Preprint CheMIn Technical Reports ISSN 3061-1075

Digitising Polyoxometalate Syntheses

Jan Leodolter[1], Ella Duvanova[2,3], Annette Rompel[2], Nadiia I. Gumerova[2], Aleksandar Kondinski[1]

released: 1 November 2026

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

[2] Fakultät für Chemie,
Institut für
Biophysikalische Chemie
Universität Wien
Josef-Holaubek-Platz 2, 1090 Wien, Austria

[3] Vienna Doctoral School in Chemistry (DoSChem)
University of Vienna
Währingerstraße 42, 1090 Vienna, Austria

Preprint No. 10



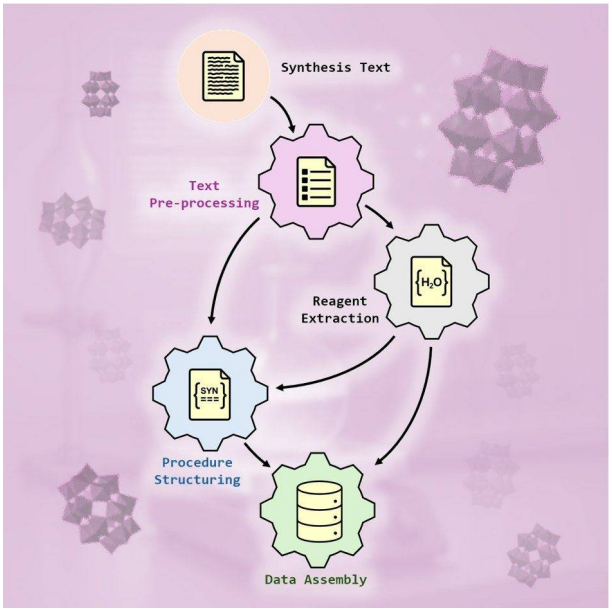
Keywords: polyoxometalates, large language models, text mining, synthesis digitisation, materials informatics

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 (POMs) are structurally diverse metal-oxo clusters whose syntheses are governed by a delicate interplay of parameters, including pH, counter-cations, and concentration. To convert this synthetic knowledge into a machine-readable format, we devised a 4-stage automated pipeline. The pipeline employs a series of Large Language Model (LLM) agents for text pre-processing, reagent extraction, and procedure structuring, followed by a final data assembly module to consolidate the outputs. This workflow transformed 405 unstructured synthesis descriptions into a strict JavaScript Object Notation (JSON) corpus containing structured metadata, reagents, and operational steps. The resulting database is ready for rapid querying and data-driven optimisation of new POM materials.



Highlights

- An LLM pipeline converts 405 polyoxometalate syntheses into a strict JSON corpus.
- Competency questions validate schema coverage of the digitised procedures.
- Unsupervised learning reveals 13 latent families of synthetic procedures.
- A RAG web explorer regenerates synthesis protocols in multiple languages.

Contents

1 Introduction **3**

2 Methodology **4**

2.1 Domain Complexity of POM Syntheses 5

2.2 Formalising Tacit POM-Synthesis Knowledge 8

2.3 Knowledge Engineering and Competency Questions 9

2.4 Automated Synthesis Data Extraction 9

2.5 Machine Learning Analysis 11

3 Results and Discussion **12**

3.1 Competency Questions Validation 12

3.2 Overview on Synthetic Procedures 13

3.3 Synthetic Data Analytics 13

3.4 Retrieval-Augmented Generation from Structured Synthesis Data 14

3.5 Limitations and Error Analysis 15

4 Conclusion and Outlook **17**

1 Abbreviation List **19**

2 Competency Questions **20**

3 Competency Question Analysis Report **20**

4 Independent Text-Generation Cross-Check **21**

5 Extraction Prompts **22**

6 Synthetic Data Overview **25**

References **33**

1 Introduction

Polyoxometalates (POMs) constitute a chemically versatile family of polynuclear, early-transition-metal oxo clusters whose electronic, structural and catalytic richness has inspired sustained interest for more than a century.[42] Constructed from high-valent molybdenum, tungsten, vanadium or niobium centres bridged by μ-oxo ligands, POMs adopt atomically precise architectures ranging from the classical Keggin and Dawson anions to giant wheel, hollow cage frameworks, and densely packed clusters, each enabling precise modulation of molecular size, shape, and functionality.[35] As these molecular metal-oxide frameworks combine structural robustness with a dense manifold of accessible redox states, they have been explored in heterogeneous and homogeneous catalysis [34], life sciences [2, 4–6, 25], and nanoelectronics [28, 29]. Despite these opportunities, the targeted synthesis and functional diversification of POMs remain largely empirical. Thus, subtle changes in pH, heteroatom identity, counter-cation size or solvent often dictate whether a reaction affords the desired anion, an intractable mixture or nothing at all.[23] Compounding this complexity, many foundational procedures for synthesising key archetypes and precursors are scattered across older literature, sometimes in multiple, inconsistent variants.[22] As such, unifying these protocols in a central, accessible framework is essential to enable reproducibility, streamline synthetic efforts, and support data-driven advances in POM chemistry. Accordingly, there is a pressing need for rational design principles and data-driven discovery tools that can accelerate the development of next-generation POM-based catalysts, materials, and bioagents.

The core challenge is that this accumulated knowledge is locked in unstructured synthesis descriptions present in scientific papers. This text is often ambiguous, relies on unstated tacit knowledge (e.g., "heat until clear," "wash with ether"), and is published in many different languages, creating a significant barrier to large-scale analysis. Whilst powerful materials informatics approaches like knowledge graphs (KGs) and data-driven models exist, they require structured, machine-readable data (such as vectors or database entries) as input. [30, 31] Manually curating this data is impractically slow, creating a bottleneck that has largely isolated POM synthesis from modern data-driven discovery.

Established rule-based and statistical chemical text-mining tools such as OSCAR4[27] and ChemDataExtractor[51] have long provided named-entity recognition and relation extraction for chemical literature, and remain effective for well-templated tasks such as identifying compound names, properties and simple reaction roles. Their underlying approach, however, is built on hand-engineered grammars and part-of-speech patterns designed to locate and tag explicitly stated entities. It is not designed to resolve the tacit, context-dependent procedural language that dominates synthesis descriptions, where the correct action is often implied rather than stated in a fixed syntactic form (e.g. “heat until clear” or “add the boiling water” implicitly requiring two separate, ordered actions – heating, then adding – neither of which is a taggable named entity). This is a difference in task, not merely in performance: these tools were built to tag entities within text, not to reconstruct an ordered, schema-conformant sequence of actions and quantities from underspecified prose, which is why we do not report a head-to-head extraction-accuracy benchmark against them in this work – doing so would compare two systems solving different problems rather than isolate a single, controlled variable, and any such comparison

is more meaningfully made against other large language models (LLM)-based structured-extraction pipelines once a shared benchmark corpus and task definition exist for this specific problem, which is not yet the case for POM synthesis digitisation. Chemistry-specific reviews of digitalisation trends reach the same conclusion: transformer-based approaches are displacing rule-based text-mining specifically because they can resolve this class of implicit, context-dependent instruction.[33] Recent advances in LLMs offer a transformative solution to this extraction problem.[17, 48] Unlike rigid parsers, LLMs possess the advanced reasoning needed to interpret the complex context and intent of chemical procedures. Foundational efforts in this domain have demonstrated LLM-driven digitisation, often by developing formal ontologies for deep semantic integration into pre-existing knowledge graphs.[45] While such an approach is powerful, it is inherently ecosystem-specific and has thus far been demonstrated on a limited scale. This work addresses the complementary challenges of scalability and immediate, practical utility. We present a lightweight, modular pipeline that deliberately prioritises the generation of a pragmatic, flexible, and universally usable JSON corpus over platform-dependent ontology integration. This JSON-first approach is what enables operation at a significantly larger scale, and it makes the structured data immediately available for downstream applications, such as the Retrieval-Augmented Generation (RAG) tool we demonstrate to improve accessibility.[21]

This effort aligns with the broader digitalisation of inorganic chemistry, where transformer-based models are rapidly superseding traditional text-mining approaches.[33] Recent studies have successfully deployed LLM agents to extract synthesis variables for Metal-Organic Frameworks (MOFs)[55] and solid-state semiconductors,[16] demonstrating that complex procedural constraints can be effectively mapped to structured formats. Furthermore, the emergence of autonomous research agents capable of planning and executing experiments[8, 9] underscores the critical need for such structured, domain-specific datasets to serve as the grounding truth for automated discovery.

This paper introduces an automated multi-stage LLM pipeline to digitise POM synthesis at scale. We demonstrate this framework by applying it to 405 unstructured synthesis routes, converting them into a queryable JSON corpus that adheres to a novel POM synthesis schema. We then leverage this structured dataset for two proof-of-concept applications. First, an unsupervised machine learning analysis to identify latent families of synthetic procedures. Second, an interactive RAG-based web application that regenerates protocols in multiple languages. In doing so, we aim to reduce reliance on trial-and-error methods, enhance the standardisation and reusability of reported procedures, and enable the automated discovery of new synthesis-structure relationships.

2 Methodology

Our methodology addresses two distinct challenges: the domain-specific complexity of POM chemistry and the general linguistic ambiguity of scientific texts. This section is therefore organised as follows: First, we detail the domain complexity of POM syntheses that necessitates a structured data approach. Second, we outline the challenge of formalising tacit POM-synthesis knowledge from procedural texts. Building on these foundations,

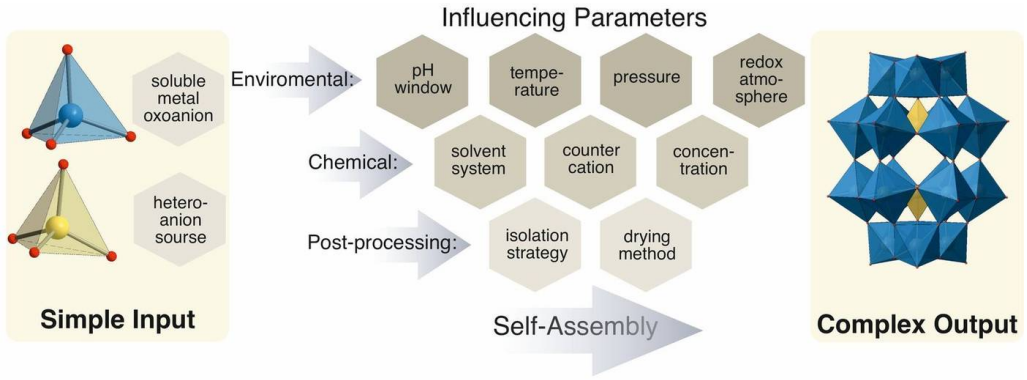


Figure 1: *POM formation depends on a densely connected set of experimental parameters.*

we then develop the competency questions and JSON schema that define our data model. Finally, we describe the automated four-stage pipeline developed to perform the data extraction and the machine learning methods used for analysis.

2.1 Domain Complexity of POM Syntheses

The synthesis of POMs is driven by acid-induced condensation of early transition metal oxoanions and is sensitive to a wide set of physicochemical variables (Figure 1), so even small deviations can steer a reaction toward undesired species or amorphous precipitates. Although the formation of these clusters often appears straightforward, reproducibility and product selectivity depend on subtle variations in pH, concentration, redox environment and ionic composition. For the present work, we assembled a curated set of 405 experimentally reported procedures, each assigned an internal identifier to support traceability across the pipeline. Candidate procedures were initially identified from classical multi-volume preparative-chemistry compilations, including the *Inorganic Syntheses* series and a Chinese-language inorganic synthesis reference work, which served as an index of established preparations rather than as the cited source itself: for each candidate, the authors subsequently traced and verified the original primary-literature article or book chapter, and it is this original source – not the compilation – that is cited and deposited alongside every procedure (Supporting Information). This is a targeted, expert-curated corpus rather than the product of a single systematic database query with a fixed search string, and it is not an exhaustive or statistically representative sample of the published POM synthesis literature: it necessarily reflects the sources the authors identified and traced, is skewed towards addenda/heteroatom families (W, Mo, V) that recur across classical references, and should not be read as a census of POM synthesis as a whole. The collection includes classical syntheses such as the preparation of $(NH_4)_6[ZnW_{12}O_{40}] \cdot 15H_2O$[11] and more complex molybdenum oxo assemblies such as $(H_3O)(C_3H_5N_2)[Mn(OH)_6Mo_6O_{18}] \cdot 3.5H_2O$.[49]

Two general strategies are employed in POM synthesis. One pot routes combine soluble addenda sources such as $Na_2WO_4 \cdot 2H_2O$ or $(NH_4)_6[Mo_7O_{24}] \cdot 4H_2O$ with suitable heteroatom precursors, for example H_3PO_4 or Na_2SiO_3, in aqueous medium followed by con-

trolled acidification. A classical case is the formation of $[\alpha\text{-SiW}_{12}O_{40}]^{4-}$ from a stoichiometric mixture of tungstate and silicate in water, which is obtained on acidification under the reported conditions.[52] In contrast, stepwise routes rely on the isolation and subsequent transformation of reactive intermediates such as lacunary or substituted fragments. A preformed framework $[PW_{11}O_{39}]^{7-}$ can for instance react with K_2IrF_6 in water under hydrothermal conditions to give the targeted substituted cluster $[PW_{11}O_{39}Ir(H_2O)]^{4-}$, which shows the flexibility of the stepwise approach for functionalised systems.[10, 50] These two routes are summarised in Figure 2.

Among the many variables influencing POM synthesis, *pH* is the most critical. The condensation of oxoanions to form discrete POM clusters typically takes place within narrow pH windows, sometimes with tolerances as small as ± 0.2 pH units. Even minor deviations can trigger hydrolysis, formation of lacunary species or uncontrolled polymerisation. The preparation of the Wells Dawson anion $[P_2W_{18}O_{62}]^{6-}$ was reported to succeed only in the very narrow pH interval from 1.2 to 1.4, and a small drift from this range already produced material that could not be characterised.[1] This high sensitivity is consistent with the analysis of Graham and Finke,[22], who showed that vague instructions for pH adjustment in older procedures were a frequent reason for failure to reproduce pure $K_6[\alpha P_2W_{18}O_{62}] \cdot 14H_2O$. An example of this diversification is shown in Figure 3.

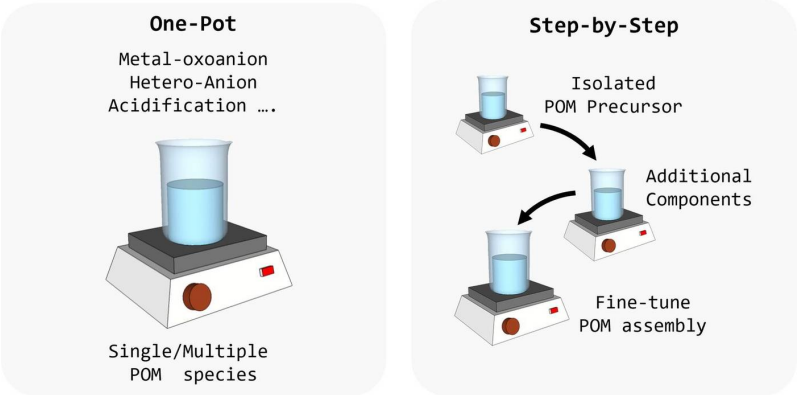


Figure 2: *Overview of one-pot and step-by-step synthesis strategies for polyoxometalates. A common set of interdependent factors, such as pH, counter-cation identity, concentration, redox conditions, solvent environment, temperature, pressure and isolation strategy, governs both routes.*

The role of counter-cations in directing POM structure and solubility is equally essential. Large and weakly coordinating tetra-n-butylammonium (Bu_4N^+) cations were shown to stabilise highly charged species such as $[SiW_9Nb_3O_{40}]^{7-}$ and to enable their transfer into organic media.[20] In contrast, potassium cations promote the formation of dimeric assemblies of the type $[Ln(\beta_2\,SiW_{11}O_{39})_2]^{13-}$ where Ln denotes a lanthanoid cation by templating contacts between lacunary subunits $[\beta_2\,SiW_{11}O_{39}]^{8-}$.[3] These examples agree with the established cation size effects on nuclearity and aggregation in POM chemistry.[36, 41]

The redox environment is another determining factor, especially for molybdenum-based and vanadium-based systems that undergo comproportionation with ease. Electrochemical reduction of $H_4[SiMo_{12}O_{40}] \cdot 15H_2O$ in a mixed aqueous and acetic acid medium un-

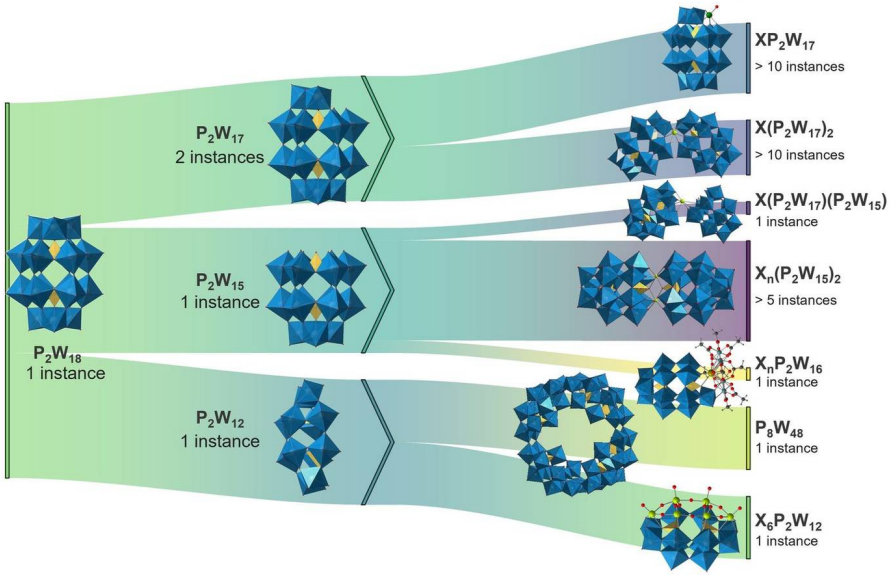


Figure 3: *Synthetic diversification of the Wells-Dawson archetype through lacunary precursors.*

der inert atmosphere produced the reduced species $H_6[Si_2Mo_2^{V}Mo_{16}^{VI}O_{40}] \cdot 15H_2O$ as a red solid, which shows how important strict control of redox potential and exclusion of oxygen is in such systems.[18] Chemical reduction gives a similar picture. Use of sodium dithionite in acidic sodium molybdate solution can lead to large mixed-valence molybdenum oxide clusters that appear as a deep blue solid after standing for 24 h. Together, these examples show that both electrochemical and chemical control of the redox state are decisive for steering speciation and nuclearity in reduced POMs and that even mild reductants or traces of oxygen can change the outcome. Concentration effects are not linear. Below about 10^{-3} M in addenda metal, the condensation does not proceed, while above about 10^{-1} M, rapid polycondensation gives gel-like material. Scale-up therefore has to be done in small steps with adjustment of volume and addition rate instead of a simple proportional increase.[12, 24, 26]

Temperature and pressure also play critical roles in directing both crystallisation and speciation in POM synthesis. The vanadotungstate cluster $Na_4(NH_4)_2[\alpha_2H_2VW_{17}O_{54}(VO_4)_2]$ was obtained by hydrothermal treatment at 160 °C for 72 h followed by slow evaporation, which gave well defined purple crystals in 89% yield.[39] This shows that elevated temperature and the autogenous pressure of a sealed system can give access to POM structures that are kinetically inaccessible or not stable under ambient conditions.[40]

Solvent composition and ionic strength play a central role in shaping POM speciation, solubility and crystallisation behaviour. Water is the main solvent, but its combination with organic cosolvents or its replacement by polar aprotic media is often required for isolation or structural control. The Keggin type heteropolyacid $H_5[\alpha/\beta AlW_{12}O_{40}] \cdot 15H_2O$ was obtained from a precursor solution of $Na_6[Al(AlOH)_2W_{11}O_{39}]$ through acidification and prolonged refluxing in aqueous H_2SO_4 and the product was then isolated by liquid-liquid extraction.[15, 53] Addition of diethyl ether induced phase separation and allowed selective recovery of the hydrated heteropolyacid from the lower aqueous phase, which

shows how solvent polarity and phase behaviour can be used to separate POMs from interfering ions such as Na^+ or Cl^-.[54]

Non-aqueous synthesis is also possible under strictly anhydrous conditions.[43] A substituted Keggin type niobotungstate $(nBu_4N)_4[PNbW_{11}O_{40}]$ was prepared entirely in dry acetonitrile under inert atmosphere.[44] Here, the use of organic soluble tetraalkylammonium salts ensured solubility in the organic medium and enabled high-yield crystallisation by solvent evaporation and ether-based recrystallisation. This approach is useful for isolating hybrid or highly charged POM species that are unstable or inconvenient in water.

The isolation and purification of POMs often present additional challenges. Many syntheses end with a precipitate or a crystalline product, but post-synthetic steps such as redissolution, ion exchange or reprecipitation can still be needed to remove by-products or to stabilise labile species. A giant mixed-valence molybdenum oxide cluster was obtained as blue rod-like crystals by controlled dropwise mixing at $4\,mL\,h^{-1}$ followed by reaction under stirred flow conditions.[19] The crude material nevertheless required a rediffusion crystallisation step to give an analytically pure product, which shows that even well formed crystals can profit from recrystallisation in order to improve phase selectivity and purity. A different strategy was used for $K_8[\beta_2SiW_{11}O_{39}]\cdot 14H_2O$, which was selectively precipitated by adding KCl to an acidified reaction mixture.[47] The solid obtained in this way was redissolved and reprecipitated to improve homogeneity. This classical purification approach shows how counter-ion exchange and salting-out can promote selective crystallisation of POM frameworks, especially when slowly transforming or metastable species are present.[7, 35, 36]

Finally, rigorous characterisation is indispensable for establishing reliable structure-property relationships in POM chemistry: multinuclear NMR, IR, Raman spectroscopy, and ICP-OES are routinely combined to confirm the elemental composition and purity of the final product. Taken together, these examples show the intricate interplay of synthetic parameters that govern POM formation and stability. Reproducible and scalable synthesis depends not only on precise control of individual variables but also on an integrated view of how these variables interact in practice. This complexity highlights the value of centralised, annotated databases, such as the one developed in this study, for capturing synthetic diversity and guiding rational design.

2.2 Formalising Tacit POM-Synthesis Knowledge

A major challenge in digitising chemical literature is the prevalence of *tacit knowledge*—implied actions, unstated conditions, and unstructured free text phrasing such as *heat until clear*, *let it settle*, or *wash with ether*. Human chemists infer the correct actions from this context, while simple text-matching algorithms fail. To capture this implied intent, we employ a reasoning-based approach. Instead of matching only keywords, we use a LLM to interpret the *intent* of the procedure. This allows conversion of ambiguous text into structured, machine-readable data. For example, a phrase like *settle for 12 h at 25 °C* is transformed into structured `time` and `target` objects. This JSON schema forms the foundation of the entire analysis.

2.3 Knowledge Engineering and Competency Questions

Competency questions (CQs) express in natural language the types of queries that an ontology or data model should be able to answer. They capture the scientific information needs that motivated the schema and therefore act both as design requirements and as test cases. By showing that each CQ can be answered with the structured data we generate, we give evidence that the representation supports typical literature mining, synthetic planning and data retrieval tasks in POM chemistry.

To make every competency question answerable directly from the captured data, each POM synthesis is stored as a rigorously structured JSON record with clearly defined blocks and fields. Each element of the top-level `procedures` array is a self-contained object centred on a procedure node that branches into six thematic blocks. The `procedure_information` section records the identifier, title, DOI, provenance, reference, extraction date and the extraction agent so that every route can be traced unambiguously. All chemical inputs appear exactly once in the reagents list, where a unique `reagent_id` is paired with name, role, amount, units and where reported concentration. These identifiers are then referenced by the workflow steps, which prevents duplication. Equipment that is used is listed in a dedicated apparatus block. The core of the record is the chronologically ordered `procedure_steps` array. Each step carries a unit operation label, numeric parameters for time and temperature, any linked `reagent_id`, a `source_line_idx` field recording exactly which sentence of the normalised Phase I text the step was derived from, and, where the source text reports it, further condition fields such as pH, atmosphere and apparatus, so that any structured field can be traced back to the specific line of source text that justifies it, and so that tacit instructions become machine actionable without losing nuance. Analytical readouts such as NMR peaks, IR bands or elemental analysis values are grouped under characterisation, while final yield and other target-specific information are summarised in `product_summary` to enable rapid filtering. By enforcing a fixed key order, explicit null placeholders for missing values, and unit retention exactly as reported, the schema combines strict machine-readability with the flexibility needed to capture the tacit knowledge that chemists infer from experimental synthesis texts.

2.4 Automated Synthesis Data Extraction

To convert the raw procedural texts into the coherent and queryable JSON corpus, we implemented an automated, multi-stage pipeline orchestrated by a main Python controller (Figure 4). This pipeline employs a series of specialised LLM agents, each executing a constrained task via schema-aligned prompting. This modular design is deliberate: instead of a single, complex request prone to error, we chain multiple, simple-to-verify tasks. This separation of concerns (e.g., extracting reagents is a separate task from structuring procedural steps) is the core of the pipeline's strategy for minimising errors and grounding the final output. The LLM-driven modules (Phases I–III) are powered by the DeepSeek Reasoner model, selected for its combination of strong instruction-following on schema-constrained extraction tasks, an explicit reasoning/chain-of-thought capability well suited to inferring tacit procedural intent (e.g. resolving an implied action rather than only matching stated keywords, as discussed below), and cost-effectiveness at the scale

of 405 procedures processed across three sequential phases each – a material, practical consideration, since API-hosted proprietary LLMs are billed per token and a lower-cost model directly determines how large a corpus of this kind can be extracted, audited, and, where necessary, repaired (Section 3.5) within a realistic project budget. We did not conduct a systematic model-selection benchmark across alternative LLMs, and note this as a design choice made for practical throughput and cost rather than one established by comparative evaluation.

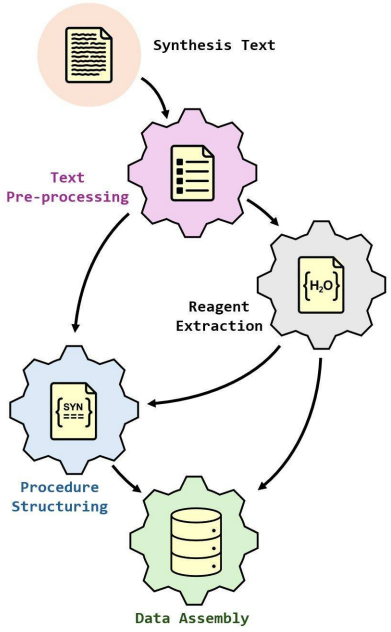


Figure 4: *Automated synthesis data extraction pipeline. (I) A Text Normalisation Agent cleans and structures the raw text. (II) A Reagent Extraction Agent builds a chemical manifest from the cleaned text. (III) A Procedure Structuring Agent maps the normalised text onto the reagent manifest to create a step-by-step workflow. (IV) A final Data Assembly script merges these artefacts into the final JSON record.*

Phase I, the Text Normalisation Agent, ingests the raw `.txt` files. Its primary role is not to extract, but to act as an "extractive technical writer." It is prompted to *rewrite* the unstructured free text into a strict, one-action-per-line format. This includes standardising verbs (e.g., "shake" to "stir") and resolving contextual actions (e.g., a single sentence "add 50 mL boiling water" is intelligently expanded into two discrete steps: "Heat 50 mL water to boiling." and "Add 50 mL water."). This agent also uses an optional disambiguation map (provided as a CSV file) to disambiguate files containing multiple syntheses by targeting only the product formula specified in the map. The output is a clean, focused, and machine-actionable text file for each procedure.

Phase II, the Reagent Extraction Agent, receives the normalised text from Phase I. It is given a highly constrained prompt to *only* extract chemical names and their total amounts. This agent is designed to aggregate quantities (e.g., "100 mL water" and "50 mL water" mentioned in different steps are correctly listed as a total "150 mL water") and capture complex descriptions (e.g., "10 M HCl"). This step deliberately ignores all procedu-

ral logic, focusing solely on creating a per-procedure chemical manifest. Aggregating quantities to the procedure level, rather than attempting to resolve which step consumed which specific sub-quantity of a repeatedly-used reagent, was a deliberate scope decision: many source texts state a reagent's total amount once and then refer to it by name in later steps without repeating the quantity, so per-step attribution would in many cases require the extraction agent to infer an unstated split rather than read a stated value, trading a small number of well-defined, verifiable outputs (total amounts) for a larger number of inferred ones. The trade-off is real and is quantified directly in the Supporting Information: an independent text-generation cross-check detected that this aggregation caused mis-attributed step-level quantities in a substantial fraction of procedures when the full record was later used to regenerate free text (Supporting Information, "Independent Text-Generation Cross-Check"), which is why the Retrieval-Augmented Generation (RAG) explorer (Section 3) is grounded in a subsequently step-level-corrected version of the corpus rather than the raw Phase II output. This manifest is also used to update a global reagent consistency database to ensure reagent name consistency across the corpus.

Phase III, the Procedure Structuring Agent, is the core reasoning step. This agent is prompted with three key inputs: 1) the normalised, one-action-per-line text from Phase I, 2) the reagent manifest from Phase II (which provides a definitive list of reagent IDs, e.g., `R1: Water`), and 3) a strict JSON schema for `procedure_steps`. The agent's task is to map the unstructured free text actions onto the JSON schema, parsing instructions (e.g., "settle under 25 °C for 12 h") into structured data. Critically, it is instructed to link each action to the specific reagents from the manifest by their IDs (e.g., `"reagent_usage": ["R1"]`). This schema-aligned mapping of two pre-processed inputs ensures the final output is both structured and grounded in the source materials.

Phase IV, the Data Assembly Module, is the final non-LLM script. It acts as a deterministic assembler, programmatically reading the JSON artefacts from Phase II (the reagent list) and Phase III (the step list) for a given synthesis. It merges them into the single, complete JSON object that follows the schema defined in this work (e.g., `procedure_information`, `reagents`, `procedure_steps`). These individual objects are then collected, sorted by their identifier, and written into the final aggregated JSON corpus. The orchestrator script also archives all intermediate files from all phases into a time-stamped run folder to ensure full traceability and reproducibility.

2.5 Machine Learning Analysis

To categorise the 405 synthesis procedures, we employed an unsupervised machine learning pipeline, computed on the step-level-corrected corpus described in Section 2.4 and Section 3.5 rather than on the raw Phase II/III output, so that the reported clusters reflect the audited, corrected step data rather than the small fraction of procedures affected by the reagent-quantity aggregation issue. First, each procedure was converted into a high-dimensional feature vector (1089 features) by integrating (1) Term Frequency-Inverse Document Frequency (TF–IDF) vectorised reagent names, (2) TF–IDF vectorised operational sequences using 1 and 2 grams, and (3) standard scaled numerical and categorical parameters. Missing numerical values (unreported temperatures or times, present for a minority of steps where the source text did not state a condition) were imputed using the

column median rather than dropped, so that procedures with an incomplete but otherwise informative record were not excluded from clustering entirely; median imputation was chosen over mean imputation for its robustness to the outlier-prone, right-skewed distributions typical of reaction times (a small number of multi-day aged syntheses otherwise dominate a mean) [46]. Given the high dimensionality, this sparse feature matrix was then reduced to a 10-dimensional dense embedding using Uniform Manifold Approximation and Projection (UMAP) [38]. This embedding was then analysed using Hierarchical Density Based Spatial Clustering of Applications with Noise (HDBSCAN) [13, 14, 37], a density-based approach that identified 13 distinct synthesis families and classified 14 procedures as outliers. A separate 2-dimensional UMAP embedding was generated from the original 1089-dimensional feature space for visualisation, with points coloured by their final cluster assignments.

3 Results and Discussion

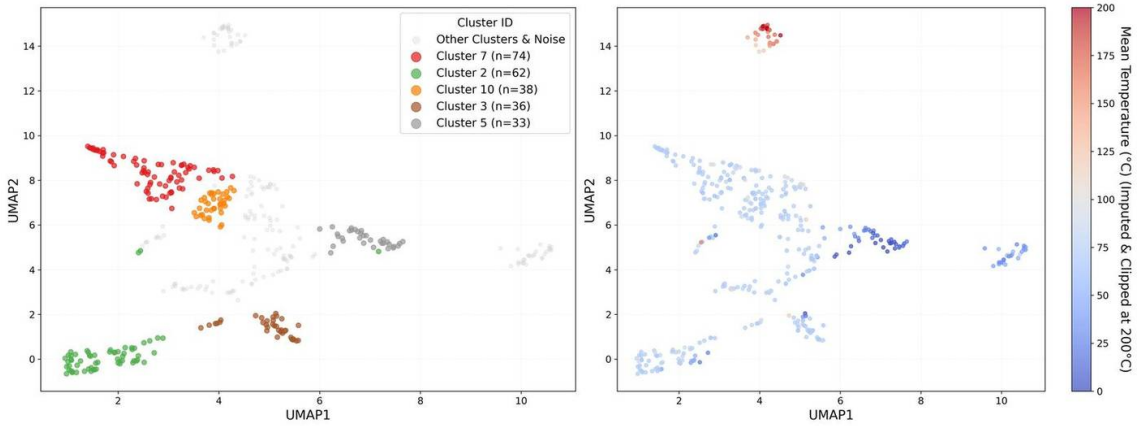


Figure 5: *Visualisation of the reaction space using Uniform Manifold Approximation and Projection (UMAP). The left panel highlights the five most populous clusters, while the right panel colours each procedure by its mean temperature. Points clustered closer together share greater similarity in reagents and operations.*

3.1 Competency Questions Validation

To validate the structured JSON schema, which was designed based on the competency questions (CQs) defined in Section 2.3, the 405-procedure corpus was first tested. The complete set of CQs was executed as automated queries against the dataset. The JSON-based dataset successfully answered every query (including two that legitimately returned zero matching procedures), confirming that the schema is queryable and internally consistent enough to retrieve specific procedural details reliably. This demonstrates the utility of the data model as a retrieval and filtering layer; it is a test of schema coverage and consistency, not a measurement of extraction accuracy against the original texts, which is

addressed separately by the independent text-generation cross-check reported in the Supporting Information. For example, queries successfully identified 86 procedures using `Sodium Tungstate Dihydrate` as a main reagent, 121 procedures involving a target temperature of at least 80°C, and 184 procedures that include a `Filter` operation. A comprehensive validation report for all CQs and their corresponding result counts is provided in the Supporting Information (Section S2).

3.2 Overview on Synthetic Procedures

Following the successful data extraction from the 4-stage pipeline (Section 2.4), the resulting corpus enables a high-level statistical analysis of the synthetic POM landscape. This task was previously intractable with unstructured text. Analysis of the addenda metal sources (Supporting Information, Table S1) confirms the dataset's focus on classical POM chemistry: Tungsten-based POMs are the most common (199 procedures, 49.1%), followed by Molybdenum (75 procedures, 18.5%) and Vanadium (40 procedures, 9.9%). Mixed-metal systems, such as V/W (30) and Mo/V (28), are also well-represented.

Solvent usage statistics (Supporting Information, Table S2) highlight the dominance of aqueous routes, with Water appearing in 309 procedures. Critically for purification and non-aqueous routes, the use of organic solvents is also quantitatively captured, including Diethyl Ether (53), Ethanol (45), and Acetonitrile (32). The frequency of unit operations (Supporting Information, Table S3) provides a digital fingerprint of POM synthesis. The five most common operations by total instance count are Add (1200), Dissolve (367), Heat (307), Filter (276), and Stir (236), reflecting the typical "one-pot" acidification and precipitation/crystallisation workflows discussed in the methodology.

3.3 Synthetic Data Analytics

To uncover latent patterns within this validated corpus, we applied the unsupervised machine learning pipeline described in Section 2.5. The featurised dataset, representing each synthesis by its reagents, operations, and parameters, was projected into a 2-dimensional space using UMAP. Subsequent clustering via Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) successfully partitioned the dataset, revealing 13 chemically-distinct families of procedures (a full interpretation of all clusters is available in the Supporting Information, Table S4).

Figure 5 visualizes this UMAP projection, which maps complex procedures into a 2D space where synthetically similar procedures are placed near each other. Figure 5a highlights the five largest clusters, which represent the most common procedural archetypes in the corpus. The separation of these clusters demonstrates that the dataset contains distinct, well-populated groups of related syntheses.

This abstract clustering corresponds to tangible chemical parameters. For example, Figure 5b overlays the same UMAP projection with the mean procedure temperature, revealing clear "hot" and "cold" regions that correlate strongly with cluster locations. This is validated by inspecting the cluster definitions (Supporting Information, Table S4): Cluster 0,

located in the hot region, is a distinct hydrothermal-type family defined by its high mean temperature (167.0 °C) and long reaction times, featuring copper(II) nitrate and trimesic acid among its defining reagents. In contrast, Cluster 5 represents a "cold" synthesis family (13.6 °C), defined by a `Filter`/`Dry` sequence and a hydrochloric acid/sodium hydroxide reagent profile. Other clusters are defined by solvent environments, e.g. Cluster 3, a largely "Aqueous-Alcohol" wash/filter workup family built around ethanol and diethyl ether. The ability to automatically group procedures by these key chemical features validates that the LLM-extracted data retains the nuanced information required for data-driven analysis; full addenda-composition fractions for every cluster, including a UMAP projection coloured by addenda set, are given in the Supporting Information (Table S4 and Figure S1).

3.4 Retrieval-Augmented Generation from Structured Synthesis Data

Finally, to demonstrate a practical application of the curated corpus, we developed and deployed a browser-based *Polyoxometalate Explorer* (Figure 6). The application functions as an interactive knowledge layer, loading the entire structured dataset directly into the client's browser. This static-first design enables millisecond-fast faceted filtering by a range of metadata, including procedure identifier, target compound, elemental composition, source DOI, reagent identity, unit operation, step count, and starting addenda source; molecular mass and formal charge are not filtered on, as they are not consistently available across this corpus. This client-side approach requires no server-side search computation, which ensures both low latency and complete user data privacy, as search terms never leave the local machine. The tool's primary innovation is its use of RAG to translate structured data into human-readable text. When a user selects a POM, the entire structured JSON object for that entry—not the original unstructured text—is sent as the grounding context to a server-side endpoint backed by a GPT-4o-mini model. This model was chosen for this web-facing application due to its optimal balance of speed, cost, and high-fidelity synthesis text generation. This "structured-data-to-text" workflow ensures the generated protocol is grounded exclusively in the validated, extracted data, preventing the LLM from re-inserting ambiguities or hallucinating steps; steps for which the source recorded no specific reagent or condition are explicitly annotated as such before generation, so the model reports them as unspecified rather than inventing plausible-sounding detail. Furthermore, the user can select a target language from a dropdown menu, which modifies the prompt to regenerate the synthesis protocol in any desired language. This feature makes the curated procedures globally accessible, directly addressing and overcoming language barriers in experimental science. The explorer is deployed as a containerised web application (https://digital-chemistry.io/POMSyn/), with the static corpus and search interface served alongside a lightweight backend that mediates the generation requests. Fluent, readable output is not by itself evidence that a generated protocol is factually correct, so the generation step was independently audited rather than assumed reliable: every one of the 405 records was regenerated and automatically cross-checked, sentence by sentence, against both the structured JSON and the original source text for reagent identities and quantities, conditions, operation order and work-up steps, with any factual discrepancy flagged explicitly rather than averaged away (full methodology and results in the Supporting Information). Researchers can rapidly interrogate hundreds of

POM compositions, generate synthesis protocols on demand in their language of choice, and copy either the human-readable text or the machine-readable JSON directly into their own workflows, significantly lowering the barrier to protocol adaptation and comparative analysis.

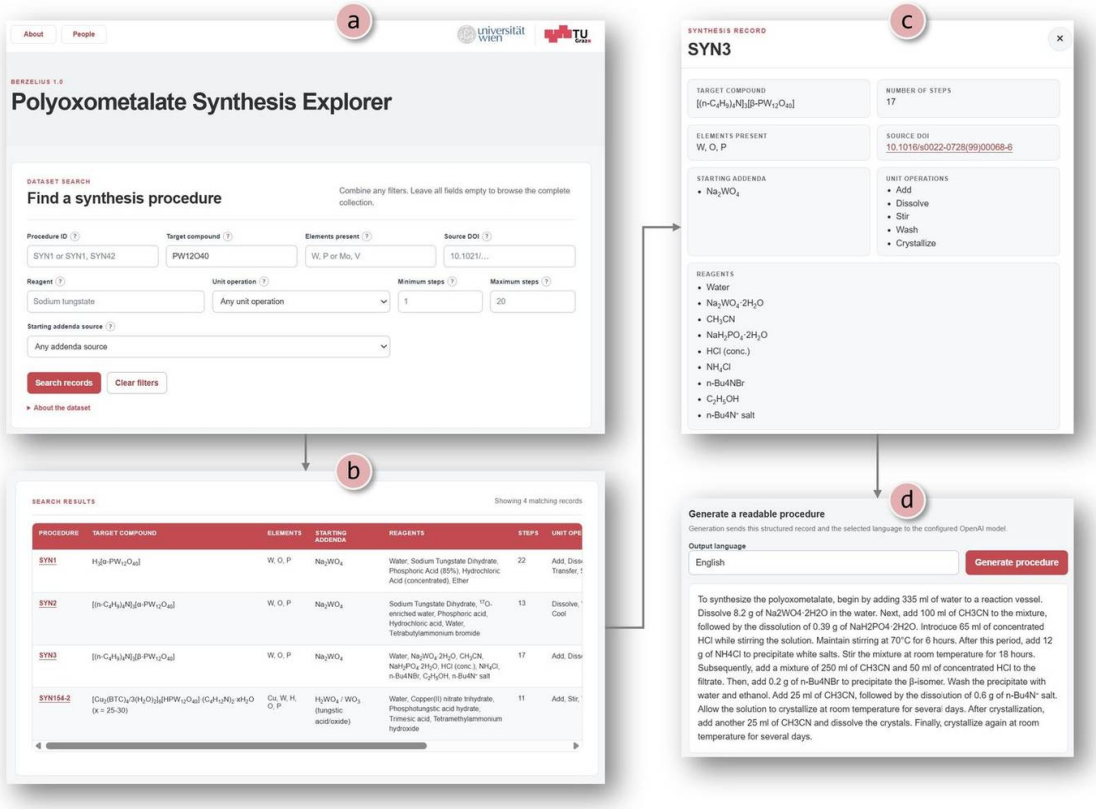


Figure 6: *Web interface of the* Polyoxometalate Explorer. *(a) Faceted, in-browser search form (procedure identifier, target compound, elemental composition, source DOI, reagent, unit operation, step count, and starting addenda source). (b) Search results table for the query* $PW_{12}O_{40}$. *(c) Structured synthesis record for the selected procedure (SYN3), showing target compound, step count, source DOI, starting addenda, unit operations, and reagents. (d) Modal card generated on demand via retrieval-augmented generation (RAG), displaying a regenerated, human-readable synthesis description for the selected entry.*

3.5 Limitations and Error Analysis

The pipeline and corpus presented here have specific, identifiable limitations, which we report explicitly rather than leave implicit.

Corpus scope. As noted in Section 2, the 405-procedure corpus is an expert-curated collection built from sources the authors identified and traced to their original publication, not an exhaustive or systematically sampled census of the POM synthesis literature; it should

not be used to draw quantitative conclusions about the relative prevalence of synthesis strategies in the literature at large, only about the strategies present within this corpus.

Extraction reliability, quantified. Rather than asserting extraction accuracy, we measured it directly: every one of the 405 records was used to regenerate a free-text synthesis description, which was then automatically checked, claim by claim, against both the structured JSON and the original source text (reagent identities and quantities, conditions, operation order, work-up steps). On the initial pipeline output, this cross-check returned 69 procedures with an exact match, 181 with an incomplete-but-uncontradicted match, and 155 with an identifiable factual discrepancy – predominantly the reagent-quantity aggregation issue discussed in Section 2.4, i.e. a structural limitation of the schema's design rather than a random error rate. Rather than report this as a static accuracy figure, we treated it as an error signal to act on: the flagged records were repaired at the JSON level (per-step quantities, temperatures and timings re-attached from the source text with line-level provenance) and re-audited against the same source text, after which 0 records retained a factual discrepancy. We regard this extract-then-audit-then-repair workflow as a defensible design in the sense that it is demonstrated here to converge, verifiably, to zero detected discrepancies, which a claim of first-pass perfection would not itself demonstrate; separating extraction from an explicit, independently-checkable audit-and-repair pass also keeps each stage individually simple and matches common practice in large-scale data curation more broadly. We do not, however, claim this is the only or the best design: a small ablation experiment reported in the Supporting Information, comparing this pipeline's first-pass output against a single-pass extraction baseline on the same procedures, provides concrete evidence that prompting the extraction step to record amounts at the step level directly – rather than aggregating to a procedure-level total and repairing afterward – would reduce how often the repair stage is needed in the first place, and we report that finding plainly as a concrete direction for a future pipeline revision rather than treat the current design as final. Full methodology, per-record verdicts, and the repair procedure are given in the Supporting Information. Critically, this automated check was not the only validation applied: the corpus was reviewed by practising synthetic POM chemists among the authors (co-authors experienced in exactly this experimental chemistry), who manually verified extracted reagent identities, addenda-source assignments, and target-compound formulas against the original primary-literature source for every record flagged as ambiguous or inconsistent, and confirmed the corpus's chemical validity. Several genuine extraction and formula errors were identified and corrected specifically through this domain-expert review, not by the automated cross-check alone. This human-in-the-loop verification by domain experts is, together with the corpus-wide automated cross-check above, our evidence for the reliability of the deposited data. We note for completeness that this expert review was not run as a separate, formally blinded annotation exercise reporting a single corpus-wide precision/recall statistic; that additional, fully independent statistical benchmark, at full corpus scale, remains a natural extension. A smaller-scale but concrete comparison against an alternative LLM-based extraction strategy was, however, run rather than left undone: the single-pass ablation described above and detailed in the Supporting Information. Independent verification by readers, however, is not merely possible in principle but actively enabled: every record in the *Polyoxometalate Explorer* (Section 3) links directly to its source DOI, so any reader can check any individual extraction against the original publication immediately, without needing to obtain the raw JSON

or re-run the pipeline; this open, ongoing, corpus-wide verifiability by the community is precisely why the deposited data and the interactive explorer are released openly rather than only summarised in this manuscript.

Schema and model choices. In addition to the prose and by-example description of the JSON schema (Section 2.3), a formal, machine-validatable JSON Schema document is deposited alongside the corpus (Supporting Information), against which every record in the deposited corpus validates; the deposited license terms are likewise stated explicitly there. Similarly, as noted in Section 2.4, the choice of DeepSeek Reasoner for the extraction agents and GPT-4o-mini for the RAG explorer reflect practical throughput and cost considerations for a corpus of this size rather than a systematic comparison across candidate models; the extraction agents (Phases I–III) are run with sampling temperature 0 throughout, to maximise run-to-run reproducibility of the extraction itself, while the RAG explorer's generation endpoint uses a modest temperature of 0.2 to allow natural language variation in the regenerated prose without departing from the grounded content. We would expect the pipeline's general architecture, though not necessarily its precise error rates, to transfer to other sufficiently capable instruction-following LLMs. Both models were accessed via their respective hosted commercial APIs rather than a locally pinned, fixed model checkpoint; unlike providers that expose dated snapshot identifiers, neither API used here guarantees that a named model string resolves to a byte-identical checkpoint indefinitely, so exact reproduction of this extraction at a much later date may reflect a since-updated version of the underlying model rather than the one originally used, a general reproducibility caveat of building on hosted, continuously-updated commercial LLM APIs rather than a limitation specific to this pipeline.

Clustering parameter sensitivity. The number of clusters recovered by HDBSCAN depends on its `min_cluster_size` parameter; we quantified this directly with a parameter sweep (Supporting Information) rather than presenting the reported 13-cluster result as parameter-independent. The cluster count ranges from 9 to 25 across a reasonable parameter range, but is stable at 13 across `min_cluster_size` = 8–10, indicating a genuine plateau in the underlying structure rather than a value tuned to a target count.

4 Conclusion and Outlook

This work demonstrated an automated pipeline that leverages LLMs to convert 405 unstructured polyoxometalate synthesis descriptions into a structured JSON corpus. This validated, machine-readable dataset enabled immediate data-driven applications, including statistical analysis, unsupervised clustering of synthetic families, and a RAG web tool. The methodology provides a practical bridge between ambiguous legacy literature and modern materials informatics, contributing to the digitalisation of POM chemistry[32] and significantly lowering the barrier to reproducible, data-driven POM discovery.

The current dataset, while rich in procedural parameters, lays the groundwork for a truly predictive model. Instead of only elucidating known relationships, the critical next step is to integrate this synthesis corpus with structural information, creating a framework for generative and predictive chemistry. This combined dataset will be the foundation for models capable of predicting synthesis parameters for novel, target POMs not yet in the

database and, ultimately, designing new, plausible synthetic routes from scratch. This work shifts the paradigm from archival data mining towards an AI-assisted system for POM discovery.

Acknowledgments

AK thanks the Field of Expertise - Advanced Materials Science of TU Graz for Competitive Start-up Funding support (AF22-635-1). This research was funded in part by the Austrian Science Fund (FWF) [DOI 10.55776/PAT4299925 (to A.R.); DOI 10.55776/PAT2849824 (to N.I.G.)]. E.D. acknowledges support from the Vienna Doctoral School in Chemistry (DoSChem). For open access purposes, the authors have applied a CC BY public copyright license to any author-accepted manuscript version arising from this submission.

Data Availability Statement

All data generated and analysed in this study are openly available. The structured JSON corpus and the archived extraction runs are deposited in the TU Graz Research Data Management repository of the authors under the project title *Digitising Polyoxometalate Syntheses*. The complete source code together with example notebooks for reproducing the 4-stage pipeline is available in the public Digital Chemistry Git repository at https://github.com/digital-chemistry/pomsyn. An interactive web explorer for browsing and searching the deposited corpus is available at https://digital-chemistry.io/POMSyn/. Further access details and file organisation are provided in the Supporting Information.

Conflicts of Interest

The authors declare no conflicts of interest.

Author contributions

Jan Leodolter and Ella Duvanova contributed equally to this work.

Correspondence

Corresponding authors: Nadiia I. Gumerova (nadiia.gumerova@univie.ac.at) and Aleksandar Kondinski (kondinski@tugraz.at).

1 Abbreviation List

API	Application Programming Interface
Bu_4N^+	tetra-*n*-butylammonium cations
CQs	Competency Questions
CSV	Comma-Separated Values
GPT-4o-mini	Generative Pre-trained Transformer 4o mini
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
h	hour
ICP-OES	Inductively Coupled Plasma Optical Emission Spectroscopy
ID	Identifier
IR	Infrared Spectroscopy
JSON	JavaScript Object Notation
KGs	Knowledge Graphs
Ln	Lanthanoid
LLM	Large Language Model
LLMs	Large Language Models
MOF	Metal-Organic Framework
MOFs	Metal-Organic Frameworks
NMR	Nuclear Magnetic Resonance
POM	Polyoxometalate
POMs	Polyoxometalates
RAG	Retrieval-Augmented Generation
TF-IDF	Term Frequency–Inverse Document Frequency
UMAP	Uniform Manifold Approximation and Projection

2 Competency Questions

Example Questions

- Show me all procedures listing `Sodium Tungstate Dihydrate` in the main `reagents` list.
- Show me all procedures that use `n-Bu`$_4$`NBr` in `reagents` or `procedure_steps`.
- Show me all procedures that use `Water` as a reagent in a `Dissolve` step.
- Show me procedures where a reagent `concentration` is greater than 1.0 mol/L.
- Show me procedures where the total procedure-level `amount` of `Hydrochloric acid` exceeds 100 mL.
- Show me procedures where the total procedure-level `amount` of `n-Bu`$_4$`NOH` exceeds 0.5 mol.
- Show me procedures executed in more than 10 elementary steps.
- Show me procedures that involve a `Stir` operation for longer than 5 min.
- Show me procedures where a `target.temperature.value` reached at least 80 °C.
- Show me all procedures that include a `Filter` `unit_operation`.
- Show me procedures with a `Wait` step with a `time` under 30 min.
- Show me procedures that include an `Adjust pH` step to a `target.pH` of 2.5.

3 Competency Question Analysis Report

CQ Analysis Report

Summary

Successfully loaded 405 procedures. Precalculation of reagent variant sets completed. Reagent-level quantities (amount, concentration) are recorded once per procedure as an aggregated total (Phase II), not separately for every step in which a reagent is used; the two amount/concentration CQs below are therefore evaluated against the procedure-level total rather than a single step.

- Procedures listing `Sodium Tungstate Dihydrate` in the main `reagents` list found 86 procedures.
- Procedures using `n-Bu`$_4$`NBr` in `reagents` or `procedure_steps` found 26 procedures.
- Procedures using `Water` as a reagent in a `Dissolve` step found 183 procedures.
- Procedures where a reagent `concentration` is greater than 1.0 mol/L found 32 procedures.
- Procedures where the total procedure-level amount of `Hydrochloric acid` exceeds 100 mL found 15 procedures.
- Procedures where the total procedure-level amount of `n-Bu`$_4$`NOH` exceeds 0.5 mol found 0 procedures.
- Procedures executed in more than 10 elementary steps found 120 procedures.
- Procedures with a `Stir` operation longer than 5 min found 85 procedures.

- Procedures where a `target.temperature.value` reached at least 80 °C found 121 procedures.
- Procedures that include a `Filter unit_operation` found 184 procedures.
- Procedures with a `Wait` step with a `time` under 30 min found 12 procedures.
- Procedures with an `Adjust pH` step to a `target.pH` of 2.5 found 0 procedures.

Analysis complete.

4 Independent Text-Generation Cross-Check

As an additional, independent check of corpus fidelity beyond the competency-question audit above, each of the 405 structured procedure records was used to regenerate a free-text synthesis paragraph, which was then automatically compared against both the structured JSON and the original extracted source text for that procedure (reagent identities and quantities, temperature/time/pH/atmosphere, operation order, and work-up conditions). This yielded 69 procedures whose generated text matched the source exactly (PASS), 181 with an ambiguous or incomplete match not amounting to a contradiction (REVIEW), and 155 with a genuine discrepancy (FAIL) – predominantly cases where a reagent quantity recorded once per procedure (see the aggregation note above) was attributed to the wrong step on regeneration, rather than an error in the underlying extracted data itself.

A repaired version of the 155 FAIL and select REVIEW records was subsequently produced – attaching each reagent's correct per-step quantity, temperature and timing directly from the original source text, with line-level provenance – and independently re-validated against that source text (0 remaining FAIL; residual REVIEW flags are limited to literal unit-normalisation differences, e.g. "2 h" vs. "120 min"). Two small schema defects introduced during this repair (a misplaced `target` field in 3 steps; 51 temperature/time values recorded as bare strings rather than the standard `{value, unit}` structure) were found and corrected before promotion. Because this repair changes step counts, unit-operation sequencing, and recorded temperatures for a subset of procedures – exactly the fields the clustering analysis in the main text's Machine Learning Methodology section is computed from – the clustering was re-run on the repaired corpus rather than left describing a superseded dataset. This repaired corpus is what is deposited and what every count, table, and figure in this work (including the cluster analysis) is derived from; the original, uncorrected extraction is archived separately alongside the deposited data for transparency.

Ablation: single-pass extraction vs. the decomposed pipeline

To provide a concrete, run data point against a simpler LLM-based extraction strategy, the same model (DeepSeek Reasoner, temperature 0) was prompted to extract a full structured record directly from the raw source text in a single pass, with no Phase I/II/III decomposition, for a fixed sample of the first 30 procedures by identifier (`SYN1–SYN63`). Each single-pass output was audited with the identical generate-then-cross-check methodology used above. The result should be reported plainly rather than framed to favour the pipeline presented in this work: on this sample, the single-pass baseline's first-pass extraction showed

1/30 FAIL (3%), compared with **12/30 FAIL (40%)** for the same 30 procedures in the original, pre-repair output of the decomposed pipeline. Inspection of the 12 pipeline failures confirms they are essentially all instances of the reagent-quantity aggregation issue described above: because Phase II is deliberately prompted to record one procedure-level total per reagent rather than a per-step amount, the decomposed pipeline structurally cannot get step-level quantity attribution right on first pass, whereas a single-pass extraction has no such artificial separation and tends to write step-level amounts directly.

This is not evidence that single-pass extraction is superior overall: the single-pass baseline was not run through an equivalent independent audit-and-repair cycle, so this is a first-pass-only comparison, not a comparison of final, deployed quality, and the deposited corpus (built via the decomposed pipeline plus repair) reaches 0/405 FAIL after that cycle, a claim not established here for the single-pass baseline at any scale. It is, however, informative and reported honestly rather than omitted: it confirms that the specific procedure-level aggregation boundary chosen in Phase II is the direct, mechanistic cause of the majority of first-pass discrepancies, and that a future pipeline revision prompting Phase II to record step-level amounts directly – effectively adopting what the single-pass baseline does implicitly – is a concrete, evidence-backed direction for reducing reliance on the post-hoc repair stage, rather than a hypothetical one.

5 Extraction Prompts

This section reproduces the system prompts used by the three LLM-driven stages of the extraction pipeline described in the Methodology section of the main text (Phase I text normalisation, Phase II reagent extraction, Phase III procedure structuring), exactly as implemented in the released codebase. Phase I is split into two sub-steps: an identification pass that isolates the target product name(s) named in a source text, and a rewrite pass that normalises the procedure into one action per line. Phase IV (data assembly) is a deterministic Python script with no LLM call and therefore has no associated prompt; it programmatically merges the Phase II reagent manifest and Phase III step list into the final JSON record.

Phase I(a) – Product Identification Prompt

```
You are an extractive agent.
Read the source and return final isolated product names that appear
    verbatim in the text.
Copy names exactly (including Unicode and hydrates).
Ignore intermediates, generic words like "product", and anion-only
    formulas (e.g., bare [X]^n- with no counterion/hydrate).

Return JSON ONLY:
{"products":["<name>", "..."]}

If none are present, return {"products":[]}.

[SOURCE TEXT]
```

```
{{source_text_here}}
```

Phase I(b) – Text Normalisation Prompt

```
You are an extractive technical writer.

TARGETS
- Write a procedure only for names in {{product_names}}, in that order.
- If {{product_names}} is empty, write one cleaned, self-contained
    procedure.

FORMAT (STRICT)
- One action per line; end with a period.
- Each line must start with exactly one of:
  Add, Heat, Cool, Stir, Filter, Wash, Dry, Dissolve, Evaporate,
    Crystallize, Precipitate, Adjust pH, Separate, Decant, Centrifuge,
    Extract, Transfer, Purge, Grind, Wait
- Keep original order except the two normalizations below.

ONLY ALLOWED NORMALIZATIONS
1) If a step uses "boiling <solvent>" and a numeric <VOL> is given for
    that solvent, insert just before it:
   Heat <VOL> <solvent> to boiling.
2) If a solid is added to a solvent not yet present and a numeric <VOL>
    is given, insert just before it:
   Add <VOL> <solvent>.

NEVER
- Invent reagents, volumes, temps, times, pH, or steps.
- Use Charge or Shake (use Add and Stir).
- Output analytics, headings, equations, or meta text.
- Output fragments (lines that aren't actions).

PRESERVE
- Exact quantities/units and reagent phrases (e.g., "concentrated acetic
    acid"); normalize only mL, g, degC.

MULTIPLE TARGETS
- Separate procedures with: ---PROCEDURE-BREAK---

OUTPUT
- Only the action lines (and the separator if needed).

[SOURCE TEXT]
{{source_text_here}}
```

Phase II – Reagent Extraction Prompt

```
You are a highly precise data extraction agent. Your only task is to read
     the provided [SOURCE TEXT] and extract a de-duplicated list of
    unique "reagents".

CRITICAL INSTRUCTIONS:
1. Identify Chemicals: Find every unique chemical substance used in the
    synthesis.
2. Use Canonical Names: You MUST use the most common chemical name or
    formula for the "name" key (e.g., "Sodium Tungstate Dihydrate", "
    Na2WO4.2H2O", "Water").
3. Aggregate Amounts: If the same chemical is used multiple times, sum
    its total amount.
4. Output Format: Your response MUST be a single JSON object with one key
    : "reagents". Each item in the list must be a JSON object with "name"
     and "amount" keys.

GROUNDING RULE: Your output must ONLY contain chemicals explicitly found
    in the [SOURCE TEXT]. Do not add any chemicals from your external
    knowledge.

---
[SOURCE TEXT]:
{{source_text_here}}
```

Phase III – Procedure Structuring Prompt

```
You are an extractive chemist
Output {"procedure_steps":[...]} only
No assumptions. Omit any key that is not stated in the source
unit_operation must be one of ['Charge','Add','Stir','Heat','Cool','Wait
    ','Filter','Wash','Dry','Dissolve','Evaporate','Crystallize','
    Precipitate','Adjust pH','Separate','Decant','Centrifuge','Extract','
    Transfer','Purge','Grind']
When new reagents are introduced include "reagent_usage":[{"reagent_id":"
    R#"}] using only IDs from the list
Map every reagent ID to the first step where it enters the process
Use {"time":{"value":N,"unit":"min"}} or {"time":{"end_condition":"until
    ..."}} for durations only
Use {"target":{"temperature":{"value":N,"unit":"degC"}}} for explicit
    target temperatures only
Never invent placeholders like Solution A or boiling water or Na2WO4
    solution. Refer to reagents only via reagent_id in reagent_usage
For each step include "source_line_idx": the 1-based index of the
    sentence in CLEANED TEXT that justifies this step

[LIST OF REAGENTS]
{{reagents_list_here}}

[CLEANED TEXT]
```

```
{{source_text_here}}
```

6 Synthetic Data Overview

Table S1: *Addenda counts*

Addenda	Count
Mo	75
V	40
W	199
Mo/V	28
Mo/V/W	1
Mo/W	10
V/W	30

Table S2: *Most frequently used solvents*

Solvent	No. of procedures
Water	309
Diethyl Ether	53
Ethanol	45
Acetonitrile	32
Methanol	19
Acetone	15
Dichloromethane	6
DMF	3
n-Butanol	2
Benzene	1

Prior to UMAP embedding and HDBSCAN clustering (main text, Machine Learning Analysis section), missing values in the quantitative features (mean/standard-deviation temperature, mean/standard-deviation time, reagent and step counts) were imputed using

Table S3: *Most frequently used unit operations*

Unit step	Total instances
Add	1200
Dissolve	367
Heat	307
Filter	276
Stir	236
Cool	219
Wash	214
Dry	167
Wait	166
Evaporate	142

the column median across the corpus, so that procedures with an incompletely reported temperature or time profile could still be embedded. Cluster-level mean temperatures reported below and in the main text therefore reflect this imputation and should be read as descriptive of each cluster's dominant thermal regime rather than as a strict average of explicitly reported values alone.

Figure S1 shows the same UMAP projection used in the main text (Figure 5), here coloured by addenda set rather than by cluster membership or temperature. It is provided so that the addenda-composition fractions quoted in Table S4 (e.g. Cluster 2, Cluster 6) can be inspected directly against the underlying data.

Table S4: *Summary and interpretation of the 13 identified synthesis clusters ($N = 391$). Fourteen outliers were excluded.*

Cluster	N	Key Defining Features (Measures)	Interpretation
7	74	• **Addenda:** Other (50/74) / W (11/74) • **Heteroatom:** P (10/74), B (2/74) • **Solvent:** Aqueous (74/74) • **Temp:** 73.1 ± 7.1 °C • **Reagent:** Water, Sodium Tungstate Dihydrate, Guanidine Hydrochloride, Hydrochloric Acid • **Operation:** add, dissolve, add add	Largest cluster: an exclusively aqueous, tungstate-leaning family defined by dissolution and crystallisation steps. It carries the largest phosphorus-heteroatom presence of any cluster (10/74).

Table S4 – continued from previous page

Cluster	N	Key Defining Features (Measures)	Interpretation
2	62	• **Addenda:** Other (48/62) / Mo (6/62) • **Heteroatom:** Ge (3/62), B (2/62) • **Solvent:** Other (62/62) • **Temp:** 69.7 ± 16.4 °C • **Reagent:** Potassium Hydroxide, Methanol, Cesium Chloride, Hydrochloric Acid • **Operation:** add, add add, evaporate	Second-largest cluster: an exclusively non-aqueous, non-standard ("Other"-classified) solvent family built on methanolic KOH, distinguished by an evaporation step. It is molybdate-leaning among the assigned addenda.
10	38	• **Addenda:** Other (16/38) / W (9/38) • **Heteroatom:** P (4/38), Si (3/38) • **Solvent:** Aqueous (38/38) • **Temp:** 73.2 ± 7.0 °C • **Reagent:** Water, Potassium Chloride, Sodium Molybdate Dihydrate, Hydrochloric Acid • **Operation:** add, add add, dissolve	Aqueous molybdate/tungstate family defined by long wait or ageing periods and a varied heteroatom profile, with P, Si, Ge, and B all represented.
3	36	• **Addenda:** Other (15/36) / Mo (9/36) • **Heteroatom:** P (3/36), Si (1/36) • **Solvent:** Aqueous-Alcohol (30/36) • **Temp:** 69.3 ± 18.9 °C • **Reagent:** Ethanol, Diethyl Ether, Water, Potassium Chloride • **Operation:** add, wash, add add	Aqueous-alcohol wash/dry workup family (30/36 Aqueous-Alcohol), defined by repeated wash and filter operations around a Dawson-type lacunary tungstophosphate.
5	33	• **Addenda:** Other (20/33) / W (7/33) • **Heteroatom:** P (2/33), Si (2/33) • **Solvent:** Aqueous (28/33) • **Temp:** 13.6 ± 14.1 °C • **Reagent:** Water, Hydrochloric Acid, Sodium Hydroxide, Acetone • **Operation:** add, dissolve, filter	A distinct cold synthesis family (mean 13.6°C), defined by a dissolve→filter→dry sequence and a hydrochloric acid/sodium hydroxide reagent profile.
0	30	• **Addenda:** Other (20/30) / W (4/30) • **Heteroatom:** Si (3/30), P (2/30) • **Solvent:** Aqueous (24/30) • **Temp:** 167.0 ± 25.9 °C • **Reagent:** Water, Sodium Tungstate Dihydrate, Copper(II) Nitrate Trihydrate, Tetramethylammonium Hydroxide • **Operation:** add, add add, transfer heat	Hydrothermal-type family (mean 167.0°C) with long, variable reaction times (mean ∼46 h). It features copper(II)/trimesic-acid ligand chemistry alongside conventional tungstate reagents.

Table S4 – continued from previous page

Cluster	N	Key Defining Features (Measures)	Interpretation
1	28	• **Addenda:** Other (17/28) / V (6/28) • **Heteroatom:** Si (3/28), P (2/28) • **Solvent:** Aqueous (18/28) • **Temp:** 43.8 ± 14.3 °C • **Reagent:** Water, Potassium Chloride, Sodium Metavanadate, $Na_8H[PW_9O_{34}]$ • **Operation:** add, cool, add add	Vanadate/lacunary-phosphotungstate family with moderate, variable temperature (43.8 ± 14.3°C) and frequent heat/cool cycling.
4	22	• **Addenda:** Other (14/22) / Mo (4/22) • **Heteroatom:** P (3/22) • **Solvent:** Non-Aqueous (21/22) • **Temp:** 76.4 ± 8.7 °C • **Reagent:** Acetonitrile, Pyridine, Ethanol, Tetrabutylammonium Bromide • **Operation:** add, add add, wash	Non-aqueous organic-solvent family (acetonitrile/pyridine) with repeated wash steps. It is nearly exclusively non-aqueous (21/22).
9	18	• **Addenda:** Other (8/18) / W (8/18) • **Heteroatom:** P (7/18) • **Solvent:** Aqueous (18/18) • **Temp:** 79.4 ± 8.9 °C • **Reagent:** Hydrochloric Acid, Water, Potassium Chloride, Sodium Tungstate Dihydrate • **Operation:** add, adjust, adjust pH	pH-adjustment-defined phosphotungstate family (7/18 phosphorus-heteroatom) with fast reaction times (mean 6.5 h) and the hottest mean temperature (79.4°C) among the non-hydrothermal clusters.
8	15	• **Addenda:** Other (6/15) / W (5/15) • **Heteroatom:** Ge (2/15) • **Solvent:** Aqueous (15/15) • **Temp:** 77.3 ± 12.9 °C • **Reagent:** Water, Potassium Chloride, Sodium Hydroxide, Hydrochloric Acid • **Operation:** add, dissolve, add add	Hot, exclusively aqueous tungstate family with germanium-heteroatom representation, including germanium dioxide among defining reagents. It is defined by dissolve/stir/heat sequencing.
12	15	• **Addenda:** W (7/15) / Mo+V (3/15) • **Heteroatom:** P (5/15), Si (2/15) • **Solvent:** Aqueous (15/15) • **Temp:** 62.0 ± 23.6 °C • **Reagent:** Water, Sodium Metavanadate, Sodium Tungstate, Sulfuric Acid • **Operation:** add, filter, dissolve	The only cluster where tungsten is the single most common assigned addenda class (7/15). It is associated with sulfuric acid and phosphorus-heteroatom chemistry and has the second-highest mean step count of any cluster (19.8).

Table S4 – continued from previous page

Cluster	N	Key Defining Features (Measures)	Interpretation
6	10	• **Addenda:** Other (5/10) / Mo+V (2/10) • **Solvent:** Aqueous (6/10) • **Temp:** 73.2 ± 36.7 °C • **Reagent:** Water, Tetramethylammonium Chloride, Sodium Molybdate Dihydrate, Sodium Chloride • **Operation:** add, dissolve dissolve, dissolve	Smallest-but-one cluster, defined by the longest mean reaction or ageing time of any cluster by a wide margin (∼326 h, i.e. several weeks) around a tetramethylammonium/molybdate system.
11	10	• **Addenda:** Other (5/10) / W (4/10) • **Heteroatom:** P (2/10) • **Solvent:** Non-Aqueous (9/10) • **Temp:** 67.2 ± 13.9 °C • **Reagent:** Diethyl Ether, Acetonitrile, Dichloromethane, Hydrochloric Acid • **Operation:** wash, add, wash wash	Non-aqueous, wash-intensive family with the highest mean step count of any cluster (21.4) and mean reagent count (8.4), built around diethyl ether/acetonitrile/dichloromethane extraction chemistry.

Clustering Parameter Sensitivity

The number of clusters and outliers reported above depends on the HDBSCAN `min_cluster_size` parameter (10 in the reported analysis, with all other pipeline parameters held fixed). To characterise this dependence directly rather than assert robustness, we reran clustering on the identical 1089-feature representation while sweeping this parameter:

`min_cluster_size`	5	8	**10**	12	15	20
Clusters found	25	13	**13**	11	11	9
Outliers	24	14	**14**	17	17	12

The cluster count is not a fixed, parameter-independent property of the data: it ranges from 9 to 25 across this range, and a reader should not treat "13" as an intrinsic count of synthesis families. It is, however, not arbitrary either: the result is stable at 13 clusters across `min_cluster_size` = 8–10 and stable again at 11 clusters across 12–15, indicating genuine plateaus rather than a parameter value picked to hit a target number, with smaller, more granular sub-families progressively merging as the size threshold is raised. `min_cluster_size` = 10 was chosen prior to this sweep as a conventional default for a corpus of this size (avoiding both the very fine fragmentation seen at 5 and the coarse merging seen at 20) and falls within the first stable plateau.

Of the 405 procedures in the deposited JSON corpus, a starting addenda source could be identified directly from the extracted reagent data for 256 procedure-slots (count-

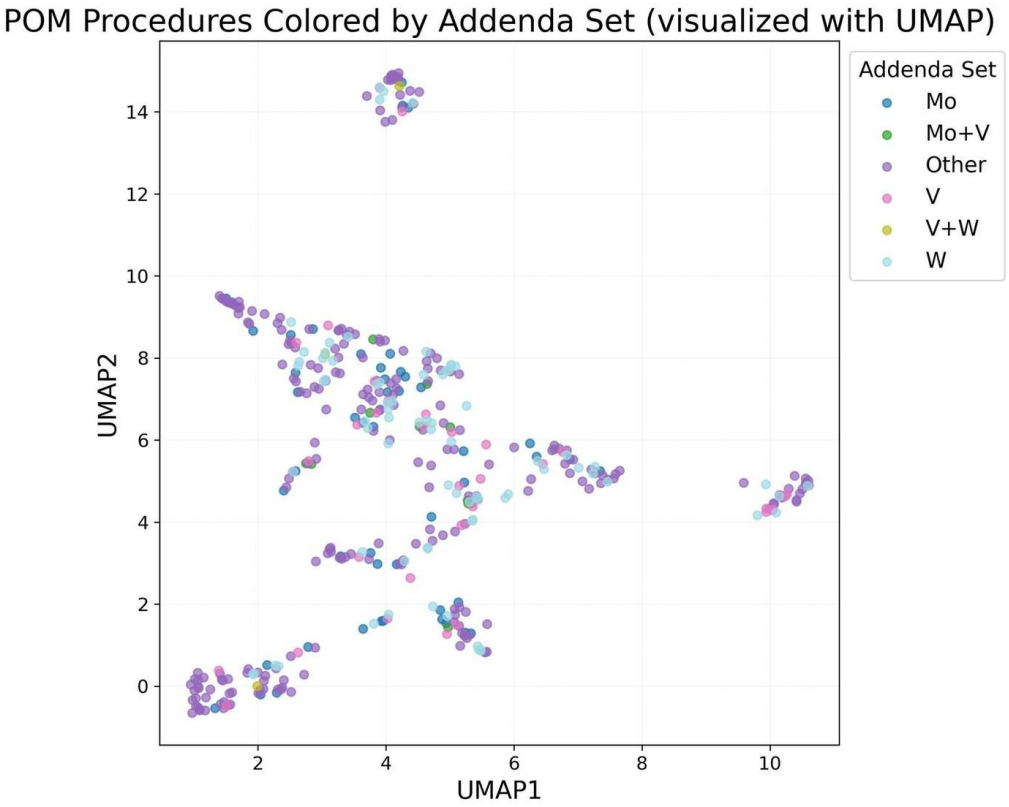


Figure S1: *UMAP projection of the 405-procedure corpus, coloured by addenda set (Mo, V, W, and their binary/ternary combinations, or "Other"). Compare with Figure 5 in the main text and the addenda fractions listed in Table S4.*

ing mixed-addenda procedures once per addenda source), grouped by named starting salt in Table S5 below. A further 153 procedures start from an already complex, preformed POM precursor (a heteropolyacid, lacunary salt, or similarly complex starting compound) identifiable in the reagent list, but for which the earlier procedure that produced that specific precursor is not cross-linked in this table. The remaining procedures are each accounted for explicitly rather than left in an undifferentiated *unclassified* bucket: 4 use a non-classical addenda metal (antimony or iron) verified directly from the reagent data; 2 are heteroatom-doping steps with no addenda-bearing reagent of their own; 17 are counterion-exchange, recrystallisation or workup steps with no addenda-bearing reagent; and 1 procedure has no reagents recorded at all. No product-family classification (e.g. Keggin, Wells–Dawson) or precursor→product synthetic lineage is asserted beyond what is directly verifiable from the deposited reagent data itself; every identifier below is present in the deposited 405-procedure corpus.

Table S5: *Starting addenda source, verified directly from deposited reagent data, for the 405 procedures in the deposited JSON corpus*

Starting addenda source	N	Procedure SYN IDs

Na2WO4	81	SYN1, SYN2, SYN3, SYN6, SYN7, SYN8-01, SYN21, SYN23, SYN31, SYN60, SYN63, SYN64, SYN66-01, SYN72-1, SYN72-2, SYN83, SYN84, SYN88, SYN91, SYN92, SYN109, SYN110, SYN112, SYN113-1, SYN113-2, SYN114, SYN115, SYN116, SYN123, SYN135, SYN136, SYN138-01, SYN148, SYN162, SYN164, SYN165, SYN166, SYN170, SYN171, SYN173, SYN194-01, SYN195, SYN199-01, SYN200, SYN202, SYN204, SYN207, SYN230, SYN263, SYN271-2, SYN271-4, SYN288, SYN289, SYN290, SYN291, SYN292, SYN296, SYN297-1, SYN297-2, SYN299, SYN417, SYN418, SYN419, SYN422, SYN423, SYN432, SYN433-01, SYN434, SYN436, SYN437, SYN438, SYN451, SYN457, SYN461, SYN487, SYN541, SYN546, SYN547, SYN548, SYN549, SYN578
K2WO4	1	SYN452-01
Cs2WO4	1	SYN413
Na2WO4-metatungstate variants	1	SYN569
H2WO4 / WO3 (tungstic acid/oxide)	9	SYN85, SYN93, SYN154-2, SYN154-6-01, SYN195, SYN269, SYN413, SYN414, SYN466
Na2MoO4	56	SYN45, SYN46, SYN50, SYN51, SYN53-01, SYN54, SYN67, SYN98-1, SYN99-1, SYN99-2, SYN99-3, SYN142, SYN144, SYN147, SYN149, SYN150-1-01, SYN150-2, SYN151, SYN153, SYN178, SYN179, SYN181, SYN208, SYN221, SYN222, SYN225, SYN226, SYN234, SYN245, SYN254, SYN256-1, SYN256-2, SYN260, SYN262, SYN265, SYN386, SYN387, SYN393, SYN395, SYN396, SYN399, SYN400, SYN404, SYN405, SYN406, SYN407, SYN408, SYN411, SYN412, SYN441, SYN443, SYN444, SYN447, SYN449, SYN450, SYN570
(NH4)6Mo7O24 (ammonium heptamolybdate)	13	SYN96, SYN209, SYN214-2, SYN214-3, SYN216-1, SYN216-2, SYN217, SYN218-1, SYN231, SYN233-1, SYN236-1, SYN261, SYN397
MoO3 (molybdenum oxide)	8	SYN257, SYN258, SYN259, SYN389, SYN390, SYN439, SYN440, SYN446
H2MoO4 (molybdic acid)	2	SYN145-1, SYN253
NaVO3	22	SYN69, SYN70, SYN71-2, SYN73, SYN76-1, SYN76-2, SYN93, SYN95-1-01, SYN149, SYN150-1-01, SYN150-2, SYN151, SYN157, SYN187, SYN188, SYN310, SYN323, SYN324, SYN443, SYN444, SYN450, SYN451
CsVO3	1	SYN337
NH4VO3	13	SYN181, SYN207, SYN208, SYN318, SYN330, SYN333, SYN335, SYN339, SYN343, SYN439, SYN440, SYN441, SYN446
KVO3	5	SYN313, SYN319, SYN325, SYN347, SYN452-01
Na3VO4	2	SYN312, SYN350
V2O5	10	SYN317, SYN327, SYN329, SYN331, SYN340-1, SYN340-2, SYN342, SYN393, SYN442, SYN453
VOSO4	11	SYN74, SYN75, SYN77, SYN114, SYN122, SYN316, SYN341, SYN447, SYN449, SYN570, SYN578
Other vanadium(V/IV) source (halide, acac, stock solution)	9	SYN72-1, SYN72-2, SYN78, SYN141, SYN142, SYN311, SYN328, SYN336-3, SYN455

Nb2O5	8	SYN58, SYN59, SYN156, SYN356, SYN357, SYN358-01, SYN359, SYN361
Other niobium source (halide, heptaniobate)	2	SYN79, SYN370
Ta2O5	1	SYN371
Pre-formed POM precursor (heteropolyacid, lacunary or other complex starting compound; identity verified from reagent data, but its own synthetic origin is not cross-linked in this table)	153	SYN16, SYN20, SYN26, SYN27, SYN29, SYN39, SYN42, SYN43, SYN44, SYN62, SYN71-1, SYN80, SYN98-2, SYN103, SYN105, SYN106, SYN111, SYN119-01, SYN120, SYN121, SYN125, SYN131, SYN132, SYN133, SYN152-2-01, SYN154-1, SYN154-3, SYN154-4, SYN154-5, SYN155, SYN159-2, SYN161, SYN172, SYN182, SYN183, SYN184, SYN189, SYN190, SYN191, SYN192, SYN193, SYN196, SYN197, SYN198-01, SYN205, SYN211, SYN214-1, SYN224, SYN233-2, SYN238, SYN239, SYN242, SYN243, SYN246, SYN247, SYN248, SYN249, SYN250, SYN266, SYN276, SYN277, SYN278, SYN279, SYN280, SYN283, SYN284, SYN285, SYN286, SYN287, SYN293, SYN294, SYN298, SYN300, SYN301, SYN302, SYN305-1, SYN308, SYN321, SYN322, SYN326, SYN345, SYN349, SYN351, SYN352, SYN355, SYN362, SYN366-1, SYN366-2-01, SYN367, SYN368, SYN369, SYN372, SYN382, SYN385, SYN391, SYN392, SYN401, SYN402, SYN403, SYN421, SYN429, SYN456, SYN462, SYN464, SYN467-01, SYN468, SYN472, SYN474, SYN476, SYN480, SYN481-01, SYN484, SYN485, SYN486, SYN488, SYN489, SYN491, SYN492, SYN495, SYN496, SYN498, SYN499, SYN501, SYN502, SYN505, SYN536, SYN537, SYN538, SYN539, SYN542, SYN543-1, SYN543-2, SYN545, SYN550, SYN551, SYN553, SYN554, SYN555, SYN556, SYN557, SYN558, SYN559, SYN560, SYN561, SYN562, SYN563, SYN552, SYN564-1, SYN564-2, SYN565, SYN566, SYN567, SYN572
Non-classical addenda system verified from reagent data (Sb (antimonate) system)	2	SYN579, SYN580
Non-classical addenda system verified from reagent data (Sb (arylantimonic acid) system)	1	SYN581
Non-classical addenda system verified from reagent data (Fe-fluoride system)	1	SYN582
Heteroatom-doping step only (no W/Mo/V/Nb/Ta addenda-bearing reagent in this entry)	2	SYN215, SYN465
Counterion-exchange, recrystallisation or workup step (no addenda-bearing reagent in this entry)	17	SYN61, SYN86, SYN117, SYN126, SYN127, SYN130, SYN159-1, SYN167, SYN281, SYN338, SYN348, SYN353, SYN426, SYN427, SYN428, SYN460, SYN477
No reagents recorded for this procedure	1	SYN118

References

[1] J. C. Bailar, Jr., H. S. Booth, and M. Grennert. Phosphotungstic acid. In *Inorganic Syntheses*, volume 1, page 132 to 133. John Wiley and Sons, 1939. doi:10.1002/9780470132326.ch49.

[2] A. Barba-Bon, N. I. Gumerova, E. Tanuhadi, M. Ashjari, Y. Chen, A. Rompel, and W. M. Nau. All-inorganic polyoxometalates act as superchaotropic membrane carriers. *Adv. Mater.*, 36(1):2309219, 2024. doi:10.1002/adma.202309219.

[3] B. S. Bassil, M. H. Dickman, B. V. D. Kammer, and U. Kortz. The monolanthanide containing silicotungstates ln beta2 siw11o39 2 13 ln equals la ce sm eu gd tb yb lu a synthetic and structural investigation. *Inorg. Chem.*, 46(7):2452 to 2458, 2007. doi:10.1021/ic061737d.

[4] A. Bijelic and A. Rompel. Ten good reasons for the use of the tellurium-centered anderson–evans polyoxotungstate in protein crystallography. *Acc. Chem. Res.*, 50 (6):1441–1448, 2017. doi:10.1021/acs.accounts.7b00109.

[5] A. Bijelic, M. Aureliano, and A. Rompel. The antibacterial activity of polyoxometalates: structures, antibiotic effects and future perspectives. *Chem. Commun.*, 54(10): 1153–1169, 2018.

[6] A. Bijelic, M. Aureliano, and A. Rompel. Polyoxometalates as potential next-generation metallodrugs in the combat against cancer. *Angew. Chem. Int. Ed.*, 58 (10):2980–2999, 2019. doi:10.1002/anie.201812022.

[7] J. Boeyens, G. McDougall, and J. van R. Smit. Crystallographic study of the ammonium/potassium 12-molybdophosphate ion-exchange system. *J. Solid State Chem.*, 18:191–199, 1976. doi:10.1016/0022-4596(76)90125-4.

[8] D. A. Boiko, R. MacKnight, B. Kline, and G. Gomes. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023. doi:10.1038/s41586-023-06792-0.

[9] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.*, 6(5):525–535, 2024. doi:10.1038/s42256-024-00832-8.

[10] C. Brevard and M. M. Tourné. Tungsten 183 nmr a complete and unequivocal assignment of the tungsten tungsten connectivities in heteropolytungstates via two dimensional tungsten 183 nmr. *J. Am. Chem. Soc.*, 105(24):7059 to 7063, 1983. doi:10.1021/ja00362a008.

[11] D. H. Brown and J. A. Mair. Heteropolytungstic acids and heteropolytungstates. part iii. 12 tungstozincic acid and its salts and some 12 tungsto 3 zincates. *J. Chem. Soc.*, page 2597 to 2599, 1958. doi:10.1039/jr9580002597.

[12] L.-X. Cai, Y.-H. Hu, L.-P. Zhou, P.-M. Cheng, X.-Q. Guo, Y.-T. Chan, and Q.-F. Sun. Polyoxometalate condensation and transformation mediated by adaptive coordination–assembled molecular flasks. *Chem. Sci.*, 16:7956–7962, Mar. 2025. doi:10.1039/D4SC08729A.

[13] R. J. G. B. Campello, D. Moulavi, and J. Sander. Density based clustering based on hierarchical density estimates. In *Pacific Asia Conference on Knowledge Discovery and Data Mining*, pages 160–172. Springer, 2013. doi:10.1007/978-3-642-37456-2_14.

[14] R. J. G. B. Campello, D. Moulavi, A. Zimek, and J. Sander. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Trans. Knowl. Discov. Data*, 10(1):5, 2015. doi:10.1145/2733381.

[15] J. J. Cowan, A. J. Bailey, R. A. Heintz, B. T. Do, K. I. Hardcastle, C. L. Hill, and I. A. Weinstock. Formation isomerization and derivatization of keggin tungstoaluminates. *Inorg. Chem.*, 40(26):6666 to 6675, 2001. doi:10.1021/ic0106120.

[16] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson, and A. Jain. Structured information extraction from scientific text with large language models. *Nat. Commun.*, 15(1):1418, 2024. doi:10.1038/s41467-024-45563-x.

[17] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson, and A. Jain. Structured information extraction from complex scientific text with fine-tuned large language models. *Nat. Commun.*, 15(1):1418, 2024. doi:10.1038/s41467-024-45563-x.

[18] C. R. Deltcheff, M. Fournier, R. Franck, and R. Thouvenot. Vibrational investigations of polyoxometalates. 2. evidence for anion anion interactions in molybdenum vi and tungsten vi compounds related to the keggin structure. *Inorg. Chem.*, 22(2): 207 to 216, 1983. doi:10.1021/ic00144a006.

[19] L.-Y. Feng, Y. Xu, Y.-H. Wang, C.-W. Hu, Y.-J. Qi, and E.-B. Wang. Synthesis and crystal structure of the first waugh-type isopolyoxomolybdate $na_4mo_{10}o_{32} \cdot 8h_2o$. *J. Mol. Struct.*, 645:231–234, 2003. doi:10.1016/S0022-2860(02)00582-3.

[20] R. G. Finke and M. W. Droege. Trisubstituted heteropolytungstates as soluble metal oxide analogs 1 the preparation characterization and reactions of organic solvent soluble forms of si2w18nb6o77(8-) siw9nb3o40(7-) and the siw9nb3o40(7-) supported organometallic complex [(c5me5)rhsiw9nb3o40](5-). *J. Am. Chem. Soc.*, 106(23): 7274–7277, 1984. doi:10.1021/ja00335a086.

[21] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang. Retrieval-Augmented Generation for Large Language Models: A Survey, 2024. URL https://arxiv.org/abs/2312.10997.

[22] C. R. Graham and R. G. Finke. The classic wells- dawson polyoxometalate, k6 [α-p2w18o62]· 14h2o. answering an 88 year-old question: what is its preferred, optimum synthesis? *Inorg. Chem.*, 47(9):3679–3686, 2008. doi:10.1021/ic702295y.

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

[24] N. I. Gumerova and A. Rompel. Polyoxometalates in solution: speciation under spotlight. *Chem. Soc. Rev*, 49(21):7568–7601, 2020. doi:10.1039/D0CS00392A.

[25] N. I. Gumerova and A. Rompel. Interweaving disciplines to advance chemistry: Applying polyoxometalates in biology. *Inorg. Chem.*, 60(9):6109–6114, 2021. doi:10.1021/acs.inorgchem.1c00125.

[26] G. Hervé, A. Tézé, and R. Contant. General principles of the synthesis of polyoxometalates in aqueous solution. In J. J. Borrás-Almenar, E. Coronado, A. Müller, and M. T. Pope, editors, *Polyoxometalate Molecular Science*, pages 33–54. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003.

[27] D. M. Jessop, S. E. Adams, E. L. Willighagen, L. Hawizy, and P. Murray-Rust. OSCAR4: a flexible architecture for chemical text-mining. *Journal of Cheminformatics*, 3(1):1–12, 2011. doi:10.1186/1758-2946-3-41.

[28] A. Kondinski. Metal–metal bonds in polyoxometalate chemistry. *Nanoscale*, 13 (32):13574–13592, 2021. doi:10.1039/D1NR03115H.

[29] A. Kondinski, A. Banerjee, and S. S. Mal. Hervé- and krebs-type magnetic polyoxometalate dimers. *Magnetochemistry*, 8(9):96, 2022. doi:10.3390/magnetochemistry8090096.

[30] A. Kondinski, J. Bai, S. Mosbach, J. Akroyd, and M. Kraft. Knowledge engineering in chemistry: From expert systems to agents of creation. *Acc. Chem. Res.*, 56(2): 128–139, 2023. doi:10.1021/acs.accounts.2c00617.

[31] A. Kondinski, N. Gumerova, and A. Rompel. A curated polyoxometalate formula dataset and its application to data-driven cluster chemistry. *Data*, 9(11):124, 2024. doi:10.3390/data9110124.

[32] A. Kondinski, N. I. Gumerova, A. Rompel, P. Falcaro, and T. Schreck. Data-driven polyoxometalate chemistry. *Chem. Eur. J.*, 31(55):e01528, 2025. doi:10.1002/chem.202501528. URL https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/chem.202501528.

[33] O. Kononova, H. Huo, T. He, Z. Rong, T. Botari, W. Sun, V. Tshitoyan, and G. Ceder. Text-mined dataset of inorganic materials synthesis recipes. *Sci. Data*, 6(1):203, 2019. doi:10.1038/s41597-019-0224-1.

[34] R. Liu and C. Streb. Polyoxometalate-single atom catalysts (pom-sacs) in energy research and catalysis. *Adv. Energy Mater.*, 11(25):2101120, 2021. doi:10.1002/aenm.202101120.

[35] D.-L. Long, E. Burkholder, and L. Cronin. Polyoxometalate clusters, nanostructures and materials: From self-assembly to designer materials and devices. *Chem. Soc. Rev.*, 36(1):105–121, 2007. doi:10.1039/B502666K.

[36] K. Y. Matsumoto and Y. Sasaki. The crystal structure of α-$k_8siw_{11}o_{39}\cdot 13h_2o$. *Bull. Chem. Soc. Jpn.*, 49(1):156–158, 1976. doi:10.1246/bcsj.49.156.

[37] L. McInnes, J. Healy, and S. Astels. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017. doi:10.21105/joss.00205.

[38] L. McInnes, J. Healy, N. Saul, and L. Grossberger. UMAP: Uniform manifold approximation and projection. *J. Open Source Softw.*, 3(29):861, 2018. doi:10.21105/joss.00861.

[39] H. N. Miras, D. Stone, D. L. Long, E. J. L. McInnes, P. Kogerler, and L. Cronin. Exploring the structure and properties of transition metal templated VM17(VO4)2 dawson like capsules. *Inorg. Chem.*, 50(17):8384 to 8391, 2011. doi:10.1021/ic200943s.

[40] H. N. Miras, D. Stone, D.-L. Long, E. J. L. McInnes, P. Kögerler, and L. Cronin. Exploring the structure and properties of transition metal templated $\{VM_{17}(VO_4)_2\}$ dawson-like capsules. *Inorg. Chem.*, 50(17):8384–8391, 2011. doi:10.1021/ic200943s.

[41] A. Misra, K. Kozma, C. Streb, and M. Nyman. Beyond charge balance: Countercations in polyoxometalate chemistry. *Angew. Chem. Int. Ed.*, 59(2):596–612, 2020. doi:10.1002/anie.201905381.

[42] 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.

[43] E. Radkov and R. H. Beer. High yield synthesis of mixed metal keggin polyoxoanions in non aqueous solvents preparation of (n-bu4n)4[pmw11o40] (m = v nb ta). *Polyhedron*, 14(15-16):2139–2143, 1995. doi:10.1016/0277-5387(95)00026-O.

[44] E. V. Radkov and R. H. Beer. High-yield synthesis of mixed-metal keggin polyoxoanions in non-aqueous solvents: preparation of $(n\text{-}Bu_4N)_4[PMW_{11}O_{40}]$ (m = v, nb, ta). *Polyhedron*, 14(15–16):2139–2143, 1995. doi:10.1016/0277-5387(95)00026-O.

[45] S. D. Rihm, F. Saluz, A. Kondinski, J. Bai, P. W. V. Butler, S. Mosbach, J. Akroyd, and M. Kraft. Extraction of chemical synthesis information using the world avatar. *Digit. Discov.*, 4:2893–2909, 2025. doi:10.1039/d5dd00183h.

[46] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM.*, 18(11):613–620, 1975. doi:10.1145/361219.361220.

[47] Y. Sari M A and M. D. Situngkir. Thermal stability of polyoxometalate compound of keggin k8[2-siw11o39]·nh2o supported with sio2. *Sci. Technol. Indones.*, 1(1): 8–15, 2016. doi:10.26554/sti.2016.1.1.8-15.

[48] M. Schilling-Wilhelmi, M. Ríos-García, S. Shabih, M. V. Gil, S. Miret, C. T. Koch, J. A. Márquez, and K. M. Jablonka. From Text to Insight: Large Language Models for Materials Science Data Extraction. *Chem. Soc. Rev.*, 54(3):1125–1150, 2025. doi:10.1039/d4cs00913d.

[49] V. Shivaiah and S. K. Das. Polyoxometalate supported transition metal complexes and their charge complementarity synthesis and characterization of [m(oh)_6mo_6o_18{Cu(phen)(H_2O)_2}_2] [m(oh)_6mo_6o_18{Cu(phen)(H_2O)Cl}_2]·5h_2o (m = al^{3+}, cr^{3+}). *Inorg. Chem.*, 44(24):8846–8854, 2005. doi:10.1021/ic050830j.

[50] M. N. Sokolov, S. A. Adonin, D. A. Mainichev, C. Vicent, N. F. Zakharchuk, A. M. Danilenko, and V. P. Fedin. Synthesis and characterization of [pw11o39ir(h2o)]4 successful incorporation of ir into polyoxometalate framework and study of the substitutional lability at the iriii site. *Chem. Commun.*, 47:7833 to 7835, 2011. doi:10.1039/c1cc10912j.

[51] M. C. Swain and J. M. Cole. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *Journal of Chemical Information and Modeling*, 56(10):1894–1904, 2016. doi:10.1021/acs.jcim.6b00207.

[52] A. Tezé, G. Hervé, R. G. Finke, and D. K. Lyon. Alpha beta and gamma dodecatungstosilicic acids isomers and related lacunary compounds. In *Inorganic Syntheses*, volume 27, page 85 to 96. John Wiley and Sons, 1990. doi:10.1002/9780470132586.ch16.

[53] I. A. Weinstock, J. J. Cowan, E. M. G. Barbuzzi, H. Zeng, and C. L. Hill. Equilibria between alpha and beta isomers of keggin heteropolytungstates. *J. Am. Chem. Soc.*, 121(19):4608 to 4617, 1999. doi:10.1021/ja982908j.

[54] I. A. Weinstock, J. J. Cowan, E. M. G. Barbuzzi, H. Zeng, and C. L. Hill. Equilibria between α and β isomers of keggin heteropolytungstates. *J. Am. Chem. Soc.*, 121 (19):4608–4617, 1999. doi:10.1021/ja982908j.

[55] Z. Zheng, O. Zhang, C. Borgs, J. T. Chayes, and O. M. Yaghi. Chatgpt chemistry assistant for text mining and the prediction of MOF synthesis. *J. Am. Chem. Soc.*, 145(32):18048–18062, 2023. doi:10.1021/jacs.3c05819.