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…

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

- Ontology automation
- Phenopackets

- chemical mutagenesis
- metabolic phenotypes

- Environmental metabolomics
- Environment characterization
- Biomanufacturing

- Metabolic models
- Fitness experiments

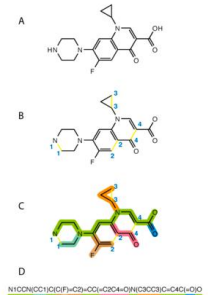
- LOINC to OBO integration
- uPheno ontology
- OBO ontology
- metabolic phenotypes

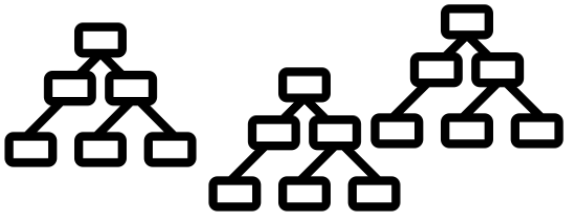
Which resource? And why use an ontology at all?

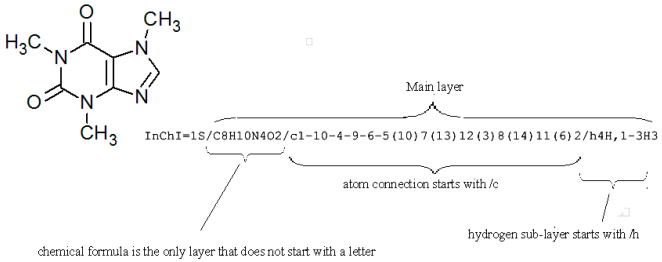
Ontologies and terminologies

Databases

Diagrams / Formulae





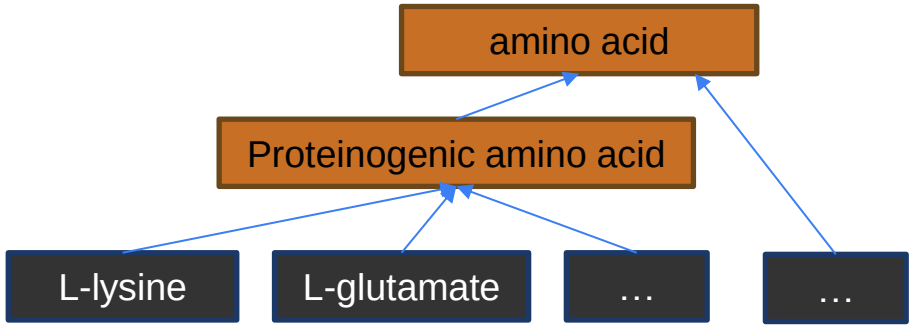


Why ontologies? Is-a hierarchies

- 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



Hastings, J. (2017). Primer on ontologies. *The gene ontology handbook*, *1446*, 3-13.

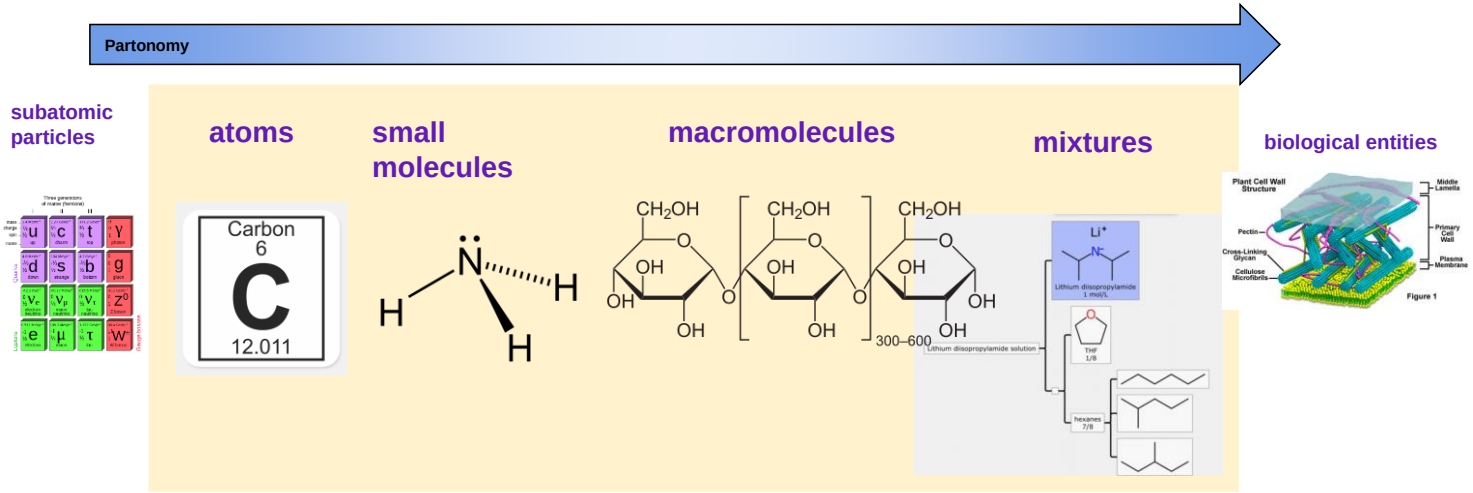
Why ontologies? Knowledge graph edges

- 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

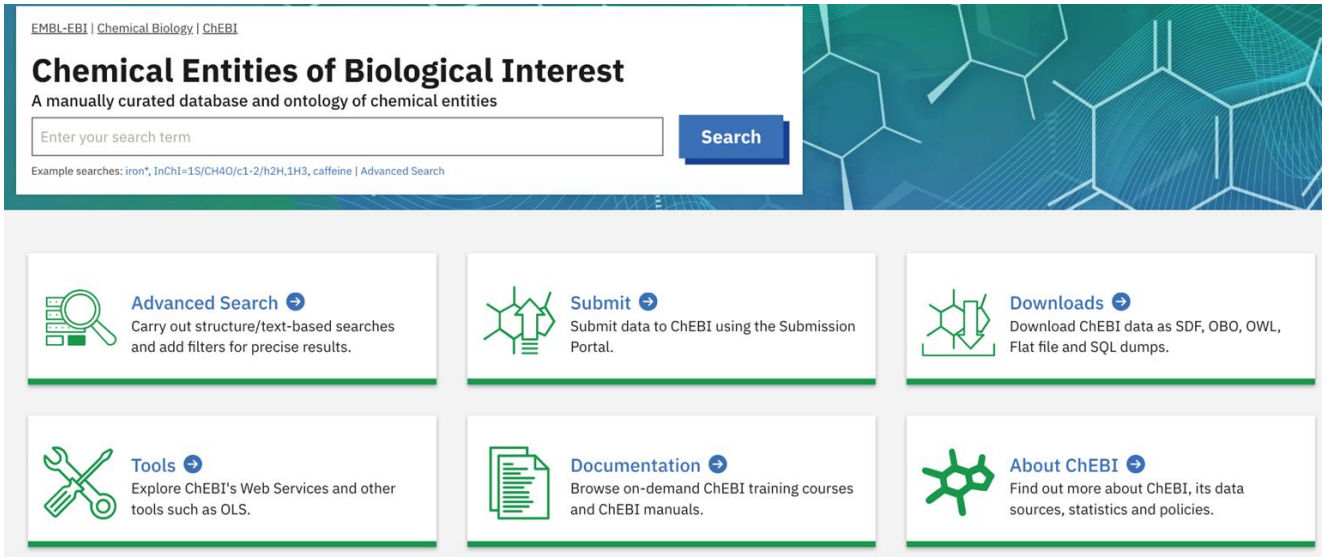
- Definitions in Relation Ontology define semantics, compositions, and rules

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



>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**

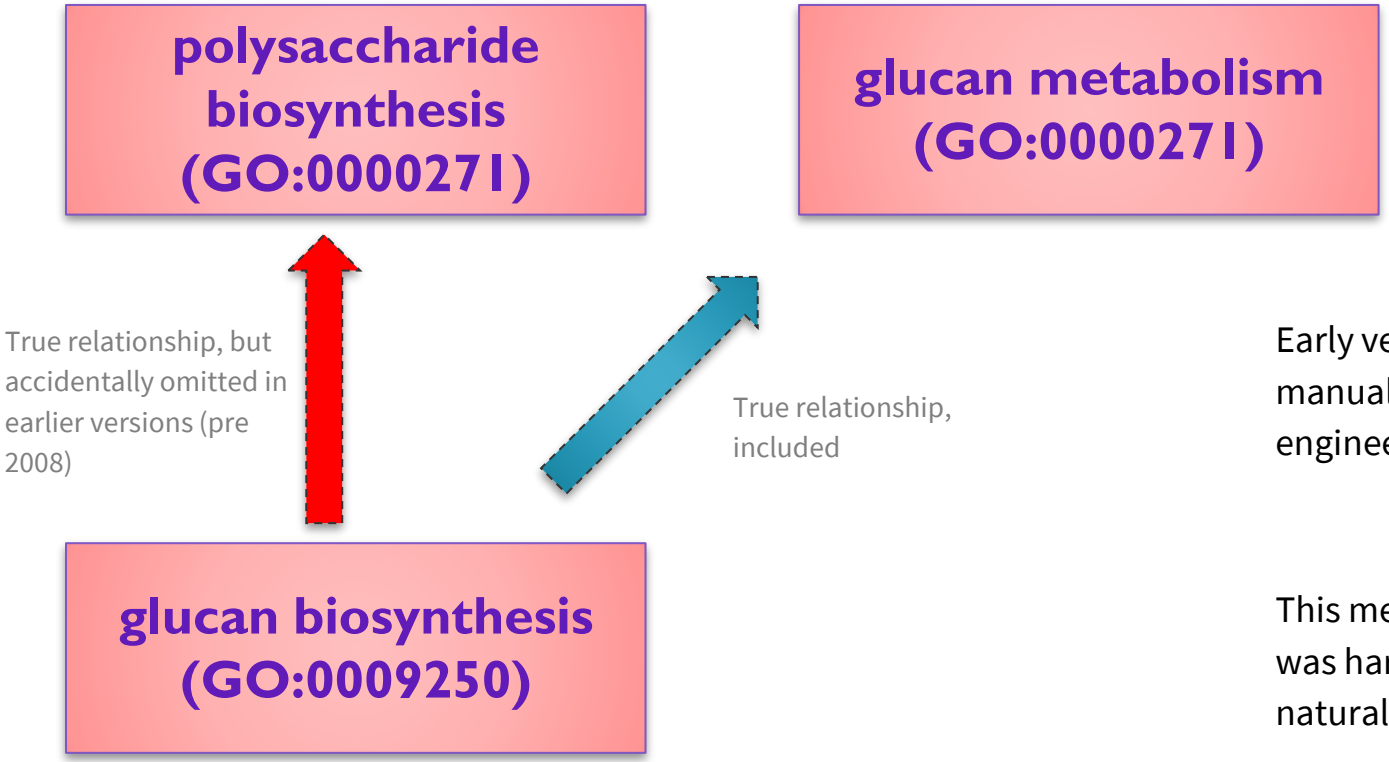
- Terms are highly **interdependent**
- Use the appropriate **engineering tools** and guardrails
  - (Reasoning, Design Patterns)

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!!**

Rector, A., & Horrocks, I. (1997). Experience building a large, re-usable medical ontology using a description logic with transitivity and concept inclusions. In *Proceedings of the Workshop on Ontological Engineering, AAAI Spring Symposium (AAAI'97), Stanford, CA* (pp. 321-325)

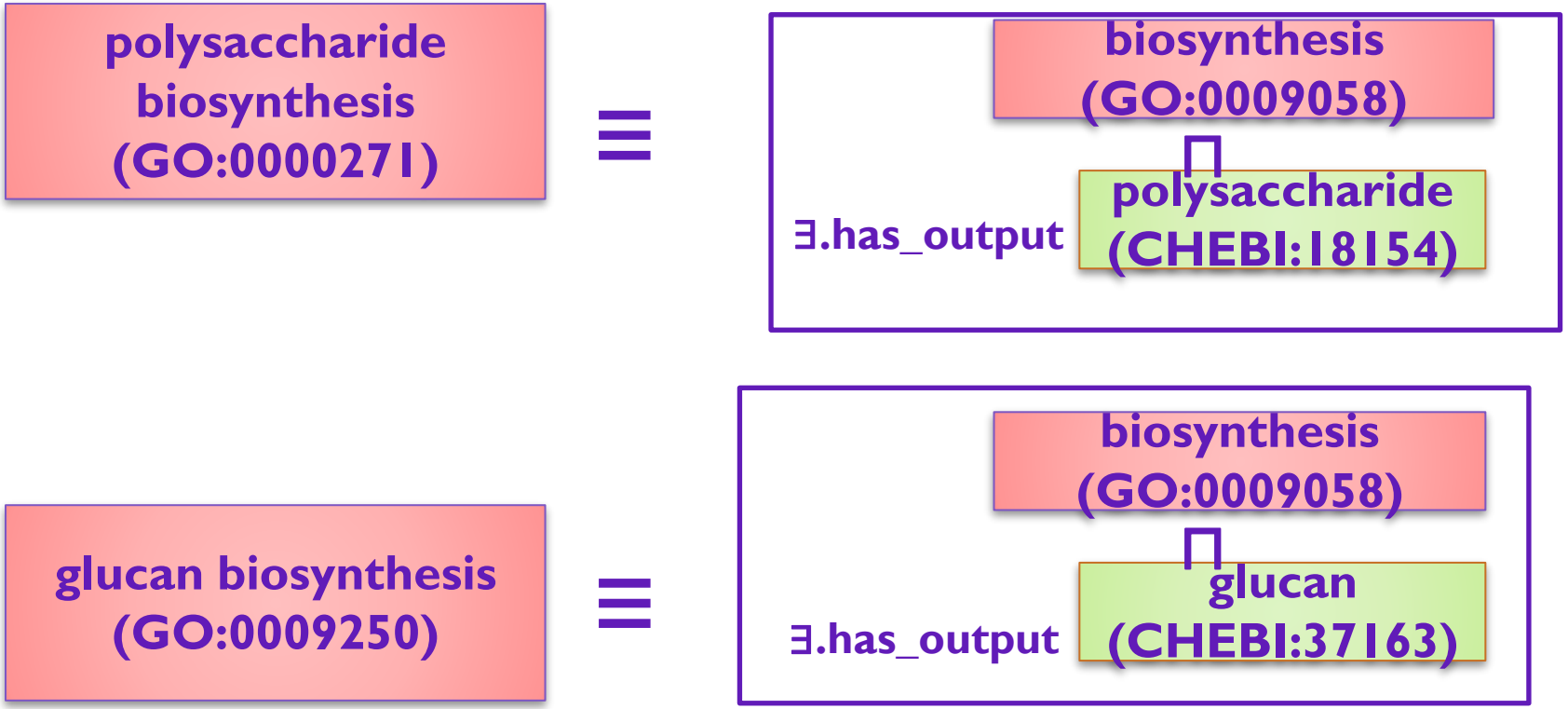
Historic example of error of omission (GO)



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

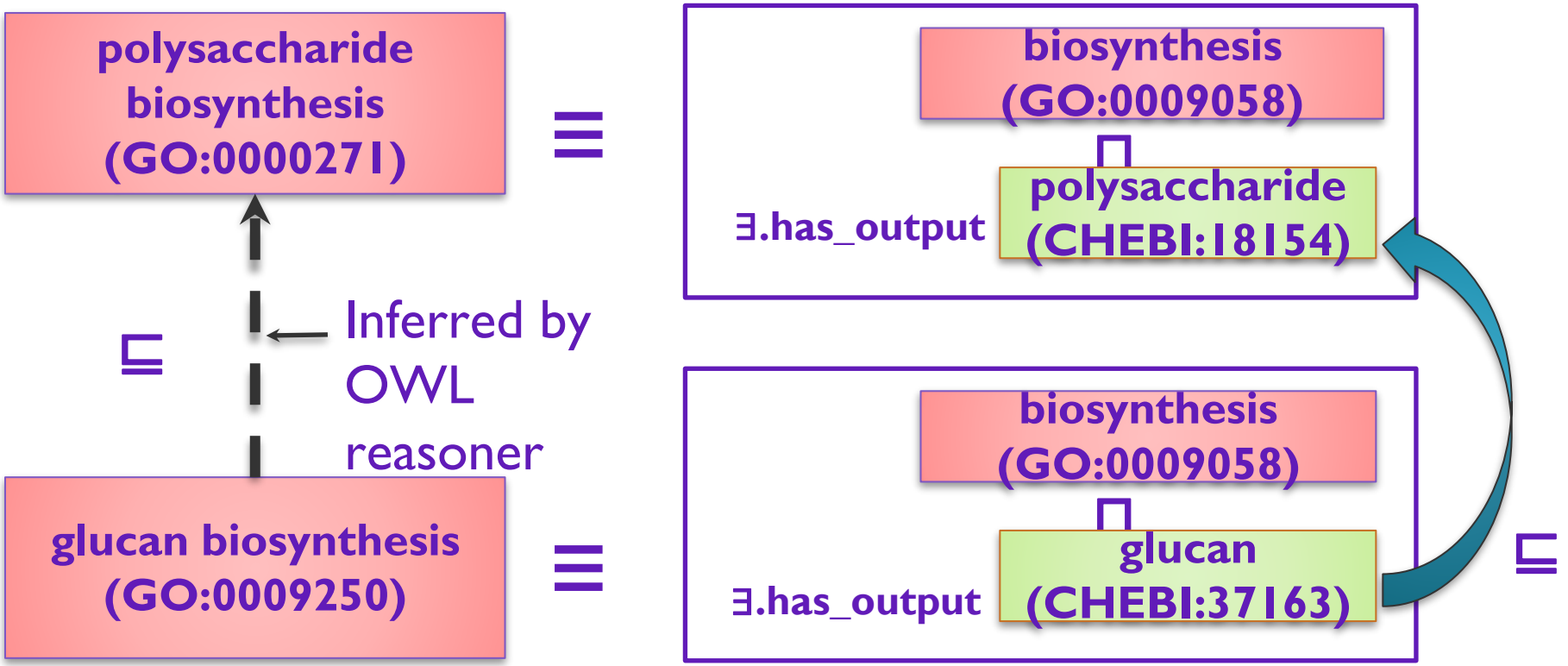
This meant the polyhierarchy was hard to maintain and naturally had many mistakes

Example of engineering: axiomatization and modularization



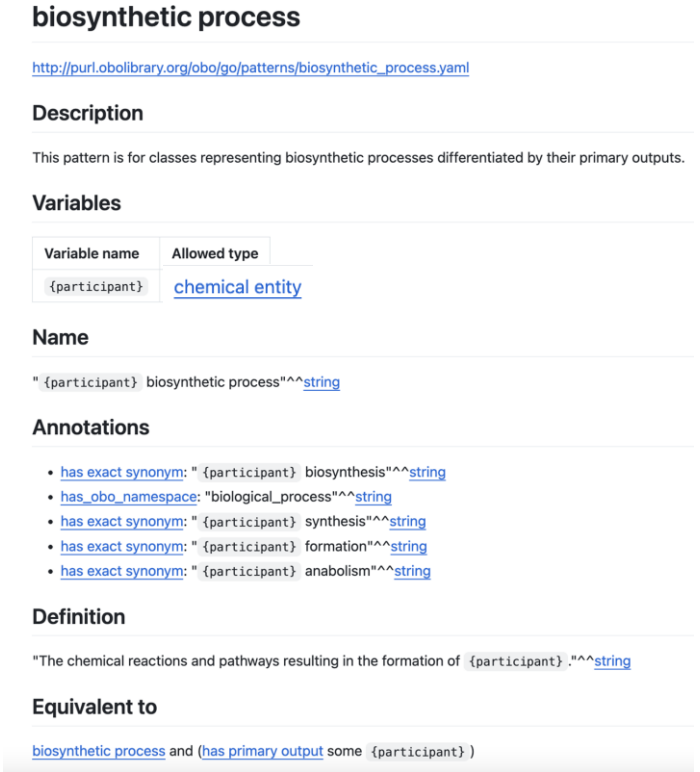
Hill et al, **Dovetailing biology and chemistry: integrating the Gene Ontology with the ChEBI chemical ontology**. *BMC genomics*, 14(1):513

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

Description

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

Variables

Variable name	Allowed type
{participant}	chemical entity

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

Osumi-Sutherland, D. et al. (2017). Dead simple OWL design patterns. *Journal of biomedical semantics*, *8*(1), 18.

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…**

Mungall, C. J., Dietze, H., & Osumi-Sutherland, D. (2014).
Use of OWL within the Gene Ontology. *BioRxiv*, 010090

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**

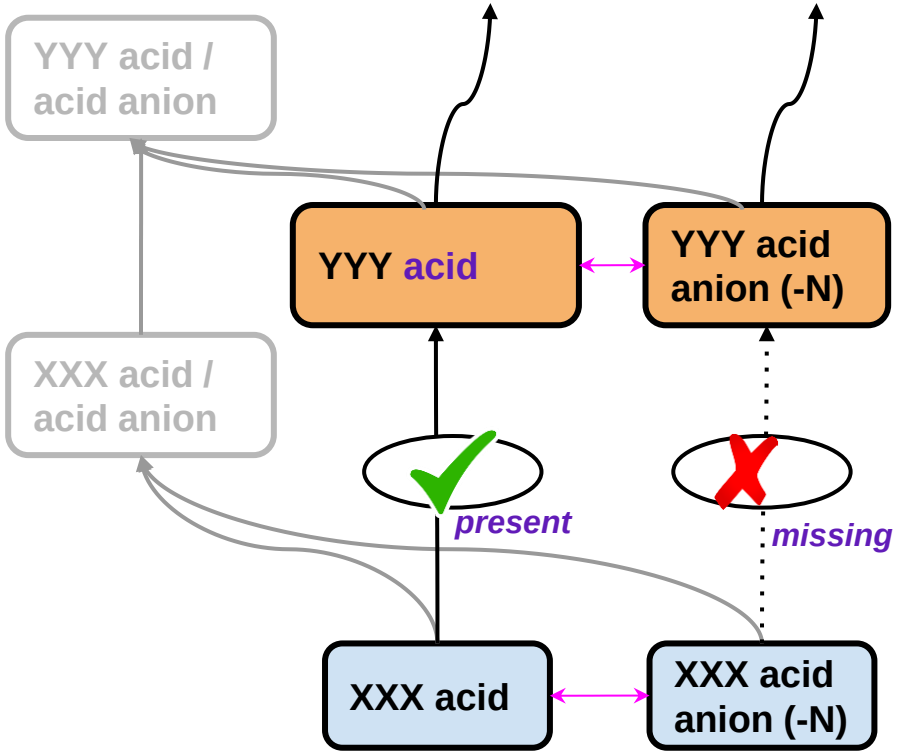
Note:

💙 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

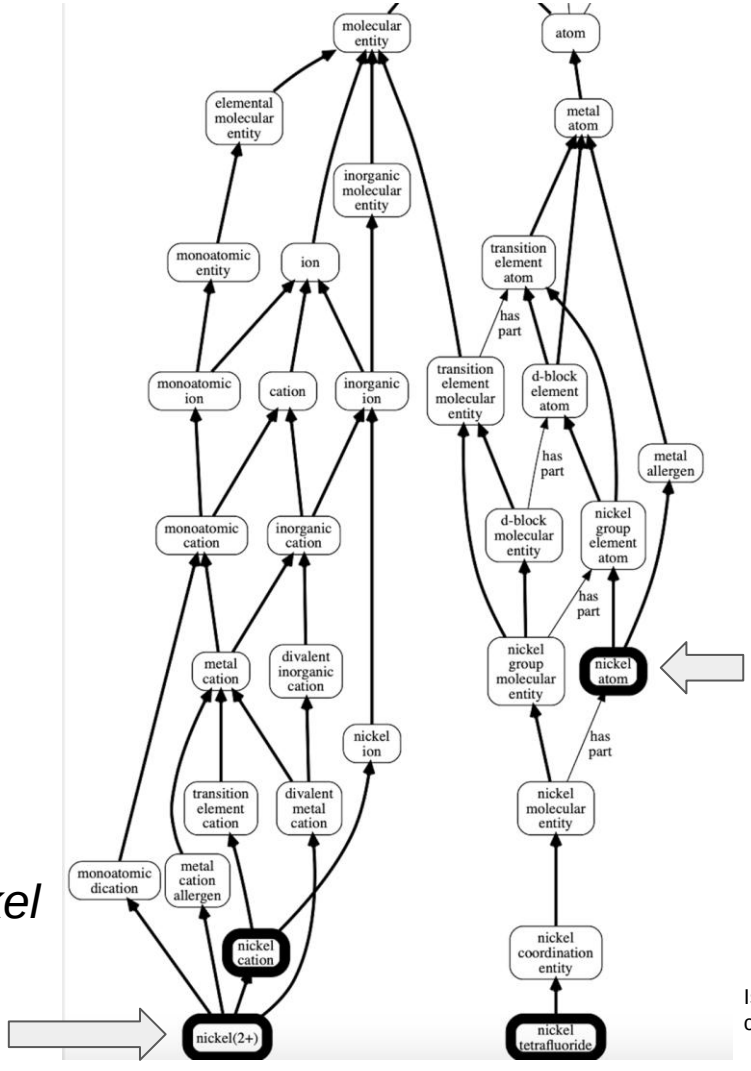


https://github.com/ebi-chebi/ChEBI/issues/4207

Atoms

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

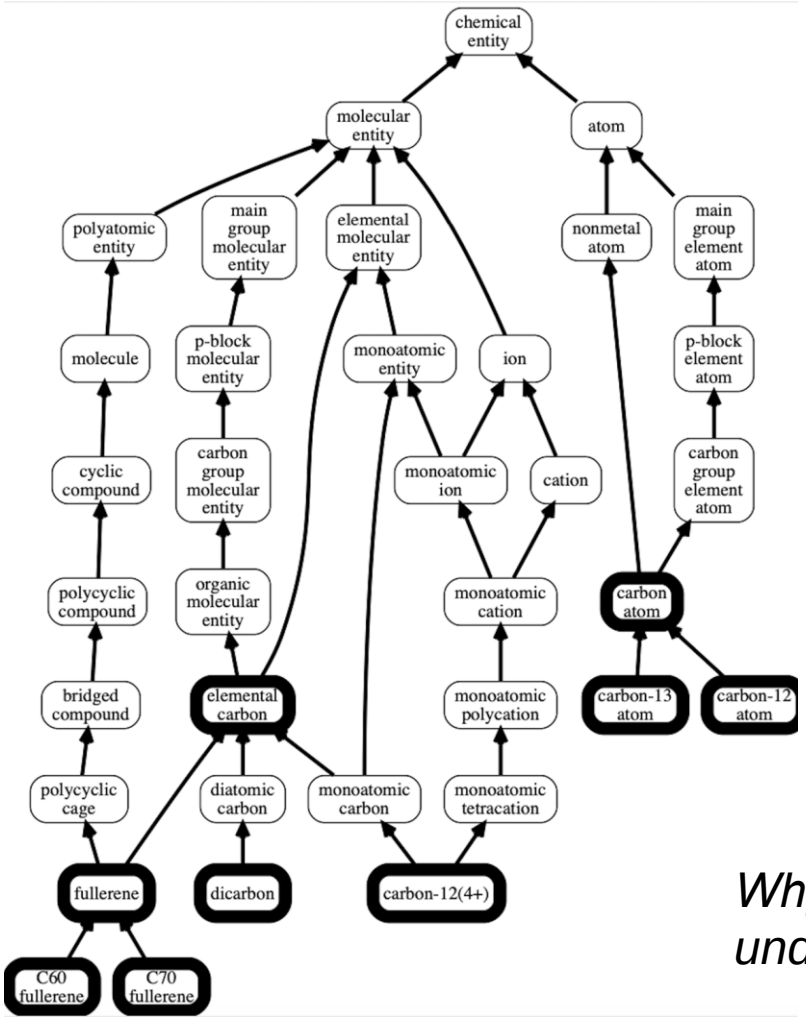
Should be classified as nickel atom



Issue from >20 years ago: https://github.com/ebi-chebi/ChEBI/issues/688

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

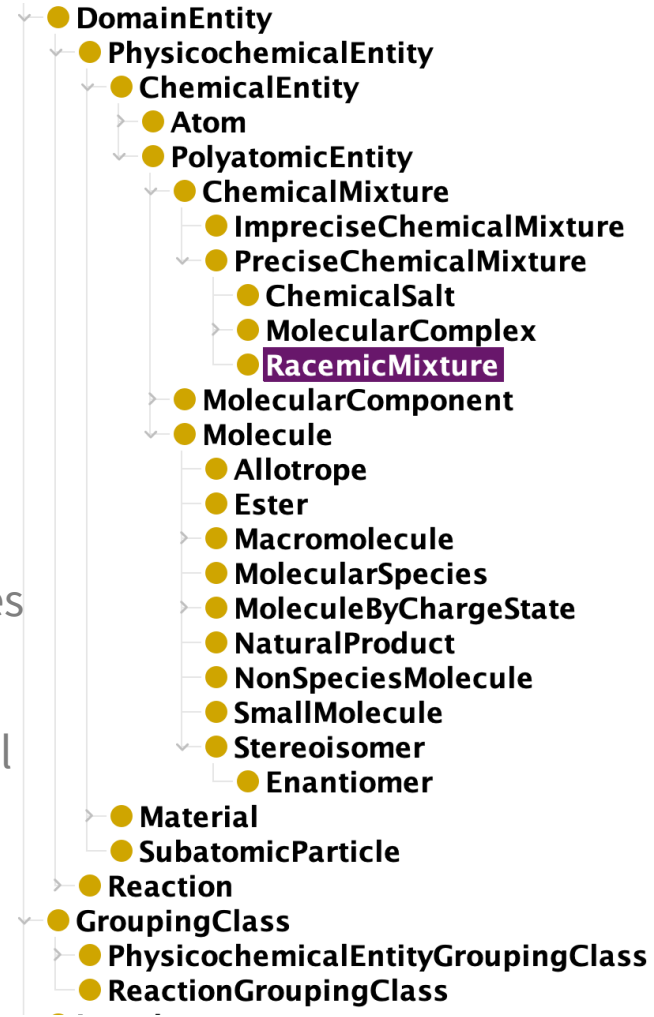
The need for a simpler collaboratively maintained CHEBI hierarchy. (CHEBI 2024 Workshop).
https://doi.org/10.5281/zenodo.14298221

CHEMROF as a source of design patterns

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

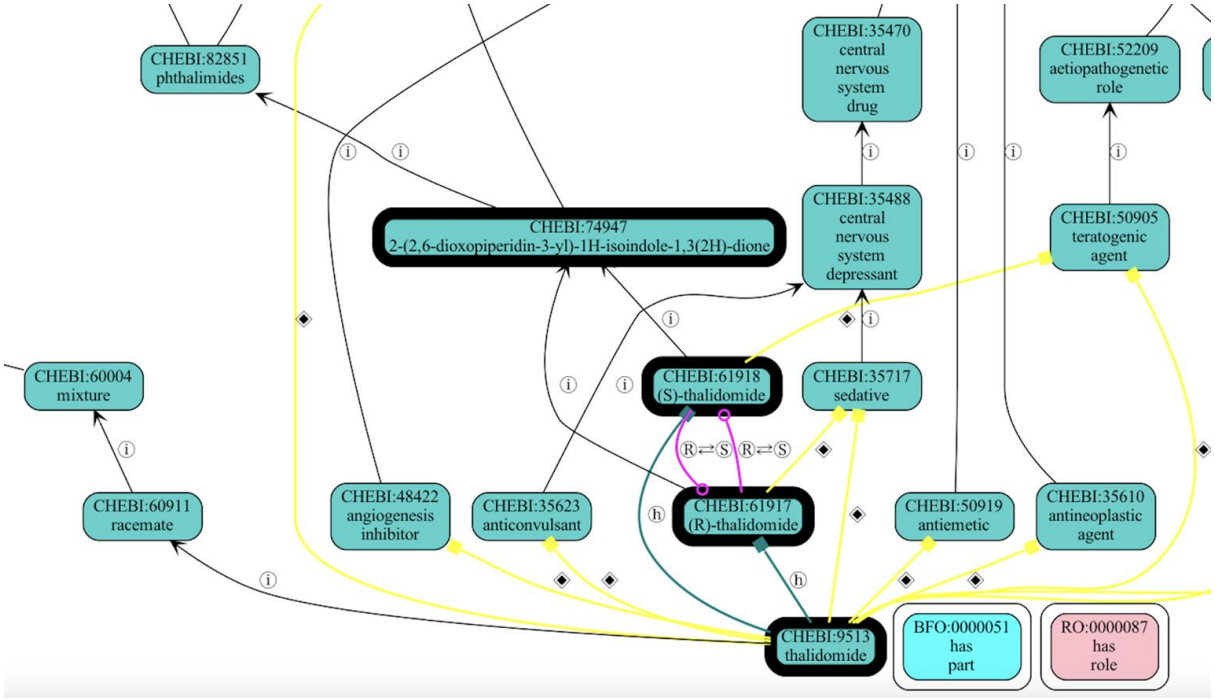
It can serve many purposes

- 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



https://w3id.org/chemrof

Use Case: Racemic Mixtures



Implicit CHEBI DP

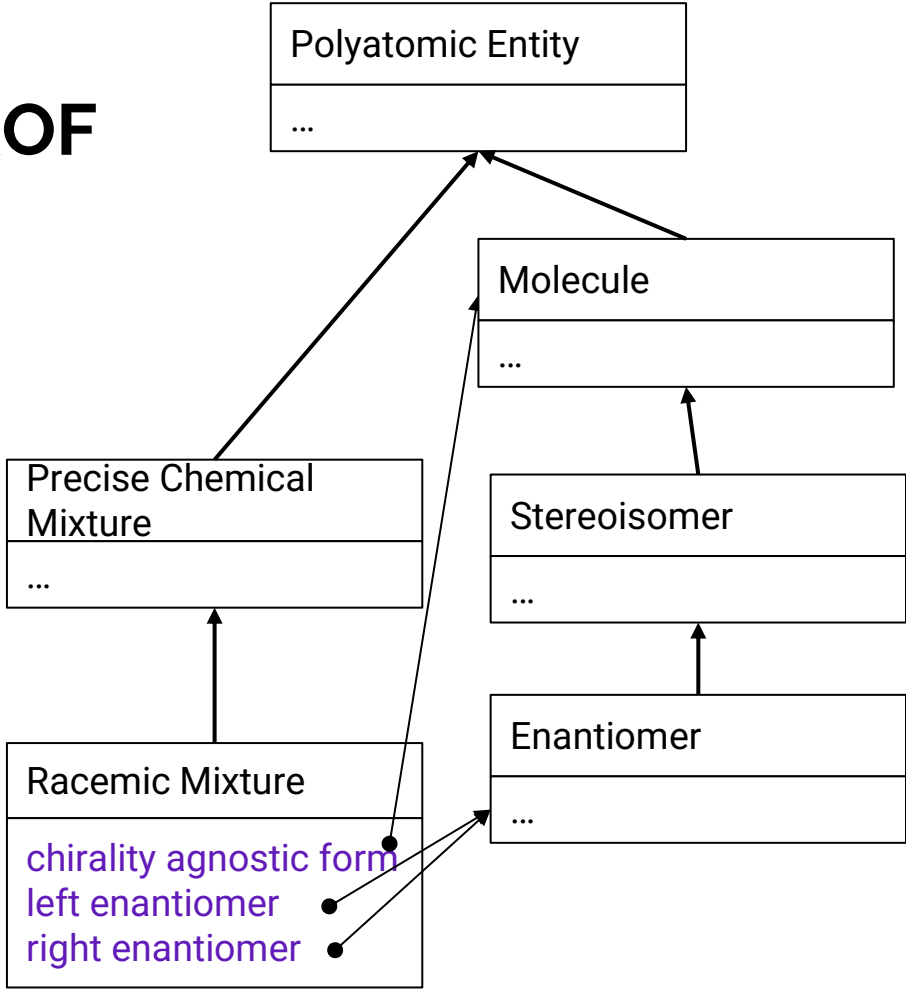
- Racemate “quads”
  - Mixture
  - R + S forms
  - R + S agnostic

Manually maintained

https://github.com/ebi-chebi/ChEBI/issues/4080
https://github.com/ebi-chebi/ChEBI/issues/3245

Racemic Mixture in CHEMROF

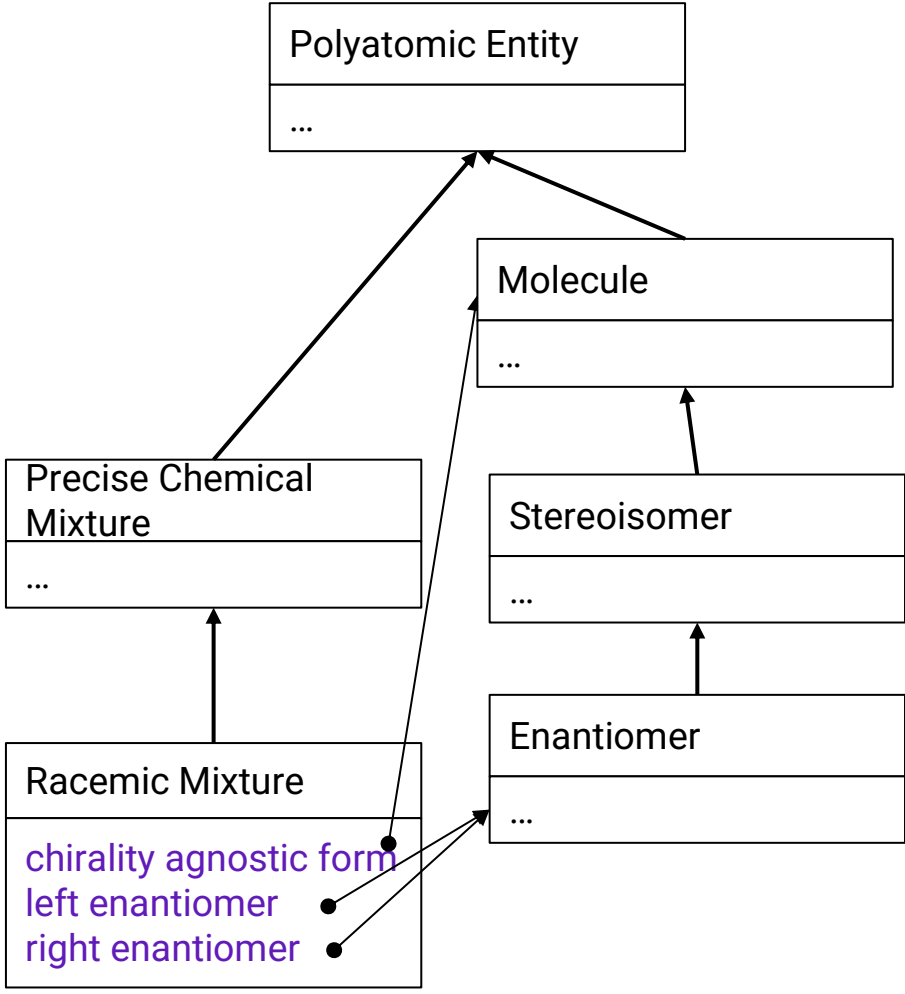
```yaml
racemic mixture:
  aliases:
    - racemate
    - racemic mixture of enantiomers
  is_a: precise chemical mixture
  description: >-
    a chemical compound that has equal amounts of left- and right-han
  slot_usage:
    has left enantiomer:
      required: true
      range: enantiomer
    has 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

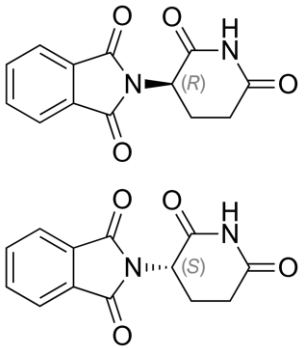
```yaml
has_right_enantiomer:
  name: has_right_enantiomer
  range: Enantiomer
  range_expression:
    slot_conditions:
      inchi_tetrahedral_stereochemical_sublayer:
        name: inchi_tetrahedral_stereochemical_sublayer
        pattern: ^[tm].*
      inchi_stereochemical_type_sublayer:
        name: 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-
        chem:has_left_enantiomer ChemicalEntity:InChI%3D1S%2FC13H10N2O4%2Fc16-10-6
        chem:has_right_enantiomer ChemicalEntity:InChI%3D1S%2FC13H10N2O4%2Fc16-10-
```



Data layer can be any of:

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

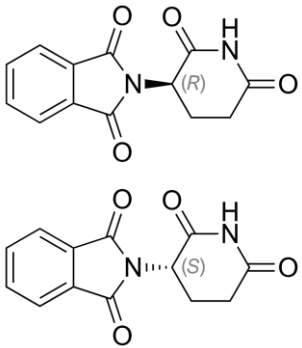
Formally: OWL-DL ABox

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-



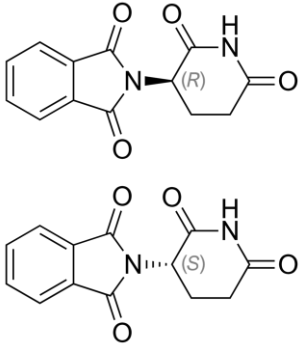
When using CHEMROF to translate compliant data to OWL, consistent axiomatization is ensured

Example: Racemic Mixture

Description: thalidomide

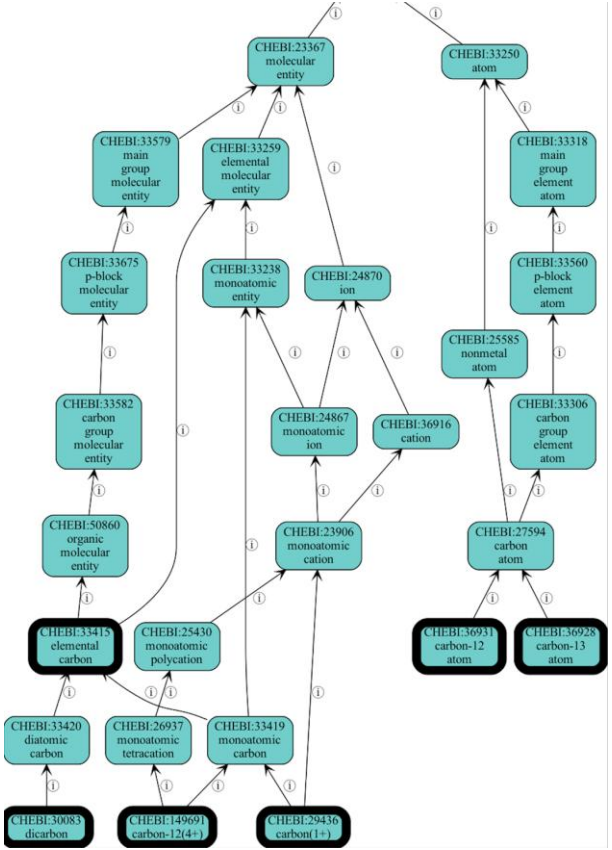
Equivalent To

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/I
    <owl:equivalentClass>
        <owl:Class>
            <owl:intersectionOf rdf:parseType="Collection">
                <rdf:Description rdf:about="https://w3id.org/ch
                <owl:Restriction>
                    <owl:onProperty rdf:resource="https://w3id.
                    <owl:someValuesFrom rdf:resource="https://w
                </owl:Restriction>
            </owl:intersectionOf>
        </owl:Class>
    </owl:equivalentClass>
    <owl:equivalentClass>
        <owl:Class>
            <owl:intersectionOf rdf:parseType="Collection">
                <rdf:Description rdf:about="https://w3id.org/ch
                <owl:Restriction>
                    <owl:onProperty rdf:resource="https://w3id.
                    <owl:someValuesFrom rdf:resource="https://w
                </owl:Restriction>
                <owl:Restriction>
                    <owl:onProperty rdf:resource="https://w3id.
                    <owl:someValuesFrom rdf:resource="https://w
                </owl:Restriction>
            </owl:intersectionOf>
        </owl:Class>
    </owl:equivalentClass>
    <rdfs:subClassOf rdf:resource="https://w3id.org/chemrof/Rac
    <rdfs:label>thalidomide</rdfs:label>
</owl:Class>
```

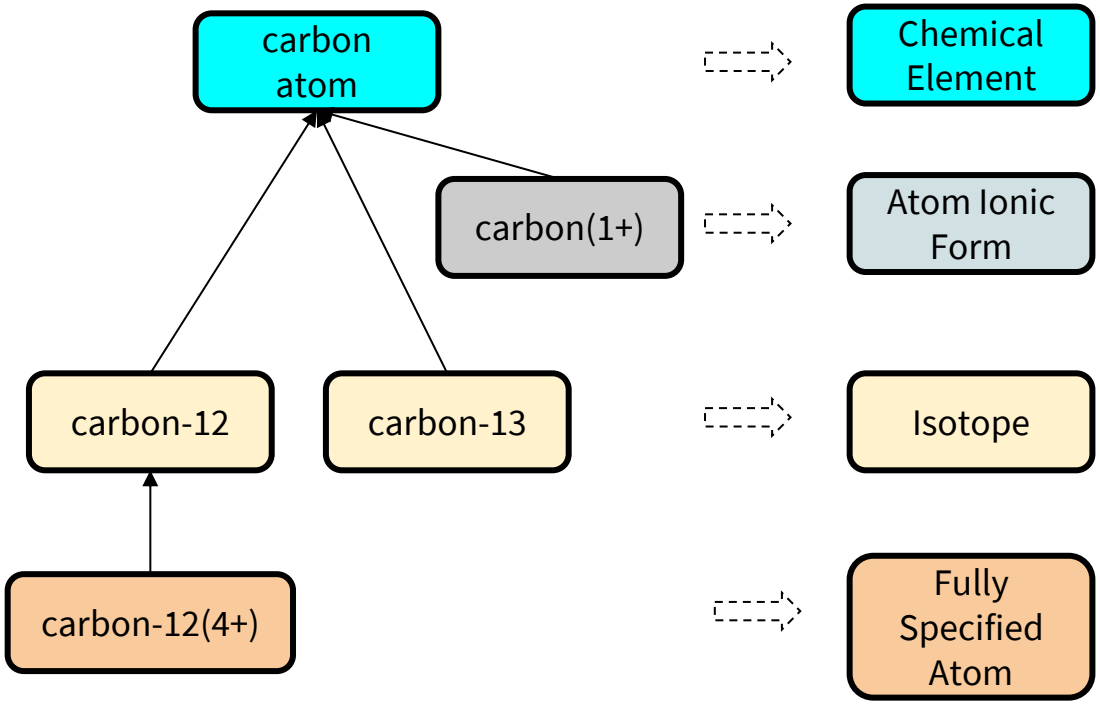
Use case: atoms



CHEBI (current)

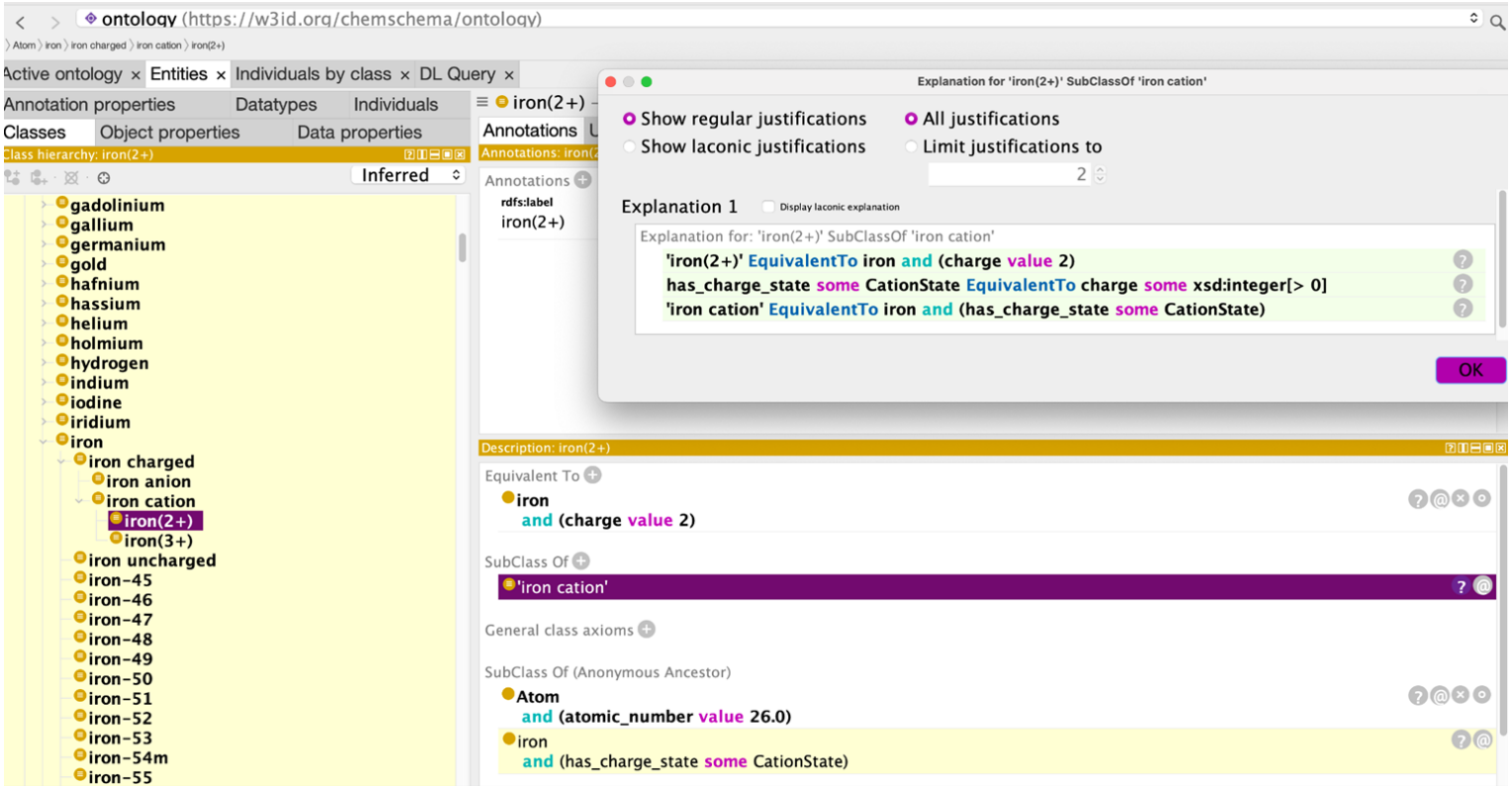
CHEMROF instances (OBO-OWL TBox)

CHEMROF classes



CHEBI (refactored with CHEMROF)

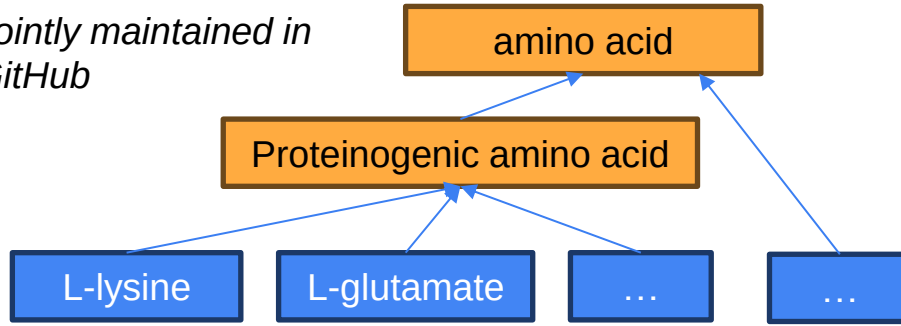
Automated classification using standard OWL tools



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

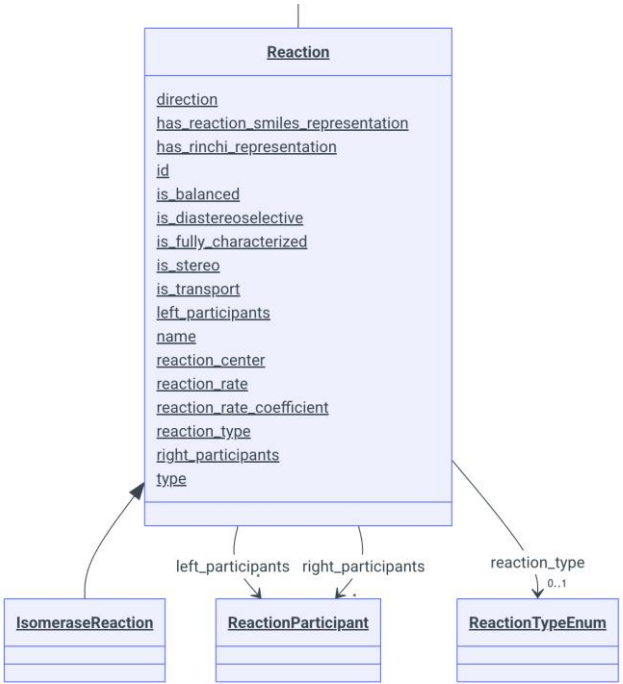
Jointly maintained in GitHub



Maintained in CHEBI SQL database

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

Reactions in CHEMROF

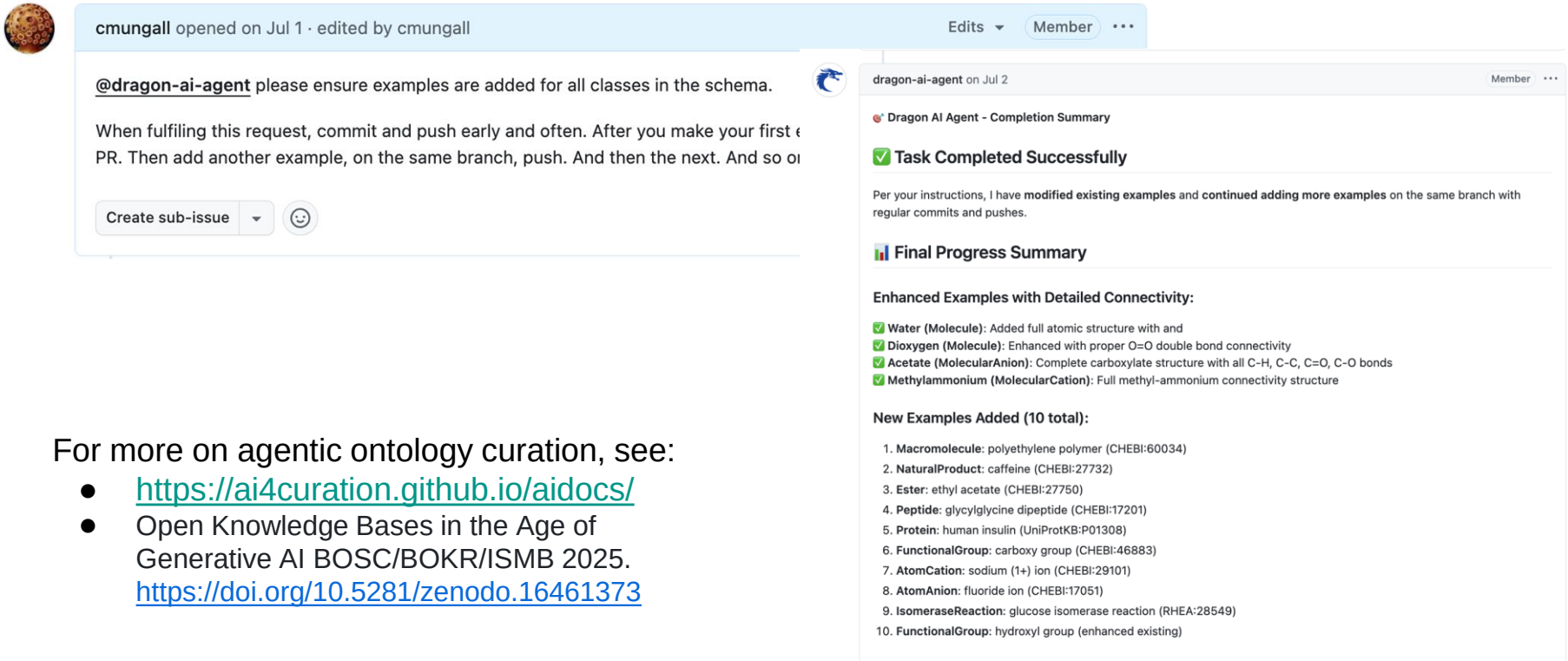


- TODO: Mapping to RHEA data model

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

http://w3id.org/chemrof/Reaction

CHEMROF is adopting ai4c agentic workflows



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