PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Case Study 8: The Role of Structure in Physical Sciences Data Management

Ian Bruno, Suzanna Ward

Cambridge Crystallographic Data Centre, Cambridge, CB2 1EZ

1 Structure Representation in the Physical Sciences

Contents

1.1 What do we mean by "structure"?

The word structure means many things and the first one that comes to mind will very much depend on the context in which someone is operating. For example, in a civil engineering context it might primarily relate to the construction of a building; for an organisation that could be housed in such a building, it might reflect how people and groups are arranged and interrelate; for an individual within that organisation, it might mainly refer to the way they organise their work and any reports they produce.

For the purposes of this report, "structure" refers to the structure of molecules and materials, primarily at an atomic level. It focuses mainly on the chemical and materials sciences and has in mind connections to the biological sciences and other related domains.

The report may at times refer to structure in other contexts such as the architecture of information and data systems, but hopefully not in an unhelpful way.

1.2 Importance of structure representation in the Physical Sciences

The structure of molecules and materials is fundamental to much of the research undertaken in the physical sciences. The primary focus of our report is the digital representation of structure and the management of structures and associated properties (collectively, "structural data"). Representation of structure in ways that are machine actionable, i.e. can be understood and processed by a computer, is critical for realisation of the FAIR Data Principles [1] across the physical sciences.

Digital representations of structures underpin information and data management systems regularly used by researchers but not all will be aware of what these representations are. It is likely that many researchers will equate digital representation of a structure to an image in a chemical structure drawing package and nothing more. Conversely, some researchers will be very aware of the nuances of digital representations of structure as these are fundamental to their research. Their experiences may, however, be limited to a particular domain and they may be unaware of, and get tripped up by, some of the broader interoperability challenges associated with structure representation.

1.3 Structure Representation as an Enabler

The following are some scenarios where representation of structure and structural data resources are key enablers of experimental and computational research.

Synthetic Chemistry

- Structure-based search of external and internal resources to identify prior art that can inform experimental design and avoid unnecessarily repeating effort.
- Search for availability of starting materials, and safety information related to these.
- Capture of reaction schemes, starting materials and products in a research information management system, linking these to specific batches or samples.
- Analysis of products and measurement of properties using a variety of analytical techniques to understand composition, configuration and characteristics of the structure of products.
- Determination of the 3D structure of products using experimental methods such as crystallography.
- Communication of research to colleagues and collaborators through reports and articles, conveying synthetic pathways and important structural features.
- Filing a patent claim that articulates a series of related novel compounds or materials discovered as a result of the research.

Computational Chemistry

- Generate starting models of structures for study using computational methods.
- Undertake computational experiments to model and simulate electronic structure and dynamic behaviour of molecules and materials using a range of software packages.
- Validate computational models of structures by comparison with experimentally derived structures.
- Calculate properties of molecules and materials and correlate these to structural features to develop predictive models.
- Store computational models of structures, share with colleagues and collaborators, and publish to enable reproducibility and reuse.

Data Science

- Data-mining of chemical structures and their properties to create datasets for AI and machine learning; often involves linking between resources based on structure.
- Train, validate and exploit models using structure representations and features either as labels on input data or as an intrinsic part of the machine learning methodology used.
- Store, share and publish successive versions of training and validation sets, including structure representations, to enable review and reuse of models.

Other approaches enabled by digital structure representation

- As part of input instructions for automated reaction processes [2].

- Performing virtual combinatorial library design and virtual high-throughput screening of potential drug molecules [3]; often involves interacting with structural biology models.
- In the *de novo* design of novel structures [4] and new materials [5] using computational and informatics approaches.

1.4 Mapping to the Research Data Lifecycle

The following summarises the role structure representation plays at different stages of the Jisc Research Data Lifecycle [6]:

- **Plan and design:** Retrieving information to inform experimental design. Designing inputs into experimental and computational studies.
- **Collect and capture:** Capturing reaction schemes and outcomes. Modelling 3D structure. Mining data resources and building predictive models.
- **Collaborate and analyse:** Sharing results with collaborators. Refinement and validation of models.
- **Manage, store and preserve:** Associating chemical identity, molecular structure and atomic composition with data, samples, and models.
- **Share and publish:** Reliably communicating structures and related data in forms understandable by researchers and machines. Linking data across domains and resources.
- **Discover, reuse, and cite:** Reliably retrieving data and models associated with specific substances to provide a foundation for future research.

1.5 Structure of this Report

The sections of this report are as follows:

Section 1: This introduction.

Section 2: The Challenges of Structure Representation

- The many ways that a structure may be digitally represented.
- Why multiple ways of representing structure should be supported.
- The scientific challenges of structure representation.
- The importance of capturing provenance of structures and their representation.
- Comments on culture and lessons we can learn from other domains.

Section 3: Workflow and Infrastructure

- How workflows and infrastructure should handle structure representation and the importance of capturing this as metadata.
- The social and technical constructs that can enable and encourage adoption of best practices by researchers.
- Considerations for the reliable storage, discovery and exchange of structural data across systems.

Section 4: Storage and Curation of Structural Data

- Where structural data can be found today.
- The importance of investing in data curation.
- The role of data repositories in supporting high quality data publication.
- A review of research data repository operating models.
- Guiding principles for sustainable and trusted Open Science infrastructure.

Section 5: Related Initiatives

- An overview of associated domain-specific and general research data initiatives.
- The importance of engaging with industry to help shape priorities and establish support.

Section 6: A summary of recommendations.

Within each section, key points that feed into final recommendations are highlighted in a box.

Context drawn from experiences specifically relating to management and publication of data in crystallography have a grey shaded background.

Wherever appropriate and possible, attempts have been made to support assertions and statements made in the report with references to articles and other supporting information. If there are no supporting references, then the report will be drawing primarily on the collective experiences and insights of its authors.

1.6 Authors of this report

The primary authors of this report have between them almost 50 years of experience working in the field of structural chemistry, research data publication and management, development of scientific information systems and software, and engagement with related initiatives worldwide. They are also drawing on the past and current activities of the Cambridge Crystallographic Data Centre (CCDC) [7] which has been providing data and software services to research communities in academia and industry for over 55 years.

1.6.1 The Cambridge Crystallographic Data Centre

The CCDC is a non-profit organisation that compiles and disseminates the Cambridge Structural Database (CSD) [8] , a trusted database of fully curated organic and metal-organic crystal structures. The widespread use of structural data worldwide more than 55 years after the CSD was established and the reliance on the CSD by industry and academia are evidence that the collation and curation of individual experiments enables generation of new scientific insights and knowledge.

Alongside curating the CSD, the CCDC develops scientific software to enable researchers to apply knowledge derived from the data to practical research problems. It also collaborates in research projects with academia and industry to advance the application of structural chemistry data. The Centre engages in the development of data standards in crystallography and chemistry and contributes to more general research data initiatives globally. It participates in scientific publishing initiatives and partnerships that are aimed at ensuring the timely and reliable publication of research data.

Since its beginnings in the University of Cambridge in 1965, the CCDC has evolved into an independent self-funding organisation that aims to balance the goals of open access to research data with the need for long-term preservation and sustainability. It is a Partner Institution of the University of Cambridge.

1.6.2 Primary Authors

Ian Bruno, Director of Data Initiatives, CCDC

Ian has a BSc in Chemistry from Durham University and a PhD in Information Science from the University of Sheffield. He has worked at the CCDC since 1993 in a variety of roles. As a Scientific Software Engineer, he contributed to the development of many of CCDC's core software products and has had management and leadership responsibilities for a range of software, data and science teams and activities.

In his current role as Director of Data Initiatives, Ian is responsible for establishing strategies that will make CCDC's data more readily discoverable and reusable by wider scientific communities. A specific focus is on the adoption and development of community principles, standards and sustainability models to support these aims. He is a regular speaker on topics relating to the sustainable stewardship and application of high-quality research data.

Ian is an active participant in a range of international initiatives relating to the standards and infrastructure required to support the sharing of research data. This includes involvement in activities of the International Union of Crystallography, the International Union of Pure and Applied Chemistry, the InChI Trust, the Research Data Alliance, FORCE11, the CoreTrustSeal Assembly of Reviewers, and DataCite Steering Groups.

Suzanna Ward, Head of Data and Community, CCDC

Suzanna has an MChem degree from the University of Southampton and began her career at the CCDC in 2006 as a Scientific Editor, working to curate and enrich crystal structures into the CSD. After taking on the responsibility for managing the team that creates the database in 2013, Suzanna became Head of Data and Community in 2019. In her current role Suzanna leads the team responsible for creating the CSD and oversees the Education and Outreach activities for the Centre.

Suzanna is an active member of the structural chemistry community, presenting, chairing and providing training sessions at many international conferences and crystallographic schools.

1.6.3 Other Contributors

The authors acknowledge their CCDC colleagues who have contributed to the content and review of this report, in particular Stephen Holgate, Philip Andrews, Paul McKeown, Clare Tovee, Seth Wiggin, David Lowe, Susan Reutzel-Edens, Jason Cole, Stewart Adcock, Carmen Nitsche, and Jürgen Harter.

Thanks are due to Leah McEwen, Chair of IUPAC's Committee on Publications and Cheminformatics Data Standards, for review of sections relating to IUPAC activities. We also thank members of the PSDI Pilot team for guidance and support, in particular Simon Coles and Nicola Knight.

Finally, we acknowledge the Engineering and Physical Sciences Research Council for funding that has supported this work (Grant number EP/W032252/1) [9].

1.7 References

1. M. D. Wilkinson *et al.* (2016) "Comment: The FAIR Guiding Principles for scientific data management and stewardship", *Scientific Data*, 3(1), https://doi.org/10.1038/sdata.2016.18.

2. A. J. S. Hammer, A. I. Leonov, N. L. Bell, and L. Cronin (2021) "Chemputation and the Standardization of Chemical Informatics", *JACS Au*, 1(10), https://doi.org/10.1021/jacsau.1c00303.

3. G. Schneider "Trends in Virtual Combinatorial Library Design", *Current Medicinal Chemistry*, 9(23), https://doi.org/10.2174/0929867023368755.

4. V. D. Mouchlis *et al.* (2021) "Advances in De Novo Drug Design: From Conventional to Machine Learning Methods", *International Journal of Molecular Sciences*, 22(4), https://doi.org/10.3390/ijms22041676.

5. J. J. de Pablo *et al.* (2019) "New frontiers for the materials genome initiative", *npj Comput Mater*, 5(1), https://doi.org/10.1038/s41524-019-0173-4.

6. Jisc (2018) "Research data management toolkit", https://www.jisc.ac.uk/full-guide/rdm-toolkit (accessed 25 April 2022).

7. "The Cambridge Crystallographic Data Centre (CCDC)", https://www.ccdc.cam.ac.uk/ (accessed 25 April 2022).

8. C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward (2016) "The Cambridge Structural Database", *Acta Cryst B*, 72(2), https://doi.org/10.1107/S2052520616003954.

9. EPSRC (2021) "Physical Sciences Data Infrastructure (PSDI) Phase 1 Pilot", https://gow.epsrc.ukri.org/NGBOViewGrant.aspx?GrantRef=EP/W032252/1 (accessed 25 April 2022).

2 The Challenges of Structure Representation

Contents

2.1 Overview

In this section we describe the different ways in which structures can be represented digitally, noting limitations and related considerations. Throughout, we make recommendations for capabilities that should be considered as part of a Physical Sciences Data Infrastructure relating to structure representation.

2.2 Everyday Interactions with Structure Representation

For many researchers in the physical sciences, a structure is something they create a diagram of in a drawing package, embed in a manuscript and later see in a static form in an article or report. How the structure is represented behind the scenes will be of little concern to them.

Others may spend more of their time interacting with 3D models of structures generated via computational or experimental means. The representation of their structures will be tied up with other output generated by modelling and analysis packages with which they may be very familiar.

For some, the structure representation may just be a label or avatar they associate with a set of measured properties, or with a more abstract model of a material divorced from the chemical composition of the substance being studied.

Those invested in data science disciplines such as cheminformatics will, by contrast, be very familiar with different structure representation formats, their limitations and the challenges of converting between these.

A general consideration for a Physical Sciences Data Infrastructure should be the creation of pathways that can channel the different ways researchers naturally interact with structures into representations that can be readily exploited by data scientists and their machines.

2.3 Structure Representation Types

There is no shortage of ways to represent structures. Open Babel [1], a toolbox designed to convert data from a range of chemistry-related domains, boasts support for over 110 chemical file formats [2]. Open Babel's categorisation of these formats indicates the different purposes for which they have been defined:

- Common cheminformatics formats (standards and de facto standards)
- Proprietary cheminformatics formats
- Computational chemistry formats
- Crystallography formats
- Reaction formats
- 2D drawing formats
- 3D viewer formats
- Kinetics and thermodynamics formats
- Molecular dynamics and docking formats
- Biological data formats

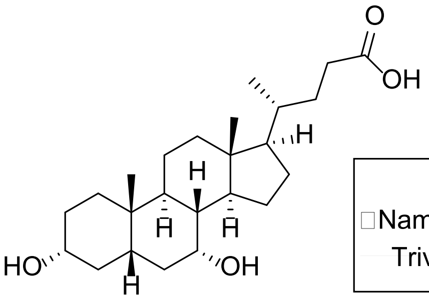
For the purposes of this analysis, we have grouped representation formats based more on the following intrinsic characteristics which we elaborate on in subsequent sections:

- Designed for humans
- Abstract identifiers

- Linear representations
- Standard identifiers
- Connection table representations
- 2D diagrams
- 3D representations
- Mixtures and reactions
- Universal formats
- Structure classifiers

2.3.1 Representations Designed for Humans

These include chemical formulae, chemical names and images of chemical diagrams and are almost certainly the main ways most structures are communicated today. Whilst intended primarily for communication of structural information to people via the printed page, they can to some degree be interpreted by software to generate representations more appropriate for analysis by machines.



Chemical Formula: $C_{24}H_{40}O_4$
□Name: 3α,7α-Dihydroxy-5β-cholanic acid
Trivial Name: Chenodeoxycholic acid

Figure 1: Formula, name and diagram of Chenodeoxycholic acid.

There are software tools available that can convert a chemical name to a more structured representation [3–8], particularly if the name conforms to IUPAC nomenclature rules [9]. Software that can reverse engineer static images of structures in graphics formats to richer representations has been available for a number of years [10] and AI/Machine Learning is enabling further innovations in this space [11,12].

There are, however, limitations to how reliably these human-oriented manifestations can be converted to representations appropriate for processing by machines. Whilst a formula can be deconstructed into a chemical composition, it cannot be guaranteed to convey a particular structural isomer. Generating a structure from a name is unlikely to be possible if a trivial name is used instead of a systematic one. Artefacts in graphical images can easily confuse deconvolution of these resulting in incorrect interpretation.

Formats intended primarily for communication with people should not be completely written-off as useless for machines, particularly if that is all that is available. It would, however, be rash to rely on representations generated from these alone without some form of subsequent validation.

2.3.2 Abstract Identifiers

We use the term "abstract identifiers" to refer to identifiers that on their own will not convey what a structure is without reference to another resource.

CAS Registry Numbers ®: A widely recognised identifier often associated with a chemical substance is the Chemical Abstracts Service (CAS) Registry Number® [13]. CAS catalogues millions of chemical substances and its Registry Number uniquely identifies these.

Regulatory Identifiers: Similar in spirit to the CAS Registry Number is the Unique Ingredient Identifier (UNII) used by the US Federal Drug Agency to identify substances in drugs, biologics, foods, cosmetics and other regulated entities. A UNII is a unique and unambiguous identifier generated using molecular structure but is ultimately non-semantic [14]. European regulatory agencies use a European Commission (EC) Number which is a sequential number assigned to a substance in official chemical inventories [15].

Accession IDs: Accession IDs uniquely identify records in data resources associated with a particular structure. However, depending on how a resource is organised, one structure may have multiple associated IDs so these cannot be assumed to be unique.

Identifier Type	Identifier	identifiers.org Link
CAS Registry Number	474-25-9	https://identifiers.org/cas:474-25-9
UNII	UNII 0GEI24LG0J	https://identifiers.org/unii:0GEI24LG0J
EC Number	207-481-8	
CCDC Accession ID	CCDC 1281373	https://identifiers.org/ccdc:1281373
ChEMBL Accession ID	CHEMBL240597	https://identifiers.org/chembl.compound:CHEMBL240597
PubChem Accession ID	Compound 10133	https://identifiers.org/pubchem.compound:10133
DrugBank Accession ID	DB06777	https://identifiers.org/drugbank:DB06777

Table 1: Abstract Identifiers associated with Chenodeoxycholic acid and identifiers.org links where registered.

Common to all of these identifiers is that the only way to translate from these to a machine representation of a structure is by looking them up either in the resource with which they are directly associated or another resource that may store these. As illustrated in Table 1, this is made easier if identifiers have been registered with a service such as identifiers.org [16]. To get to the actual structure associated with the ID will require navigation through database records which will not be organised in a consistent way across resources. If the identifier is associated with a resource that has restricted access, then comprehensive lookup of the associated substance will not be possible for everyone.

Despite their limitations, abstract identifiers associated with a chemical structure provide a potentially powerful way of linking related structural information across different data resources so should be captured when available.

2.3.3 Linear Representations

Linear representations aim to represent as a string the connectivity of a structure, i.e. which atoms are connected and by which bond types, and in so doing, unambiguously represent the molecular structure of a substance. Early linear representations such as Wiswesser Line Notation [17] were key enablers of early chemical information systems but are now largely consigned to the history books. Here we focus on SMILES which has been widely adopted as a backbone of many cheminformatics systems and workflows.

SMILES (Simplified Input Line Entry System) was originally developed by and is proprietary to Daylight Chemical Information Systems [18]. An open specification has been developed by the community and is available as OpenSMILES [19]. SMILES has a structure-based query counterpart (SMARTS) and a related reaction transformation language (SMIRKS). Despite its wide use as a *de facto* standard in cheminformatics, SMILES is not without its limitations.

Whilst the SMILES string for a well-defined organic molecule is expected to be unambiguous, it will not necessarily be unique. Not only can the order of atoms in a string vary between equivalent representations but different conventions can be chosen to represent aromaticity. For any one molecule there could be many millions of different SMILES representations [20]. This means that the temptation to compare SMILES strings as an easy way to establish chemical identity must be approached with caution. SMILES can be canonicalized to provide consistent representations but there is more than one way to canonicalize a SMILES string and no universal standard.

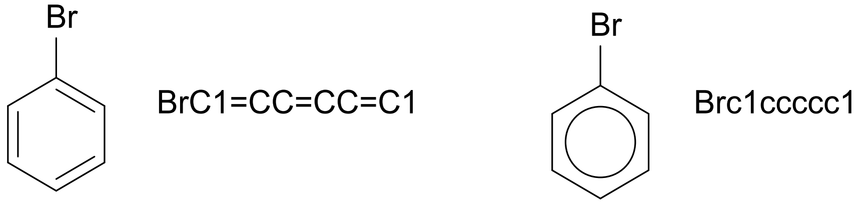


Figure 2: SMILES strings for Bromobenzene illustrating how different but equivalent representations of Bromobenzene result in different SMILES strings.

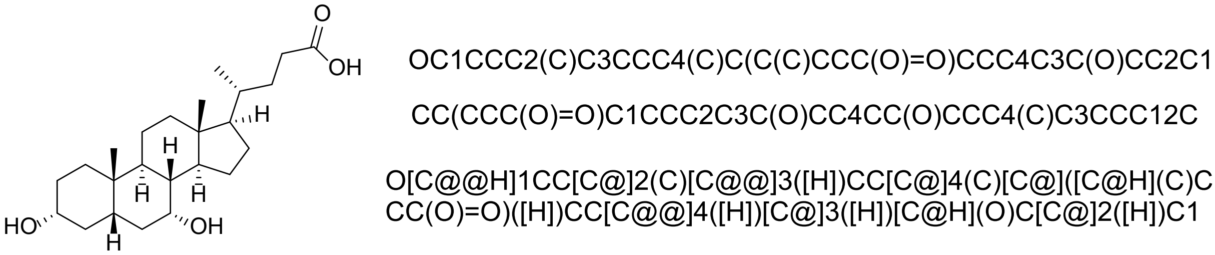

O
OH
H
H
H
H
HO
H
OH
OC1CCC2(C)C3CCC4(C)C(C(C)CCC(O)=O)CCC4C3C(O)CC2C1
CC(CCC(O)=O)C1CCC2C3C(O)CC4CC(O)CCC4(C)C3CCC12C
O[C@@H]1CC[C@]2(C)[C@@]3([H])CC[C@]4(C)[C@]([C@H](C)C
CC(O)=O)([H])CC[C@@]4([H])[C@]3([H])[C@H](O)C[C@]2([H])C1

Figure 3: SMILES strings for Chenodeoxycholic acid. The top two strings indicate that different ordering of atoms results in different SMILES strings. The bottom string is an isomeric SMILES representation which reflects atom stereochemistry.

To reliably use a representation such as SMILES, it must be interpreted rather than compared directly. Even here there are pitfalls to be aware of. The SMILES specification is not necessarily unambiguous or complete which has led to different implementations of SMILES interpreting the same string in different ways. Further, different implementors may choose bespoke extensions to accommodate missing features that end up being in conflict. An IUPAC project group is currently working to establish definitive guidelines that, if followed by all those providing support for SMILES, would avoid these pitfalls [21].

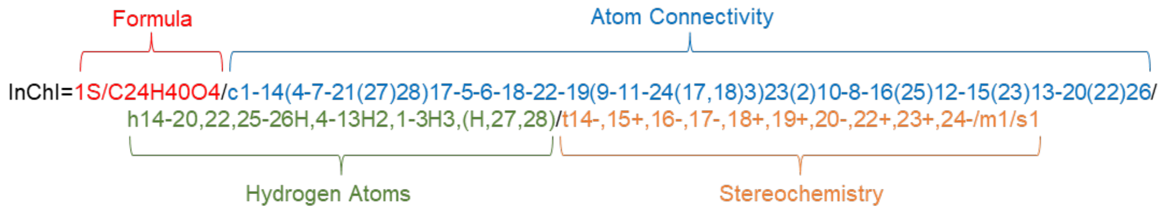
Attempts to use SMILES strings in the context of some machine learning approaches have led to the development of alternative linear representations that are similar to SMILES but less likely to become invalid when manipulated as part of model building [22,23].

2.3.4 InChI – A Unique and Unambiguous Standard Identifier

It was very much with challenges around uniqueness and ambiguity in mind that the IUPAC International Chemical Identifier (InChI) was conceived [24]. The aim of InChI is to provide a unique and unambiguous string that is identical for different but equivalent representations of the same structure. The InChI achieves this by normalising the representation provided according to rules encoded in the standard InChI generator. Key facets of InChI are:

a. InChI strings can be reliably compared to establish the identity of two different structures independent of bonding conventions chosen.
b. InChI strings can be compared at different levels of specificity because the InChI string has a layered format separating out facets such as connectivity, charge and stereochemistry.
c. InChI strings can be converted back into the input structure but because of normalisation, this may not reflect the bonding representation provided and may be one that is considered unusual by a typical chemist.

An InChI string can be hashed into an InChIKey which is a more compact form of the identifier, introduced to avoid pitfalls with search engines that might not ideally handle the more verbose InChI string [25]. An InChIKey cannot be converted directly back to the original structure so has the same characteristics as the abstract identifiers described in Section 2.4.2. It reflects InChI layering to a limited degree and will for most practical purposes uniquely distinguish different structures[a].


Formula
Atom Connectivity
InChI=1S/C24H40O4/c1-14(4-7-21(27)28)17-5-6-18-22-19(9-11-24(17,18)3)23(2)10-8-16(25)12-15(23)13-20(22)26/
h14-20,22,25-26H,4-13H2,1-3H3,(H,27,28)/t14-,15+,16-,17-,18+,19+,20-,22+,23+,24-/m1/s1
Hydrogen Atoms
Stereochemistry

InChI Key: RUDATBOHQWOJDD-BSWAIDMHSA-N

Figure 4: InChI and InChIKey for Chenodeoxycholic acid with an indication of how the InChI string is layered. These will be the same for any equivalent representation of Chenodeoxycholic acid.

InChI has been widely adopted to facilitate searching and linking across large data resources [27]. It works well for discrete organic molecules but does have limitations beyond this. These limitations are being actively worked on by the wider community through projects sponsored by IUPAC and the InChI Trust who both have oversight of the standard; this is further outlined in Section 5.

We consider it essential that a future infrastructure exposes InChIs and InChIKeys for structures where these can be meaningfully generated as a key enabler of findability and interoperability based on chemical structure.

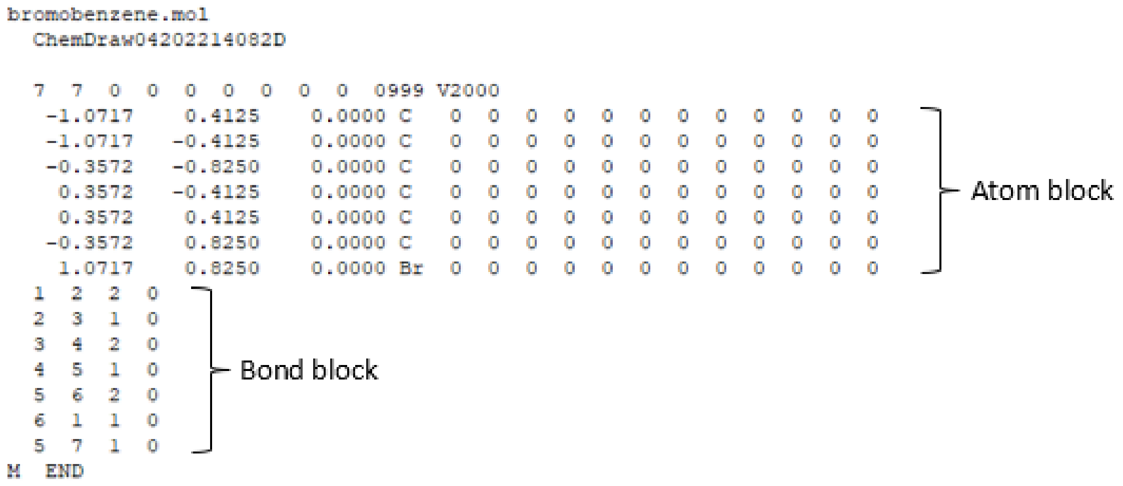
We would note that researchers should not themselves necessarily have to generate InChIs – instead they could be asked or enabled to provide structures in a format from which an InChI can be generated. The InChI Generator supports MOLfile (described in Section 2.4.5) but a format from which a MOLfile can be generated, such as SMILES, could also suffice. It is also possible to interface directly to the InChI Generator via its API (Application Programming Interface) which can help avoid limitations of intermediate file formats and generally results in better performance when generating InChIs in bulk.

2.3.5 Connection Table Representations

The “connection table” has long been a bedrock of chemical structure representation [28]. Simply put, a connection table is a list of atoms in a molecule and a list of the bonds connecting

[a] There is a finite but very small probability of different InChIs being hashed to the same InChIKey [26].

these. Mathematically, it defines nodes and edges of a graph that are annotated with attributes of atoms and bonds respectively. As such it provides a representation that is ripe for the application of methods adapted from Graph Theory that underpin many structure-based searching systems. If 2D coordinates are associated with the nodes of the graph in a connection table, you then have a machine representation of a chemical diagram.



```
bromobenzene.mol
  ChemDraw04202214082D

  7  7  0  0  0  0  0  0  0  0999 V2000
   -1.0717    0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.0717   -0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3572   -0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.3572   -0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.3572    0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3572    0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.0717    0.8250    0.0000 Br  0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0
  2  3  1  0
  3  4  2  0
  4  5  1  0
  5  6  2  0
  6  1  1  0
  5  7  1  0
M  END
```

Figure 5: A connection table format (MOLfile) for Bromobenzene. The Atom block lists atoms, their elements and coordinates and may specify other atomic properties. The Bond block defines bonds between atoms using indices that reflect the ordering of atom in the Atom block and a numerical encoding of bond type.

The most widely known file format based on a connection table representation is perhaps the MOLfile, a member of the family of CTfile formats initially published by MDL Information Systems in 1992 [29]. Today, the specification of these formats is maintained by BIOVIA [30]. In addition to the connectivity of atoms in a molecule, facets such as charge, atom stereochemistry, radical state and isotopes can all be specified. Some features of the MOLfile worthy of note:

- It is not uncommon for hydrogens to be omitted from a MOLfile with the expectation that these can be inferred based on the heavy atom count. Whilst this is reasonable for well-defined organic molecules, it does not work as well for metal-organics.
- Bond stereochemistry can only be defined based on the coordinates of the atoms – if these are missing or incorrect, bond stereochemistry cannot be reliably inferred.
- The MOLfile has good support for functional group symbols that enables both a compact diagram and a semantically correct representation of a structure to be captured simultaneously.

The most widely used version of the MOLfile is V2000. A more recent V3000 version has been specified to address limitations of V2000 but has not yet been widely adopted. Advantages of V3000 include:

- Free format rather than fixed format makes parsing easier and removes limitations imposed by fixed field widths.
- Support for enhanced stereochemistry.
- Use of templates to support better representation of large molecules.
- Broader range of bond types, in particular a “coordination” bond that if adopted could enable more reliable representation of metal-organics.

We hypothesise that the reason that V3000 has not been widely embraced is because there are many well-supported systems based on V2000 that have no compelling need to adopt the new features of V3000 at this time.

Strictly speaking, a MOLfile represents a single molecule. Multiple MOLfiles for different structures concatenated into a single file is an SDfile. An SDfile allows additional data properties to be associated with a structure, essentially enabling representation of a database record containing a structure. Often, the terms MOLfile and SDfile are used interchangeably.

The CTfile family of formats also extend to allow structural data for reactants and products of a reaction to be represented in a Rxnfile. As well as capturing the identity of the molecules involved, it is possible to map atoms across reactants and products which is an important feature for systems wanting to manage and exploit reaction information at a structural level.

The MOLfile allows various query features to be specified within the same representation.

2.3.6 2D Chemical Diagram Representations

Whilst a machine representation of a structure can be richly represented in formats such as MOLfile, this is not the entry point for most researchers. Instead, they are likely to use one of the chemical structure drawing packages listed in Table 2.

Package	Provider	URL
ChemDraw	PerkinElmer	https://perkinelmerinformatics.com/products/research/chemdraw/
BIOVIA Draw	Dassault Systèmes	https://discover.3ds.com/ctfile-documentation-request-form
ChemSketch	ACD/Labs	https://www.acdlabs.com/resources/freeware/chemsketch/
ChemDoodle	iChemLabs	https://www.chemdoodle.com/
ChemWindow	Wiley	https://sciencesolutions.wiley.com/chemwindow-chemical-structure-drawing-software/

Table 2: Common chemical structure drawing packages.

These diagram drawing packages are universally likely to support export of structures in a range of recognised representation including MOLfile, SMILES and InChI. However, researchers are much more likely to rely on copying and pasting diagrams into documents and saving their structures in formats that are specific and proprietary to the drawing package being used.

One of the most widely used drawing packages is ChemDraw, for which the proprietary format has been published [31]. This makes it possible to parse files saved in native ChemDraw format to extract chemically meaningful objects. However, if a user of these packages has chosen to bypass the chemical drawing functionality of such a package and use generic shapes to e.g. represent benzene as a hexagon and a circle, no chemical meaning will be conveyed. This may seem an unlikely thing for a self-respecting researcher to do but is used to illustrate that the meaning intended by a visual drawing may not be manifest in a digital representation of the diagram saved in a format specific to a drawing package.

2.3.7 3D Representations

If a file format can support 2D coordinates to enable layout of a 2D diagram, then it is no great stretch to envisage these adding a third coordinate and enabling representation of structures in 3D. Indeed, both MOLfile and ChemDraw files have the ability to do this.

A commonly encountered connection table format that also does this is the MOL2 file format [32]. This was originally conceived by Tripos as part of their Sybyl molecular modelling package very much with 3D representation in mind. MOL2 supports the concept of atom typing that allows the user to distinguish e.g. an sp3 carbon from an sp2 one, an sp2 oxygen from a carboxylate one, an amide nitrogen from a trigonal planar one. These distinctions are particularly important when computationally modelling structures. In addition, MOL2 can convey information required to generate a crystal packing of a structure and has features that support representation of biological macromolecules.

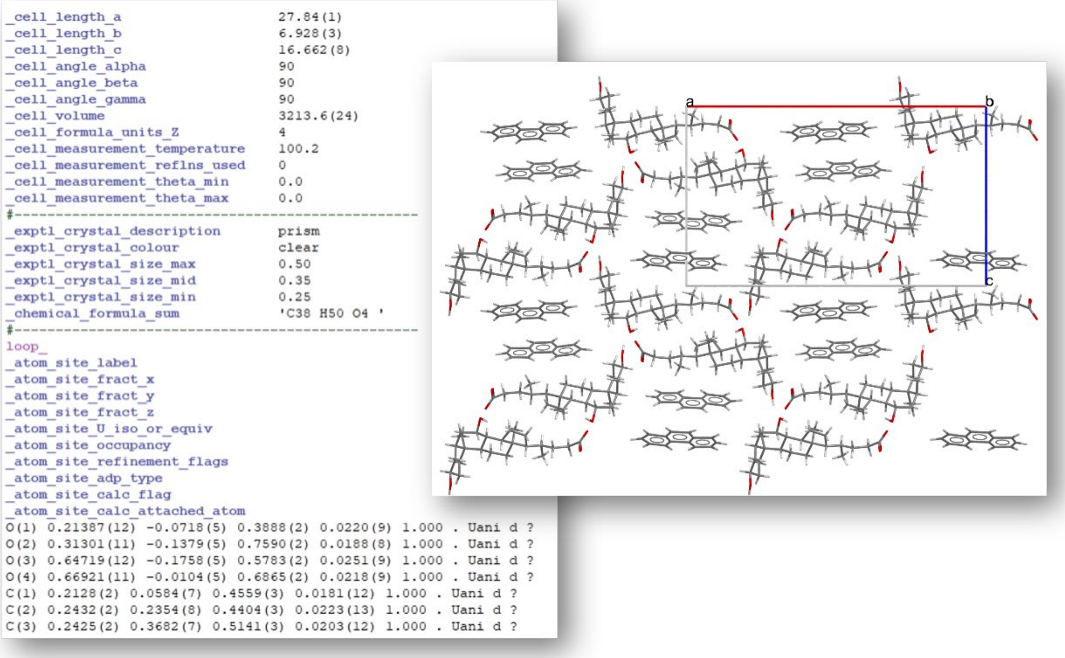
It is important not to think of 3D representation as just an extension of 2D representation, as in many areas of physical sciences the 3D structure is the primary focus of the study being undertaken. In the study of inorganic materials, for example, a chemical diagram may not offer more than a simple formula because it is the 3D arrangement of atoms in the material that is key to understanding the structural properties and behaviour.

The ability to store and display 3D structures should be baked into a future Physical Sciences Data Infrastructure from the start.

It is also important to consider the provenance of 3D structures, whether experimentally determined or computationally calculated; the experimental methods or software algorithms and parameters used; and additional metadata that might be important for a researcher to understand how the 3D model of the structure was arrived at. We say more on this in Section 2.6.1.

2.3.7.1 Experimental 3D Structures

Chemical crystallography has a well-established file format, the Crystallographic Information File (CIF), that enables many aspects of a diffraction experiment to be captured [33]. As well as the model of the structure that has been determined, a CIF can capture properties of the sample studied, the equipment and software used, and the software and parameters involved in processing data and refining the model. The format is supported by CIF dictionaries that clearly define the data items expected in the file enabling the structure and associated information to be semantically represented in a way that a machine can reliably interpret [34].



```
_cell_length_a                    27.84(1)
_cell_length_b                    6.928(3)
_cell_length_c                    16.662(8)
_cell_angle_alpha                 90
_cell_angle_beta                  90
_cell_angle_gamma                 90
_cell_volume                      3213.6(24)
_cell_formula_units_Z             4
_cell_measurement_temperature     100.2
_cell_measurement_reflns_used     0
_cell_measurement_theta_min       0.0
_cell_measurement_theta_max       0.0
#------------------------------------------------
_exptl_crystal_description        prism
_exptl_crystal_colour             clear
_exptl_crystal_size_max           0.50
_exptl_crystal_size_mid           0.35
_exptl_crystal_size_min           0.25
_chemical_formula_sum             'C38 H50 O4 '
#------------------------------------------------
loop_
_atom_site_label
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_U_iso_or_equiv
_atom_site_occupancy
_atom_site_refinement_flags
_atom_site_adp_type
_atom_site_calc_flag
_atom_site_calc_attached_atom
O(1) 0.21387(12) -0.0718(5) 0.3888(2) 0.0220(9) 1.000 . Uani d ?
O(2) 0.31301(11) -0.1379(5) 0.7590(2) 0.0188(8) 1.000 . Uani d ?
O(3) 0.64719(12) -0.1758(5) 0.5783(2) 0.0251(9) 1.000 . Uani d ?
O(4) 0.66921(11) -0.0104(5) 0.6865(2) 0.0218(9) 1.000 . Uani d ?
C(1) 0.2128(2) 0.0584(7) 0.4559(3) 0.0181(12) 1.000 . Uani d ?
C(2) 0.2432(2) 0.2354(8) 0.4404(3) 0.0223(13) 1.000 . Uani d ?
C(3) 0.2425(2) 0.3682(7) 0.5141(3) 0.0203(12) 1.000 . Uani d ?
```

Figure 6: Excerpt from a CIF deposited as CCDC 224567[b] alongside a visualisation of the crystal structure it represents. The CIF includes the atomic coordinates and cell parameters determined as part of a diffraction study alongside chemical and physical properties of the sample studied and experimental conditions and parameters. Data items in blue beginning with an underscore (_) are all precisely defined in accompanying CIF dictionaries.

Whilst the 3D arrangement of atoms in a molecule is reported, and sometimes an indication of which atoms are considered to be bonded, it is rare to find a complete chemical representation of the structure studied. CIF data items for capturing a 2D representation of the structure do exist but are rarely used. This means that a chemical representation of the 3D model must be separately generated. Whilst this can be attempted automatically from the 3D model [35], such processes cannot be guaranteed to give an accurate result and further human validation is required to ensure that the resulting representation is correct.

2.3.7.2 Computed Structures

Representation of 3D structure is key to many computational chemistry studies involving calculations and simulations of atomic and molecular properties, systems and behaviours. Close to half of the formats supported by OpenBabel relate in some way to computational chemistry which is perhaps indicative of the many methods and software packages used in a computational chemistry context.

Simulated Crystal Structures

The challenges of capturing computational methods used has been made manifest when attempting to define reporting standards for Crystal Structure Prediction (CSP) methods as part of the 7th CSP Blind Test organised by the CCDC [36]. Steps involved in CSP may include:

- Generation, optimisation and clustering of conformers
- Generation, optimisation and clustering of structures

[b] T. Fukami, K. Yamaguchi, Y. Tozuka, K. Moribe, T. Oguchi, and K. Yamamoto (2004) "CCDC 224567: Experimental Crystal Structure Determination". Cambridge Crystallographic Data Centre. https://doi.org/10.5517/CC7JP3J.

- Energy calculations to rank predicted structures.

Methods involved include force field calculations, semi-empirical methods, DFT, wavefunction calculations, evolutionary algorithms, simulated annealing, Monte Carlo sampling and approaches based on AI and Machine Learning. Different software packages with different parameters may be involved at different stages and the overall process is likely to be iterative rather than linear.

It is beyond the scope of this report to delve into solutions for the reporting of computational chemistry methods, but in Table 3 we offer an indication of some of the approaches we are aware of for reliably capturing the results of such studies and the methods used. We observe that some of these relate to Chemistry Markup Language which will be discussed in Section 2.4.9.

Title	Year	Ref.
The semantics of Chemical Markup Language (CML) for computational chemistry	2012	[37]
From data to analysis: linking NWChem and Avogadro with the syntax and semantics of Chemical Markup Language	2013	[38]
GNVC: Gainesville Core Ontology - standard for publishing results of computational chemistry	2015	[39]
Code interoperability and standard data formats in quantum chemistry and quantum dynamics: The Q5/D5Cost data model	2014	[40p. 5]
An Ontology and Semantic Web Service for Quantum Chemistry Calculations	2019	[41]
mwfn: A Strict, Concise and Extensible Format for Electronic Wavefunction Storage and Exchange	2021	[42]
TREXIO: a standard format for storing wave functions	2022	[43]
AiiDA: automated interactive infrastructure and database for computational science	2016	[44] [45]

Table 3: A selection of references to formats, ontologies and other tools designed to enable the reliable representation and exchange of results from computational chemistry studies.

2.3.8 Mixtures and Reactions

Much of what we have discussed thus far concerns the representation of individual molecules, but we must remember that many of the substances and materials studied in the physical sciences are comprised of multiple components. This includes simple salts, hydrates and solvates through to complex mixtures that are the product of a reaction or process. The ratio of components in a mixture may be precisely known, or it could be somewhat vague: for example, an unknown ratio of hexanes of unknown structural formula.

Some multi-component systems can be non-stochiometric, i.e., it is impossible to express the ratio of products using whole numbers. This can be related to an individual atom, for example an individual position in a molecule with a fractional ratio of two different elements. It can also be the charge on a molecule that is fractional, as seen in charge transfer salts where several organic molecules, each with a partial charge, are needed to balance the integer charge on a different molecule.

There are well-established methods for representing the mixture of starting materials and products that define a reaction. These include the Rxnfile referenced in Section 2.4.5, and RInChI [46], an application of InChI used to identify and catalogue reactions. More recently, a

Mixfile format and an application of InChI, MInChI, have been proposed as a means to achieve the same for mixtures more generally [47].

Consideration should be given to the representation and storage of multi-component systems including reactions and mixtures and not just individual molecular structures.

2.3.9 Universal Formats

Throughout the history of structure representation, there have been attempts to establish a universal representation format for chemical structures.

- In 1990, John Barnard published a draft specification for a Standard Molecular Data Format (SMD) that had been developed with input from European chemical and pharmaceutical companies [48].
- Five years later Barnard combined forces with the originators of CIF to propose a Molecular Information File (MIF) based on the STAR framework that underpins CIF [49].
- In a perspective article in 1996, Englebert Zass noted that standards such as SMD and MOLfile existed and were used but not universally enough [50].
- In 2004, Jan Noordik [51] drew attention to the challenges of exchanging structural information because connection formats differ from program to program despite attempts at standardization.
- In a 2009 review article [52], Peter Willett noted that there were many formats described in the literature but down-played the disadvantages observing that “it is normally possible to convert from one to another without too much trouble”. By this point, the Open Babel toolbox for converting between structures had been available for the best part of a decade.

Zass referred to both SMD and MOLfile when making his comments about lack of universal adoption of standards and it is interesting to note that MOLfile has come to dominate whilst SMD, the attempt at a universal standard, has become just a footnote in the history of structure representation. We venture that the reason for this is the tooling that was available for the MOLfile – initially through proprietary systems developed by MDL and later in more open toolkits wanting to interoperate with data from these systems. Similarly, the crystallographic community invested in tools and workflows to support use of CIF whilst no similar effort was made around support of MIF – CIF has thrived, MIF has not.

Successful adoption of structure representation formats requires well-supported tools.

Another universal attempt which recognised the importance of tooling is the XML-based Chemistry Markup Language (CML) [53–60]. This is more than just a format in two regards: (1) it is backed by an XML schema which semantically define the contents of the file and (2) it was released with tools and convertors to support the generation and manipulation of CML files. CML built in support for molecules, reactions, spectra and analytical data, computational chemistry, chemical crystallography, and materials. It set out to incorporate existing XML schema and dictionaries such as CIF rather than reinvent these. The vision was to provide a data structure for computation and a semantic infrastructure for physical science.

Whilst there are pockets of use of CML, it has never been widely adopted despite its visionary approach. One might argue that its main failure was that it was ahead of its time relative to the community it wished to support. Technology has moved on since CML was first conceived but the ideas behind it are quite relevant to the infrastructure of today.

A more recent attempt at a universal format for structure and associated data is the Unified Data Model [61], developed through the Pistoia Alliance. The current version of this has support for molecules, reactions, referencing and embedding of analytical data and samples. Like CML, it draws on existing ontologies, taxonomies and standards rather than re-defining these.

If considering semantic data models for capturing data, computations and experiments, a Physical Sciences Data Infrastructure should look to past and present initiatives for inspiration, pre-existing practical solutions, and lessons learned.

2.3.10 Structure Classification

Much of what we have discussed in this section relates to precise representation of structures. In more complex systems, structure classification can be a convenient way of categorising and grouping structures.

2.3.10.1 Semantic Classification

The ChEBI ontology [62], CO (The Chemical Ontology) [63], ChemOnt [64] and the Chemistry Vocabulary [65] all allow structures to be classified based on features such as chemical class and functional groups. Some aim towards a general classification whilst others are designed with the scope of a particular domain or data resource in mind. CO and ChemOnt are supported by tools that allow automatic classification of a structure (checkmol and ClassyFire respectively). These classifications can enable measures of semantic similarity and structural similarity to complement each other, as well as facilitate exchange of information across systems in a common language.

ChEBI Roles Classification	
Chemical Role(s)	Bronsted acid
Biological Role(s)	Human metabolite
	Mouse metabolite

ClassyFire categorisation	
Kingdom	Organic compounds
Super Class	Lipids and lipid-like molecules
Class	Steroids and steroid derivatives
Sub Class	Bile acids, alcohols and derivatives
Direct Parent	Dihydroxy bile acids, alcohols and derivatives

Figure 7: Ontological classifications of Chenodeoxycholic acid.

2.3.10.2 Systematic Classification

There are some types of structures for which it is possible to systematically generate classifiers based on their structural features.

Metal Organic Frameworks (MOFs) for example are often categorised by their topology, either considering the framework or the void space in pores and channels [66]. There are several systems currently in use for deriving descriptors for MOFs [67–70]. These lend themselves to varying degrees of automation. There are inherent challenges with the definition of these descriptors, as networks may have ambiguous topology based on the assignment of nodes and linkers. More rigorous solutions for representing the diverse characteristics present in structures such as MOFs are being explored [71,72] and a CIF dictionary for capturing the facets of topological descriptors has been proposed [73].

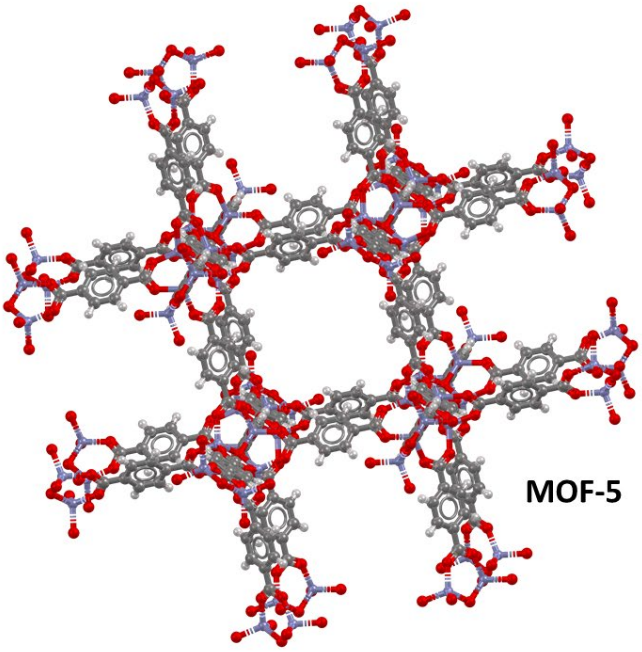


Figure 8: The 3D structure of a metal organic framework (catena-(tris(μ$_4$-Terephthalato)-(μ$_4$-oxo)-tetra-zinc)[c] *along with its MOF ID from the Reticular Chemistry Naming and Numbering Database [74].*

Automated classification approaches have aided the generation of subsets of MOFs for use in computational analysis [75,76]. Structures in these subsets are "cleaned up" to remove uncertainty associated with labile or disordered parts of a structure. A choice that has to be made here is whether to manually review the outcome of automated classification and processing to ensure that structures have been appropriately included. The main consideration is balancing the manual effort required against the impact of inappropriate structures on downstream research.

Interlocked molecules such as catenanes, rotaxanes and knots are another category of structures where systematic classification might be used to identify classes with similar topology [77]. It is also possible to generate classification of hydrogen-bonding networks based on Graph Set analysis that will distinguish motifs such as chains and rings of varying sizes [78].

2.3.10.3 Polymorphs

A substance can exist in different crystal structures and these different crystal forms are referred to as polymorphs. Different polymorphs are distinguished by labels. Typically, the first person to discover an instance of polymorphism will decide the labels to use within a polymorph family, and this will be an arbitrary categorisation or sequence. These labels are important to enable researchers to know that there are polymorphs and distinguish one from the other, yet there is no universally adopted standard for naming polymorphs. Instances have

[c] N. Lock *et al.* (2013) "CCDC 938392: Experimental Crystal Structure Determination". Cambridge Crystallographic Data Centre. https://doi.org/10.5517/CC10HGQP.

arisen where different researchers have labelled the same polymorph differently, leading to conflict and confusion.

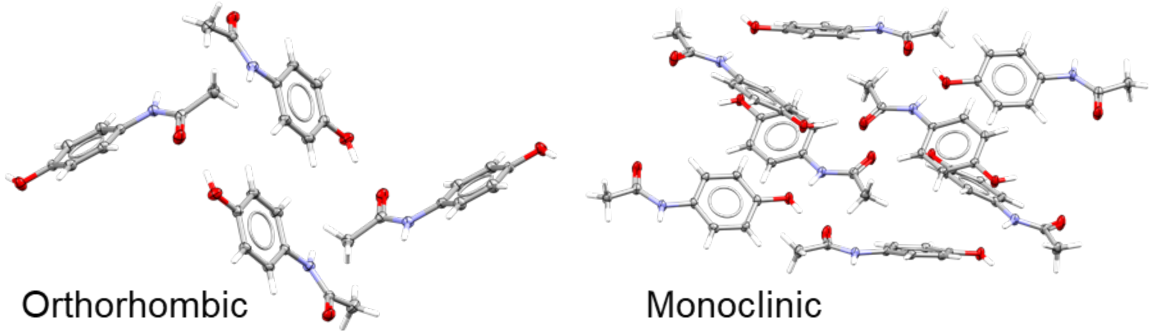


Figure 9: Different packing arrangements of Polymorph II (Orthorhombic)[d] and Polymorph I (Monoclinic)[e] of paracetamol.

2.3.10.4 Manual Classification

Generally, classifications are useful for retrieving classes of related structures where a structure-based search would not be able to unambiguously do this. Classifications that have proved useful within the Cambridge Structural Database (CSD) in this regard include carbohydrates, peptides, porphyrins, steroids, terpenes and alkaloids; assigning these typically requires manual intervention.

A future infrastructure should support the generation and storage of a range of structure classifications that enable the identification, retrieval, clustering and analysis of related structures.

[d] C. Nichols and C. S. Frampton (2000) "CCDC 135452: Experimental Crystal Structure Determination". Cambridge Crystallographic Data Centre. https://doi.org/10.5517/CC4JYF0.

[e] C. Nichols and C. S. Frampton (2000) "CCDC 135451: Experimental Crystal Structure Determination". Cambridge Crystallographic Data Centre. https://doi.org/10.5517/CC4JYDZ.

2.4 Scientific Limitations

2.4.1 Metal-containing Compounds

Many of the representation formats described in the preceding sections have been developed with well-defined organic molecules in mind. When we introduce metals into the mix, representation conventions can be found lacking because of the need to distinguish delocalised, dative and multi-centre pi bonding concepts and cope with atom valences not typically found in organic chemistry.

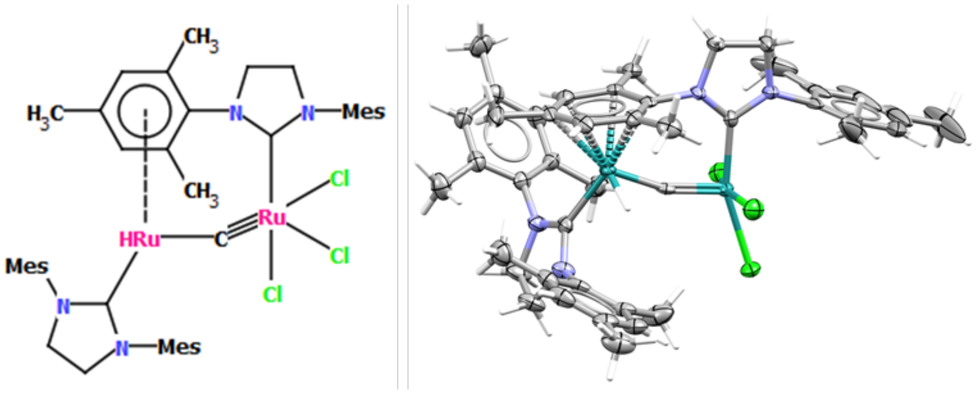


Figure 10: An example of different metal bonding modes in CSD Entry ABAFIT[f] including aromatic η6, carbene and triple bonds to metal.

Different tool providers have introduced different workarounds to adapt existing representation formats to accommodate concepts of organometallic and coordination chemistry but these are not consistently implemented. The concept of a zero-order bond [79] potentially addresses most deficiencies but this has not been universally adopted. The quest for universal conventions for the representation and identification of molecular structures containing metals is an ongoing one.

An illustration of the challenges can be seen by looking at bonding between metal centres. There are structures in the CSD where a metal-metal bond has been characterised as being a quadruple bond. There are even a handful of structures where researchers have characterised the bond as being a quintuple one, the earliest being from 2005 wherein the theoretical possibility of sextuple bonds is also hypothesised [80]. Further, some metal-metal bonds may be characterised as having partial bond orders, e.g. 2.5 [81]. The ability to handle partial bond orders and multiplicities beyond the triple bond is not universally or consistently accommodated by current representation formats.

[f] S. H. Hong, M. W. Day, and R. H. Grubbs (2004) "CCDC 223170: Experimental Crystal Structure Determination". Cambridge Crystallographic Data Centre. https://doi.org/10.5517/CC7H71Z.

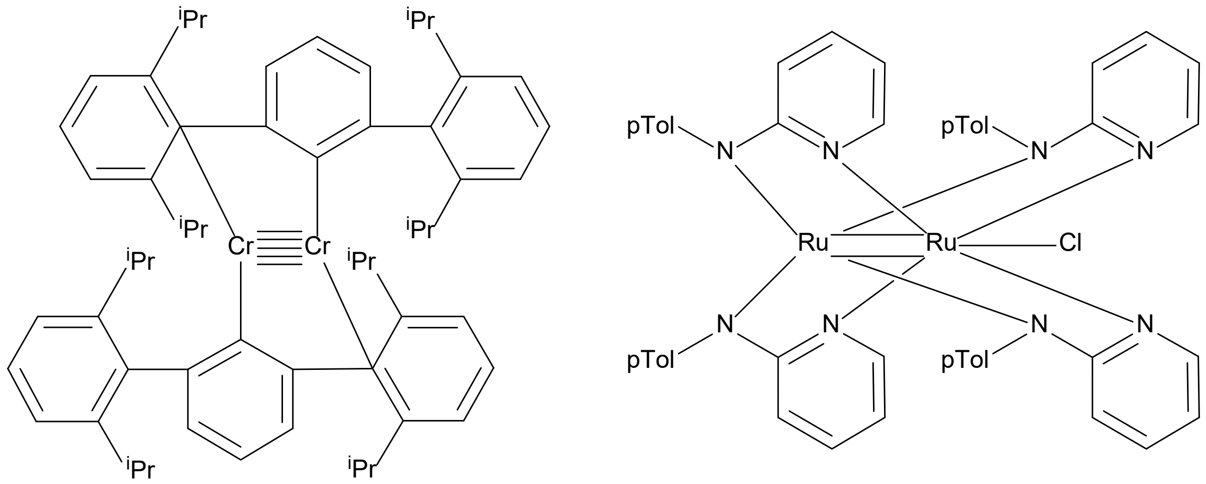


Figure 11: Left: CSD Entry PAZBOI[g] which is characterised as having a quintuple Cr-Cr metal bond. Right: CSD Entry DAVPOJ[h] which is characterised as having a Ru-Ru bond of order 2.5.

2.4.2 Polymeric Structures

Representation of polymeric structures requires the identification of a monomeric unit from which the entire polymer can be generated. This can be fraught with challenges when it comes to canonical representation, particularly if there is more than one constitutional repeating unit that could be selected as illustrated in Figure 12. These challenges have been explored as part of work undertaken to encode simple polymers using InChI [82]. BigSMILES has been proposed as a way of providing a linear representation of polymers that takes into account their stochastic nature and can be used for indexing structures in polymer databases [83].

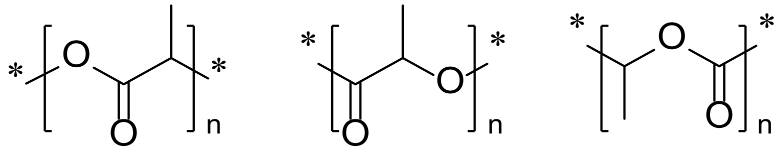


Figure 12: Different representations of polymeric unit for poly(lactic acid).

These challenges are magnified when there may be more than one dimension to the polymerisation, as is often found in metal-containing polymers. Figure 13 shows the monomer unit initially generated for a polymer in the CSD and the unit that resulted from review by an expert editor. Consideration in making the final choice was given to human readability (keeping the metal cluster intact) and representing enough of the structure to enable reliable structure-based searching. Careful consideration of the representation of polymeric structures is important to avoid pitfalls that can be encountered when searching for these [84].

[g] T. Nguyen, A. D. Sutton, M. Brynda, J. C. Fettinger, G. J. Long, and P. P. Power (2006) "CCDC 276888: Experimental Crystal Structure Determination". Cambridge Crystallographic Data Centre. https://doi.org/10.5517/CC993WK.

[h] M. D. Roy *et al.* (2022) "CCDC 2098290: Experimental Crystal Structure Determination". Cambridge Crystallographic Data Centre. https://doi.org/10.5517/CCDC.CSD.CC28FFSY.

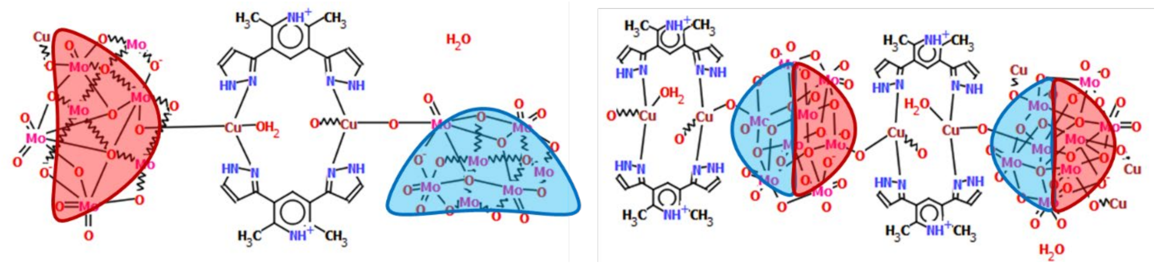


Figure 13: Two different representations of a metal-containing polymer (CSD Entry TISRUL[i]). The initial representation generated automatically (left) splits the metal cluster in two creating a more complex polymeric repeat unit. The curated representation (right) keeps the metal cluster intact making the structure clearer.

2.4.3 Interlocked Molecules

Mechanically interlocked molecular systems are those containing separate molecules that are not covalently bound but are entangled such that the individual molecules cannot be separated without breaking a covalent bond. An example of such a system is shown in Figure 14.

This is a class of compounds where the inter-relation of the separate molecules is very difficult to capture in a 2D representation and are difficult to identify using structure-based search methods. To aid reliable retrieval, these currently need to be manually classified. Procedures that would allow this class of structures to be systematically classified have been proposed but not widely adopted [77].

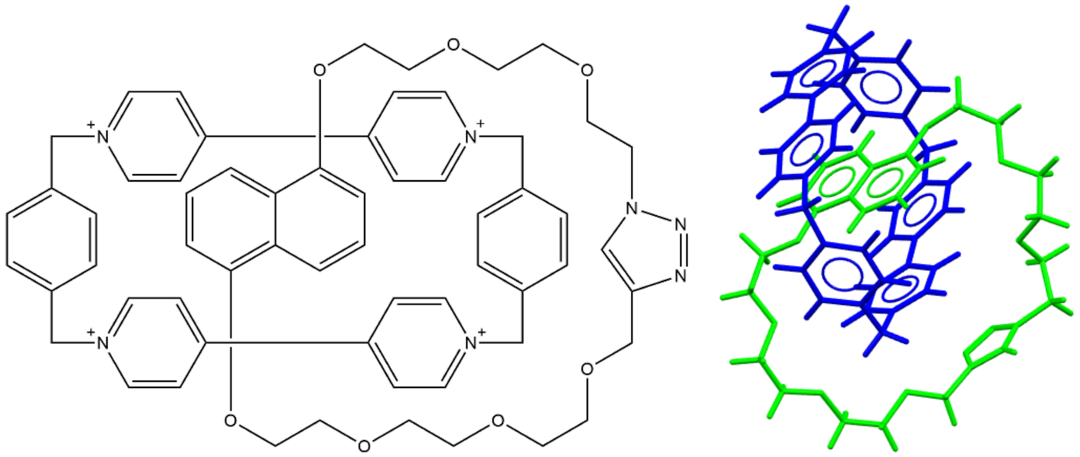


Figure 14: Example of a catenane, involving two mechanically interlocked molecules (CSD Entry ADOMUC[j]).

2.4.4 Stereochemistry

Many 2D representation formats can represent common atom and bond stereochemistry features found in organic molecules. Some extend to representation of stereochemistry at metal centres, but this is not universally supported. Other aspects of stereochemistry that can be challenging to represent include atropoisomerism; for example, axial chirality where

[i] Z. Shi *et al.* (2019) "CCDC 1817481: Experimental Crystal Structure Determination". Cambridge Crystallographic Data Centre. https://doi.org/10.5517/CCDC.CSD.CC1Z07FN.

[j] O. S. Miljanic, W. R. Dichtel, S. I. Khan, S. Mortezaei, J. R. Heath, and J. F. Stoddart (2007) "CCDC 655177: Experimental Crystal Structure Determination". Cambridge Crystallographic Data Centre. https://doi.org/10.5517/CCPZRR4.

rotation around a single bond is hindered and imposes a spatial arrangement that cannot be superimposed on its mirror image (Figure 15).

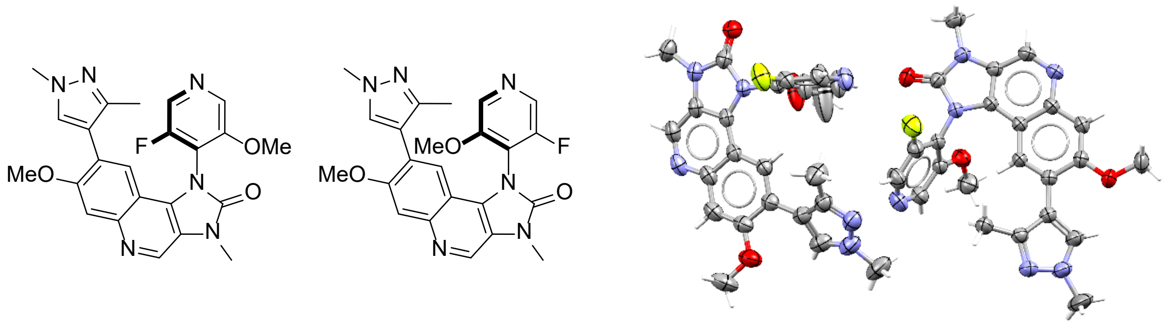


Figure 15: Example of a pair of atropisomers (CSD Entry CAQSUM[k]). The rotation of the pyridine ring is hindered affording two different configurations of the fluoro and methoxy substituents.

A feature of the more recent V3000 MOLfile is that a mixture of stereoisomers can be represented for one structure, accommodating situations where you have a sample that may contain more than one stereoisomer or where the precise stereo configuration is not known. Other representation formats may not readily accommodate this.

If you have a 3D model of a structure the stereochemistry is inherently defined, but even then caution is needed to be sure that the model provided reflects the stereoisomer studied. In crystallography, additional steps are required to determine the specific configuration of the structure, and these are not always undertaken.

When it comes to matching structures across resources where stereochemistry may be uncertain, ill-defined or a mixture, consideration needs to be given to how precisely one should match to a specific stereoisomer. Sometimes the precise configuration will matter, sometimes more fuzzy matching will be acceptable.

2.4.5 Tautomerism

Tautomers are variations of a compound that differ in the position of protons (or hydrogen atoms) and electrons. Warfarin for example can exist in solution in 40 distinct tautomeric forms [85].

Sometimes distinguishing the precise tautomer matters – sometimes it may be preferable to treat all tautomers as equivalent. InChI for example aims to be tautomer invariant. However, its current algorithm only recognises a small subset of a much larger set of possible tautomeric interconversions [86].

In the solid state, it is generally desirable to represent the specific tautomer observed as different tautomers can lead to different packing configurations [87].

[k] C. Saal, B. Axel, M. Krier, and T. Fuchß (2021) "CCDC 2101962: Experimental Crystal Structure Determination". Cambridge Crystallographic Data Centre. https://doi.org/10.5517/CCDC.CSD.CC28K87C.

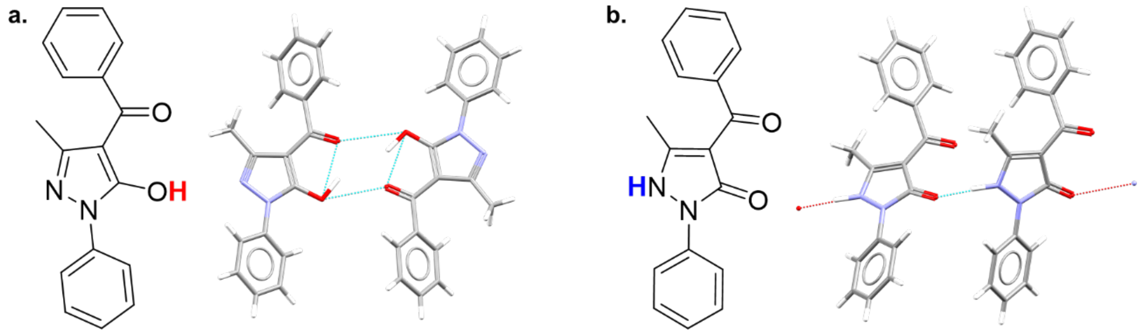
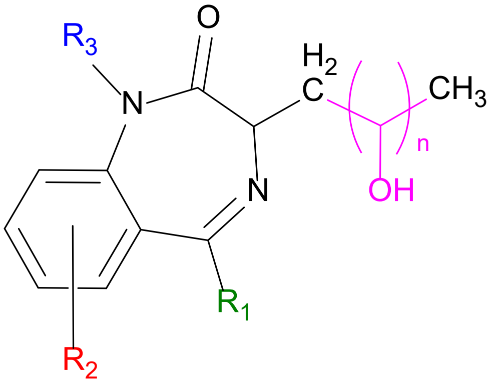


Figure 16: Example of tautomerism and effect on molecule packing. CSD Entry YUYDOL[l] (a) forms a less stable hydrogen bond network compared to CSD Entry DEBFAR[m] (b) even though (a) is the more stable tautomer.

2.4.6 Variability

Sometimes it is desirable to have a single representation that defines a class of compounds that have a common scaffold and vary only in the functional groups that are attached at a certain point. Such variability can be captured in a Markush or generic structure. Markush structures are commonly found in patents as an attempt extend the boundaries and dilute the specificity of a disclosure being made. A Markush structure can potentially represent an infinite number of specific structures.



R_1 = any of propyl / $CH_3CH(Br)CH_3$ / OMe attached at that specific point (substituent variation)

R_2 = Br attached to any unsubstituted atom in the fused Phenyl ring (position variation)

n = 2-5 repeat units of a C-OH fragment (frequency variation)

R_3 = any alkyl group containing 1-6 carbons attached to the N (homology variation - potentially infinite)

Figure 17: An example of a Markush structure which represents multiple different chemical structures using one 2D diagram and a range of variation types.

Solutions for the storage and retrieval of generic structures from patents have been long established [88] but it is not just in patents where these can be used. Markush structures have proved useful as a compact means to efficiently store and process the structures represented by combinatorial libraries used in high-throughput screening [89]. Aspects of variability exemplified by Markush structures can be accommodated by some representation formats but not all.

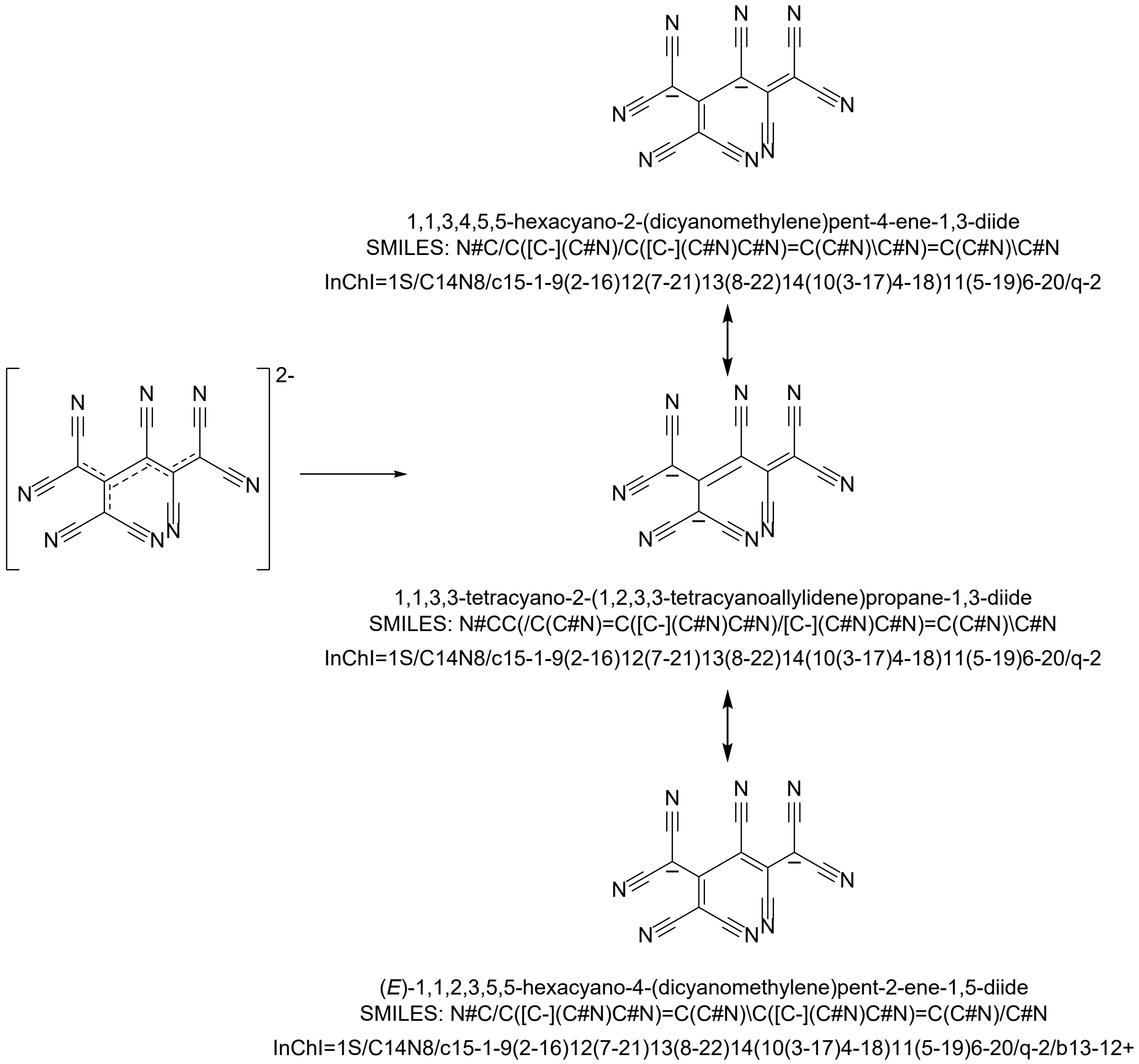
2.4.7 Delocalised Bonding

Delocalised bonds are commonly used to represent a compound that may exist in more than one resonant form, where electrons are mobile across conjugated bonds. The example in

[l] Y. Akama, M. Shiro, T. Ueda, M. Kajitani (1995) "CSD Entry YUYDOL". Cambridge Crystallographic Data Centre. https://identifiers.org/csd: YUYDOL.

[m] M.J.O'Connell, C.G.Ramsay, P.J.Steel (1985) "CSD Entry YUYDOL". Cambridge Crystallographic Data Centre. https://identifiers.org/csd: YUYDOL.

Figure 18 reflects use of delocalised bonds to represent a system that can exist in at least three different resonant forms. Each resonant form has a different IUPAC name and will result in different canonical SMILES. All share the same InChI string although for one of these, bond stereochemistry was identified. Reverse engineering the InChI reveals a different resonant form (Figure 19), one that might not be immediately considered from the initial representation of the delocalised system. This one example demonstrates many of the challenges of reliable structure representation that can be encountered with solutions available today.

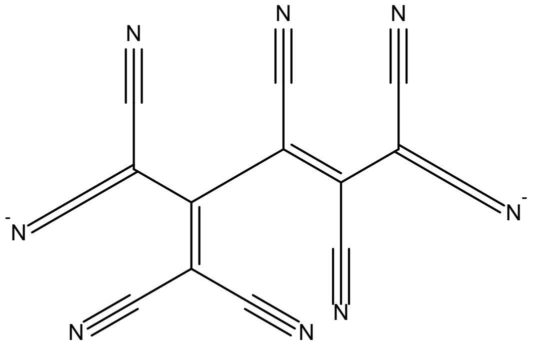


1,1,3,4,5,5-hexacyano-2-(dicyanomethylene)pent-4-ene-1,3-diide
SMILES: N#C/C([C-](C#N)/C([C-](C#N)C#N)=C(C#N)\C#N)=C(C#N)\C#N
InChI=1S/C14N8/c15-1-9(2-16)12(7-21)13(8-22)14(10(3-17)4-18)11(5-19)6-20/q-2

1,1,3,3-tetracyano-2-(1,2,3,3-tetracyanoallylidene)propane-1,3-diide
SMILES: N#CC(/C(C#N)=C([C-](C#N)C#N)/[C-](C#N)C#N)=C(C#N)\C#N
InChI=1S/C14N8/c15-1-9(2-16)12(7-21)13(8-22)14(10(3-17)4-18)11(5-19)6-20/q-2

(*E*)-1,1,2,3,5,5-hexacyano-4-(dicyanomethylene)pent-2-ene-1,5-diide
SMILES: N#C/C([C-](C#N)C#N)=C(C#N)\C([C-](C#N)C#N)=C(C#N)/C#N
InChI=1S/C14N8/c15-1-9(2-16)12(7-21)13(8-22)14(10(3-17)4-18)11(5-19)6-20/q-2/b13-12+

Figure 18: Different resonance forms of a conjugated dianionic molecule and associated IUPAC names, SMILES and InChI representations. Generated using ChemDraw 20.0.0.41.



(*E*)-(2,3,4,6-tetracyano-5-(dicyanomethylene)hepta-1,3,6-triene-1,7-diylidene)diamide
SMILES: N#CC(/C(C(C#N)=C=[N-])=C(C#N)/C#N)=C(C#N)\C(C#N)=C=[N-]
InChI=1S/C14N8/c15-1-9(2-16)12(7-21)13(8-22)14(10(3-17)4-18)11(5-19)6-20/q-2/b13-12+

Figure 19: The resonant form generated from the InChI of the conjugated dianionic molecule used in Figure 18. This indicates the representation that results from the normalisation and canonicalization undertaken by InChI algorithms. Generated using ChemDraw 20.0.0.41.

2.5 Broader Considerations of Structure Representation

2.5.1 Provenance of Structures – and their Representations

Consideration of 3D structures in Section 2.4.7 noted the importance of capturing and reporting the provenance of structures that may be available through a Physical Sciences Data Infrastructure and prompts the following recommendations:

Be sure to clearly distinguish between experimentally determined and computationally calculated structures.

For experimentally determined structures, be sure to capture relevant experimental details using standard file formats where available.

For computationally calculated structures, capture details of the methods and software packages used along with input and output files that could enable reproducibility and validation.

Work with existing communities to establish standards for the reliable reporting of computational methods.

Some of these recommendations should be considered in conjunction with the representation of 2D structures as well. We noted in Section 2.4.3 that different software packages may use varying conventions when generating representations such as SMILES. It would thus be good practice to capture details of software packages and toolkits used to generate a particular representation of a structure.

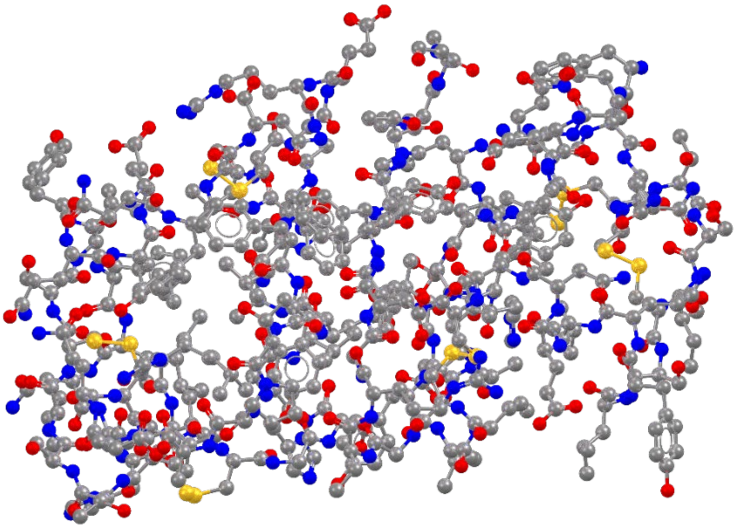
A Physical Sciences Data Infrastructure should embrace principles of software citation as they pertain to the generation, calculation, representation and analysis of structural data at all stages of the research life cycle.

Principles for software citation have been established [90] and guidance for implementation of these has been produced [91–95]. Attention is now turning to making machine learning models available in accordance with FAIR principles [96]. Proper treatment of AI and Machine Learning models within a future infrastructure will be important for the understanding and reproducibility of data-driven research. Use of ontologies for describing the chemical information entities used in the application of cheminformatics (structures, descriptors, algorithms and software libraries) [97] might also be considered.

2.5.2 Biological Structure

For Biological Macromolecular structures there are analogous representations to those found in the chemical domain but perhaps not the diversity. Proteins and nucleic acids can be represented as linear strings of codes for amino acids and bases [98]. 3D structures at atomic resolution can be represented in mmCIF files [99] or the legacy PDB format [100].

Pertinent to chemistry is the Hierarchical Editing Language for Macromolecules (HELM) [101], a notation which aims to bridge the gap between small molecules and sequences to provide a standard representation of complex biomolecules. HELM originated in Pfizer and was further developed through the Pistoia Alliance. A partnership between the Pistoia Alliance and IUPAC has now been established to address the future sustainability and support for HELM as a standard [102].



PEPTIDE1{F.V.N.Q.H.L.C.G.S.H.L.V.E.A.L.Y.L.V.C.G.E.R.G.F.F.Y.T.P.K.T}|PEPTIDE2{G.I.V.E.Q.C.C.T.S.I.C.S.L.Y.Q.L.E.N.Y.C.N}$PEPTIDE1,PEPTIDE2,7:R3-7:R3|PEPTIDE2,PEPTIDE2,6:R3-11:R3|PEPTIDE1,PEPTIDE2,19:R3-20:R3$$$

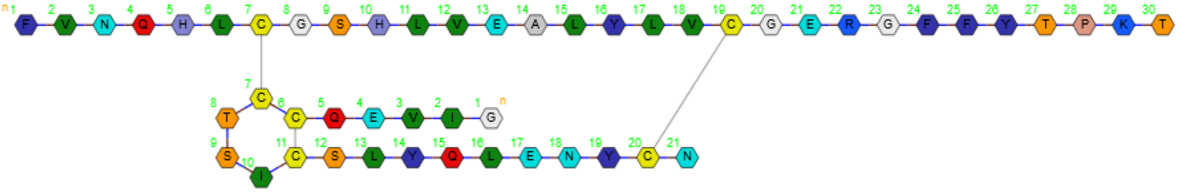

n 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T P K T
7 6 5 4 3 2 1 n
C C Q E V I G
8 9 10 11 12 13 14 15 16 17 18 19 20 21
T S I C S L Y Q L E N Y C N

Figure 20: 3D model of Insulin from PDB Entry 1TRZ[n] *and its HELM notation represented as a string and graphically*[o]*.*

[n] E. Ciszak and G. D. Smith (1994) "PDB Entry 1TRZ". https://doi.org/10.2210/pdb1trz/pdb.

[o] HELM string obtained from PubChem [103]; graphical representation generated using the HELM Web Editor [104].

3D structural data from chemistry can be used to help validate, analyse and optimise chemical components of 3D biological structures that have been modelled computationally or experimentally [105–107]. Knowledge derived from chemical structures can be used to visualise, analyse and optimise interactions between drug molecules and protein binding sites [108–110]. Being able to connect across chemical and biological resources based on structure can yield insights that help inform our understanding of biological function [111].

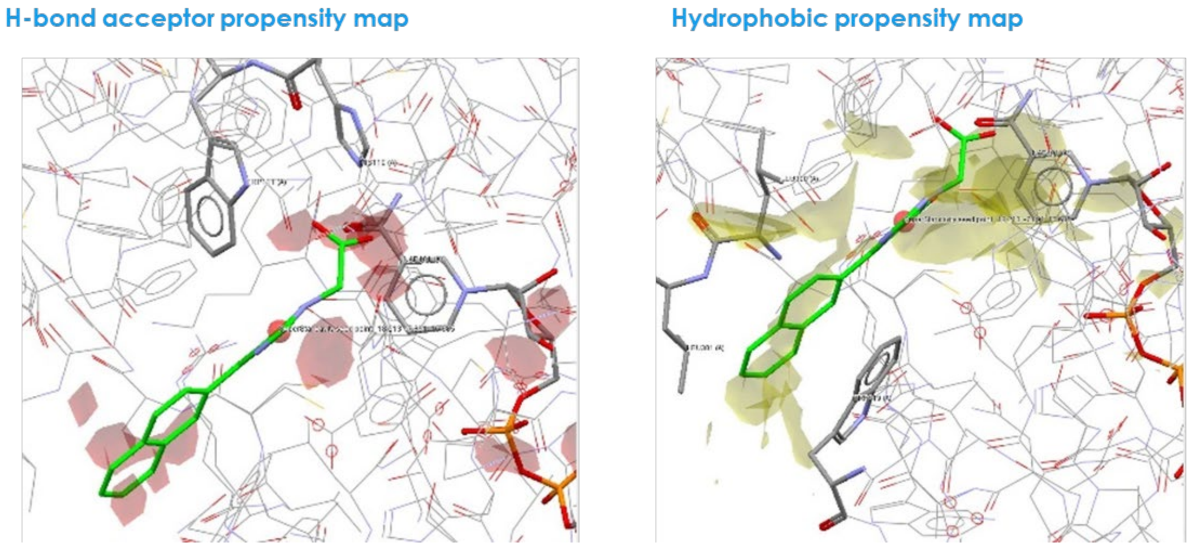


Figure 21: Contour surfaces derived from CSD data indicating the preferred positions of H-bond acceptors (left, red surface) and hydrophobic groups (right, green surface) in the active site of a PDB Entry 4IGS[p]*. Data derived from small molecule chemical structures is being used to probe the interaction of a potential inhibitor (green molecule) with the binding site of a biological macromolecule. Based on work reported by Saito et al. [112].*

Consideration should be given to the degree to which a Physical Sciences Data Infrastructure needs to handle representations of biological molecules. In any case it should enable links out to relevant biological resources based on structure. Partnership with life science organisations such as EBI is recommended as a way to establish solutions that satisfy requirements across domains.

As a side note to this section, we mention the recent headline-grabbing news related to DeepMind and their AI system AlphaFold that can predict the 3D structure from amino acid sequence alone [113]. This and comparable successes from RosTTAFold [114] would not have been possible without the body of curated experimental data found in the Protein Data Bank. It highlights the following important point:

If the promise of modern data science techniques is to be realised, the availability of compilations of quality experimental data and reliable structure representation is crucial.

2.5.3 Changing Habits – Scoping Opportunities

Another lesson we can take from structural biology relates to the time and effort required to change established researcher habits.

The mmCIF specification was first published in 1996 [115] at a time when the common exchange format for protein structures was the PDB format. As the complexity of structures

[p] A. Cousido-Siah *et al.* (2014) "PDB Entry 4IGS". https://doi.org/10.2210/pdb4igs/pdb

being determined increased, the boundaries of the PDB format were being reached, creating a need to shift from this to mmCIF. It took until 2019, a period of over 20 years, before the wwPDB felt they could mandate deposition of structures in mmCIF format and stop accepting these in PDB format [116]. One of the key reasons for inertia were the workflows and tools that had been established based around the legacy PDB format that were underpinning the discipline of structural bioinformatics. A key enabler of the transition to mmCIF was provision of workshops and resources leading up to the change in policy that helped ensure the disruption to the wider community would be minimal.

Lessons to be learnt from this include the following:

- Recognise that change takes time and that a future infrastructure may have to support legacy representation formats for longer than might be desirable.
- Be prepared to invest in tools and education that will enable communities to embrace change without fear of disruption.
- Insofar as possible, separate services, tools and workflows from underlying formats to enable these to evolve independently.

It is currently 30 years since the CIF ecosystem was first established but there remain opportunities for this to further evolve. CIF is based on STAR [117] and in 2012, the group from which STAR originated published extensions to syntax [118], a new dictionary definition language [119] and a language that allows relationships between data items to be expressed [120]. The overall aim of these changes was to enable a richer representation of scientific data and push towards the concept of a self-validating data document. Despite their potential, the adoption of these new features has been limited. One reason for this is that established workflows are working well with original versions of CIF and there is no great motivation to disrupt what isn't broken. This, however, misses out on the opportunity for doing better.

If future infrastructure is to evolve beyond limitations of technology and standards today, then there needs to be resources set aside for ongoing investment and experimentation with new paradigms and approaches.

2.5.4 Visualisation and Depiction

2.5.4.1 3D Visualisation

A further consideration from structural biology relates to visualisation of 3D structures where cartoon representations are often used to provide an indication of structural features that a ball-and-stick model would otherwise not expose. In the physical sciences the structures studied may not be as complex as biological macromolecules but features such as polyhedra, surfaces and spheres can be useful in communicating features of a structure – both what is there and what is absent, i.e. voids.

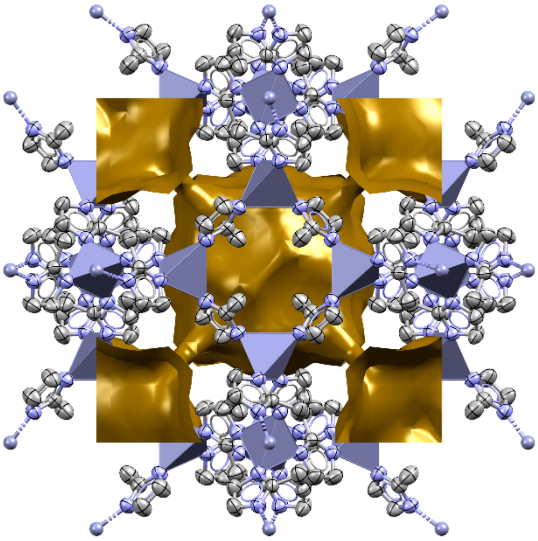


Figure 22: Example of the 3D structure of a metal-organic framework (ZIF-8) highlighting metal coordination environment (blue polyhedra) and accessible void space (gold surface). CSD Entry FAWCEN03[q]*.*

Infrastructure should look to provide 3D visualisation aids that readily allow researchers to communicate and observe key features of a structure.

A starting point for web-based 3D visualisation is JSMol which over time has incorporated a rich set of visualisation features. JSMol is written in Java and converted to JavaScript [121]. Visualisers that use more recent technology such as WebGL include NGL Viewer [122] and Mol* Viewer [123]. Originating in the structural biology community these may not be readily suited for 3D visualisation of chemical structures and materials. However, they are Open Source so could be built on in this regard.

2.5.4.2 2D Depiction

When it comes to the 2D depiction of structures, we have one simple recommendation:

If infrastructure is to render 2D representations of structures, then guidelines issued by IUPAC [124]should be adhered to as much as possible to reduce the chance of diagrams being misleading to users of the infrastructure.

[q] W. Morris, C. J. Stevens, R. E. Taylor, C. Dybowski, O. M. Yaghi, and M. A. Garcia-Garibay (2012) "CCDC 864312: Experimental Crystal Structure Determination". Cambridge Crystallographic Data Centre. https://doi.org/10.5517/CCY0D1C.

2.6 References

1. N. M. O’Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison (2011) "Open Babel: An open chemical toolbox", *Journal of Cheminformatics*, 3(1), https://doi.org/10.1186/1758-2946-3-33.

2. "Supported File Formats and Options — Open Babel v2.3.0 documentation", http://openbabel.org/docs/2.3.0/FileFormats/Overview.html (accessed 14 April 2022).

3. "Generate Chemical Names from Structure | ACD/Name", https://www.acdlabs.com/products/draw_nom/nom/name/index.php (accessed 14 April 2022).

4. "Chemical Name and Structure Conversion | ChemAxon", https://chemaxon.com/products/chemical-name-conversion (accessed 14 April 2022).

5. OpenEye Scientific Software "Lexichem Toolkit | Chemical Structures & Chemical Names Interconversion | Cheminformatics", https://www.eyesopen.com/lexichem-tk (accessed 14 April 2022).

6. J. Brecher (1999) "Name=Struct: A Practical Approach to the Sorry State of Real-Life Chemical Nomenclature", *J. Chem. Inf. Comput. Sci.*, 39(6), https://doi.org/10.1021/ci990062c.

7. D. M. Lowe, P. T. Corbett, P. Murray-Rust, and R. C. Glen (2011) "Chemical Name to Structure: OPSIN, an Open Source Solution", *J. Chem. Inf. Model.*, 51(3), https://doi.org/10.1021/ci100384d.

8. L. Krasnov, I. Khokhlov, M. V. Fedorov, and S. Sosnin (2021) "Transformer-based artificial neural networks for the conversion between chemical notations", *Sci Rep*, 11(1), https://doi.org/10.1038/s41598-021-94082-y.

9. G. J. Leigh (2011) "*Principles of chemical nomenclature: a guide to IUPAC recommendations*". Cambridge: Royal Society of Chemistry.

10. K. Rajan, H. O. Brinkhaus, A. Zielesny, and C. Steinbeck (2020) "A review of optical chemical structure recognition tools", *Journal of Cheminformatics*, 12(1), https://doi.org/10.1186/s13321-020-00465-0.

11. J. Staker, K. Marshall, R. Abel, and C. M. McQuaw (2019) "Molecular Structure Extraction from Documents Using Deep Learning", *J. Chem. Inf. Model.*, 59(3), https://doi.org/10.1021/acs.jcim.8b00669.

12. K. Rajan, A. Zielesny, and C. Steinbeck (2020) "DECIMER: towards deep learning for chemical image recognition", *Journal of Cheminformatics*, 12(1), https://doi.org/10.1186/s13321-020-00469-w.

13. "CAS REGISTRY and CAS Registry Number FAQs", *CAS*, https://www.cas.org/support/documentation/chemical-substances/faqs (accessed 14 April 2022).

14. T. Peryea *et al.* (2021) "Global Substance Registration System: consistent scientific descriptions for substances related to health", *Nucleic Acids Research*, 49(D1), https://doi.org/10.1093/nar/gkaa962.

15. "EC-No", https://www.chemeurope.com/en/encyclopedia/EC-No.html (accessed 14 April 2022).

16. S. M. Wimalaratne *et al.* (2018) "Uniform resolution of compact identifiers for biomedical data", *Sci Data*, 5(1), https://doi.org/10.1038/sdata.2018.29.

17. W. J. Wiswesser (1952) "The Wiswesser Line Formula Notation", *Chem. Eng. News Archive*, 30(34), https://doi.org/10.1021/cen-v030n034.p3523.

18. D. Weininger (1988) "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules", *J. Chem. Inf. Model.*, 28(1), https://doi.org/10.1021/ci00057a005.

19. "OpenSMILES specification", http://opensmiles.org/opensmiles.html (accessed 14 April 2022).

20. "How do I write thee? Let me count the ways – NextMove Software", https://nextmovesoftware.com/blog/2014/07/15/how-do-i-write-thee-let-me-count-the-ways/ (accessed 14 April 2022).

21. "IUPAC SMILES+ Specification", *IUPAC | International Union of Pure and Applied Chemistry*, https://iupac.org/project/2019-002-2-024 (accessed 14 April 2022).

22. N. O'Boyle and A. Dalke (2018) "DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures", https://doi.org/10.26434/chemrxiv.7097960.v1.

23. M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik (2020) "Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation", *Machine Learning: Science and Technology*, 1(4), https://doi.org/10.1088/2632-2153/aba947.

24. S. Heller, A. McNaught, S. Stein, D. Tchekhovskoi, and I. Pletnev (2013) "InChI - the worldwide chemical structure identifier standard", *Journal of Cheminformatics*, 5(1), https://doi.org/10.1186/1758-2946-5-7.

25. C. Southan (2013) "InChI in the wild: an assessment of InChIKey searching in Google", *Journal of Cheminformatics*, 5(1), https://doi.org/10.1186/1758-2946-5-10.

26. "» An InChIkey Collision is Discovered and NOT Based on Stereochemistry ChemConnector Blog", http://www.chemconnector.com/2011/09/01/an-inchikey-collision-is-discovered-and-not-based-on-stereochemistry/ (accessed 14 April 2022).

27. W. A. Warr (2015) "Many InChIs and quite some feat", *J Comput Aided Mol Des*, 29(8), https://doi.org/10.1007/s10822-015-9854-3.

28. "2.2.1: Introduction to Connection Tables", *Chemistry LibreTexts*, https://chem.libretexts.org/Courses/University_of_Arkansas_Little_Rock/ChemInformatics_(2017)%3A_Chem_4399_5399/2.2%3A_Chemical_Representations_on_Computer%3A_Part_II/2.2.1%3A_Introduction_to_Connection_Tables (accessed 14 April 2022).

29. A. Dalby *et al.* (1992) "Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited", *J. Chem. Inf. Comput. Sci.*, 32(3), https://doi.org/10.1021/ci00007a012.

30. Dassault Systèmes (2020) "CTfile Formats", *Dassault Systèmes*, https://discover.3ds.com/ctfile-documentation-request-form (accessed 12 January 2022).

31. "CDX Format Specification: General", https://www.cambridgesoft.com/services/documentation/sdk/chemdraw/cdx/General.htm (accessed 12 January 2022).

32. M. Sihag (2021) "Mol2 file format explained for beginners (Cheminformatics Part 2)", *ChemicBook*, https://chemicbook.com/2021/02/20/mol2-file-format-explained-for-beginners-part-2.html (accessed 14 April 2022).

33. S. R. Hall, F. H. Allen, and I. D. Brown (1991) "The crystallographic information file (CIF): a new standard archive file for crystallography", *Acta Crystallographica Section A Foundations of Crystallography*, 47(6), https://doi.org/10.1107/S010876739101067X.

34. S. R. Hall and B. McMahon (2016) "The Implementation and Evolution of STAR/CIF Ontologies: Interoperability and Preservation of Structured Data", *Data Science Journal*, 15(3), https://doi.org/10.5334/dsj-2016-003.

35. I. J. Bruno, G. P. Shields, and R. Taylor (2011) "Deducing chemical structure from crystallographically determined atomic coordinates", *Acta Crystallographica Section B Structural Science*, 67(4), https://doi.org/10.1107/S0108768111024608.

36. "The Seventh CSP Blind Test - 2020 to 2022 - The Cambridge Crystallographic Data Centre (CCDC)", https://www.ccdc.cam.ac.uk/Community/initiatives/cspblindtests/csp-blind-test-7/ (accessed 14 April 2022).

37. W. Phadungsukanan, M. Kraft, J. A. Townsend, and P. Murray-Rust (2012) "The semantics of Chemical Markup Language (CML) for computational chemistry : CompChem", *Journal of Cheminformatics*, 4(1), https://doi.org/10.1186/1758-2946-4-15.

38. W. A. de Jong, A. M. Walker, and M. D. Hanwell (2013) "From data to analysis: linking NWChem and Avogadro with the syntax and semantics of Chemical Markup Language", *Journal of Cheminformatics*, 5(1), https://doi.org/10.1186/1758-2946-5-25.

39. Chemical Semantics, Inc (2015) "GNVC: Gainesville Core Ontology - standard for publishing results of computational chemistry.", http://ontologies.makolab.com/gc/gc.html (accessed 21 April 2022).

40. E. Rossi *et al.* (2014) "Code interoperability and standard data formats in quantum chemistry and quantum dynamics: The Q5/D5Cost data model", *Journal of Computational Chemistry*, 35(8), https://doi.org/10.1002/jcc.23492.

41. N. Krdzavac *et al.* (2019) "An Ontology and Semantic Web Service for Quantum Chemistry Calculations", *J. Chem. Inf. Model.*, 59(7), https://doi.org/10.1021/acs.jcim.9b00227.

42. T. Lu and Q. Chen (2021) "mwfn: A Strict, Concise and Extensible Format for Electronic Wavefunction Storage and Exchange", https://doi.org/10.26434/chemrxiv-2021-lt04f-v5.

43. "TREXIO: a standard format for storing wave functions", https://trex-coe.github.io/trexio/index.html (accessed 21 April 2022).

44. G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari, and B. Kozinsky (2016) "AiiDA: automated interactive infrastructure and database for computational science", *Computational Materials Science*, 111, https://doi.org/10.1016/j.commatsci.2015.09.013.

45. A. Merkys, N. Mounet, A. Cepellotti, N. Marzari, S. Gražulis, and G. Pizzi (2017) "A posteriori metadata from automated provenance tracking: integration of AiiDA and TCOD", *Journal of Cheminformatics*, 9(1), https://doi.org/10.1186/s13321-017-0242-y.

46. G. Grethe, G. Blanke, H. Kraut, and J. M. Goodman (2018) "International chemical identifier for reactions (RInChI)", *Journal of Cheminformatics*, 10(1), https://doi.org/10.1186/s13321-018-0277-8.

47. A. M. Clark, L. R. McEwen, P. Gedeck, and B. A. Bunin (2019) "Capturing mixture composition: an open machine-readable format for representing mixed substances", *Journal of Cheminformatics*, 11(1), https://doi.org/10.1186/s13321-019-0357-4.

48. J. Barnard (1993) "The Standard Molecular Data (SMD) Format", *Chemical Structures 2*, Available: http://link.springer.com/chapter/10.1007/978-3-642-78027-1_17.

49. F. H. Allen, J. M. Barnard, A. P. F. Cook, and S. R. Hall (1995) "The Molecular Information File (MIF): Core Specifications of a New Standard Format for Chemical Data", *J. Chem. Inf. Comput. Sci.*, 35(3), https://doi.org/10.1021/ci00025a009.

50. E. Zass (1996) "From handbooks to databases on the net: New solutions and old problems in information retrieval for chemists", *Journal of chemical information and computer sciences*, Available: http://pubs.acs.org/doi/abs/10.1021/ci950249d.

51. J. Noordik (2004) "The CAOS/CAMM Center: The Dutch national academic facility for computer-assisted organic synthesis and modeling", *Cheminformatics*, Available: http://iospress.metapress.com/index/rd7e7efkrllad1tn.pdf.

52. P. Willett (2009) "Similarity methods in chemoinformatics", *Annual Review of Information Science and \ldots*, Available: http://onlinelibrary.wiley.com/doi/10.1002/aris.2009.1440430108/abstract.

53. P. Murray-Rust and H. S. Rzepa (1999) "Chemical Markup, XML, and the Worldwide Web. 1. Basic Principles", *J. Chem. Inf. Comput. Sci.*, 39(6), https://doi.org/10.1021/ci990052b.

54. P. Murray-Rust and H. S. Rzepa (2001) "Chemical Markup, XML and the World-Wide Web. 2. Information Objects and the CMLDOM", *J. Chem. Inf. Comput. Sci.*, 41(5), https://doi.org/10.1021/ci000404a.

55. G. V. Gkoutos, P. Murray-Rust, H. S. Rzepa, and M. Wright (2001) "Chemical Markup, XML, and the World-Wide Web. 3. Toward a Signed Semantic Chemical Web of Trust", *J. Chem. Inf. Comput. Sci.*, 41(5), https://doi.org/10.1021/ci000406v.

56. P. Murray-Rust and H. S. Rzepa (2003) "Chemical Markup, XML, and the World Wide Web. 4. CML Schema", *J. Chem. Inf. Comput. Sci.*, 43(3), https://doi.org/10.1021/ci0256541.

57. P. Murray-Rust, H. S. Rzepa, M. J. Williamson, and E. L. Willighagen (2004) "Chemical Markup, XML, and the World Wide Web. 5. Applications of Chemical Metadata in RSS Aggregators", *J. Chem. Inf. Comput. Sci.*, 44(2), https://doi.org/10.1021/ci034244p.

58. G. L. Holliday, P. Murray-Rust, and H. S. Rzepa (2006) "Chemical Markup, XML, and the World Wide Web. 6. CMLReact, an XML Vocabulary for Chemical Reactions", *J. Chem. Inf. Model.*, 46(1), https://doi.org/10.1021/ci0502698.

59. S. Kuhn *et al.* (2007) "Chemical Markup, XML, and the World Wide Web. 7. CMLSpect, an XML Vocabulary for Spectral Data", *J. Chem. Inf. Model.*, 47(6), https://doi.org/10.1021/ci600531a.

60. P. Murray-Rust and H. S. Rzepa (2011) "CML: Evolution and design", *Journal of Cheminformatics*, 3(1), https://doi.org/10.1186/1758-2946-3-44.

61. Pistoia Alliance (2021) "*UDM*". Accessed: 14 April 2022. Available: https://github.com/PistoiaAlliance/UDM.

62. J. Hastings *et al.* (2013) "The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013", *Nucleic Acids Research*, 41(D1), https://doi.org/10.1093/nar/gks1146.

63. H. J. Feldman, M. Dumontier, S. Ling, N. Haider, and C. W. V. Hogue (2005) "CO: A chemical ontology for identification of functional groups and semantic comparison of small molecules", *FEBS Letters*, 579(21), https://doi.org/10.1016/j.febslet.2005.07.039.

64. Y. Djoumbou Feunang *et al.* (2016) "ClassyFire: automated chemical classification with a comprehensive, computable taxonomy", *Journal of Cheminformatics*, 8(1), https://doi.org/10.1186/s13321-016-0174-y.

65. Institut de l'information scientifique et technique (Inist) - CNRS/UPS76 "Vocabulaire de chimie", https://doi.org/10.13143/LOTR.4642.

66. "Deconstruction of Crystalline Networks into Underlying Nets: Relevance for Terminology Guidelines and Crystallographic Databases | Crystal Growth & Design", 18(6), https://doi.org/10.1021/acs.cgd.8b00126.

67. M. O'Keeffe, M. A. Peskov, S. J. Ramsden, and O. M. Yaghi (2008) "The Reticular Chemistry Structure Resource (RCSR) Database of, and Symbols for, Crystal Nets", *Acc. Chem. Res.*, 41(12), https://doi.org/10.1021/ar800124u.

68. E. L. First and C. A. Floudas (2013) "MOFomics: Computational pore characterization of metal–organic frameworks", *Microporous and Mesoporous Materials*, 165, https://doi.org/10.1016/j.micromeso.2012.07.049.

69. V. A. Blatov, A. P. Shevchenko, and D. M. Proserpio (2014) "Applied Topological Analysis of Crystal Structures with the Program Package ToposPro", *Crystal Growth & Design*, 14(7), https://doi.org/10.1021/cg500498k.

70. "MOFplus Project Webpage", https://www.mofplus.org/content/show/generalnetinfo (accessed 14 April 2022).

71. B. J. Bucior *et al.* (2019) "Identification Schemes for Metal–Organic Frameworks to Enable Rapid Search and Cheminformatics Analysis", *Crystal Growth & Design*, https://doi.org/10.1021/acs.cgd.9b01050.

72. D. J. O'Hearn, A. Bajpai, and M. J. Zaworotko (2021) "The 'Chemistree' of Porous Coordination Networks: Taxonomic Classification of Porous Solids to Guide Crystal Engineering Studies", *Small*, 17(22), https://doi.org/10.1002/smll.202006351.

73. "(IUCr) CIF Definition save_TOPOLOGY", https://www.iucr.org/__data/iucr/cifdic_html/3/TOPOLOGY_CIF/CTOPOLOGY.html (accessed 14 April 2022).

74. K. E. Cordova and O. M. Yaghi "Reticular Chemistry Naming and Numbering Database", https://globalscience.berkeley.edu/database (accessed 27 April 2022).

75. Y. G. Chung *et al.* (2019) "Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019", *J. Chem. Eng. Data*, 64(12), https://doi.org/10.1021/acs.jced.9b00835.

76. P. Z. Moghadam *et al.* (2017) "Development of a Cambridge Structural Database Subset: A Collection of Metal–Organic Frameworks for Past, Present, and Future", *Chem. Mater.*, 29(7), https://doi.org/10.1021/acs.chemmater.7b00441.

77. J. Dobrowolski (2003) "Classification of Topological Isomers: Knots, Links, Rotaxanes, etc.", *Croatica Chemica Acta*, 76.

78. J. Bernstein, R. E. Davis, L. Shimoni, and N.-L. Chang (1995) "Patterns in Hydrogen Bonding: Functionality and Graph Set Analysis in Crystals", *Angewandte Chemie International Edition in English*, 34(15), https://doi.org/10.1002/anie.199515551.

79. A. M. Clark (2011) "Accurate Specification of Molecular Structures: The Case for Zero-Order Bonds and Explicit Hydrogen Counting", *J. Chem. Inf. Model.*, 51(12), https://doi.org/10.1021/ci200488k.

80. T. Nguyen, A. D. Sutton, M. Brynda, J. C. Fettinger, G. J. Long, and P. P. Power (2005) "Synthesis of a Stable Compound with Fivefold Bonding Between Two Chromium(I) Centers", *Science*, 310(5749), https://doi.org/10.1126/science.1116789.

81. M. D. Roy *et al.* (2022) "Electronic Structure of Ru26+ Complexes with Electron-Rich Anilinopyridinate Ligands", *Inorg. Chem.*, 61(8), https://doi.org/10.1021/acs.inorgchem.1c03346.

82. A. Yerin "InChI encoding of polymers current results and further tasks". Available: https://www.inchi-trust.org/wp/wp-content/uploads/2017/11/23.-InChI-Polymer-Yerin-201708.pdf.

83. T.-S. Lin *et al.* (2019) "BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules", *ACS Cent. Sci.*, 5(9), https://doi.org/10.1021/acscentsci.9b00476.

84. D. Seebach *et al.* (2009) "Polymer Backbone Conformation-A Challenging Task for Database Information Retrieval", *Angewandte Chemie International Edition*, 48(51), https://doi.org/10.1002/anie.200904422.

85. L. Guasch, M. L. Peach, and M. C. Nicklaus (2015) "Tautomerism of Warfarin: Combined Chemoinformatics, Quantum Chemical, and NMR Investigation", *The Journal of Organic Chemistry*, 80(20), https://doi.org/10.1021/acs.joc.5b01370.

86. D. K. Dhaked, W.-D. Ihlenfeldt, H. Patel, V. Delannée, and M. C. Nicklaus (2020) "Toward a Comprehensive Treatment of Tautomerism in Chemoinformatics Including in InChI V2", *J. Chem. Inf. Model.*, 60(3), https://doi.org/10.1021/acs.jcim.9b01080.

87. A. J. Cruz-Cabeza and C. R. Groom (2010) "Identification, classification and relative stability of tautomers in the cambridge structural database", *CrystEngComm*, 13(1), https://doi.org/10.1039/C0CE00123F.

88. M. F. Lynch and J. D. Holliday (1996) "The Sheffield Generic Structures Projecta Retrospective Review", *J. Chem. Inf. Comput. Sci.*, 36(5), https://doi.org/10.1021/ci950173l.

89. J. M. Barnard, G. M. Downs, A. von Scholley-Pfab, and R. D. Brown (2000) "Use of Markush structure analysis techniques for descriptor generation and clustering of large combinatorial libraries", *Journal of Molecular Graphics and Modelling*, 18(4), https://doi.org/10.1016/S1093-3263(00)80091-1.

90. A. M. Smith, D. S. Katz, and K. E. Niemeyer (2016) "Software citation principles", *PeerJ Comput. Sci.*, 2, https://doi.org/10.7717/peerj-cs.86.

91. D. S. Katz *et al.* (2019) "Software Citation Implementation Challenges", *arXiv:1905.08674 [cs]*, Accessed: 19 April 2022. Available: http://arxiv.org/abs/1905.08674.

92. N. P. Chue Hong *et al.* (2019) "Software Citation Checklist for Authors", Zenodo. https://doi.org/10.5281/zenodo.3479199.

93. N. P. Chue Hong *et al.* (2019) "Software Citation Checklist for Developers", Zenodo. https://doi.org/10.5281/zenodo.3482769.

94. Task Force on Best Practices for Software Registries *et al.* (2020) "Nine Best Practices for Research Software Registries and Repositories: A Concise Guide", *arXiv:2012.13117 [cs]*, Accessed: 19 April 2022. Available: http://arxiv.org/abs/2012.13117.

95. D. S. Katz *et al.* (2021) "Recognizing the value of software: a software citation guide". F1000Research. https://doi.org/10.12688/f1000research.26932.2.

96. Katz, Daniel S., Pollard, Tom, Psomopoulos, Fotis, Huerta, Eliu, Erdmann, Chris, and Blaiszik, Ben (2020) "FAIR principles for Machine Learning models", https://doi.org/10.5281/ZENODO.4271995.

97. J. Hastings, L. Chepelev, E. Willighagen, N. Adams, C. Steinbeck, and M. Dumontier (2011) "The Chemical Information Ontology: Provenance and Disambiguation for Chemical Data on the Biological Semantic Web", *PLOS ONE*, 6(10), https://doi.org/10.1371/journal.pone.0025513.

98. "FASTA", *Wikipedia*. Accessed: 19 April 2022. Available: https://en.wikipedia.org/w/index.php?title=FASTA&oldid=1074507962.

99. P. E. Bourne, H. M. Berman, B. McMahon, K. D. Watenpaugh, J. D. Westbrook, and P. M. D. Fitzgerald (1997) "Macromolecular crystallographic information file", in *Methods in Enzymology*, 277, Academic Press. https://doi.org/10.1016/S0076-6879(97)77032-0.

100. "Atomic Coordinate Entry Format Version 3.3", https://www.wwpdb.org/documentation/file-format-content/format33/v3.3.html (accessed 19 April 2022).

101. T. Zhang, H. Li, H. Xi, R. V. Stanton, and S. H. Rotstein (2012) "HELM: A Hierarchical Notation Language for Complex Biomolecule Structure Representation", *J. Chem. Inf. Model.*, 52(10), https://doi.org/10.1021/ci3001925.

102. "IUPAC Subcommittee on HELM", *IUPAC | International Union of Pure and Applied Chemistry*, https://iupac.org/who-we-are/committees/committee-details/ (accessed 19 April 2022).

103. PubChem "Insulin human", https://pubchem.ncbi.nlm.nih.gov/compound/118984375 (accessed 29 April 2022).

104. "HELM Web Editor", http://webeditor.openhelm.org/hwe/examples/App.htm (accessed 28 April 2022).

105. P. D. Adams *et al.* (2016) "Outcome of the First wwPDB/CCDC/D3R Ligand Validation Workshop", *Structure*, 24(4), https://doi.org/10.1016/j.str.2016.02.017.

106. O. S. Smart *et al.* (2018) "Validation of ligands in macromolecular structures determined by X-ray crystallography", *Acta Cryst D*, 74(3), https://doi.org/10.1107/S2059798318002541.

107. J. Liebeschuetz, J. Hennemann, T. Olsson, and C. R. Groom (2012) "The good, the bad and the twisted: a survey of ligand geometry in protein crystal structures", *J Comput Aided Mol Des*, 26(2), https://doi.org/10.1007/s10822-011-9538-6.

108. M. L. Verdonk, J. C. Cole, and R. Taylor (1999) "SuperStar: a knowledge-based approach for identifying interaction sites in proteins.", *Journal of molecular biology*, 289, https://doi.org/10.1006/jmbi.1999.2809.

109. D. R. Boer, J. Kroon, J. C. Cole, B. Smith, and M. L. Verdonk (2001) "Superstar: comparison of CSD and PDB-based interaction fields as a basis for the prediction of protein-ligand interactions11Edited by R. Huber", *Journal of Molecular Biology*, 312(1), https://doi.org/10.1006/jmbi.2001.4901.

110. J. W. M. Nissink and R. Taylor (2004) "Combined use of physicochemical data and small-molecule crystallographic contact propensities to predict interactions in protein binding sites", *Org. Biomol. Chem.*, 2(22), https://doi.org/10.1039/B405205F.

111. "BioChemGRAPH - an integrated knowledge graph to facilitate basic and translational research", https://gtr.ukri.org/projects?ref=BB%2FT019778%2F1 (accessed 19 April 2022).

112. R. Saito, M. Hoshi, A. Kato, C. Ishikawa, and T. Komatsu (2017) "Green fluorescent protein chromophore derivatives as a new class of aldose reductase inhibitors", *European Journal of Medicinal Chemistry*, 125, https://doi.org/10.1016/j.ejmech.2016.10.016.

113. J. Jumper *et al.* (2021) "Highly accurate protein structure prediction with AlphaFold", *Nature*, 596(7873), https://doi.org/10.1038/s41586-021-03819-2.

114. M. Baek *et al.* (2021) "Accurate prediction of protein structures and interactions using a three-track neural network", *Science*, 373(6557), https://doi.org/10.1126/science.abj8754.

115. "Dictionary Index mmcif_std.dic", http://mmcif.rcsb.org/dictionaries/mmcif_std.dic/Index/ (accessed 19 April 2022).

116. P. D. Adams *et al.* (2019) "Announcing mandatory submission of PDBx/mmCIF format files for crystallographic depositions to the Protein Data Bank (PDB)", *Acta Cryst D*, 75(4), https://doi.org/10.1107/S2059798319004522.

117. S. R. Hall (1991) "The STAR file: a new format for electronic data transfer and archiving", *Journal of Chemical Information and Modeling*, 31(2), https://doi.org/10.1021/ci00002a020.

118. N. Spadaccini and S. R. Hall (2012) "Extensions to the STAR File Syntax", *J. Chem. Inf. Model.*, 52(8), https://doi.org/10.1021/ci300074v.

119. N. Spadaccini and S. R. Hall (2012) "DDLm: A New Dictionary Definition Language", *Journal of Chemical Information and Modeling*, 52(8), https://doi.org/10.1021/ci300075z.

120. N. Spadaccini, I. R. Castleden, D. du Boulay, and S. R. Hall (2012) "dREL: A Relational Expression Language for Dictionary Methods", *Journal of Chemical Information and Modeling*, 52(8), https://doi.org/10.1021/ci300076w.

121. R. M. Hanson, J. Prilusky, Z. Renjian, T. Nakane, and J. L. Sussman (2013) "JSmol and the Next-Generation Web-Based Representation of 3D Molecular Structure as Applied to Proteopedia", *Israel Journal of Chemistry*, 53(3–4), https://doi.org/10.1002/ijch.201300024.

122. "NGL Viewer: a web application for molecular visualization | Nucleic Acids Research | Oxford Academic", https://academic.oup.com/nar/article/43/W1/W576/2467902 (accessed 19 April 2022).

123. D. Sehnal *et al.* (2021) "Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures", *Nucleic Acids Research*, 49(W1), https://doi.org/10.1093/nar/gkab314.

124. J. Brecher (2008) "Graphical representation standards for chemical structure diagrams (IUPAC Recommendations 2008)", *Pure and Applied Chemistry*, 80(2), https://doi.org/10.1351/pac200880020277.

3 Workflow and Infrastructure

Contents

3.1 Overview

This section identifies considerations for data management workflows and infrastructure that relate to structure. It outlines enablers that can help result in the richest possible representation of structure alongside data, and the services that should be provided to enable structure-based discovery of data and information.

3.2 Data or Metadata

An important consideration is whether structure representation should be considered to be “data” or “metadata”, i.e. is it the primary result of an experiment or is it a descriptive characteristic of a data set. Consider the following scenarios:

An experimental or computational study of 3D structure: Here, a model of the structure being studied will be central to the data ultimately reported. However, even when modelling a structure at an atomic level, the data may not inherently represent an unambiguous representation of the substance or material being studied. Data for an experimentally determined crystal structure for example may only be the positions of atoms in 3D space which is insufficient as a precise and unambiguous representation of structure.

A spectroscopic study of a substance: The result of a spectroscopic study will be a spectrum defined by lines and peaks. Features of this spectrum will indicate the presence and properties of specific fragments or functional groups within a molecule. It will not necessarily provide a complete representation of the structure studied.

A measured or calculated property of a structure: This will most likely be just a number – hopefully with uncertainties, units and details of the method used – but very divorced from the representation of the structure of the substance studied.

Given the above, we contend the following:

Representations of structure should primarily be considered metadata that is required as part of reporting the results of any experimental or computational study of physical substances and their properties.

3.3 Required Level of Representation

In Section 2 we outlined the strengths and the limitations of a range of different structure representations: none may be perfect but some are more useful than others. We would encourage capturing a range of representations where possible and offer the following steer as to what might be considered desirable:

- Representations designed for human consumption and abstract identifiers have an important role in the communication and linking of structural data but should be considered insufficient on their own and ideally be accompanied by a richer form of representation.
- If the precise structural isomer of a structure matters then this is best captured by linear representations such as SMILES or connection tables such as MOLfile.
- A goal of the infrastructure should be to provide standard identifiers for all structures for which this is possible. Provision of standard identifiers such as InChI by researchers should be encouraged but equally valid is a representation in a format from which an InChI can be reliably generated.
- 3D models of structures will invariably need to be accompanied by a more precise representation of the structure being studied.

3.4 Enablers of Best Practice

Ideally, a researcher would need to know little about many of the technical considerations covered in this report as tools and workflows would take care of these behind the scenes. This is largely the position reached within crystallography where:

- Software tools output data in standard formats and include all the data required for subsequent validation and review.
- Validation tools highlight aspects of the data that might need closer inspection with an expectation that severe alerts will be addressed or explained.
- Repositories provide workflows that enable deposition of data with required metadata and validation reports.
- Publishers promote policies and provide workflows that require data to be validated and deposited in standard formats prior to manuscript submission.
- Publishers and repositories provide workflows and services that enable data and validation reports to be used as part of the article peer review process, and ensure links between data and article are tracked from manuscript submission to publication and beyond.

Beyond crystallography, tools and workflows to support publication of structural data are less well established. Because of this it becomes important that researchers are aware of expectations that should be met so they can do their best to satisfy these and help create demand for improvements to services they rely on.

3.5 Structure Validation

Having a tool that can feedback on the validity of a digital representation of a structure could aid stakeholders across the data lifecycle. In particular:

- Researchers exporting representations from drawing tools can check that these do not represent something other than the structure(s) intended.
- Journal publishers can build confidence that datasets are properly identified.
- Data infrastructure providers can ensure consistency in representation across datasets.

A platform for validation and standardization of chemical structures according to sets of systematic rules has been described and used against large datasets [1]. This issued alerts of varying levels of severity, similar to the checkCIF service widely used in crystallography [2]. The service is unfortunately no longer directly accessible, although the code is available in GitHub [3].

As an enabler of best practice and an arbiter of the reliability of digital structure representations, a future service should contribute to the specification and development of a community chemical structure validation service.

3.6 Digital Lab Notebooks

Many of the challenges of reliably capturing structural data in digital forms could be minimised through adoption of digital lab notebooks throughout the research data lifecycle. Systems that can link directly to the instrumentation and software used to generate data offer a particular advantage as they reduce the need for manual transcription of information by researchers and should ensure more reliable capture of data. Of most relevance to this report are systems that include rich support for the storage and management of chemical structures.

3.6.1 Targeted Solutions

In the pharma industry the use of electronic lab notebooks, informatics platforms and inventory systems is well-established as a means of storing and tracking information for regulatory purposes and future exploitation. Many commercial solutions in this space have sophisticated support for the management and mining of structural data to support scientific discovery. The adoption of these solutions in academia is limited, perhaps in part because they are too specialised to provide a generic solution that can satisfy the breadth of research undertaken in a university environment, but also because the overall cost may exceed available budgets.

More generic lab notebook solutions have been emerging in recent years, several of which have specific support for management of compounds, reactions and molecules. Further, some academic laboratories have invested in developing lab notebooks and repositories targeted towards structural chemistry needs, most notably Chemotion from KIT [4]. Studies that explore barriers to the adoption of Electronic Lab Notebooks [5] and compare their various features [6] have been undertaken. A comprehensive list of ELN products compiled by the University of Cambridge currently includes over 30 available solutions [7].

A key consideration for a Physical Sciences Data Infrastructure is to what extent it should invest in developing solutions for the storage and search of structural data and information. This is particularly important when there are many pre-existing solutions with structure-based intelligence available already. If the infrastructure requires such a solution, it should look to build on one of these existing ones.

3.6.2 Generic Solutions

Part of a survey undertaken as part of this pilot[a] asked researchers in the physical sciences about their current use of digital notebook solutions. Responses barely mentioned the targeted solutions referred to in Section 3.7.1. The most favoured solutions were generic ones such as OneNote and Jupyter Notebooks. Whilst it is possible to do manipulation and display of chemical structure within Jupyter Notebooks, their suitability as a long-term store of data is perhaps unproven. Whilst one can connect OneNote to chemical structure drawing tools [8–10], or copy and paste structures from drawing packages, OneNote does not provide the sophistication of searching and storage offered by more targeted solutions.

This tendency towards generic solutions for the storage of scientific information in an unstructured form has resonances of a trend in the early 2000s for enterprises to adopt SharePoint as their central information system. This sparked the development of solutions that aimed to make SharePoint chemically intelligent [11–13]. Of note in this context also is Chem4Word, an add-in for Microsoft Word that enables chemical information inserted into a Word document to be stored and manipulated in a semantic way [14].

3.6.3 Notebook and Repository Interoperability

Given the range of platforms that might be adopted for management and storage of research data, an aspect a future infrastructure might usefully focus on is establishing generic mechanisms for the reliable exchange of structure-related data between these. We note here some generic solutions that have been proposed for the exchange of data and metadata between systems that could form the basis of solutions for this.

BagIt: This is a standard specification for organising related files [15] and forms the basis of a RDA recommendation for a standard format to enable exchange of data and metadata between repositories at a generic level [16].

dTool: A tool for packaging data and metadata developed to support exchange of data across decentralised research groups using different storage technologies [17].

RO-Crate: An emerging lightweight approach to packaging data and metadata in a way that can be used independently of infrastructure. Uses Schema.org annotations and JSON-LD to enable metadata to be described in an accessible and practical way for use in a wide variety of situations [18,19].

In developing a Physical Sciences Data Infrastructure, it will be important to give consideration to the structure-related capabilities required to enable reliable exchange of data and metadata between generic and specialised repositories and notebook solutions.

[a] The PSDI Case Study survey asked researchers in the physical sciences about their working practices and interaction with technology across the research data lifecycle. There is an intention to publish anonymised results on the STFC web site at some point.

3.7 Metadata Registration

The FAIR Data Principles [20] place a heavy emphasis on the availability of metadata alongside data and note that metadata should survive even if the actual data becomes unavailable. A way to ensure this aim is met, is to register metadata for a digital object in a public repository. Registration of metadata will typically occur when a Digital Object Identifier (DOI) is assigned to a research object. If the impact of this is to be meaningful then it is important for there to be more than the minimally required metadata.

Particularly important is taking advantage of standard structure representations and identifiers to reflect the substances and materials that might relate to a dataset. This can be done through subject fields that reflect compound name, SMILES and ideally InChIs [21–23]. Also important are links to related research objects expressed using standard identifiers. These enable the construction of Knowledge Graphs that can link research outputs and actors across organisations and domains [24]. Relevant here is the RDA Open Science Graphs for FAIR Data Interest Group which is focused on the interoperability between services and information models of various Open Science Graph initiatives [25].

However, not all research objects are going to get a DOI and alternative metadata registries will be required. A question is whether a Physical Sciences Data Infrastructure could provide a metadata catalogue for datasets that will support connections to other established registries.

A Physical Sciences Data Infrastructure should encourage and enable the registration of metadata in open registries, ensuring that this includes appropriate structure identifiers and links to related objects as a key enabler of findability and interoperability.

3.8 Samples and Structures

We have primarily confined ourselves in this report to the digital representation of substances and materials. We should not forget these will often have a physical manifestation and any one sample or specimen will be unique. We note here the existence of identifier schemes designed to uniquely identify samples, specifically IGSN rooted in the Earth Sciences community [26] and RRID which has grown out of the life sciences community [27].

InChI can also have a role in uniquely identifying the components of a sample and work has been undertaken to establish a specification for QR codes that could be used on sample bottles to help identify their contents, and in associated documentation required when transporting chemicals and materials [28].

A Physical Sciences Data Infrastructure should understand requirements for identification of samples in the physical sciences and how existing initiatives can aid with this.

3.9 Bulk Export of Structures

Not every operation that a researcher is going to want to perform on a structure will be possible within any one environment.

Software Interoperability at the CCDC

The CCDC produces a suite of software tools that themselves inter-operate with software packages from other vendors, including modelling suites, workflow tools, compound name generation software and third-party chemical drawing tools. Common to implementing these cross-vendor integrations is the need to exchange structures individually or in bulk. Over the years a range of formats referenced in Section 2 have been used to achieve this, including MOLfiles and SDfiles, MOL2 files, SMILES/SMARTS and even a blend of CIF and MIF. Sometimes these need to capture query features as well as a definition of a particular structure and other times they need to be augmented with features bespoke to a specific interchange. Often it is necessary for a starting representation of a structure to be manipulated to meet the conventions expected by third party packages.

It will be essential for a Physical Sciences Data Infrastructure to provide support for export of individual or subsets of structures accessible through a future infrastructure in a range of recognised formats. This should not just be the format in which the structure was provided but in formats that will enable interoperability of structures and data with other tools and software packages.

Citation is important for ascribing provenance and credit. If a future infrastructure is to provide access to a range of resources that may change over time and allow bulk export then the ability to cite subsets that may be diverse in their origin is a challenge that should be confronted. Initiatives that may help enable citation of subsets include:

RDA Dynamic Data Citation Recommendations: Guidelines for how to cite arbitrary subsets of dynamic data by ensuring in particular that data can be versioned and time-stamped and queries stored [29].

A Data “Reliquary”: A concept being developed by members of the earth science community to enable citation of a large number of individual datasets [30].

If bulk export of structures is provided as an option, consideration should be given to how the chosen subset can be reliably cited in any further work

3.10 Structure-based Discovery

If a future infrastructure becomes an aggregator or host of structure-based information and data, then it will be critical to provide the ability to query these based on structure composition and connectivity. Many existing systems and toolkits enable this and the main point we would emphasise here is that a future infrastructure should look to identify partnerships that can address this need through pre-existing solutions rather than developing such discovery services from scratch.

We emphasise in particular the importance of InChI for providing a standard way to enable both discovery and linking across distributed systems. How to resolve an InChI in a standard way has long been a consideration within the InChI community given the multiplicity of end points an InChI might have. The solution to this probably rests in having a standard framework that is adopted by resources with structural data to link to so that the APIs connecting these can speak a common language [31].

A Physical Sciences Data Infrastructure should support systematic discovery of data based on structure composition, connectivity and identity. It should look to reuse existing solutions to enable this. Importantly it should expose interfaces that allow the lookup and linking of data based on InChI.

3.11 References

1. K. Karapetyan, C. Batchelor, D. Sharpe, V. Tkachenko, and A. J. Williams (2015) "The Chemical Validation and Standardization Platform (CVSP): large-scale automated validation of chemical structure datasets", *J Cheminform*, 7(1), https://doi.org/10.1186/s13321-015-0072-8.

2. A. L. Spek (2009) "Structure validation in chemical crystallography", *Acta Crystallogr D Biol Crystallogr*, 65(2), https://doi.org/10.1107/S090744490804362X.

3. "Chemical Validation and Standardization Platform (CVSP) – ChemSpider Blog", https://blogs.rsc.org/chemspider/2018/11/30/chemical-validation-and-standardization-platform-cvsp/ (accessed 11 April 2022).

4. P. Tremouilhac *et al.* (2017) "Chemotion ELN: an Open Source electronic lab notebook for chemists in academia", *Journal of Cheminformatics*, 9(1), https://doi.org/10.1186/s13321-017-0240-0.

5. S. Kanza *et al.* (2017) "Electronic lab notebooks: can they replace paper?", *J Cheminform*, 9(1), https://doi.org/10.1186/s13321-017-0221-3.

6. Harvard Longwood Medical Area Research Data Management Working Group (2021) "Electronic Lab Notebook Comparison Matrix", https://doi.org/10.5281/ZENODO.4723752.

7. L. Cadwallader (2020) "Electronic Research Notebook Products", https://www.data.cam.ac.uk/data-management-guide/electronic-research-notebooks/electronic-research-notebook-products (accessed 21 April 2022).

8. "JChem for Office | Chemaxon Docs", https://docs.chemaxon.com/display/docs/jchem-for-office.md (accessed 12 April 2022).

9. "Tools for Microsoft Office | Scilligence", https://www.scilligence.com/web/scilligence-touchmol4office/ (accessed 11 April 2022).

10. Coderik (2021) "Add-in for recognition of hand-drawn chemical structures", *r/OneNote*, https://www.reddit.com/r/OneNote/comments/l0ls4c/addin_for_recognition_of_handdrawn_chemical/ (accessed 11 April 2022).

11. "JChem for SharePoint | ChemAxon", https://chemaxon.com/products/jchem-for-sharepoint (accessed 12 April 2022).

12. P. J. Skinner, P. M. D. Phil, R. Potenzone, K. Blanchard, and M. Schoenberg (2011) "Bridging the Knowledge Gap: Searching SharePoint, E-Notebook, Chromatography Data Systems and Unstructured Documents for Chemical and other Scientific Information to Enable Cooperation, Collaboration and Improved Decision Making". Available: http://www.cambridgesoft.com/code_land/pdf/Molecular_Medicine_TriConference.pdf.

13. K. Tallapragada, J. Chewning, D. Kombo, and B. Ludwick (2012) "Making SharePoint® Chemically Aware^{TM}", *Journal of Cheminformatics*, 4(1), https://doi.org/10.1186/1758-2946-4-1.

14. "Chem4Word", *Chemistry Add-In for Microsoft Word*, https://www.chem4word.co.uk/ (accessed 12 April 2022).

15. J. A. Kunze, J. Littman, L. Madden, J. Scancella, and C. Adams (2018) "The BagIt File Packaging Format (V1.0)", Internet Engineering Task Force, Request for Comments RFC 8493. https://doi.org/10.17487/RFC8493.

16. RDA Research Data Repository Interoperability WG (2018) "Research Data Repository Interoperability WG Final Recommendations". https://doi.org/10.15497/RDA00025.

17. T. S. G. Olsson and M. Hartley (2019) "Lightweight data management with dtool", *PeerJ*, 7, https://doi.org/10.7717/peerj.6562.

18. S. Soiland-Reyes *et al.* (2022) "Packaging research artefacts with RO-Crate", *Data Science*, Preprint(Preprint), https://doi.org/10.3233/DS-210053.

19. P. Sefton *et al.* (2022) "RO-Crate Metadata Specification 1.1.2", https://doi.org/10.5281/zenodo.5841615.

20. M. D. Wilkinson *et al.* (2016) "Comment: The FAIR Guiding Principles for scientific data management and stewardship", *Scientific Data*, 3(1), https://doi.org/10.1038/sdata.2016.18.

21. M. J. Harvey, N. J. Mason, A. McLean, and H. S. Rzepa (2015) "Standards-based metadata procedures for retrieving data for display or mining utilizing persistent (data-DOI) identifiers", *Journal of Cheminformatics*, 7(1), https://doi.org/10.1186/s13321-015-0081-7.

22. H. S. Rzepa, A. Mclean, and M. J. Harvey (2016) "InChI as a Research Data Management Tool", *Chemistry International*, 38(3–4), https://doi.org/10.1515/ci-2016-3-408.

23. M. J. Harvey, A. McLean, and H. S. Rzepa (2017) "A metadata-driven approach to data repository design", *Journal of Cheminformatics*, 9(1), https://doi.org/10.1186/s13321-017-0190-6.

24. H. Cousijn *et al.* (2021) "Connected Research: The Potential of the PID Graph", *Patterns*, 2(1), https://doi.org/10.1016/j.patter.2020.100180.

25. "Open Science Graphs for FAIR Data IG", *RDA*, https://www.rd-alliance.org/groups/open-science-graphs-fair-data-ig (accessed 12 April 2022).

26. J. Klump *et al.* (2021) "Towards Globally Unique Identification of Physical Samples: Governance and Technical Implementation of the IGSN Global Sample Number", *Data Science Journal*, 20, https://doi.org/10.5334/dsj-2021-033.

27. A. Bandrowski *et al.* (2015) "The Resource Identification Initiative: A cultural shift in publishing", *F1000Res*, 4, https://doi.org/10.12688/f1000research.6555.2.

28. "Specification of International Chemical Identifier (InChI) QR Codes for Labels on Chemical Samples", *IUPAC | International Union of Pure and Applied Chemistry*, https://iupac.org/recommendation/specification-of-international-chemical-identifier-inchi-qr-codes-for-labels-on-chemical-samples/ (accessed 12 April 2022).

29. A. Rauber, A. Asmi, D. van Uytvanck, and S. Proell (2015) "Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC)", https://doi.org/10.15497/RDA00016.

30. S. Stall *et al.* (2021) "Data Citation Community of Practice - 29 October 2021 Workshop". https://doi.org/10.5281/zenodo.5641236.

31. "*InChI Resolver*". InChI Resolver. Accessed: 12 April 2022. Available: https://github.com/inchiresolver/inchiresolver.

4 Storage and Curation of Structural Data

Contents

4.1 Overview

This section addresses considerations associated with the storage and curation of structural data. It looks not only at structures but also data that are associated with these and leads into more general considerations associated with enabling the availability of high-quality scientific data.

4.2 Sources of Structural Data

4.2.1 Journal Articles

A study undertaken in 2015 compared the degree to which articles in biology, chemistry, mathematics, and physics published in 2014 used and shared data [1]. Findings suggested that around 80% of articles in chemistry *use* data whilst only around 6% *shared* original data. However, when it came to articles *reporting* original data, 70% of chemistry articles did this, more than any of the other fields compared.

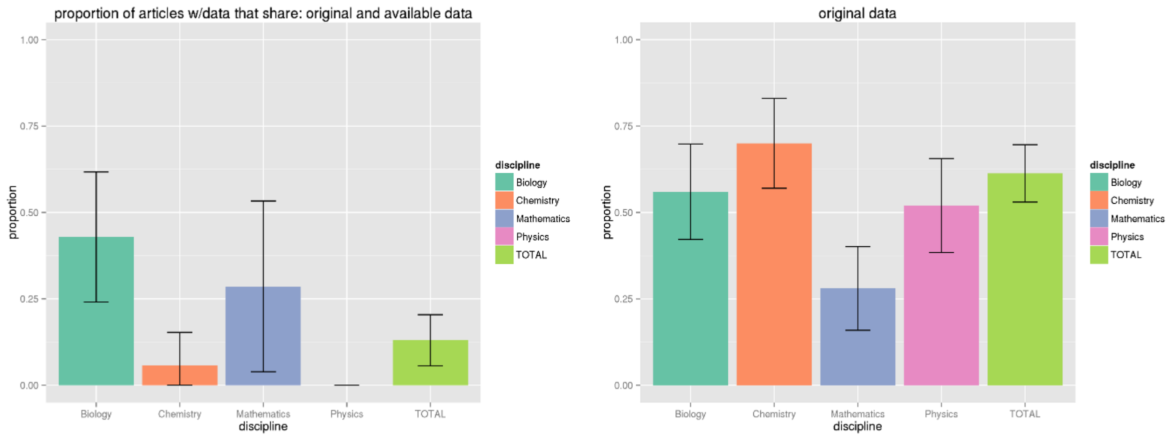


Figure 23: Charts from Womack RP (2015) [1]: On the left, proportion of articles by discipline that share data, among articles with original research data. On the right, proportion of articles by discipline with original data generated by the research described in the article. http://creativecommons.org/licenses/by/4.0/.

Another 2015 study profiled where common types of research data were published by organic synthetic chemists at the University of Michigan [2]. Of around 1,300 data occurrences, most (946) were reported in a PDF, with graphical image formats being the next most common (310). Just 24 were in what might be described as a structured format, with 17 of those being CIF. Most data occurrences were reported either in the body of the article or in Supporting Information.

There is often a tendency to believe that chemists in particular do not share data. We contend that they do and have a strong tradition of doing so: they just don't always do it in formats that are amenable to the digital age of data science.

A goal of a Physical Sciences Data Infrastructure should be to address barriers that prevent structural data being published in forms that are readily interpretable by machines.

4.2.2 Secondary Sources of Structural Information

Despite publishing practices that could be considered less than ideal, the rich body of structural data that has been published over the centuries is not completely lost to current generations of researchers. Reference volumes such as the Gmelin Handbook of Inorganic Chemistry first published in 1817 and the abstracting activities of the Chemical Abstracts Service (CAS) that commenced in the early 20th Century have laid the foundation of rich electronic resources such as Reaxys and SciFinder respectively. These include a wealth of curated structural information covering millions of substances abstracted from journals, patents, dissertations, and seminal reference works, all of which can be retrieved using structure-based search.

The CAS Registry [3] covers over 194 million organic and inorganic substances, including alloys, coordination compounds, minerals, mixtures, polymers, and salts disclosed in publications since the early 1800s and boasts that these have been enriched with ~8 billion experimental and predicted property data points and spectra.

Reaxys [4] has data on over 170 million organic, inorganic and organometallic substances with over 500 million published experimental facts, including substance property, spectral and reaction data.

A problem with these comprehensive resources is that they may not be available to all researchers in all institutions [5]. If just one resource is available, a researcher may be missing out on structures or properties that can only be found in others. Interfaces to these resources have primarily been targeted at interactive use by researchers. Both boast APIs enabling machine access although these are not typically available by default. There are, however, indications of a willingness to make APIs accessible for academic research purposes [6].

The EPSRC has a long history of enabling access to richly curated sources of data relevant to the physical sciences, initially through the Chemical Database Service [7] and latterly through the Physical Sciences Data-science Service [8]. A future Physical Sciences Data Infrastructure undoubtedly could have a continuing role to play here and one that enables machine access to a more comprehensive range of structural data resources in support of UK academic research endeavours.

Consideration should be given to building on the tradition of EPSRC support for UK academic access to highly curated data resources by enabling access to a wider range of resources, particularly those that offer extensive and comprehensive coverage of structural data and properties.

For any third-party resource that is made available through the infrastructure, the ability for UK researchers to access data through machine APIs should be secured.

4.2.3 Publicly Accessible Sources of Structural Information

In the absence of access to subscription resources, researchers can turn to publicly accessible resources of structural data that in recent years have evolved to be comparable with the more traditional resources described in Section 4.3.2. These include:

PubChem: Data for 111 million compounds drawn from 844 data sources with 293 million bioactivities. Funded by US government. [9,10]

ChemSpider: Access to over 100 million structures from hundreds of data sources including data on physical properties and spectra. Originally an individual enterprise, now owned and maintained by the Royal Society of Chemistry. [11,12]

ChEMBL : 2.1 million compounds with data on 1.4 million assays; manually curated, bringing together chemical, bioactivity and genomic data [13,14]. A sibling database, **SureChEMBL**, provides access to at least 17 million chemical entities mentioned in 14 million patents published since 1970 [15,16]. Both are maintained by EMBL-EBI.

All have APIs that enable machine accessibility to core data based on chemical structure although systematic access to all properties may be limited. Further, some data may not be available in a consistent and comparable form and may require processing before being used in data science activities. The values associated with physical properties are rarely accompanied by uncertainties; however, given a range of reported values from different sources these could potentially be derived.

There is opportunity for a Physical Sciences Data Infrastructure to save time for researchers and make data-driven research more efficient by investing in activities that make data in public resources available in more structured forms and through richer APIs.

Such an enterprise might best be done in partnership with those providing these resources: here we look to the model reflected by PubMed Central and Europe PMC where separate institutions have invested in providing complementary services built on a common core of publicly available content for the benefit of researchers worldwide. Could, for example PSDI be the core of a hypothetical PubChem UK or even PubChem Europe?

4.2.4 Theses and Dissertations

In 2017/2018, there were a reported 100,275 doctorate research postgraduate enrolments in the UK[a]. The number of first year starters was 29,025. A total of 70,045 Postgraduate Research studies were in science subject areas with 13,160 being in the physical sciences. There were 4,585 first year enrolments for physical science postgraduate research studies and 3,680 postgraduate research qualifications awarded. This suggests that around 4,000 theses are produced annually in physical science areas, many capturing 3-4 years of research. A fair proportion of these will involve the synthesis and study of structures and materials and the measurement and analysis of their properties. If articles are generated as a result of this research then some of this data may make it into the public domain, but much of it will be hidden from researchers and inaccessible to their machines. Conceivably the data in these theses will be unique to the studies undertaken and unlikely to be reproduced elsewhere in the scientific ecosphere.

In 2014 a report was published [18] describing efforts undertaken to mine PhD theses available through the British Library Electronic Theses Online Service (EThOS) for chemical compounds that could be included in a National Compound Collection – an initiative of the Royal Society of Chemistry. An initial pilot had been undertaken by a team of data collectors working across the country to manually redraw molecules to add them to the collection, an approach that would be difficult to scale [19]. The report examined more automated text and data mining approaches and the challenges encountered identifying theses of relevance, extracting structures and associated properties, and considerations regarding rights. It noted in its concluding section the benefits that could be accrued were data underpinning results reported in theses to be separately preserved in their native forms. A question this raises is what role a Physical Sciences Data Infrastructure could play in helping to realise this promise.

To be able to strongly encourage or even require physical sciences data generated as part of publicly funded PhD research to be made openly available, there need to be platforms that enable this to be done with minimal barriers to the actual researcher. Requiring this of PhD students would introduce them to best practices in research data management. It would also provide opportunity for them to contribute a fresh perspective on data publication approaches traditionally adopted and help identify improvements that may not be obvious to those for whom current practices have become engrained.

A future infrastructure should consider how it can provide the means, motive and opportunity for publication of structural data associated with doctoral theses and dissertations in machine-accessible and reusable forms – not just to ensure data are available for future research, but also to train and engage the next generation of researchers in best practices for data management and publication.

4.3 Research Data Repositories

The data publication platforms outlined so far are primarily focused on data published within or alongside journal articles and aggregations of data compiled from the literature or other

[a] Figures taken from Higher Education Student Statistics: UK, 2017/18 – Statistical Bulletin SB252 issued by HESA on 17 January 2019 [17].

data resources. There is another group of data publishing platforms that enable a researcher to publish their datasets more directly and deliberately, specifically data repositories. These can be classed into 3 distinct categories: Domain Repositories, Institutional Data Repositories, and General Data Repositories.

4.3.1 Domain Repositories

Some of the oldest domain repositories date back to the 1960s although in their early years they were similar to the abstracting services discussed in Section 4.3.2. The advent of the World Wide Web combined with the development of standards and tools for digital representation of data fuelled a change in expectations around how data should be made available and catalysed a transformation in how repositories operate.

4.3.1.1 Data Publishing Workflows

The development of data publishing paradigms in crystallography

Crystal structure data has been published in journals since the early 20th Century[b]. Some journal articles included tables of derived atom coordinates and supporting processed data[c]. Others might report just the atomic coordinates embedded in a paragraph within an article[d]. In the early days of crystallographic databases in the 1960s, a central activity was for coordinates published in articles to be retyped into database records. The early versions of these databases were in fact printed volumes generated from an electronic file of data [23]. Electronic access to data started to be established in the 1970s [24].

It wasn't until the 1990s when practices started to change significantly with the advent of digital data standards for capturing crystal structure data (CIF) and the increasing ubiquity of the Internet and the World Wide Web. It took time for these changes to have an impact, however. Initially it was not uncommon for CIFs to be published in PDFs or Word Documents as Supplementary Information rather than in their native form and manual effort was still required to transcribed data into a database. It took until 1999, 8 years after the CIF specification was published, for the number of structures being submitted to the Cambridge Structural Database (CSD) electronically in CIF form to exceed 80% [25]. Early digital submission to repositories was based on email with Web-based deposition processes becoming established in 2009 [26].

A combination of publisher policies, repository workflows and standards-compliant software tools means that today it is almost universal that a crystal structure will be deposited with a domain repository prior to submission of a manuscript. Accession IDs enable structures to be linked and tracked with the article through to publication. Assignment of DOIs to a dataset supports further linking and discovery. Structures that are not destined to be associated with an article can be published directly through databases: currently around 10% of structures added annually to the CSD are published independently as *CSD Communications* [27].

It is perhaps interesting to reflect where we are on this journey for other structural data types. Take spectra for example – it is not uncommon for these to be published as static images in

[b] The earliest structure in the CSD is of Graphite and was published in 1924 [20].

[c] Articles in the first volume of Acta Crystallographica published in 1948 included tables of atomic coordinates (derived data) and both observed and calculated structure-factor amplitudes (processed data). See e.g. [21].

[d] The fractional coordinates of four atoms observed in the asymmetric unit of a crystal structure of polyvinyl alcohol were reported in a single sentence of a paragraph in a 1948 Nature article [22].

supplementary information with key datapoints described in paragraph form. There is rarely any deposition of data in digital or semantic form. Repositories capable of storing spectral data do exist [28,29] but data deposition is not routinely required as part of publication workflows.

There are various initiatives that are aiming to change the way in which spectroscopy data are published [30–33] and it is hoped that these can learn from experiences in other domains such as crystallography to reach a much-improved endpoint in quicker time. Making the data available in digital form is one thing – curation and enrichment to enable it to be readily reusable for data science studies is the next step.

4.3.1.2 Domain-specific Data Curation

The more mature domain repositories have invested in and continue to develop workflows and expertise that enable data to be readily reused by researchers and machines across domains and ensure high levels of quality and trust. The key benefits this investment offers to the research data lifecycle include:

- Ensuring as comprehensive coverage of published data as possible – not just journal articles but theses, patents and other sources too. (*Discover and cite*, *Plan and design*)
- Reducing technical barriers to data publication through joined-up workflows with journal publishers. (*Share and publish*)
- Providing a platform for publication of datasets that might not be published alongside an article. (*Share and publish*)
- Ensuring data conforms to required standards of representation and quality at the point it is deposited. (*Reuse*)
- Providing a long-term home for data and providing trusted access for future generations of researchers. (*Manage, store, preserve*)
- Enriching data on an ongoing basis to ensure it is of sufficient completeness and quality for reuse beyond the context in which it was generated. (*Discover and reuse*)
- Tools to enable search, analysis and visualisation of aggregated collections of data. (*Collaborate and analyse, Discover and reuse*)

More general benefits of curation include:

- Enabling data to be reused in multidisciplinary contexts to address relevance to grand challenges and sustainable development goals.
- Supporting the application of data generated in an academic context to help address industry priorities.
- Seeding new disciplines – Bioinformatics, Cheminformatics and Solid Form Informatics have all evolved because of the availability of well-managed biological and chemical structure data that can be exploited by computational tools.
- Enabling individuals and communities with a drive and passion for aggregating and mining data to generate insights that transcend individual experiments.

Curation is best applied across the research data lifecycle, from when a dataset is first produced through to publication and beyond. Sometimes it can be automated but to achieve maximum value, involvement of experts in the domain is key. Fundamentally this is about enacting the FAIR Data Principles [34] to enable the cost savings and opportunities that reports have estimated as being significant for research-driven economies [35]. Domain repositories do this once to make it unnecessary for researchers around the globe to repeat data management processes in which they may not be expert and which have been estimated to account for 80% of the effort required for data-intensive analysis [36,37].

Key to the reusability and interoperability of chemistry data is having a reliable representation of the chemical substances to which the data or properties relate, and this is a core focus of curation activities in chemistry data repositories. To see the value of this – and the challenges of doing it well, we can compare two resources in crystallography – the CSD and the Crystallography Open Database (COD).

The benefits of investing in curation exemplified

The CSD [26] is supported by an organisation comprising scientists, software developers and other staff who facilitate the operation of data curation workflows that combine automated processing with review by structural chemistry experts. The COD [38] is a more limited operation relying primarily on automated processing. At this present time, the CSD has over 1 million structures[e] and the COD under half a million[f]. Pretty much every structure in the CSD has a reliable representation of the chemical structure associated with the data that has been reviewed by an expert allowing structure-based search across the collection. Of the 486k structures in the COD only 214k can be searched by a substructure query[g]. Anyone wishing to use the COD in a cheminformatics context would need to generate missing representations for the other 272k structures. Likely these would be the ones not readily handled by automated processes where intervention by an expert would be required to ensure a reliable representation. The CSD aims to avoid this burden having to be duplicated by researchers worldwide by investing time and resources in ensuring that structures are FAIR-ready for all.

An advantage of the COD over the CSD is that all of its data is made available in the public domain without restriction. All entries in the CSD are freely accessible on an individual basis but there are restrictions on free services when it comes to systematic access. The reason for this relates to the CSD's cost recovery model, highlighting that there is a cost to quality data curation that must be covered somehow. This is an aspect explored further in Section 4.5.

4.3.2 Institutional Data Repositories

In recent years, many research institutions have developed research data repositories to provide a place for their researchers to publish data in line with funder and journal policies. These may be implemented on platforms such as DSpace [39] or DataVerse [40] or through services provided by organisations such as FigShare [41]. The operation of these repositories is often overseen by support staff who are knowledgeable in best practices for research data management and able to advise researchers on general topics relating to data publication. These support staff are often generalists, however, and the extent to which they will be knowledgeable enough to provide domain-specific advice may be limited. A consequence of this can be that the formats in which data are published or the metadata provided through institutional repositories may fall far short of what is required to ensure data are discoverable and reusable at anything other than a very general level. This challenge has been recognised and there are various approaches being taken to improve the domain applicability of data stored in institutional repositories.

[e] With the update released 16 March 2022 the CSD contained over 1,156,000 structures (Archived).

[f] As at 1 April 2022, http://crystallography.net/cod/ reported 486,757 entries in the COD (Archived).

[g] As at 1 April 2022, http://www.crystallography.net/cod/search.html reported substructure search was possible for a subset of the COD containing 214,406 structures (Archived).

4.3.2.1 The Data Curation Network

The Data Curation Network [42] are producing data curation primers [43] that are peer-reviewed, living documents detailing a specific subject, disciplinary area or curation task and that can be used as a reference to curate research data. At the time of writing, there do not appear to be many primers targeted towards the physical sciences.

4.3.2.2 Embedded Data Stewards

Some institutions have established Data Stewards who are disciplinary experts with knowledge of data management and work within faculties in order to advise researchers and faculty members on the various aspects of research data management specific to their domain [44].

4.3.2.3 Data Champions

A variation on the above are Data Champions such as those found at the University of Cambridge [45]. Data Champions may be PhD Students, Post Docs or PIs who advise members of their research communities on proper handling of research data, producing guidance and forming networks across schools and departments. In chemistry, Data Champions have produced guides and commissioned talks to raise awareness of issues pertinent to chemistry [46].

We note here also the importance of individual advocates willing to invest time in pushing the boundaries of existing data infrastructures to accommodate new expectations and requirements, demonstrate these in action and highlight where improvements may be needed [47].

4.3.3 General Data Repositories

When there is neither a domain repository nor an institutional one, researchers can fall back on general data repositories such as FigShare [48] or Zenodo [49]. Here the reusability of the data and the quality of metadata will be very much in the hands of this researcher. If published in a domain-specific format and accompanied by appropriate domain-specific metadata, the dataset may be discoverable and reusable, but there is unlikely to be the consistency and completeness that can be imposed by domain repositories and, to a degree, institutional ones. An approach that exploits use of general repositories for publishing data from high performance computing jobs whilst ensuring rich metadata has recently been proposed [50].

4.3.4 Recommendations Relating to Research Data Repositories

Don't reinvent the wheel – if there are domain repositories out there, consider how best to connect to these to avoid duplication of data publishing and curation effort and resources.

Champion the importance of and requirements for quality data publishing and curation in the physical sciences.

Adopt existing curation guidelines where these are available – work with wider communities to develop new ones where they are not.

Promote the importance of machine interpretable data formats to support efficient publication workflows and enable data reuse – advocate against practices that result in non-semantic publication of data that cannot be reliably interpreted by machines.

Where possible and appropriate, require an indication of the chemical identity of a substance to be provided alongside the data, ideally in a machine-interpretable form. Encourage the use of standard identifiers to enable discovery and interoperability across data publication platforms.

Identify and highlight gaps in domain-specific data storage and curation and cultivate communities to address these. Consider providing a publication platform for physical sciences data that does not have a natural domain-specific home.

Become a centre of expertise and best practice in research data management and curation in the physical sciences alongside provision of any technical infrastructure.

4.4 Research Data Repository Sustainability

As noted in Section 4.4.1, data curation comes at a cost and this section outlines sustainability models that enable these activities and which a future infrastructure may want to consider.

4.4.1 Research Data Repository Operating Models

Table 1 indicates some operational characteristics of data resources mentioned in this report along with some other exemplars. The operating models of these resources are elaborated on further below.

Data Resource	Operating Model	Data Availability	URL
Chemical Abstracts Service	Society owned	Mostly closed	https://www.cas.org/
ChemSpider		Free access	https://www.chemspider.com/
Cambridge Structural Database	Value-added services	Some restrictions apply	https://www.ccdc.cam.ac.uk/
Inorganic Crystal Structure Database			https://icsd.products.fiz-karlsruhe.de/
Powder Diffraction File			https://www.icdd.com/
Crystallography Open Database	Project Funding	Free access	http://crystallography.net/cod/
Dryad	Data Publishing Charges	Open	https://datadryad.org/
Protein Data Bank	Publicly funded	Open	http://www.wwpdb.org/
ChEMBL			https://www.ebi.ac.uk/chembl/
PubChem			https://pubchem.ncbi.nlm.nih.gov/
NERC Data Centres			https://eds.ukri.org/

Table 4: Operating characteristics of a selection of scientific data resources.

4.4.1.1 Society Owned

The **Chemical Abstracts Service** (CAS) is a Division of the American Chemical Society. It is largely a closed resource but has recently expanded its openly available CAS Common Chemistry collection [51] to include nearly 500,000 substances. However, this is just 0.2% of the 182 million substances in its registry.

ChemSpider is owned by the Royal Society of Chemistry. It can be freely searched through web and programmatic interfaces, but data cannot be easily and comprehensively downloaded in bulk.

4.4.1.2 Self-financing through Value-added Services

The **Cambridge Structural Database** (CSD), the **Inorganic Crystal Structure Database** (ICSD), and the **Powder Diffraction File** are all funded by revenue generated by non-profit organisations through the provision of value-added services built on top of the data that they provide. All structures from the CSD and ICSD are freely available on an individual basis. Revenue is primarily generated through software and expert services. Geographically sensitive pricing along with programmes such as FAIRE [52] and FACE [53] aim to ensure all data and tools are accessible to academic research communities around the globe. Associated with the Powder Diffraction File is a Grant-in-Aid programme that extends financial support to investigators for the preparation of powder diffraction data [54]. There is no charge to deposit data in any of these resources.

4.4.1.3 Data Publishing Charges

Dryad is included as an example of a resource that recovers its costs through Data Publishing Charges. The base price for depositing a dataset at time of writing is $120 USD[h] although this cost may be hidden from the researcher through institutional or publisher arrangements.

4.4.1.4 Project Funded

The **Crystallography Open Database** (COD) has primarily been funded through fixed-term project funding but has also received contributions from community and commercial users [56].

4.4.1.5 Publicly Funded

4.4.1.5.1 Global Consortia

The **Protein Data Bank** (PDB) is sustained through a global consortium of organisations in the US, Europe and Japan (the wwPDB). Member organisations are funded regionally and collaborate on data deposition and archive activities whilst independently providing services for data consumption. This takes advantage of diversified funding sources[i] and results in a complementary, sometimes competing, services for data consumers. The wwPDB has also established a Foundation [58] to raise funds in support of additional outreach activities not covered by other grants.

4.4.1.5.2 Funded through Host Institution

ChEMBL is funded as part of the EMBL-EBI, an initiative funded by EU member states and others providing data and computational resources to support life science research in academia and industry. EMBL-EBI hosts the hub for Elixir which unites Europe's leading life science organisations to ensure the ongoing stewardship of publicly funded research data.

A study on the value and impact of the EBI published in 2021 [59] estimated that the value users place on EBI data resources in a 12-month period is £1.25 billion for an annual investment of £110 million. Users of EBI services reported efficiency savings estimated to be the equivalent of at least £2.6 billion and possibly up to £11 billion per annum worldwide. The contribution of services to realising wider research impacts was estimated as £2.2 billion per year.

4.4.1.5.3 Directly Funded

PubChem is funded by the Intramural Research Programme of the National Library of Medicine, one of the National Institutes of Health in the US [9].

[h] From Dryad FAQ, March 2022 [55] (Archived).

[i] Nine funders worldwide acknowledged at time of writing [57] (Archived).

NERC Data Centres are specifically funded to support the data storage and curation needs of projects funded by the National Environment Research Council in the UK. As well as providing infrastructure they also provide expert guidance at all stages of a research project, not just at the point of publication. To some degree, they blend the Data Steward approach with the Domain Repository one.

An analysis of the economic value of one of the NERC data centres, the British Atmospheric Data Centre, published in 2013 [60] suggested the centre enabled a significant increase in research efficiency estimated to be worth at least £10 million per annum and between a 4-fold to 12-fold return on investment.

4.4.2 General Principles of Research Data Repository Sustainability

4.4.2.1 Business Models for Sustainable Research Data Repositories

An OECD report on Business Models for Sustainable Research Data Repositories published in 2017 provides a comprehensive overview of different models that could be adopted to sustain long-term stewardship of data in support of open research [61]. Some key points from this report include:

- The design and sustainability of research data repository business models depend on many factors. A clearly articulated business model is indispensable for all research data repositories.
- Cost optimisation efforts can help ensure the effective and sustainable management of digital assets over time. Taking advantage of economies of scale is also important.
- A successful business model has to align with a repositories mission and be sensitive to the context in which it operates – these may be subject to change over time.

The report notes that project funding can provide a mechanism to test the need for a data repository and the seed money needed to create one but that as the repository matures, a different funding model is likely to be needed.

4.4.2.2 The Principles of Open Science Infrastructure

The Principles of Open Science Infrastructure (POSI) [62] reinforce messages from the OECD report by discouraging use of time-limited funds to fund day-to-day activities needed to support long term infrastructure. They also encourage revenue sources consistent with mission and based on services not data

Many will argue that the ideal funding model is one which allows data and services to be made freely and openly available for all. Where this is not possible, business models that support free and open data while charging some or all users for access to value-adding services are next preferred. It must be recognised that the need to monetise services will likely lead to some restrictions on breadth, depth or timeliness of access to data.

As well as addressing sustainability, the Principles of Open Science Infrastructure also note requirements relating to insurance and governance. Insurance requirements can be fulfilled if software and data related to the resource are made openly available. Good governance should ensure that the needs of user communities are met by infrastructure.

Several research infrastructure organisations have recently signed up publicly to the Principles of Open Science Infrastructure, noting how they meet them and where they currently fall short [63].

4.4.3 Research Data Repository Sustainability Recommendations

We make the following recommendations regarding the funding and sustainability of curation and storage activities for structural data, and for data infrastructure generally.

Recognise that investment in data infrastructure requires investment in expertise as well as technology.

Learn from the experiences of existing data repositories when considering benefits of investment in data storage and curation.

Consider studies that demonstrate the return on investment of established physical sciences data resources for the wider economy.

Invest for the long term, not the short term and ideally to enable all data and services to be openly available.

Recognise that if data storage and curation activities have to be self-sustaining then some access restrictions or barriers may be required.

Consider adopting Principles of Open Science Infrastructure to guide the governance and sustainability of data infrastructure.

Consider where there may be opportunity for national and international cooperation to pool resources and minimise costs.

4.5 Trust Frameworks for Structural Data Resources

CoreTrustSeal is an international organisation that assesses and certifies the trustworthiness of data repositories [64,65]. Certification is based on self-assessment by a repository which is reviewed by members of the CoreTrustSeal Assembly of Reviewers [66]. The criteria against which a repository is judged [67] are:

- Organisational Infrastructure: Clear mission; sufficient funding and continuity plans; compliance with legal and ethical norms; governance and guidance.
- Digital Object Management: Assessing integrity, authenticity and quality; having documented preservation plans and workflows; supporting discovery and reuse.
- Technology: Well-supported, secure, infrastructure.

CoreTrustSeal essentially implements the TRUST Principles relating to Transparency, Reuse, User-centricity, Sustainability, Technology [68].

Re3data [69] is a registry of data repositories that can be searched by subject and certification. In the Natural Sciences there are 43 disciplinary CoreTrustSeal repositories. Excluding those that serve the Geosciences, there are really just two certified repositories that could be considered to serve the structural science community – the CSD and the PDB.

Certification aside, identifying resources where one can deposit structural data is no easy task. The most recent ACS style guide identifies a list of 17 chemistry-friendly repositories spanning crystallography, spectroscopy, geochemical data, and properties of materials and assay data [70]. Five of these might be considered more biological than chemical. Re3data and FAIRsharing, a curated collection of data metadata standards, repositories and polices [71], both allow subject-based searches for data repositories associated with chemistry. A key challenge here is narrowing this down to a list that support data publication generally and aren't tied to a particular project and thus unavailable for open deposition.

In 2016, the Elixir project published criteria for what they consider to be Core Data Resources fundamental for the life sciences [72]. These included qualitative and quantitative indicators that reflect:

- Scientific focus and quality of science.
- Community served by the resource.
- Quality of service.
- Legal and funding infrastructure, and governance.
- Impact and translational stories.

The aim of this exercise was to establish the most valuable, used and useful resources for their user communities. They initially identified around 20 European data resources that met these criteria [73] and a further list of Deposition Databases [74] that either met or nearly met the criteria and accept deposition of data from researchers worldwide. The concepts behind this initiative are now being extended globally through the Global Biodata Coalition [75].

Establishing criteria that characterise a trusted repository in the physical sciences that will accept deposition of data would be of benefit to the wider scientific community but also important in establishing gaps that an infrastructure might wish to fill or partners it should work with.

Finally, we note here frameworks for assessing the FAIRness of individual datasets of which there are many emerging [76]. These will be general in their approach but might benefit from a disciplinary take on what it means for e.g., a dataset associated with structural chemistry or materials science to have a high level of FAIR maturity, particularly where metadata is concerned.

Considering criteria that make a structural dataset in the physical sciences FAIR will help to establish the requirements an infrastructure will need to support and expect. Foremost amongst these must be a unique identifier of each substance associated with the dataset.

4.6 References

1. R. P. Womack (2015) "Research Data in Core Journals in Biology, Chemistry, Mathematics, and Physics", *PLoS ONE*, 10(12), https://doi.org/10.1371/journal.pone.0143460.

2. Y. Li and J. Thielen (2015) "Profiling common types of research data and methods published by organic synthesis chemists at the University of Michigan", Accessed: 13 April 2022. Available: http://deepblue.lib.umich.edu/handle/2027.42/111832.

3. "CAS REGISTRY", *CAS*, https://www.cas.org/cas-data/cas-registry (accessed 13 April 2022).

4. "Reaxys - An expert-curated chemistry database", *Elsevier.com*, https://www.elsevier.com/en-gb/solutions/reaxys (accessed 13 April 2022).

5. chemlibrarian (2020) "SciFinder-n will be cancelled with effect from 1 October 2020", *Chemistry Library blog*, https://cambridgechemlib.wordpress.com/2020/09/02/scifinder-n-will-be-cancelled-with-effect-from-1-october-2020/ (accessed 13 April 2022).

6. "CAS opens data vault to MIT scientists", *Chemical & Engineering News*, https://cen.acs.org/physical-chemistry/computational-chemistry/CAS-opens-data-vault-MIT/98/web/2020/11 (accessed 13 April 2022).

7. B. McMeeking and D. Fletcher (2004) "The United Kingdom Chemical Database Service: CDS", *Cheminformatics*, 1(1–2).

8. "The EPSRC Physical Sciences Data-science Service", https://www.psds.ac.uk/ (accessed 13 April 2022).

9. S. Kim *et al.* (2021) "PubChem in 2021: new data content and improved web interfaces", *Nucleic Acids Research*, 49(D1), https://doi.org/10.1093/nar/gkaa971.

10. "PubChem", https://pubchem.ncbi.nlm.nih.gov/ (accessed 13 April 2022).

11. H. E. Pence and A. Williams (2010) "ChemSpider: An Online Chemical Information Resource", *J. Chem. Educ.*, 87(11), https://doi.org/10.1021/ed100697w.

12. "ChemSpider | Search and share chemistry", http://www.chemspider.com/ (accessed 13 April 2022).

13. A. Gaulton *et al.* (2017) "The ChEMBL database in 2017", *Nucleic Acids Research*, 45(D1), https://doi.org/10.1093/nar/gkw1074.

14. "ChEMBL Database", https://www.ebi.ac.uk/chembl/ (accessed 13 April 2022).

15. G. Papadatos *et al.* (2016) "SureChEMBL: a large-scale, chemically annotated patent document database", *Nucleic Acids Research*, 44(D1), https://doi.org/10.1093/nar/gkv1253.

16. "SureChEMBL", https://www.surechembl.org/ (accessed 13 April 2022).

17. "Higher Education Student Statistics: UK, 2017/18 | HESA", https://www.hesa.ac.uk/news/17-01-2019/sb252-higher-education-student-statistics (accessed 20 April 2022).

18. K. McNeice (2014) "Text and Data Mining in EThOS", British Library Research Repository. https://doi.org/10.23636/1162.

19. D. M. Andrews *et al.* (2016) "The creation and characterisation of a National Compound Collection: the Royal Society of Chemistry pilot", *Chem. Sci.*, 7(6), https://doi.org/10.1039/C6SC00264A.

20. O. Hassel and H. Mark (1924) "Über die Kristallstruktur des Graphits", *Z. Physik*, 25(1), https://doi.org/10.1007/BF01327534.

21. E. M. Archer (1948) "The crystal structure of p-chloriodoxybenzene", *Acta Cryst*, 1(2), https://doi.org/10.1107/S0365110X48000193.

22. C. W. Bunn (1948) "Crystal Structure of Polyvinyl Alcohol", *Nature*, 161(4102), https://doi.org/10.1038/161929a0.

23. O. Kennard, D. G. Watson, F. H. Allen, and S. Bellard (1971) "*Molecular Structures and Dimensions, (Vol 1–15)*". Reidel, Dordrecht, The Netherlands.

24. F. H. Allen *et al.* (1979) "The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information", *Acta Crystallographica Section B*, 35(10), https://doi.org/10.1107/S0567740879009249.

25. "Building a complete picture - The Cambridge Crystallographic Data Centre (CCDC)", https://www.ccdc.cam.ac.uk/Community/blog/building-a-complete-picture/ (accessed 13 April 2022).

26. C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward (2016) "The Cambridge Structural Database", *Acta Cryst B*, 72(2), https://doi.org/10.1107/S2052520616003954.

27. "ISSN 2631-9888 (Online) | CSD Communications | The ISSN Portal", https://portal.issn.org/resource/ISSN/2631-9888 (accessed 13 April 2022).

28. S. Kuhn and N. E. Schlörer (2015) "Facilitating quality control for spectra assignments of small organic molecules: nmrshiftdb2 – a free in-house NMR database with integrated LIMS for academic service laboratories", *Magnetic Resonance in Chemistry*, 53(8), https://doi.org/10.1002/mrc.4263.

29. H. Horai *et al.* (2010) "MassBank: a public repository for sharing mass spectral data for life sciences", *Journal of Mass Spectrometry*, 45(7), https://doi.org/10.1002/jms.1777.

30. A. M. Hunter, E. M. Carreira, and S. J. Miller (2020) "Encouraging Submission of FAIR Data at the Journal of Organic Chemistry and Organic Letters", *Journal of Organic Chemistry*, 85(4), https://doi.org/10.1021/acs.joc.0c00248.

31. B. C. Sorkin, J. M. Betz, and D. C. Hopp (2020) "Toward FAIRness and a User-Friendly Repository for Supporting NMR Data", *Org. Lett.*, 22(8), https://doi.org/10.1021/acs.orglett.0c01143.

32. P. Tremouilhac *et al.* (2021) "Chemotion Repository, a Curated Repository for Reaction Information and Analytical Data", *Chemistry–Methods*, 1(1), https://doi.org/10.1002/cmtd.202000034.

33. R. M. Hanson *et al.* (2022) "IUPAC specification for the FAIR management of spectroscopic data in chemistry (IUPAC FAIRSpec) – guiding principles", *Pure and Applied Chemistry*, 0(0), https://doi.org/10.1515/pac-2021-2009.

34. M. D. Wilkinson *et al.* (2016) "Comment: The FAIR Guiding Principles for scientific data management and stewardship", *Scientific Data*, 3(1), https://doi.org/10.1038/sdata.2016.18.

35. "Cost-benefit analysis for FAIR research data: Cost of not having FAIR Research Data". Available: https://op.europa.eu/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1/language-en.

36. M. Brodie (2015) "Understanding Data Science: An Emerging Discipline for Data Intensive Discovery", Obninsk, Russia. Available: http://ceur-ws.org/Vol-1536/paper32.pdf.

37. M. Brodie (2015) "Understanding Data Science: An Emerging Discipline for Data-Intensive Discovery", in *Getting Data Right*, O'Reilly Media, Inc.

38. S. Gražulis *et al.* (2012) "Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration", *Nucleic Acids Research*, 40(D1), https://doi.org/10.1093/nar/gkr900.

39. M. Smith *et al.* (2003) "DSpace: An Open Source Dynamic Digital Repository", *D-Lib Magazine*, 9(1), https://doi.org/10.1045/january2003-smith.

40. M. Crosas (2011) "The Dataverse Network®: An Open-Source Application for Sharing, Discovering and Preserving Data", *D-Lib Magazine*, 17(1/2), https://doi.org/10.1045/january2011-crosas.

41. "figshare - research data portal for institutions to showcase", https://knowledge.figshare.com/institution/products/data-repository (accessed 13 April 2022).

42. "Data Curation Network – Ethical. Reusable. Better.", https://datacurationnetwork.org/ (accessed 13 April 2022).

43. "*data-primers*". DataCurationNetwork. Accessed: 13 April 2022. Available: https://github.com/DataCurationNetwork/data-primers.

44. M. Teperek, M. J. Cruz, E. Verbakel, J. Böhmer, and A. Dunning (2018) "Data Stewardship addressing disciplinary data management needs", *IJDC*, 13(1), https://doi.org/10.2218/ijdc.v13i1.604.

45. "Establishing, Developing, and Sustaining a Community of Data Champions", https://datascience.codata.org/articles/10.5334/dsj-2019-023/ (accessed 13 April 2022).

46. "Chemistry Data Champions - Achievements | Chemistry Library", https://www-library.ch.cam.ac.uk/chemistry-data-champions-achievements (accessed 13 April 2022).

47. H. S. Rzepa (2022) "The Long and Winding Road towards FAIR Data as an Integral Component of the Computational Modelling and Dissemination of Chemistry", *Israel Journal of Chemistry*, 62(1–2), https://doi.org/10.1002/ijch.202100034.

48. "figshare - credit for all your research", https://figshare.com/ (accessed 13 April 2022).

49. "Zenodo - Research. Shared.", https://zenodo.org/ (accessed 13 April 2022).

50. C. Cave-Ayland, M. Bearpark, C. Romain, and H. S. Rzepa (2022) "CHAMP is a HPC Access and Metadata Portal", *Journal of Open Source Software*, 7(70), https://doi.org/10.21105/joss.03824.

51. "CAS Common Chemistry", https://commonchemistry.cas.org/ (accessed 13 April 2022).

52. "Frank H. Allen International Research and Education Programme - The Cambridge Crystallographic Data Centre (CCDC)", https://www.ccdc.cam.ac.uk/Community/FAIRE/ (accessed 13 April 2022).

53. "FIZ Karlsruhe Advancing Crystallography Education Programme | ICSD", https://icsd.products.fiz-karlsruhe.de/en/support/fiz-karlsruhe-advancing-crystallography-education-programme (accessed 13 April 2022).

54. "Grant-in-Aid – ICDD", https://www.icdd.com/grant-in-aid/ (accessed 13 April 2022).

55. "Dryad FAQ - Publish and Preserve your Data", https://datadryad.org/stash/faq (accessed 14 March 2022).

56. I. Bruno, S. Gražulis, J. R. Helliwell, S. N. Kabekkodu, B. McMahon, and J. Westbrook (2017) "Crystallography and Databases", *Data Science Journal*, 16, https://doi.org/10.5334/dsj-2017-038.

57. "wwPDB: FAQ", https://www.wwpdb.org/about/faq (accessed 13 April 2022).

58. "wwPDB Foundation", https://foundation.wwpdb.org/ (accessed 13 April 2022).

59. Neil Beagrie and John Houghton (2021) "Data-driven discovery: The value and impact of EMBL-EBI managed data resources". Accessed: 13 April 2022. Available: https://www.embl.org/documents/document/embl-ebi-impact-report-2021.

60. Neil Beagrie and John Houghton (2013) "The Value and Impact of the British Atmospheric Data Centre". Accessed: 13 April 2022. Available: https://repository.jisc.ac.uk/5382/1/BADCReport_Final.pdf.

61. "Business models for sustainable research data repositories", OECD Science, Technology and Industry Policy Papers 47. https://doi.org/10.1787/302b12bb-en.

62. G. Bilder, J. Lin, and C. Neylon (2020) "The Principles of Open Scholarly Infrastructure", https://doi.org/10.24343/C34W2H.

63. "Posse", *The Principles of Open Scholarly Infrastructure*, https://openscholarlyinfrastructure.org/posse/ (accessed 13 April 2022).

64. "CoreTrustSeal", *CoreTrustSeal*, https://www.coretrustseal.org/ (accessed 13 April 2022).

65. H. L'Hours, M. Kleemola, and L. de Leeuw (2019) "CoreTrustSeal: From academic collaboration to sustainable services", *IASSIST Quarterly*, 43(1), https://doi.org/10.29173/iq936.

66. "Assembly of Reviewers", *CoreTrustSeal*, https://www.coretrustseal.org/about/assembly-of-reviewers/ (accessed 13 April 2022).

67. CoreTrustSeal Standards and Certification Board (2019) "CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022", https://doi.org/10.5281/zenodo.3638211.

68. D. Lin *et al.* (2020) "The TRUST Principles for digital repositories", *Sci Data*, 7(1), https://doi.org/10.1038/s41597-020-0486-7.

69. "re3data.org", https://doi.org/10.17616/R3D (accessed 13 April 2022).

70. Leah McEwen (2019) "3.1.5 Preparing Your Data for Publication", in *The ACS Guide to Scholarly Communication*, American Chemical Society. https://doi.org/10.1021/acsguide.30105.

71. S.-A. Sansone *et al.* (2019) "FAIRsharing as a community approach to standards, repositories and policies", *Nat Biotechnol*, 37(4), https://doi.org/10.1038/s41587-019-0080-8.

72. C. Durinx *et al.* (2017) "Identifying ELIXIR Core Data Resources". F1000Research. https://doi.org/10.12688/f1000research.9656.2.

73. "ELIXIR Core Data Resources", *ELIXIR*, https://elixir-europe.org/platforms/data/core-data-resources (accessed 13 April 2022).

74. "ELIXIR Deposition Databases for Biomolecular Data", *ELIXIR*, https://elixir-europe.org/platforms/data/elixir-deposition-databases (accessed 13 April 2022).

75. "Global Biodata Coalition", https://globalbiodata.org/ (accessed 13 April 2022).

76. "FAIRassist.org", https://fairassist.org/#!/ (accessed 13 April 2022).

5 Collaboration With Other Initiatives

Contents

5.1 Overview

This section provides an overview of organisations and initiatives actively working on issues relevant to the representation of structure and the management of structural data. There are two over-riding recommendations to make concerning this:

A Physical Sciences Data Infrastructure should avoid as much as possible reinventing the wheel and look to align with others working to address similar requirements whether through repurposing or collaboration.

A Physical Sciences Data Infrastructure should invest in helping to improve existing wheels by contributing expertise and other support to efforts aimed at the development of shared standards and infrastructure.

Standards bodies are not financially well-endowed, so activities rely a lot on voluntary effort which can dramatically impact on the pace at which these can advance. Creating opportunities for practitioners to contribute to both the development and coordination of community standards activities as an embedded part of their role could help advance these significantly.

We also note in this section the importance of establishing strong relationships with industry in order to ensure a future infrastructure delivers value for the UK industrial sector and economy.

5.2 Domain Initiatives

5.2.1 International Union of Pure and Applied Chemistry

The International Union of Pure and Applied Chemistry (IUPAC) [1] has a track record of over a century in establishing standards for reliably communicating entities and concepts related to the chemical sciences, including chemical identity and representation, standard property data and definitive terminology. Many of these emerged from an analogue age but there has been a digital focus for some years, including JCAMP-DX – a format for exchange of spectra information [2], InChI – the International Standard Chemical Identifier [3], and the GoldBook – an online digital aggregation of IUPAC terminologies [4]. Table 1 lists current IUPAC projects that are of particular relevance to physical sciences data infrastructure.

IUPAC Project	URL
IUPAC SMILES+ Specification	https://iupac.org/project/2019-002-2-024
Gold Book Developments	https://iupac.org/project/2019-032-1-024
Machine-Accessible Periodic Table	https://iupac.org/project/2019-020-2-024
A Standard for FAIR Data Management of Spectroscopic Data	https://iupac.org/project/2019-031-1-024
A Metadata Schema for Critically Evaluated Solubility Measurement Data	https://iupac.org/project/2020-018-1-024
Development of a Machine Accessible Kinetic Databank for Radical Polymerizations	https://iupac.org/project/2019-045-1-400
Guidance for the Compilation, Critical Evaluation and Dissemination of Chemical Data	https://iupac.org/project/2018-009-2-500
Standardized Reporting of Gas Adsorption Isotherms	https://iupac.org/project/2021-016-1-024
Revisions to ThermoML	https://iupac.org/project/2017-016-3-100
Projects to develop guidelines for the graphical representation of structures	https://iupac.org/project/2017-039-2-800 https://iupac.org/project/2017-036-2-800

Table 5: IUPAC projects with relevance to physical sciences data infrastructure.

IUPAC also supports a range of InChI-related projects detailed in Section 5.3.2 and has recently established a partnership with the Pistoia Alliance (Section 5.6.1) to formalise and advance HELM as a standard. It collaborates with organisations such as CODATA (Section 5.5.2) and the Bureau International des Poids et Mesures [5] on matters concerning the digital representation of units and physical constants.

A general challenge for IUPAC is how to build momentum and drive collaborations that will enable more rapid progress with digital standards on a broader front [6,7]. To this end, it has recognised the strategic opportunity to develop a Centre of Excellence for Digital Chemistry Standards which aims to address community-wide needs including:

- Standards that support effective data publication and reuse.
- Use cases that bridge industry, academic and global communities.

- Training in Digital Chemistry for the future scientific workforce.
- Coordination, communication and facilitation across domains and initiatives.

This initiative will leverage IUPAC's successful model of distributed volunteer expertise and community governance and planning is underway to develop a roadmap and assess infrastructure needs for a broader program of digital standards development. A consideration for a Physical Sciences Data Infrastructure should be the degree to which it might be able to help facilitate such an endeavour.

5.2.2 The InChI Trust

The InChI Trust [8] is responsible for supporting the development, maintenance and adoption of InChI, a task it undertakes in collaboration with IUPAC. The Trust and IUPAC coordinate scientific development of InChI through project groups that focus on specific areas [9]. Table 2 provides a summary of current InChI project areas.

InChI Project	URL
InChI Requirements for Organometallic and Coordination Compound structures	https://iupac.org/project/2009-040-2-800
Handling of Inorganic Compounds for InChI	https://iupac.org/project/2012-046-2-800
Redesign of Handling of Tautomerism in InChI	https://iupac.org/project/2012-023-2-800
Implementation of InChI for Chemically Modified Large Biomolecules	https://iupac.org/project/2013-010-1-800
QR Codes and Industry Applications	https://iupac.org/project/2015-019-2-800
InChI Extension for mixture composition	https://iupac.org/project/2015-025-4-800
Open Education Resources for InChI	https://iupac.org/project/2018-012-3-024
Enhanced Recognition and Encoding of Stereoconfiguration by InChI Tools	https://iupac.org/project/2019-017-2-800
Nanomaterials [10]	https://codata.org/initiatives/task-groups/extension-of-inchi-for-nanomaterials/
Isotopologues [11]	https://github.com/MSI-Metabolomics-Standards-Initiative/inchi-isotopologue-extension
Markush and Variability	https://github.com/Goodman-lab/InChiMarkush
InChI Resolver Reference Implementation	https://github.com/inchiresolver/inchiresolver

Table 6: Active InChI project groups.

The InChI Trust is further investing in work that will enable wider practical contributions from the community to development of the technology required to support InChI as a standard.

Engagement in and support for the future development and sustainability of InChI should be a key consideration for a Physical Sciences Data Infrastructure.

5.2.3 International Union of Crystallography

The IUCr [12] has undoubtedly been a pioneer in the effective communication of structural information and in establishing standards, tools and workflows that support this in crystallography. This has supported and enabled disciplinary data repositories that ensure the availability of FAIR data across sub-domains. Lessons from these experiences have been woven into this report and should continue to be a reference point for future planning of a Physical Sciences Data Infrastructure.

A current consideration of the IUCr Committee on Data concerns the retention of raw data: when is this necessary and how it should be made public [13]. Connected to this are questions surrounding the value of retaining this data and whether the costs of doing so are justified. As these discussions develop, they are likely to lead to recommendations that are pertinent to a Physical Sciences Data Infrastructure.

5.3 Regional Initiatives

5.3.1 NFDI (Germany)

NFDI [14] is a significant long-term investment by the German government to fund a National Data Infrastructure for Germany (up to 90 million Euros per year over 9 years[a]). It covers all areas of science and has established around 20 consortia [15] including a number relevant to the physical sciences indicated in Table 3.

Domain	Consortium	URL
Catalysis-Related Sciences	NFDI4Cat	https://nfdi4cat.org/
Chemistry	NFDI4Chem	https://www.nfdi4chem.de/
Engineering	NFDI4Ing	https://nfdi4ing.de/
Data from Photon and Neutron Experiments	DAPHNE4NFDI	https://www.daphne4nfdi.de/
Condensed-Matter Physics and the Chemical Physics of Solids	FAIRmat	https://www.fair-di.eu/fairmat/
Data Science and Artificial Intelligence	NFDI4DataScience	https://www.nfdi4datascience.de/
Earth Sciences	NFDI4Earth	https://www.nfdi4earth.de/
Materials Science & Engineering	NFDI-MatWerk	https://nfdi-matwerk.de/
Particles, Universe, Nuclei and Hadrons	PUNCH4NFDI	https://www.punch4nfdi.de/

Table 7: NFDI Consortia relevant to the physical sciences.

Of most relevance to structural data in the chemical sciences will be the activities of NFDI4Chem [16] which include:

- Identifying the need for new research data repositories, establishing these and linking them to international repositories.
- Minimum information standards for data and machine-readable metadata, open data standards, in order to support the FAIR principles for research data.
- Fostering use of Electronic Laboratory Notebooks, tools and APIs between instrumentation and software to establish an embedded, digital information architecture.

NFDI4Chem has expressed an openness to connecting to international activities and collaboration with this NFDI consortium is certainly advised.

5.3.2 European Open Science Cloud

We highlight here the findings of a report on the readiness levels of various disciplines for engagement in the European Open Science Cloud undertaken by the RDA as part of

[a] See article 8 of https://www.gwk-bonn.de/fileadmin/Redaktion/Dokumente/Papers/NFDI.pdf. See also https://www.rwth-aachen.de/cms/root/Forschung/Forschungsdatenmanagement/~hdlbr/Aufbau-einer-nationalen-Forschungsdateni/lidx/1/.

RDA4EOSC [17]. A mapping carried out for this study identified the Materials and Chemical Sciences as being under-represented within the EOSC [18,19].

On Materials Science, the report noted:

> Developments are fragmented due to the funding mostly national or regional and traditional or discipline specific gap between industrial and academic interests and agendas. The EOSC offer the opportunity of addressing these issues while leveraging structures and work like the FAIR-DI in Germany, the European Commission Joint Research Center materials database (MatDB) and connecting that to international initiatives like the Materials Genome Initiative (MGI) and MaterialsCommons.org, and the Materials Research Data Alliance.

On Chemical Sciences, it noted:

> “There is a demonstrable need for coordinated development of updated and scaled infrastructures, hard and soft, for enabling chemical data exchange and connecting data providers with data users across sources and applications.” The EOSC aims to provide a framework for such a coordinated approach and support disciplines in strengthening their value propositions across the research landscape, as it builds on cross community interoperability and federation efforts.

We would recommend understanding how the EOSC initiative can help connect a UK Physical Sciences Data Infrastructure to European and other research infrastructures.

5.3.3 FAIRsFAIR

The European-funded FAIRsFAIR project [20] set out to foster FAIR data practices in Europe by establishing an overall knowledge infrastructure on academic quality data management, procedures, standards, metrics and related matters, based on the FAIR principles. It has produced recommendations and guidance on a range of matters relating to FAIR data and services covering data practices, data policy, certification, training and education. The outputs of this initiative will undoubtedly be useful input to the planning of a national physical sciences data infrastructure.

5.4 Global Initiatives

5.4.1 GO FAIR

GO FAIR [21] is a bottom-up stakeholder-driven initiative that aims to implement the FAIR Data Principles specifically through development of the Internet of FAIR Data and Services. Initially aligned with EOSC, it now has a presence in other regions including the US. Activity within GO FAIR is undertaken by Implementation Networks [22] of which there are instances relating to Nanomaterials, Novel Materials Discovery, Materials Science and Chemistry [23]. Activity within GO FAIR has led to a proposal for a FAIR Digital Object for Molecular Structure (or a FAIR Molecule) [24] which has yet to see adoption beyond a limited proof of concept.

5.4.2 CODATA

The Committee on Data of the International Science Council [25] promotes international collaboration to advance Open Science and improve the availability and usability of data for all disciplines. It works across the Scientific Unions and partners with other global research data initiatives. Current task groups include one on the digital representation of units of measurement (DRUM) [26] and another looking to extend InChI for nanomaterials [27]. This latter task group continues a theme of past activities around nanomaterials which resulted in a Uniform Description System for Materials on the Nanoscale [28].

5.4.3 Research Data Alliance

The Research Data Alliance (RDA) [29] has a mission to build the social and technical bridges needed to enable the open sharing and reuse of data. It does this through Working and Interest groups that attract experts in infrastructure, research data management and publication from across sectors and domains. Outputs from RDA working groups and current areas of focus have been referred to throughout this report. The RDA offers an ideal place for regional physical sciences data infrastructures to meet and share experiences and challenges and draw on best practices from across the wider research communities.

5.5 Industry Initiatives

5.5.1 Pistoia Alliance

The mission of the Pistoia Alliance [26] is to lower the barriers to innovation in Life Sciences R&D through pre-competitive collaboration. It benefits from support from industry organisations who contribute to prioritisation and funding.

Projects of relevance to data infrastructure that the Pistoia Alliance has sponsored include the Unified Data Model and HELM, both discussed in Section 2. We also note the Pistoia Alliance Chemical Safety Library [30], a crowd-sourced database of hazardous reactions shared with the community. This has been enhanced and is now hosted by CAS, a Division of the American Chemical Society [31].

Pistoia has communities of practice focused on FAIR implementation, AI and Machine Learning, Data Governance, Lab of the Future, and User Experience Design, all of which will have some relevance to physical sciences data infrastructure.

5.5.2 FAIRplus

FAIRplus [32] is an initiative led by Janssen and Elixir that aims to make data from Innovative Medicine Initiative (IMI) projects and industry partners available in a FAIR manner. It is doing this by establishing guidelines and a maturity assessment framework. A question to consider is what similar initiatives with industry it might be appropriate for a Physical Sciences Data Infrastructure to align with.

5.5.3 European Materials Modelling Council

The European Materials Modelling Council (EMMC) [33] was initially an EC funded project that now runs as a non-profit association and brings together modelling experts from industry and academia to support the development of digital strategies in materials discovery and manufacture. It has focus areas on Model Development, Interoperability, Digitalisation, Software and Industry Impact all of which may be relevant to a Physical Sciences Data Infrastructure.

5.6 Industry Engagement

The CCDC and Industry Partnership

The CCDC has benefited greatly through partnership with and guidance from industry over the years, leading to the development and practical application of Solid-form Informatics [34] and Particle Informatics [35]. Development of the CSD-Theory suite used by Case Study 2 of the PSDI Pilot was primarily sponsored by a partnership between industry and the CCDC.

More generally, industry has contributed funds for value-added software and services that draw on CCDC's science and expertise, typically providing 60-70% of the revenue required to operate the centre.

As well as engaging with industry through existing networks and initiatives, a UK Physical Sciences Data Infrastructure should consider how it can best engage with industry to help shape priorities that will deliver value for the UK industrial sector and economy. This could be by providing and helping to prioritise use cases as well as contributing to the funding required to realise these in an open and sustainable way.

5.7 References

1. "International Union of Pure and Applied Chemistry", *IUPAC | International Union of Pure and Applied Chemistry*, https://iupac.org/ (accessed 14 April 2022).

2. J. G. Grasselli (1991) "JCAMP-DX, a standard format for exchange of infrared spectra in computer readable form (Recommendations 1991)", *Pure and Applied Chemistry*, 63(12), https://doi.org/10.1351/pac199163121781.

3. S. Heller, A. McNaught, S. Stein, D. Tchekhovskoi, and I. Pletnev (2013) "InChI - the worldwide chemical structure identifier standard", *Journal of Cheminformatics*, 5(1), https://doi.org/10.1186/1758-2946-5-7.

4. V. Gold, Ed. "*The IUPAC Compendium of Chemical Terminology*". Research Triangle Park, NC: International Union of Pure and Applied Chemistry (IUPAC). https://doi.org/10.1351/goldbook.

5. "Welcome - BIPM", https://www.bipm.org/en/home (accessed 14 April 2022).

6. J. G. Frey (2014) "Digital IUPAC", 36(1), https://doi.org/10.1515/ci.2014.36.1.14.

7. I. Bruno, S. Coles, W. Koch, L. McEwen, F. Meyers, and S. Stall (2021) "FAIR and Open Data in Science: The Opportunity for IUPAC", *Chemistry International*, 43(3), https://doi.org/10.1515/ci-2021-0304.

8. "InChI Trust – InChI: open-source chemical structure representation algorithm", https://www.inchi-trust.org/ (accessed 14 April 2022).

9. "InChI Working Groups – InChI Trust", https://www.inchi-trust.org/inchi-working-groups/ (accessed 21 April 2022).

10. I. Lynch *et al.* (2020) "Can an InChI for Nano Address the Need for a Simplified Representation of Complex Nanomaterials across Experimental and Nanoinformatics Studies?", *Nanomaterials*, 10(12), https://doi.org/10.3390/nano10122493.

11. H. Moseley, R. Salek, M. Arita, E. Schymanski, and P. Rocca-Serra (2018) "Inchl Isotopologue and Isotopomer Proposal". figshare. https://doi.org/10.6084/M9.FIGSHARE.7150964.V1.

12. "International Union of Crystallography", https://www.iucr.org/ (accessed 14 April 2022).

13. S. Coles and A. Sarjeant (2021) "IUCr workshop on ‘When should small molecule crystallographers publish raw diffraction data?’", *ACA RefleXions*, 4.

14. "nfdi | Nationale Forschungsdateninfrastruktur e. V.", https://www.nfdi.de/ (accessed 14 April 2022).

15. "Konsortien | nfdi", https://www.nfdi.de/konsortien/ (accessed 14 April 2022).

16. "Chemistry Consortium in the National Research Data Infrastructure". Available: https://nfdi4chem.de/.

17. "Launching RDA4EOSC: a new initiative from the Research Data Alliance supporting the European Open Science Cloud", *RDA*, https://www.rd-alliance.org/launching-rda4eosc-new-initiative-research-data-alliance-supporting-european-open-science-cloud (accessed 14 April 2022).

18. "RDA4EOSC: Supporting the engagement of disciplinary research communities with the European Open Science Cloud", *RDA*, https://www.rd-alliance.org/rda4eosc-supporting-engagement-disciplinary-research-communities-european-open-science-cloud (accessed 14 April 2022).

19. "RDA4EOSC: Report on Disciplinary engagement in EOSC", *Google Docs*, https://docs.google.com/document/d/1k27DNgoxqcrjzGh5IVweAUHMKKYnF-N5XZIEqWdwUbI/edit (accessed 14 April 2022).

20. "FAIRsFAIR", https://www.fairsfair.eu/ (accessed 14 April 2022).

21. "GO FAIR initiative: Make your data & services FAIR". Available: https://www.go-fair.org/.

22. "Current Implementation Networks", *GO FAIR*, https://www.go-fair.org/implementation-networks/overview/ (accessed 14 April 2022).

23. S. J. Coles, J. G. Frey, E. L. Willighagen, and S. J. Chalk (2020) "Taking FAIR on the ChIN: The Chemistry Implementation Network", *Data Intelligence*, 2(1–2), https://doi.org/10.1162/dint_a_00035.

24. E. A. Schultes *et al.* (2019) "FAIR Molecule Concept", https://doi.org/10.17605/OSF.IO/B7JWG.

25. "CODATA, The Committee on Data for Science and Technology", https://codata.org/ (accessed 14 April 2022).

26. "Digital Representation of Units of Measurement (DRUM)", *CODATA, The Committee on Data for Science and Technology*, https://codata.org/initiatives/task-groups/drum/ (accessed 14 April 2022).

27. "Extension of InChI for Nanomaterials", *CODATA, The Committee on Data for Science and Technology*, https://codata.org/initiatives/task-groups/extension-of-inchi-for-nanomaterials/ (accessed 14 April 2022).

28. "Uniform Description System for Materials on the Nanoscale (v.1.0) Published", *CODATA, The Committee on Data for Science and Technology*,

https://codata.org/uniform-description-system-for-materials-on-the-nanoscale-v-1-0-published/ (accessed 14 April 2022).

29. "RDA | Research Data Sharing without barriers", https://www.rd-alliance.org/ (accessed 14 April 2022).

30. "Chemical Safety Library", https://safescience.cas.org/ (accessed 14 April 2022).

31. "Launch of Pistoia Alliance Chemical Safety Library powered by CAS platform facilitates information sharing between scientists to improve laboratory safety", *CAS*, https://www.cas.org/resources/press-releases/chemical-safety-library (accessed 14 April 2022).

32. "FAIRplus | Home page", https://fairplus-project.eu/ (accessed 14 April 2022).

33. "The European Materials Modelling Council", *The European Materials Modelling Council*, https://emmc.info/ (accessed 14 April 2022).

34. N. Feeder *et al.* (2015) "The integration of solid-form informatics into solid-form selection", *Journal of Pharmacy and Pharmacology*, 67(6), https://doi.org/10.1111/jphp.12394.

35. M. J. Bryant *et al.* (2019) "'Particle Informatics': Advancing Our Understanding of Particle Properties through Digital Design", *Crystal Growth & Design*, 19(9), https://doi.org/10.1021/acs.cgd.9b00654.

6 Recommendations for a UK Physical Sciences Data Infrastructure

6.1 Key Messages

This report has aimed to provide an overview of:

- The standards and solutions available to support the representation of molecular structure and materials, their strengths and current limitations.
- Infrastructure considerations relating to the management, discovery and reuse of structures and associated properties.
- The importance of access to high quality structural data and the curation effort required to achieve this.
- Opportunities for global collaboration to establish effective solutions for the management of structures and their associated data.

The key messages we distil out of this are captured in Table 1. A compilation of more specific recommendations highlighted throughout this report is provided in following sections.

<table>
<tr><th>Embrace standards for structure representation</th><th>Support the discovery and interoperability of structures and associated data</th></tr>
<tr><td>➢ Support a range of digital structure representations recognising strengths and weaknesses.
➢ Capture the provenance of structures and their digital representations.</td><td>➢ Consider a digital structure representation to be required metadata whenever appropriate.
➢ Adopt standard structure identifiers to link between resources and aid discoverability.</td></tr>
<tr><th colspan="2">Advocate for and enable access to curated and trusted structural data</th></tr>
<tr><td>➢ Establish structure-based API access to commercial and public structural data resources.
➢ Invest in cleaning up data in public structural data resources.</td><td>➢ Invest in expertise to support data curation and development of guidance and standards.
➢ Establish physical sciences criteria for categorising resources as trusted and datasets as FAIR.</td></tr>
<tr><th colspan="2">Invest in and support change in partnership with global communities</th></tr>
<tr><td colspan="2">➢ Partner with organisations, initiatives and solution providers across regions and domains.
➢ Cultivate relationships with industry to guide priorities and provide support for sustainability.</td></tr>
</table>

Table 8: Summary of key messages regarding the role of structure in physical sciences data management.

6.2 Detailed Recommendations

6.2.1 Embrace Standards for Structure Representation

Consistent and reliable representation of structure is a critical enabler of data reuse and discovery throughout the research data lifecycle, across the physical sciences and beyond. Structure representation should thus be a core consideration in the design of a Physical Sciences Data Infrastructure.

Current research practices tend towards structure representations that are not conducive to interpretation by machines and use in data-driven science. A Physical Sciences Data Infrastructure should enable and encourage researchers to provide representations of structure that can be readily interpreted by machines.

A plurality of structure representations should be supported and encouraged. Traditional representations of structure (diagrams, names) are helpful but should ideally be accompanied by more machine-readable representations.

Existing machine-readable representations work well for organic structures but even then, may be prone to ambiguity of interpretation. Consideration should be given to the technical and scientific limitations of current structure representation methodologies.

Structure-based classifications that enable semantic interoperability and faceted search should be supported. Identifiers such as registry and database accession IDs are also important for enabling links between structural data resources.

Wherever possible, a standard International Chemical Identifier (InChI) should be generated and stored for all structures as a key enabler of findability and interoperability. This does not necessarily mean requiring researchers to provide InChIs but does require a representation from which an InChI can be reliably generated.

IUPAC guidelines for the depiction of 2D chemical diagrams should be followed where appropriate.

Consideration should be given to the representation and storage of multi-component systems including reactions and mixtures and not just individual molecular structures.

Accommodation must be made for the handling of 3D structures that have been determined computationally or experimentally. The provenance of a 3D structure – whether experimental or computational – and the methods used to generate it should be clearly indicated. For aspects of the infrastructure that involve human interaction, the ability to easily visualise key features of 3D structure should be provided.

The software packages involved in the generation of structure representations and models should be captured in line with community recommendations for software citation.

There is significant intersection between the physical sciences and biological sciences, particularly when it comes to understanding biological structure and mechanisms. A Physical Sciences Data Infrastructure should consider storage of biological macromolecular structures and/or links to biological resources based on structure. Partnership with relevant bioinformatics organisations is advised.

Recognise that change takes time and that a future infrastructure may have to support legacy representation formats for longer than might be desirable. Be prepared to invest in tools and education that will enable communities to embrace change without fear of disruption.

Insofar as possible, separate out services, tools and workflows from underlying formats to enable these to evolve independently. Provide opportunity for ongoing investment and experimentation with new paradigms and approaches for structure representation.

6.2.2 Support the Discovery and Interoperability of Structures and Associated Data

A Physical Sciences Data Infrastructure should facilitate the future reuse of structural data by providing technical enablers that support the discovery and interoperability of relevant data and metadata across resources.

Recognise the digital representation of the structure of a substance studied or a material modelled as an essential piece of metadata that should be stored and tracked alongside physical sciences datasets.

Contribute to the development of a community chemical structure validation service to support and encourage best practice and enable assessment of the reliability of a digital structure representation. Help to develop benchmarking reference sets that can be used to judge compliance of tools with structure representation standards.

If sophisticated storage, search and analysis and management of structural data is required as part of the infrastructure, partner with organisation(s) who have developed solutions to satisfy these needs.

Encourage and enable the registration of metadata in open registries, including appropriate structure identifiers and links to related objects in order to contribute to wider networks of open science knowledge graphs.

Connect structure representation to sample identification, taking advantage of standard identifiers for samples and structures.

Provide services that enable the retrieval of data from across resources based on structure identity, composition, and connectivity, partnering with existing solution-providers where possible. In particular expose interfaces that enable the lookup and linking of data based on standard identifiers such as InChI.

Support the ability to faithfully exchange individual datasets and aggregated subsets of structure-based data and metadata between systems within and external to a future infrastructure. Enable citation of datasets and aggregated subsets to support reproducibility, provenance and credit.

6.2.3 Advocate For and Enable Access to Curated and Trusted Structural Data

A Physical Sciences Data Infrastructure should promote the importance of high-quality curated data, enable access to existing sources of curated structural data in support of UK academic research, and cultivate expertise and criteria that can encourage the increased availability of FAIR and trusted structural data.

Consideration should be given to how a future infrastructure can provide access to high quality curated structural data that is available in third party resources as well as within the infrastructure itself.

Building on the tradition of EPSRC support for access to highly curated data resources through the Physical Sciences Data-science Service and its predecessors, explore how this can be extended to incorporate increased access to richly curated sources of structural data and properties, crucially ensuring access via machine APIs.

Identify opportunities to invest in making data in public resources more valuable and available in more structured forms through richer APIs. Also consider APIs that enable federated access to data across public and proprietary resources based on common languages of structure representation.

Champion the importance of and requirements for high quality data publishing and curation in the physical sciences. Adopt existing curation guidelines where available – work with wider communities to develop new ones where these are not available.

Promote the importance of machine interpretable data formats to support efficient publication workflows and enable data reuse. Advocate against practices that result in non-semantic publication of data that cannot be reliably interpreted by machines.

Consider how a future infrastructure can provide the motivation and means for publication of structural data associated with doctoral theses and dissertations in machine-accessible and reusable forms – not just to ensure data are available for future research, but also to train the next generation of researchers in best practices for data management and publication.

Identify and highlight gaps in domain-specific data storage and curation and cultivate communities to address these. Consider providing a publication platform for physical sciences data that does not have a natural domain-specific home.

Become a centre of expertise and best practice in research data management and curation in the physical sciences alongside provision of any technical infrastructure.

Recognise that investment in data infrastructure requires investment in expertise as well as technology. Invest for the long term and ideally to enable all data and services to be openly available.

Learn from the experiences of existing data repositories when considering benefits of investment in data storage and curation. Consider commissioning case studies that demonstrate the return on investment of established physical sciences data resources for the wider economy. Consider where there may be opportunity for national and international cooperation to pool resources and minimise costs.

Recognise that if data storage and curation activities must be self-sustaining then some restrictions or barriers are likely. Adopt the Principles of Open Science Infrastructure to guide the governance and sustainability of data infrastructure.

Identify criteria for characterising a physical sciences data repository as trusted and a dataset stored within that as FAIR from a domain perspective. Base this on existing frameworks for establishing trustworthiness and assessing FAIR maturity.

6.2.4 Invest In and Support Change in Partnership with Global Communities

A Physical Sciences Data Infrastructure should partner with and invest in international initiatives that aim to develop the standards, infrastructure and guidelines needed to advance the management of structural data specifically and research data generally.

Align with other organisations and initiatives looking to address data representation, publication and management challenges relevant to the physical sciences. Be willing to contribute time to efforts aimed at the development of shared infrastructure and standards.

Cultivate partnerships with industry to inform priorities and identify funding opportunities for development of a Physical Sciences Data Infrastructure that will deliver value for industrial and research sectors in the UK.