



Project Title	Global cooperation on FAIR data policy and practice
Project Acronym	WorldFAIR
Grant Agreement No	101058393
Instrument	HORIZON-WIDERA-2021-ERA-01
Topic, type of action	HORIZON-WIDERA-2021-ERA-01-41 HORIZON Coordination and Support Actions
Start Date of Project	2022-06-01
Duration of Project	24 months
Project Website	http://worldfair-project.eu

D3.3 Utility services for Chemistry Standards

Work Package	WP03 - Chemistry
Lead Author (Org)	Paul Thiessen (U.S. National Center for Biotechnology Information)
Contributing Author(s) (Org)	Evan Bolton (U.S. National Center for Biotechnology Information), Antony Williams (U.S. Environmental Protection Agency), Leah McEwen (Cornell University), Fatima Mustafa (International Union of Pure and Applied Chemistry)

Due Date	29.02.2024
Date	06.12.2023
Version	1.1 DRAFT NOT YET APPROVED BY THE EUROPEAN COMMISSION
DOI	https://doi.org/10.5281/zenodo.10514901 (this version 1.1) https://doi.org/10.5281/zenodo.10289785 (concept DOI)

Dissemination Level

<input checked="" type="checkbox"/>	PU: Public
<input type="checkbox"/>	PP: Restricted to other programme participants (including the Commission)
<input type="checkbox"/>	RE: Restricted to a group specified by the consortium (including the Commission)
<input type="checkbox"/>	CO: Confidential, only for members of the consortium (including the Commission)

Versioning and contribution history

Version	Date	Authors	Notes
0.7	17.10.2023	Paul Thiessen	First draft
0.8	06.11.2023	Leah McEwen	Draft for internal review
0.9	28.11.2023	All authors	Draft incorporating review comments
1.0	06.12.2023	All authors	Content ready
1.1	06.12.2023	Leah McEwen, Laura Molloy (editor)	Final proofed version including attribution details

Disclaimer

WorldFAIR has received funding from the European Commission's WIDERA coordination and support programme under Grant Agreement no. 101058393. The content of this document does not represent the opinion of the European Commission, and the European Commission is not responsible for any use that might be made of such content.

Abbreviations and Acronyms

AI	Artificial Intelligence
API	Application Programming Interface
CCD	CompTox Chemicals Dashboard
CGI	Common Gateway Interface
ChEMBL	Chemical database of the European Molecular Biology Laboratory
CIF	Crystallographic Information File
CSD	Cambridge Structural Database
ELN	Electronic Laboratory Notebook
EPA	US Environmental Protection Agency
FAIR	Findable, Accessible, Interoperable, Reusable
GSRS	Global Ingredient Archival System
HTTP	HyperText Transfer Protocol
InChI	IUPAC International Chemical Identifier
InChIKey	International Chemical Identifier key
IUPAC	International Union of Pure and Applied Chemistry
JSON	JavaScript Object Notation
LLM	Large Language Model
ML	Machine Learning
MOL	Molfile extension (molecular file format)
PDB	Protein DataBank
PNG	Portable Network Graphics
SDF	Structure-Data File
SMILES	Simplified Molecular-Input Line-Entry System
URL	Universal Resource Locator
W3C	World Wide Web Consortium
XML	eXtensible Markup Language

Executive summary

The International Union of Pure and Applied Chemistry (IUPAC) has initiated a community project through the WorldFAIR initiative to define a common protocol for programmatic exchange of chemical representations. Representing chemical substances in the form of distinct chemical structures is fundamental to communicating chemical information. Validation of chemical structure description is an essential requirement for the re-usability of FAIR chemical data. The outcome of this work includes a specification that articulates a shared data model for chemical information exchange through an API that can be implemented by any system that manages chemical structure records. This deliverable outlines a conceptual framework and provides a demo prototype to engage community input and adoption.

This deliverable aims to describe criteria for web-based services that participating organisations can implement based on their existing and/or preferred technologies (e.g., toolkits, programming languages). The services are intended to confirm chemical identity and provide real time feedback on the machine-readability of chemical data and metadata representations based on IUPAC standard rule sets and community best practices. The goal is to support a range of stakeholders engaging in chemical data exchange online, including providers of chemical databases, curators of chemical repositories, chemistry application developers, chemical toolkit developers, and researchers sharing, searching and analysing chemical information programmatically. The initial specification focuses on resolving chemical entities and validating chemical structure representations.

The overarching goal of the WorldFAIR Chemistry Work Package (WP03) is to support the use of chemical data standards in research and curation workflows, between and across disciplines. This will enable downstream data reuse through provision of practical direction and resources. Other deliverables developed under the WorldFAIR Chemistry (WP03) case study further demonstrate and facilitate the use of chemistry data standards, including a framework that can be used by policymakers and developers of services and tools to support FAIR reporting of chemical data (D3.1 Digital guidance), and a digital ‘cookbook’ of interactive recipes demonstrating how to handle digital chemical data (D3.2 Training package).

WP03 activities are coordinated through IUPAC, the world authority on chemical nomenclature and terminology that constitute a common global language for communicating chemistry. In the context of the formal IUPAC process for reviewing and adopting consensus standards in chemistry, this work should be regarded as provisional guidance. Complete review and adoption of standards through IUPAC to reach the status of “Recommendation”, which has a specific meaning in the IUPAC lexicon, will occur after WP03 is complete.

This work was supported [in part] by the U.S. National Center for Biotechnology Information of the National Library of Medicine (NLM), U.S. National Institutes of Health.

This manuscript has been reviewed by the Center for Computational Toxicology and Exposure, United States Environmental Protection Agency and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the Agency nor does mention of trade names or commercial products constitute endorsement or recommendation for use. The authors declare no conflict of interest.

Table of contents

Executive summary	4
1. Introduction	6
2. Use cases	7
2.1 Global search	8
2.2 Cross-exchange	9
2.3 Validation	9
2.4 Interoperability	10
3. Common protocol	11
3.1 Global chemical information resolver	11
3.2 Chemical structure representation validator	12
4. Global chemical information resolver	13
4.1 Architecture concept	13
4.1.1 Common data model	15
4.2 Implementation	15
4.2.1 Local resolver	15
4.2.2 Meta resolver	20
5. Chemical structure representation validator	20
5.1 Prototype and examples	21
5.2 Sample web application	26
6. Available infrastructure	26
7. Community engagement	29
7.1 Protocol adoption	29
7.2 Specification development	30
8. Conclusion	31
9. Appendices	32
9.1 Oral presentation: Standardised programmatic access to chemical information (Protocol Services)	32
9.2 Workshop Session: Doc-a-thon: Chemical representation best practices for humans and machines	32
9.3 Poster presentation: IUPAC WorldFAIR Chemistry: Managing Chemical Data Digitally	32
9.4 Structure validator demo	32
Bibliography	33

1. Introduction

Many different (scientific) disciplines, (sub)domains, organisations and users need to access and integrate chemical information, including from oceanography, meteorology, astronomy, metabolomics, proteomics, bioinformatics, geology, biomedical informatics, chemical vendors, pharmaceutical companies, regulatory agencies, scientific publishers, and many others. A substantial barrier to this chemical information exchange is that each chemical information resource that indexes chemical structures lacks standardised system-to-system interoperability. This lack of interoperability means that each chemistry-oriented resource interprets chemical structure data differently and, as a result, heterogeneity across databases and resources exists and can cause confusion when disseminating chemical data.

Representing chemical substances in structure form is one of the most critical functions in communicating chemistry, including sharing FAIR and machine-readable chemical data. There are a range of approaches for articulating chemical substance information depending on the scientific nature and context. Different digital motifs and models used in chemical structure databases and chemical structure software present additional layers of complexity. Chemical interpretation can thus vary between data systems and directly impact downstream reuse, including analysis of associated data. Therefore validation of chemical description is an essential requirement for re-usability of chemical data, from discovery to modelling and predictive artificial intelligence or machine learning (AI/ML) applications.

We aim to describe criteria for web-based services to confirm chemical identity and provide real-time feedback on the machine readability of chemical data and metadata representation based on IUPAC standard rule sets and community best practices. The intent is that chemical data resources and software can implement the specified protocols based on their existing/preferred toolkits, programming languages, etc.

The goal is to support the following stakeholders and functions:

- **Chemical database providers** implementing and providing these web services to the public.
- **Chemistry application developers** that would directly use these web services.
- **Chemists and other researchers** accessing these web services indirectly, through a chemical drawing program, electronic laboratory notebook (ELN) or other applications.
- **Chemical toolkit developers** whose toolkits could be used to implement the web services described here.

This report describes a programmatic interface for services that provide the functionality described. A prototype implementation has been developed as a demo. It is not intended as a final product but to encourage participating organisations to implement the services using whatever technology is convenient to them.

2. Use cases

In the greater environment of Open Science and data-driven approaches to global challenges, access to chemical information is critical for many cross-domain research areas. Ascertaining chemical composition is a common need in characterising research samples in many fields, as has been explored in WorldFAIR Deliverable 3.1, ‘Digital recommendations for Chemistry FAIR data policy and practice’.¹ Further integration with other chemical property data can enrich the scientific knowledge base and enable additional analyses and activities; for example, environmental monitoring and chemicals risk assessment.

Most chemical data resources^{2,3,4,5} and reported chemical data⁶ are described and organised around chemical entities⁷ (among other possible information). Chemicals are often classified by composition and molecular characteristics.⁸ Naming conventions for organising and searching chemical information very often incorporate these aspects, particularly representations of chemical structure.^{9,10,11} There are various approaches to interpreting underlying chemical theories associated with molecular structures. How these are implemented in models may vary between systems, depending on scientific and/or business scope, creating substantial challenges for system-to-system interoperability and chemical structure data exchange. See further discussion and references in WorldFAIR Deliverable 3.1, ‘Digital recommendations for Chemistry FAIR data policy and practice’.¹²

¹ McEwen, L., & Bruno, I. (2023). WorldFAIR Project (D3.1) Digital recommendations for Chemistry FAIR data policy and practice (Version 1). Zenodo. <https://doi.org/10.5281/zenodo.7887282>

² Hastings J, Owen G, Dekker A, et al. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research*. 2016 Jan;44(D1):D1214-9. DOI: 10.1093/nar/gkv1031

³ Sinclair G, Thillainadarajah I, Meyer B, et al. Wikipedia on the CompTox Chemicals Dashboard: Connecting Resources to Enrich Public Chemical Data. *J Chem Inf Model*. 2022 Oct 24;62(20):4888-4905. DOI: 10.1021/acs.jcim.2c00886

⁴ Jacobs A, Williams D, Hickey K, et al. CAS Common Chemistry in 2021: Expanding Access to Trusted Chemical Information for the Scientific Community. *J Chem Inf Model*. 2022 Jun 13;62(11):2737-2743. DOI: 10.1021/acs.jcim.2c00268

⁵ Richard AM, Huang R, Waidyanatha S, et al. The Tox21 10K Compound Library: Collaborative Chemistry Advancing Toxicology. *Chem Res Toxicol*. 2021 Feb 15;34(2):189-216. DOI: 10.1021/acs.chemrestox.0c00264

⁶ Lowe CN, Williams AJ. Enabling High-Throughput Searches for Multiple Chemical Data Using the U.S.-EPA CompTox Chemicals Dashboard. *J Chem Inf Model*. 2021 Feb 22;61(2):565-570. DOI: 10.1021/acs.jcim.0c01273

⁷ Heller SR, McNaught A, Pletnev I, et al. InChI, the IUPAC International Chemical Identifier. *J Cheminform*. 2015 May 30;7:23. DOI: 10.1186/s13321-015-0068-4

⁸ Williams AJ, Gaines LGT, Grulke CM, et al. Assembly and Curation of Lists of Per- and Polyfluoroalkyl Substances (PFAS) to Support Environmental Science Research. *Front Environ Sci*. 2022 Apr 5;10:1-13. DOI: 10.3389/fenvs.2022.850019

⁹ Bento AP, Hersey A, Félix E, et al. An open source chemical structure curation pipeline using RDKit. *J Cheminform*. 2020 Sep 1;12(1):51. DOI: 10.1186/s13321-020-00456-1

¹⁰ Karapetyan K, Batchelor C, Sharpe D, et al. The Chemical Validation and Standardization Platform (CVSP): large-scale automated validation of chemical structure datasets. *J Cheminform*. 2015 Jun 19;7:30. DOI: 10.1186/s13321-015-0072-8

¹¹ Hähnke VD, Kim S, Bolton EE. PubChem chemical structure standardisation. *J Cheminform*. 2018 Aug 10;10(1):36. DOI: 10.1186/s13321-018-0293-8

¹² McEwen, L., & Bruno, I. (2023). WorldFAIR Project (D3.1) Digital recommendations for Chemistry FAIR data policy and practice (Version 1). Zenodo. <https://doi.org/10.5281/zenodo.7887282>

There are a number of scenarios where it is useful to navigate across distributed data resources using programmatic methods - for example, a global search for specific chemicals, cross-exchange of chemical information between data repositories, validation of converted or predicted chemical representations, or integration of distributed data for compiled meta-analysis. Key to supporting these various scenarios is to confirm the chemical identities associated with the query results and to minimise the technical challenges associated with the inherent inconsistency in chemical representation and potential misinterpretation of chemical structure.

2.1 Global search

Imagine a scenario where a researcher could easily ask the general question, “Which resources or organisations around the world have information about this chemical?” but without needing to query each resource individually or to be familiar with each site’s search interface. Systems that provide federated searching or aggregator services across chemistry-related data collections have been developed to support different research domains - for example, the Virtual Atomic and Molecular Data Centre portal¹³, the NORMAN substance database^{14,15}, and the Universal Protein resource (UniProt)^{16,17}. Such systems also often provide additional features and scope to the searching interface relevant to the particular subject of the resource.

Broader, more general searches to discover chemical information can also be usefully supported through a distributed system. Rather than attempting to provide complete records for a chemical from every different database, such a search would simply provide links back to individual resources where users can get more detail. For example, a user could input a SMILES string or an InChIKey, and in a single click of a button, quickly discover if a record exists for this chemical in one (or several) of the online database resources: for example, PubChem^{18,19} or ChEMBL (Chemical database of the European Molecular Biology Laboratory)^{20,21}. It would then be up to the user to follow links to those sites to get more information as desired.

¹³ https://portal.vamdc.eu/vamdc_portal

¹⁴ <https://www.norman-network.com/nds/susdat>

¹⁵ Mohammed TH, Aalizadeh R, Alygizakis N, et al. The NORMAN Suspect List Exchange (NORMAN-SLE): facilitating European and worldwide collaboration on suspect screening in high resolution mass spectrometry. *Environ Sci Eur.* 2022;34(1):104. DOI: 10.1186/s12302-022-00680-6

¹⁶ <https://www.uniprot.org>

¹⁷ UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* 2023 Jan 6;51(D1):D523-D531. DOI: 10.1093/nar/gkac1052

¹⁸ <https://pubchem.ncbi.nlm.nih.gov>

¹⁹ Kim S, Chen J, Cheng T, Gindulyte A, et al. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* 2021 Jan 8;49(D1):D1388-D1395. DOI: 10.1093/nar/gkaa971

²⁰ <https://www.ebi.ac.uk/chembl>

²¹ Zdrazil B, Felix E, Hunter F, et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.* 2023 Nov 2;gkad1004. DOI: 10.1093/nar/gkad1004

2.2 Cross-exchange

There are a number of chemical structure-related data repositories within chemistry. Each caters to or is specific to a particular community, use case, or sub-branch of the chemical sciences or neighbouring disciplines. These chemical structure repositories include such diverse resources as GlyTouCan for glycans^{22,23}, the Protein Data Bank (PDB) for large molecules (and their small molecule bound ligands)^{24,25}, the Cambridge Structural Database (CSD) for chemical crystal structures^{26,27}, and PubChem for small molecules and their bioactivities, among others. In addition, many chemical substance vendors provide a data access portal. These various resources often provide crosslinking and exchange chemical information with each other and with other resources that integrate these repositories; for example, curated data collections such as the Global Ingredient Archival System (GIRS)^{28,29} and specialised analysis-based resources like the Pharos Project³⁰. Furthermore, these chemistry-oriented resources are accessed by users who query and download information. Users often need to consider a number of chemical structure-oriented resources as a part of their everyday work.

Each resource has its own system requirements for data access interfaces and chemical structure interpretation. The lack of standardisation between these systems creates interoperability challenges. A consistent approach for chemistry resources to expose information about the chemical representations used in their system through a common protocol would facilitate navigation across these resources. Individual data systems providing their associations between InChIs and their record identifiers could foster data hubs that provide one-stop shops for links to all resources relevant to a chemical structure using InChI as the cross-link. This could enable resources to push and pull links without dependencies on mapping to localised rules, and could scale to various scenarios of many-to-one relationships.

2.3 Validation

Machine readability of chemical structure information is a critical part of modern chemical publications, but there is sometimes a disconnect between how a chemist conceives and draws a

²² <https://glytoucan.org>

²³ Fujita A, Aoki NP, Shinmachi D, et al. The international glycan repository GlyTouCan version 3.0. *Nucleic Acids Res.* 2021 Jan 8;49(D1):D1529-D1533. DOI: 10.1093/nar/gkaa947

²⁴ <https://www.rcsb.org>

²⁵ Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. *Nat Struct Biol.* 2003 Dec;10(12):980. DOI: 10.1038/nsb1203-980

²⁶ <https://www.ccdc.cam.ac.uk/solutions/software/csd>

²⁷ Ferrence GM, Tovee CA, Holgate SJW, et al. CSD Communications of the Cambridge Structural Database. *IUCrJ.* 2023 Jan 1;10(Pt 1):6-15. DOI 10.1107/S2052252522010545

²⁸ <https://gsrs.ncats.nih.gov>

²⁹ Peryea T, Southall N, Miller M, et al. Global Substance Registration System: consistent scientific descriptions for substances related to health. *Nucleic Acids Res.* 2021 Jan 8;49(D1):D1179-D1185. DOI: 10.1093/nar/gkaa962

³⁰ <https://pharosproject.net>

chemical structure versus how that structure is stored electronically and interpreted by different chemical information systems. Researchers and other users may not be experts in how the computer interprets their structure as it is exchanged across multiple databases and programs.

There are several areas where challenges may arise in consistent representation and interpretation of chemical structures when aggregating or exchanging data between systems, including:

- Inconsistent syntax, due to errors or different conventions.
- Implausible structures according to the models and rule-sets used in a given system.
- Different approaches to structuring links and exports.
- Different interfaces for searching and structuring queries.

What if there were a way to check machine-readable structures, to view diagrams generated from different representations in different systems and review common issues such as valence, implicit hydrogens, stereocenters, etc? A checking service could ideally review and compare representations, provide visual feedback and flag potential errors or inconsistencies, comparable to the checkCIF service³¹ used with the standard Crystallographic Information File (CIF)³².

Confirming the identity of chemical substances is an important part of tracking provenance and reusability of chemical data. As it is quite possible that machine representations could become altered in the course of automated processes, validation sequences should ideally be applied at every point in a workflow where a different system is involved: when chemistry datasets are first ingested into a repository, when data are being integrated from different sources, when running queries against different systems, when using chemical notation in large language models (LLMs) or other AI/ML applications, etc.

2.4 Interoperability

Importing data from an external resource is fraught with issues, and workflows that implement data ingestion processes need to be designed with detailed rulesets to mimic a human importing and checking data as it is processed. In chemical data, the most important thing to check and align is the chemical substance that the data represent. If data are imported and accidentally assigned to the wrong substance they will not be available to the community where they should be and potentially will confuse or mislead researchers that find them associated with the wrong substance. The opportunity for users to check their representations of structure, and if needed correct them, will go a long way toward mitigating this issue. In addition, repositories where the data are contributed

³¹ <https://checkcif.iucr.org>

³² Spek AL. checkCIF validation ALERTS: what they mean and how to respond. Acta Crystallogr E Crystallogr Commun. 2020 Jan 1;76(Pt 1):1-11. DOI: 10.1107/S2056989019016244

can also use the validation process to automate the checking of the data alignment and alert developers only when needed.

3. Common protocol

The primary mechanism proposed is a standard programming interface that accepts common chemical representations to query for a chemical substance. Passing the query through a common protocol would enable a distributed search to match the chemical representation to other resources that use the same representation and are therefore likely to hold further information about that chemical.

Two related objectives are proposed to address the following questions programmatically:

- **Global chemical information resolver:** does a particular resource have any information about this particular chemical?
- **Chemical structure representation validator:** how does this resource interpret my chemical structure? Does it match my expectations?

These functions represent a two-step process necessary to facilitate FAIR data exchange:

- The global resolver provides a mechanism to programmatically **Find** and **Access** data related to a particular chemical.
- The structure validator provides information to assess the **Interoperability** and **Reusability** of a chemical representation, and by proxy the suitability of the data associated with it.

The focus of this approach is on exposing the associations of broadly-used chemical representations to records in individual databases, to enable users (people and machines) to navigate a complex, distributed chemical data landscape. SMILES and InChI are demonstrated as two well-developed, community-driven linear notations for chemical representation that are frequently used in programmatic queries and lend themselves to use in APIs. As a canonical identifier, InChI further provides a programmatic cross-walk between data records that are chemically related to some granularity.

3.1 Global chemical information resolver

Many public databases have programmatic interfaces that allow users to query for a particular chemical through some sort of web service. However, the interfaces are all unique to individual organisations; for example, the query interface for PubChem works differently from the EPA

CompTox Chemicals Dashboard (CCD) application^{33,34}. Application developers who want to gather information from multiple resources currently have to write separate, specialised code for each database: a heavy burden.

A common web service interface used by multiple database providers would enable developers to easily add the ability to query all these databases using a single code layer. One could even imagine a “global search” web site where a user could enter a single query, such as an InChIKey or SMILES string, and have that go out in real-time to PubChem, EPA CCD, ChEMBL, etc., to see if they currently have a corresponding record in their databases.

Such a system requires a standard programmatic interface - for example, to a Common Gateway Interface (CGI) using the W3C standard HTTP protocol (essentially any programming language that can send HTTP calls over the internet). This involves participating organisations agreeing on the interface and implementing a service that adheres to it. The advantage to standardising the interface alone is that the web service itself could be implemented in any technology an organisation chooses (*e.g.*, C++, Java, Python, etc.).

3.2 Chemical structure representation validator

Machine readability of a chemical structure is a significant issue in cheminformatics. While there are some common community-standard formats for chemical structures, different major chemical databases may interpret and process a chemical structure record somewhat (or very) differently, depending on the context and models used. This aspect of our work is not about creating a full standard for chemical structure interpretation, but rather providing a common way for different resources to provide feedback on their processing regarding a user’s chemical structure representation.

Imagined here is a system by which a chemist’s application – such as a structure drawing program or ELN software – could be connected to validation services provided by major databases/institutions. The chemist could then ask, with the push of a button in their application, “What does this database or repository ‘think’ of the structure as I’ve drawn it here?” and “Does this structure representation match what I know about this chemical?”

Having a common web service API to send the user’s structure and get feedback like this would allow application developers, and by extension their users, to easily check their structures against multiple resources, each of which most often have their own rules for chemical processing, to see whether the chemicals are processed as the user expects or whether there are ambiguities. In other

³³ <https://www.epa.gov/chemical-research/comptox-chemicals-dashboard>

³⁴ Williams AJ, Grulke CM, Edwards J, et al. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *J Cheminform.* 2017 Nov 28;9(1):61. DOI: 10.1186/s13321-017-0247-6

words, this is a response to the question, “Is the computer handling my chemical structure the same way I (the expert chemist) believe that it should?” This is fundamental to machine readability.

The following sections further describe the conceptual data model to support these proposed services for finding and reusing chemical information. Implementation examples based on the PubChem data model illustrate the approach with a large, open, public chemical resource that aggregates across many data sources. These demonstrations are provided to inform further collaboration on the model with other participating organisations, and are not intended to serve as complete reference implementations.

4. Global chemical information resolver

The concept described here for a global chemical information resolver is a distributed system that would help address the question, “Does this resource know anything about this particular chemical?” By using a common and reasonably simple communications protocol, any number of separate resources would be able to register themselves as part of this “global search” and be able to provide results back to a single central query host. Each individual resource would implement a search for the input chemical representation within their own infrastructure, presumably using existing internal methodology.

Importantly, this frees the central query manager from doing any actual cheminformatics work (e.g., parsing SMILES) or needing to know the details of each resource’s search API. This simplifies the cross-site query system and makes it easy to add new databases to the search. It also spreads the computational burden across all organisations, rather than putting a heavy load on the central query server.

4.1 Architecture concept

The proposed design of the global resolver is fairly straightforward, and has three main components:

- a shared data model for information exchange,
- a local resolver implemented and hosted by each participating resource,
- and a central “meta resolver” that communicates with the local resolvers and the end user (which may be a human or a machine), as diagrammed in Figure 1 below. While humans need not be in the loop, it also provides a web page interface that the user can interact with directly.

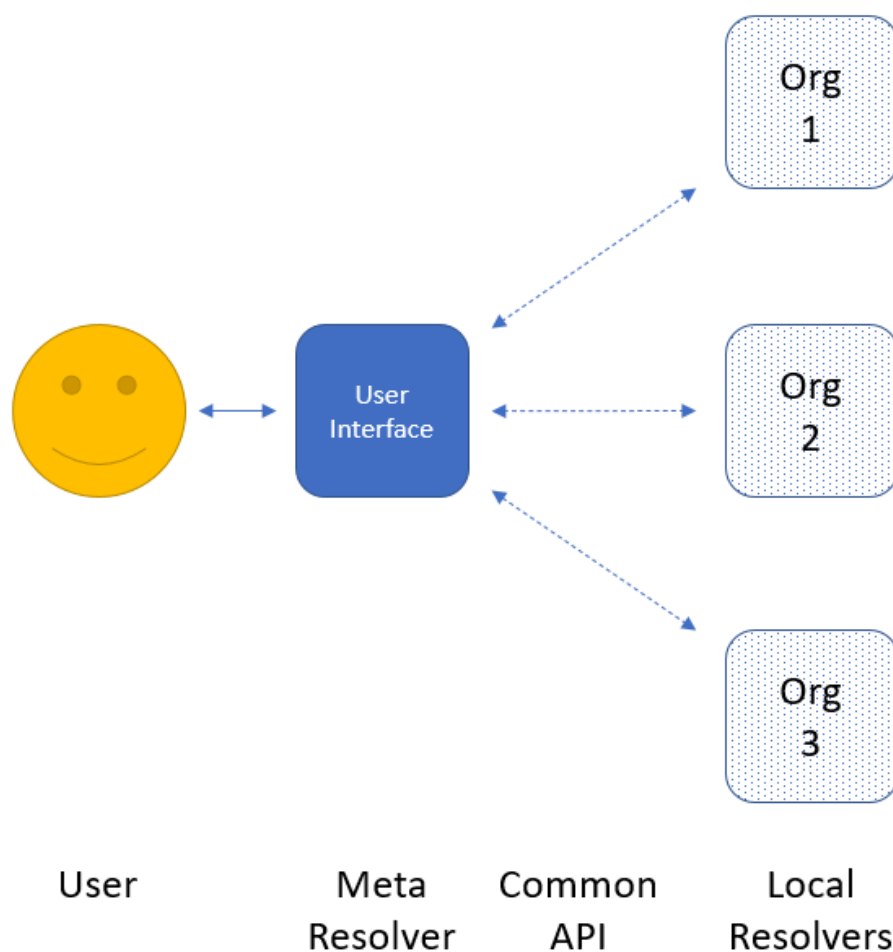


Figure 1. The proposed design of the global resolver.

In brief, the user (machine or human) would provide a chemical structure query, and the meta resolver would send that query to each local resolver in its list (presumably in parallel) to provide a distributed query. Each local resolver would perform the lookup within their system, and respond back to the meta resolver, which would gather all the responses to present to the user. The most important part would be for the user to be able to clearly understand which organisations are responding, and to have links to follow through to those resources for more detail. In this way, the meta resolver does not need to know what each resource's full record contains, or to try to present each resource's heterogeneous record in a common way. All the meta resolver needs to know is, "this organisation has information on this chemical" and it is then up to the user to decide where to go from there.

4.1.1 Common data model

If this system is to be practical to implement across multiple databases all over the world, with new participants to be added at any time and without great effort, there must be a common “language” that is used for communications between the central meta resolver and the individual organisations’ local resolvers. It should use, for communications and the data model, generic common standards that are not dependent on any particular host platform, programming language or database system. This simplifies the meta resolver because it uses exactly the same method to communicate with each organisation, and frees individual organisations to implement their local component in whatever way is most convenient for them given their existing infrastructure.

What is proposed here – but is certainly open to change – uses the HTTP protocol for communications, with normal common gateway interface (CGI) arguments for input, and then a simple JSON construct for the response data. The HTTP and CGI standards are universal enough that programmers using just about any platform should have access to the web data parsers and web communications tools needed to implement this service.

As a parenthetical note, PubChem has a very simple data model for this written in XML schema, as NCBI has (publicly available) software to read and write equivalent JSON or XML files based on a single schema. This schema, in its prototype form, is used to construct the examples in the next section, and is available at https://pubchem.ncbi.nlm.nih.gov/resolver/resolver_data.xsd

A more formal implementation, if it uses JSON, should probably use a JSON-native format like JSON Schema (<https://json-schema.org/draft/2020-12/json-schema-core.html>, although it is still under development at the time of this writing). Alternatively, the data format of choice for this whole system could be XML instead. Regardless, there should be some way to make it clear to developers what the structure of the data should be and make it possible to validate the JSON data automatically. These details, and whether to use HTTP or some other communication protocol, are not inherently critical to the overall design of this system but would be important to work out before individual organisations begin work on their implementations.

4.2 Implementation

4.2.1 Local resolver

Each individual organisation that wants to participate in this global search would have to implement their own resolver, hosted on their local system. As stated earlier, exactly how they implement it is not relevant, so long as it follows the common rules of the communications protocol. The local resolver would need to be able to interpret the query chemical structure and perform the search in their system, essentially responding with either, “Here is a summary of what I know about this chemical”, or “I don’t have this chemical in my database.”

There is a sample implementation available from PubChem. For example, for the hydrocarbon **butane**, each of the URLs in Table 1 produce the same output (shown in Scheme 1), from different representations of the input structure:

Table 1. Various representations of **butane** used as input to the prototype resolver.

Chemical representation (input by user)	Input URL (base URL implemented by local resource)
SMILES	https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?smiles=CCCC
InChIKey	https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?inchikey=IJDNQMDRQITEOD-UHFFFAOYSA-N
InChI (full string)	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)
chemical name	https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?name=butane

These all give the JSON response as shown in Scheme 1, below.

Scheme 1. Output from PubChem with information available about **butane**, in response to various representations of this chemical.

```
{
  "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=1&cid=7843",
    "IUPACName": "butane",
    "SMILES": "CCCC",
    "InChI": "InChI=1S/C4H10/c1-3-4-2/h3-4H2,1-2H3",
    "InChIKey": "IJDNQMDRQITEOD-UHFFFAOYSA-N"
  }
}
}
}

```

The key parts of this response data include:

- the identity of the organisation (Resource and **ResourceURL**);
- the identifier used by this organisation for this chemical (**ResourceIdentifier** and **ResourceIdentifierType**);
- a link to the organisation's full detail page (**RecordURL**);
- a link to an image of the input substance (**ImageURL**);
- and other standard identifiers for this chemical according to this organisation (**IUPACName**, **SMILES**, **InChI**, **InChIKey**).

The meta resolver (described in more detail below) would then be able to take this simple response from each organisation and present the information to the user, presumably in some sort of tabular format.

It is possible that a database may not contain the given input representation, in which case an empty list could be produced, or else some to-be-determined explicit "not found" message as shown in Scheme 2 for the input URL (PubChem example):

Input URL with InChIKey:

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

Scheme 2. Example response from a database such as PubChem that does not contain the given input identifier.

```
{
  "Result": {}
}
```

Some reasonable (human-readable) error message should be produced if the input cannot be interpreted – erroneous SMILES, invalid InChI format, and such. But the details of these error reports should be left to the individual responder, as it is too complex to enforce a common reporting standard in such cases due to different software toolkits, algorithms, etc.

There are many different ways to provide chemical structures as inputs, and not every resource would be able to handle every type. For example, some organisations may not have the ability to interpret SMILES as input as some degree of chemical information processing is necessary (i.e., it is not a simple string match). There should be some way for each organisation to indicate which input methods it supports. Similarly, some databases may not have individual record web pages for each chemical or may not have IUPAC-style names for each chemical, etc., so available outputs should also be provided in some way. For example, the prototype response presented in Scheme 3 reflects what happens if the protocol is called with no parameters at all, whereby the resource returns a list of available inputs and outputs:

Input URL:

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

Scheme 3. The resource returns a list of available inputs and outputs if the protocol is called with no input parameters specified.

```
{
  "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
    }
  }
}

```

The **AvailableInputs** and **AvailableOutputs** fields indicate what inputs and outputs this local resolver can handle. In this case, PubChem is saying that it can't take an image (PNG file) as input, but other resources may have the ability to convert a chemical drawing in an image to a chemical structure. This does not mean that every record will necessarily have all of these output fields – some records in PubChem for instance may have a SMILES but not a (computed) IUPAC name. At least one of the outputs should be specified for each record, otherwise there would be nothing to show to the user.

Another key field here is the **ResolverURL**, which is the base URL of the organisation's CGI that implements this protocol; a key detail that will be important to the meta resolver.

4.2.2 Meta resolver

The last major component of this design, and the only part the user (machine or human) would interact with directly, is the “meta resolver”: the web page that takes user input, performs the search, and presents results back to the user. The general idea is that the meta resolver would have built into it a list of all available local resolvers. When the user submits a query, the meta resolver would (presumably in parallel) send requests to each local resolver based on that input, wait for results to be returned, and then provide or show a table of responding organisations with hyperlinks back to their record pages.

This is feasible for any number of participating organisations because the HTTP request is *exactly the same* for each resource in the list, varying only by the base URL of the request (**ResolverURL** above). As the response data from each local resolver is in *exactly the same* JSON format, the meta resolver would easily be able to parse the results from each organisation, without needing to know anything about the details of that organisation's internal data structures, algorithms, and so on. This makes the programming of the meta resolver relatively straightforward. Of course, the meta resolver would need to know that if, for example, the user supplies SMILES as input, that it would only send requests to local resolvers that implement SMILES as an input method. (While it is possible that the input type could be optional and autodetected, this is not foolproof, as some SMILES could be interpreted as a chemical name and vice versa.)

While there is not currently a prototype implementation of the “meta resolver”, the “machine and human” interface would provide the means to perform the aforementioned actions described in this section. In addition, it would be able to obtain the full list of resolvers, search the list of resolvers, and filter resolvers to just those classified as a particular type or set of types. In addition, it would enable the metadata about each resolver (name, description, classification, owner organisation, and other appropriate metadata) to be obtained. These additional capabilities could be similar to the PubChem Data Sources page (<https://pubchem.ncbi.nlm.nih.gov/source/>) for example, which provides a human-friendly interface to search, subset, download, and locate desired data source metadata.

5. Chemical structure representation validator

Every major chemical information system has its own rules for how to process and validate chemical structures. The goal here is not to create a standard for those rules, but rather to provide a common way for each institution to provide their own validation feedback to a chemical structure represented in an industry standard format (such as SMILES or MOL/SDF). If each institution uses

the same web service API specification for this validation, then an application such as ChemDraw could easily let the user choose from a number of resources, then provide a “validate” button that would send their structure to that resource over the internet and get back a standard response with feedback on that chemical, which the application can interpret and present back to the user.

These validators would presumably answer basic chemical informatics questions such as “Are all of the atoms in this structure interpreted as valid chemical elements and isotopes?” or “Is there an appropriate valence for each element or are some in a hypervalent state?” or “Are all of the stereocenters present in this molecule fully recognized and defined?” If the validator can produce an image, then the chemist could check whether the automatically-produced image matches what they have drawn and that the input they have provided has been interpreted as expected by the system they are interacting with.

The following section further describes a conceptual approach to the API with some examples of how these might be implemented. Detailed technical specification needs further development.

5.1 Prototype and examples

PubChem has, for demonstration purposes, created a prototype of this chemical structure validation service. In its simplest use, it takes chemical structure input (e.g., SMILES) and produces a JSON message. For example, Scheme 4 shows the output from the URL containing the SMILES of n-butane:

Input URL with SMILES:

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

Scheme 4. *Prototype of chemical structure validation service with output from PubChem demonstrating the response to a URL containing the SMILES of n-butane as an example.*

```
{
  "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"
}
]
}

```

The example service above provides some basic cheminformatics properties, such as the number of atoms, stereocenters, and so on. Again, the goal is to make sure that a machine-interpreted result from a specific resource (in this case PubChem) matches the chemist's expectations.

(Technical note: the data model used for this response is available as an XML schema here: https://pubchem.ncbi.nlm.nih.gov/resolver/resolver_data.xsd; PubChem has tools that simplify XML and JSON input/output based on XML Schema, so it was more convenient to implement this prototype from the XML Schema. If JSON is designated as the default format of this service, it would probably be better ultimately to use a JSON Schema if/when that, or something like it, becomes a more widely recognized standard. But it is important to have a fully defined data model so that any application that uses this service knows what to expect, and how to parse the results.)

Here is a slightly more complicated structure, with both atom and bond stereochemistry, with output as shown in Scheme 5:

Input URL with SMILES:

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

Scheme 5. Output of an example with both atom and bond stereochemistry.

```
{
  "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"
      }
    ]
  }
}
```

From this output, the chemist can easily see that the stereocenters have been perceived and fully defined as per PubChem's model.

If the responding database perceives an inconsistency or ambiguity in the chemical structure representation relative to their local rules, there should be some feedback as to what the issue is; for example, the response from PubChem to input for a pentavalent carbon as shown in Scheme 6.

Input URL with SMILES:

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

Scheme 6. PubChem feedback to a perceived error in the input structure representation.

```
{
  "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"
    ]
  }
}
```

Here is another example, with the PubChem response for what it perceives as an invalid isotope shown in Scheme 7:

Input URL with SMILES:

[https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&smiles=C\[5H\]](https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&smiles=C[5H])

Scheme 7. PubChem response to an invalid isotope.

```
{
  "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"
    ]
  }
}
```



```
1
}
}
```

While 5H does exist, the PubChem resource has a policy to reject anything with isotopes with <1ms half life. Other institutions may have different policies. It is not the intention here to define the precise validation rules, but rather to provide a way for each organisation to give feedback on the structure according to their own existing internal rules, but in a standard format.

Chemists are used to looking at chemical structure drawings and so they may prefer to see an image rather than the data fields shown above in Scheme 7. For example, Figure 2 shows the output from the input URL with SMILES:

[https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&format=png&smiles=C\[C@H\]\(CCCC\(C\)C\)\[C@H\]1CC\[C@@H\]2\[C@@\]1\(CC\[C@H\]3\[C@H\]2CC=C4\[C@@\]3\(CC\[C@@H\]\(C4\)O\)C\)C](https://pubchem.ncbi.nlm.nih.gov/resolver/resolver.cgi?action=validate_structure&format=png&smiles=C[C@H](CCCC(C)C)[C@H]1CC[C@@H]2[C@@]1(CC[C@H]3[C@H]2CC=C4[C@@]3(CC[C@@H](C4)O)C)C)

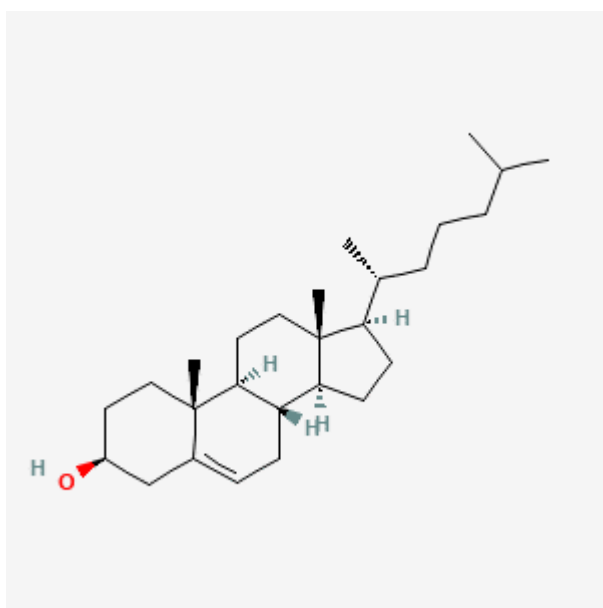


Figure 2. PubChem output in an image format.

5.2 Sample web application

The examples above show the actual HTTP URLs used by this proposed service. Of course, chemists are not generally going to be typing URLs into their web browser, but would instead be using this as part of some other application. PubChem has a (very simplistic) example of this as a web page service that uses a simple form input to the validation CGI:

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

This sample web interface lets the user select an input type (e.g., SMILES, InChiKey, etc.), and choose whether to get validation details (as JSON) or an image. It also demonstrates the possibility of accepting MOL/SDF as input, which is a multi-line format not amenable to a simple URL syntax as in the examples above.

Two different implementations are provided to illustrate distinctions in validation rules: one using PubChem's existing but internal standardisation software, and another using RDKit – an open-source chemical information toolkit³⁵. This demonstration is provided in the form of a web page in PubChem (see link above), and also via an interactive Jupyter notebook:

<https://iupac.github.io/WFChemProtocols/IUPACProtocolsDemo.html>

6. Available infrastructure

In addition to the prototype HTTP service as described above, documentation and examples for this work are available on GitHub, as shown in Figure 3. The primary repository is at:

<https://github.com/IUPAC/WFChemProtocols>.

³⁵ <https://www.rdkit.org>

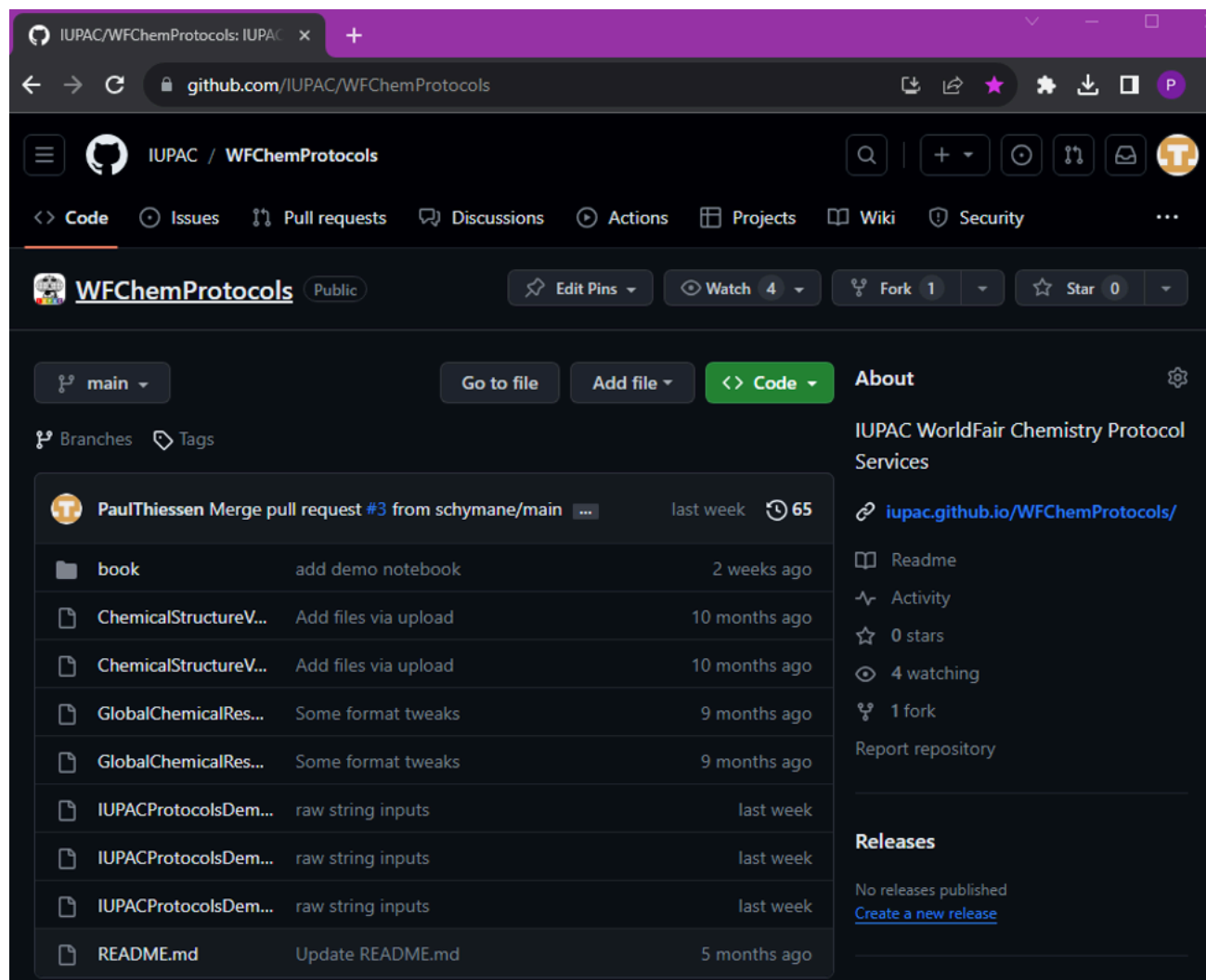


Figure 3. IUPAC GitHub website for WFChemProtocols.

The GitHub repository contains the raw markdown files (containing formatted textual documentation) for a Jupyter notebook with the project documentation, published at <https://iupac.github.io/WFChemProtocols/intro.html> (as shown in Figure 4).

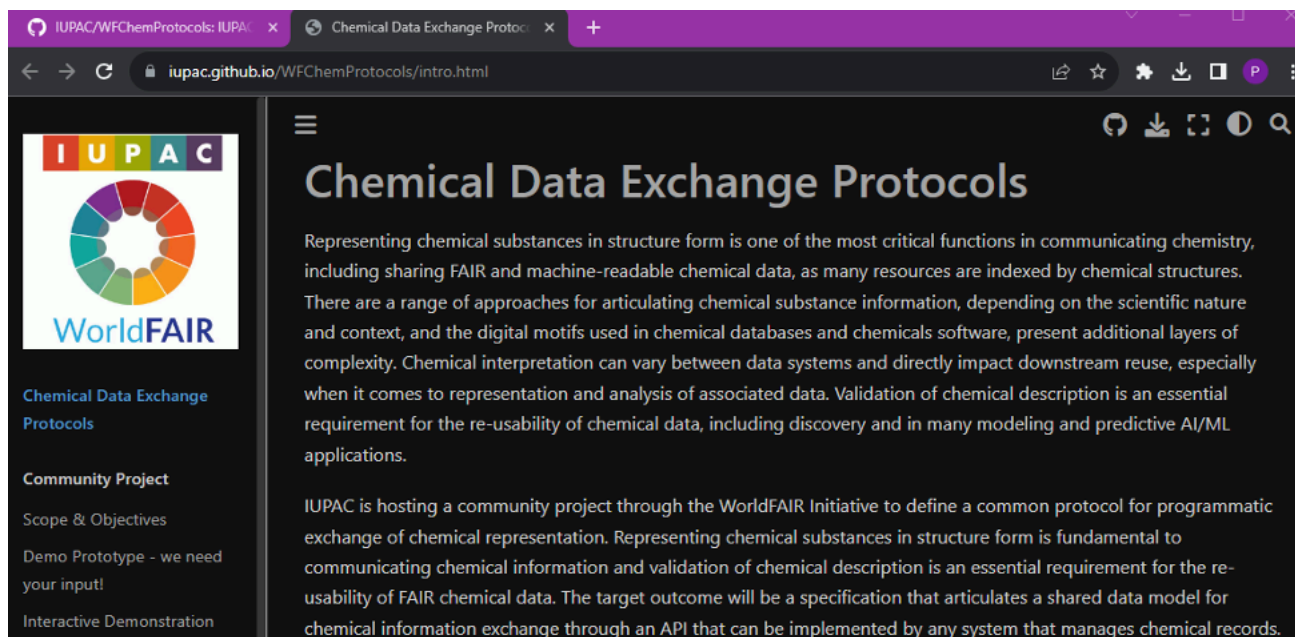


Figure 4. IUPAC GitHub.io website for the WFChemProtocols Jupyter Notebook.

For users wanting to dive into this more deeply, there is also an interactive demonstration available at <https://iupac.github.io/WFChemProtocols/IUPACProtocolsDemo.html>.

This is a Jupyter Notebook that contains live code, visualisations and explanatory text that can be downloaded directly, or can be launched through Google's Colab site (see Figure 5) via the rocket icon in the upper right-hand side of the page (as shown in Figure 5).

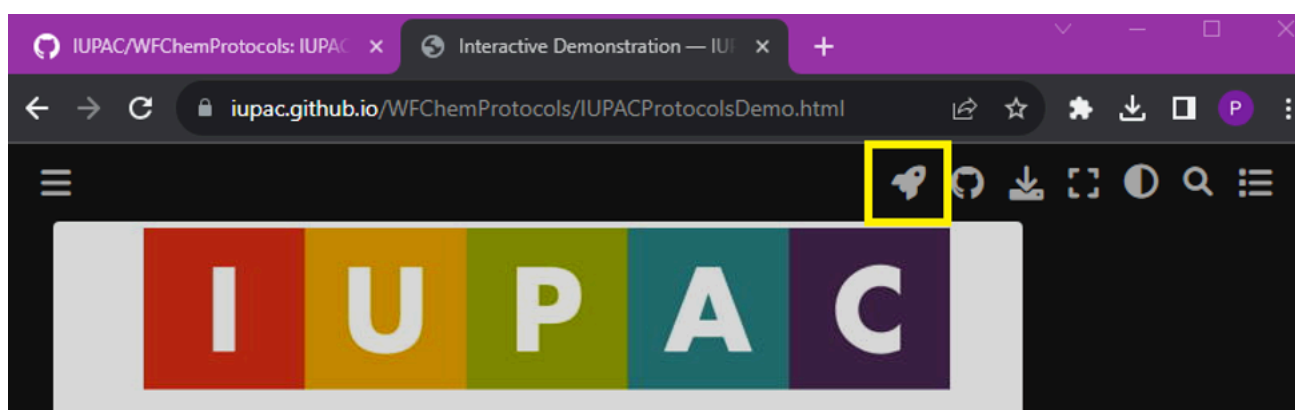


Figure 5. IUPAC Jupyter notebook showing the location of the Google Colab site link using a yellow square.

The purpose of the notebook is to demonstrate the expected functionality of the resolver being proposed by this IUPAC-WorldFAIR activity. In this interactive resource, users can run code examples (in the Python programming language) and make changes, for example to include their own SMILES strings for chemical structures, in order to see how different inputs affect the results of the prototype web service.

This can also be invaluable to users who are less familiar with APIs, scripting in Python, or using the JSON format. Seeing how this is implemented and being able to play with the code is important to researchers in understanding what aligning the exchange protocol does for them and how they can take advantage of the approach. Additional notebooks demonstrating functionalities of chemical structure representations and other functions working with chemical data are under development through the WorldFAIR Chemistry D3.2 work (see the [IUPAC FAIR Chemistry Cookbook](#)³⁶).

7. Community engagement

Community engagement is ongoing to formalise this approach to uniform system-to-system interoperability. There are four key aspects to drive adoption: promoting use of the same protocol, requesting help to improve the specification, drawing in help to maintain and evolve the specification as a function of time, and communicating and engaging with the broader community through the online materials, outreach activities and other approaches. With IUPAC as the key forum for the specification, there will always be the auspices of a standards organisation to further the needs of the community.

7.1 Protocol adoption

As an open-source protocol, the difficulty for software developers to implement the protocol is minimal. Most existing web-based, data-focused resources already have basic protocols to do what this protocol does; however, these non-standard protocols are different for each resource. Most existing user-focused resources already have interfaces to one or more data-focused resources and would welcome a unified protocol, as it opens up any protocol-compliant system to their users.

Adoption of the protocol entails a given user-focused organisation being able to properly formulate queries and interpret responses of the protocol. Examples of user-focused organisations include software drawing tool providers (such as ChemDraw), ELN software providers (such as those from Scilligence, Indigo and others), and programming toolkit providers (such as those from OpenEye, ChemAxon or the open source variants such as RDKit, CDK).

³⁶ <https://iupac.github.io/WFChemCookbook>

Adoption also entails a given data-focused organisation handling user-focused protocol queries and spending appropriate responses. Example data-focused organisations would include online chemistry data systems (such as PubChem, EPA CCD, GSRS, ChEMBL, DrugBank).

Participating organisations at the most basic level would need to be compliant with the protocol. More advanced participation would entail helping to improve the protocol, developing software that utilises the protocol, and enhancing adoption of the protocol. Adoption by a small number of popular chemistry resources such as chemistry drawing packages, open source chemistry APIs, and open/free databases is sufficient to create critical mass for protocol adoption.

Community outreach is ongoing for protocol adoption. The Jupyter notebook provides the information necessary to adopt the protocol. PubChem has developed demonstration projects. Feedback mechanisms are in place to capture community input via a Google form³⁷ and via GitHub discussions³⁸ and bug report mechanisms. At workshops and conferences, data-focused and user-focused organisations are being engaged with talks, discussions, and a one-page flyer³⁹ with QR codes for easy access to further information. Engagement by means of online meetings and discussions are ongoing with key stakeholders of user-focused and data-focused resources.

7.2 Specification development

It is envisioned that ongoing work to develop the specification and additional prototypes will be coordinated by IUPAC through a continuation of the project initiated for this deliverable. Formalisation of the proposed approach will involve articulation and agreement of participating organisations on the interface. Initial requirements outlined in this report include the following parameters:

- Shared data model for information exchange
 - HTTP protocol for communications
 - CGI arguments for input
 - JSON construct for response data (such as JSON Schema <https://json-schema.org/draft/2020-12/json-schema-core.html>).
- Local resolvers
 - Each individual organisation implements their own local resolver that follows the common rules of the communications protocol.
- Common specification and syntax for the input and response formats
 - Input: HTTP request is consistent for each resource in the list, varying only by the base URL of the request (ResolverURL)
 - Response: response data from each local resolver is in a consistent JSON format.

³⁷ <https://docs.google.com/forms/d/e/1FAIpQLScub8Joj8CzvXJXDA5-OA1ix7h9p09MjUDFhmVIZtafzG0rCQ/viewform>

³⁸ <https://github.com/IUPAC/WFChemProtocols/discussions>

³⁹ <https://doi.org/10.5281/zenodo.8322966>

- Meta resolver
 - Essential fields
 - **ResolverURL** (base URL of the organisation's CGI that implements this protocol)
 - **AvailableInputs** (indicate what inputs this local resolver can handle)
 - **AvailableOutputs** (indicate what outputs this local resolver can handle).
 - Response data
 - **Resource** and **ResourceURL** (the identity of the organisation)
 - **ResourceIdentifier** and **ResourceIdentifierType** (the local identifier used by this organisation for this chemical);
 - **RecordURL** (a link to the organisation's full detail page)
 - **ImageURL** (a link to an image of the input substance)
 - **IUPACName, SMILES, InChI, InChIKey** (and other standard identifiers for this chemical according to this organisation).
- Validator
 - Flags for potential omission, ambiguity, inconsistency (may vary across systems).

Central to the meta resolver is the list of local resolvers, those organisations participating in the “global search.” To facilitate participation, a self-registration mechanism with a low barrier of human intervention will need to be provided, coupled with a basic vetting process to ensure that each resource is legitimate. This information content will be stewarded by IUPAC as the formal governance body.

8. Conclusion

Presented in this report is a two-fold mechanism to facilitate FAIR chemical data exchange, involving a global chemical information resolver to programmatically **Find** and **Access** data related to a particular chemical, and a chemical representation validator to assess the **Interoperability** and **Reusability** of a chemical representation associated with a data record. The proposed communications protocol would enable any data resource that indexes chemicals to register as part of this global search service, utilising their existing query infrastructure. Standardising the API enables navigation between systems without dependencies on bespoke mappings and can help expose inconsistencies in chemical representation.

This approach provides broadly applicable system-to-system interoperability for chemicals across domains and use cases. Effectiveness of the service is enhanced through the use of machine-readable chemical representations in controlled vocabularies, ontologies and metadata schema. IUPAC will continue to engage with the community to further develop the service and increase the adoption of InChI and other machine-readable chemical representations.

9. Appendices

9.1 Oral presentation: Standardised programmatic access to chemical information (Protocol Services)

This is an overview⁴⁰ of the [WorldFAIR Chemistry](#) sub-project “[Protocol Services](#)”. It was presented at the workshop titled “Advancing FAIR Chemistry: Developing New Services for Sharing Chemical Data”. The event took place during the ACS spring meeting in Indianapolis, US, on March 27, 2023.

9.2 Workshop Session: Doc-a-thon: Chemical representation best practices for humans and machines

This is a recording⁴¹ of a 90 min discussion session. It was presented at the workshop titled “Advancing FAIR Chemistry: Developing New Services for Sharing Chemical Data” organised by the [WorldFAIR Chemistry](#) Case Study. The event took place during the ACS spring meeting in Indianapolis, US, on March 27, 2023. The Github direct link is [here](#).

9.3 Poster presentation: IUPAC WorldFAIR Chemistry: Managing Chemical Data Digitally

WorldFAIR Chemistry poster⁴² presented at the IUPAC CHAINS. The meeting took place in the Hague, the Netherlands, Aug 18-26 2023.

9.4 Structure validator demo

View a video of the Structure Validator demo [here](#).⁴³ Try it [here](#).⁴⁴ Feedback is welcomed via the [Questionnaire](#) or [Feedback & suggestions](#) links.

⁴⁰ Bolton, E. (2023, April 5). Standardised programmatic access to chemical information (Protocol Services). Zenodo. <https://doi.org/10.5281/zenodo.7803871>

⁴¹ Scalfani, V., McEwen, L., & Bolton, E. (2023, April 6). Doc-a-thon: Chemical representation best practices for humans and machines. Zenodo. <https://doi.org/10.5281/zenodo.7803914>

⁴² Mustafa, F., McEwen, L., Bruno, I., Chalk, S., & Bolton, E. (2023). IUPAC WorldFAIR Chemistry: Managing Chemical Data Digitally. Zenodo. <https://doi.org/10.5281/zenodo.8322967>

⁴³ https://www.youtube.com/watch?v=8VRodQPCI_U

⁴⁴ <https://iupac.github.io/WFChemProtocols/demo.html>

Bibliography

- Bento AP, Hersey A, Félix E, Landrum G, Gaulton A, Atkinson F, Bellis LJ, De Veij M, Leach AR. An open source chemical structure curation pipeline using RDKit. *J Cheminform.* 2020 Sep 1;12(1):51. <https://doi: 10.1186/s13321-020-00456-1>.
- Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. *Nat Struct Biol.* 2003 Dec;10(12):980. <https://doi: 10.1038/nsb1203-980>.
- Bolton, E. (2023, April 5). Standardised programmatic access to chemical information (Protocol Services). Zenodo. <https://doi.org/10.5281/zenodo.7803871>
- Chalk, S., Andres, A., Bloodworth, S., Coles, S. Cuadros, J. Herris-Pawlis, S., Joliffe, J., Kim, S., Knight, N., Kroenlin, K., Li, Y., McEwen, L., Munday, S., Mustafa, F., & Scalfani, V. (2023). IUPAC WorldFAIR Cookbook for FAIR chemical data. Jupyter book, <https://iupac.github.io/WFChemCookbook> (accessed, 20231105).
- Ferrence GM, Tovee CA, Holgate SJW, Johnson NT, Lightfoot MP, Nowakowska-Orzechowska KL, Ward SC. CSD Communications of the Cambridge Structural Database. *IUCrJ.* 2023 Jan 1;10(Pt 1):6-15. <https://doi: 10.1107/S2052252522010545>.
- Fujita A, Aoki NP, Shinmachi D, Matsubara M, Tsuchiya S, Shiota M, Ono T, Yamada I, Aoki-Kinoshita KF. The international glycan repository GlyTouCan version 3.0. *Nucleic Acids Res.* 2021 Jan 8;49(D1):D1529-D1533. <https://doi:10.1093/nar/gkaa947>.
- Hähnke VD, Kim S, Bolton EE. PubChem chemical structure standardisation. *J Cheminform.* 2018 Aug 10;10(1):36. <https://doi:10.1186/s13321-018-0293-8>.
- Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research.* 2016 Jan;44(D1):D1214-9. <https://DOI: 10.1093/nar/gkv1031>.
- Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. InChI, the IUPAC International Chemical Identifier. *J Cheminform.* 2015 May 30;7:23. <https://doi:10.1186/s13321-015-0068-4>.
- Jacobs A, Williams D, Hickey K, Patrick N, Williams AJ, Chalk S, McEwen L, Willighagen E, Walker M, Bolton E, Sinclair G, Sanford A. CAS Common Chemistry in 2021: Expanding Access to Trusted Chemical Information for the Scientific Community. *J Chem Inf Model.* 2022 Jun 13;62(11):2737-2743. <https://doi:10.1021/acs.jcim.2c00268>.
- Karapetyan K, Batchelor C, Sharpe D, Tkachenko V, Williams AJ. The Chemical Validation and Standardization Platform (CVSP): large-scale automated validation of chemical structure datasets. *J Cheminform.* 2015 Jun 19;7:30. <https://doi:10.1186/s13321-015-0072-8>.
- Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* 2021 Jan 8;49(D1):D1388-D1395. <https://doi: 10.1093/nar/gkaa971>.
- Lowe CN, Williams AJ. Enabling High-Throughput Searches for Multiple Chemical Data Using the U.S.-EPA CompTox Chemicals Dashboard. *J Chem Inf Model.* 2021 Feb 22;61(2):565-570. <https://doi:10.1021/acs.jcim.0c01273>.

- McEwen, L., & Bruno, I. (2023). WorldFAIR Project (D3.1) Digital recommendations for Chemistry FAIR data policy and practice (Version 1). Zenodo. <https://doi.org/10.5281/zenodo.7887282>
- Mohammed Taha H, Aalizadeh R, Alygizakis N, Antignac JP, Arp HPH, Bade R, Baker N, Belova L, Bijlsma L, Bolton EE, Brack W, Celma A, Chen WL, Cheng T, Chirsir P, Čirka L, D'Agostino LA, Djoumbou Feunang Y, Dulio V, Fischer S, Gago-Ferrero P, Galani A, Geueke B, Głowacka N, Glüge J, Groh K, Grosse S, Haglund P, Hakkinen PJ, Hale SE, Hernandez F, Janssen EM, Jonkers T, Kiefer K, Kirchner M, Koschorreck J, Krauss M, Krier J, Lamoree MH, Letzel M, Letzel T, Li Q, Little J, Liu Y, Lunderberg DM, Martin JW, McEachran AD, McLean JA, Meier C, Meijer J, Menger F, Merino C, Muncke J, Muschket M, Neumann M, Neveu V, Ng K, Oberacher H, O'Brien J, Oswald P, Oswaldova M, Picache JA, Postigo C, Ramirez N, Reemtsma T, Renaud J, Rostkowski P, Rüdel H, Salek RM, Samanipour S, Scheringer M, Schliebner I, Schulz W, Schulze T, Sengl M, Shoemaker BA, Sims K, Singer H, Singh RR, Sumarah M, Thiessen PA, Thomas KV, Torres S, Trier X, van Wezel AP, Vermeulen RCH, Vlaanderen JJ, von der Ohe PC, Wang Z, Williams AJ, Willighagen EL, Wishart DS, Zhang J, Thomaidis NS, Hollender J, Slobodnik J, Schymanski EL. The NORMAN Suspect List Exchange (NORMAN-SLE): facilitating European and worldwide collaboration on suspect screening in high resolution mass spectrometry. *Environ Sci Eur.* 2022;34(1):104. <https://doi:10.1186/s12302-022-00680-6>.
- Mustafa, F., McEwen, L., Bruno, I., Chalk, S., & Bolton, E. (2023). IUPAC WorldFAIR Chemistry: Managing Chemical Data Digitally. Zenodo. <https://doi.org/10.5281/zenodo.8322967>
- Peryea T, Southall N, Miller M, Katzel D, Anderson N, Neyra J, Stemmann S, Nguyễn ĐT, Amugoda D, Newatia A, Ghazzaoui R, Johanson E, Diederik H, Callahan L, Switzer F. Global Substance Registration System: consistent scientific descriptions for substances related to health. *Nucleic Acids Res.* 2021 Jan 8;49(D1):D1179-D1185. <https://doi:10.1093/nar/gkaa962>.
- Richard AM, Huang R, Waidyanatha S, Shinn P, Collins BJ, Thillainadarajah I, Grulke CM, Williams AJ, Lougee RR, Judson RS, Houck KA, Shobair M, Yang C, Rathman JF, Yasgar A, Fitzpatrick SC, Simeonov A, Thomas RS, Crofton KM, Paules RS, Bucher JR, Austin CP, Kavlock RJ, Tice RR. The Tox21 10K Compound Library: Collaborative Chemistry Advancing Toxicology. *Chem Res Toxicol.* 2021 Feb 15;34(2):189-216. <https://doi:10.1021/acs.chemrestox.0c00264>.
- Scalfani, V., McEwen, L., & Bolton, E. (2023, April 6). Doc-a-thon: Chemical representation best practices for humans and machines. Zenodo. <https://doi.org/10.5281/zenodo.7803914>
- Sinclair G, Thillainadarajah I, Meyer B, Samano V, Sivasupramaniam S, Adams L, Willighagen EL, Richard AM, Walker M, Williams AJ. Wikipedia on the CompTox Chemicals Dashboard: Connecting Resources to Enrich Public Chemical Data. *J Chem Inf Model.* 2022 Oct 24;62(20):4888-4905. <https://doi:10.1021/acs.jcim.2c00886>.
- Spek AL. checkCIF validation ALERTS: what they mean and how to respond. *Acta Crystallogr E Crystallogr Commun.* 2020 Jan 1;76(Pt 1):1-11. <https://doi:10.1107/S2056989019016244>.
- Thiessen, P., Bolton, E., Williams, A., Mustafa, F., & McEwen, L. (2023). IUPAC WorldFair Chemistry Protocol Services. Jupyter book, <https://iupac.github.io/WFChemProtocols> (accessed, 20231105).

- Thiessen, P. (2023, May 29). How do I know that a structure representation is correct?, https://www.youtube.com/watch?v=8VROdQPCI_U (accessed 20231105).
- Thiessen, P. (2023, October 17). Demo Prototype, <https://iupac.github.io/WFChemProtocols/demo.html> (accessed 20231105).
- Thomaidis NS, Hollender J, Slobodnik J, Schymanski EL. The NORMAN Suspect List Exchange (NORMAN-SLE): facilitating European and worldwide collaboration on suspect screening in high resolution mass spectrometry. *Environ Sci Eur.* 2022;34(1):104. <https://doi:10.1186/s12302-022-00680-6>.
- UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* 2023 Jan 6;51(D1):D523-D531. <https://doi:10.1093/nar/gkac1052>.
- Williams AJ, Gaines LGT, Grulke CM, Lowe CN, Sinclair GFB, Samano V, Thillainadarajah I, Meyer B, Patlewicz G, Richard AM. Assembly and Curation of Lists of Per- and Polyfluoroalkyl Substances (PFAS) to Support Environmental Science Research. *Front Environ Sci.* 2022 Apr 5;10:1-13. <https://doi:10.3389/fenvs.2022.850019>.
- Williams AJ, Grulke CM, Edwards J, McEachran AD, Mansouri K, Baker NC, Patlewicz G, Shah I, Wambaugh JF, Judson RS, Richard AM. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *J Cheminform.* 2017 Nov 28;9(1):61. <https://doi:10.1186/s13321-017-0247-6>.
- Zdrazil B, Felix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, de Veij M, Ioannidis H, Lopez DM, Mosquera JF, Magarinos MP, Bosc N, Arcila R, Kizilören T, Gaulton A, Bento AP, Adasme MF, Monecke P, Landrum GA, Leach AR. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.* 2023 Nov 2:gkad1004. <https://doi:10.1093/nar/gkad1004>.