The Open Reaction Database

Connor Coley (ccoley@mit.edu) and
Steven Kearnes (kearnes@google.com)

RDKit UGM | 7 October 2020 | open-reaction-database.org

Governing Committee

- Connor Coley (MIT)
- Abby Doyle (Princeton)
- Spencer Dreher (Merck)
- Joel Hawkins (Pfizer)
- Klavs Jensen (MIT)
- Steven Kearnes (Google)

Advisory Board

- Juan Alvarez (Merck)
- Alán Aspuru-Guzik (Toronto, MADNESS)
- Tim Cernak (Michigan)
- Lucy Colwell (Cambridge, SynTech, Google)
- Werngard Czechtizky (AstraZeneca)
- Matthew Gaunt (Cambridge, SynTech)
- Mimi Hii (Imperial, ROAR)
- Greg Landrum (T5 Informatics)
- Fabio Lima (Novartis)
- Christos Nicolaou (Lilly)
- Sarah Reisman (Caltech)
- Matthew Sigman (Utah)
- Sarah Trice (MilliporeSigma)
- Matt Tudge (GSK)

Design considerations

From the documentation: "support machine learning and related efforts in reaction prediction, chemical synthesis planning, and experiment design"

Design considerations

From the documentation: "support machine learning and related efforts in reaction prediction, chemical synthesis planning, and experiment design"

Goals:

- Provide a structured data format for chemical reaction data
- Provide an interface for easy browsing and downloading of data
- Make reaction data freely and publicly available for anyone to use
- Encourage sharing of precompetitive proprietary data, especially HTE data

Design considerations

From the documentation: "support machine learning and related efforts in reaction prediction, chemical synthesis planning, and experiment design"

Non-goals (at least initially):

- Capture reaction processes as action sequences for robotic execution
- Store processed, structured analytical data or other inputs that are not directly related to the machine learning efforts described above
- Integrate model building or other use of the data as part of the database

Primary use cases: synthetic organic chemistry

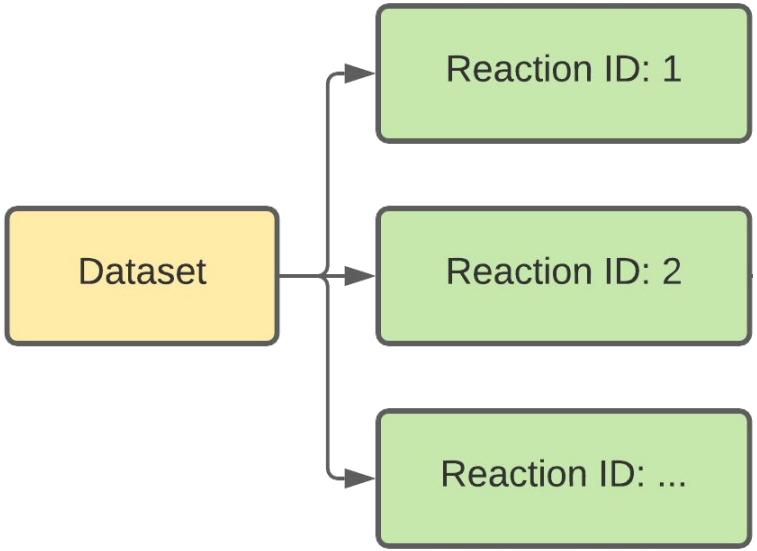
1. High-throughput experimentation
    a. Data are recorded in spreadsheet formats including only varied parameters;
    b. One template Reaction is defined to specify all aspects held constant;
    c. The Dataset is defined by iterating over the spreadsheet and creating one Reaction entry per experimental condition.
2. “Traditional” bench chemistry
    a. A chemist uses a graphical webform to define the settings and outcomes of all reactions used within a paper or project;
    b. The structured Dataset is saved, uploaded to the Open Reaction Database, and used as part of their supporting information;
    c. A list of reactions is exported from the Dataset in an SI-like text format.

Goals for the schema

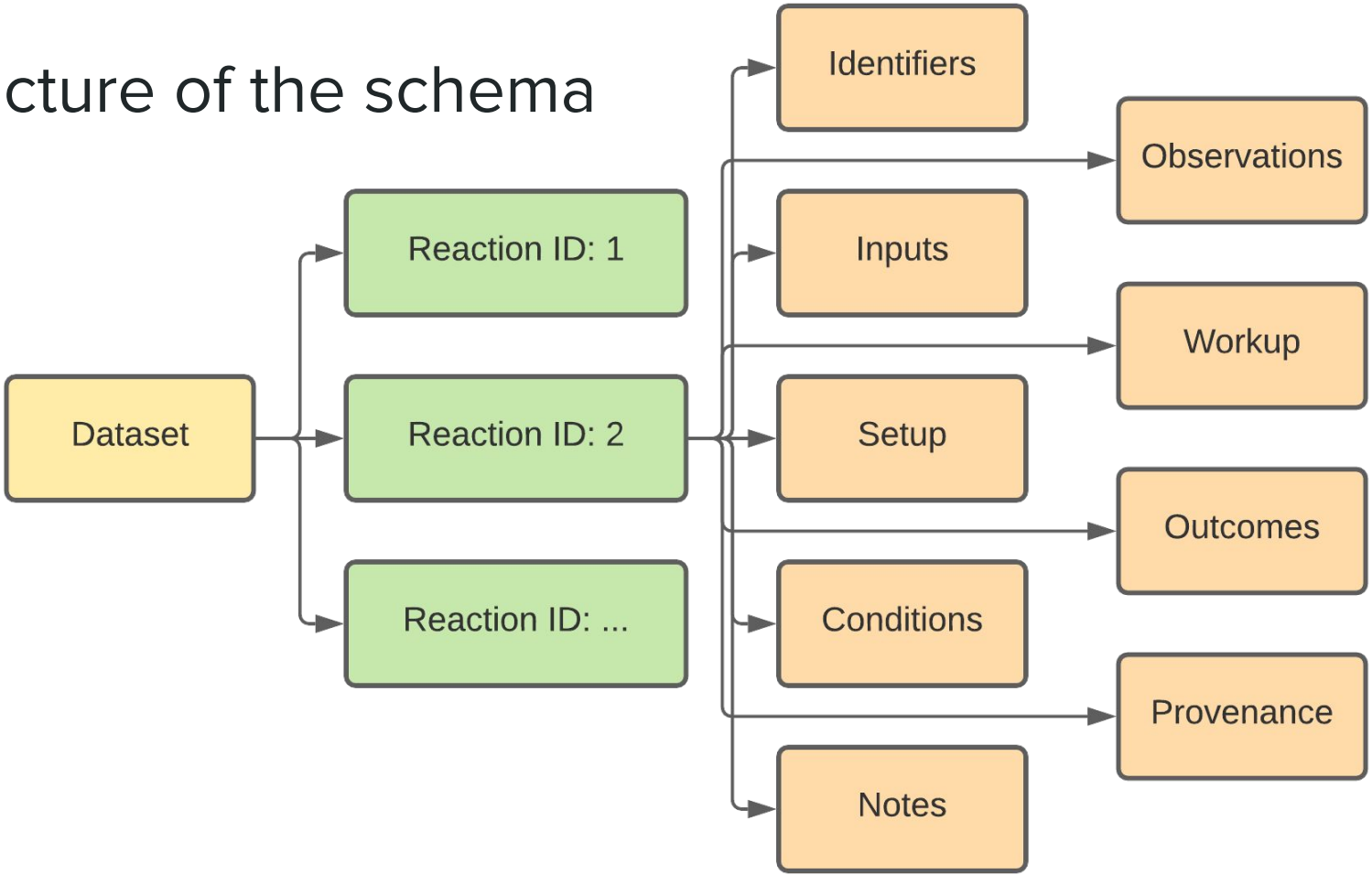
- Capture the most important aspects of reactions in a *structured* format
  - Guided by our survey last winter, the focus is on single-step batch reactions
  - Structured data enables downstream ML applications
- Allow additional details in a flexible, *unstructured* format
- Match chemist expectations around structure and nomenclature
- Record what physically occurred in a chemical reaction; de-emphasize recording of a chemist's intent
  - e.g., record the actual masses and volumes that were used to create a stock solution, not the target concentration
- Be human readable

Structure of the schema

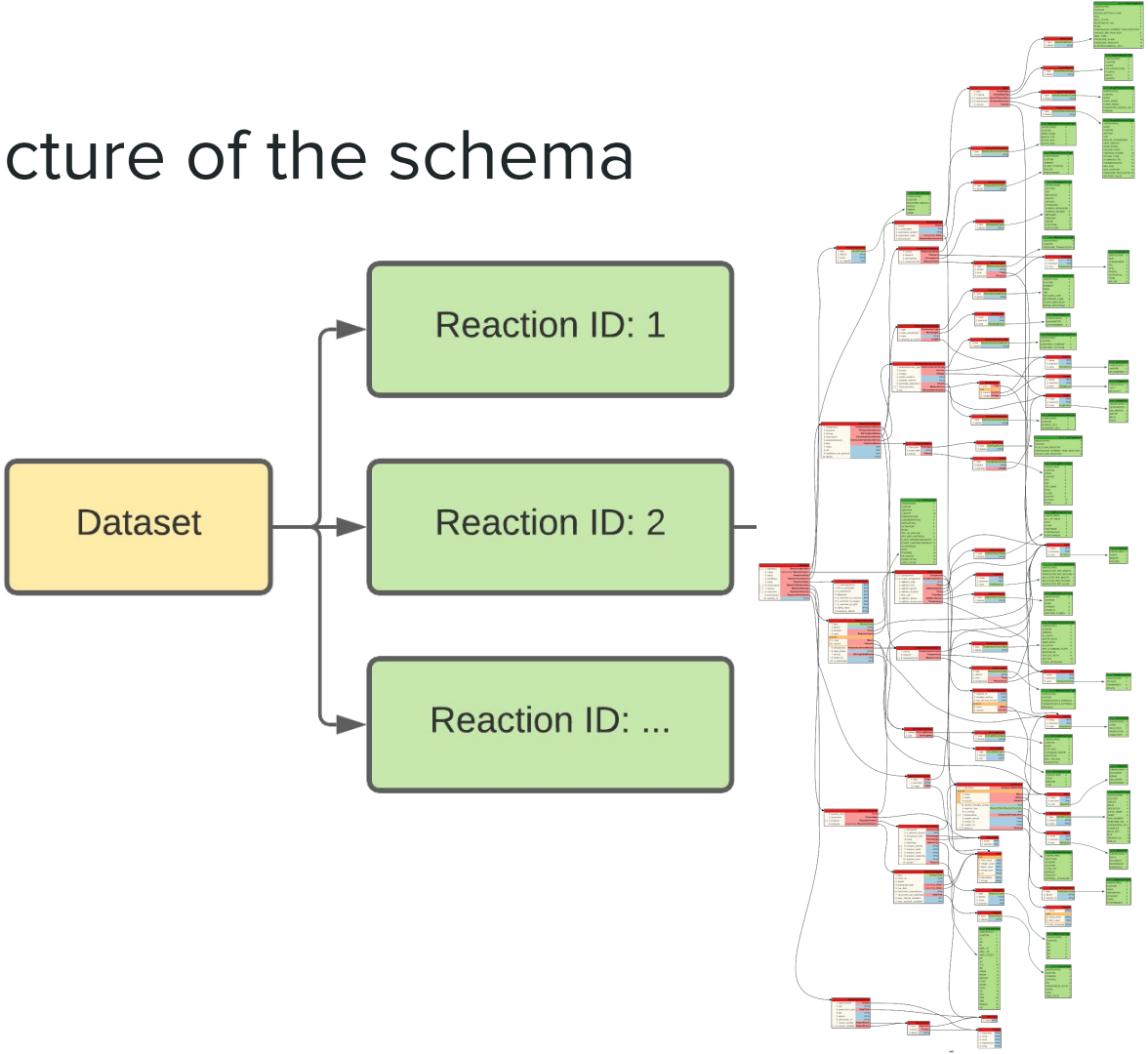
Structure of the schema



Structure of the schema



Structure of the schema



Protocol buffers

```
message Mass {
  enum MassUnit {
    UNSPECIFIED = 0;
    KILOGRAM = 1;
    GRAM = 2;
    MILLIGRAM = 3;
    MICROGRAM = 4;
  }
  float value = 1;
  // Precision of the measurement (with the same units as `value`).
  float precision = 2;
  MassUnit units = 3;
}
```

Protocol buffers

```
mass = schema.Mass(value=1.25, units='GRAM')
```

Protocol buffers

```
mass = schema.Mass(value=1.25, units='GRAM')

resolver = units.UnitResolver()
mass = resolver.resolve('1.25 g')
```

Protocol buffers

```
mass = schema.Mass(value=1.25, units='GRAM')

resolver = units.UnitResolver()
mass = resolver.resolve('1.25 g')

mass_json = """{
  "value": 1.25,
  "units": "GRAM"
}"""
mass = json_format.Parse(mass_json, schema.Mass)
```

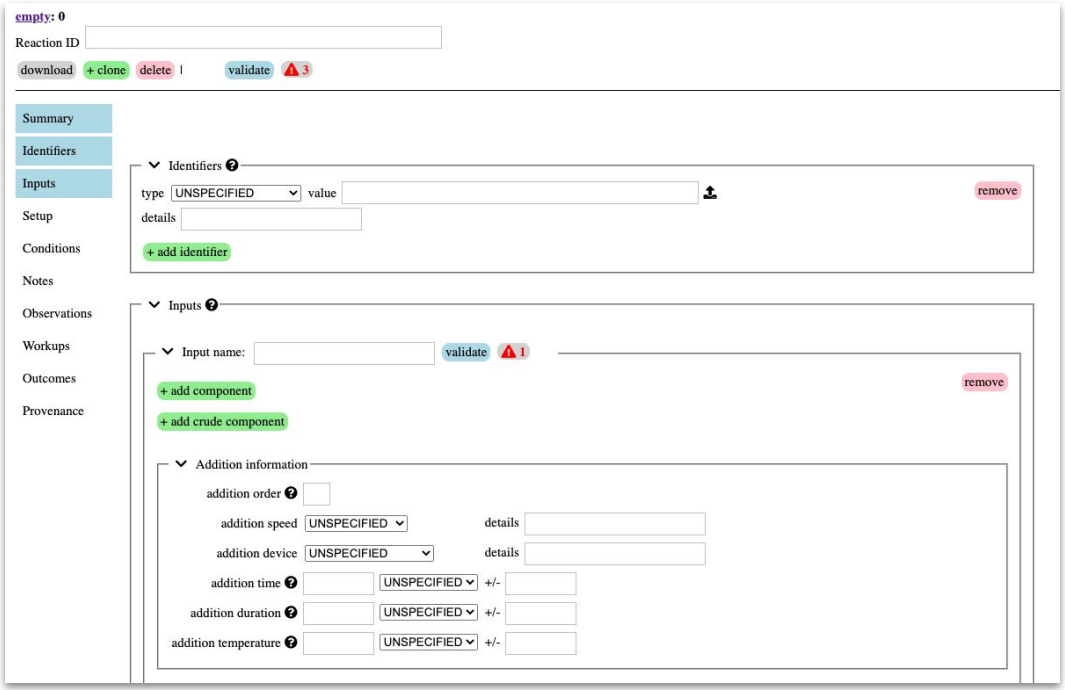
```
reaction = schema.Reaction()
reaction.identifiers.add(value=r'deoxyfluorination', type='NAME')

# Input 1a is a stock solution of alcohol in THF
reaction.inputs['alcohol in THF'].addition_order = 1
solute = reaction.inputs['alcohol in THF'].components.add()
solvent = reaction.inputs['alcohol in THF'].components.add()

solute.reaction_role = schema.Compound.ReactionRole.REACTANT
solute.identifiers.add(value=r'c1ccccc1CCC(O)C', type='SMILES')
solute.moles.CopyFrom(unit_resolver.resolve('0.1 mmol'))
solute.is_limiting = True

solvent.reaction_role = schema.Compound.ReactionRole.SOLVENT
solvent.identifiers.add(value=r'THF', type='NAME')
solvent.identifiers.add(value=r'C1CCCO1', type='SMILES')
solvent.volume.CopyFrom(unit_resolver.resolve('125 uL'))
solvent.preparations.add(type='DRIED')

...
```

```
reaction = schema.Reaction()
reaction.identifiers.add(value=r'deoxyfluorination', type='NAME')

# Input 1a is a stock solution of alcohol in THF
reaction.inputs['alcohol in THF'].addition_order = 1
solute = reaction.inputs['alcohol in THF'].components.add()
solvent = reaction.inputs['alcohol in THF'].components.add()

solute.reaction_role = schema.Compound.ReactionRole.REACTANT
solute.identifiers.add(value=r'c1ccccc1CCC(O)C', type='SMILES')
solute.moles.CopyFrom(unit_resolver.resolve('0.1 mmol'))
solute.is_limiting = True

solvent.reaction_role = schema.Co
solvent.identifiers.add(value=r'
solvent.identifiers.add(value=r'C
solvent.volume.CopyFrom(unit_reso
solvent.preparations.add(type='DR

...
```

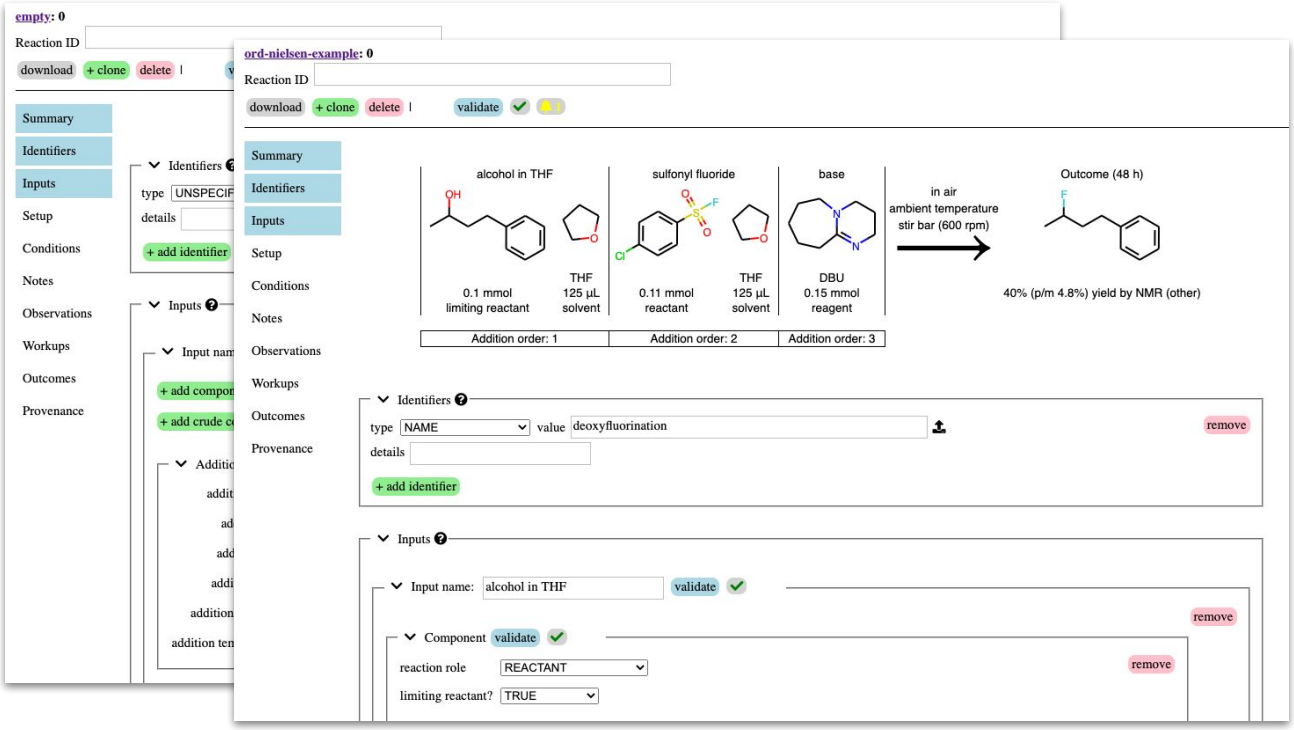
The RDKit is used in the data validation process:

- Reaction SMILES are checked for validity
- Compound structural identifiers are checked for validity and consistency

Web interface

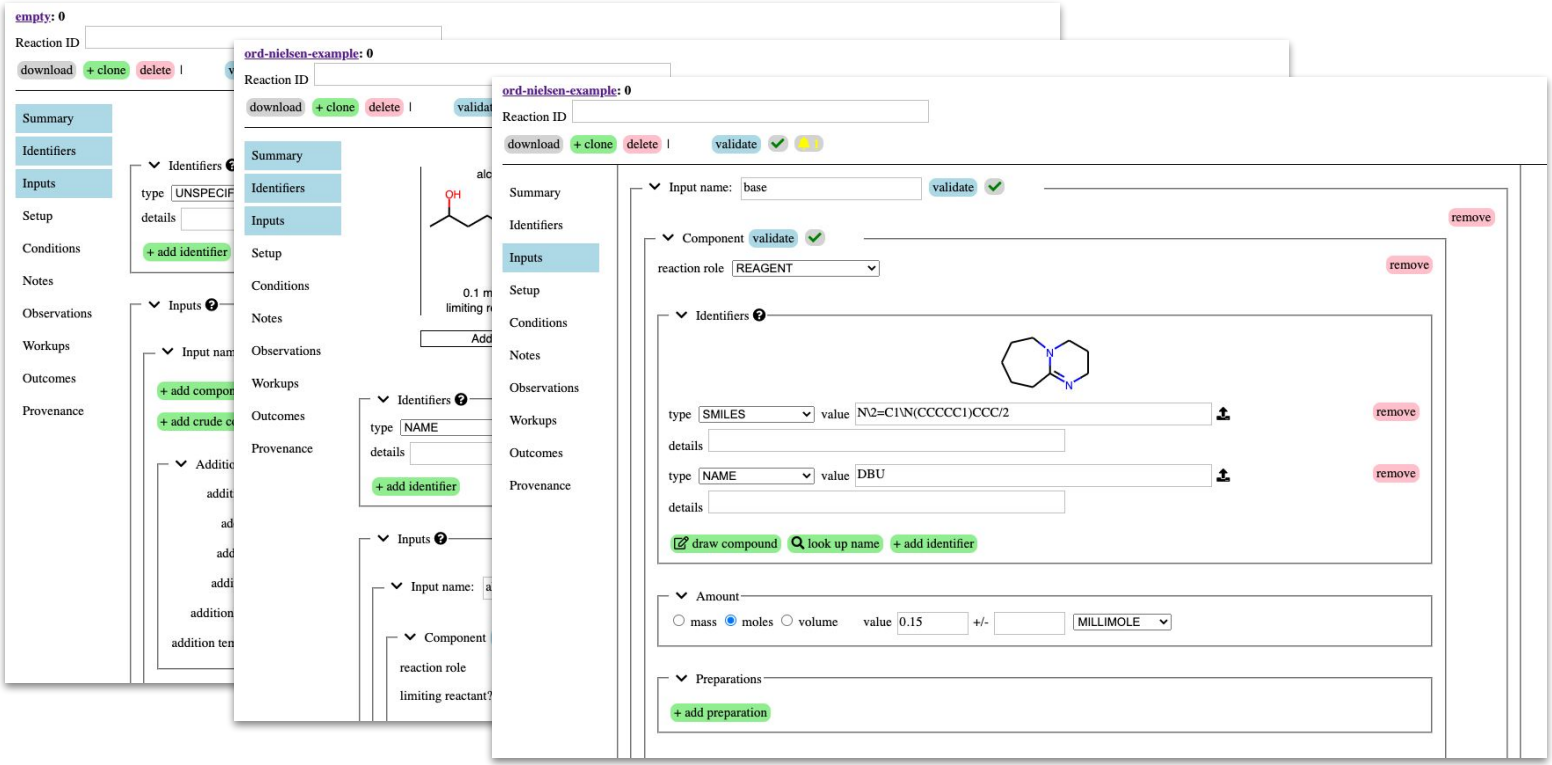


Web interface



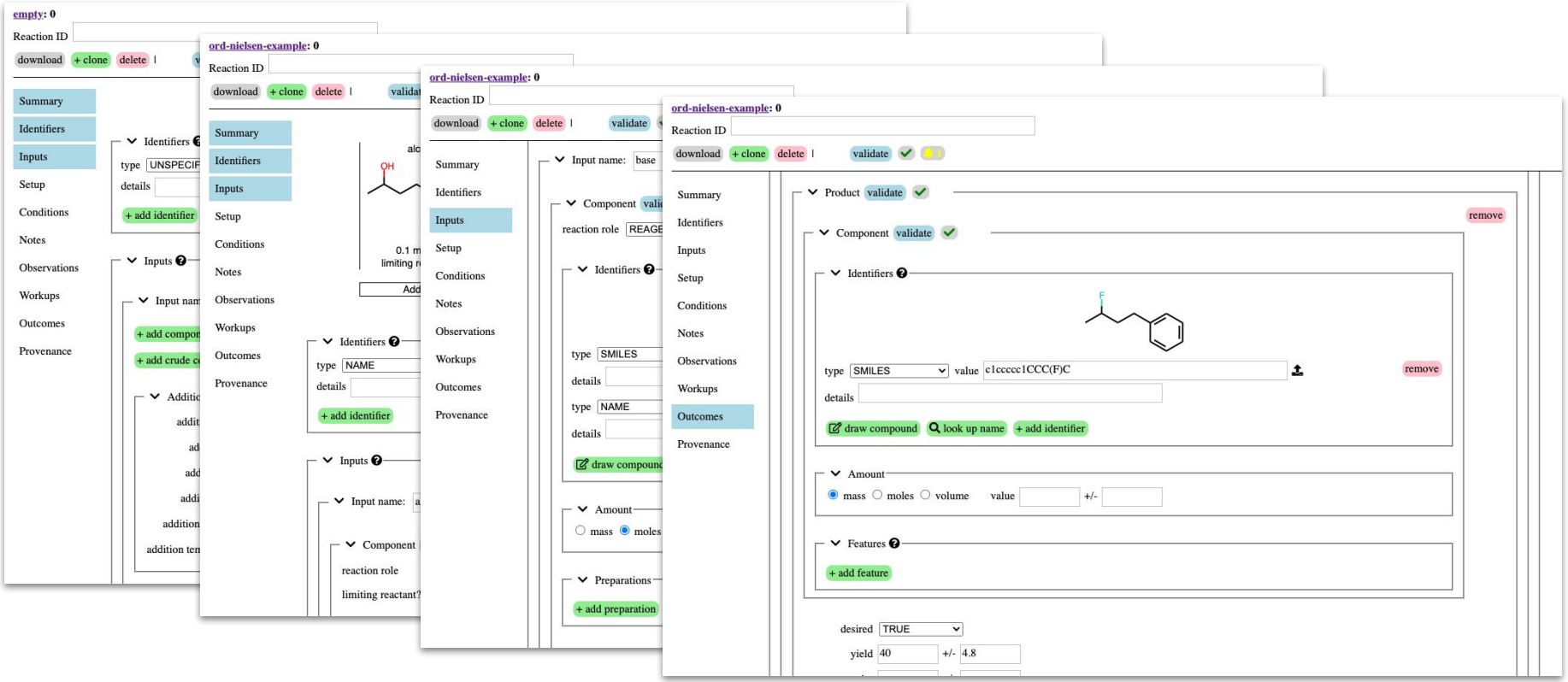
Contributors: Nathan Kim (MIT), Anton Kast (Google)

Web interface



Contributors: Nathan Kim (MIT), Anton Kast (Google)

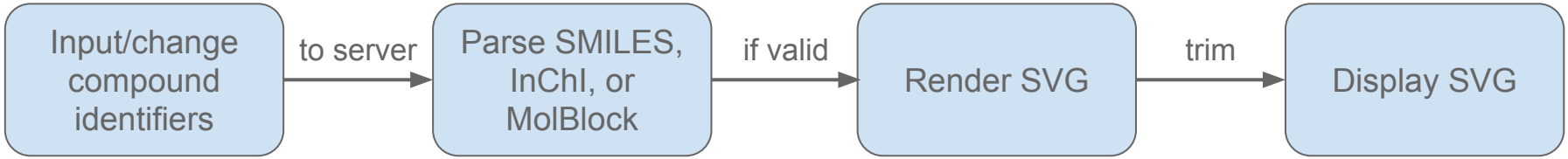
Web interface

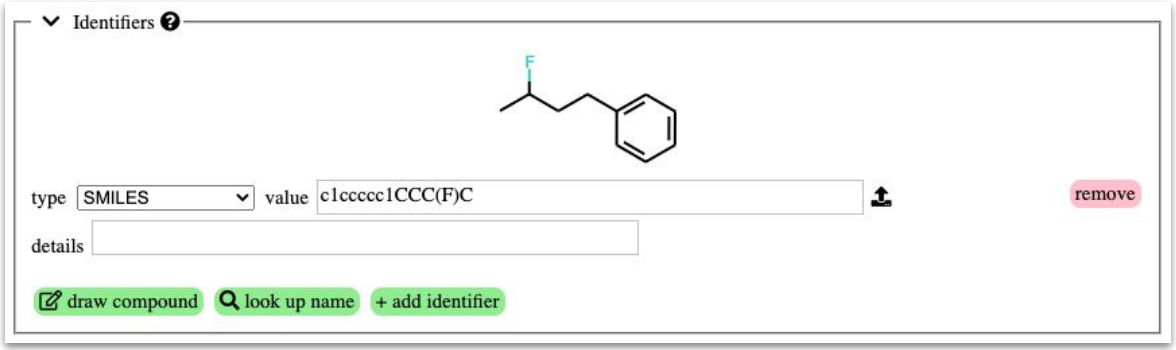


Contributors: Nathan Kim (MIT), Anton Kast (Google)

Web editor compound visualization

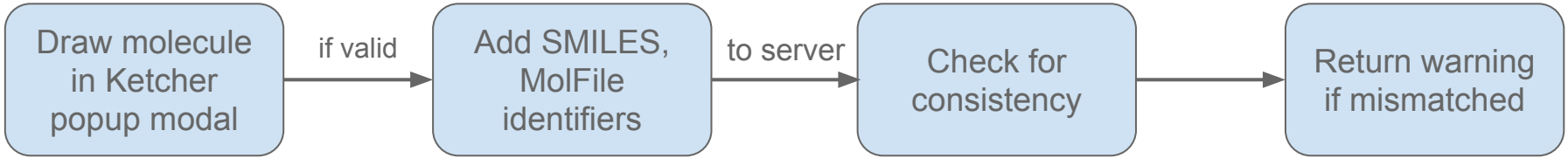
When viewing/editing a Reaction record, it's much easier to look at a chemical rendering than a SMILES or InChI string:

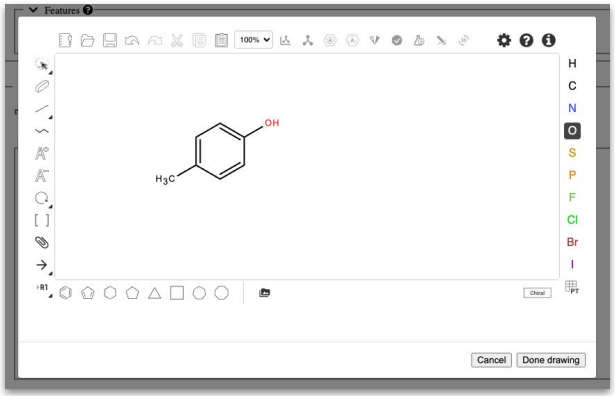


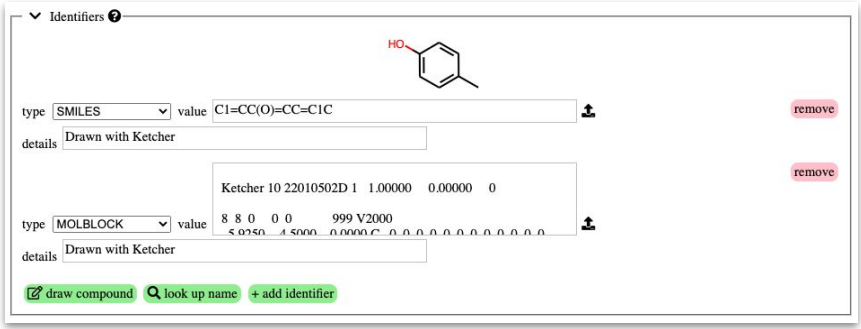


Web editor Ketcher integration

When defining a new molecule, it is handy to have an embedded drawing tool; we use Ketcher.

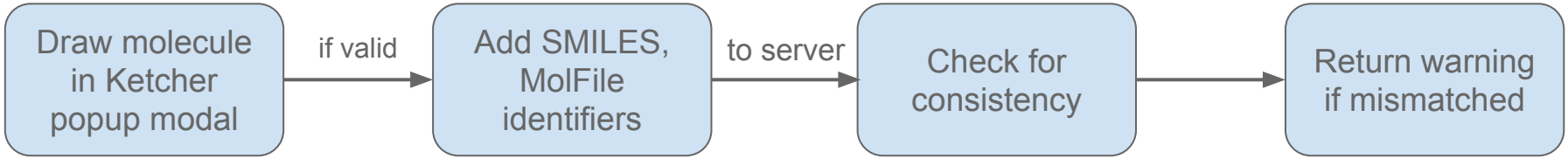


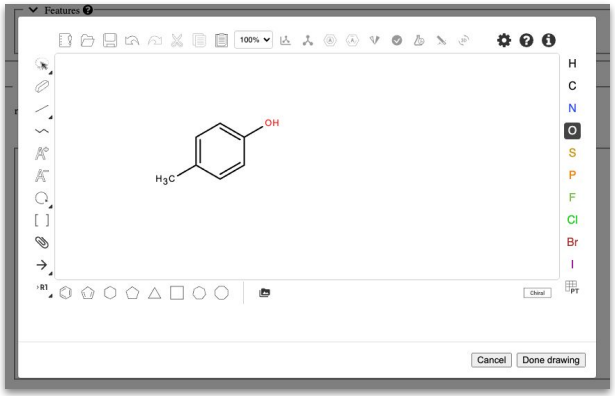




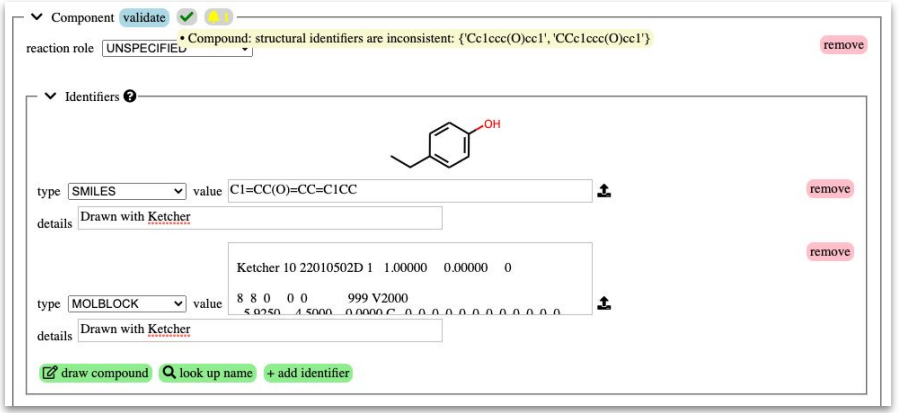
Web editor Ketcher integration

When defining a new molecule, it is handy to have an embedded drawing tool; we use Ketcher.





change SMILES

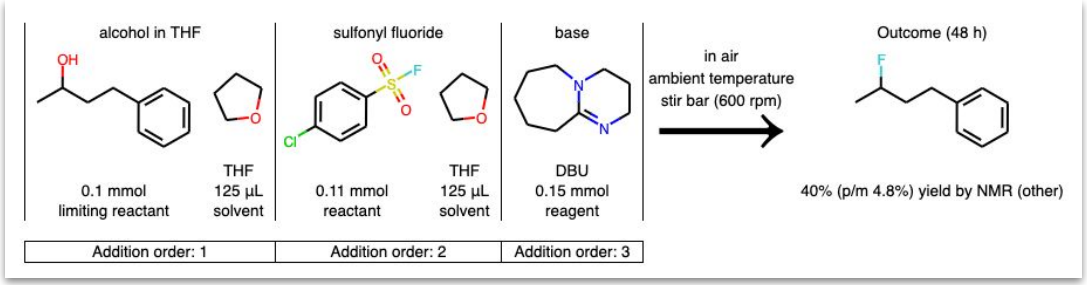


Web editor reaction visualization

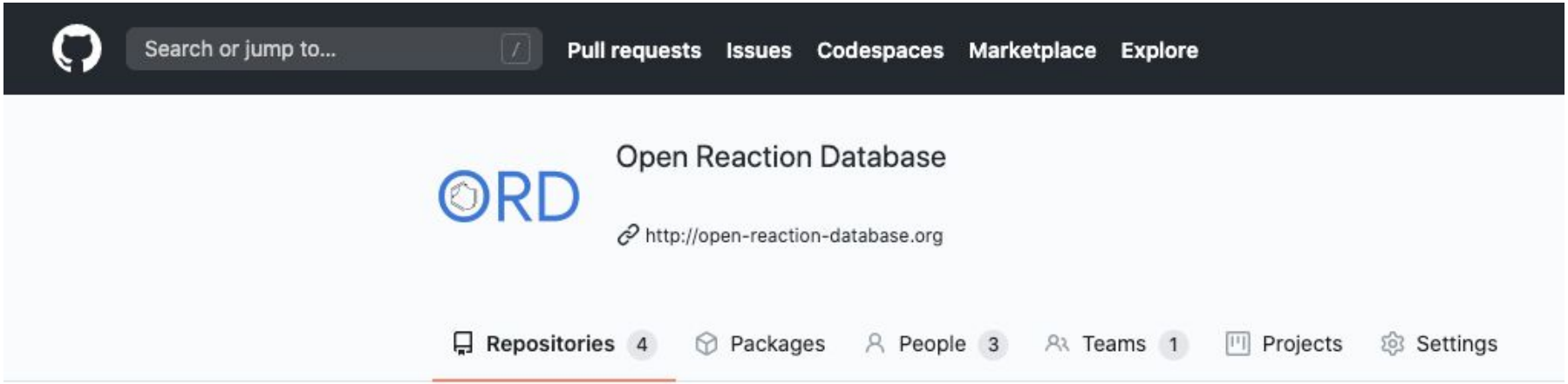
A tabular view of the overall reaction provides a concise summary. This is an HTML table produced using a Jinja2 template and RDKit-generated SVGs.

```
…
{% for key, input in reaction.inputs|sort_addition_order %}
    {% for compound, border in input.components|get_input_borders %}
        <td class="{{ border }}">{{ compound|compound_svg }}</td>
    {% endfor %}
{% endfor %}
…
```

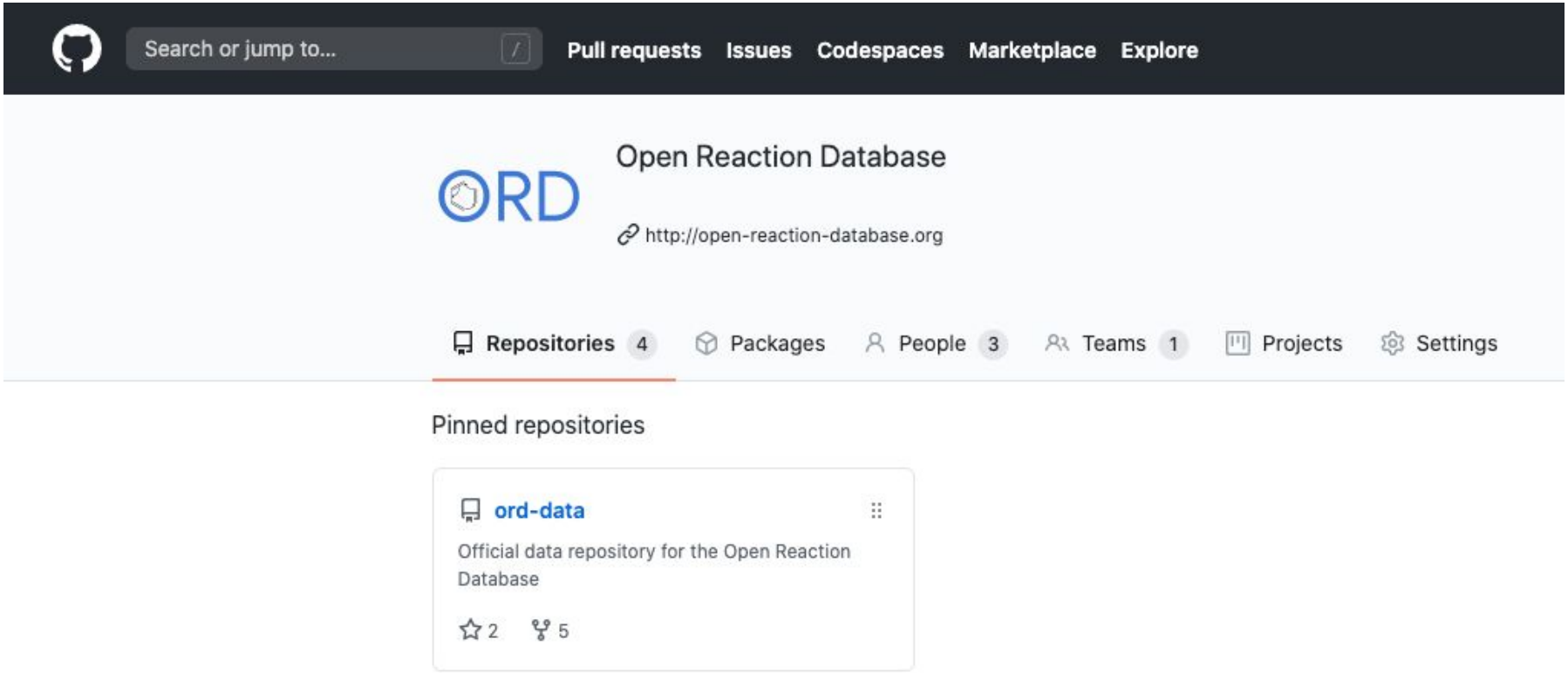
```
mol = message_helpers.mol_from_compound(compound)
if mol:
    return drawing.mol_to_svg(mol)
```



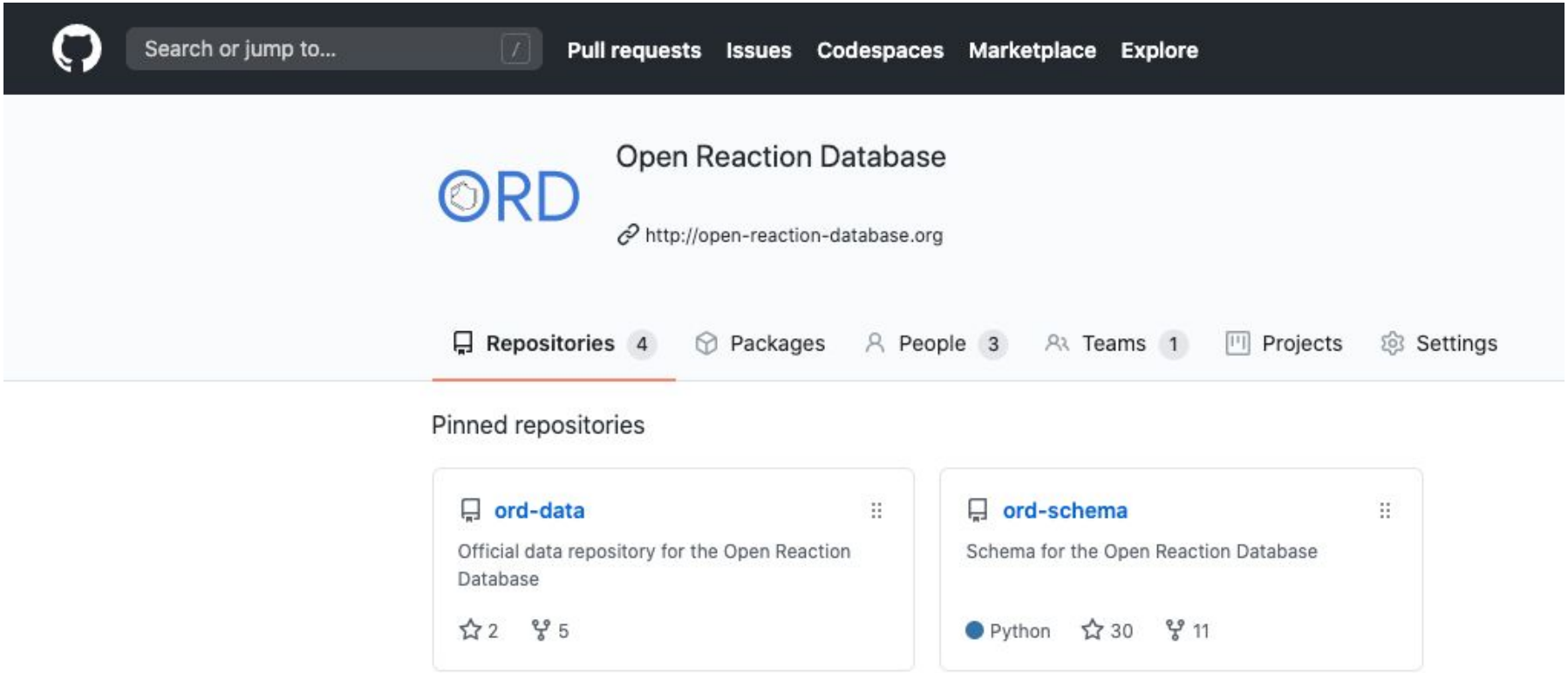
Where does the ORD live?



Where does the ORD live?

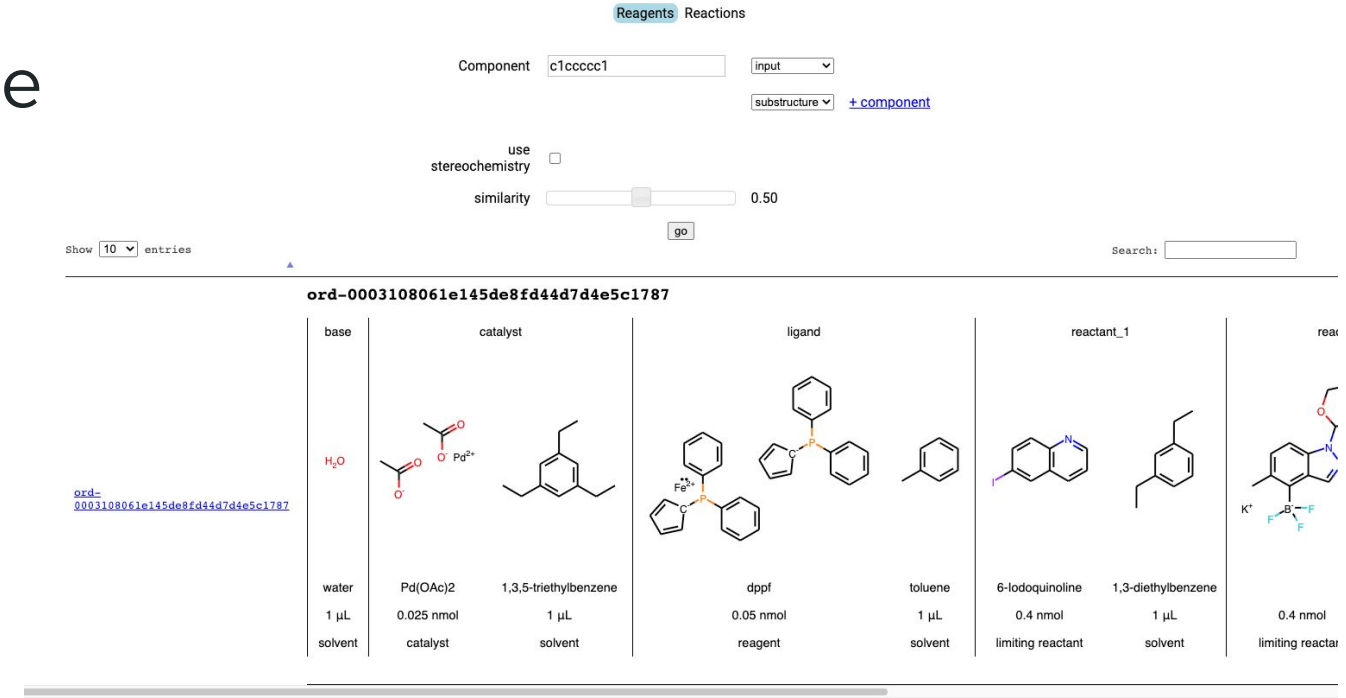


Where does the ORD live?



Search interface

Coming soon!



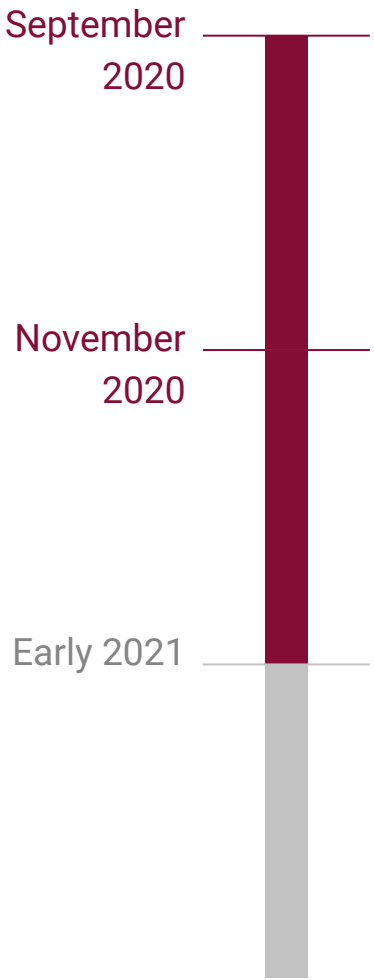
The interface makes heavy use of the RDKit postgres extension for compound and reaction searching by identity, substructure, and similarity.

Legal considerations

- Code is available under the Apache 2.0 license
- Data is available under the CC-BY-SA 4.0 license
- Terms of Use (drafted in kind by Google lawyers)

More info

Roadmap



Alpha testing

- Proof of concept for record creation and submission workflows
- Small group of invited contributors

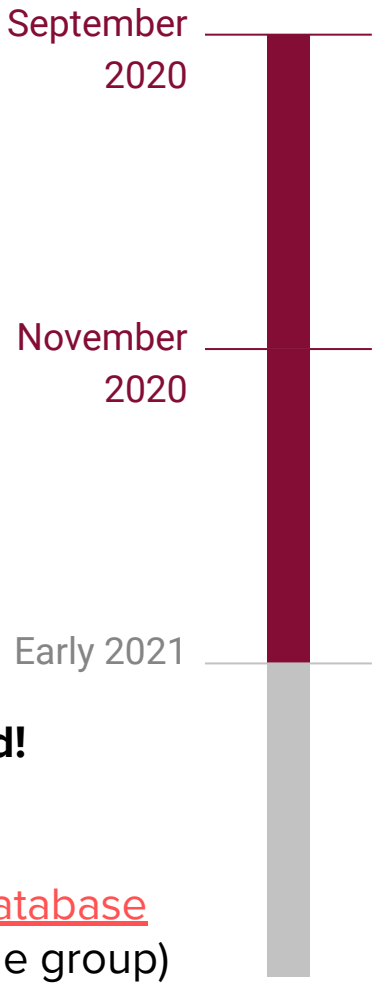
Beta testing / pre-launch expansion

- Expand the database prior to public launch
- Refine review and maintenance procedures
- Larger group of invited contributors

Public launch

- Open submissions to all contributors
- Invite specific contributions from industry and academia
- Solicit downstream use in ML and other applications

Roadmap



Alpha testing

- Proof of concept for record creation and submission workflows
- Small group of invited contributors

Beta testing / pre-launch expansion

- Expand the database prior to public launch
- Refine review and maintenance procedures
- Larger group of invited contributors

Public launch

- Open submissions to all contributors
- Invite specific contributions from industry and academia
- Solicit downstream use in ML and other applications

Stay tuned!

Join the
open-reaction-database
mailing list (Google group)

ORD