Standardized programmatic access to chemical information (Protocol Services)

WorldFAIR Chemistry
Evan Bolton, Ph.D.

ACS National Meeting, 27 March 2023

https://iupac.org/project/2022-029-1-024/

What are protocol services?

The IUPAC FAIR Chemistry Protocols are intended to support the bro
community in adoption of standards for programmatic exchange of
chemical data.

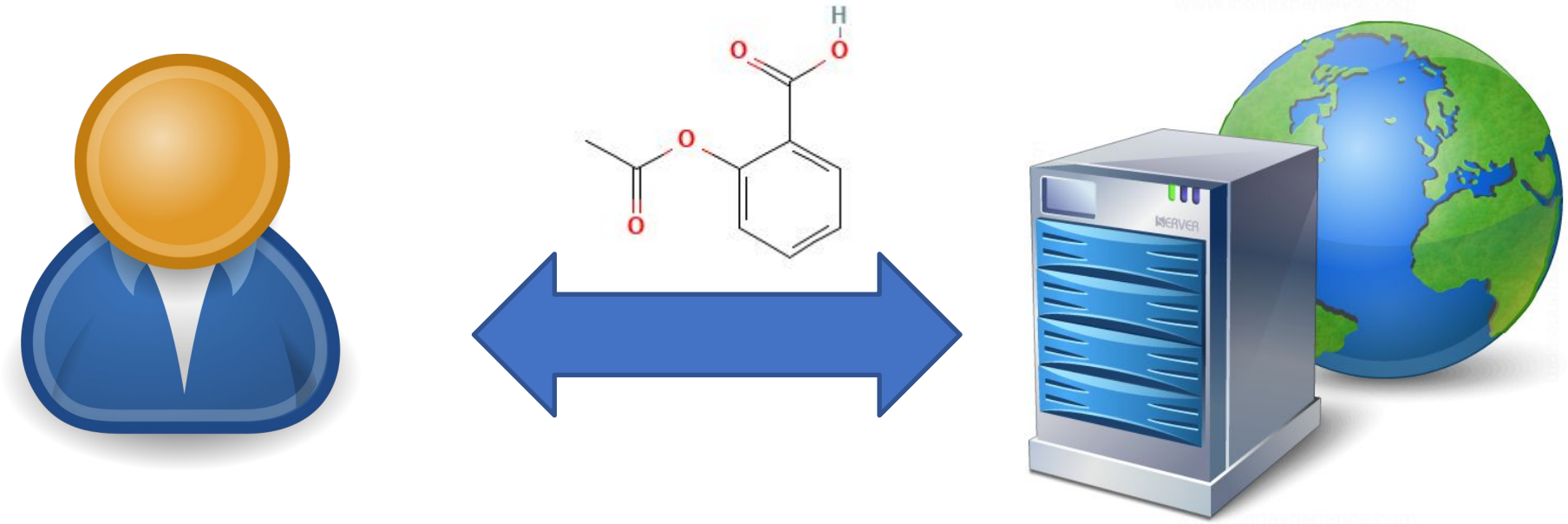
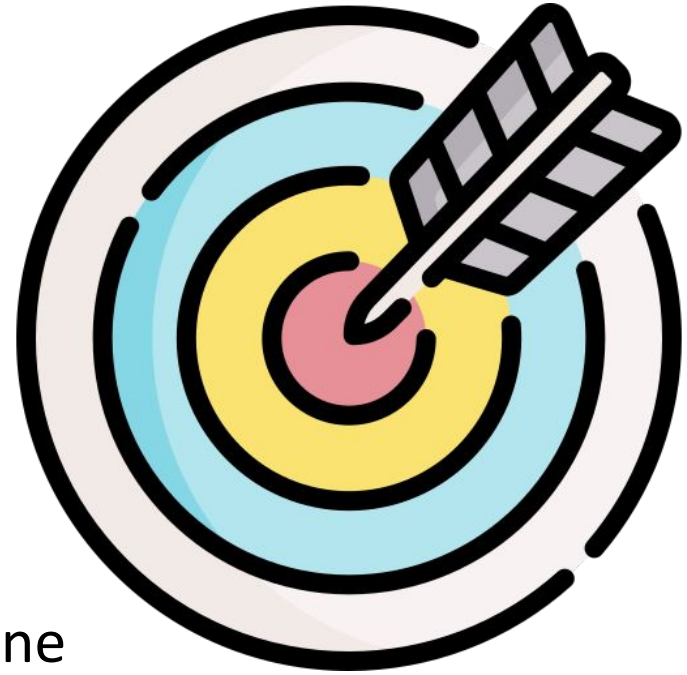


Image credits:
https://upload.wikimedia.org/wikipedia/commons/thumb/1/12/User_icon_2.svg/800px-User_icon_2.svg.png
https://pubchem.ncbi.nlm.nih.gov/image/imgsrv.fcgi?cid=2244&t=l
https://d1nhio0ox7pgb.cloudfront.net/_img/v_collection_png/512x512/shadow/server_earth.png

Protocol Services Aims

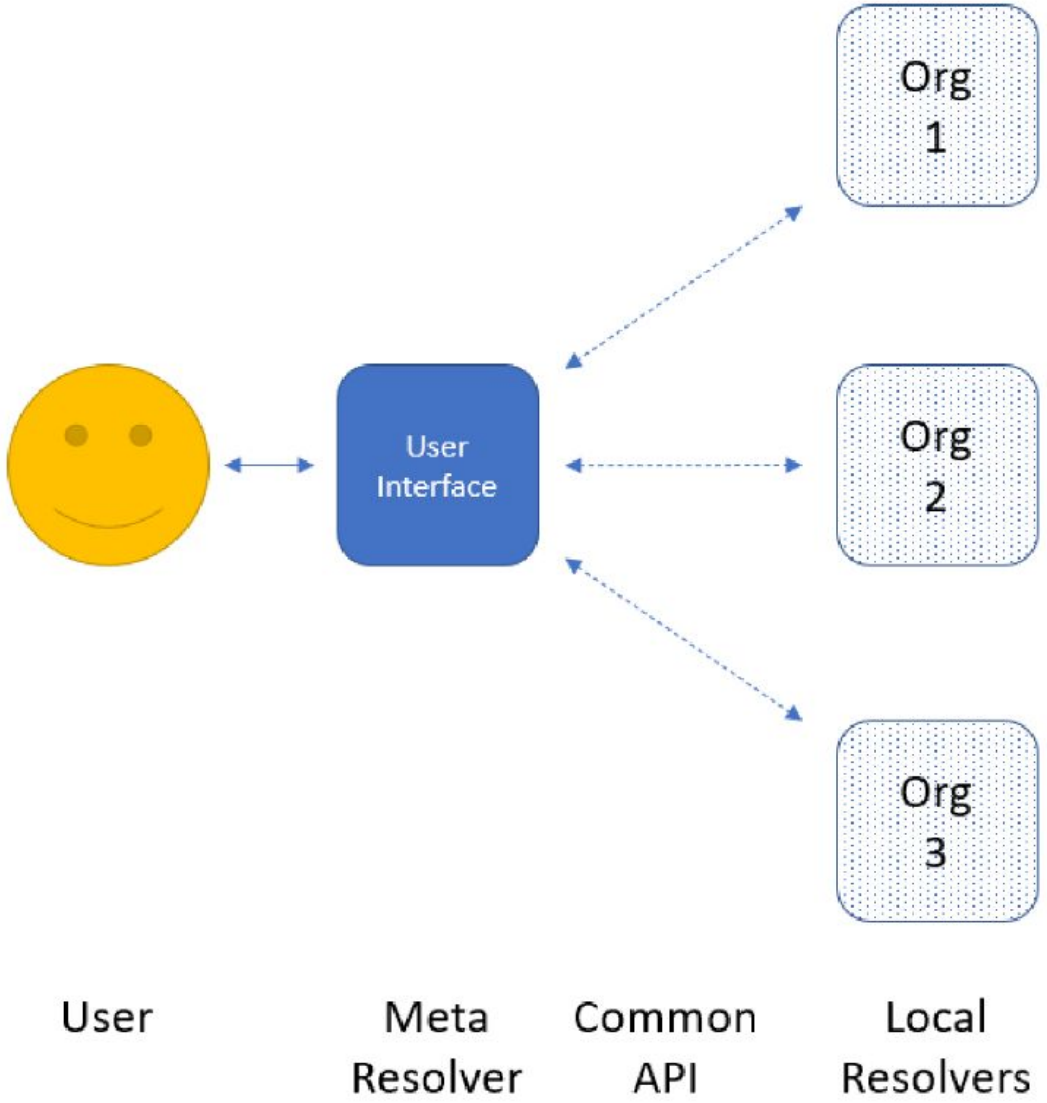


- **Common schema**: Define a common interface for chemistry websites resolving different chemical representations.
- **Shared data model**: The specification articulates shared data model for information exchange through an API that can be implemented by any system that manages chemical records.
- **Help find data**: Users will query registered resources with different scopes and capabilities at a general level to determine where they can find additional data and information.
- **Interprets chemical representations**: The model will cover a range of chemical representations, including structural diagrams, nomenclature and other linear notations.
- **Standard responses**: The protocol will provide a standard set of error codes for flagging ambiguous or conflicting representations.

Image credit:
https://cdn-icons-png.flaticon.com/512/4185/4185501.png

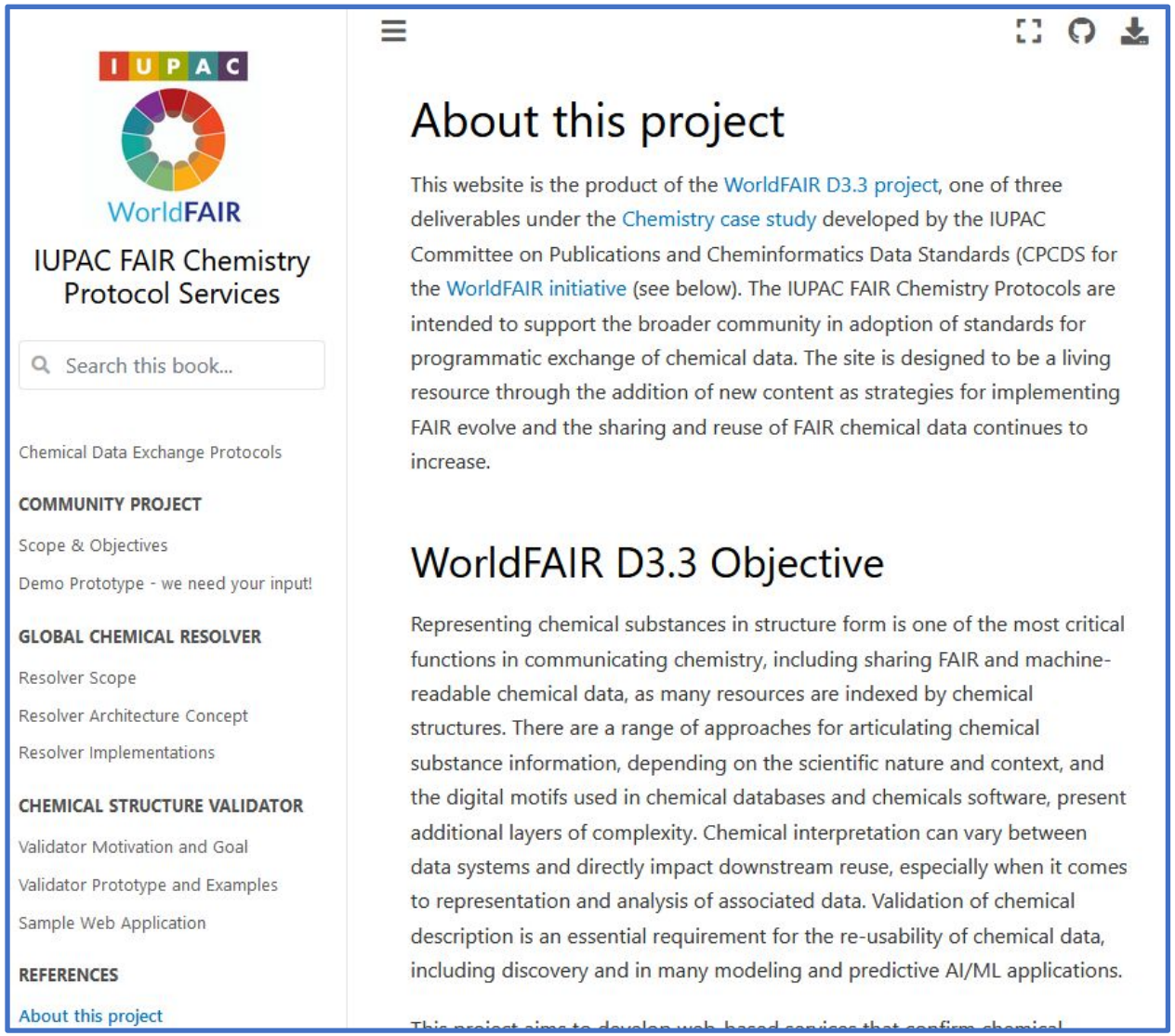
Why protocol services?

- Reduce barriers to access chemical information
- Standard schema for chemistry services
- Simplicity of same web-based API for chemistry websites
- Potential to use “all” (participating) chemistry websites



IUPAC FAIR Protocol Services

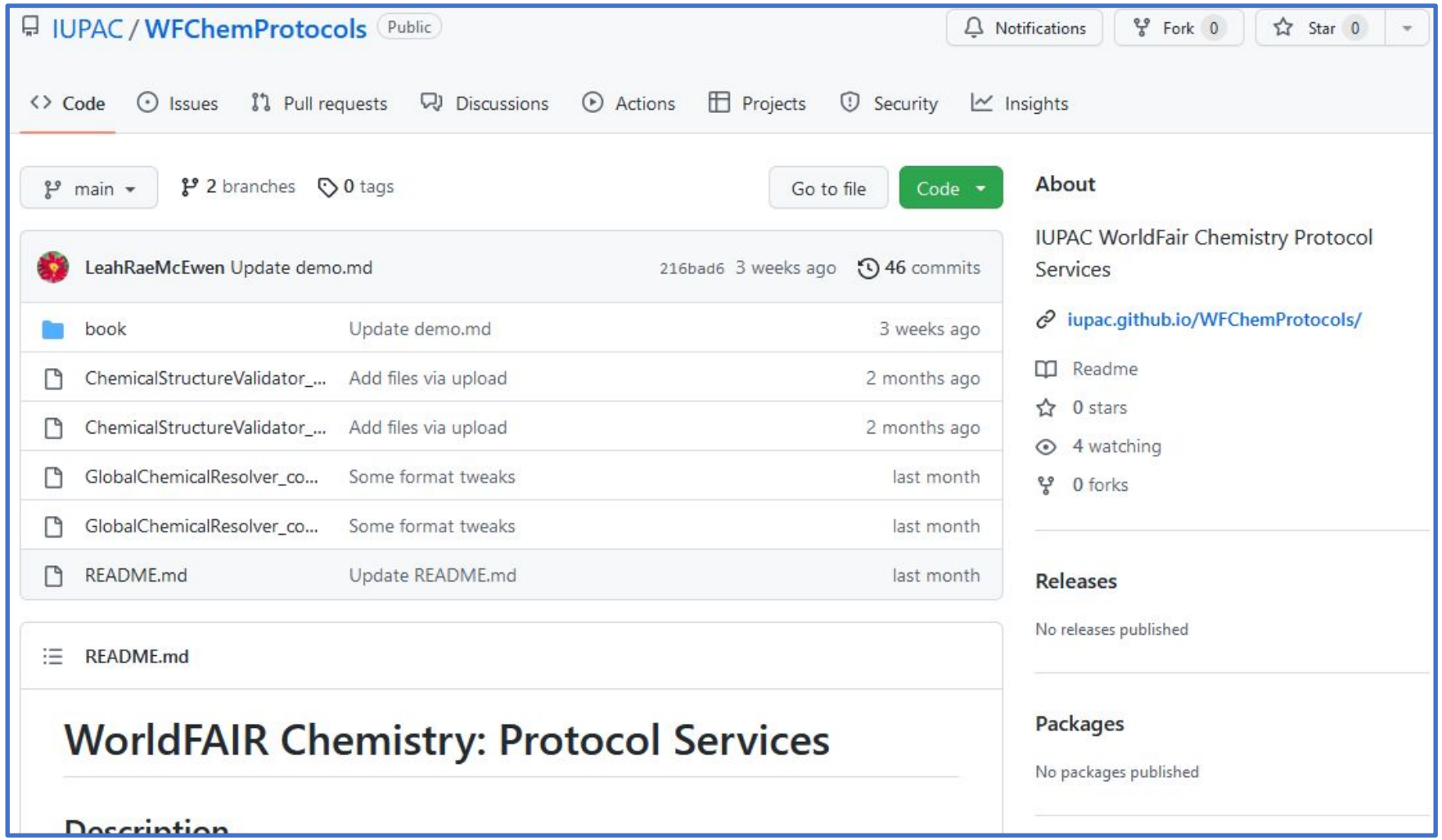
https://iupac.github.io/WFChemProtocols/about.html



https://github.com/IUPAC/WFChemProtocols

IUPAC FAIR Protocol Services

https://iupac.github.io/WFChemProtocols/about.html



https://github.com/IUPAC/WFChemProtocols

Introducing Protocol Services

https://iupac.github.io/WFChemProtocols/about.html

Two exemplar services:

- Chemical Structure Validator – Takes input of a chemical structure, interprets the structure, *determines if the structure is valid*, and returns a message to the user.
- Global Chemical Resolver – Takes input of a chemical structure, interprets the structure, *checks if the structure is found*, and returns a message to the user.

https://github.com/IUPAC/WFChemProtocols

Why target these two services?

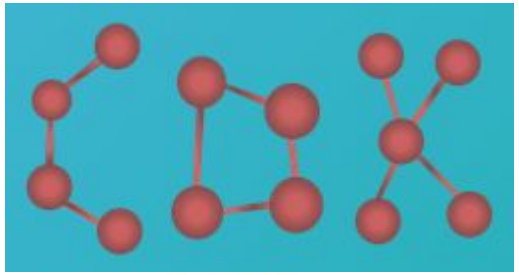
Chemical Structure Validator
Global Chemical Resolver

Many chemistry data providers, chemistry toolkits, chemical drawing packages, chemistry application providers

Each has its own way of chemical structure handling

Improve machine understanding of chemical information

Open new opportunities for cross-science data access



Examples of chemistry-oriented resources. Not an endorsement.

Example protocols implementation on PubChem

Schema available (XML):

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver_data.xsd

```
<xs:element name="Fault">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="Code" type="xs:string"/>
      <xs:element name="Message" type="xs:string"/>
      <xs:element name="Details" type="xs:string" minOccurs="0" maxOccurs="unbounded"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="Result">
                                    Occurs="0" maxOccurs="unbounded"/>
                                    type="xs:string"/>
      <xs:element name="ResourceURL" type="xs:string"/>
      <xs:element name="ResourceIdentifier" type="xs:string"/>
      <xs:element name="ResourceIdentifierType" type="xs:string" minOccurs="0"/>
      <xs:element name="URL" type="xs:string" minOccurs="0"/>
      <xs:element name="IUPACName" type="xs:string" minOccurs="0"/>
      <xs:element name="SMILES" type="xs:string" minOccurs="0"/>
      <xs:element name="InChI" type="xs:string" minOccurs="0"/>
      <xs:element name="InChIKey" type="xs:string" minOccurs="0"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
</xs:schema>
```

Currently, using XML schema but intend to use JSON schema approaches

JSON allowed options (JSON):

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi

```
{
  "Result": {
    "ServiceDetails": [
      {
        "Resource": "PubChem",
        "ResourceURL": "https://pubchem.ncbi.nlm.nih.gov",
        "ResolverURL": "https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi",
        "AvailableInputs": {
          "SDF": true,
          "SMILES": true,
          "InChI": true,
          "InChIKey": true,
          "PNG": false,
          "Name": true
        },
        "AvailableOutputs": {
          "IUPACName": true,
          "SMILES": true,
          "InChI": true,
          "InChIKey": true,
          "ResourceIdentifier": true,
          "RecordURL": true,
          "ImageURL": true
        }
      }
    ]
  }
}
```

(Global) Chemical Resolver in action

*Takes input of a chemical structure, interprets the structure, checks if the structure is **found**, returns a message to the user*

Example Chemical Resolver U

(different representations of butane u

Yellow is constant and specific to each resource

https://pubchem.ncbi.nlm.nih.gov/resolver/res

https://pubchem.ncbi.nlm.nih.gov/resolver/res

https://pubchem.ncbi.nlm.nih.gov/resolver/res

https://pubchem.ncbi.nlm.nih.gov/resolver/res

(note that the full InChI is URL-encode


```
{
  "Result": {
    "Match": [
      {
        "Resource": "PubChem",
        "ResourceURL": "https://pubchem.ncbi.nlm.nih.gov",
        "ResourceIdentifier": "7843",
        "ResourceIdentifierType": "CID",
        "RecordURL": "https://pubchem.ncbi.nlm.nih.gov/compound/7843",
        "ImageURL": "https://pubchem.ncbi.nlm.nih.gov/image/imgsrv.fcgi?t=l&cid=7843",
        "IUPACName": "butane",
        "SMILES": "CCCC",
        "InChI": "InChI=1S/C4H10/c1-3-4-2/h3-4H2,1-2H3",
        "InChIKey": "IJDNQMDRQITEOD-UHFFFAOYSA-N"
      }
    ]
  }
}
```


Each of these URLs produces the same JSON output, from different representations of the input structure of butane (SMILES, name, InChIKey, InChI, respectively)

Not found example: https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?smiles=CNOS

(Global) Chemical Resolver in action

*Takes input of a chemical structure, interprets the structure, checks if the structure is **found**, returns a message to the user*

Example Chemical Resolver URLs:

(different representations of butane using SMILES, name, InChIKey, InChI, respectively)

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?smiles=CCCC

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?name=butane

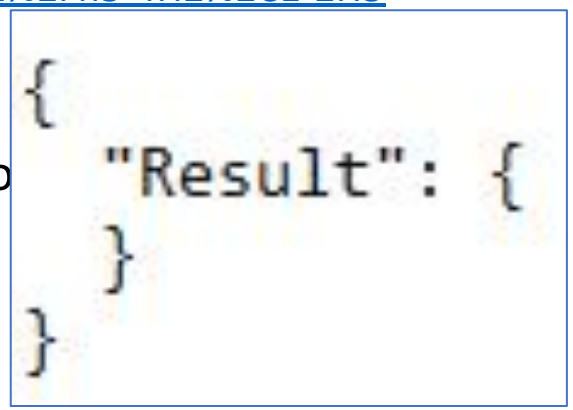
https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?inchikey=IJDNQMDRQITEOD-UHFFFAOYSA-N

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?inchi=InChI%3D1S%2FC4H10%2Fc1-3-4-2%2Fh3-4H2%2C1-2H3

(note that the full InChI is URL-encoded because it contains special characters)

Each of these URLs produces the same JSON output, from different representatio
structure of butane (SMILES, name, InChIKey, InChI, respectively)

Not found example: https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?smiles=CNOS



Chemical structure validator in action

*Takes input of a chemical structure, interprets the structure, checks if the structure is **valid**, returns a message to the user*

Example Chemical Structure Validator URLs:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&smiles=CCCC

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&smiles=C/C=C/[C@H]1[C@H](O1)C

Invalid carbon valence:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&smiles=CC(C)(C)(C)C

Invalid isotope:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&smiles=C%5b5H%5d

Output an image:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&format=png&smiles=C%5bC@H%5d(CCCC(C)C)%5bC@H%5d1CC%5bC@@H%5d2%5bC@@%5d1(CC%5bC@H%5d3%5bC@H%5d2CC=C4%5bC@@%5d3(CC%5bC@@H%5d(C4)O)C)C

Chemical structure validator in acti

Takes input of a chemical structure, interprets the str
*structure is **valid**, returns a message to the user*

Example Chemical Structure Validator URLs:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&smile

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&smile

Invalid carbon valence:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&smile

Invalid isotope:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&smile

Output an image:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&forma
C(C)C)%5bC@H%5d1CC%5bC@@H%5d2%5bC@@%5d1(CC%5bC@H%5d3%5bC@H%5d2
H%5d(C4)O)C)C


```
{
  "Result": {
    "Message": "Structure is valid",
    "Statistics": [
      {
        "Type": "DefinedAtomStereo",
        "Value": "0"
      },
      {
        "Type": "UndefinedAtomStereo",
        "Value": "0"
      },
      {
        "Type": "DefinedBondStereo",
        "Value": "0"
      },
      {
        "Type": "UndefinedBondStereo",
        "Value": "0"
      },
      {
        "Type": "HeavyAtoms",
        "Value": "4"
      },
      {
        "Type": "IsotopeAtoms",
        "Value": "0"
      },
      {
        "Type": "CovalentUnits",
        "Value": "1"
      }
    ]
  }
}
```

Chemical structure validator in actio

Takes input of a chemical structure, interprets the struc
*structure is **valid**, returns a message to the user*

Example Chemical Structure Validator URLs:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&smiles=C

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&smiles=C/

Invalid carbon valence:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&smiles=C

Invalid isotope:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&smiles=C

Output an image:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&format=p
C(C)C)%5bC@H%5d1CC%5bC@@H%5d2%5bC@@%5d1(CC%5bC@H%5d3%5bC@H%5d2CC=
H%5d(C4)O)C)C


```
{
  "Result": {
    "Message": "Structure is valid",
    "Statistics": [
      {
        "Type": "DefinedAtomStereo",
        "Value": "2"
      },
      {
        "Type": "UndefinedAtomStereo",
        "Value": "0"
      },
      {
        "Type": "DefinedBondStereo",
        "Value": "1"
      },
      {
        "Type": "UndefinedBondStereo",
        "Value": "0"
      },
      {
        "Type": "HeavyAtoms",
        "Value": "7"
      },
      {
        "Type": "IsotopeAtoms",
        "Value": "0"
      },
      {
        "Type": "CovalentUnits",
        "Value": "1"
      }
    ]
  }
}
```

Chemical structure validator in action

*Takes input of a chemical structure, interprets the structure, checks if the structure is **valid**, returns a message to the user*

Example Chemical Structure Validator URLs:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&smiles=CCCC

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&smiles=C/C=C/[C@H]1[C@H](O1)C

Invalid carbon valence:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&smiles=CC(C)(C)(C)C

Invalid isotope:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?act

Output an image:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?act
C(C)C)%5bC@H%5d1CC%5bC@@H%5d2%5bC@@%5d1(CC
H%5d(C4)O)C)C


```
{
  "Fault": {
    "Code": "Invalid",
    "Message": "Structure is not valid",
    "Details": [
      "Record 0: Warning: \"pcData/pubchem_valence.cpp\", line 290: Detected
illegal valence for element \"C\": 5 sigma bonds, 0 pi bonds, 0 charge",
      "Exception: Valence validation failed"
    ]
  }
}
```

Chemical structure validator in action

*Takes input of a chemical structure, interprets the structure, checks if the structure is **valid**, returns a message to the user*

Example Chemical Structure Validator URLs:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cg

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cg

Invalid carbon valence:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cg

Invalid isotope:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&smiles=C%5b5H%5d

Output an image:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&format=png&smiles=C%5bC@H%5d(CCCC(C)C)%5bC@H%5d1CC%5bC@@H%5d2%5bC@@%5d1(CC%5bC@H%5d3%5bC@H%5d2CC=C4%5bC@@%5d3(CC%5bC@@H%5d(C4)O)C)C


```
{
  "Fault": {
    "Code": "Invalid",
    "Message": "Structure is not valid",
    "Details": [
      "Record 0: Info: \"OpenEye/pubchem_compound.cpp\", line 3121: Atom ID \"2\" has illegal isotope (5) for atomic number 1 (\"H\")",
      "Exception: Element validation failed"
    ]
  }
}
```

Chemical structure validator in action

*Takes input of a chemical structure, interprets the structure, checks if the structure is **valid**, returns a message to the user*

Example Chemical Structure Validator URLs:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&smiles=CCCC

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&smiles=C/C=C/[C@H]1[C@H](O1)C

Invalid carbon valence:

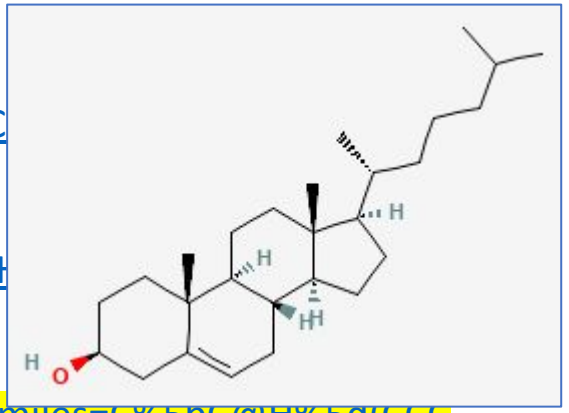
https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&smiles=CC(C)(C

Invalid isotope:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&smiles=C%5b5H

Output an image:

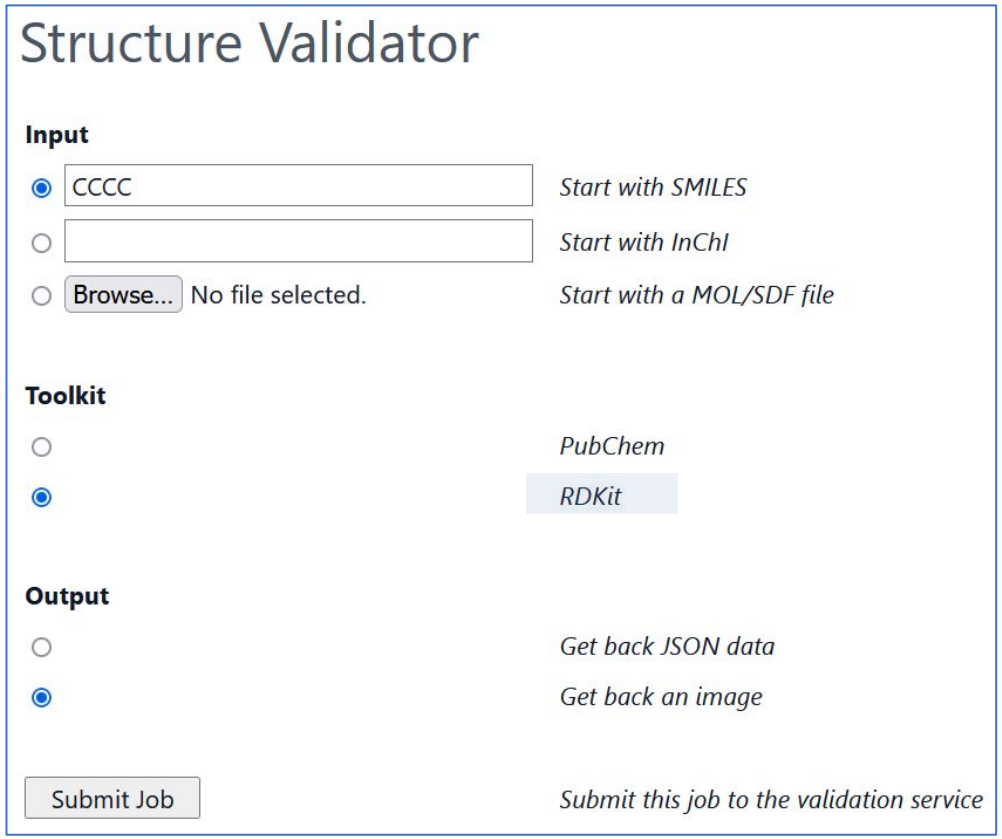
https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&format=png&smiles=C%5bC@H%5d(CCC(C)C)%5bC@H%5d1CC%5bC@@H%5d2%5bC@@%5d1(CC%5bC@H%5d3%5bC@H%5d2CC=C4%5bC@@%5d3(CC%5bC@@H%5d(C4)O)C)C

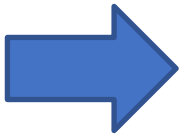


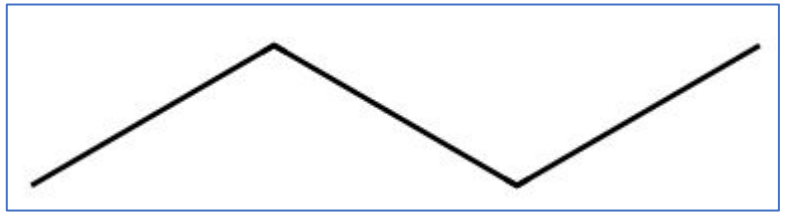
Example protocols implementation on PubChem

Interactive interface available:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=input_form



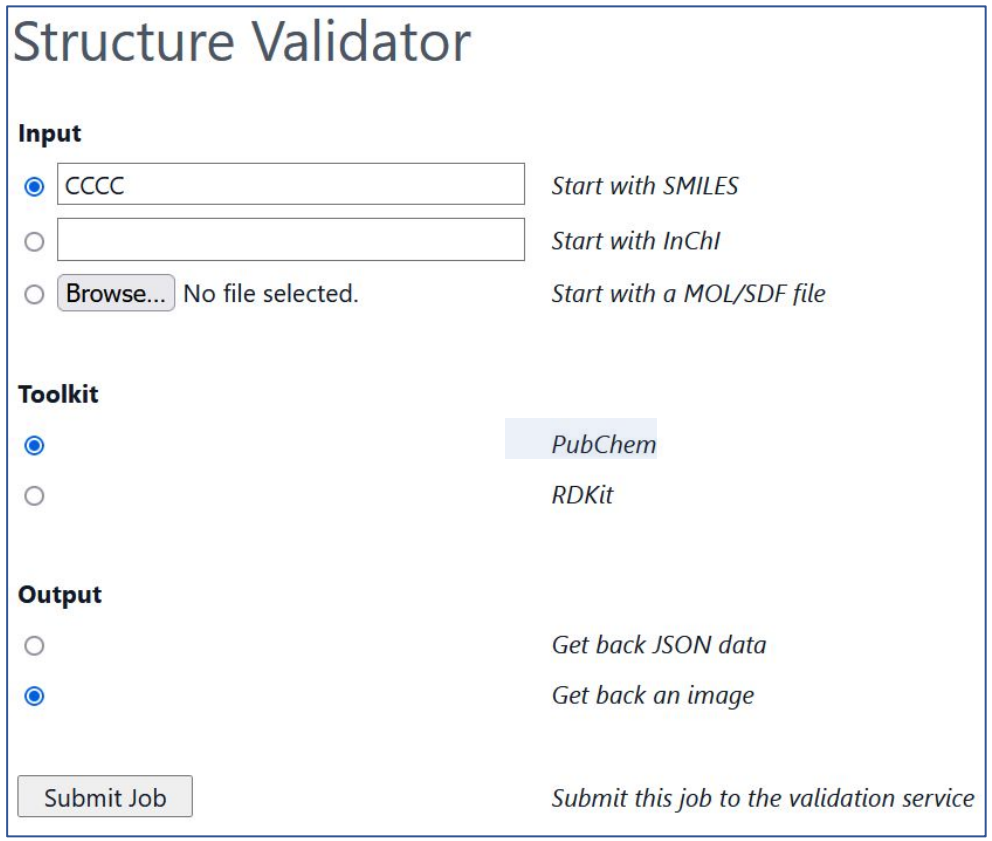


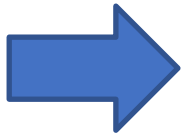


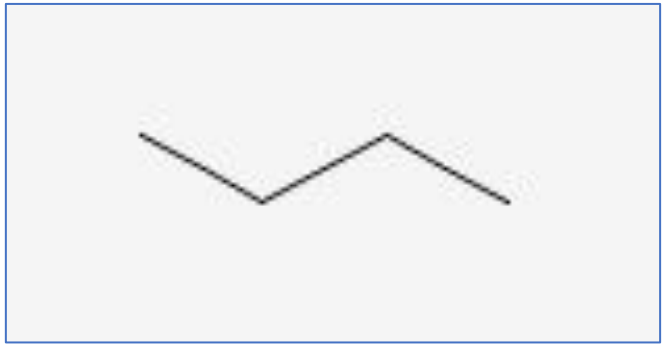
Example protocols implementation on PubChem

Interactive interface available:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=input_form





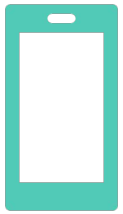


Protocol Services Stakeholders

https://iupac.github.io/WFChemProtocols/about.html

The goal is to support the following stakeholders and functions:

- **Scientific database providers that support chemical data**: organizations that would implement and provide these web services to the public.



- **Scientific application and web service developers**: people writing applications that would directly use the web services described here.

- **Chemists and other scientists**: end users, who would be accessing these web services indirectly, through a chemical drawing program or ELN notebook or such.

- **Chemical toolkit developers**: people writing chemical toolkits that could be used to implement the web services described here.

https://github.com/IUPAC/WFChemProtocols

Timeline for Protocol Services

https://iupac.github.io/WFChemProtocols/about.html

- White paper(s) created
- Github and ".io" websites are setup
- Demonstration implementation available
- Expanding involvement to key stakeholders

https://github.com/IUPAC/WFChemProtocols

Next major milestone: **Oct. 2023**

Deliver an MVP ("minimum viable product") for:

Chemical Structure Validator and Global Chemical Resolver

What is needed for Protocol Services?

- Expand community involvement:
  - **Scientific database providers that support chemical data**
  - **Scientific application and web service developers**
  - **Chemists and other scientists**
  - **Chemical toolkit developers**
- Develop a standard set of chemical structure error messages
- Develop Global Chemical Resolver “Meta Resolver” protocol
- Migrate XML schema to JSON schema
- Agree on protocol service parameters and functionality

How to get involved with Protocol Services?

Try out the demo:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=input_form

Provide some feedback:

https://forms.gle/6m5Jrfj4gAJTuYEL9

GitHub site Discussions (and Issue tracker):

https://github.com/IUPAC/WFChemProtocols/discussions

IUPAC WorldFAIR Chemistry Acknowledgments

Evan Bolton

Paul Thiessen

Tony Williams

<u>IUPAC WorldFAIR</u>

Leah McEwen

Fatima Mustafa

<u>Special thanks</u>

Sunghwan Kim

Stuart Chalk

This research was supported [in part] by the Intramural Research Program of the NIH, National Library of Medicine

WorldFAIR

Standardized Programmatic Access to Chemical Information (Protocol Service)

WorldFAIR Chemistry

ACS National Meeting, 27 March 2023

Discussion

Shared notes link:

https://bit.ly/ACSChemNotes

WorldFAIR "Global cooperation on FAIR data policy and practice" is funded by the EC HORIZON-WIDERA-2021-ERA-01-41 Coordination and Support Action under Grant Agreement No. 101058393.

Standardized Programmatic Access to Chemical Information (Protocol Service)

WorldFAIR Chemistry

ACS National Meeting, 27 March 2023

Contact & Follow US

iupac.org

@FAIRChemistry
@iupac

tinyurl.com/iupacchan

FAIRChemistry@iupac.org

@iupac.org

FAIRChemistry Community

WorldFAIR “Global cooperation on FAIR data policy and practice” is funded by the EC HORIZON-WIDERA-2021-ERA-01-41 Coordination and Support Action under Grant Agreement No. 101058393.

How to get involved with Protocol Services?

Try out the demo:

https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=input_form

Provide some feedback:

https://forms.gle/6m5Jrfj4gAJTuYEL9

GitHub site Discussions (and Issue tracker):

https://github.com/IUPAC/WFChemProtocols/discussions