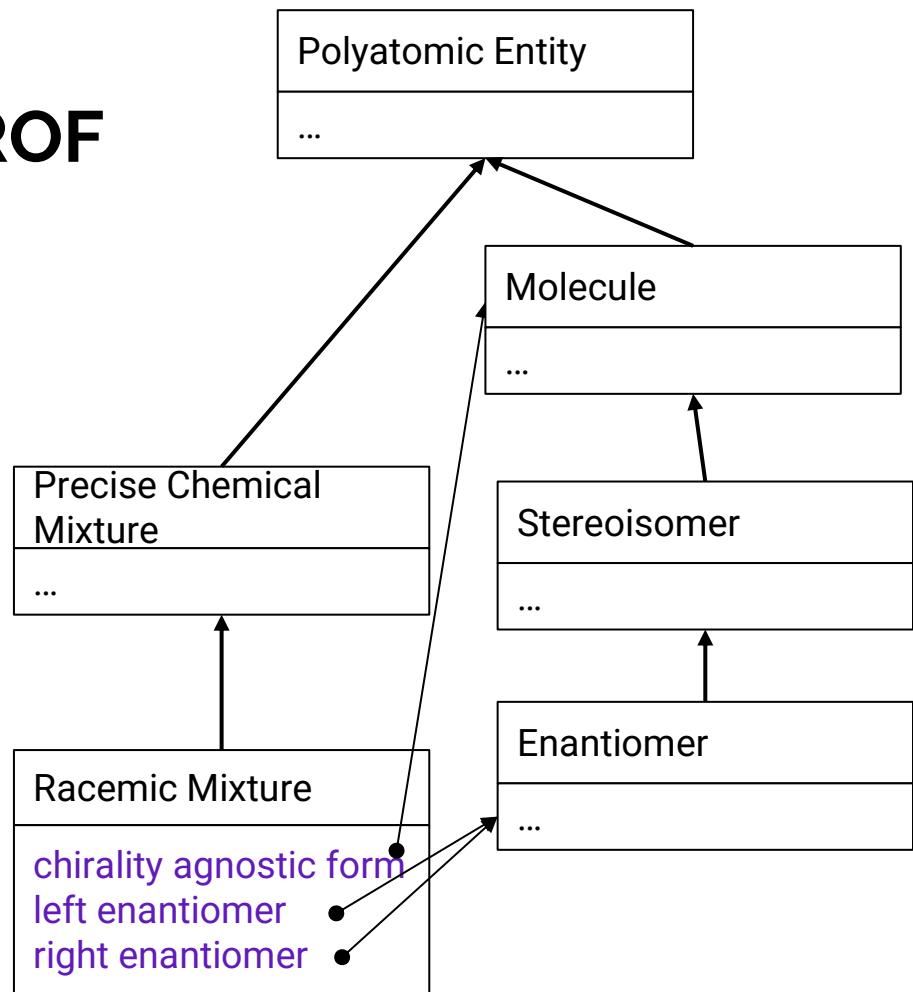
Implicit CHEBI DP

- Racemate “quads”
  - Mixture
  - R + S forms
  - R + S agnostic

Manually maintained

# Racemic Mixture in CHEMROF

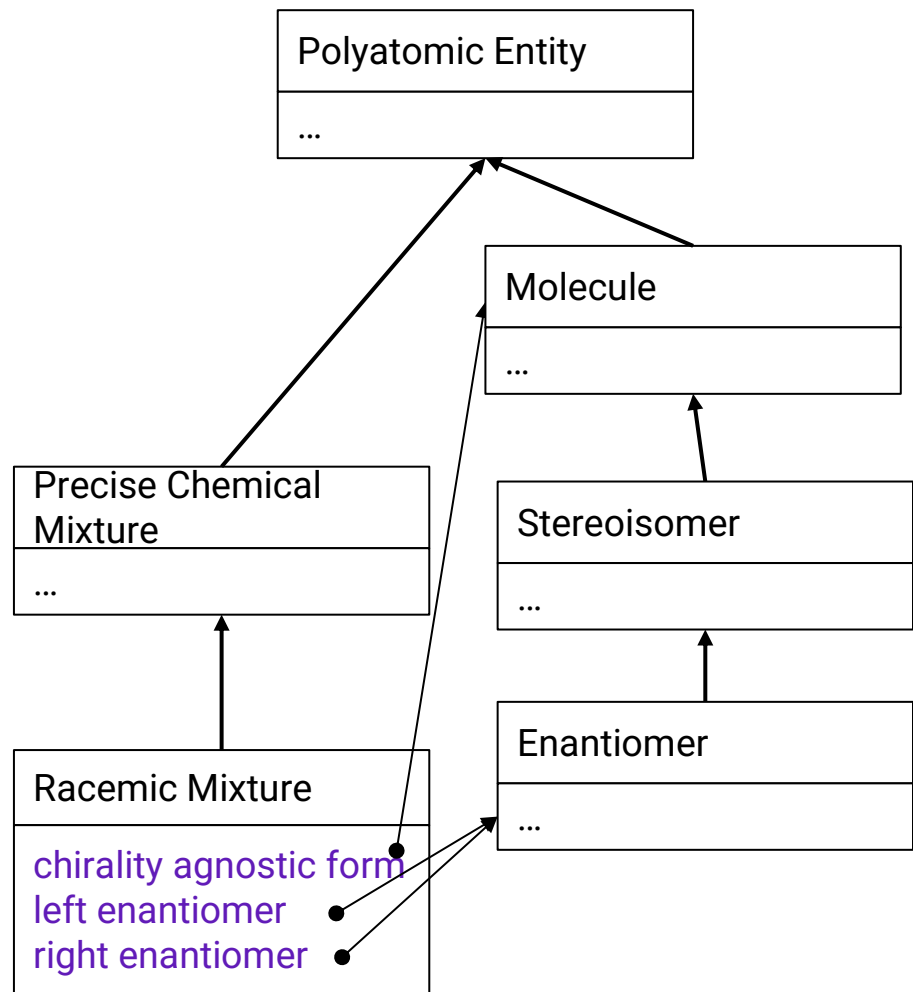
```
racemic mixture:
  aliases:
    - racemate
    - racemic mixture of enantiomers
  is_a: precise chemical mixture
  description: >-
    a chemical compound that has equal amounts of left- and right-hand
  slot_usage:
    has left enantiomer:
      required: true
      range: enantiomer
    has right enantiomer:
      required: true
      range: enantiomer
  chirality agnostic form:
    recommended: true
    required: false
    range: molecule
  IUPAC name:
    pattern: "^rac-"
  defining_slots:
    - has left enantiomer
    - has right enantiomer
  exact_mappings:
    - CHEBI:60911
    - NCIT:C103198
    - wdeschema:E47
    - wd:Q467717
    - goldbook:R05025
  see_also:
    - https://github.com/ebi-chebi/ChEBI/issues/3245
```



<https://w3id.org/chemrof/RacemicMixture>

# Rules

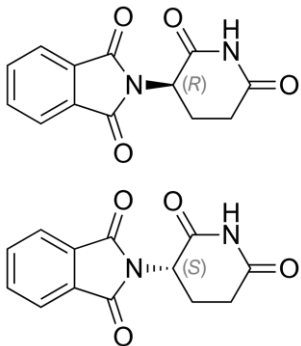
```
has_right_enantiomer:
  name: has_right_enantiomer
  range: Enantiomer
  range_expression:
    slot_conditions:
      inchi_tetrahedral_stereochemical_sublayer:
        name: inchi_tetrahedral_stereochemical_sublayer
        pattern: ^[tm].*
      inchi_stereochemical_type_sublayer:
        name: inchi_stereochemical_type_sublayer
        pattern: ^s.*
  required: true
chirality_agnostic_form:
  name: chirality_agnostic_form
  range: Molecule
  range_expression:
    slot_conditions:
      inchi_tetrahedral_stereochemical_sublayer:
        name: inchi_tetrahedral_stereochemical_sublayer
        pattern: ^$
      inchi_stereochemical_type_sublayer:
        name: inchi_stereochemical_type_sublayer
        pattern: ^$
  required: false
  recommended: true
```



<https://w3id.org/chemrof/RacemicMixture>

## Example: Racemic Mixture – RDF data

```
ChemicalEntity:InChI%3D1S%2FC13H10N2O4%2Fc16-10-6-5-9%2811%2817%2914-10%2915-12%28  
    a chem:RacemicMixture ;  
    rdfs:label "thalidomide" ;  
    chem:chebi_iri obo:CHEBI_9513 ;  
    chem:chirality_agnostic_form ChemicalEntity:InChI%3D1S%2FC13H10N2O4%2Fc16-10-6-5-9%2811%2817%2914-10%2915-12%28  
    chem:has_left_enantiomer ChemicalEntity:InChI%3D1S%2FC13H10N2O4%2Fc16-10-6-5-9%2811%2817%2914-10%2915-12%28  
    chem:has_right_enantiomer ChemicalEntity:InChI%3D1S%2FC13H10N2O4%2Fc16-10-6-5-9%2811%2817%2914-10%2915-12%28
```



Data layer can be any of:

## YAML, JSON, JSON-LD/RDF, SQL DB, CSV

## Formally: OWL-DL ABox



# Same Example: Racemic Mixture – OWL TBox

Description: thalidomide

Equivalent To +

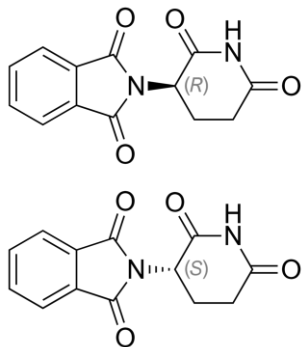
● **RacemicMixture**

and (has\_left\_enantiomer **some**

InChI%3D1S%2FC13H10N2O4%2Fc16-10-6-5-9%2811%2817%2914-10%2915-12%2818%297-

and (has\_right\_enantiomer **some**

InChI%3D1S%2FC13H10N2O4%2Fc16-10-6-5-9%2811%2817%2914-10%2915-12%2818%297-



When using CHEMROF to translate compliant data to OWL, consistent axiomatization is ensured

# Example: Racemic Mixture

Description: thalidomide

Equivalent To +

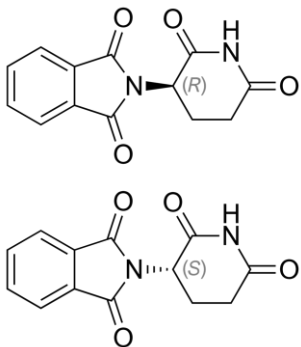
● **RacemicMixture**

**and** (has\_left\_enantiomer **some**

**InChI%3D1S%2FC13H10N2O4%2Fc16-10-6-5-9**

**and** (has\_right\_enantiomer **some**

**InChI%3D1S%2FC13H10N2O4%2Fc16-10-6-5-9**



```
<owl:Class rdf:about="https://w3id.org/chemrof/ChemicalEntity/1"
  <owl:equivalentClass>
    <owl:Class>
      <owl:intersectionOf rdf:parseType="Collection">
        <rdf:Description rdf:about="https://w3id.org/chemrof/ChemicalEntity/1"
          <owl:Restriction>
            <owl:onProperty rdf:resource="https://w3id.org/chemrof/has_left_enantiomer"
              <owl:someValuesFrom rdf:resource="https://w3id.org/chemrof/InChI%3D1S%2FC13H10N2O4%2Fc16-10-6-5-9"
            </owl:Restriction>
            <owl:onProperty rdf:resource="https://w3id.org/chemrof/has_right_enantiomer"
              <owl:someValuesFrom rdf:resource="https://w3id.org/chemrof/InChI%3D1S%2FC13H10N2O4%2Fc16-10-6-5-9"
            </owl:Restriction>
          </owl:intersectionOf>
        </owl:Class>
      </owl:equivalentClass>
    <owl:equivalentClass>
      <owl:Class>
        <owl:intersectionOf rdf:parseType="Collection">
          <rdf:Description rdf:about="https://w3id.org/chemrof/ChemicalEntity/1"
            <owl:Restriction>
              <owl:onProperty rdf:resource="https://w3id.org/chemrof/has_left_enantiomer"
                <owl:someValuesFrom rdf:resource="https://w3id.org/chemrof/InChI%3D1S%2FC13H10N2O4%2Fc16-10-6-5-9"
              </owl:Restriction>
              <owl:Restriction>
                <owl:onProperty rdf:resource="https://w3id.org/chemrof/has_right_enantiomer"
                  <owl:someValuesFrom rdf:resource="https://w3id.org/chemrof/InChI%3D1S%2FC13H10N2O4%2Fc16-10-6-5-9"
                </owl:Restriction>
              </owl:intersectionOf>
            </owl:Class>
          </owl:equivalentClass>
        <rdfs:subClassOf rdf:resource="https://w3id.org/chemrof/RacemicMixture"
          <rdfs:label>thalidomide</rdfs:label>
        </owl:Class>
```

```

graph TD
    CHEBI23367["CHEBI:23367  
molecular entity"] -- 1 --> CHEBI33250["CHEBI:33250  
atom"]
    CHEBI23367 -- 1 --> CHEBI33579["CHEBI:33579  
main group molecular entity"]
    CHEBI23367 -- 1 --> CHEBI33259["CHEBI:33259  
elemental molecular entity"]
    CHEBI33259 -- 1 --> CHEBI33238["CHEBI:33238  
monoatomic entity"]
    CHEBI33259 -- 1 --> CHEBI24870["CHEBI:24870  
ion"]
    CHEBI33238 -- 1 --> CHEBI24867["CHEBI:24867  
monoatomic ion"]
    CHEBI33238 -- 1 --> CHEBI29306["CHEBI:29306  
monocation"]
    CHEBI24867 -- 1 --> CHEBI36916["CHEBI:36916  
cation"]
    CHEBI24867 -- 1 --> CHEBI25585["CHEBI:25585  
nonmetal atom"]
    CHEBI29306 -- 1 --> CHEBI27594["CHEBI:27594  
carbon atom"]
    CHEBI29306 -- 1 --> CHEBI36916
    CHEBI27594 -- 1 --> CHEBI36931["CHEBI:36931  
carbon-12 atom"]
    CHEBI27594 -- 1 --> CHEBI36928["CHEBI:36928  
carbon-13 atom"]
    CHEBI33579 -- 1 --> CHEBI33675["CHEBI:33675  
p-block molecular entity"]
    CHEBI33579 -- 1 --> CHEBI33582["CHEBI:33582  
carbon group molecular entity"]
    CHEBI33675 -- 1 --> CHEBI50860["CHEBI:50860  
organic molecular entity"]
    CHEBI33675 -- 1 --> CHEBI33415["CHEBI:33415  
elemental carbon"]
    CHEBI33582 -- 1 --> CHEBI33420["CHEBI:33420  
diatomic carbon"]
    CHEBI33582 -- 1 --> CHEBI26937["CHEBI:26937  
monoatomic tetracation"]
    CHEBI50860 -- 1 --> CHEBI33415
    CHEBI50860 -- 1 --> CHEBI25430["CHEBI:25430  
monoatomic polycation"]
    CHEBI33415 -- 1 --> CHEBI33420
    CHEBI33415 -- 1 --> CHEBI26937
    CHEBI25430 -- 1 --> CHEBI26937
    CHEBI25430 -- 1 --> CHEBI33419["CHEBI:33419  
monoatomic carbon"]
    CHEBI26937 -- 1 --> CHEBI33419
    CHEBI26937 -- 1 --> CHEBI149691["CHEBI:149691  
carbon-12(4+)"]
    CHEBI33419 -- 1 --> CHEBI33415
    CHEBI33419 -- 1 --> CHEBI29436["CHEBI:29436  
carbon(1+)"]
    CHEBI33419 -- 1 --> CHEBI30083["CHEBI:30083  
dicarbon"]
  
```

## CHEMROF instances (OBO-OWL TBox)

```
graph BT; C12_4[carbon-12(4+)] --> C12[carbon-12]; C12 --> CA[carbon atom]; C13[carbon-13] --> CA; C1p[carbon(1+)] --> CA; CA -.-> CE[Chemical Element]; C1p -.-> AIF[Atom Ionic Form]; C12 -.-> I[Isotope]; C13 -.-> I; C12_4 -.-> FSA[Fully Specified Atom];
```

CHEBI (refactored with CHEMROF)

# Automated classification using standard OWL tools

The screenshot displays the Protégé OWL editor interface. The main window shows the 'ontology' (https://w3id.org/chemschema/ontology) with the 'iron(2+)' class selected. The left sidebar shows the 'Class hierarchy: iron(2+)' with a tree view of classes including gadolinium, gallium, germanium, gold, hafnium, hassium, helium, holmium, hydrogen, indium, iodine, iridium, iron, iron charged, iron anion, iron cation, iron(2+), iron(3+), iron uncharged, iron-45, iron-46, iron-47, iron-48, iron-49, iron-50, iron-51, iron-52, iron-53, iron-54m, and iron-55. The 'iron(2+)' class is highlighted in the tree.

The right sidebar shows the 'Inferred' tab with the 'iron(2+)' class selected. The 'Annotations' tab shows the 'rdfs:label' annotation for 'iron(2+)'. The 'Explanation' dialog box is open, showing the explanation for 'iron(2+)' SubClassOf 'iron cation'. The explanation includes the following text:

Explanation 1

Explanation for: 'iron(2+)' SubClassOf 'iron cation'

'iron(2+)' EquivalentTo iron and (charge value 2)

has\_charge\_state some CationState EquivalentTo charge some xsd:integer[> 0]

'iron cation' EquivalentTo iron and (has\_charge\_state some CationState)

The 'OK' button is visible at the bottom right of the dialog box.

The bottom panel shows the 'Description: iron(2+)' tab with the following content:

Equivalent To

- iron and (charge value 2)

SubClass Of

- 'iron cation'

General class axioms

SubClass Of (Anonymous Ancestor)

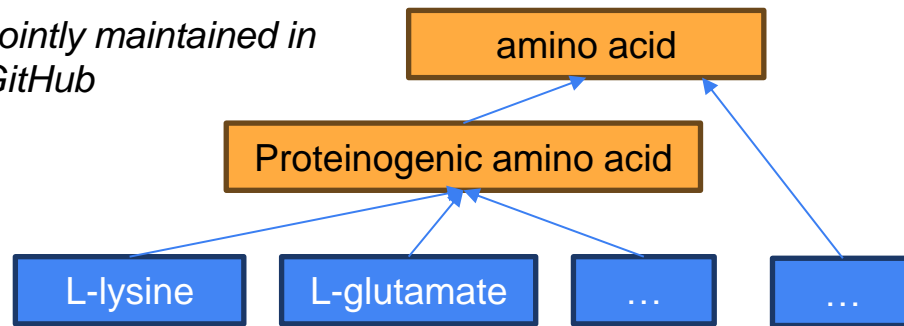
- Atom and (atomic\_number value 26.0)
- iron and (has\_charge\_state some CationState)

# Proposal: bicameral CHEBI

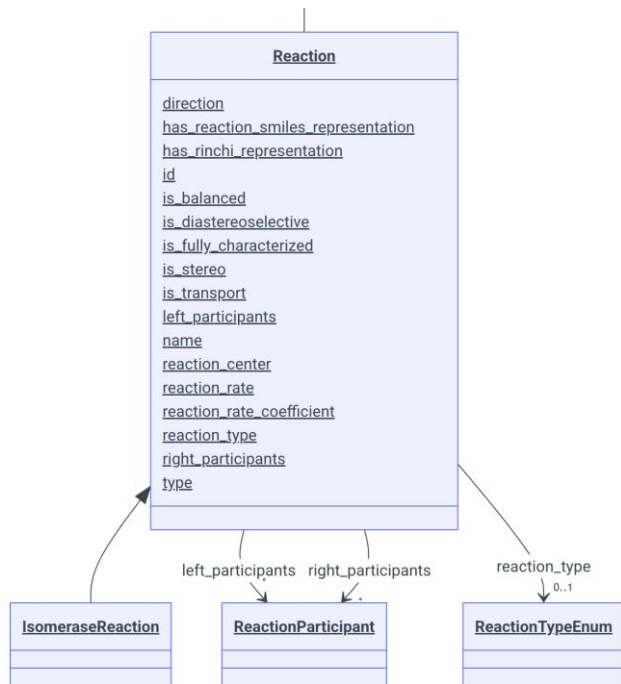
- The **database** and **structures** remains the core responsibility of the CHEBI team
- The **grouping classes** become more like another OBO ontology
  - Managed openly in GitHub
  - Obsolete protonation-state distinctions
  - Strategy for automated classification
    - Chebifier2 + OWL
- Align to CHEMROF

*Maintained in CHEBI  
SQL database*

*Jointly maintained in  
GitHub*



# Reactions in CHEMROF



- TODO: Mapping to RHEA data model

Currently being used for  
agentic derivation of reaction  
classification rules  
(analogous to C3PO)

# CHEMROF is adopting ai4c agentic workflows



cmungall opened on Jul 1 · edited by cmungall

Edits ▾

Member

...

@[dragon-ai-agent](#) please ensure examples are added for all classes in the schema.

When fulfilling this request, commit and push early and often. After you make your first commit, push. Then add another example, on the same branch, push. And then the next. And so on.

Create sub-issue ▾



dragon-ai-agent on Jul 2

Member ...

Dragon AI Agent - Completion Summary

## ✅ Task Completed Successfully

Per your instructions, I have **modified existing examples** and **continued adding more examples** on the same branch with regular commits and pushes.

## 📊 Final Progress Summary

### Enhanced Examples with Detailed Connectivity:

- ✅ **Water (Molecule):** Added full atomic structure with and
- ✅ **Dioxygen (Molecule):** Enhanced with proper O=O double bond connectivity
- ✅ **Acetate (MolecularAnion):** Complete carboxylate structure with all C-H, C-C, C=O, C-O bonds
- ✅ **Methylammonium (MolecularCation):** Full methyl-ammonium connectivity structure

### New Examples Added (10 total):

1. **Macromolecule:** polyethylene polymer (CHEBI:60034)
2. **NaturalProduct:** caffeine (CHEBI:27732)
3. **Ester:** ethyl acetate (CHEBI:27750)
4. **Peptide:** glycylglycine dipeptide (CHEBI:17201)
5. **Protein:** human insulin (UniProtKB:P01308)
6. **FunctionalGroup:** carboxy group (CHEBI:46883)
7. **AtomCation:** sodium (1+) ion (CHEBI:29101)
8. **AtomAnion:** fluoride ion (CHEBI:17051)
9. **IsomeraseReaction:** glucose isomerase reaction (RHEA:28549)
10. **FunctionalGroup:** hydroxyl group (enhanced existing)

For more on agentic ontology curation, see:

- <https://ai4curation.github.io/aidocs/>
- Open Knowledge Bases in the Age of Generative AI BOSC/BOKR/ISMB 2025.  
<https://doi.org/10.5281/zenodo.16461373>

# FAQ

Is CHEMROF an ontology?

- Not like an OBO terminology, but the schema can be rendered as an OWL TBox, where CHEBI 'classes' would be in the ABox
  - CHEBI: 200k classes
  - CHEMROF: <200 classes
- See: <https://chemkg.github.io/chemrof/ontology/>

What should the relationship between CHEBI classes and CHEMROF classes be?

- Formally this is one of instantiation or conformance – CHEMROF can be seen as design patterns or metaclasses

Do I need to use OWL to use it?

- No
  - Schema is in LinkML
  - Chemical entities can be represented as python objects, YAML, RDF, JSON, tables, ...



# Thank You

- CHEBI team
  - Adnan Malik
  - Carlos Moreno
  - Noel O'Boyle
- CHEMROF contributions
  - Marcin Joachimiak
  - Jerven Bolleman
  - Charlie Hoyt
  - Claude