Preprint CheMIn Technical Reports ISSN 3061-1075

Computation-ready Experimental Polyoxometalate Structural Dataset

Aleksandar Kondinski [1]

released: May 25, 2026

[1] Institute of Physical and Theoretical Chemistry
Graz University of Technology
Stremayrgasse 9, 8010 Graz, Austria

Preprint No. 2

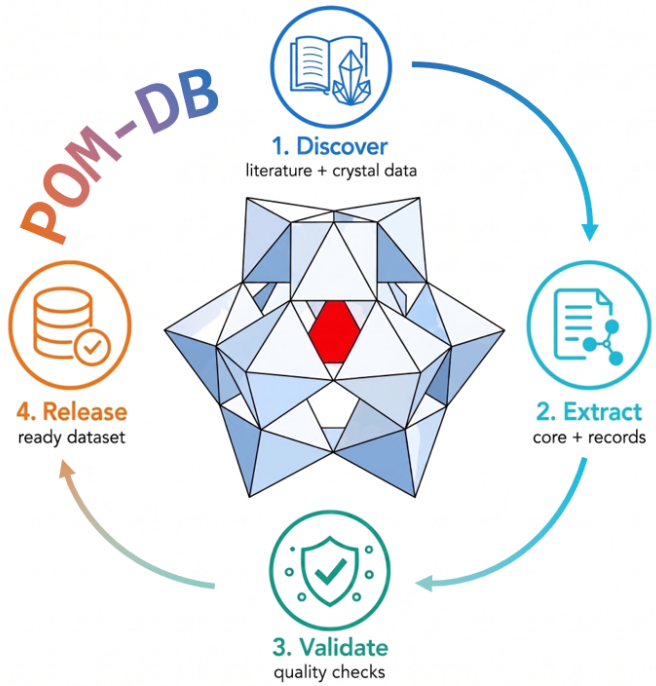
Keywords: Polyoxometalates, structural data, digital chemistry, cheminformatics, FAIR data

Edited by
Chemical Modelling and Informatics Group
Institute of Physical and Theoretical Chemistry
Graz University of Technology
Stremayrgasse 9, 8010 Graz
Austria (Europe)

E-Mail: kondinski@tugraz.at
World Wide Web: http://chemin.at/

Abstract

Polyoxometalates are metal–oxo clusters whose charges, counter-ion environments, structural forms, and reported material compositions are difficult to represent consistently in digital form. This Data Descriptor presents POM-DB, a structure-aware and computation-ready dataset that links POM-level formulae with elemental compositions, molecular masses, charges, classification labels, structural identifiers, material records, CAS Registry Numbers, literature provenance, released CML structure files, and Weisfeiler–Lehman graph-kernel topology annotations. The dataset is provided as a hierarchical JSON release with associated CML files, a schema, validation code, and an exploratory web interface. It is intended for structure inspection, format conversion, descriptor generation, structural classification, and data-driven reuse. Technical validation confirms schema conformity, formula parsing, mass recalculation, charge completeness, identifier formatting, CML linkage, and material-level provenance coverage.



Highlights

- POM-DB contains 1936 curated POM-level structural records.
- Over 6300 material records capture salts, hydrates, and counter-ions.
- All 1936 POM records link to released CML structure files.
- Topology annotations are provided for all 1936 released records.

Contents

1 Background & Summary

Polyoxometalates are discrete metal–oxo clusters formed mainly from early transition metals such as molybdenum, tungsten, and vanadium.[23,24] Their chemistry spans a wide range of nuclearities, compositions, charges, heteroatom substitutions, organic components, counter-ion environments, hydration states, and redox states.[8,17] This diversity underpins applications in catalysis, energy-related chemistry, molecular electronics, and biological or biomedical contexts.[1,16] It also makes POM chemistry difficult to represent consistently in digital form. Recent work on data-driven POM chemistry has emphasised that progress in this area requires curated records, structure-aware representations, and reusable digital resources for chemical reasoning and discovery.[14]

In neighbouring areas of materials chemistry, curated computation-ready datasets have become enabling infrastructure for high-throughput modelling and data-driven discovery. For metal–organic frameworks, the CoRE MOF database converted experimentally derived MOF structures into a form suitable for molecular simulations, initially reporting more than 4700 porous structures and later expanding to more than 14000 structures in CoRE MOF 2019.[3,4] For covalent organic frameworks, the CURATED COF workflow provided cleaned, uniform, and reproducibly refined COF structures, including literature-derived and DFT-optimised datasets for downstream adsorption modelling.[20] Related curated resources have also been reported for zeolite and open-framework materials, and for specialised CSD subsets targeting MOF structures.[18,28,29] These examples show that structure curation, format harmonisation, and provenance tracking are prerequisites for reliable computational reuse.

Comparable dataset development in POM chemistry is more challenging because POM records often combine cluster-level entities, full crystalline materials, counter-ions, solvent or water content, charge states, and literature provenance in a single reported composition. A publication may describe a POM anion, a full crystalline material, a hydrated salt, a counter-ion-containing composition, a CAS-indexed material, and one or more coordinate files. These layers are connected, but they are not identical. The difficulty is further amplified by the information density of inorganic crystal structures and by long-standing challenges in POM formula designation and nomenclature.[11,15] Standard chemical identifiers such as InChI and InChIKey are valuable for chemical indexing, but inorganic and coordination compounds can present representation challenges that are not always encountered for small organic molecules.[2,9] For POMs, formula, charge, bonding, protonation, metal substitution, counter-ion environment, and coordinate representation may all be relevant to interpretation. Without an explicit data structure linking these levels, formula search, charge comparison, molecular-mass calculation, provenance tracking, structural comparison, and computational reuse become difficult.

A previous Data Descriptor reported a formula-centred polyoxometalate dataset compiled from literature and database searches.[13] Such formula-level curation is useful for inspecting composition, charge, mass, and broad chemical coverage, but it cannot by itself support structure-level reuse, topology analysis, or coordinate-based descriptor generation. POM-DB addresses this gap by linking POM-level molecular entities to CAS-indexed material occurrences, DOI-linked literature provenance, crystallographic source evidence, topology annotations, and released CML structure files derived from validated crystal-

lographic records (Fig. 1). The current release contains 1936 POM-level records, 6329 nested material records across a wide range of salt forms, hydration states, and counter-ion variants, linked released CML files for all 1936 POM-level records, and deposited geometry and symmetry annotations for all released records. Aggregate structural statistics were computed on a coordinate-bearing working subset. The dataset is intended to support reuse in computational screening, formula parsing, data-driven modelling, cheminformatics, structural classification, and provenance-aware inorganic chemistry applications. Detailed compositional, charge, mass, nuclearity, and structural statistics are reported in the Data Records section.

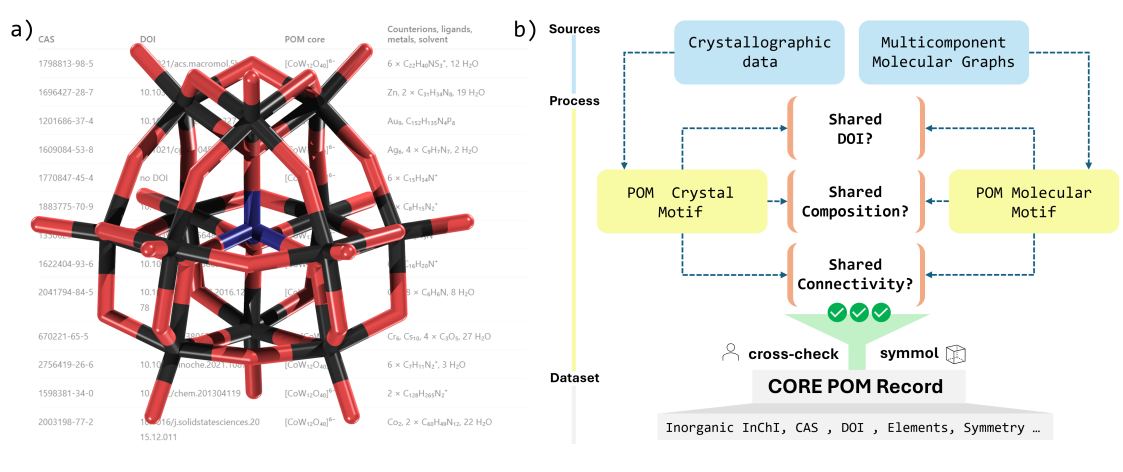


Figure 1: ***Dataset construction and record linking.*** ***a****, POM-level clusters are linked to reported material records, including CAS-indexed entries, DOI information, formulae, counter-ions, ligands, metals, and solvent components.* ***b****, Crystallographic motifs and molecular graph records are matched using DOI, composition, and connectivity checks to produce POM-DB records with identifiers, provenance, composition, annotations, and released CML structure-file links.*

2 Methods

Input data and discovery workflow

Candidate POM-containing records were identified through CAS SciFinder searches using compound-class queries targeting metal–oxo cluster compositions with molybdenum, tungsten, vanadium, niobium, or tantalum as addenda elements.[5,21] For each entry, the POM-level component was extracted from the full material composition, and an inorganic InChI string was generated to encode its molecular graph and charge.[2] Cross-comparison of these InChI strings reduced the candidate pool to a set of unique POM entities. The molecular graph of each unique POM was then matched against graphs generated from crystallographic entries. A successful match required agreement in element composition, graph connectivity, and shared Digital Object Identifier. Records passing this multi-constraint matching were subject to manual verification.

The curation workflow combined SciFinder-derived molecular records with crystallographic structure evidence in a multi-stage pipeline. Raw SciFinder RDF exports were merged, parsed into structured records, and separated into individual MOL files. Free-text bibliographic entries were normalised to DOI-linked citations using Crossref queries. POM products were deduplicated using reconnected inorganic InChI representations, assigned stable identifiers, and parsed to extract formulae, charges, bonding information, and coordination descriptors. Crystallographic structures were retrieved from the Crystallography Open Database using DOI-based lookup, from the Cambridge Structural Database using refcode-based retrieval where required, and from DOI-linked online crystallographic sources where available.[6,7] POM clusters were extracted from CIF files as isolated XYZ working structures using graph traversal and distance-based bonding rules. SciFinder MOL graphs were then matched to these coordinate-bearing working structures using composition checks, graph matching, and geometry-aware alignment. Confirmed matches were subsequently converted into the released CML-linked hierarchical JSON index. The full script-level workflow, including file names, intermediate artefacts, and matching criteria, is documented in the deposited workflow notes.

Coordinate-bearing structures used for symmetry analysis were processed through the SYMMOL program to automatically determine the highest recognisable point-group symmetry for each cluster geometry.[22] Records were included when they contained a polyoxometalate or polyoxometalate-related metal–oxo cluster entry with sufficient information to assign a POM-level formula and at least partial provenance. Records were excluded or flagged when the formula could not be parsed, the composition was not sufficiently resolved, the structure file could not be associated with a POM-level entry, or the record could not be linked to meaningful provenance.

Record model

The dataset uses a hierarchical record model. The top level is the POM entry, identified by a stable record identifier such as `POM_X1000`. Each POM entry contains formula-level, structure-level, topology-level, and identifier metadata. The nested `pom_material` object is a keyed collection of one or more material records associated with the POM entry. This structure allows one POM-level composition to be linked to reported material records, for example different salts, hydration states, counter-ion combinations, solvent-containing forms, or repeated database records. At the POM level, the dataset stores a typographic display formula, a plain formula string for computational processing, element counts, molecular mass, charge, classification labels, InChI, InChIKey, released CML structure-file references, and a `topology_clustering` block. Topology grouping is based on a Weisfeiler–Lehman graph-kernel representation, which converts labelled molecular graphs into graph features suitable for exact grouping and low-dimensional comparison.[25] The `topology_clustering` block stores the WL hash, exact cluster assignment, cluster size, and PCA embedding coordinates. At the material level, the dataset stores reported material formulae, product formulae, product element counts, product molecular weights, CAS Registry Numbers, literature citations, DOI information where available, and a primary `material_reference` metadata block. The DOI is an important part of the material-level record because it provides a direct link to the original

literature supporting the reported material and allows users to trace a POM-level entry to the publication context in which the material was described. Released POM-level records always include formula, composition, charge, identifiers, CML references, `geometry`, `symmetry`, and a `topology_clustering` block.

Formula processing was designed to retain both human-readable and machine-readable representations. The field `pom_formula_display` stores a display formula with typographic subscripts, superscripts, and charge notation. The field `pom_formula_plain` stores a plain formula string suitable for parsing and downstream processing. Element counts were extracted from the plain formula representation and stored as dictionaries. Molecular masses were calculated from the parsed element-count representation. Product-level element counts and molecular weights were calculated separately where material-level formula information was available, because reported materials may include counter-ions, waters of crystallisation, solvent molecules, or organic components in addition to the POM unit. Net charge was stored at the POM level where available or assigned during curation. Charge values are represented numerically to support filtering, comparison, and computational workflows. Labels were assigned to describe broad POM classes, such as `polyoxotungstate`, `polyoxomolybdate`, and `polyoxovanadate`. These labels are intended as practical classification aids.

Where available, POM-level records include InChI and InChIKey strings. These identifiers were generated using the IUPAC InChI reference implementation.[10] The generated identifiers were used to support comparison, matching, and retrieval of POM-level entities. The InChI and InChIKey values are retained as calculated identifiers, and they may be recalculated and updated in future releases as the inorganic InChI implementation develops further. These identifiers should be interpreted with care for inorganic cluster chemistry, where formula, charge, bonding, counter-ion environment, and structural context may not always be captured with the same reliability as in small organic molecules.[2,9] For this reason, the dataset treats InChI and InChIKey as useful retrieval aids. The dataset also includes released CML structure files. These structure files support downstream workflows such as structure inspection, conversion, descriptor generation, topology analysis, and preparation for computational chemistry.

The primary dataset is exported as a hierarchical JSON file that preserves the nested relation between POM-level and material-level records. The JSON representation is the canonical released record structure. A flat relational representation is used internally to support the web interface and summary analyses, but the archived JSON repository remains the authoritative versioned source of record for citation, download, and reproducible reuse. The dataset is organised to support findable, accessible, interoperable, and reusable data practices.[27] The JSON structure is also compatible with later mapping to richer markup or ontology-based representations, including Chemical Markup Language and chemistry ontology frameworks.[19,26] For easier exploration, selected records can also be searched through the Digital Chemistry POM Explorer. The web interface supports interactive browsing and filtering by POM identifier, formula, label, InChI/InChIKey, molecular mass, charge, material formula, product formula, CAS Registry Number, DOI, citation, and elemental composition.

By "computation-ready" we mean that the deposited records and released CML structure files are suitable for coordinate inspection, molecular descriptor generation, chemical

format conversion, structural classification, and pipeline input without further manual processing of the data format. The term does not imply that all structures are automatically validated for quantum-chemical or force-field simulations, which may require additional expert preparation steps such as proton placement, charge assignment, counter-ion selection, or geometry optimisation. Users intending to run electronic-structure or molecular-dynamics calculations should verify the suitability of individual records for their specific use case.

Data Records

The deposited dataset contains a computation-ready index of polyoxometalate records and associated released CML structure files. Each POM-level record links formula, composition, molecular mass, charge, classification labels, InChI, InChIKey, material occurrences, CAS Registry Numbers, DOI-linked literature provenance where available, topology annotations, and references to associated CML files. The current release contains 1936 POM-level records, 6329 nested material records in the released JSON, linked released CML files for all 1936 POM-level records, and `topology_clustering` annotations for all released records. Geometry and symmetry annotations are present for matched subsets of records. A processed subset of 1616 coordinate-bearing working structures was used for the aggregate structural summaries shown in Fig. 2**e–g**. This working subset should therefore be interpreted as a source for descriptive structural statistics rather than as a separate count of deposited JSON annotations.

The dataset is deposited on Zenodo.[12] The repository contains the curated hierarchical JSON file, the released CML files, the JSON schema, validation code, topology-analysis code, dependency information, code and data licences, and a README file describing the deposited resources. The JSON file is the canonical data record. The CML files are the released coordinate representation linked from the JSON and provide structures for inspection, conversion, descriptor generation, and computational preparation. Internal XYZ working structures were used in parts of the processing workflow and in some aggregate geometry and symmetry analyses, but they are not the primary released structure-file representation.

The released records show the expected chemical breadth of a POM-focused structural dataset. Polyoxomolybdates and polyoxotungstates form the largest families, followed by polyoxovanadates, with smaller numbers of polyoxoniobates and polyoxotantalates (Fig. 2**a**). The element distribution reflects both the inorganic POM cores and the reported material environments around them. Oxygen is present in all 1936 records, while carbon and nitrogen occur in a large fraction of records, consistent with the frequent presence of organic ligands, organic counter-ions, and other organic components in reported POM materials (Fig. 2**b**). The dataset also spans a wide charge and mass range, with charges from -40 to $+4$ and molecular masses from 265 to 11486 Da. Overall, the release contains 249 neutral species, 1659 anionic species, and 28 cationic entries.

The metal–oxygen archetype plot gives a compact view of the structural range covered by the coordinate-bearing subset (Fig. 2**c**). It maps structures by addenda-metal count and oxygen count, with the colour scale indicating the number of metal–oxygen edges.

The plot shows that the dataset covers many different core sizes. As expected, larger POM cores generally contain more oxygen atoms and more metal–oxygen connections, although the spread within the plot also shows that nuclearity alone is not sufficient to describe the structural diversity. Nuclearity (n_M), defined here as the number of addenda metal atoms per cluster, ranges from 3 to 54. Hexametalates ($n_M = 6$) are the most frequent group with 497 entries, followed by dodecametalates ($n_M = 12$, 252 entries) and octadecametalates ($n_M = 18$, 200 entries). The processed coordinate-bearing subset corresponds to 566 distinct structural archetypes, defined here by unique addenda-metal and oxygen count combinations, with the number of metal–oxygen edges used as an additional descriptor of connectivity density.

The central topology projection provides a complementary graph-based view of the released POM records (Fig. 2**d**). The WL-derived representation organises records by connectivity pattern. The resulting point cloud forms a structured but partially overlapping landscape, which is chemically reasonable because topology is not determined only by addenda-metal identity. For example, POMs containing different addenda metals can still share related connectivity patterns, while records within the same broad family can occupy different regions if their graph structures differ. The representative exact topology clusters placed around the projection illustrate recurring connectivity motifs and their member counts. These examples illustrate how the WL representation can organise repeated structural patterns across the dataset without relying on manual assignment of traditional named archetypes.

The lower panels summarise geometry-derived properties for the 1616 coordinate-bearing working structures (Fig. 2**e–g**). Average metal coordination values are concentrated near six, consistent with the frequent occurrence of octahedrally coordinated addenda-metal environments in POM structures. The metal-to-oxygen ratio distribution is centred in the range expected for oxo-bridged metal–oxygen frameworks. The radius of gyration distribution provides a compact measure of overall cluster size and spatial extent, with larger values corresponding to more spatially extended clusters. Together, these distributions describe the local coordination environment, metal–oxygen composition, and size range of the coordinate-bearing working subset, and should be read as descriptive structural statistics.

Repository structure

The repository is organised as a compact release package containing the canonical dataset, schema, released structure files, and the small set of code files needed for validation and topology analysis, as summarised in Table 1.

Data fields

The principal fields stored in the deposited records are summarised in Table 2, covering the POM-level record, nested material-level entries, and the deposited topology annotations. Geometry and symmetry annotations are included in the released records, and exact machine-readable field names are defined in `schema.json`.

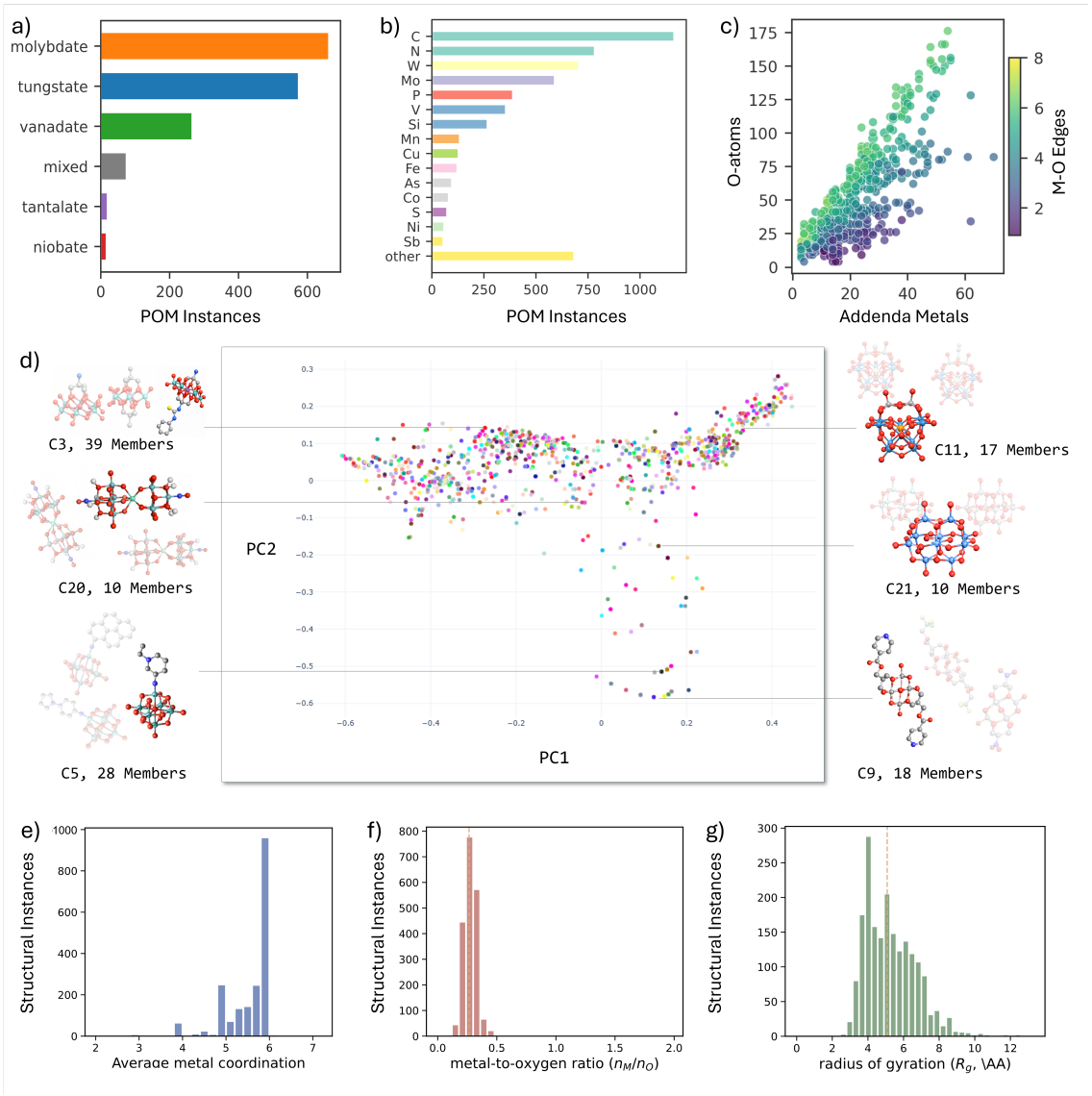


Figure 2: ***Structural, compositional, and topological overview of POM-DB.*** ***a****, Distribution of POM families.* ***b****, Element prevalence across the released records.* ***c****, Metal–oxygen archetype landscape, plotted by addenda-metal count and oxygen count, with colour indicating the number of metal–oxygen edges.* ***d****, Central WL-derived topology projection of the released POM records, with representative exact topology clusters shown around the embedding.* ***e****, Average metal coordination numbers for the 1616 coordinate-bearing working structures.* ***f****, Metal-to-oxygen ratios for the same working subset.* ***g****, Radius of gyration distribution for the same working subset.*

Table 1: ***Deposited files.***

File or folder	Format	Contents
`README.md`	Markdown	Repository guide, reuse notes, and example queries.
`pom-db.json`	JSON	Canonical hierarchical POM records
`schema.json`	JSON Schema	Field definitions and validation rules
`cml/`	CML	Released structure files referenced by the dataset records
`codes/`	Python / text	Validation script, dependency list, optional topology workflow, and code license
`LICENSE`, `codes/LICENSE`	Text	Dataset/documentation license and code license

Table 2: ***Principal deposited fields.***

Field	Description
Record ID	Stable identifier for each POM-level entry, for example `POM_X1000`.
Display formula	Formula shown with typographic subscripts, superscripts, and charge notation.
Plain formula	Plain formula string used for machine processing.
Element counts	Element-count dictionary for the POM-level composition.
Molecular mass	Molecular mass calculated for the POM-level composition.
Charge	Net charge assigned to the POM-level record.
Labels	Broad classification labels such as polyoxotungstate, polyoxomolybdate, or polyoxovanadate.
InChI	InChI string associated with the POM-level structure, where available.
InChIKey	InChIKey associated with the POM-level structure, where available.
Released CML files	Released CML structure files associated with the POM-level entry.
CCDC refcode	Refcode in the Cambridge Structural Database for entries retrieved from CSD.
CCDC reference	URL to the corresponding CSD record, where available.
COD reference	Identifier or URL for the corresponding Crystallography Open Database entry, where available.
Material records	Nested material-level records linked to the POM-level entry.
Material formula	Reported material formula including counter-ions, waters, solvent molecules, or organic components where present.
Product formula	Plain formula representation of the full reported material composition.
Product element counts	Element-count dictionary for the full reported material composition.
Product molecular weight	Molecular weight calculated for the full reported material composition.
CAS number	CAS Registry Number associated with the material record, where available.
Citations	Bibliographic provenance including citation text and DOI where available.
Material reference	Primary reference block containing title, author, citation, and DOI information.
Topology annotations – deposited for all released records	
WL hash	Weisfeiler–Lehman graph hash used for exact topology grouping after atom-type coarsening.
Cluster index	Integer index of the exact topology cluster.
Cluster label	Human-readable topology cluster label, for example `C275`.

Field	Description
Topology size	Number of records in the exact topology cluster.
Core composition	Compact coarsened core composition string for the topology group.
PCA coordinate 1	First PCA coordinate from the WL subgraph-feature embedding.
PCA coordinate 2	Second PCA coordinate from the WL subgraph-feature embedding.

Example record structure

A typical POM-DB entry is organised as a hierarchical record, with the POM-level entity connected to one or more associated material records. The POM-level record stores the curated cluster information, including formula representations, element counts, molecular mass, charge, identifiers, crystallographic source metadata, linked structure files, topology annotations, and deposited geometry and symmetry descriptors. The material records store the reported material context, including material formulae, product formulae, molecular weights, CAS Registry Numbers, and reference information. This general structure keeps the curated POM entity and the experimentally reported material forms connected while preserving them as distinct data layers, as shown in Fig. 3**a**.

This distinction is important because a POM-level entity and a reported crystalline material are not necessarily the same chemical object. The record model can be explored through the Digital Chemistry POM Explorer at https://digital-chemistry.io/pom/, which allows users to search the dataset, inspect matching POM-level records, view molecular structures, and access associated material-level information, as shown in Fig. 3**b**. The web interface is intended as an access and inspection tool, while the deposited repository remains the authoritative versioned dataset. A complete example record is provided in the deposited README file, and formal field definitions are provided in `schema.json`.

3 Technical Validation

Technical validation was designed to assess whether the released package is internally consistent, structurally well formed, traceable to source provenance, and reusable in downstream workflows. The deposited validation script performs release-level checks covering schema conformity, record-identifier integrity, formula parsing, element-count consistency, molecular-mass recalculation, charge completeness, identifier formatting, CML linkage and structure parsing, material-level provenance coverage, and annotation coverage. The principal release-level outcomes are summarised in Table 3.

All 1936 released records conform to the deposited schema and carry internally consistent identifier, formula, charge, and annotation fields. Formula strings parsed successfully for all records, the extracted element-count dictionaries matched the stored `contains_elements` values in every case, and recalculated molecular masses agreed with stored values to within 0.1 Da. All released `cml/` references resolved and parsed with matching record identifiers, non-empty atom arrays, and unique atom identifiers. All 6329 material records carry citation, CAS, and reference metadata, with 5997 including

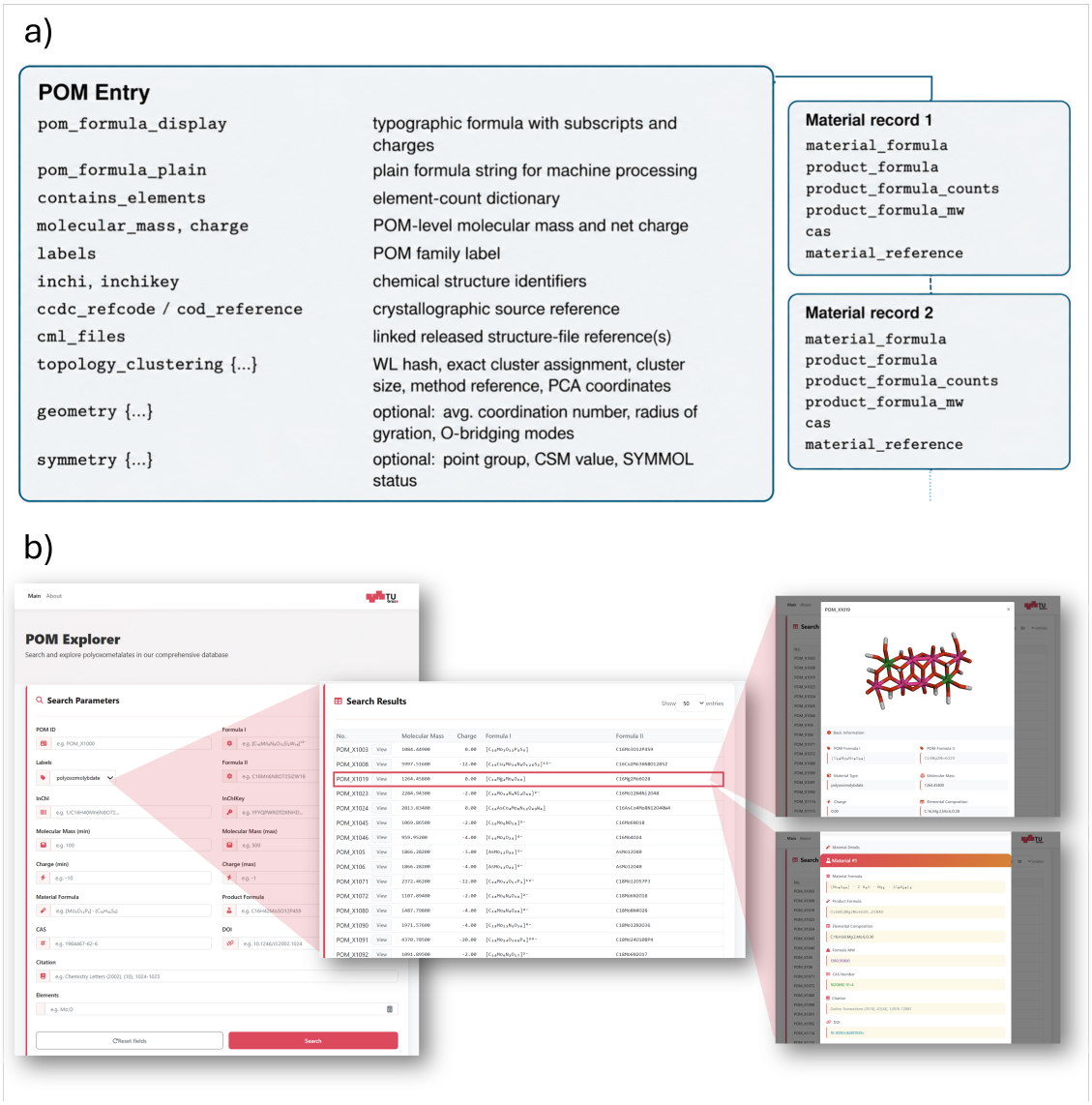


Figure 3: ***Record structure and interactive exploration.*** ***a****, Hierarchical POM-DB record model connecting a POM-level entry to material-level records, source metadata, structure-file links, topology annotations, and deposited geometry and symmetry descriptors.* ***b****, Digital Chemistry POM Explorer interface for search, filtering, record inspection, molecular visualisation, and material-level provenance lookup.*

at least one DOI. The `geometry`, `symmetry`, and `topology_clustering` blocks are present for all 1936 released records.

Table 3: ***Technical validation outcomes for the deposited POM-DB dataset.*** *Results are reported for the released JSON/CML package.*

Check	Outcome
Schema conformity	1936 / 1936 records conform to the released schema
Record identifier integrity	1936 / 1936 record identifiers are present, unique, and aligned with the top-level JSON keys
Formula parsing	All 1936 POM-level formula strings parsed successfully
Element-count consistency	1936 / 1936 parsed POM-level element-count dictionaries match the stored `contains_elements` values
Mass recalculation	1936 / 1936 stored masses agree with recalculated values to within 0.1 Da
Charge completeness	1936 / 1936 records carry an explicit numeric charge value
InChI / InChIKey formatting	1936 / 1936 records carry correctly formatted InChI and InChIKey strings
CML linkage and structure check	1936 / 1936 CML references resolve and parse with matching record identifiers, non-empty atom arrays, and unique atom identifiers
Material provenance metadata	6329 / 6329 material records carry citation, CAS, and reference metadata
DOI coverage	5997 / 6329 material records include at least one DOI
Geometry annotation coverage	`geometry` is present for all 1936 / 1936 released records
Symmetry annotation coverage	`symmetry` is present for all 1936 / 1936 released records
Topology annotation coverage	`topology_clustering` is present for all 1936 / 1936 released records

4 Usage Notes

Users should distinguish POM-level records from material-level records. POM-level formulae describe the curated cluster entity, whereas material-level records may include counter-ions, waters of crystallisation, solvent molecules, organic components, and repeated literature occurrences. The released JSON links CML structure files for all 1936 POM-level records and includes `geometry`, `symmetry`, and `topology_clustering` annotations for the full released record set. Aggregate geometry and symmetry summaries in Fig. 2**e–g** were computed on a processed subset of 1616 coordinate-bearing working structures and should therefore be read as descriptive structural statistics rather than as separate release-count totals. Complete examples for loading and querying the dataset in

Python are provided in the deposited README file, together with the SciFinder query details and a minimal validation script reproducing the technical validation table.

The released CML files provide the public POM-level coordinate representation for structure inspection, descriptor generation, format conversion, structural classification, and workflow input. They should not be interpreted as geometry-optimised, solution-phase, or directly simulation-ready structures without further expert preparation. Electronic-structure or molecular-dynamics calculations may require proton placement, charge assignment, counter-ion selection, solvent treatment, oxidation-state and spin-state assessment, and geometry optimisation. POMs are polynuclear metal–oxo clusters that can undergo partial reduction, with charge often delocalised across the metal–oxo framework. As a result, related redox or charge states may display only modest geometric differences. Experimentally derived POM geometries can therefore provide useful starting structures for modelling related charge states, provided that protonation, counter-ion environment, oxidation-state distribution, spin state, and geometry relaxation are checked case by case.

5 Code Availability

A validation script (`validate_dataset.py`) and dependency list (`requirements.txt`) are included in the public repository alongside `pom-db.json`, `schema.json`, and the released CML files. The validation script reproduces the released package-level counts, schema checks, formula-parsing checks, identifier-format checks, CML linkage and structure checks, and material-level provenance coverage checks. The validation code was tested with Python 3.11, and the tracked package requirements are listed in `requirements.txt`. Workflow documentation describes the RDF parsing, DOI normalisation, InChI-based deduplication, CIF retrieval, MOL–structure matching, and final JSON assembly steps. Some upstream acquisition steps depend on subscription resources, including SciFinder and CSD, and are therefore documented for transparency rather than provided as fully rerunnable public workflows. Crystallographic structure retrieval from the Cambridge Structural Database requires a CSD licence.

Acknowledgments

A.K. thanks the Field of Expertise – Advanced Materials Science of Graz University of Technology for Competitive Start-up Funding support (AF22-635-1). The author thanks members of the molecular inorganic InChI community for helpful discussions on chemical identifiers and their interpretation. The Cambridge Crystallographic Data Centre (CCDC) reviewed and approved the sharing of limited geometric information for POM motifs derived from CSD entries.

References

[1] A. Bijelic, M. Aureliano, and A. Rompel. The antibacterial activity of polyoxometalates: structures, antibiotic effects and future perspectives. *Chem. Commun.*, 54: 1153–1169, 2018. doi:10.1039/C7CC07549A.

[2] G. Blanke et al. Making the inchi fair and sustainable while moving to inorganics. *Faraday Discuss.*, 256:503–519, 2025. doi:10.1039/D4FD00145A.

[3] Y. G. Chung et al. Computation-ready, experimental metal–organic frameworks: a tool to enable high-throughput screening of nanoporous crystals. *Chem. Mater.*, 26: 6185–6192, 2014. doi:10.1021/cm502594j.

[4] Y. G. Chung et al. Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: Core mof 2019. *J. Chem. Eng. Data*, 64:5985–5998, 2019. doi:10.1021/acs.jced.9b00835.

[5] S. W. Gabrielson. Scifinder. *J. Med. Libr. Assoc.*, 106:588–590, 2018. doi:10.5195/jmla.2018.515.

[6] S. Gražulis et al. Crystallography open database: an open-access collection of crystal structures. *J. Appl. Crystallogr.*, 42:726–729, 2009. doi:10.1107/S0021889809016690.

[7] C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward. The cambridge structural database. *Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater.*, 72:171–179, 2016. doi:10.1107/S2052520616003954.

[8] N. I. Gumerova and A. Rompel. Synthesis, structures and applications of electron-rich polyoxometalates. *Nat. Rev. Chem.*, 2:0112, 2018. doi:10.1038/s41570-018-0112.

[9] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi. Inchi, the iupac international chemical identifier. *J. Cheminform.*, 7:23, 2015. doi:10.1186/s13321-015-0068-4.

[10] IUPAC InChI. Iupac inchi reference implementation. GitHub repository, 2025. URL https://github.com/IUPAC-InChI/InChI.git. tag v.1.07.4_RC4, commit 629f704.

[11] Y. P. Jeannin. The nomenclature of polyoxometalates: how to connect a name and a structure. *Chem. Rev.*, 98:51–76, 1998. doi:10.1021/cr960397i.

[12] A. Kondinski. Computation-ready experimental polyoxometalate structural dataset, 2026. URL https://doi.org/10.5281/zenodo.20370381.

[13] A. Kondinski, N. I. Gumerova, and A. Rompel. Curated polyoxometalate formula dataset. *Data*, 9:124, 2024. doi:10.3390/data9110124.

[14] A. Kondinski, N. I. Gumerova, A. Rompel, P. Falcaro, and T. Schreck. Data-driven polyoxometalate chemistry. *Chem. Eur. J.*, 31:e202501528, 2025. doi:10.1002/chem.202501528.

[15] S. V. Krivovichev. Which inorganic structures are the most complex? *Angew. Chem. Int. Ed.*, 53:654–661, 2014. doi:10.1002/anie.201304374.

[16] R. Liu and C. Streb. Polyoxometalate-single atom catalysts in energy research and catalysis. *Adv. Energy Mater.*, 11:2101120, 2021. doi:10.1002/aenm.202101120.

[17] X. López, J. J. Carbó, C. Bo, and J. M. Poblet. Structure, properties and reactivity of polyoxometalates: a theoretical perspective. *Chem. Soc. Rev.*, 41:7537–7571, 2012. doi:10.1039/C2CS35168D.

[18] P. Z. Moghadam et al. Development of a cambridge structural database subset: a collection of metal–organic frameworks for past, present, and future. *Chem. Mater.*, 29:2618–2625, 2017. doi:10.1021/acs.chemmater.7b00441.

[19] P. Murray-Rust, H. S. Rzepa, and M. Wright. Development of chemical markup language as a system for handling complex chemical content. *New J. Chem.*, 25: 618–634, 2001. doi:10.1039/B008780G.

[20] D. Ongari, A. V. Yakutovich, L. Talirz, and B. Smit. Building a consistent and reproducible database for adsorption evaluation in covalent–organic frameworks. *ACS Cent. Sci.*, 5:1663–1675, 2019. doi:10.1021/acscentsci.9b00619.

[21] N. Paskin. Digital object identifier system. In *Encyclopedia of Library and Information Sciences*, volume 3, pages 1586–1592. 2010. doi:10.1081/E-ELIS3-120044418.

[22] T. Pilati and A. Forni. Symmol: a program to find the maximum symmetry in an atom cluster, given as a z-matrix or as cartesian coordinates. *J. Appl. Crystallogr.*, 31:503–504, 1998. doi:10.1107/S0021889898002763.

[23] M. T. Pope. *Heteropoly and Isopoly Oxometalates*, volume 8 of *Inorganic Chemistry Concepts*. Springer, Berlin and Heidelberg, 1983.

[24] M. T. Pope and A. Müller. Polyoxometalate chemistry: an old field with new dimensions in several disciplines. *Angew. Chem. Int. Ed. Engl.*, 30:34–48, 1991. doi:10.1002/anie.199100341.

[25] N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn, and K. M. Borgwardt. Weisfeiler–lehman graph kernels. *J. Mach. Learn. Res.*, 12:2539–2561, 2011.

[26] P. Strömert, J. Hunold, A. Castro, S. Neumann, and O. Koepler. Ontologies4chem: the landscape of ontologies in chemistry. *Pure Appl. Chem.*, 94:605–622, 2022. doi:10.1515/pac-2021-2007.

[27] M. D. Wilkinson et al. The fair guiding principles for scientific data management and stewardship. *Sci. Data*, 3:160018, 2016. doi:10.1038/sdata.2016.18.

[28] S. Yang, M. Lach-Hab, I. I. Vaisman, E. Blaisten-Barojas, X. Li, and V. L. Karen. Framework-type determination for zeolite structures in the inorganic crystal structure database. *J. Phys. Chem. Ref. Data*, 39:033102, 2010. doi:10.1063/1.3432459.

[29] C. Zheng, Y. Li, and J. Yu. Database of open-framework aluminophosphate structures. *Sci. Data*, 7:107, 2020. doi:10.1038/s41597-020-0452-4.