Preprint CheMIn Technical Reports ISSN 3061-1075

Digitalisation of Inorganic Chemistry with LLMs

Boshko Koloski [1], Senja Pollak [1], Sašo Džeroski [1], Aleksandar Kondinski [2]

released: 24 June 2026

[1] Department of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana, Slovenia

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

Preprint No. 4



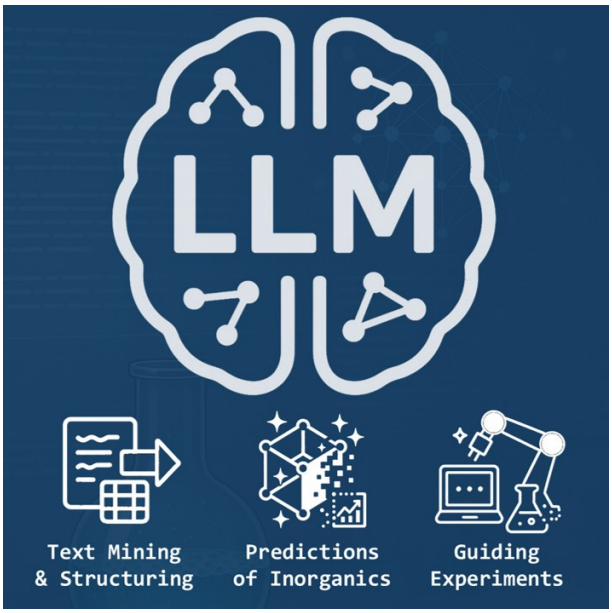
Keywords: large language models, inorganic chemistry, text mining, materials informatics, autonomous experimentation

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

Over the past few years, large language models have become technologically ubiquitous and now offer a powerful route to accelerate discoveries in chemistry. In this article, we highlight current impactful applications of large language models in inorganic chemistry, from smart text mining of the inorganic literature through the proposal and discovery of new materials to real-time experimentation. We also discuss ongoing developments and their potential future impact on the field.



Highlights

- LLMs extract structured, machine-readable data from inorganic literature.
- Language models propose new compositions, structures, and synthesis routes.
- LLM agents guide closed-loop, autonomous experimentation.
- Benchmarks show LLMs complement, not replace, physics-based models.

Contents

1 Introduction 3

2 LLM-driven text mining and structuring 4

3 LLM-based predictions for inorganics 6

4 LLMs supporting real-time guidance 8

5 Summary and Outlook 10

References 12

1 Introduction

Artificial intelligence (AI) encompasses all techniques that enable machines to exhibit intelligent behaviour.[30] Machine learning (ML) is the data-driven branch of AI that is based on the development of statistical regularities directly from data.[30] Deep learning (DL), a sub-branch of ML that employs deep neural networks, now dominates AI applications in perception (*i.e.* visual analytics) and natural language processing (NLP).[9] Namely, its self-supervised implementation produces foundation models, which are very large networks pre-trained on broad, often multimodal corpora and later adapted to diverse downstream problems.[9, 30, 35] Large language models (LLMs) are a text-centred family of foundation models with nearly every state-of-the-art LLM relying on the transformer architecture, whose self-attention mechanism, as introduced by Vaswani *et al.*, has been shown to scale efficiently to billions of parameters, enabling advanced NLP applications.[35] Multimodal variants commonly employ cross-attention to fuse non-text modalities. Retrieval augmented generation (RAG) adds a further safeguard by letting the model fetch primary literature or database records at run time and ground each claim in a cited passage. Recent agentic RAG systems like PaperQA2 report strong performance on literature review/QA tasks by retrieving full-text sources and enforcing citation-backed claims.[21] This lowers hallucination frequency, though it cannot remove it entirely.[22]

Building on this foundation, LLMs support text categorisation, keyword extraction, and the automatic conversion of free prose into structured data relevant to chemistry.[5, 41] Internally, LLMs treat text as a sequence of tokens, usually whole words or smaller subword pieces, which pass through stacked layers of self-attention (where each token attends to all other tokens in the sequence).[35] As their tokenisers are trained on generic web corpora, chemical information, such as oxidation states (*e.g.* Fe^{III}), unusual ligand labels, and even unicode arrows can be fragmented into several subtokens, a mismatch that can erode numerical fidelity and alter the chemical meaning. On average, terms were segmented into tokens only 4–6 characters long, producing fragmented inputs and eroding structured chemical semantics at the embedding layer.[1, 38] Likewise, Tarasova notes that chemical named-entity recognition is strongly affected by tokenization.[33] Training typically lies in two stages: first, pretraining, where the network reads an internet-scale corpus and learns to predict unseen tokens from context, and second, fine-tuning or instruction tuning, where the same parameters are aligned with task-specific prompts used by chemists. Decoder-style models learn autoregressively to predict the next token, whereas encoder-style models recover a subset of intentionally masked tokens, as demonstrated by ChemBERTa and MolBERT.[4, 7] From an architectural standpoint, the community distinguishes encoder-only systems that yield compact semantic embeddings useful for similarity search and property prediction, decoder-only systems that excel at continuous text generation and dialogue, and newer encoder–decoder hybrids that aim to combine or balance between both strengths.[27] These fixed-length vector embeddings cluster chemically related sentences, reactions, and coordination environments in latent space, enabling k-nearest-neighbour screening for catalysts, ligands, or structure–property relationships without any natural-language output.

While these architectures unlock powerful representations, their open-ended training also introduces new sources of error. As a standalone transformer lacks an authenticated

knowledge base, it can generate confident yet unfounded statements, a phenomenon known as "hallucination". Reinforcement learning from human feedback (RLHF) reduces the most obvious errors by steering the model toward expert judgment, but the same procedure can miscalibrate token probabilities, so a persuasive answer may still be wrong.[25] Retrieval augmented generation (RAG) adds a further safeguard by letting the model fetch primary literature or database records at run time and ground each claim in a cited passage, which lowers hallucination frequency, though it cannot remove it entirely.[22] A third approach appears in ReAct agents (Reason and Act agent framework), which explicitly alternates a natural-language "Reason" step with an "Act" step that calls an external tool such as a crystallographic search, a quantum-chemical calculation, or a laboratory robot. By publishing every intermediate thought and the evidence it retrieves, systems like ChemCrow allow chemists to inspect the whole chain of logic before accepting a recommendation.[3] In general, LLMs are already reshaping research.[37] In particular, this article, therefore, aims to survey and highlight the latest LLM-driven advances in inorganic chemistry, outlining directions their forthcoming implications on the field.

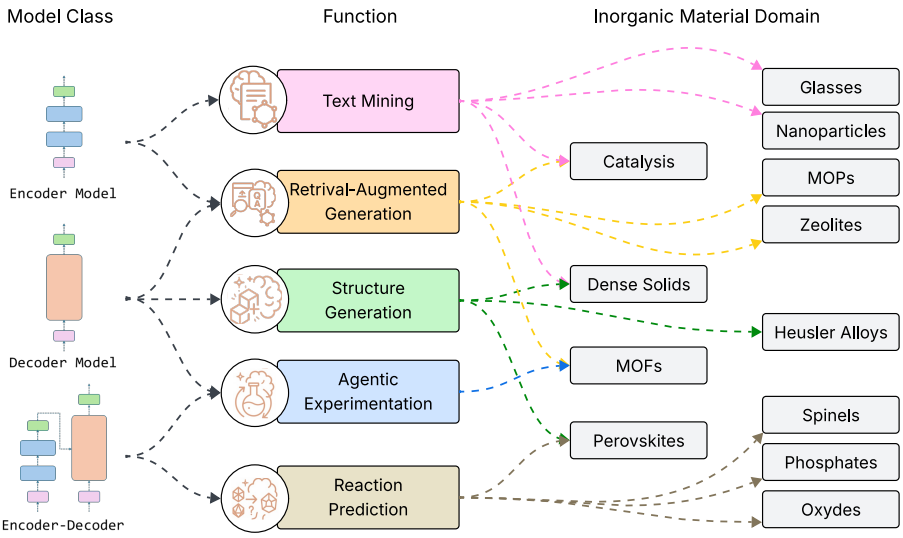


Figure 1: *Transformer-based encoder and decoder models power a range of LLM applications in inorganic chemistry. Encoder networks mainly handle text mining and retrieval, and augmented question answering. Decoder networks drive generative exploration of compositions and crystal structures, steer agent-controlled autonomous experiments, and predict reactions or synthesis pathways. Representative material classes for each task appear on the right.*

2 LLM-driven text mining and structuring

Automated chemistry literature data extraction has been gradually gaining a lot of interest over the past decade.[32] Before transformers, traditional NLP was applied by Kononova *et al.*[18] who converted a corpus of 53,538 inorganic synthesis paragraphs into 19,488 balanced solid-state reactions through a combination of topic modelling, random forests

and sequence tagging. Their approach has reached 93 % chemistry accuracy in terms of predicting correct precursors, targets, reactions; attributing the errors to parsing of paragraphs and the aforementioned tokenisation problem, resulting in failure to parse chemical composition. However, a significant challenge in training on extracted data is synthesis reproducibility; multiple attempts of the same literature procedure can yield different outcomes due to subtle, unreported variations, making it difficult for models to identify critical procedural details.[26] Random forests approach was also undertaken by Zaki *et al.* who have applied the idea on extracting the annealing temperatures and indentation loads from a small set of glass-mechanics papers. The authors combined the textual data with compositional records to create a 102-point set, which improved Vickers-hardness prediction to a test $R^2 = 0.89$ after processing the data.[36] Recent advances have expanded this to multimodal curation. For example, Chan and coworkers developed EXSCLAIM!, an automated pipeline that extracts and labels microscopy and spectroscopy data from the primary literature to bridge the gap between text and visual evidence.[31]

Gupta *et al.* present MatSciBERT, a BERT model continued pretraining on about 150,000 materials papers about 285 million words. It sets a new state of the art on the Matscholar NER benchmark with test Macro F1 of 0.8638; improves relation classification for synthesis procedures; and classifies glass versus non glass abstracts with 96% accuracy. The glass abstracts cover inorganic glass topics including bioactive and rare earth doped systems, while metallic glasses appear in the pretraining corpus. Checkpoints are public and run on standard GPUs.[11] Trewartha *et al.*[34] introduce MatBERT trained on about two million materials papers. Across solid state, doping, and gold nanoparticle morphology tasks it exceeds general baselines by about one to four F1 points, and a compact BiLSTM with Mat2Vec still beats untuned BERT in these settings. The nanoparticle corpus concerns gold nanoparticles such as rods and spheres. Code and weights are openly available.

Transformer-based text miners now dominate inorganic chemistry curation, and two workflows have emerged. The first keeps the language model frozen and steers it with prompts over a focused corpus. Zheng *et al.*[39] guided the GPT (Generative Pretrained Transformer) framework through 228 metal organic framework (MOF) papers. GPT-3.5 and GPT-4 first isolated the synthesis sections and then extracted 26,257 reaction variables for roughly 800 frameworks, giving F1 scores between 0.90 and 0.99, where F1 is the harmonic mean of precision and recall. The second workflow fine-tunes a lighter model on a few hundred labelled examples and then applies it more broadly. Dagdelen *et al.* fine-tune GPT-3 and open access Llama-2 on a few hundred annotated passages to jointly extract entities and relations.[6] The models emit JSON records that link inorganic host compounds (for example, doped oxides and chalcogenides in solid state semiconductors), their dopants, crystal structure or phase labels, and guest species in metal-organic frameworks. On a solid state doping task they reach F1 of about 0.82 for host to dopant links; on general and MOF extractions, exact match relation F1 typically lies in the 0.3 to 0.6 range (with higher manual scores reflecting normalisation and error correction). A human in the loop setup reduced annotation time by approximately 57%.[6]

3 LLM-based predictions for inorganics

Prior to LLMs, data-driven property prediction in inorganic chemistry relied on descriptor-based machine learning, in which compositions are encoded as fixed-length feature vectors derived from curated elemental properties (e.g. the 98-feature Oliynyk property list[19] or Magpie statistics) or structural fingerprints (SOAP, crystal graphs), and mapped to targets by conventional algorithms. These approaches require manual feature engineering but encode explicit physical knowledge, whereas LLMs learn representations implicitly from text—more flexible but less physically grounded.

Beyond information extraction and structuring, LLMs can act as powerful approximators and predictor of chemical properties. Inorganic synthesis spans a vast composition space, thus the expectation of current AI models is only to unveil promising starting conditions. Before the transformer models, Kim *et al.* mined synthesis text with recurrent nets and trained a conditional variational autoencoder, a generative model that samples outputs given the target to propose action sequences and precursors.[14] From about 51,000 action sequences and 116,000 precursors, the model suggested plausible precursor sets for $InWO_3$ and $PbMoO_3$ with training only up to 2005, and it screened 83 predicted ABO_3 perovskites to 19 with at least one route using commercially available precursors. More recently, Okabe *et al.* fine-tuned a distilled GPT on about nineteen thousand balanced reactions from the previously published corpus.[18, 24] One variant maps reactants to products, another maps products to reactants, and a third writes a full equation from only a target formula. Using a Tanimoto similarity on element counts, these models keep roughly *ca.* 90 % chemical fidelity even when prompts include extra verbs such as *heat*, *mix* or *quench*. Demonstrations cover $BaTiO_3$, $SrTiO_3$, the high-T_c cuprate $YBa_2Cu_3O_7$, $BiFeO_3$, $LiMn_2O_4$, $Ni_{0.6}Zn_{0.4}Fe_2O_4$ and $Co_3Sn_2S_2$.[24] However, these models face several limitations. The Tanimoto similarity metric based on element counts is relatively coarse and cannot distinguish between different oxidation states or structural motifs. The models are constrained to compositions well-represented in their training data and cannot reliably extrapolate to novel chemistries. Critically, they predict thermodynamically plausible reactions but provide no information about kinetic barriers, reaction conditions, or practical feasibility. A suggested precursor set may be chemically valid yet experimentally impractical due to cost, availability, or incompatible processing requirements.

Kim and coworkers extended this idea with SYNTHGPT, a GPT 3.5 model fine-tuned on positive unlabelled compositions from the Materials Project.[15] Given only a formula, the model classifies each composition as likely to be synthesised or as unlabelled which is unknown or unlikely. They recalibrated the decision threshold using the estimated proportion of true positives in the unlabelled pool, which balanced recall and precision and matched or exceeded a stoichiometric graph fingerprint.[15] When prompted with a full reaction such as $LiFePO_4 \longleftarrow Li_2CO_3 + FeC_2O_4 + (NH_4)_2HPO_4$, the same model selected precursors with top 1 accuracy on par with the Elementwise template formulation method and top 5 accuracy that slightly exceeded that method's notional limit because it can choose valid reagents outside that template, for example phosphate or ammonium salts.[15] The Elementwise template assumes one precursor per metal element in the target which constrains its outputs, whereas SYNTHGPT is not bound by that rule.

In a recent study, Kim *et al.* developed STRUCTGPT, a GPT-4o-mini model trained to

predict the synthesisability of inorganic materials from textual descriptions of their crystal structures.[16] The performance of the model was shown to be comparable to established crystal-graph convolutional neural networks (CGCNNs), while a related approach using the model's text embeddings in a positive-unlabeled classifier set a new benchmark for the F_1 score at a lower computational cost. The model demonstrates high structural sensitivity, correctly lowering its synthesisability score in response to small, symmetry-breaking coordinate perturbations. Crucially, StructGPT successfully identified twelve hypothetical, near-hull compounds (within 0.01 eV atom^{-1}) as non-synthesizable, a prediction that aligns with failed experimental attempts and suggests the model captures kinetic barriers missed by simple energy screens. A key feature is the ability of the model to explain its reasoning, generating rules based on factors like size mismatch and coordination strain, with the authors noting distinct explanatory themes for major inorganic material classes like perovskites, Heuslers, and spinels. However, the model's text-based representation may miss geometric details and cannot reliably handle defects or disorder, and its explanations may represent post-hoc rationalisations rather than true mechanistic understanding. The chemical origin of these predictions remains unclear. SynthGPT is trained on Materials Project compositions, so its predictions likely reflect both computed thermodynamic stability and patterns of experimental accessibility. StructGPT shows sensitivity to symmetry-breaking coordinate perturbations and identifies some near-hull compounds as non-synthesizable, which suggests that it may also capture aspects of kinetic inaccessibility beyond simple energy-based screening.[16] At the same time, both signals are affected by data bias. Compositions that dominate the training set may appear more synthesizable regardless of their true thermodynamic or kinetic status. Addressing this limitation will require datasets that include failed as well as successful synthesis attempts.

Gruver *et al.* finetune a large language model to write *inorganic* crystal structures as strings and to propose stable candidates.[10] Training uses the Materials Project MP 20 subset of stable inorganic crystals with at most twenty atoms per unit cell and an extended set of about one hundred twenty seven thousand Materials Project entries for text conditioning and infilling. The strongest model reaches 99.6 % structural validity and 95.4 % compositional validity, and 49.8 % of samples are predicted metastable with $\hat{E}_{\text{hull}} < 0.1\,\text{eV}\,\text{atom}^{-1}$ versus 28.8 % for a diffusion baseline. The domain covers diverse inorganic families such as oxides, nitrides, carbides, borides, halides, and intermetallics, including perovskites and spinels.

Despite these advances, LLM-based prediction models face several common limitations. Synthesis prediction models rely on coarse metrics that cannot distinguish oxidation states or structural motifs, and while they suggest thermodynamically plausible reactions, they provide no information about kinetic barriers, reaction conditions, or practical feasibility. Synthesizability classifiers may misclassify thermodynamically stable but kinetically inaccessible compositions or miss realizable metastable phases. Text-based structure representations may lose geometric nuances and struggle with defects, disorder, or non-stoichiometry, while model explanations may represent post-hoc rationalizations rather than mechanistic understanding. Crystal structure generators, despite high structural validity, cannot guarantee experimental synthesizability beyond energy criteria, are typically restricted to small unit cells, and provide no synthesis guidance. Fundamentally, all models are constrained by training data and struggle to extrapolate to novel chemistries or underexplored composition spaces.

A comparison between LLM-based and established materials-informatics approaches is given in Table 1. Recent benchmarking efforts comparing LLM-based property predictors against crystal graph convolutional neural networks (CGCNN) show that neither approach universally dominates.[16, 23] For property prediction tasks, CGCNN outperforms LLM-based models on five out of ten benchmark datasets, particularly those with complex structural features requiring detailed geometric understanding.[23] Conversely, LLMs excel on datasets with shorter textual descriptions or composition-focused tasks, where linguistic context provides advantages over graph-based representations. For synthesizability prediction, fine-tuned LLMs such as StructGPT achieve F_1 scores comparable to PU-CGCNN methods, with the best performance obtained by combining LLM-derived embeddings with traditional positive-unlabelled learning classifiers.[16] These benchmarks highlight that LLMs currently complement rather than replace physics-based and graph-based methods. The primary advantages of LLMs lie in their natural language interfaces, their ability to integrate unstructured literature knowledge, and their explainability through text generation, rather than in superior predictive accuracy on well-defined numerical tasks. Future progress will likely require hybrid architectures that combine the geometric precision of graph neural networks with the contextual reasoning of language models.

The reliability of LLM predictions increases substantially when they are coupled with first-principles validation. Several frameworks now implement autonomous "propose-and-test" cycles where LLMs generate hypotheses that are validated through DFT calculations. For instance, the AtomAgents framework employs physics-aware multimodal agents to orchestrate the entire pipeline: generating alloy structures, executing simulations, and integrating deep learning potentials with physics-based validation.[8] Similarly, the MatPC framework demonstrates this synergy in practice, using LLMs to semantically screen photovoltaic candidates before rigorous DFT confirmation.[20] In future, going beyond structure generation, LLMs can potentially act as technical assistants for computational workflows, suggesting optimal DFT functionals (e.g. identifying when Hubbard U corrections are necessary for specific transition metals) or selecting appropriate basis sets for heavy f-block elements.

4 LLMs supporting real-time guidance

Natural-language queries are central to day-to-day research, and large language models paired with curated, domain-wide knowledge graphs such as *The World Avatar* (TWA) make those queries both fluent and reliable. TWA stores chemistry as RDF (Resource Description Framework) triples, which can be retrieved using the SPARQL query language. The retrieval-augmented generation (RAG) then enables an LLM to retrieve relevant triples at runtime, frame them in readable text, and attach precise citations. With this framework, *Rihm et al.* ingested more than 1,500 metal–organic polyhedra into TWA and launched the agent MARIE, which turns a plain-English question into a SPARQL query, returns a verified table, and displays an interactive 3-D model for each cage.[28] *Kondinski et al.* have applied the RAG architecture to zeolites, covering over 200 frameworks, thousands of material variants, and their reference X-ray patterns.[17]

Table 1: *Comparison of LLM-based and established materials-informatics approaches for inorganic chemistry tasks. Metrics are reported as given in the cited references.*

Method	Input repr.	Task	Key metric	Advantages	Limitations
LLM-based approaches					
MatSciBERT[11]	Text (150 k papers)	NER, classification	Macro $F_1 = 0.86$ on Matscholar NER	Open weights and standard GPU use	Encoder-only model with no generation
SYNTHGPT[15]	Composition string	Synthesisability prediction; precursor selection	Top-1 accuracy similar to the Elementwise template	Not restricted by the one-precursor-per-metal rule	Coarse metric and no kinetic or reaction-condition information
STRUCTGPT[16]	Textual crystal structure	Synthesisability prediction	F_1 comparable to PU-CGCNN at lower computational cost[16]	Sensitive to structural perturbations and provides text-based explanations	May miss defects, disorder, and non-stoichiometry, and explanations may be post hoc[16]
Gruver *et al.*[10]	CIF-as-string	Crystal structure generation	99.6% structural validity, with 49.8% of samples predicted to be metastable at $E_{\text{hull}} <$ 0.1 eV atom^{-1}	High validity and better performance than the diffusion baseline	Limited to small unit cells and provides no synthesis route
Established approaches					
CGCNN[23]	Crystal graph (3D structure)	Property prediction	Better than LLMs on 5 of 10 benchmark datasets[23]	Geometrically precise with physics-informed message passing	Requires known 3D structures and has no language interface
Descriptor-based ML (e.g. Oliynyk property list[19])	Compositional descriptors (98 elemental properties)	Structure classification; property prediction	Effective on small datasets (50–1000 points) and experimentally validated[19]	Interpretable, works with limited data, and does not require a 3D structure	Requires manual feature engineering and contains no explicit geometric information

Beyond retrieval, LLM-based agents are also actively used to guide live experiments. While systems like MOLSIMPLIFY have long automated the generation of transition metal complexes (Kulik et al.),[12] LLM agents now offer higher-level orchestration. Zheng *et al.* built a "ChatGPT Research Group" of seven role specific assistants that communicate with a single chemist, process the literature, write Python code, operate a robotic platform and a programmable microwave system, and guide a Bayesian optimiser.[40] The loop searched a space of about six million possible microwave reaction conditions and converged in about 120 experiments to yield highly crystalline metal organic frameworks such as MOF 321 and MOF 322, with surface areas and pore volumes close to the theoretical values and with high water uptake.[40] Recent developments, such as the interconnection of LLMs with an RDF framework, report the simultaneous involvement of up to six GPT-4 agents that process papers, queue high-throughput batches, read spectra, and design scale-ups—all inside one chat window.[29] A GPT-4 model in combination with eighteen chemistry tools has likewise been shown to work smoothly in organic chemistry, planning reaction routes and designing new chromophoric materials.[3] This combination of tools and chain-of-thought reasoning within the ReAct framework makes these implementations superior to a plain GPT model and likely to be applied in inorganic domains soon.[3] Similarly, Jihan Kim and coworkers introduced CHATMOF, an autonomous multi-agent system that leverages specialised tools to predict and generate metal-organic frameworks with high fidelity.[13]

5 Summary and Outlook

Large language models are rapidly transitioning from proof-of-concept demonstrations to practical tools in inorganic chemistry. They already extract facts from the literature with high accuracy, propose plausible new compositions, and guide closed-loop experiments for material synthesis. The next significant advance will come from better practice rather than bigger models. To build truly dependable systems, the community should curate robust and reusable data by sharing negative and null results and by preserving chemical meaning during tokenisation. Compact open source LLMs also merit consideration for algorithmic transparency, as their modest computational needs enable localised privacy-preserving deployment where appropriate.[2] Trust and safety must be central, with outputs grounded in curated databases with clear citations, with the full chain of logic recorded when models control instruments, and with rule-based validation before any action. More broadly, we view language models as one part of a broader foundation model ecosystem that includes vision, geometry-aware, generative, spectroscopy, and multimodal models, with language models serving as the coordinating layer that explains decisions and connects tools and hardware. With this practice-first approach and a balanced foundation model stack, we can build reliable AI agents that plan inorganic syntheses, explain structures, and accelerate discovery, while moving toward more autonomous and trustworthy chemistry. At the same time, language models are likely to be most useful in well-defined, high-throughput workflows such as literature triage, protocol drafting, and tool-calling pipelines for screening. They are less reliable when small electronic or geometric differences determine the outcome, as in strongly correlated d- and f-electron systems, metal–insulator transitions, iron(II) complexes, or magnetic anisotropy in lan-

thanides. In such cases, coupling language models with explicit physics-based validation remains essential.

Author contributions

All authors contributed to the conceptualisation and writing of this manuscript.

Conflicts of interest

There are no conflicts to declare.

Data availability

There is no new chemical data associated with this article.

Acknowledgements

BK, SP and SDž thank the Slovenian Research and Innovation Agency (ARIS) for financial support under GC–0001, P2-0103 and PR-12394. AK acknowledges competitive internal funding from TU Graz (AF22-635-1).

References

[1] Note: This tokenisation mismatch is particularly problematic for coordination chemistry. If a model fragments symbols like Fe^{III} such that the oxidation state is decoupled from the metal centre, the underlying neural network loses the ability to learn fundamental chemical constraints such as charge balance, redox logic, and ligand field stabilisation.

[2] X. Bai, Y. Xie, X. Zhang, H. Han, and J.-R. Li. Evaluation of open-source large language models for metal–organic frameworks research. *J. Chem. Inf. Model.*, 64 (13):4958–4965, 2024. doi:10.1021/acs.jcim.4c00065.

[3] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *Nature Machine Intelligence*, 6:525–535, 2024. doi:10.1038/s42256-024-00832-8.

[4] S. Chithrananda, G. Grand, and B. Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction, 2020.

[5] K. Choudhary and M. L. Kelley. Chemnlp: A natural language-processing-based library for materials chemistry text data. *J. Phys. Chem. C.*, 127(35):17545–17555, 2023. doi:10.1021/acs.jpcc.3c03106.

[6] 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.

[7] B. Fabian, T. Edlich, H. Gaspar, M. Segler, J. Meyers, M. Fiscato, and M. Ahmed. Molecular representation learning with language models and domain-relevant auxiliary tasks, 2020.

[8] A. Ghafarollahi and M. J. Buehler. Automating alloy design and discovery with physics-aware multimodal multiagent ai. *Proceedings of the National Academy of Sciences*, 122(4):e2414074122, 2025. doi:10.1073/pnas.2414074122.

[9] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. ISBN 9780262035613. URL http://www.deeplearningbook.org.

[10] N. Gruver, A. Sriram, A. Madotto, A. G. Wilson, C. L. Zitnick, and Z. Ulissi. Fine-tuned language models generate stable inorganic materials as text. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. doi:10.48550/arXiv.2402.04379.

[11] T. Gupta, M. Zaki, N. M. A. Krishnan, and Mausam. Matscibert: A materials-domain language model for text mining and information extraction. *npj Comput. Mater.*, 8(1):102, 2022. doi:10.1038/s41524-022-00784-w.

[12] J. P. Janet and H. J. Kulik. Strategies and software for machine learning accelerated discovery in transition metal chemistry. *Chem. Sci.*, 8(8):5137–5152, 2017. doi:10.1039/C7SC01247K.

[13] Y. Kang and J. Kim. Chatmof: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models. *Nat. Commun.*, 15(1):461, 2024. doi:10.1038/s41467-024-48998-4.

[14] E. Kim, Z. Jensen, A. van Grootel, K. Huang, M. Staib, S. Mysore, H.-S. Chang, E. Strubell, A. McCallum, S. Jegelka, and E. Olivetti. Inorganic materials synthesis planning with literature-trained neural networks. *J. Chem. Inf. Model.*, 60(3):1194–1201, 2020. doi:10.1021/acs.jcim.9b00995.

[15] S. Kim, Y. Jung, and J. Schrier. Large language models for inorganic synthesis predictions. *J. Am. Chem. Soc.*, 146(29):19654–19659, 2024. doi:10.1021/jacs.4c05840.

[16] S. Kim, J. Schrier, and Y. Jung. Explainable synthesizability prediction of inorganic crystal polymorphs using large language models. *Angew. Chem. Int. Ed.*, 64(19): e202423950, 2025. doi:10.1002/anie.202423950.

[17] A. Kondinski, P. Rutkevych, L. Pascazio, D. N. Tran, F. Farazi, S. Ganguly, and M. Kraft. Knowledge graph representation of zeolitic crystalline materials. *Digit. Discov.*, 3:2070–2084, 2024. doi:10.1039/D4DD00166D.

[18] 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.

[19] S. Lee, C. Chen, G. Garcia, and A. O. Oliynyk. Machine learning descriptors in materials chemistry used in multiple experimentally validated studies: Oliynyk elemental property dataset. *Data in Brief*, 53:110178, 2024. doi:10.1016/j.dib.2024.110178.

[20] X. Liang et al. Matpc: Prompting large language model, crystal structure prediction, and first-principles for semantic-driven material design. *ACS Applied Materials & Interfaces*, 2025. doi:10.1021/acsami.5c08809. Published online July 28, 2025.

[21] J. Lála et al. Paperqa2: Real-time scientific agent. *arXiv preprint*, 2024. arXiv:2409.13741.

[22] P. R. Maharana and K. Joshi. Retrieval augmented generation for building datasets from scientific literature. *ChemRxiv*, pages 1–19, 2024. doi:10.26434/chemrxiv-2024-qjx32. Preprint.

[23] A. Niyongabo Rubungo, K. Li, J. Hattrick-Simpers, and A. Bousso Dieng. Llm4mat-bench: benchmarking large language models for materials property prediction. *Machine Learning: Science and Technology*, 6(2):020501, may 2025. doi:10.1088/2632-2153/add3bb. URL https://doi.org/10.1088/2632-2153/add3bb.

[24] R. Okabe, Z. West, A. Chotrattanapituk, M. Cheng, D. C. Carrizales, W. Xie, R. J. Cava, and M. Li. Large language model-guided prediction toward quantum materials synthesis, 2024. URL https://arxiv.org/abs/2410.20976. Preprint, cond-mat.mtrl-sci.

[25] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback, 2022.

[26] P. Raccuglia et al. Machine-learning-assisted materials discovery using failed experiments. *Nature*, 533:73–76, 2016.

[27] M. C. Ramos, C. J. Collison, and A. D. White. A review of large language models and autonomous agents in chemistry. *Chem. Sci.*, 16:2514–2572, 2025. doi:10.1039/D4SC03921A.

[28] S. D. Rihm, D. N. Tran, A. Kondinski, L. Pascazio, F. Saluz, X. Deng, S. Mosbach, J. Akroyd, and M. Kraft. Natural language access point to digital metal–organic polyhedra chemistry in the world avatar. *Data-Centric Eng.*, 6:e22, 2025. doi:10.1017/dce.2025.12.

[29] Y. Ruan, C. Lu, N. Xu, Y. He, Y. Chen, J. Zhang, J. Xuan, J. Pan, Q. Fang, H. Gao, X. Shen, N. Ye, Q. Zhang, and Y. Mo. An automatic end-to-end chemical synthesis development platform powered by large language models. *Nat. Commun.*, 15(1): 10160, 2024. doi:10.1038/s41467-024-54457-x.

[30] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, fourth edition, 2020. ISBN 9781292401133. URL http://aima.cs.berkeley.edu/.

[31] E. Schwenker, W. Jiang, T. Spreadbury, N. Ferrier, O. Cossairt, and M. K. Y. Chan. Exsclaim!: Harnessing materials science literature for self-labeled microscopy datasets. *Patterns*, 4(11):100843, 2023. doi:10.1016/j.patter.2023.100843.

[32] M. C. Swain and J. M. Cole. Chemdataextractor: A toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.*, 56(10): 1894–1904, 2016. doi:10.1021/acs.jcim.6b00207.

[33] O. A. Tarasova, A. V. Rudik, N. Y. Biziukova, D. A. Filimonov, and V. V. Poroikov. Chemical named entity recognition in the texts of scientific publications using the naïve bayes classifier approach. *Journal of Cheminformatics*, 14(1):55, Aug 2022. ISSN 1758-2946. doi:10.1186/s13321-022-00633-4. URL https://doi.org/10.1186/s13321-022-00633-4.

[34] A. Trewartha, N. Walker, H. Huo, S. Lee, K. Cruse, J. Dagdelen, A. Dunn, K. A. Persson, G. Ceder, and A. Jain. Quantifying the advantage of domain-specific pretraining on named-entity recognition tasks in materials science. *Patterns*, 3(4): 100488, 2022. doi:10.1016/j.patter.2022.100488.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Proc. of the 31st Adv. Neural Inf. Process. Syst:*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[36] M. Zaki, Jayadeva, and N. M. A. Krishnan. Extracting processing and testing parameters from materials science literature for improved property prediction of glasses. *Chem. Eng. Process.*, 180:108607, 2022. doi:10.1016/j.cep.2021.108607.

[37] Y. Zhang, S. A. Khan, A. Mahmud, H. Yang, A. Lavin, M. Levin, J. Frey, J. Dunnmon, J. Evans, A. Bundy, S. Džeroski, J. Tegner, and H. Zenil. Exploring the role of large language models in the scientific method: from hypothesis to discovery. *npj Artif. Intell.*, 1(1):14, Aug 2025. doi:10.1038/s44387-025-00019-5.

[38] H. Zhao, X. Tang, Z. Yang, X. Han, X. Feng, Y. Fan, S. Cheng, D. Jin, Y. Zhao, A. Cohan, and M. Gerstein. Chemsafetybench: Benchmarking llm safety on chemistry domain. *arXiv preprint*, 2024. arXiv:2411.16736.

[39] 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.

[40] Z. Zheng, O. Zhang, H. L. Nguyen, N. Rampal, A. H. Alawadhi, Z. Rong, T. Head-Gordon, C. Borgs, J. T. Chayes, and O. M. Yaghi. Chatgpt research group for optimising the crystallinity of MOFs and COFs. *ACS Cent. Sci.*, 9(11):2161–2170, 2023. doi:10.1021/acscentsci.3c01087.

[41] H. Öztürk, A. Özgür, P. Schwaller, T. Laino, and E. Ozkirimli. Exploring chemical space using natural language processing methodologies for drug discovery. *Drug Discov. Today*, 25(4):689–705, 2020. ISSN 1359-6446. doi:10.1016/j.drudis.2020.01.020.