

Review

From Scientific Copilots to Tool-Grounded Autonomy: AI Agents in Simulation-Driven Materials Discovery

Anibal Alviz-Meza ^{1,*} , Alejandro Valencia-Arias ², Segundo Rojas-Flores ³ and Felix Diaz ⁴ 

¹ Escuela de Posgrado, Universidad Continental, Lima 15302, Peru

² Vicerrectoria de Investigacion y Posgrado, Universidad de Los Lagos, Osorno 5311157, Chile; alejandro.valenciaarias@ulagos.cl

³ Institutos y Centros de Investigacion de la Universidad Cesar Vallejo, Universidad Cesar Vallejo, Trujillo 13001, Peru; srojasf@ucv.edu.pe

⁴ Facultad de Ingenieria y Arquitectura, Escuela Profesional de Ingenieria de Sistemas, Universidad Autonoma del Peru, Lima 15842, Peru; felix.diaz@autonoma.pe

* Correspondence: aalviz@continental.edu.pe

Abstract

Artificial intelligence (AI) agents and large language model (LLM) agents are beginning to move materials discovery beyond isolated prediction tasks and toward tool-grounded workflows that can retrieve prior knowledge, configure simulations, launch calculations, inspect outputs, and decide what to do next. However, adjacent reviews on materials informatics, self-driving laboratories, natural-language processing in materials science, and autonomous chemistry have not isolated simulation-driven materials workflows as a distinct evidence base. This review addresses that gap through PRISMA-guided searches in Scopus (8 May 2026) and Web of Science (15 June 2026) for English-language journal articles published between 2022 and 2026. The combined search returned 232 records; 27 full texts were assessed and 26 studies were included in the final qualitative synthesis after one full-text exclusion. No eligible study was published in 2022 or 2023, indicating that the field emerged only in 2024 and expanded rapidly in 2025–2026. Catalysis and adsorption tasks ($n = 6$) and alloy design or evaluation ($n = 5$) dominated the corpus, while specialized multi-agent architectures were the most common pattern ($n = 13$). Across the included studies, agentic reasoning was most often coupled to workflow orchestration or integration tools, molecular-dynamics or atomistic simulation environments, and materials-data or machine learning screening pipelines; public repositories or archival artifacts were reported in 18 of 26 studies, experimental validation in six, and robotic closed-loop execution in only one study. The strongest evidence came from workflows that grounded language model decisions in simulators, structured databases, or experimentally verifiable outputs rather than in free-form text alone. This review therefore establishes AI agent workflow orchestration as a distinct analytical category within materials discovery and identifies the reporting, validation, and reproducibility conditions required for these systems to function as credible scientific infrastructure rather than as conversational demonstrations.



Academic Editors: Tzanko Donchev and Mihail Kolev

Received: 28 May 2026

Revised: 17 June 2026

Accepted: 17 June 2026

Published: 21 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

Keywords: agentic AI; large language models; multi-agent systems; materials discovery; materials design; simulation workflows; materials informatics; density functional theory; molecular dynamics

1. Introduction

Materials discovery is increasingly limited not by the absence of predictive models, but by the difficulty of coordinating the full sequence of scientific work that surrounds them: retrieving prior knowledge, constructing candidate structures, preparing simulation inputs, selecting toolchains, interpreting outputs, and deciding whether another iteration is justified. Machine learning and materials informatics have already transformed candidate prioritization, property prediction, and inverse design [1,2]. At the same time, broader acceleration agendas now integrate artificial intelligence, high-performance computing, automation, and robotics into multi-stage discovery pipelines rather than isolated modeling tasks [2–4]. In practice, the bottleneck is therefore increasingly procedural and organizational. The central challenge is not only to predict a property, but to manage the sequence of reasoning and action that connects a scientific objective to executable computational or experimental steps [2–10].

Recent reviews document important parts of this transition, but they do not synthesize the same object of study. Reviews of self-driving laboratories emphasize closed-loop experimentation, hardware integration, and laboratory automation [3]. Reviews of AI-driven materials design focus more broadly on generative models, inverse design, reinforcement learning, and materials-discovery acceleration [4]. Other reviews concentrate on natural-language processing in materials science, large language models for materials and chemistry, reticular chemistry, or generative crystal-structure modeling [5,11–14]. In parallel, chemistry-oriented reviews have shown that large language models can move beyond text generation toward tool use, retrieval, planning, code execution, and multi-agent coordination [5–10]. Taken together, this literature shows that autonomous scientific systems are becoming technically plausible. It does not yet provide a focused systematic synthesis of studies in which agents are used to orchestrate simulation-centered materials workflows.

That distinction matters because a property-prediction model, a literature assistant, a generative model, and an agent that configures and executes a simulator are not equivalent scientific objects. Once the unit of analysis shifts from the model to the workflow, different evaluation criteria become necessary. In this review, an agentic workflow was understood minimally as a system that selects tools, parameterizes a computational task, triggers execution of the relevant toolchain, inspects the resulting outputs, and determines whether and how the next iteration should proceed. In simulation-driven materials discovery, the key issue is whether the system can connect reasoning to action across tools such as density functional theory packages, molecular dynamics engines, adsorption and process simulators, CALPHAD environments, crystallographic curation routines, and programmable analysis pipelines. Under that view, the relevant questions are no longer limited to accuracy of prediction. They include which material domains are actually being targeted, which tool families are being orchestrated, how autonomy is bounded in practice, how validation is performed, and how often prompts, traces, repositories, and failure cases are reported in reproducible form [5–10].

To the best of our knowledge, no prior review has isolated simulation-driven, tool-orchestrating AI agent workflows for materials discovery as a distinct evidence base with study-level extraction of autonomy, validation, and reproducibility [3–14]. This is not a minor gap in categorization. Without that distinction, studies that execute atomistic simulations, studies that manage databases and ranking pipelines, and studies that merely expose a conversational interface to materials knowledge can all be collapsed under the same label of agentic AI, even though they differ substantially in scientific function and evidentiary strength. The novelty of the present review therefore lies in treating agentic AI not as a generic synonym for artificial intelligence in materials science, but as a work-

flow architecture whose scientific value depends on planning, coordinating, executing, and evaluating discovery tasks across tool-integrated computational environments.

Accordingly, the present study provides a PRISMA-guided systematic review of journal articles retrieved from Scopus and Web of Science and published between 2022 and 2026, with staged screening, cross-database deduplication, full-text assessment, and qualitative synthesis of the final corpus. No eligible article was ultimately retained from 2022 or 2023; one was identified in 2024, nine in 2025, and sixteen in the partial year 2026, which underscores how quickly this literature is emerging and why a workflow-level synthesis has become timely. The review maps the literature across application domains, agent architectures, tool-grounding strategies, autonomy levels, validation practices, and reproducibility signals. Its purpose is not simply to inventory papers, but to show that the field is currently defined by orchestration-heavy designs, credibility through external tool grounding, and reporting practices that still lag behind technical enthusiasm.

2. Materials and Methods

2.1. Review Design and Reporting Framework

This study was designed as a systematic review and reported in accordance with PRISMA 2020 [15,16]. Because the expected evidence base was methodologically heterogeneous, a pooled meta-analysis was not pursued. The retained studies span catalysts, alloys, porous materials, polymers, crystal-generation tasks, atomistic simulation workflows, and robotic materials research, with markedly different validation bases and reporting styles. The synthesis was therefore structured as a comparative qualitative review focused on workflow architecture, computational role, autonomy, validation, and reproducibility. Searches and screening were conducted through Scopus (<https://www.scopus.com>) and Web of Science Core Collection (<https://www.webofscience.com>); software and repository names reported in the evidence tables are characteristics extracted from the cited studies.

2.2. Database, Search Strategy, and Search Date

Scopus and Web of Science Core Collection were selected because, together, they provide broad coverage across materials science, chemistry, chemical engineering, physics, and computational research while supporting transparent screening and cross-database auditability. The Scopus search was executed on 8 May 2026, and the equivalent Web of Science topic search was executed on 15 June 2026. The exact Scopus search equation applied to the title, abstract, and keyword fields was: TITLE-ABS-KEY(("AI agent*" OR "artificial intelligence agent*" OR "LLM agent*" OR "large language model agent*" OR "agentic AI" OR "multi-agent system*" OR "multiagent system*" OR "autonomous agent*" OR "tool-using agent*" OR "ReAct agent*" OR "planner-executor" OR "agentic workflow*") AND ("materials discover*" OR "material discover*" OR "materials design" OR "material design" OR "materials informatics" OR "computational materials" OR catalyst* OR alloy* OR polymer* OR zeolite* OR "metal-organic framework*" OR "crystalline material*" OR "energy material*" OR "functional material*" OR "high entropy alloy*" OR "phase diagram*") AND (simulation OR computational OR DFT OR "density functional theory" OR "molecular dynamics" OR LAMMPS OR CALPHAD OR "high-throughput screening" OR "workflow orchestration" OR "simulation tool*" OR "phase-field" OR MOOSE OR "Quantum ESPRESSO" OR "atomistic simulation")) AND PUBYEAR > 2021 AND PUBYEAR < 2027 AND DOCTYPE(ar) AND LANGUAGE(English).

The query combined three conceptual blocks within the title, abstract, and keyword fields. The first captured agentic AI terminology, including AI agents, LLM agents, autonomous agents, multi-agent systems, tool-using agents, ReAct agents, and agentic workflows. The second captured the materials-discovery scope, including materials discovery,

materials design, materials informatics, computational materials, catalysts, alloys, polymers, zeolites, metal–organic frameworks, crystalline materials, energy materials, functional materials, and high-entropy alloys. The third captured simulation and computation, including DFT, molecular dynamics, CALPHAD, high-throughput screening, workflow orchestration, simulation tools, phase-field methods, Quantum ESPRESSO, and atomistic simulation.

2.3. Eligibility Criteria

Records were eligible when they were indexed in Scopus or Web of Science as English-language journal articles published between 2022 and 2026; described an AI agent, LLM agent, autonomous agent, or multi-agent workflow as an operational part of the study; addressed materials discovery, design, optimization, or closely related computational materials tasks; and included a simulation, computational modeling, screening, code execution, or domain tool component that was materially relevant to the workflow. Reviews, perspectives, editorials, conference papers, book chapters, and notes were excluded, as were studies that discussed AI in materials science without an explicit agentic workflow. Conference papers were excluded to maintain a homogeneous, peer-reviewed, fully indexed evidence base with comparable reporting depth, since the workflow-level extraction used here typically requires the detail of a full journal article; this restriction may nonetheless bias coverage toward subfields that publish primarily in journals.

2.4. Selection Process and Full-Text Assessment

To probe recall beyond the primary query, the agentic terminology block was re-run with the synonyms self-driving, closed-loop, autonomous workflow, automated simulation pipeline, materials acceleration platform, and active learning over the same window in both databases. The additional records were self-driving laboratory or active learning studies without an explicit tool-orchestrating agent layer and did not change the eligible set once the unchanged criteria were applied; the well-known in-scope systems (e.g., AtomAgents, VASPIlot, Adsorb-Agent) are all captured by the primary query. The cross-database design was deliberately limited to Scopus and Web of Science, the two principal curated, peer-reviewed bibliometric databases.

The screening process followed a staged dual-database logic. The combined searches yielded 232 records, corresponding to 189 from Scopus and 43 from Web of Science. These records were reduced to 127 after the 2022–2026 time window, 95 after restriction to journal articles, and 93 after exclusion of non-English records. Cross-database duplicate control using DOI and normalized titles removed 17 overlapping records, leaving 76 unique English-language articles for thematic screening. First, 30 records were excluded at the metadata level because titles and abstracts indicated review articles or clearly off-scope uses of terms such as agent-based, autonomous, simulation, or multi-agent in non-material domains or without an explicit AI agent workflow coordinating materials computations. Second, the remaining 46 records underwent title-and-abstract assessment, which excluded 19 additional articles because they still lacked a qualifying agentic materials-discovery workflow, addressed non-material domains, or used agent-based modeling in a non-AI sense. Twenty-seven full texts were therefore assessed. One article was excluded at full text because the acronym DFT referred to design-for-test automation in VLSI benchmarking rather than density functional theory or materials discovery. Record-level reasons for the pre-full-text exclusions are provided in Appendix A.

2.5. Data Extraction and Qualitative Synthesis

Data extraction was performed by only one researcher while the other approved the final collected corpus. This process was structured around the review question rather than bibliographic description alone. For each eligible study, the extraction framework

captured the application domain, material class, agent architecture, tool family, workflow autonomy, validation strategy, reported quantitative evidence, reproducibility signals, and the main limitation acknowledged by the study. The classification categories were anchored to the published literature rather than defined ad hoc: autonomy was coded using the established degree-of-autonomy scale for self-driving laboratories, whose piecewise, semi-closed-loop, and closed-loop regimes correspond to the medium, medium-high, and high bands used here, and the architecture categories follow established taxonomies of LLM-based agent systems that separate single-agent and planner-executor designs from role-specialized multi-agent systems. The tool-grounding distinction between simulator-native and discovery-manager workflows is an operational categorization proposed here. This structure was selected to distinguish papers that merely mention agentic AI from papers that actually use agents to plan, execute, evaluate, or iteratively refine computational discovery workflows.

2.6. Methodological Scope and Limitations

Several scope conditions should be considered when interpreting the final corpus. First, the formal review was limited to Scopus and Web of Science Core Collection, which improves reproducibility but can still omit relevant studies indexed elsewhere. Second, the corpus was restricted to English-language journal articles. Third, agentic AI terminology remains unstable, so no single search equation can fully exhaust a rapidly evolving vocabulary. Fourth, the 2026 segment is necessarily partial because the searches were executed on 8 May 2026 and 15 June 2026. Finally, the retained studies differ substantially in material domain, toolchain, and evidence maturity, which is why the review is framed as a qualitative synthesis rather than a pooled effect-size analysis. All reported proportions derive from a corpus of $n = 26$, so each inclusion or exclusion shifts a proportion by roughly 3.8 percentage points and the right-censored 2026 segment should not be read as a growth rate. Although the two principal curated databases were searched, studies indexed only in other repositories or released as preprints may be under-represented. A sensitivity re-run augmenting the agentic block with terms such as self-driving, closed-loop, and active learning surfaced mostly self-driving laboratory studies without an explicit tool-orchestrating agent layer and did not change the eligible set.

3. Results

3.1. Corpus Profile and Publication Dynamics

The final synthesis comprised 26 studies. Figure 1 shows that the main attrition occurred before full-text assessment rather than after it. Of the 76 deduplicated English-language records, 30 were removed at metadata level as clearly off-scope and 19 more were excluded at title-and-abstract screening, leaving 27 full texts and 26 final included studies after one full-text exclusion. Although the formal review window covered 2022–2026, no paper published in 2022 or 2023 met the eligibility criteria after screening and full-text assessment. This does not mean that early candidate records were absent. Instead, the candidate records from 2022–2023 that reached thematic screening were excluded because they concerned service operation simulations, elastomer aging, biofilm morphology, urban electrification, wastewater metabolism, or other non-target uses of agent-based modeling rather than AI agent orchestration for materials discovery. The four WoS-unique records that required full-text audit all proved eligible and strengthened the 2025–2026 segment, but they did not change the absence of any eligible 2022–2023 study. The eligible evidence base therefore begins effectively in 2024, grows in 2025, and continues into the partial year 2026; because the searches were executed in 2026 the latest-year count is right-censored and is not evidence of an accelerating annual rate. This emergence pattern is consistent

with the view that agentic materials workflows constitute a genuinely new research front rather than a mature body of literature that had simply been overlooked.

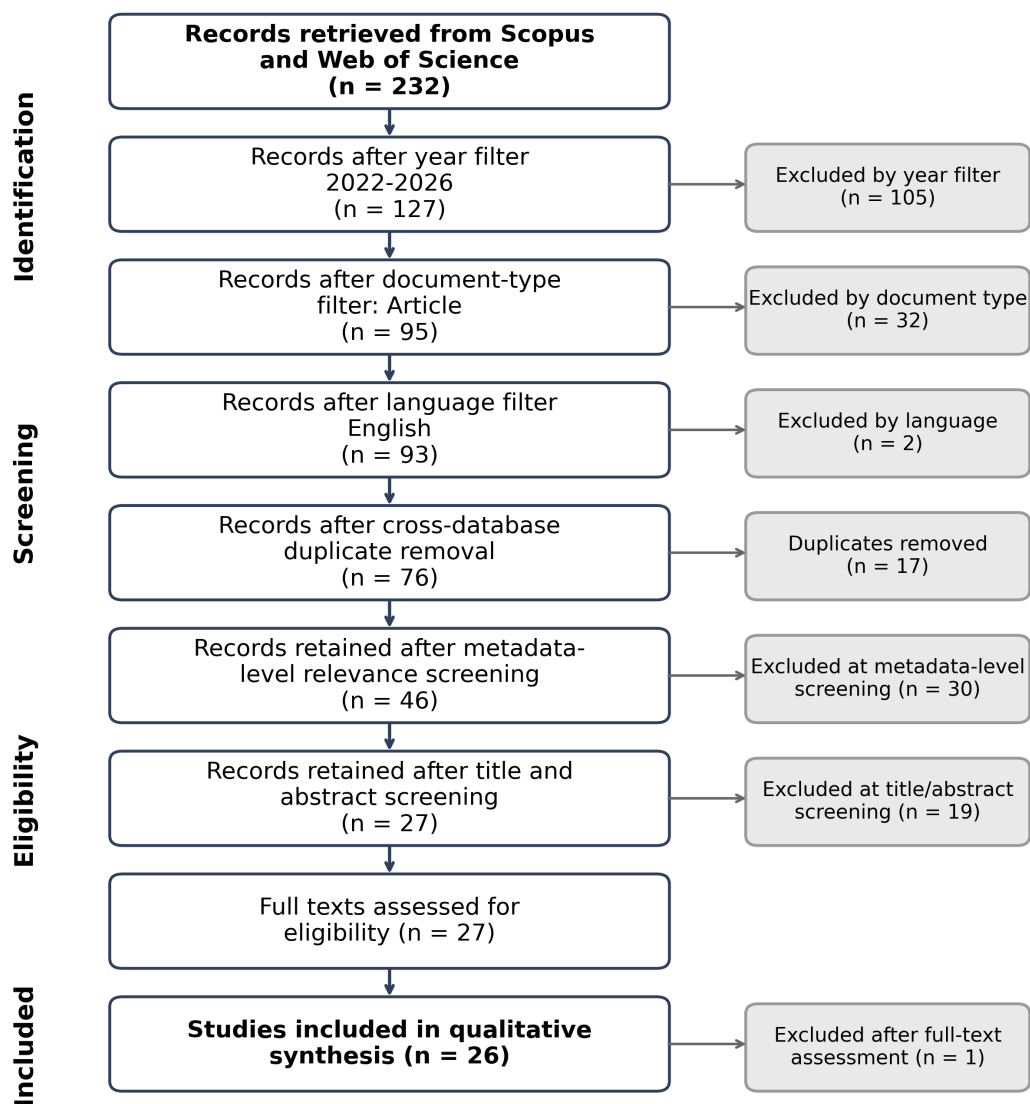


Figure 1. PRISMA flow diagram of the combined Scopus and Web of Science identification, screening, full-text assessment, and final inclusion process.

Outlet distribution was dispersed. *Digital Discovery* published three included articles, while *ACS Nano*, *Communications Materials*, and *Scripta Materialia* each published two; the remaining 17 studies were distributed across 17 different journals with a single contribution each. This pattern suggests that the area has not yet consolidated around a single disciplinary home. Instead, it spans catalysis, materials engineering, atomistic simulation, porous materials, robotics, and computational chemistry venues.

The study-level characteristics of the included corpus are summarized in Table 1.

Table 1. Characteristics of the included combined corpus, including application domain, architecture pattern, tool families, and validation basis.

Ref.	Year	Application Domain	Architecture	Tools	Validation Basis
[17]	2024	General materials design workflows	Specialized multi-agent	MechGPT; retrieval agents; PySCF; autonomous code execution	Physics-grounded case studies and code execution demonstrations

Table 1. Cont.

Ref.	Year	Application Domain	Architecture	Tools	Validation Basis
[18]	2025	Alloy design and discovery	Specialized multi-agent	AtomAgents; atomistic workflows; EAM and moment-tensor potentials; NEB; multimodal image analysis	Four physics-grounded alloy-design experiments with literature and mechanistic checks
[19]	2025	Electrocatalyst data mining and design	Specialized multi-agent	GPT-4o/mini agents; GraphRAG; random forest; agglomerative clustering; Optuna; CSV/code handlers	Manual extraction audit, train–test evaluation, and interactive design demonstrations
[20]	2025	MOF catalyst optimization	Hybrid human-in-the-loop	MOFSyn; XGBoost/SHAP; BM25/FAISS retrieval; RAGAS; natural-language protocol navigation	Agent-guided experimental optimization, expert review, and MOF knowledge exam
[21]	2025	Custom molecular-simulation workflows	Specialized multi-agent	LAMMPS; Moltemplate; PACKMOL; PLUMED; WHAM; MDAnalysis; PyLAT; PDF-backed RAG	End-to-end workflow demonstrations for setup, execution, and analysis
[22]	2025	Thermoelectric-material design	Hybrid human-in-the-loop	DeepSeek-R1; DeepSeek-Prover-V2; literature RAG; comparative design logic	Human–machine comparative design plus reported experimental confirmation
[23]	2025	Accelerated alloy design	Specialized multi-agent	AtomAgents; GNN surrogates; NEB ground truth; solute-strengthening theory; coding and plotting agents	Held-out composition testing and comparison to experimental alloy trends
[24]	2025	Polymer and general simulation automation	Single-agent or planner–executor loop	Python; Skyfield; Astropy; Runge-Kutta N-body simulation; ML workflows	Case-based task execution against reference theory and outputs
[25]	2025	Autonomous electronic-structure workflows	Specialized multi-agent	VASP; Slurm; structure-retrieval tools; plotting utilities; MCP servers	Routine and advanced DFT benchmark missions
[26]	2026	MOF screening for CO ₂ capture	Specialized multi-agent	GCMC adsorption simulations; TSA process modeling; Pareto sorting; NSGA-II/III refinement; consensus ranking	Case study on screened MOFs with process metrics and semantic review

Table 1. Cont.

Ref.	Year	Application Domain	Architecture	Tools	Validation Basis
[27]	2026	Electrocatalyst design with multiscale analysis	Hybrid human-in-the-loop	VASP with DFT + U; VASPKIT; Materials Studio/Forcite MD; operando Raman and ATR-SEIRAS; EXAFS/XAS	Combined experimental characterization with DFT and MD analysis
[28]	2026	Adsorption-configuration search	Single-agent or planner-executor loop	Open Catalyst demo; EquiformerV2 relaxation; adsorption-configuration generation	Twenty-system benchmark against exhaustive algorithmic enumeration
[29]	2026	Radiation-tolerant alloy discovery	Single-agent or planner-executor loop	LAMMPS; Mie + ZBL potentials; AXIS vacancy analysis; Claude-Sonnet API	Six 50-iteration optimization trajectories using 100 eV PKA simulations
[30]	2026	Biomass-valorization catalyst screening	Hybrid human-in-the-loop	Mistral 7B; SQLite; ChromaDB; sentence-transformer embeddings; LangChain; Optuna; AutoSklearn	Model benchmarking plus two experimental deoxygenation validations
[31]	2026	Additive-manufacturing alloy evaluation	Hybrid human-in-the-loop	Thermo-Calc; TC-Python; CALPHAD; lack-of-fusion solver; MCP	Literature comparison and candidate-alloy process-map evaluation
[32]	2026	Programmatic crystal-structure generation	Single-agent or planner-executor loop	Fine-tuned code-generation agent; CIF/API output modes; StructureMatcher evaluation	Crystal-generation benchmarking against an untuned code model
[33]	2026	Autonomous alloy optimization	Single-agent or planner-executor loop	LAMMPS tensile simulations; trajectory logging; composition update loop	Six fifty-iteration composition-optimization trajectories
[34]	2026	Single-atom-alloy catalyst discovery	Single-agent or planner-executor loop	Digital catalysis database; adsorption-energy analysis; descriptor modeling; DFT interpretation; experimental CO2RR validation	Mechanistic descriptor development plus experimental verification
[35]	2026	Autonomous materials research and robotics	Specialized multi-agent	Nineteen LLM agents; sixteen tools; robotic synthesis platform; AutoGen-based coordination	Robotic closed-loop optimization and expert-scored benchmark comparisons

Table 1. Cont.

Ref.	Year	Application Domain	Architecture	Tools	Validation Basis
[36]	2026	End-to-end atomistic simulation	Specialized multi-agent	Atomsk; automated potential discovery; LAMMPS/HPC execution; OVITO; Phonopy	Static and dynamic atomistic simulation benchmarks across elemental and alloy systems
[37]	2026	MOF CIF preprocessing and curation	Specialized multi-agent	Context-aware CIF parsing and curation routines for adsorption-oriented preprocessing	MOF and adsorption benchmark set comparing context-aware to topology-only preprocessing
[38]	2026	Coarse-grained polymer simulation	Specialized multi-agent	LAMMPS-based bead-spring polymer MD; topology-specific configuration and analysis routines	Case studies across linear, ring, brush, star-polymer, and dendrimer systems
[39]	2026	AI agent workflow assembly for electronic-structure and molecular-design tasks	Single-agent or planner-executor loop	Matty AI-Agent; Qwen3; MCP services; VASP; VASPKIT; RDKit; CVAE; Gaussian; Turbomole/COSMO-RS; QM9/C2DB	Representative task demonstrations in electronic-structure calculation and inverse molecular design
[40]	2025	Prompt-driven material property prediction	Single-agent or planner-executor loop	MatAgent; prompt-engineered LLM; Materials Project retrieval; PWDFT and first-principles tools; Python REPL	Three 100-question rounds for property retrieval and property prediction tasks
[41]	2026	Multi-task computational materials science	Specialized multi-agent	MatSciAgent; Materials Project; MatWeb; crystal-generation agent; continuum simulation tools; molecular-dynamics simulation agents	Reliability analysis across repeated extraction and simulation tasks plus multi-case demonstrations
[42]	2026	Autonomous DFT protocol generation and repair	Specialized multi-agent	GENIUS; Quantum ESPRESSO; smart knowledge graph; finite-state error-recovery machine; tiered LLM stack	Benchmark of 295 free-form simulation requests with automated validation and recovery analysis

3.2. Application Domains and Workflow Roles

Catalysis and adsorption tasks formed the largest application block, represented by six studies on electrocatalyst design, adsorption-configuration search, biomass-valORIZATION catalysis, MOF catalyst optimization, and single-atom-alloy discovery [19,20,27,28,30,34]. Alloy design and evaluation formed the second-largest block with five studies [18,23,29,31,33]. The remaining literature was distributed across general materials-workflow papers, atomistic simulation infrastructures, MOF and porous-material screening, and polymer simulation [17,21,22,24–26,32,35–42]. This distribution shows that the field is presently strongest where agents can either reduce the cost of iterative search or orchestrate heterogeneous

computational steps that are otherwise labor-intensive. The annual distribution, outlet concentration, and domain grouping are shown in Figures 2–4, respectively.

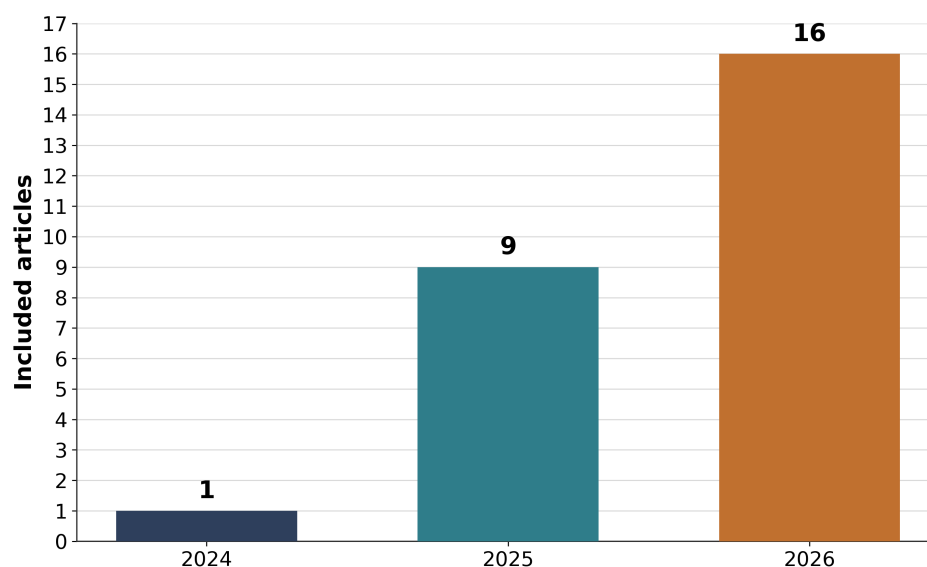


Figure 2. Annual publication distribution of the final combined corpus after full-text assessment.

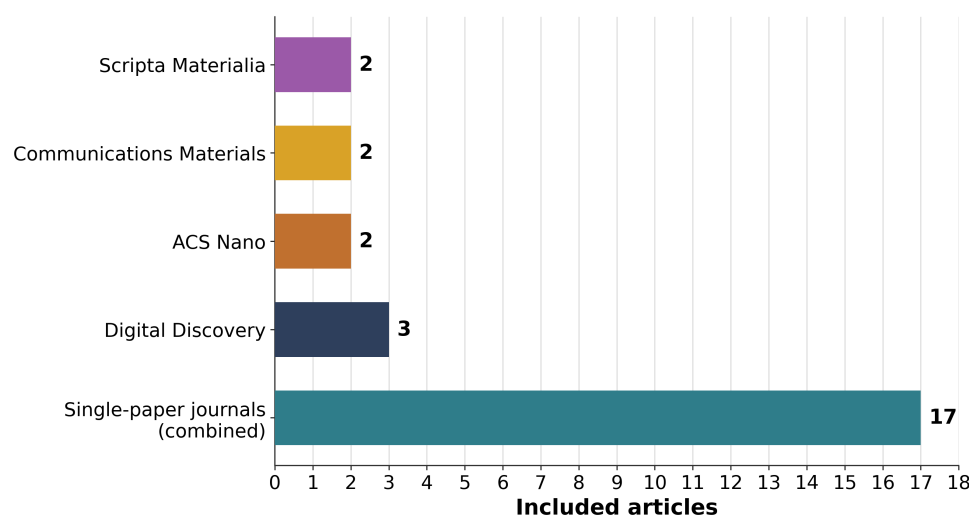


Figure 3. Outlet concentration of the final corpus; seventeen studies were dispersed across seventeen single-paper journals, while four outlets contributed more than one study.

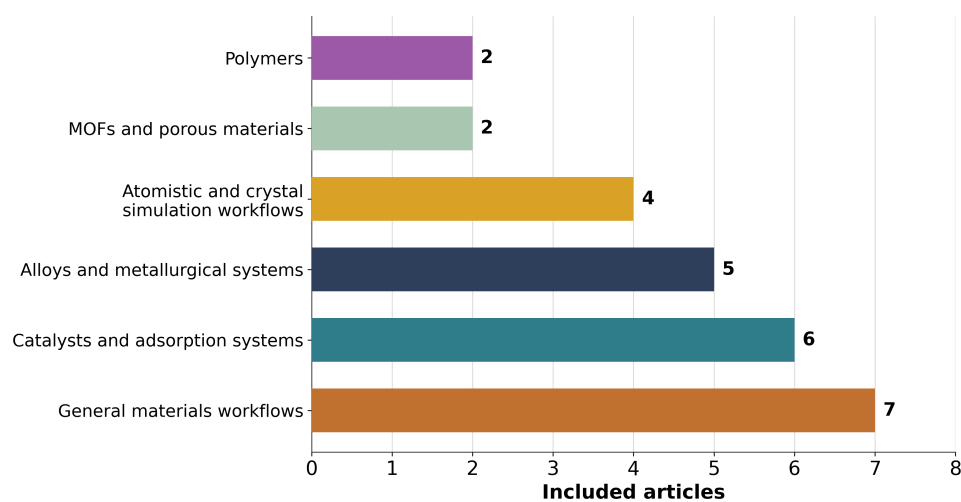


Figure 4. Distribution of the final corpus by primary material or application group.

3.3. Agent Architectures and Tool Integration

Specialized multi-agent architectures were the dominant pattern, appearing in 13 of the 26 included studies. These papers typically split the workflow into role-specific agents for planning, structure retrieval, simulation execution, analysis, and quality control [18,19,21,25,26,35–38,41,42]. A second group of studies used hybrid human-in-the-loop designs, where the agentic layer guided design or analysis but final judgment remained partly manual [20,22,27,30,31]. A third group relied on more compact planner–executor or single-agent loops that repeatedly called tools and updated decisions from the observed outputs [24,28,29,32–34,39,40]. Across all three groups, the common pattern was not unrestricted natural-language reasoning; it was tool-grounded orchestration.

Tool grounding was broad but not evenly distributed. Workflow orchestration or integration appeared in 21 of the 26 included studies, molecular dynamics or atomistic simulation environments in ten, materials data or machine learning screening pipelines in eleven, explicit DFT or electronic-structure tools in seven, porous-material curation or screening in two, and robotic execution in one. Several of the strongest simulation-centered papers were explicit about this grounding. VASPIlot autonomously configured and executed VASP studies across convergence, band-structure, and lattice-optimization tasks [25]. The end-to-end atomistic framework in *Digital Discovery* linked structure generation, potential discovery, LAMMPS execution, OVITO analysis, and phonon post-processing in a single agentic pipeline [36], while GENIUS benchmarked autonomous Quantum ESPRESSO protocol generation and repair at scale [42]. By contrast, papers such as the electrocatalyst-informatics, biomass-screening, and modular multi-task materials agent studies [19,30,41] were agentic in a broader but still relevant sense: they coordinated retrieval, structured extraction, ranking, simulation setup, and prediction rather than a single first-principles simulator. Figure 5 summarizes tool-family use across the main architecture patterns.

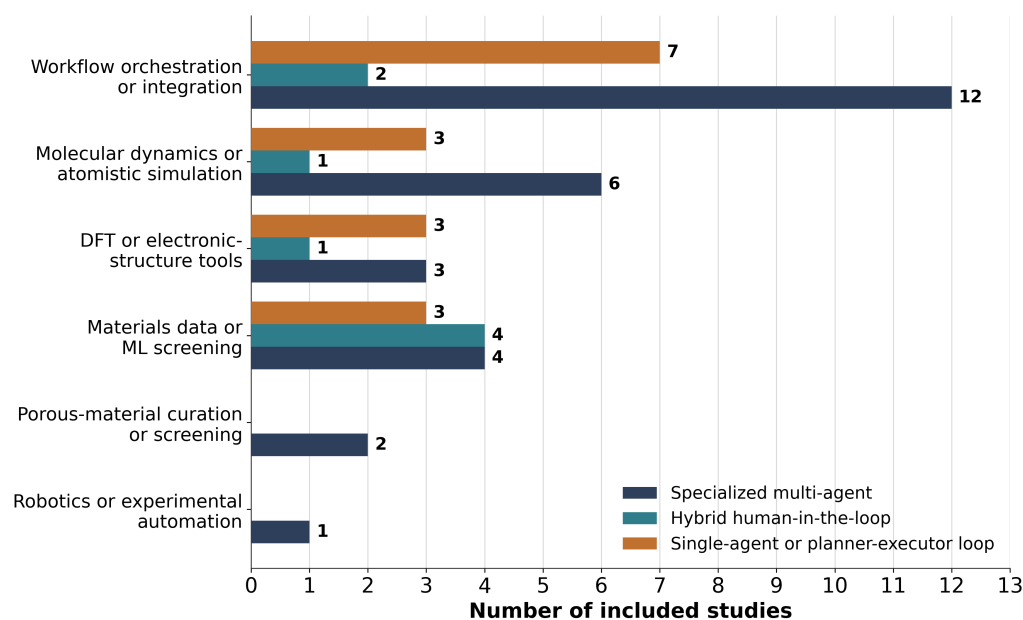


Figure 5. Distribution of tool-family use across the major architecture patterns observed in the included studies.

3.4. Validation Patterns, Quantitative Evidence, and Reproducibility

Validation quality was heterogeneous. Some papers offered benchmark-style quantitative evidence. Adsorb-Agent recovered comparable adsorption energies in 83.7% of benchmark systems and better energies in 35.0% while using only 26.9% of the candidate configurations on average [28]. The crystal-generation worker–supervisor frame-

work reduced execution-level hallucinations from 100% to as low as 5% and reached 82% structural match accuracy for fully specified prompts [32]. The alloy-optimization loop improved tensile-strength estimates from starting values of 8.26–9.81 GPa to maxima of 13.31–18.70 GPa [33]. The single-atom-alloy catalyst agent linked a structural descriptor to adsorption with $R^2 = 0.94$ and RMSE = 0.20 eV, and then experimentally validated a YCu-SAA catalyst at 60% FE for C2+ products [34]. GENIUS passed early execution validation for about 80% of 295 benchmark prompts and autonomously recovered 76.3% of initially failed cases [42], while MatAgent completed 94–96% of property-retrieval tasks and 89–91% of property-prediction tasks across three 100-question rounds [40]. Other papers were more demonstrative, emphasizing workflow feasibility, design navigation, or case-based success rather than tightly standardized metrics [17,18,21,22,35,37–39,41].

Reproducibility signals were present but uneven. Public repositories or archival artifacts were identified in 18 of the 26 included studies. Experimental validation appeared in six studies, while high-autonomy workflows were limited to six; the remaining corpus was concentrated in medium-high (twelve) and medium (eight) autonomy. Only one article demonstrated genuine robotic closed-loop execution [35]. In several papers, the strongest reproducibility evidence came from public code, tutorials, datasets, or supplementary traces [18–21,23,25,28,30,31,35–42]. In others, the workflow could be reconstructed only partially from supporting information, and prompt logs, failure traces, or decision histories remained incomplete [17,22,24,26,27,29,33,34].

Figure 6 summarizes validation, reproducibility, and deployment-maturity signals observed across the final corpus. Table 2 summarizes the main quantitative evidence, reproducibility artifacts, and study-level limitations reported by the included corpus.

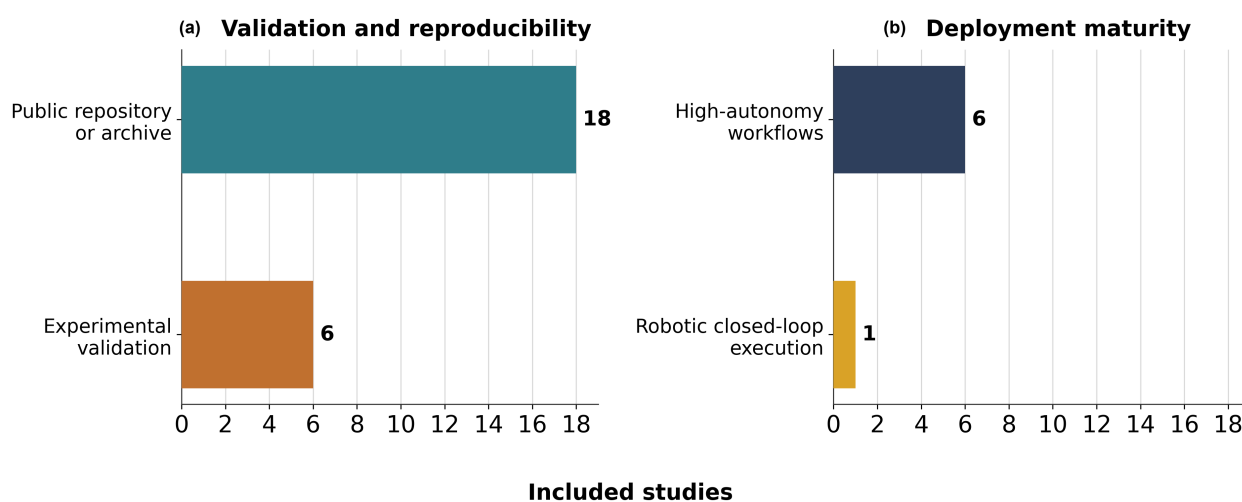


Figure 6. Validation, reproducibility, and deployment-maturity signals observed across the final combined corpus. (a) Validation basis and autonomy profile. (b) Reproducibility and deployment signals.

Table 2. Quantitative highlights, reproducibility artifacts, and main limitations reported by the included studies.

Ref.	Quantitative Highlight	Reproducibility Artifacts	Main Limitation
[17]	Autonomously executed an O ₂ scan from 0.7 to 1.8 A in 0.1 A steps and identified a minimum near 1.3 A at −148.0842 Hartree.	Conversation traces, JSON/SVG artifacts, reusable generated code; no dedicated public repository identified.	Demonstration-oriented validation rather than benchmarking against a standardized materials-discovery corpus.

Table 2. Cont.

Ref.	Quantitative Highlight	Reproducibility Artifacts	Main Limitation
[18]	Explored Nb concentration from 0 to 100 at 20-point intervals; used five random configurations by default and up to twenty for NEB tasks; identified NbMo alloys with >60% Nb as ductile in one crack-emission system.	Public repository (lamm-mit/AtomAgents), workflow figures, supplementary plans, and named tools.	Several reasoning steps still required user correction or in-context patching, especially for multimodal interpretation.
[19]	Analyzed 2321 papers; manual validation on 50 abstracts yielded precision 0.97, recall 0.95, and F1 0.96; final ML set contained 1117 entries and 159 descriptors; yield-classifier test F1 was 0.63 and FE regression R2 was 0.12.	Public GitHub repository/library (eNRRCrew), open supplementary prompts, scripts, and workflow details.	Faradaic-efficiency prediction remained weak because of dataset heterogeneity and limited high-performance examples.
[20]	Baseline campaign covered 30 synthesis conditions with conversion from about 5% to 100%; XGBoost achieved R2 = 0.83; optimized Ni@UiO-66(Ce)-R2T1 reached 100% conversion and 100% selectivity at 70 C and 2 MPa in 2 h; GPT-4o obtained the highest expert score (84.08).	Public GitHub code, curated question–answer dataset, 100-question exam set, and extensive supporting information.	Evidence is centered on one catalyst family, so cross-chemistry generalization remains uncertain.
[21]	Framework extension typically required only three targeted code modifications; one case built a system with one IRMOF-1 and twenty ethanol molecules; a water-MD example ran stably for 5 ns.	Public GitHub repository with source code, tutorials, and instructions.	Automation remains partial for fully custom LAMMPS inputs and still benefits from template support.
[22]	Recommended an Sn:Ge:Sb:Te ratio window of 0.6–0.7:0.3–0.4:0.01–0.16:1 and Cu ₂ Te alloying of 2–5 mol%; reported zT = 1.2, about 40% above the Sb-doped SnTe baseline.	Extensive supporting information with prompts, ablation notes, and comparison tables; no public code repository identified.	Evidence is SI-heavy and recommendation-driven, so autonomous execution is less explicit than in code-first agent papers.
[23]	Each composition used 50 random solute realizations; the Peierls-barrier model reached MAE = 37 meV and R2 = 0.97; the potential-energy-change model reached MAE = 60 meV and R2 = 0.89; ternary exploration covered 231 compositions at 5% intervals.	Public code/data repository, named software stack, and detailed supplementary analysis traces.	The paper reports inconsistencies in some multimodal plot interpretations, so the reasoning layer still lags behind the physics surrogate.

Table 2. Cont.

Ref.	Quantitative Highlight	Reproducibility Artifacts	Main Limitation
[24]	For the viscosity task, the neural network achieved $R^2 = 0.9589$ and $RMSE = 55.7754$, versus $R^2 = 0.9236$ and $RMSE = 76.0998$ for random forest; the freely jointed chain study used $N = 10$ – 1000 with 1000 chains per N and recovered the expected scaling law.	Supplementary AI-generated reports, named AutoProg script, and detailed research plans and figures.	The extracted full text is largely supplementary and some plotting still required manual intervention.
[25]	Completed band structure, density of states, ENCUT convergence, lattice optimization, and cross-material band-gap tasks without manual intervention; ENCUT benchmark covered 300–500 eV; reported band gaps were 1.213 eV (MoS ₂), 1.166 eV (MoSe ₂), 1.242 eV (WS ₂), and 1.065 eV (WSe ₂).	Open-source GitHub repository and documentation.	Multi-turn execution carries substantial token overhead, and the authors still emphasize the need for human oversight for scientific correctness.
[26]	PASMUT reached recovery = 0.886, purity = 0.848, and exergy efficiency = 0.068; AZIXUD reached purity = 0.846, recovery = 0.815, and efficiency = 0.083; RAHBUZ reached purity = 0.825, recovery = 0.831, and efficiency = 0.069.	Explicit top-ten scoring workflow, figures, agent roles, and supplementary outputs; data on request.	The workflow depends on simplified process models and prompt-weighted semantic arbitration rather than experimental verification.
[27]	Reported continuous operation for >100 h; DFT used a 400 eV plane-wave cutoff; MD ran for 50 ps with three independent repeats; EXAFS fitting gave Mn–O $\sim 1.88 \pm 0.02$ Å and Mn–Ni $\sim 3.08 \pm 0.02$ Å.	Extensive supplementary workflow figures, spectra, EXAFS tables, and Materials Project IDs; no public repository identified.	The catalytic performance uplift is harder to reconstruct from the extracted full text than the mechanistic workflow itself.
[28]	Found comparable adsorption energies in 83.7% of cases and lower energies in 35.0%; used only 6.8–63.6% of the initial configurations required by the algorithmic baseline, with an average of 26.9%; lower-energy discovery on intermetallics rose to 46.7%.	Public GitHub repository, Zenodo archive, and explicit benchmark metric definitions.	Evaluation is benchmarked against GNN-based relaxation with a 0.1 eV tolerance rather than full DFT endpoints.
[29]	Vacancies were reduced from 109 to 23 in a Co-rich case, 177 to 42 in a Cr-rich case, 162 to 20 in an Fe-rich case, and 203 to 15 for the equiatomic starting alloy.	trajectory.csv traces and Supplementary S1–S6 decision logs from the cited study.	Only one PKA direction per composition was used and optimization targeted vacancy count alone.

Table 2. Cont.

Ref.	Quantitative Highlight	Reproducibility Artifacts	Main Limitation
[30]	Mistral reached a 100% feature-match rate versus 60% for Gemma 12B; reasoning-chain prompting improved precision from 0.600 to 0.818; RAG improved F1 from 0.364 to 0.880 while shortening responses from 274 to 74 words; AutoSklearn achieved $R^2 = 0.7379$ and $RMSE = 1.1513$.	Partial GitHub demo dataset/repository and detailed benchmarking figures in the SI.	The full dataset is not public because of NDA restrictions and human review remains part of the decision loop.
[31]	Produced suitable lack-of-fusion maps for 11/12 known-alloy prompts, 8/9 property-driven comparisons, and 8/10 novel-alloy prompts; predicted windows for Inconel 625 encompassed literature conditions from $175\text{ W}/500\text{ mm s}^{-1}$ to $1000\text{ W}/600\text{ mm s}^{-1}$.	Public GitHub dataset link, appendices documenting tool implementation, and detailed process maps.	The physics only covers lack of fusion and the paper documents at least one LLM misinterpretation of returned tuples.
[32]	Fine-tuning reduced hallucination rates from 100% to 5–14%; best overall structural match was 36%; fully specified prompts reached 82% match accuracy with 2% execution failure.	GitHub/Hugging Face release path, supplementary material, and described code/data availability.	Performance is sensitive to prompt specificity and still far from universally reliable crystal generation.
[33]	Improved tensile strength from starting values of 8.26–9.81 GPa to maxima of 13.31–18.70 GPa; the system reported nearly 50% strength gains and adaptively changed strategy within the same optimization run.	Detailed full-text trajectory descriptions and supplementary figures; no dedicated public repository identified.	High strain-rate MD means the reported strengths indicate trends rather than directly transferable macroscopic values.
[34]	The structural descriptor ϕ correlated with CO^* adsorption at the guest metal, where * denotes an adsorbed surface species, with $R^2 = 0.94$ and $RMSE = 0.20\text{ eV}$; the experimentally tested YCu-SAA reached 60% FE for C2^+ products at -1.04 V vs. RHE and 15.2 mA cm^{-2} .	Article and supplementary data availability statements; no dedicated public code repository reported.	The broader agent workflow is stronger on design-principle extraction than on standardized benchmarking of autonomy.
[35]	Optimized perovskite nanocrystal synthesis to blue emission within 10 iterations and 3.5 h versus conventional timelines of 4–6 months; achieved 60.02% and 61.89% benchmark win rates in two comparisons.	Public GitHub repository for MARS, source data, and Python scripts.	The system is exceptional rather than typical of the corpus, and its experimental focus extends beyond simulation-centric workflows.

Table 2. Cont.

Ref.	Quantitative Highlight	Reproducibility Artifacts	Main Limitation
[36]	Recovered exact lattice constants for 5/6 benchmark systems and exact cohesive energies for 5/6; reported only 0.018% error for Au-Cu lattice parameter prediction; NiCu melting point was predicted around 1604 K.	GitHub plus Zenodo release with code, source data, and representative outputs.	The paper still represents a controlled benchmark set rather than an unrestricted materials-discovery environment.
[37]	Benchmarked 65 MOFs and 321 adsorption points; context-aware preprocessing reduced distortions from asymmetric cluster removal and improved agreement with reported adsorption metrics.	Public code support reported by the authors and supplementary data package.	The paper focuses on preprocessing and curation rather than on full downstream discovery loops.
[38]	Demonstrated both interactive and autonomous operation across five polymer topology classes and multiple solvent/simulation conditions; emphasis was on workflow completeness rather than a single benchmark score.	Public GitHub repository with DOI-linked Zenodo archive.	Quantitative benchmarking is lighter than in the strongest atomistic or catalyst papers.
[39]	Predicted a direct band gap of 2.34 eV for monolayer MoS ₂ using a 500 eV cutoff and HSE06 refinement; inverse-design cases generated 28, 25, and 28 molecules, with 11, 8, and 10 novel candidates, respectively.	Code demo available on GitHub, structured JSON outputs, and supplementary workflow descriptions.	Validation is demonstration-oriented and spans only two representative use cases rather than a standardized benchmark suite.
[40]	Across three 100-question runs, property-retrieval success was 95%, 94%, and 96%, while property-prediction success was 90%, 89%, and 91%; the study also reports precision, recall, and F1 evaluation across both task types.	Open GitHub repository (Xiaobu-USTC/MatAgent), article data availability statement, and task comparison figures.	The evaluation focuses on controlled prompt tasks rather than open-ended discovery campaigns or multi-step benchmark workflows.
[41]	Parameter extraction succeeded in 100% of five repeated runs, materials extraction was consistent in 9 of 10 runs, and one reliability analysis reported 80% output-format consistency with 100% content accuracy.	Supplementary data from the cited study for case studies and reliability tests plus public code repository (cakshat/MatSci-LLM-Agents).	Validation centers on modular task reliability rather than end-to-end closed-loop discovery under a shared benchmark.
[42]	Passed early execution validation for about 80% of 295 benchmark prompts; zero-shot success was 14.2%, and 76.3% of initially failed cases were autonomously recovered; the knowledge graph encoded 247 parameters and 330 dependency edges.	Public GitHub repository with code, data, and knowledge-graph schema (KIT-Workflows/agent-workflow-framework).	The benchmark targets Quantum ESPRESSO protocol generation and repair, so transfer to other electronic-structure ecosystems remains to be demonstrated.

3.5. Metadata-Derived Thematic Emphasis

Because author-keyword fields were inconsistently populated across the included records, an auxiliary metadata-derived term-frequency profile was constructed from normalized title terms supplemented by any available author-keyword fields. Figure 7 shows that alloy- and catalyst-centered studies remain the dominant domain-oriented signals, while MOF-, polymer-, atomistic-, and crystal-related terms form smaller but visible thematic niches. This profile is consistent with the domain analysis above and helps distinguish where the field is already concentrated from where it is still exploratory.

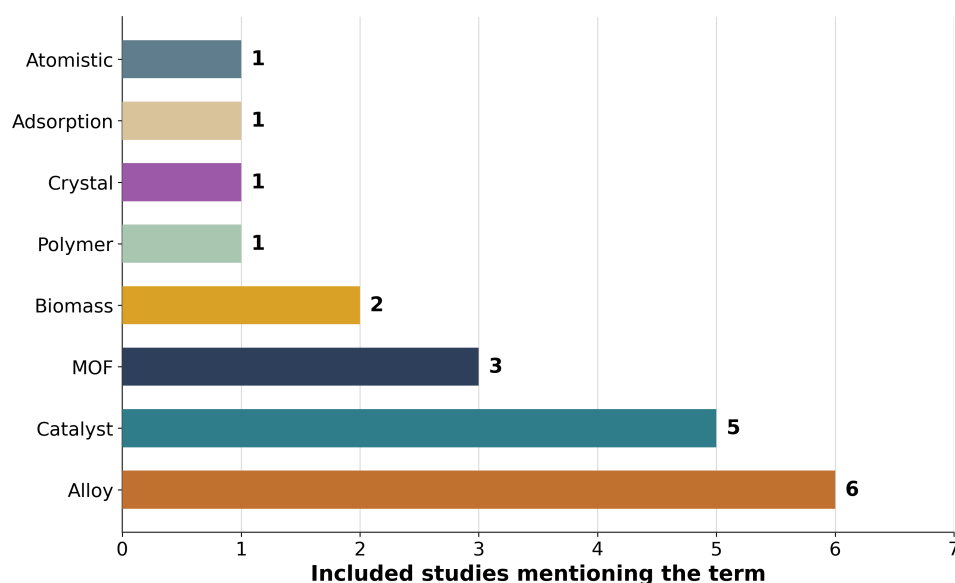


Figure 7. Frequency of normalized thematic terms derived from titles and any available author-keyword metadata in the included corpus.

4. Discussion

4.1. From Model-Centric AI to Workflow-Centric AI

The clearest message from the corpus is that the field is moving from model-centric AI toward workflow-centric AI. Earlier materials acceleration agendas emphasized prediction, inverse design, or generalized automation [2–4]. The literature synthesized here shows something more specific: the scientific contribution increasingly lies in how agents coordinate heterogeneous steps across databases, code, simulators, schedulers, and analysis routines, not simply in what a single model predicts. Twenty-one of the twenty-six included studies contained an explicit orchestration or integration layer, and no eligible study was found before 2024, which indicates that this literature is not a long-standing subfield but a recent architectural shift. That pattern aligns with broader discussions of tool-augmented language models and autonomous chemistry agents [5–10], but the final corpus is more operationally concrete than much of the perspective literature because it documents how agentic systems are embedded inside executable materials workflows and judged against tangible computational or experimental outputs. Across the included studies, agentic functions fall into three recurrent roles: infrastructure and orchestration layers that manage heterogeneous software stacks [21,25,36–42], search-space reduction loops that accelerate candidate evaluation or optimization [28,29,33], and discovery systems that carry computational reasoning far enough to support experimentally relevant decisions [20,22,27,34,35,39].

4.2. *Why Tool Grounding Matters*

A second important pattern is that tool grounding remains the main credibility mechanism. A significant selection effect is observed: the eligibility criteria require a simulation, computational, or domain-specific tool component, resulting in a corpus composed solely of tool-based studies. The importance of tool-based grounding as a mechanism for establishing credibility is, therefore, partly a consequence of the inclusion criteria and should not be interpreted as indicating that tool-based grounding is more prevalent than non-tool-based methods in the general literature. Our claim is more specific: within tool-based agent materials workflows, credibility is based on the tool interface, rather than on unstructured text. The most persuasive papers did not treat the language model as a stand-alone scientific authority. Instead, they linked it to DFT, molecular dynamics, adsorption screening, process simulation, CALPHAD, crystallographic curation, or robotic execution [18,21,25,28,29,31–36,39–42]. This grounding matters because it converts abstract reasoning into constrained action and creates natural checkpoints at which the workflow can be inspected, corrected, or benchmarked. It also reveals that not all agentic materials papers should be interpreted in the same way. Some studies are simulator-native: the agent configures or executes atomistic or thermodynamic tools directly [21,25,28,29,31,33,36,38–42]. Others act more as discovery managers around retrieval, ranking, surrogate models, or protocol navigation [19,20,26,30,41]. Both modes are scientifically relevant, but they rely on different evidentiary anchors. In practical terms, that distinction is what separates a scientific workflow from a fluent assistant, and it explains why explicit DFT-coupled systems, although still a minority, are among the most convincing demonstrations in the corpus. It also brings into view a form of hidden infrastructure that is easy to underestimate: CIF cleaning, input-file generation, job submission, error recovery, post-processing, and software-to-software translation. Papers such as Smart-CIF, DynaMate, VASPIlot, Matty, GENIUS, and ToPolyAgent show that a substantial share of the current advance lies not in replacing scientific judgment with text generation, but in making these brittle infrastructural steps programmable and recoverable [21,25,37–39,42].

4.3. *Autonomy Is Advancing, but Still Uneven*

At the same time, autonomy remains uneven. Only six studies met the criteria for high autonomy, whereas twelve were classified as medium–high and eight as medium. Only one study demonstrated full robotic closed-loop execution [35]; a single case is reported as an existence proof rather than a field-level measure. The distribution is also domain-sensitive. Hybrid human-in-the-loop systems accounted for four of the six experimentally validated papers, which suggests that human oversight appears to be retained more often where the downstream consequences of an incorrect decision are more costly or irreversible, although the small sample precludes strong inference [20,22,27,30,31]. This should not be read as a failure of the field. It indicates that, at present, agentic materials discovery is strongest when agents handle search, coordination, and execution while experts retain authority over framing, feasibility, and interpretation. A central tension in the corpus is that operational autonomy and materials-level validation do not peak in the same studies: the clearest simulator orchestration papers benchmark execution, search efficiency, or error reduction with unusual precision, whereas the papers with the clearest experimental transfer often remain hybrid and domain-scaffolded. The literature therefore shows more evidence of competent orchestration than of fully self-driving scientific autonomy, and it is best interpreted along separate axes of operational autonomy, scientific fidelity, and discovery transfer.

4.4. Reproducibility Remains a Methodological Fault Line

Reproducibility is another dividing line between promising demonstrations and reusable research infrastructure. Public repositories or archival artifacts were present in 18 of the 26 included studies, but their depth varied substantially. Repository support was more common in specialized multi-agent systems than in compact single-agent loops, yet even in the better-documented papers, prompt logs, intervention histories, scheduler settings, failure traces, and environment specifications were often incomplete. For tool-grounded systems, reproducibility cannot be reduced to source code alone. It also depends on the versioned interfaces through which the agent calls external tools, the templates used to prepare inputs, the repair logic used after failures, the stopping criteria for iterative loops, and the computational budgets under which the workflow was evaluated. The broader literature on autonomous scientific systems has already argued for more rigorous evaluation protocols and transparent tool interfaces [3,5,7–12]. The present review reinforces that argument from a materials perspective: without stronger reporting of trajectories, interventions, benchmark criteria, and environment-level dependencies, it remains difficult to compare systems fairly or to distinguish genuine autonomy from carefully curated demonstrations. A useful minimum reporting matrix would separate operational metrics such as task completion, recovery from tool failure, and iteration cost; scientific-fidelity metrics such as agreement with DFT, MD, process simulation, or experiment; and discovery-utility metrics such as candidate improvement, screening efficiency, or successful experimental transfer.

4.5. Research Agenda and Domain Implications

The domain distribution helps explain where the field currently gains traction. Catalysis and adsorption systems ($n = 6$) and alloys and metallurgical systems ($n = 5$) dominate because both areas offer relatively clear quantitative loops between structured data, simulation, mechanistic hypotheses, and optimization. Four of the six experimentally validated studies belonged to catalysis and adsorption, whereas alloy studies more often demonstrated computational optimization than experimental closure. We hypothesise, rather than conclude, that this asymmetry may reflect more mature descriptor cultures, stronger mechanistic priors, and more established simulation-to-experiment pathways in catalysis; given the small denominators this remains a small-sample observational hypothesis. Alloy-design papers, by contrast, often achieve impressive autonomous optimization within molecular dynamics or CALPHAD-like environments, but their translation to experimentally confirmed materials performance remains less common. Smaller subliterations in polymers, MOF curation, and end-to-end atomistic workflows show that the field is broadening, but they have not yet reached the same density of validation or methodological convergence. The next stage of progress is therefore unlikely to come from generic claims about AI agents in materials science. It is more likely to come from deepening a smaller set of tool-rich domains in which the agent can repeatedly act on quantitative feedback and where domain-specific benchmarks can be standardized. Importantly, the corpus already demonstrates several concrete successes: autonomous VASP execution across routine electronic-structure tasks [25], large-scale Quantum ESPRESSO protocol generation and repair [42], substantial reduction of adsorption-search effort without collapsing energy quality [28], adaptive alloy optimization under repeated MD feedback [33], catalyst discovery with experimental confirmation [34], and one genuine multi-agent robotic closure [35]. These are not yet a mature ecosystem, but they do show that the field has progressed beyond speculative proof-of-concept rhetoric.

Several priorities follow from this synthesis. First, benchmark tasks should be defined more explicitly, especially for simulator setup, recovery from failure, adsorption search, multistep optimization, and cross-tool orchestration under realistic compute constraints.

Second, reproducibility packages should include prompts, tool traces, code, representative failures, and environment specifications rather than source code alone. Third, papers should report autonomy boundaries more carefully, including when humans intervene, which decisions remain manual, and how much wall time, API usage, or high-performance-computing effort is required to complete a workflow. Fourth, the field would benefit from a clearer taxonomy that distinguishes literature assistants, workflow managers, simulator executors, and genuinely closed-loop optimizers rather than placing them all under a single agentic label. Fifth, opportunities for the next wave of work lie in portability across tool environments, broader coverage of underrepresented material classes, and agent designs that can transfer beyond one benchmark workflow or one research group's software stack. Finally, experimental validation must become more common, because computational success without materials-level confirmation is not enough to establish scientific utility. These priorities do not weaken the present literature. They identify the conditions under which agentic workflows could mature from impressive case studies into trusted research instruments.

5. Conclusions

This review identifies a young but rapidly expanding literature on agentic workflows for simulation-driven materials discovery. Within a dual-database five-year window, the final evidence base emerged only in 2024 and accelerated sharply in 2025–2026, with catalysis, adsorption, and alloy design as the dominant application areas. The strongest papers were not those making the broadest claims about artificial intelligence. They were the ones that coupled language model reasoning to concrete computational tools, explicit validation logic, and at least partially reproducible artifacts. The field is therefore already past the stage of purely speculative discussion, but it has not yet reached methodological standardization. The most defensible conclusion is that agentic AI is becoming a practical architecture for orchestrating materials workflows, especially where search, simulation, and analysis can be iterated in a tool-grounded loop and evaluated against physically meaningful endpoints. Future progress will depend less on larger models alone than on stronger benchmarking, sharper reporting discipline, clearer autonomy boundaries, deeper physical grounding, and more frequent translation from computational success to experimentally meaningful materials outcomes.

Author Contributions: Conceptualization, A.A.-M.; methodology, A.A.-M.; software, A.A.-M.; validation, A.A.-M., A.V.-A., S.R.-F. and F.D.; formal analysis, A.A.-M., A.V.-A., S.R.-F. and F.D.; investigation, A.A.-M., A.V.-A., S.R.-F. and F.D.; resources, A.A.-M., A.V.-A., S.R.-F. and F.D.; data curation, A.A.-M.; writing—original draft preparation, A.A.-M.; writing—review and editing, A.A.-M., A.V.-A., S.R.-F. and F.D.; visualization, A.A.-M.; supervision, A.A.-M.; project administration, A.A.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The complete data extraction and selection decisions underlying this systematic review are documented in the manuscript, including Appendix A and Table A1. Additional derived files generated during the analysis can be made available by the corresponding author upon reasonable request.

Acknowledgments: During the preparation of this manuscript, the authors used cloud-based AI-assisted language tools for final language polishing, including typographical, wording, and grammatical revisions. The authors reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Record-Level Exclusions Prior to Full-Text Synthesis

Table A1 reports the records excluded before full-text assessment, together with the stage at which they were removed and the main reason for exclusion. Appendix A is included to make the transition from the deduplicated English-language combined set to the final assessed full-text set fully auditable.

Table A1. Record-level exclusions before full-text assessment.

Screening Stage	Year	Record Title	Main Exclusion Reason
Metadata-level relevance screening	2022	5G network deployment and the associated energy consumption in the UK: A complex systems' exploration	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2022	Complexity Modeling of Corrosion in Carbon Fiber-Reinforced Polymer-Retrofitted Concrete Bridge	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2022	Hybrid simulation modelling of networks of heterogeneous care homes and the interfacility spread of COVID-19 by sharing staff	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2022	Incorporating environmental variability in a spatially-explicit individual-based model of European sea bass	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2022	Low-Level Cefepime Exposure Induces High-Level Resistance in Environmental Bacteria: Molecular Mechanism and Evolutionary Dynamics	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2022	Microclimate Drives Shelter-Seeking Behaviour in Lambing Ewes	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2022	Phase diagram in a one-dimensional civil disorder model	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2022	Slow nucleosome dynamics set the transcriptional speed limit and induce RNA polymerase II traffic jams and bursts	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.

Table A1. *Cont.*

Screening Stage	Year	Record Title	Main Exclusion Reason
Metadata-level relevance screening	2022	Spatial groups and cyclic oscillations induced by positive correlation between moving direction and phase of mobile oscillators	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2022	Thermal Performance of Concrete Slabs Reinforced with Glass Fiber-Reinforced Polymer Bars	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2023	A multiscale orchestrated computational framework to reveal emergent phenomena in neuroblastoma	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2023	Discontinuous Tension-Controlled Transition between Collective Actuations in Active Solids	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2023	Failure Mode-Dependent Behavior of Carbon Fiber-Reinforced Polymer Prestressed Concrete Girders	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2023	Hierarchical defect-induced condensation in active nematics	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2023	Machine Learning-based Secure 5G Network Slicing: A Systematic Literature Review	Review article rather than an original journal study.
Metadata-level relevance screening	2023	Virtual Design of 3D-Printed Bone Tissue Engineered Scaffold Shape Using Mechanobiological Modeling: Relationship of Scaffold Pore Architecture to Bone Tissue Formation	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2024	Artificial intelligence as a catalyst for combustion science and engineering	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2024	Buffering effects of nonspecifically DNA-bound RNA polymerases in bacteria	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.

Table A1. *Cont.*

Screening Stage	Year	Record Title	Main Exclusion Reason
Metadata-level relevance screening	2024	Phase equilibrium in canonical cubic structure I (sI) and II (sII) and hexagonal (sH) gas hydrate solid solutions	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2024	Protective effects of antibiotic resistant bacteria on susceptibles in biofilm: Influential factors, mechanism, and modeling	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2025	AI-Guided Molecular Simulations in VR: Exploring Strategies for Imitation Learning in Hyperdimensional Molecular Systems	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2025	Control Methods for Multi-lane Merging on Arterial Roads in Mountainous Cities Under Connected Vehicle Environments	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2025	Reinforcement learning for traffic signal control: advancing efficiency through hybrid exploration strategies	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2025	The impact of data sharing on the sustainable development of aged care service platforms	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2026	A spatiotemporal cell theory for cooperative emergence in multi-agent reinforcement learning systems	Cooperative reinforcement learning dynamics in generic multi-agent systems, without materials discovery/design or simulation-tool orchestration.
Metadata-level relevance screening	2026	Circular economy in polyisocyanurate-based insulation: lifecycle cost and glycolysis recycling modelling	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2026	Developing a personalised end-to-end optimisation algorithm for smart parking systems	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2026	Discovering new materials knowledge from “old data”	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.

Table A1. *Cont.*

Screening Stage	Year	Record Title	Main Exclusion Reason
Metadata-level relevance screening	2026	Filling the cellulosic bio-economy gap by utilizing a wedge approach combined with stakeholder collaboration	Title and abstract indicated agent-based, autonomous, or simulation language outside the target concept of AI agent orchestration for materials discovery.
Metadata-level relevance screening	2026	LLMs, RAG systems, and agents in crystalline materials discovery and characterization: A systematic review	Systematic review rather than an original journal study.
Title/abstract screening	2022	A physics-informed multi-agents model to predict thermo-oxidative/hydrolytic aging of elastomers	Physics-informed agent-based elastomer aging model; not an AI/LLM agentic discovery workflow.
Title/abstract screening	2022	Agent-Based Simulations Using Genetic Algorithm Calibration: A Children's Services Application	Title and abstract did not provide sufficient evidence of an AI agent workflow coordinating materials discovery with a relevant simulation or computational toolchain.
Title/abstract screening	2023	An Agent-Based Decision Support Framework for a Prospective Analysis of Transport and Heat Electrification in Urban Areas	Urban energy-system agent-based decision support; outside materials discovery/design.
Title/abstract screening	2023	Modeling versatile and dynamic anaerobic metabolism for PAOs/GAOs competition using agent-based model and verification via single cell Raman Micro-spectroscopy	Wastewater metabolism agent-based model; outside materials discovery/design.
Title/abstract screening	2023	Simulation of various biofilm fractal morphologies by agent-based model	Biofilm agent-based morphology simulation; not materials discovery/design.
Title/abstract screening	2024	Augmenting large language models with chemistry tools	Chemistry-tool LLM article; retained as contextual literature, not a materials-discovery corpus article.
Title/abstract screening	2024	Exploring coevolution in the diffusion of green products between consumers and enterprises—An agent-based model of two-layer heterogeneous networks	Green-product diffusion agent-based model; outside materials discovery/design.
Title/abstract screening	2024	Physics-informed models of domain wall dynamics as a route for autonomous domain wall design via reinforcement learning	Autonomous reinforcement learning design article; relevant background but no explicit AI agent workflow in metadata.
Title/abstract screening	2024	ProtAgents: protein discovery via large language model multi-agent collaborations combining physics and machine learning	Protein discovery article; outside materials discovery/design corpus.

Table A1. *Cont.*

Screening Stage	Year	Record Title	Main Exclusion Reason
Title/abstract screening	2025	A Novel Multi-Agent-Based Approach for Train Rescheduling in Large-Scale Railway Networks	Railway-network rescheduling; outside materials discovery/design.
Title/abstract screening	2025	Empowering Financial Decision-Making With AI Agents: A Case Study on Critical Thinking Development	Title and abstract did not provide sufficient evidence of an AI agent workflow coordinating materials discovery with a relevant simulation or computational toolchain.
Title/abstract screening	2025	Generative Design of Functional Metal Complexes Utilizing the Internal Knowledge and Reasoning Capability of Large Language Models	LLM-driven generative design without explicit agentic workflow or simulator orchestration in metadata.
Title/abstract screening	2025	Large language models for reticular chemistry	Reticular chemistry LLM review/context article; not a primary study for PRISMA corpus.
Title/abstract screening	2025	Multi-agent systems for optimising smart energy clusters: A case study on cost and emission reduction in industrial seaport facilities	Smart energy-cluster optimization; outside materials discovery/design.
Title/abstract screening	2026	Agent-Based Discrete Element Modeling of Microbial-Induced Carbonate Precipitation	Discrete-element MICP model; agent-based modeling but not AI agent materials discovery.
Title/abstract screening	2026	Agent-driven multi-scale simulation for predicting the catalytic activity of complexes	Title and abstract did not provide sufficient evidence of an AI agent workflow coordinating materials discovery with a relevant simulation or computational toolchain.
Title/abstract screening	2026	Assessing the impact of rider-only escalator etiquette using agent-based modelling	Escalator etiquette agent-based model; outside materials discovery/design.
Title/abstract screening	2026	Autonomous Agentic AI for Adaptive Cure Optimization and Defect Prevention in Thermoset Polymer Composite Manufacturing	Agentic control of composite curing and defect prevention addresses manufacturing process optimization rather than materials discovery/design.
Title/abstract screening	2026	HSMS-Based Event-Triggered Adaptive Dynamic Programming for Pursuit-Evasion Differential Games of Multiagent Systems	Title and abstract did not provide sufficient evidence of an AI agent workflow coordinating materials discovery with a relevant simulation or computational toolchain.

References

- Butler, K.T.; Davies, D.W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555. [\[CrossRef\]](#) [\[PubMed\]](#)
- Pyzer-Knapp, E.O.; Pitera, J.W.; Staar, P.W.J.; Takeda, S.; Laino, T.; Sanders, D.P.; Sexton, J.; Smith, J.R.; Curioni, A. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Comput. Mater.* **2022**, *8*, 84. [\[CrossRef\]](#)
- Tom, G.; Schmid, S.P.; Baird, S.G.; Cao, Y.; Darvish, K.; Hao, H.; Lo, S.; Pablo-García, S.; Rajaonson, E.M.; Skreta, M.; et al. Self-Driving Laboratories for Chemistry and Materials Science. *Chem. Rev.* **2024**, *124*, 9633–9732. [\[CrossRef\]](#) [\[PubMed\]](#)
- Cheng, M.; Fu, C.-L.; Okabe, R.; Chotrattanapituk, A.; Boonkird, A.; Hung, N.T.; Li, M. Artificial intelligence-driven approaches for materials design and discovery. *Nat. Mater.* **2026**, *25*, 174–190. [\[CrossRef\]](#) [\[PubMed\]](#)
- Ramos, M.C.; Collison, C.J.; White, A.D. A review of large language models and autonomous agents in chemistry. *Chem. Sci.* **2025**, *16*, 2514–2572. [\[CrossRef\]](#) [\[PubMed\]](#)
- Yoshikawa, N.; Skreta, M.; Darvish, K.; Arellano-Rubach, S.; Ji, Z.; Kristensen, L.B.; Li, A.Z.; Zhao, Y.; Xu, H.; Kuramshin, A.; et al. Large language models for chemistry robotics. *Auton. Robot.* **2023**, *47*, 1057–1086. [\[CrossRef\]](#)
- Bran, A.M.; Cox, S.; Schilter, O.; Baldassari, C.; White, A.D.; Schwaller, P. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **2024**, *6*, 525–535. [\[CrossRef\]](#) [\[PubMed\]](#)
- Boiko, D.A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous chemical research with large language models. *Nature* **2023**, *624*, 570–578. [\[CrossRef\]](#) [\[PubMed\]](#)
- Jablonka, K.M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. Leveraging large language models for predictive chemistry. *Nat. Mach. Intell.* **2024**, *6*, 161–169. [\[CrossRef\]](#)
- Jablonka, K.M.; Ai, Q.; Al-Feghali, A.; Badhwar, S.; Bocarsly, J.D.; Bran, A.M.; Bringuier, S.; Brinson, L.C.; Choudhary, K.; Circi, D.; et al. 14 examples of how LLMs can transform materials science and chemistry: A reflection on a large language model hackathon. *Digit. Discov.* **2023**, *2*, 1233–1250. [\[CrossRef\]](#) [\[PubMed\]](#)
- Jiang, X.; Wang, W.; Tian, S.; Wang, H.; Lookman, T.; Su, Y. Applications of natural language processing and large language models in materials discovery. *npj Comput. Mater.* **2025**, *11*, 79. [\[CrossRef\]](#)
- Pei, Z.; Yin, J.; Zhang, J. Language models for materials discovery and sustainability: Progress, challenges, and opportunities. *Prog. Mater. Sci.* **2025**, *154*, 101495. [\[CrossRef\]](#)
- Zheng, Z.; Rampal, N.; Inizan, T.J.; Borgs, C.; Chayes, J.T.; Yaghi, O.M. Large language models for reticular chemistry. *Nat. Rev. Mater.* **2025**, *10*, 369–381. [\[CrossRef\]](#)
- De Breuck, P.-P.; Wang, H.-C.; Rignanese, G.-M.; Botti, S.; Marques, M.A.L. Generative AI for crystal structures: A review. *npj Comput. Mater.* **2025**, *11*, 370. [\[CrossRef\]](#)
- Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [\[CrossRef\]](#) [\[PubMed\]](#)
- Page, M.J.; Moher, D.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ* **2021**, *372*, n160. [\[CrossRef\]](#) [\[PubMed\]](#)
- Buehler, M.J. Generative Retrieval-Augmented Ontologic Graph and Multiagent Strategies for Interpretive Large Language Model-Based Materials Design. *ACS Eng. Au* **2024**, *4*, 241–277. [\[CrossRef\]](#) [\[PubMed\]](#)
- Ghafarirollahi, A.; Buehler, M.J. Automating alloy design and discovery with physics-aware multimodal multiagent AI. *Proc. Natl. Acad. Sci. USA* **2025**, *122*, e2414074122. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hu, X.; Chen, S.; Chen, L.; Wang, H.; Zhang, X.; Zhou, Z. Automating structure-activity analysis for electrochemical nitrogen reduction catalyst design through multi-agent collaborations. *Natl. Sci. Rev.* **2025**, *12*, nwaf372. [\[CrossRef\]](#) [\[PubMed\]](#)
- Lin, J.; Zhao, D.; Lu, S.; Li, R.; Xu, X.; Wang, Z.; Li, W.; Ji, Y.; Zhang, C.; Shi, L.; et al. Conversational Large-Language-Model Artificial Intelligence Agent for Accelerated Synthesis of Metal-Organic Frameworks Catalysts in Olefin Hydrogenation. *ACS Nano* **2025**, *19*, 23840–23858. [\[CrossRef\]](#) [\[PubMed\]](#)
- Mendible-Barreto, O.A.; Díaz-Maldonado, M.; Esteva, F.J.C.; Torres, J.E.; Córdova-Figueroa, U.M.; Colón, Y.J. DynaMate: Leveraging AI-agents for customized research workflows. *Mol. Syst. Des. Eng.* **2025**, *10*, 585–598. [\[CrossRef\]](#)
- Su, H.; Lei, S.; Chen, Y.; Lv, Y.; Zhang, Q. Human-Machine Collaborative Design of SnTe-Based Thermoelectric Materials via a Multiagent Framework Leveraging Large Language Models. *ACS Appl. Mater. Interfaces* **2025**, *17*, 63572–63584. [\[CrossRef\]](#) [\[PubMed\]](#)
- Ghafarirollahi, A.; Buehler, M.J. Rapid and automated alloy design with graph neural network-powered large language model-driven multi-agent AI. *MRS Bull.* **2025**, *50*, 1309–1324. [\[CrossRef\]](#)
- Liu, Z.; Chai, Y.; Li, J. Toward Automated Simulation Research Workflow through LLM Prompt Engineering Design. *J. Chem. Inf. Model.* **2025**, *65*, 114–124. [\[CrossRef\]](#) [\[PubMed\]](#)

25. Liu, J.; Zhu, T.; Ye, C.; Fang, Z.; Weng, H.; Wu, Q. VASPIlot: MCP-facilitated multi-agent intelligence for autonomous VASP simulations. *Chin. Phys. B* **2025**, *34*, 117106. [[CrossRef](#)]
26. Li, Y.; Zhao, J.; Cao, X.E.; Li, S. A large language model-based multi-agent methodology for intelligent materials screening: A case study on MOFs for CO₂ capture. *Sep. Purif. Technol.* **2026**, *394*, 137462. [[CrossRef](#)]
27. Zhao, P.; Zhang, Y.; Chen, B.; Zhao, W.; Chen, Z.; Guan, L.; Sun, S.; Zhang, S.; Li, Y.; Peng, W.; et al. AI-Agent-Guided Design of Dual-Scale Modulated Nickel-Based Catalyst with Built-In Electric Field for Enhanced Biomass Electrooxidation. *ACS Nano* **2026**, *20*, 10640–10650. [[CrossRef](#)] [[PubMed](#)]
28. Ock, J.; Meda, R.S.; Vinchurkar, T.; Jadhav, Y.; Farimani, A.B. Adsorb-Agent: Autonomous identification of stable adsorption configurations via a large language model agent. *Digit. Discov.* **2026**, *5*, 617–629. [[CrossRef](#)]
29. Roy, A.; Devanathan, R.; Soulam, A.; Sushko, M.L. Agentic AI for autonomous discovery of radiation-tolerant CoCrFeMnNi high entropy alloys. *Scr. Mater.* **2026**, *279*, 117286. [[CrossRef](#)]
30. Liu, A.; Kim, G.; Song, H. Agentic AI framework for accelerated catalyst design in biomass valorization via small language models and automated machine learning. *Chem. Eng. J.* **2026**, *535*, 175701. [[CrossRef](#)]
31. Pak, P.; Chandrasekhar, A.; Farimani, A.B. Agentic additive manufacturing alloy evaluation. *Addit. Manuf. Lett.* **2026**, *31*, 100355. [[CrossRef](#)]
32. Baibakova, V.; Serov, A. Agentic framework for programmatic crystal structure generation using a fine-tuned worker-supervisor large language model. *Energy AI* **2026**, *24*, 100710. [[CrossRef](#)]
33. Roy, A.; Devanathan, R.; Nartu, M.S.K.K.Y.; Joshi, V.V. Autonomous alloy composition optimization using molecular dynamics guided by a large language model. *Scr. Mater.* **2026**, *272*, 117048. [[CrossRef](#)]
34. Wang, X.; Li, Z.; Zhang, D.; Li, H.; Xu, H.; Cheng, D. Catalysis AI Agent Guides Discovering the Universal Design Principle of Cu-Based Single-Atom Alloy Catalysts for CO₂ Electroreduction. *Angew. Chem. Int. Ed.* **2026**, *65*, e24612. [[CrossRef](#)] [[PubMed](#)]
35. Shi, T.; Li, Y.; Wang, Z.; Xu, W.; Jiang, G.; Dai, D.; Zhou, J.; Huang, H.; He, R.; Ramakrishna, S.; et al. Knowledge-driven autonomous materials research via collaborative multi-agent and robotic system. *Matter* **2026**, *9*, 2102577. [[CrossRef](#)]
36. Vriza, A.; Kornu, U.; Koneru, A.; Chan, H.; Sankaranarayanan, S.K.R.S. Multi-agentic AI framework for end-to-end atomistic simulations. *Digit. Discov.* **2026**, *5*, 440–452. [[CrossRef](#)]
37. Zhang, Q.; Zhang, C.; Wang, L. SmartCIF: A context-aware multi-agent system for automated preprocessing and curation of MOF CIFs. *Phys. Chem. Chem. Phys.* **2026**, *28*, 9566–9575. [[CrossRef](#)] [[PubMed](#)]
38. Ding, L.; Carrillo, J.; Do, C. ToPolyAgent: AI agents for coarse-grained bead-spring topological polymer simulations. *Digit. Discov.* **2026**, *5*, 901–909. [[CrossRef](#)]
39. Wang, X.; Zeng, Q.; Xu, D.; Zhang, L.; Jiang, G.; Yang, M. Accelerating materials discovery via AI-Agent integration of large language models and simulation tools. *J. Mater. Inform.* **2026**, *6*, 9. [[CrossRef](#)]
40. Lv, S.; Peng, L.; Wu, W.; Yao, Y.; Jiao, S.; Hu, W. Bridging language models and computational materials science: A prompt-driven framework for material property prediction. *Mater. Genome Eng. Adv.* **2025**, *3*, e70013. [[CrossRef](#)]
41. Chaudhari, A.; Ock, J.; Farimani, A.B. Modular large language model agents for multi-task computational materials science. *Commun. Mater.* **2026**, *7*, 131. [[CrossRef](#)]
42. Soleymanibrojani, M.; Aydin, R.; Guedes-Sobrinho, D.; Dias, A.C.; Piotrowski, M.J.; Wenzel, W.; Rêgo, C.R.C. GENIUS: An agentic AI framework for autonomous design and execution of simulation protocols. *Commun. Mater.* **2026**, *7*, 115. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.