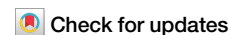


<https://doi.org/10.1038/s43246-026-01105-0>

AI-powered open-source infrastructure for accelerating materials discovery and advanced manufacturing

Mariangeles Salas¹, Anand Singh^{2,3}, Carlos Pignataro^{3,4} & Lokendra Pal^{1,5}✉

Recent advances in artificial intelligence (AI) offer significant opportunities to drive industrial transformation by addressing growing societal demands for products, techno-economic efficiency, and reduced carbon footprints. This review presents a structured framework for building transparent, scalable, and sustainable AI-driven infrastructures spanning conceptualization to commercialization for materials discovery and advanced manufacturing. The framework traces the evolution of materials development from empirical approaches toward integrated AI-enabled platforms, emphasizing open-source tools that unify data acquisition, modeling, simulation, and deployment to democratize access, foster collaboration, and enhance reproducibility. Key enabling components include self-driving laboratories for real-time optimization, advanced computational approaches for high-fidelity data, and blockchain-based mechanisms for secure data sharing, provenance, and supply-chain traceability. The review further discusses the importance of machine learning for materials property prediction, synthesis and process optimization, together with scalable cloud–edge architectures that improve efficiency and reduce latency. Emphasis is placed on lifecycle-aware design, techno-economic analysis, and ethical AI principles to align industrial development with global sustainability goals.

The discovery of new materials has been fundamental to technological advancement. From the Bronze Age to the modern era, materials with novel properties have driven advancements that have shaped human civilization¹. However, the methods for discovering and designing these materials have undergone a remarkable evolution, transitioning through several scientific paradigms: empirical science, model-based discovery, computational science, big data-driven science², and now, generative AI (Fig. 1). This transformation has redefined how researchers explore and optimize the complex design spaces of materials, paving the way for faster and more efficient innovations.

In its earliest stages, materials discovery was rooted in empirical science, where progress was achieved through observation, experimentation, and intuition¹. Researchers relied on trial-and-error methodologies, often requiring decades to identify materials with desirable properties. The advent of theoretical science in the 20th century (Fig. 1) introduced model-based material discovery, where scientific principles such as thermodynamics and quantum mechanics provided a foundation for predicting material behavior^{3,4}. Researchers began using equations and simplified models to guide experiments, reducing reliance on purely empirical methods². This era

saw significant advancements in understanding phase diagrams, crystallography, and other theoretical tools that enabled targeted material design.

The rise of computational science in the latter half of the 20th century marked a turning point². Advances in computing power allowed researchers to simulate material properties at atomic and molecular scales, enabling systematic exploration of complex systems. Computational tools like density functional theory (DFT) or molecular dynamics (MD) became standard, drastically accelerating discovery by testing hypotheses and refining designs virtually before experimental validation^{3,5–8}. However, the scalability of computational methods was limited by the time and resources required for simulations, particularly for high-dimensional problems.

The 21st century led the era of big data-driven science, transforming materials discovery with machine learning (ML) and data analytics². The availability of large experimental and computational datasets allowed researchers to train ML models to predict material properties, optimize synthesis routes, and identify patterns hidden within massive datasets^{9–12}. Unlike earlier methods, data-driven approaches exploit correlations and relationships that are not explicitly captured in theoretical models, enabling rapid screening of vast chemical design spaces. For instance, ML has been

¹Department of Forest Biomaterials, North Carolina State University, Raleigh, NC, USA. ²IBM Corporation; Raleigh, Durham, NC, USA. ³Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC, USA. ⁴Blue Fern Consulting LLC, Raleigh, NC, USA. ⁵Department of Mechanical and Aerospace Engineering, North Carolina State University, Raleigh, NC, USA. ✉e-mail: lpal@ncsu.edu

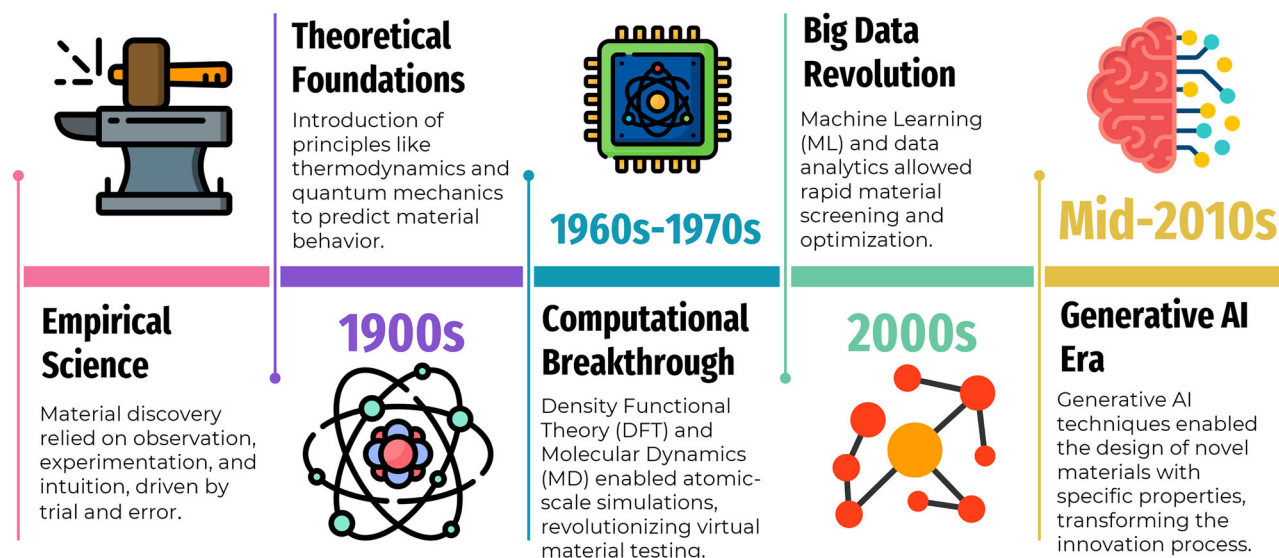


Fig. 1 | Timeline of materials discovery paradigms. Schematic timeline illustrating the evolution of materials discovery, from empirical and theoretical approaches in the early twentieth century, through computational methods and first-principles

modeling in the 1960s–1970s, the emergence of data-driven materials science in the 2000s, and the rise of generative AI-enabled discovery in the mid-2010s.

used to identify promising candidates for renewable energy^{13,14} materials, high-performance alloys^{15,16}, and advanced composites^{17–19}. This data-centric methodology significantly reduced the time and cost associated with traditional and computational approaches.

The most recent breakthrough in materials discovery is the integration of generative AI¹⁵, which moves beyond prediction to actively design new materials. Leveraging generative adversarial networks (GANs) and variational autoencoders (VAEs)^{15,20}, generative AI creates novel material compositions and structures that satisfy specific criteria, such as stability, performance, or sustainability. This represents a fundamental shift in materials discovery, where AI not only accelerates the process but also expands the possibilities for innovation.

Alongside these advancements, open-source platforms have emerged as a transformative force in materials discovery. Open-source platforms provide a collaborative framework where researchers, industries, and governments can access, share, and develop tools, datasets, and methodologies^{21,22}. These platforms foster global collaboration, enabling a diverse community of stakeholders to contribute to materials discovery efforts without being hindered by proprietary barriers. For example, databases like the Materials Project²³, Open Quantum Materials Database (OQMD)²⁴, and NOMAD²⁵ provide extensive repositories of material properties and simulation results, freely accessible to researchers worldwide. These resources serve as the foundation for training ML models and validating generative AI designs.

The significance of open-source platforms extends beyond accessibility, as they promote transparency, reproducibility, and standardization, ensuring that scientific discoveries are reproducible and accessible for independent verification and further advancement^{26,27}. Additionally, open-source platforms align sustainability and ethical responsibility by addressing the environmental impact of energy-intensive workflows through the sharing of optimized algorithms and workflows that reduce computational overhead, thus promoting energy-efficient practices^{28,29}. They also facilitate the development of community-driven guidelines for ethical data use, tackling issues such as bias, privacy, and intellectual property (IP) to ensure that AI-driven materials discovery remains aligned with societal values.

The integration of open-source platforms with the frameworks of Industry 4.0 and Industry 5.0 promises a significant advancement in materials discovery (Fig. 2). Industry 4.0, defined by digitalization, automation, and interconnected systems, enables seamless incorporation of AI into experimental workflows^{30–32}. Innovations like high-throughput

robotics, IoT-enabled devices, cloud computing, and digital twins now support real-time monitoring, precision experimentation, and global collaboration^{33,34}. These tools have transformed the traditional linear discovery pipeline into a dynamic, adaptive system capable of solving complex challenges.

In contrast, Industry 5.0 adopts a human-centric approach, emphasizing collaboration between advanced technologies, such as AI, and human creativity to deliver meaningful solutions³². This paradigm shift in materials discovery not only prioritizes speed and efficiency but also integrates ethical, environmental, and social objectives. For instance, Industry 5.0 emphasizes developing materials that are high-performing yet biodegradable, ethically sourced, and recyclable, underscoring the need to align innovation with sustainability and human values³⁵.

The convergence of these technological advancements brings both transformative opportunities and critical challenges. While AI and Industry 4.0 technologies revolutionize discovery processes, they also raise ethical and environmental concerns, such as biases in large datasets, IP issues, and data privacy risks^{36,37}. Moreover, the computational demands of these systems and their final use can contribute to significant carbon emissions, creating a paradox where tools aimed at solving sustainability problems may inadvertently harm the environment^{38,39}. Industry 5.0 addresses these issues by promoting energy-efficient workflows, transparent data practices, and responsible innovation, offering a pathway to harmonize technological progress with societal and ecological goals^{32,35}.

Taken together, these developments indicate that contemporary materials discovery is governed by the coordinated interaction of data-centric practices, physics-based simulation, and ML, rather than by any single methodological paradigm^{40,41}. Each approach is grounded in distinct assumptions about how knowledge is generated, validated, and generalized. Making these foundations explicit is essential for understanding both the opportunities enabled by AI-driven materials discovery and the structural constraints that limit its applicability.

At a fundamental level, data-centric methods in materials science rest on the premise that data representation, curation, and uncertainty management, rather than model complexity alone, determine what relationships can be learned and trusted^{40,42,43}. Typical data-processing workflows therefore follow a structured lifecycle that begins with data generation from experiments, simulations, literature, and increasingly synthetic sources⁴¹. These raw data undergo preprocessing steps such as cleaning, normalization, labeling, and harmonization to mitigate heterogeneity arising from

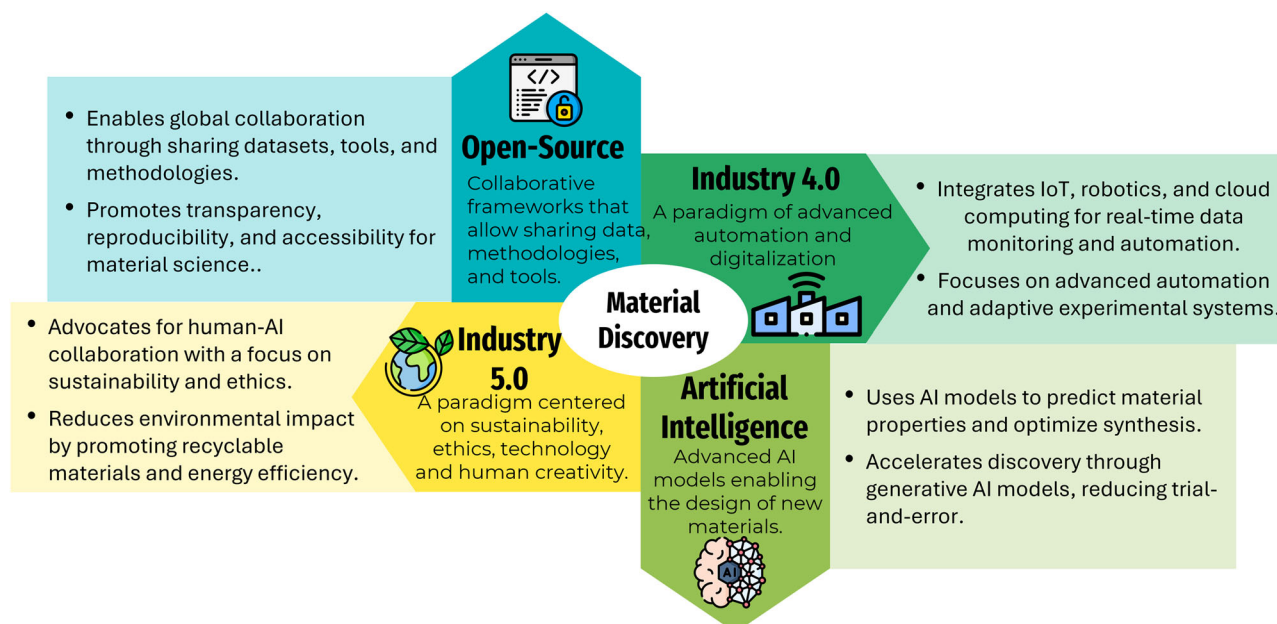


Fig. 2 | Key enablers of modern materials discovery. Conceptual overview of the principal enablers shaping contemporary materials discovery, including open-source frameworks for collaborative research, Industry 4.0–driven automation and

data integration, artificial intelligence for accelerated discovery and generative design, and Industry 5.0 principles emphasizing human-centric and sustainable innovation.

disparate protocols and assumptions^{43,44}. Feature construction and representation implicitly define the hypothesis space accessible to modeling, while uncertainty, bias, and data sparsity impose epistemic limits on downstream simulation and learning-based tasks⁴².

First-principles and atomistic simulation methods occupy a complementary role by explicitly encoding physical laws, enabling causal interpretation, mechanistic insight, and theory-informed validation of material behavior^{45,46}. Techniques such as DFT and MD provide controlled access to energetic landscapes, structural evolution, and kinetic pathways, often serving as reference standards for experimental design and model calibration^{45,47–49}. At the same time, their predictive scope is intrinsically constrained by simplifying assumptions, including exchange–correlation approximations, force-field parameterizations, finite-size effects, and restricted time scales^{50–53}. Combined with the exponential growth of configurational space, these factors limit the scalability of simulation-based exploration to complex compositional and processing regimes^{54–56}.

ML extends this framework by introducing scalable approximation and inference across high-dimensional design spaces, enabling rapid property prediction, surrogate modeling of expensive simulations, and generative exploration beyond enumerative search. Predictive models reduce computational and experimental cost, while generative and agentic approaches support inverse design and adaptive optimization. However, the generalization behavior of ML models remains tightly coupled to data quality, representativeness, and distributional bias, and does not inherently guarantee physical consistency or reliable extrapolation beyond the training domain^{57–61}. Their effective use, therefore, depends on careful integration with curated data pipelines and physics-informed validation rather than standalone deployment.

Viewed collectively, data-processing workflows, simulation methods, and ML form a coupled methodological system in which each paradigm compensates for the limitations of the others. Data infrastructures define the accessible information space, simulations anchor discovery in physical causality, and learning-based models provide the scalability required to navigate vast materials design spaces.

An integrated, systems-level perspective provides the conceptual foundation for the present work, framing materials discovery as a systems-level problem in which data, models, and computational workflows must be

co-designed rather than treated in isolation. By emphasizing the interdependence between methodological choices and infrastructural decisions, it highlights why reproducibility, scalability, and long-term sustainability cannot be addressed at the level of individual algorithms alone, but must instead be embedded within the overall research framework.

This paper offers a comprehensive overview of the technological tools and methodologies available for researchers to design AI-powered, open-source infrastructures for materials discovery. The goal is to provide a clear guide through the key stages of the data lifecycle, spanning data collection, preprocessing, storage, organization, AI modeling, and deployment. By addressing each of these components, the paper equips researchers with the insights needed to create scalable, efficient, and sustainable platforms, ensuring the acceleration of innovation in materials science while aligning with the principles of openness and reproducibility.

The discussion begins by examining traditional and modern data collection techniques, emphasizing the transition from manual experimental approaches to automated, digital systems, including in-silico simulations, synthetic data generation, and web scraping. It then explores preprocessing strategies to standardize and clean diverse datasets, followed by advanced storage solutions that leverage cloud and edge computing for scalability and real-time analysis. Data organization and indexation frameworks are highlighted for their critical role in enabling efficient retrieval and interoperability across various datasets.

Subsequent sections focus on the application of cutting-edge AI modeling and data processing methodologies, emphasizing the integration of open-source frameworks that democratize access to computational tools and foster collaboration. The paper also delves into the deployment of these technologies in real-world scenarios, demonstrating their ability to address pressing challenges in sustainability, energy efficiency, and resource optimization. Additionally, the environmental impact of computationally intensive workflows is analyzed, with a particular focus on strategies to minimize the carbon footprint of AI-driven materials discovery. Together, these elements form a roadmap for researchers seeking to develop innovative, sustainable, and accessible platforms for advancing materials science. Figure 3 presents a structured overview of the manuscript, outlining the scope and organization of the proposed end-to-end AI infrastructure for materials discovery.

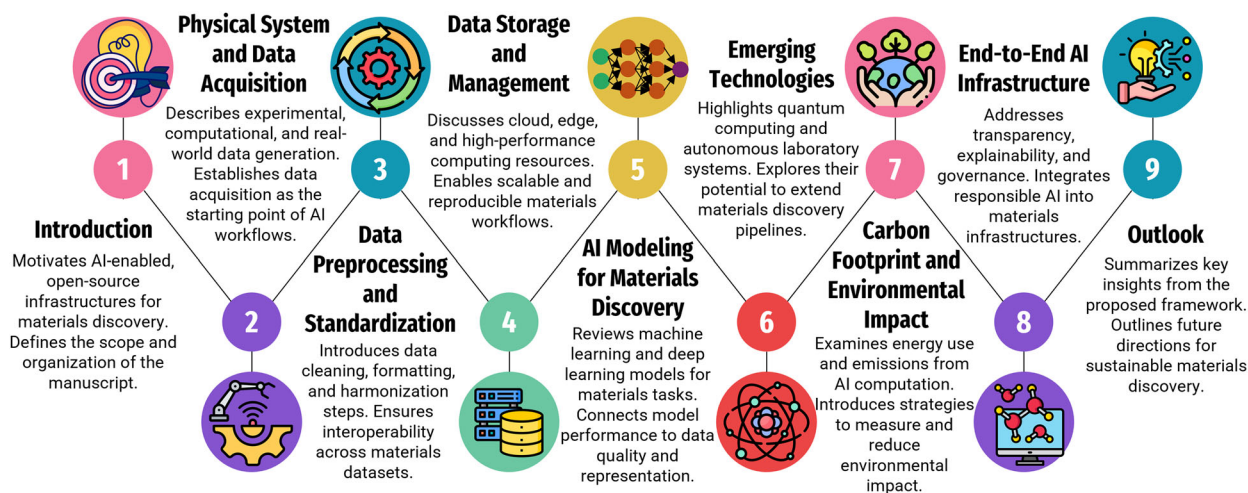


Fig. 3 | Overview of the manuscript structure. Schematic overview of the proposed end-to-end AI infrastructure for materials discovery, illustrating the role of each major section of the manuscript.

Physical system: data collection

Traditional data collection

Traditional data collection in materials discovery has largely relied on experimental methods, where scientists synthesize materials in the lab and observe their properties through trial-and-error approaches. This process involves testing different chemical compositions, processing conditions, and performance metrics like mechanical strength, thermal stability, and environmental impact. Often, these experiments are time-consuming, resource-intensive, and dependent on the expertise of the researchers conducting them^{62,63}. Large datasets from these experiments are compiled over years of iterative trials, making material discovery slow and costly. Additionally, these experiments typically produce small-scale data with limited diversity, hindering the exploration of a wider range of material possibilities.

Moreover, traditional data collection methods often suffer from a lack of standardization across different research groups, which can lead to inconsistencies and difficulties in comparing data⁶⁴. Experimental conditions, such as temperature, pressure, or purity of the chemicals, can vary, affecting the reproducibility of results. Data from academic or industrial research may also remain siloed, limiting collaborative efforts and the sharing of knowledge that could accelerate sustainable materials discovery. Although traditional methods have led to breakthroughs, they lack the speed and scale needed to address modern challenges like climate change, resource scarcity, and environmental degradation in a timely manner. As a result, there is a growing push towards integrating computational tools and large-scale databases to complement traditional experimental methods for more efficient and scalable discovery.

Open-source datasets and databases play a crucial role in the research and discovery of new materials. In the fields of chemistry and materials science, these databases provide key information on the chemical, biological, and structural properties of a wide range of compounds, granting access to experimental data that might otherwise be difficult to obtain. These databases enable researchers to perform complex analyses, run simulations, and conduct virtual screenings of compounds for the development of new materials, pharmaceuticals, and other applications.

There are various open and specialized databases widely used in materials research (Table 1). PubChem⁶⁵ is a comprehensive database that offers detailed information on the chemical properties and biological activities of small molecules, covering over 113 million compounds. Similarly, ChEMBL⁶⁶ provides data on bioactive molecules with drug-like properties, while the Crystallography Open Database (COD)⁶⁷ offers access to crystal structure data for organic, inorganic, and metal-organic compounds. Each database has a specific focus, whether on organic molecules,

crystal structures, or inorganic compounds, and their open access fosters significant collaboration and progress in the field of sustainable materials.

Digital and automated data collection

Smart systems. Self-driving laboratories (SDLs) are revolutionizing scientific research by integrating automation, ML, and robotics to accelerate discoveries in fields such as chemistry and materials science⁶⁸. These cutting-edge systems automate the entire experimental process, enabling the creation of a closed-loop framework where experiments are designed, executed, and optimized without human intervention^{69,70}. This transformative capability allows researchers to navigate vast chemical and material landscapes with unprecedented speed and precision, achieving breakthroughs that would have required significantly more time and effort using traditional laboratory approaches⁶⁸.

As depicted in Fig. 4, SDLs operate within an iterative closed-loop system driven by user-defined goals, predictive modeling through AI and ML, advanced experimental designs, automated synthesis, real-time characterization, and digital feedback for optimization. This workflow not only accelerates discovery cycles but also enhances resource efficiency, sustainability, data integrity, and reproducibility. Moreover, SDLs facilitate global collaboration and knowledge sharing, which are critical for advancing innovation in materials science.

The integration of digital twins, virtual representations of physical laboratory equipment, has further extended the capabilities of SDLs. These digital models enable remote monitoring and adjustment of experiments, fostering seamless collaboration across geographical boundaries. For example, SDLs in laboratories located in different countries, such as Cambridge and Singapore, can collaborate to optimize chemical reactions in real time³⁴. On the other hand, adherence to the FAIR guiding principles (Findable, Accessible, Interoperable, and Reusable) within SDL frameworks not only enhances transparency but also supports effective teamwork, both of which are essential for contemporary research practices⁴⁴.

One of the most significant advantages of SDL technology lies in the reduced cost of automation components, such as robotics and 3D printing, which have made these systems more accessible. Remote control capabilities allow laboratories to operate SDLs over the internet, eliminating the need for substantial infrastructure investments. This democratization of technology accelerates innovation (Fig. 4b) and extends the transformative potential of SDLs to a broader range of research institutions and disciplines⁷¹.

Advancements in digital and automated data collection have also highlighted the critical role of open-source software in laboratory automation. Tools such as Chemsyde, an open-source Python-based interface for Chemspeed Technologies' robotic platforms, exemplify this trend.

Table 1 | Key chemical and biomolecular databases

Dataset/database	Open source	Description	Number of entries	Reference
PubChem	Yes	Comprehensive database providing detailed information on the chemical properties, structures, and biological activities of small molecules, including extensive bioassay data.	113 million compounds and 286 million bioactivities	65
CHEMBL	Yes	Manually curated repository of bioactive molecules with drug-like properties, featuring detailed pharmacological data and bioactivity records.	2.1 million compounds and 17.5 million bioactivities	66
Crystallography Open Database (COD)	Yes	Open-access collection of crystal structure data for organic, inorganic, and metal-organic compounds, supporting research and educational purposes.	461,000 entries	67
ZINC database	Yes	Comprehensive database of commercially available compounds for virtual screening, including purchasable compounds and annotated properties.	230 million purchasable compounds	317
ChemSpider	Yes	An extensive chemical structure database offering fast text and structure search access to chemical entities sourced from hundreds of data providers.	67 million structures	318
Cambridge Structural Database (CSD)	No	An authoritative repository of small-molecule organic and metal-organic crystal structures, widely used in chemical and pharmaceutical research.	1.1 million crystal structures	319
ICSD (inorganic crystal structure database)	No	Specialized database for fully determined inorganic crystal structures, providing detailed crystallographic information for materials science research.	210,000 entries	320
PDB (protein data bank)	Yes	Premier database for the three-dimensional structural data of large biological molecules, including proteins and nucleic acids, crucial for bioinformatics and structural biology.	190,000 structures	321

Summary of representative chemical and biomolecular databases, indicating their open-source status, defining characteristics, and origins in experimental research. The table highlights their role as foundational resources across chemistry, structural biology, pharmacology, and materials science.

Chemspyd offers dynamic control over automated experiments, integrating seamlessly into existing workflows to enhance flexibility in experimental design. This innovation fosters modular laboratory environments, increases the speed and reproducibility of experiments, and enables continuous, 24/7 laboratory operations⁷².

SDLs are playing a pivotal role in the development of clean energy technologies by enabling rapid testing of materials such as advanced photovoltaics and thermoelectrics^{13,73}. Researchers can explore diverse material combinations at an accelerated pace, identifying energy-efficient and cost-effective solutions far faster than traditional methods^{73,74}. The discovery of sustainable materials through SDLs supports global efforts to transition to low-carbon technologies, contributing to long-term sustainability and reduced environmental impact⁷⁵.

A notable example of this transformative capability is the Ada SDL⁷⁶, which leverages automated synthesis and digital feedback loops like those depicted in Fig. 4 to optimize clean energy materials like thin-film solar cells. Ada significantly reduces the need for resource-intensive experimentation, enhancing the performance of energy-harvesting technologies while conserving resources⁷⁶. By facilitating faster innovation in clean energy, SDLs like Ada are propelling the transition to a more sustainable and energy-efficient future.

In parallel with the development of SDLs, the field of intelligent manufacturing has emphasized the use of digital twins and ML to enable predictive, adaptive, and resource-efficient systems. Digital twins are widely employed to represent physical processes and assets, supporting monitoring, optimization, and decision-making under real-world operating constraints. However, when digital twins rely exclusively on data-driven learning models, their predictive reliability may degrade under sparse data conditions, extrapolative regimes, or dynamically changing environments. In such cases, purely statistical models can yield physically inconsistent predictions, including violations of conservation laws or thermodynamic constraints, limiting their robustness and long-term applicability^{77–79}.

Physics-informed ML (PIML), including physics-informed neural networks (PINNs), has emerged as a principled approach to address these limitations by embedding governing physical laws directly into the learning process^{77,78,80,81}. Through constrained architectures, physics-aware feature engineering, or regularized loss functions, PIML enables hybrid modeling paradigms in which experimental data, numerical simulations (e.g., MDs or finite-element methods), and first-principles equations jointly inform model training. These approaches have demonstrated improved data efficiency, interpretability, and robustness across materials-relevant domains, including thermal transport, chemical reaction systems, additive processes, and structure-property modeling under limited or noisy data availability^{78,79,82–84}. By enforcing thermodynamic, mechanical, or transport constraints, PIML-based models reduce nonphysical predictions and improve predictive stability^{85–87}.

These developments are directly relevant to SDLs, where digital models increasingly guide experiment selection, parameter optimization, and closed-loop decision-making. By incorporating physics-informed learning into SDL digital twins, autonomous experimental workflows can achieve greater stability, interpretability, and resilience while reducing failed experiments and unnecessary resource consumption. In this sense, SDLs provide a controlled laboratory-scale environment in which concepts from intelligent manufacturing, such as physics-informed digital twins and sustainability-oriented optimization, can be developed, validated, and refined before translation to larger-scale systems. This positioning allows SDLs to benefit from advances in intelligent manufacturing and physics-informed modeling while maintaining their primary focus on materials discovery and experimental research, and aligns naturally with emerging priorities related to sustainable, reliable, and human-centered autonomous systems.

Synthetic data. High-fidelity computational modeling, or *in silico* simulation, has emerged as a cornerstone for data generation in materials discovery workflows. By providing atomic- to mesoscale-level insights

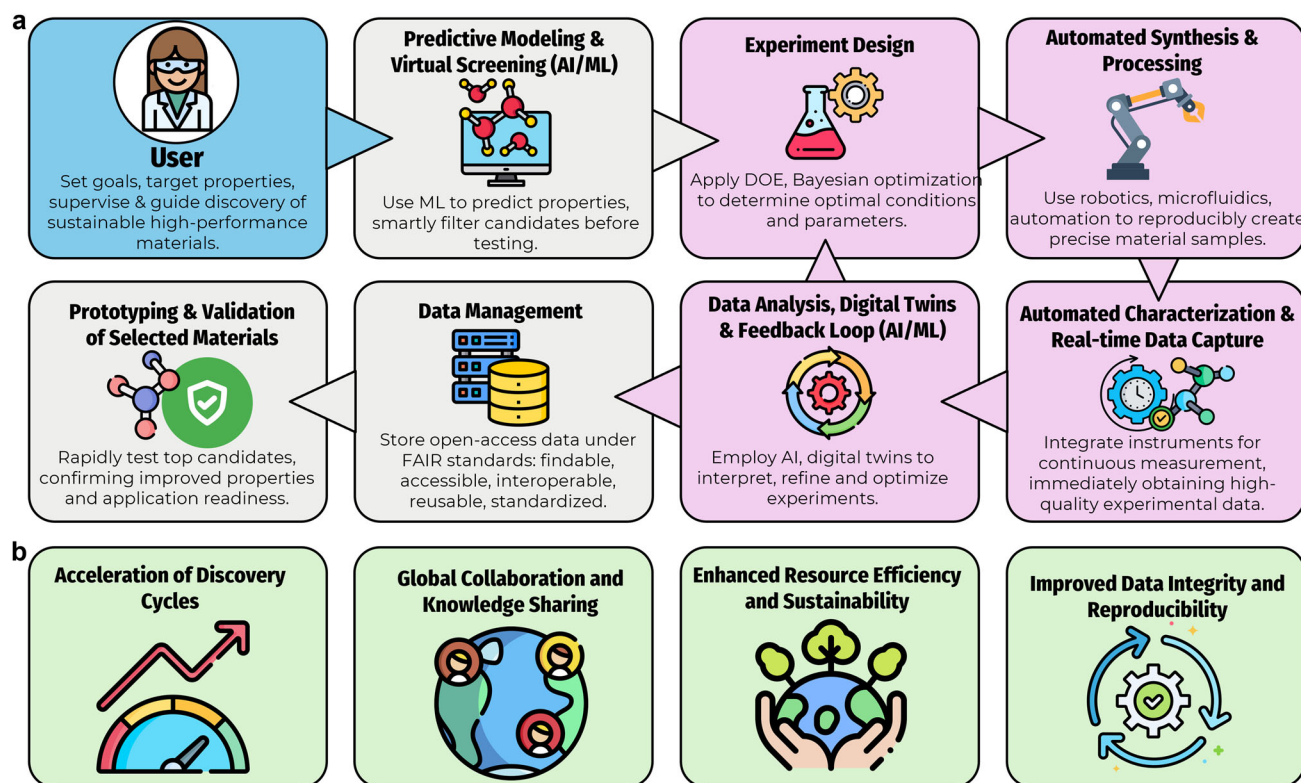


Fig. 4 | Workflow and impacts of self-driving laboratories for materials discovery. **a** Schematic representation of a self-driving laboratory (SDL) operating as a closed-loop system, integrating user-defined objectives, AI/ML-based predictive modeling, experimental design, automated synthesis, real-time characterization,

and digital feedback for iterative optimization. **b** Summary of the primary impacts associated with SDL deployment, including accelerated discovery cycles, improved resource efficiency, enhanced data integrity and reproducibility, and support for distributed collaboration in materials research.

into material properties and behaviors, these simulations complement experimental efforts and greatly expand the accessible design space. Techniques such as MD, first-principles calculations based on DFT, and advanced quantum chemistry methods enable researchers to virtually screen candidate materials, predict phase stability and electronic structure, and characterize interfacial phenomena. The resulting datasets serve as critical inputs for AI models, informing materials property prediction, design optimization, and guiding subsequent experimental validation^{55,88–90}.

A robust ecosystem of computational tools underpins these efforts. Widely adopted MD packages such as LAMMPS⁹¹, GROMACS⁹², and NAMD⁹³ provide scalable, community-driven platforms for simulating molecular interactions in complex systems. Meanwhile, first-principles frameworks like Quantum ESPRESSO⁹⁴, ABINIT⁹⁴, and VASP⁹⁵ offer quantum-mechanical precision for evaluating electronic structures, and quantum chemistry packages like ORCA⁷² and Gaussian⁸² support high-accuracy electronic structure calculations. These software tools, along with others, are summarized in Table 2, which categorizes them by simulation type and accessibility, aiding researchers in identifying appropriate tools for their projects.

Leveraging these computational tools, researchers have tackled a range of sustainability-relevant challenges. For example, MD simulations have been employed to study such as the degradation mechanisms of battery electrolytes, the properties of perovskites as electron transport layers in solar cells, green surfactants for enhanced oil recovery, and superhydrophobic cellulose derivatives for sustainable packaging^{96–99} and also they support improvements in material processing, such as optimizing energy-intensive processes like drying in the forest products industry by studying interactions at the molecular level⁵⁶. Additionally, first-principles simulations have supported the solvation properties of lignin fragments for renewable fuel production, novel materials for renewable energy applications, hybrid

nanotubes for next-generation optoelectronics, and hydrogen storage materials^{100–104}. These sustainability-driven applications exemplify the versatility and impact of *in silico* methods.

Alongside these simulation tools, open-access materials databases, described in Table 3, compile extensive thermodynamic, structural, and electronic information gleaned from both experimental results and DFT calculations. Resources like the Materials Project²³, the OQMD²⁴, and Aflowlib¹⁰⁵ provide standardized reference datasets, enabling researchers to benchmark their simulations, optimize materials more efficiently, and train AI models for property prediction.

In parallel with direct *ab initio* simulations, ML has increasingly been integrated into first-principles and atomistic workflows as a means of accelerating data generation while retaining a physics-based foundation. In this approach, machine-learning models are trained on reference electronic-structure calculations, most commonly DFT, to learn potential energy surfaces and interatomic forces, enabling MD simulations at length and time scales inaccessible to direct quantum-mechanical methods. This paradigm has been comprehensively reviewed in the context of both materials science and quantum chemistry, highlighting its ability to extend first-principles accuracy to large-scale simulations while maintaining explicit links to the underlying physical theory^{106–108}. Representative frameworks include high-dimensional neural network potentials¹⁰⁹, Gaussian approximation potentials¹¹⁰, and more recent symmetry-equivariant graph neural networks that enforce rotational and translational invariances in atomic environments¹¹¹.

These studies consistently emphasize that the reliability of ML-accelerated simulations remains fundamentally constrained by the diversity and coverage of the training data. Predictive performance degrades under extrapolative conditions, and physically inconsistent behavior may emerge outside the sampled chemical or configurational domain^{106–108}. Consequently, ML-enhanced atomistic simulations are best understood as

Table 2 | Computational tools for in silico materials simulations

Program	Use	Open source	Description	Link
LAMMPS	Molecular dynamics	Yes	Widely used for MD simulations in material research and biology.	91
GROMACS	Molecular dynamics	Yes	Known for its speed and efficiency in MD simulations.	92
NAMD	Molecular dynamics	Yes	Designed for large biomolecular systems simulations.	93
AMBER	Molecular dynamics	No	Suitable for molecular dynamics, especially for proteins and nucleic acids.	322
VASP	First principles	No	Used for DFT calculations.	95
Quantum ESPRESSO	First principles	Yes	Open-source suite for first principles calculations using DFT.	94
ABINIT	First principles	Yes	Allows for electronic structure calculations from first principles.	323
WIEN2k	First principles	No	Electronic structure calculations using the APW+lo method.	324
Gaussian	Quantum mechanics	No	Widely used for electronic structure calculations in computational chemistry.	325
ORCA	Quantum mechanics	Yes	Quantum chemistry program for electronic structure calculations.	326
Q-Chem	Quantum mechanics	No	Focused on electronic structure calculations in quantum chemistry.	327
GAMESS	Quantum mechanics	Yes	Program for electronic structure calculations in quantum chemistry.	328
NWChem	Quantum mechanics	Yes	Program for electronic structure and MD calculations.	329
CP2K	First principles/MDs	Yes	For first principles calculations and MD simulations.	330

Overview of commonly used computational tools for in silico materials simulations, categorized by simulation approach (e.g., molecular dynamics, first-principles, and quantum-mechanical methods) and by accessibility (open-source or proprietary). The table summarizes the modeling capabilities supported by each tool for materials property analysis.

Table 3 | Open-source databases for materials science

Dataset/database	Open source	Description	References
Materials project	Yes	Provides open web-based access to computed information on known and predicted materials.	23
Open quantum materials database (OQMD)	Yes	A database of DFT-calculated thermodynamic and structural properties of materials.	24
Aflowlib	Yes	Repository of calculated and experimental materials data.	105
NOMAD	Yes	Offers a comprehensive repository of materials data, including computational and experimental datasets, to support materials science research.	25

Overview of open-source databases providing computational and experimental data on thermodynamic, structural, and electronic properties relevant to materials science research. The table highlights their role as reference resources for simulation validation, materials optimization, and the training of predictive AI models.

enabling infrastructure for simulation-driven materials discovery rather than autonomous discovery algorithms, requiring ongoing validation, uncertainty-aware sampling strategies, and periodic retraining against high-fidelity quantum-mechanical reference calculations.

ML-generated synthetic data has emerged as a vital strategy to address challenges such as small or biased datasets in materials science. Unlike in silico simulations, which model real-world physical processes based on theoretical principles, generated data mimics statistical patterns in experimental datasets. ML techniques, such as VAEs, GANs, and data augmentation, play a central role in generating synthetic datasets. These approaches create novel molecular representations and expand the diversity of existing datasets, enhancing predictive capabilities^{112–115}.

Beyond accelerating discovery, in silico simulations and generative AI can reduce the environmental and logistical burdens associated with experimental work by minimizing the need for extensive physical testing. However, this shift to computational methods does not guarantee sustainability. Large-scale, high-fidelity simulations demand substantial computational resources, incurring notable energy consumption and carbon footprints. As simulations grow in complexity, the associated energy costs can become significant.

Data scraping from publicly available sources. Data scraping is a method used to automatically extract large volumes of data from publicly available sources such as websites, allowing researchers to gather information that is not easily accessible through conventional means¹¹⁵. It involves parsing HTML content, interacting with dynamic elements, and collecting structured or unstructured data. In materials science, where

data is essential for understanding material properties, synthesis methods, and experimental outcomes, web scraping holds significant potential. By automating the extraction of relevant data from scientific journals, specialized databases, and research repositories, web scraping can facilitate the compilation of large datasets. This is crucial for data-driven research and ML applications, where the quality and volume of data directly impact the accuracy of predictive models.

The application of web scraping techniques (Table 4) in materials science could transform the way researchers access and analyze scientific data. Extracting experimental details, property measurements, and synthesis conditions from journal articles or database entries can feed into comprehensive datasets, helping to identify trends, predict material behaviors, and inform the design of novel compounds. Tools like Beautiful Soup¹¹⁶ and Scrapy¹¹⁶ can parse structured HTML data from simpler websites, while Selenium¹¹⁷ and Puppeteer¹¹⁸ are valuable for scraping dynamically loaded content from JavaScript-heavy pages. Visual tools such as Octoparse¹¹⁹ and ParseHub¹²⁰ offer a more user-friendly approach, making them accessible for quick extractions, though they might lack the flexibility required for complex tasks. For enterprise-level needs, advanced solutions like Diffbot¹²¹ and Content Grabber¹²² provide automated, AI-driven data extraction capabilities, which could be leveraged for large-scale data collection in material science applications.

Automated data extraction from scientific literature. Advances in automated data extraction and ML are reshaping materials science, speeding up discovery processes that were previously thought impossible. AI, particularly large language models (LLMs) like BERT, GPT-3, and

Table 4 | Web scraping tools and frameworks for materials science

Tool/framework	Open source	Description	Integration with data analysis tools	Suitability for materials science	Reference
Beautiful soup	Yes	A library for parsing HTML and XML documents, creating parse trees from page source.	Integrates well with Python (pandas, NumPy)	Good for small-scale HTML parsing tasks	116
Scrapy	Yes	An open-source and collaborative web crawling framework for Python.	Seamless integration with Python data tools	Ideal for extensive, multi-page data extraction	331
Selenium	Yes	A suite of tools for automating web browsers.	Compatible with Python	Suitable for dynamic JavaScript-heavy content	117
Puppeteer	Yes	A Node library that provides a high-level API to control Chrome or Chromium.	Node.js and Python integration is possible	Effective for interactive websites, slower for large tasks	118
Octoparse	No	A powerful, easy-to-use web scraping tool.	Limited integration; CSV export	Accessible for simple tasks, limited customization for complex data	119
ParseHub	No	A visual data extraction tool to extract data from any website.	Export to CSV, JSON formats	Suitable for simpler, structured data extraction	120
WebHarvy	No	A point-and-click web scraping software.	No direct Python integration	Best for structured tables, limited for complex data	332
Portia (by Scrapinghub)	Yes	A visual scraping tool that allows you to scrape websites without any programming.	Integrates with Scrapy pipelines	Good for structured, predictable sites	121
Diffbot	No	An AI web scraping tool that automatically extracts data from websites.	API integration with analysis tools	Effective for large-scale, structured data extraction, costly	121
Content grabber	No	A powerful web scraping tool for enterprises.	Database integration	Powerful for complex data, requires licensing	122
Helium	Yes	A Python library for web scraping and browser automation.	Python-compatible	Suitable for straightforward tasks, lacks advanced capabilities	333
MechanicalSoup	Yes	A Python library for automating interaction with websites.	Python integration	Best for simple, static websites, not scalable	334

Overview of representative web scraping tools and frameworks, indicating their open-source status, core functionality, integration with data analysis workflows, and suitability for use in materials science research.

GPT-4, researchers have built powerful tools to compile extensive databases, significantly reducing the time, cost, and manual effort traditionally involved in data collection. These AI-driven methods have opened transformative opportunities for innovation, enabling a shift towards data-centric approaches in materials science¹²³.

A key development in this field is the use of transformer-based LLMs to extract valuable insights from scientific literature. Tools like BatteryDataExtractor and ChatExtract^{124,125} utilize these advanced models to process unstructured text, extracting meaningful material-property relationships with impressive accuracy. The NEMAD project¹²⁶ exemplifies this synergy, combining LLM-based data parsing with predictive ML models to demonstrate the effectiveness of hybrid approaches. Additionally, recent work on large-scale polymer data extraction has shown that domain-specific models, such as MaterialsBERT¹²⁷, outperform generic LLMs in recognizing materials science terminology. MaterialsBERT, trained on 2.4 million materials science abstracts, has enabled the extraction of over 300,000 material property records from approximately 130,000 polymer-related abstracts within 60 hours¹²⁷. This large-scale automated extraction significantly accelerates knowledge retrieval compared to manual curation, providing structured material property datasets crucial for developing predictive models.

Recent advancements have further expanded the scale of AI-driven data extraction in materials science. A newly developed LLM-based framework successfully extracted over one million polymer-property records from 681,000 full-text journal articles, encompassing 24 material properties and over 106,000 unique polymers¹²⁸. This study employed both commercially available (GPT-3.5) and open-source (LLaMA 2) models in conjunction with the named entity recognition (NER)-based MaterialsBERT model. The extracted dataset, made publicly available through the Polymer Scholar website, represents the most extensive polymer-property dataset compiled to date. The study also provided key insights into the computational efficiency, accuracy, and cost-effectiveness of different AI models, demonstrating that GPT-3.5 extracted significantly more records than previous methods while maintaining high precision¹²⁸.

Another notable example is the use of AI-driven methods to extract and analyze data on polymer solar cells. A recent study employed a natural language processing (NLP) pipeline to extract donor-acceptor combinations and their corresponding power conversion efficiencies from over 3300 research papers¹²⁹. This dataset, nearly five times larger than previous manually curated collections, enabled the training of ML models to predict the efficiency of new donor-acceptor pairs. By leveraging active learning strategies, these models demonstrated a potential 75% reduction in material discovery time, equivalent to accelerating innovation by 15 years¹²⁹. This showcases the transformative potential of AI in accelerating materials discovery, particularly in emerging fields like organic photovoltaics.

Similarly, a dataset comprising 3943 molecular mutations for metal-organic framework (MOF) linkers¹³⁰. In this effort, GPT-assisted tools proposed and edited chemical structures, while human reviewers validated outputs to ensure reliability. A fully automated pipeline was also developed to extract synthesis procedures from scientific literature, transforming textual descriptions into flow graphs through deep learning and rule-based systems¹³¹. Extending automation further, GPT-4V's vision capabilities have been employed to analyze images from 346 articles, automating the classification and labeling of scientific images with minimal human intervention¹³². Meanwhile, the MaterialsBERT pipeline has demonstrated its ability to recover non-trivial insights across various material applications, such as fuel cells, supercapacitors, and polymer solar cells, through large-scale literature mining^{127,133}. These efforts illustrate how AI-driven approaches are transforming dataset creation, from partially supervised methods to fully autonomous systems.

Diverse techniques have been implemented for automating data extraction. Traditional NLP methods, such as entity recognition used in Snowball 2.0¹³⁴, are being combined with advanced strategies like prompt engineering in projects such as *ChatExtract* and *NEMAD*. Another innovative approach involves uncertainty quantification, such as Shannon

entropy, as demonstrated in the Uncertainty-Informed Screening study, which assesses the reliability of AI predictions¹³⁵. Despite these successes, challenges persist in scaling up models to handle larger datasets and extracting information from complex formats like tables and figures. Moreover, many tools depend on specialized datasets, which can restrict their applicability across diverse subfields within materials science.

The wider impact of these AI-powered tools is significant. By automating the extraction of material properties and building accurate, comprehensive databases, materials science is shifting toward a data-centric paradigm that minimizes manual curation. However, realizing the full potential of AI in this field requires addressing significant challenges, including scalability, dataset diversification, and the ability to process multimodal information sources.

To overcome these hurdles, future efforts should focus on enhancing the generalizability and scalability of AI models. Developing multimodal extraction capabilities to accommodate diverse data formats and designing specialized LLMs tailored to subfields within materials science could help bridge existing gaps. Embracing continual learning and incorporating robust uncertainty quantification across a broader range of models will further enhance the practical utility of these tools for large-scale applications.

Data preprocessing, storage and organization

Data preprocessing

Data preprocessing is a crucial first step in building AI-powered infrastructure, especially in the context of materials discovery, where vast and diverse datasets need to be integrated, cleaned, and standardized. The quality of data directly influences the performance of AI models, making preprocessing essential for extracting meaningful insights. In materials science, data often comes in heterogeneous formats, creating unique challenges that require robust preprocessing techniques.

While tools such as Excel¹³⁶ are frequently used for initial, straightforward data cleaning due to their intuitive interface and widespread familiarity among researchers, they are limited in scalability and automation capabilities. These limitations make them unsuitable for the large and complex datasets typically encountered in materials science projects, where robust and scalable preprocessing is essential.

To address these needs, more advanced tools are widely utilized, including Python's Pandas library^{137,138} and OpenRefine, or Dplyr for R^{139–141}. Pandas is a versatile and powerful library designed for data manipulation, offering extensive functions for handling missing values, filtering datasets, and merging complex data structures. Its strength lies in its integration with Python's broader ecosystem, making it particularly effective for preprocessing materials science data, which often needs extensive transformation. On the other hand, OpenRefine provides a more user-friendly, interactive approach for handling messy data, particularly useful for initial data cleaning tasks like identifying inconsistencies and standardizing entries. However, its lack of scalability limits its use for large datasets typical in materials science.

Normalization and data transformation are critical for standardizing diverse datasets, ensuring that the input features are comparable and suitable for ML algorithms. NumPy is the go-to tool for numerical operations in Python, offering efficient methods for array manipulation and normalization. Its simplicity and speed are advantageous when dealing with the large, numeric datasets often encountered in materials science, such as high-throughput experimental data. In comparison, tools like Apache Spark¹⁴² and Talend¹⁴³ offer more scalable solutions for large-scale data processing, enabling distributed computing that can handle vast amounts of data typical in AI-powered infrastructure projects. Spark's integration with Python and its in-memory processing capabilities make it well-suited for big data applications, though it requires a more complex setup and greater expertise compared to simpler, standalone libraries like NumPy.

A unique aspect of materials discovery is the need to preprocess chemical data, such as molecular structures and properties. This is where RDKit¹⁴⁴ plays a significant role. RDKit is an open-source toolkit designed

for cheminformatics, providing functions to process and analyze chemical structures. It excels in tasks like molecule standardization, fingerprint generation, and chemical feature extraction, making it invaluable for working with data on compounds and materials. When integrated into a Python-based preprocessing pipeline alongside Pandas and NumPy, RDKit allows seamless extraction and transformation of molecular descriptors, which are essential for building predictive models in materials science. Its open-source nature and extensive functionality make it a preferred tool for researchers looking to include chemical information in their AI models.

For data transformation workflows, tools like KNIME¹⁴⁵ and Talend offer a more visual approach, allowing researchers to build complex preprocessing pipelines without extensive coding. KNIME's integration capabilities make it particularly useful for handling diverse data types, including numerical, text, and chemical data processed by RDKit. However, the reliance on graphical user interfaces can limit flexibility, especially when automation is needed for large-scale projects. Alternatively, R provides strong statistical and data manipulation functions, which can be particularly advantageous for materials researchers who are comfortable with its syntax.

While commercial tools like Alteryx¹⁴⁶ offer advanced features for data transformation and blending, their closed-source nature and licensing fees can be a barrier in academic and open-source projects. By integrating tools like RDKit for cheminformatics with scalable frameworks such as Apache Spark and user-friendly libraries like Pandas and NumPy, researchers can create robust preprocessing pipelines capable of handling diverse data formats. This integrated approach allows for the inclusion of complex molecular descriptors in AI models, improving predictive accuracy and accelerating materials discovery. The ability to preprocess large, heterogeneous datasets effectively not only streamlines the research process but also supports sustainability goals by reducing inefficiencies in data preparation, aligning with the broader vision of AI-powered materials discovery.

Data storage in the cloud and edge computing

Cloud and edge computing offer unique advantages for AI-powered, open-source infrastructure in materials discovery. Cloud computing stands out for its scalability and processing power, making it well-suited for handling large datasets and running complex AI algorithms¹⁴⁷. Its centralized architecture provides extensive computational resources but often faces issues with higher latency, which can be problematic for real-time analysis. On the other hand, edge computing processes data closer to its source, reducing latency and enabling faster, real-time analysis¹⁴⁸. This makes it particularly useful for applications needing immediate feedback, such as in experimental setups. However, edge devices generally have less computational capacity and storage compared to cloud platforms, limiting their ability to manage extensive datasets or highly complex models.

In the context of AI-driven, open-source infrastructure for materials discovery, data storage and computational resources are critical, especially when utilizing both cloud and edge computing. Major cloud service providers like Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP) have played a key role in enabling high-performance computing (HPC) workflows. Their infrastructure supports scaling in computational materials science, providing the flexibility needed to handle the large datasets generated during material screening and optimization processes. However, while cloud computing offers significant benefits in scalability and accessibility, it also brings challenges related to energy consumption, carbon footprint, and long-term sustainability.

To illustrate the practical applications and effectiveness of cloud platforms, recent studies have demonstrated the scalability of cloud platforms in materials discovery. For example, one study utilized up to 1000 virtual machines (VMs) on Microsoft Azure to screen over 32 million material candidates, achieving computational throughput that would be difficult to replicate with traditional on-premises HPC resources¹⁴⁹. Similarly, a decentralized, cloud-based approach enabled globally dispersed laboratories to collaborate asynchronously, highlighting the potential of distributed research infrastructure¹⁵⁰. These examples showcase the capability of cloud

providers to democratize access to advanced computational tools, lowering barriers for researchers and enabling a more inclusive approach to computational materials discovery.

Despite these advantages, environmental concerns persist. Cloud providers are making efforts to reduce the environmental impact of large-scale computing by utilizing renewable energy sources and developing carbon-neutral data centers^{151–153}. However, the carbon footprint of extensive cloud computing remains a significant concern, especially in fields like materials science, where computational tasks are both resource-intensive and lengthy. In this regard, edge computing offers a promising alternative or complement to traditional cloud storage.

Companies such as Cisco Systems, Intel Corporation, and NVIDIA are leading the development of edge computing solutions tailored for scientific and industrial applications. Edge computing minimizes the need for frequent data transfers between local devices and central cloud servers, helping to alleviate network latency and bandwidth issues. By processing data closer to its source, edge computing can reduce the energy costs associated with data movement and potentially decrease the overall carbon footprint. However, edge computing also has its limitations, particularly in terms of reduced computational power and storage capacity compared to centralized cloud platforms¹⁵⁴.

Given these considerations, balancing cloud and edge computing for AI-powered materials discovery requires careful consideration of these trade-offs. Cloud platforms excel in tasks requiring extensive computational resources and seamless tool integration. In contrast, edge computing is more suitable for real-time data processing in localized experiments, where reducing latency and avoiding data transfer delays are crucial. A hybrid approach that combines cloud and edge solutions could offer a practical way forward, optimizing task allocation based on computational needs and energy efficiency¹⁵⁵.

Furthermore, industry initiatives are addressing these sustainability concerns. New partnerships and initiatives are emerging to tackle the carbon footprint associated with data storage and cloud computing. Microsoft's collaboration with OpenAI includes commitments to building carbon-negative data centers, while Google aims to power its cloud infrastructure entirely with carbon-free energy by 2030¹⁵³, signaling a broader industry shift towards sustainability. However, the actual impact of these initiatives on the environmental footprint of computationally intensive materials discovery workflows remains to be thoroughly evaluated.

Lastly, data transfer bottlenecks remain a critical issue, given the large volumes of data typical in computational materials science. Companies like IBM and Oracle are integrating advanced storage solutions and content delivery networks (CDNs) to help address these challenges. However, widespread adoption and detailed analysis of these tools are still limited in the current literature, highlighting an area that requires further exploration and refinement. Addressing the environmental impacts and data transfer challenges through sustainable practices and innovative technologies will be crucial for the long-term success and sustainability of these infrastructures. By leveraging the strengths of both cloud and edge solutions, the materials science community can enhance collaboration, improve efficiency, and drive significant breakthroughs in materials discovery.

Data organization and indexing

Organizing data in materials science repositories presents several key challenges. The diversity of data types from crystal structures to electronic properties makes developing a unified schema challenging. Scalability is also a major issue, as current systems struggle to handle the surge in data without sacrificing performance. Furthermore, a lack of interoperability due to the absence of universal standards complicates the integration of datasets from various sources, hindering the potential for discovery. Many repositories also fail to adequately record data provenance, which hampers the verification and reproducibility of results.

To address these issues, new approaches in data organization are being explored. For example, the Materials Project has implemented flexible data models that can easily incorporate new data types without extensive

restructuring²³. Initiatives like the European Materials Modelling Ontology (EMMO)¹⁵⁶ aim to establish a common framework for describing materials and processes, though broad adoption remains limited. AiiDA¹⁵⁷ leverages a graph-based data structure to capture complex relationships between computations and data, enhancing reproducibility; however, this method can be resource-intensive for large datasets.

In this context, the FAIR principles⁴⁴ have become a widely recognized standard for organizing scientific data. These guidelines advocate for creating data repositories that are easily searchable, accessible via standardized protocols, interoperable with other systems, and reusable for future research. Although the adoption of FAIR principles is increasing in materials science^{158,159}, challenges remain, particularly with achieving interoperability between platforms and standardizing metadata.

Effective indexing is vital for the performance and usability of materials repositories, but poses its own challenges. Maintaining multidimensional indexes to support efficient searches across various properties can be computationally demanding. There is often a trade-off between speed and accuracy, as approximate indexing methods offer faster searches but may compromise precision. Additionally, keeping indexes current in dynamic repositories presents ongoing technical challenges.

Several innovative indexing strategies are under development. Hierarchical indexing, used in systems like AFLOW¹⁰⁵, supports efficient searches across multiple levels of detail but can be complex to implement in smaller systems. Semantic indexing, adopted by repositories like OMDb¹⁶⁰, uses NLP to categorize materials based on textual descriptions, though these techniques still require further refinement. Adaptive indexing, which modifies its structure based on query patterns, shows potential but has not yet been scaled in materials databases.

The impact of efficient data organization and indexing extends directly to sustainability efforts. Better data management reduces the computational load of searches and analyses, thereby lowering energy consumption. Identifying similar materials more efficiently can eliminate redundant calculations, saving valuable computational resources. Additionally, faster and more precise searches can accelerate the discovery of environmentally friendly materials, indirectly contributing to a lower carbon footprint across industries.

Despite these advancements, significant work remains in optimizing data organization and indexing for materials discovery. The lack of universal standards and the computational challenges of advanced solutions are persistent barriers. The research community must address these issues with a focus on reducing the environmental impact of the discovery process. The future of AI-powered open-source infrastructures will depend on developing systems that prioritize both performance and energy efficiency. Only then can we create a comprehensive framework that effectively tackles sustainability and carbon footprint challenges in materials discovery.

Data and AI pipeline

Data processing

Effective data processing in materials discovery necessitates the integration of diverse methodologies. The predictive modeling stage involves converting raw experimental or computational data into meaningful features and building models to estimate material properties. Essential tools in this process include Pymatgen¹⁶¹ and Matminer¹⁶².

Pymatgen is a comprehensive Python library for materials analysis, offering robust tools for parsing structural data, performing symmetry analysis, and calculating physical properties such as band gaps and formation energies. Its integration with databases like the Materials Project facilitates the rapid acquisition of high-quality materials data, which can then be preprocessed into numerical descriptors. Building on Pymatgen's capabilities, Matminer provides a rich set of feature extraction functions tailored for materials science, automating the generation of descriptors based on composition, structure, and electronic properties for use in statistical analysis and ML models.

For statistical processing, Scikit-learn¹⁶³ is preferred due to its extensive array of algorithms for regression, classification, and clustering. This stage

typically begins with exploratory data analysis to identify patterns and correlations between material descriptors and target properties. Techniques such as principal component analysis (PCA) and random forest regression in Scikit-learn help determine feature importance and reduce dimensionality, enhancing both interpretability and model performance. For example, features extracted by Matminer can be used with a Scikit-learn random forest regressor to predict mechanical properties like hardness or elasticity, serving as a baseline model for further refinement through deep learning techniques.

Handling large datasets is a common requirement in materials discovery, given the vast number of potential chemical compositions and structural configurations. The Atomic Simulation Environment (ASE)¹⁶⁴ provides essential tools for managing big data from high-throughput simulations. ASE facilitates the automation of atomistic simulations, supports various computational chemistry codes (including Quantum ESPRESSO), and is designed for scalability in high-performance computing environments¹⁶⁵.

Real-time data analysis is increasingly important in materials discovery, enabling adaptive learning systems and dynamic optimization of experimental conditions. Deep learning frameworks such as TensorFlow¹⁶⁶ and PyTorch¹⁶⁷ are pivotal for implementing models that process real-time data streams. TensorFlow supports the development of neural networks capable of handling large, complex datasets and deploying models in real-time environments, which is useful for integrating ML models into experimental setups. PyTorch's dynamic computation graph offers flexibility for building experimental neural architectures and supports reinforcement learning for real-time optimization tasks.

In cheminformatics, RDKit complements the real-time processing capabilities of TensorFlow and PyTorch by providing tools for handling chemical information and molecular structure analysis. RDKit enables efficient preprocessing of chemical data, including molecular fingerprinting and the calculation of descriptors related to molecular shape, electronic properties, and reactivity, as previously discussed. These features can be combined with deep learning frameworks to develop predictive models that process molecular data in real time.

Building on the foundation of robust data processing and real-time analysis, AI modeling has emerged as a transformative approach in materials discovery. By leveraging advanced ML and deep learning techniques, AI enables the development of predictive models capable of uncovering complex patterns and relationships in high-dimensional datasets. These models not only enhance the efficiency of property predictions and material screening but also facilitate the exploration of novel materials through generative frameworks and reinforcement learning. The integration of AI into materials research paves the way for accelerated innovation, bridging the gap between experimental data, computational simulations, and practical applications.

AI modeling

Traditional ML models in materials science. ML has become a transformative tool in materials science and engineering (MSE), enabling innovative solutions in material discovery, process optimization, and sustainability. Figure 5 provides a comprehensive overview of a typical ML workflow in MSE, illustrating the diverse data inputs, ML methodologies, and their practical applications.

Traditional ML algorithms, such as support vector machines (SVMs), artificial neural networks (ANNs), and Decision Trees, have demonstrated their effectiveness in optimizing material synthesis processes, including nanofiltration and biomaterial development^{12,168}. Predictive modeling techniques, such as Bayesian optimization and random forests (RF), have minimized experimental requirements while advancing research on cellulose composites, ionic liquids, and stable metallic glasses. These advancements have enabled breakthroughs in smart packaging and biocompatible hydrogels^{17,169,170}.

Deep learning models for material property prediction. Deep learning models extend the capabilities of traditional ML by extracting hierarchical features from large, complex datasets. Advanced neural networks, such as polymer-unit graph neural networks (PU-MPNNs), have revolutionized material property predictions for organic semiconductors, reducing training times by 98% while maintaining over 80% accuracy¹⁷¹. Similarly, Universal Interatomic Potentials (UIP) models, including MACE and CHGNet, have achieved remarkable accuracy with RMSEs as low as $\sim 0.3\text{--}0.7\text{ J/m}^2$, driving innovation in catalysis and nanotechnology applications¹⁷².

In materials characterization, convolutional architectures like CNNs and DenseNets have automated microstructure classification with accuracies exceeding 94%. Tools like DeepXRD have further enhanced structural analysis by predicting X-ray diffraction (XRD) spectra with unprecedented precision^{173–175}. Graph-based and multimodal learning frameworks have also advanced the field by integrating diverse datasets to enhance predictive accuracy. Models like SA-GNN and GeoCGNN have refined property predictions, achieving MAE values as low as 0.044 eV/atom for formation energy. Meanwhile, the MultiMat framework has reduced MAE by 10% for elastic tensors through multimodal data integration^{176–178}.

For imaging tasks, convolutional approaches like multi-scale refocusing networks (MRNs) and Mask R-CNN have optimized material image analysis, automating processes such as defect detection, nanoparticle recognition, and structural classification with accuracies exceeding 90%^{18,175,179–181}. CNN-based methods have also facilitated damage detection in sandwich composites with accuracies over 94%, while AI models like Cellpose have promoted innovations in biofouling-resistant materials, reducing bacterial adhesion by 98%^{18,182}. These advancements underscore the pivotal role of graph and image-based learning in sustainable materials science.

Deep neural networks (DNNs) have outperformed traditional methods like Random Forest Regression in predicting lattice constants and band gaps for inorganic perovskites, achieving RMSE values below 0.1 Å and 0.7 eV, respectively. These improvements have significantly expedited research in optoelectronics¹⁴. Hybrid deep learning–ensemble learning (DL-EL) models, optimized with Genetic Algorithms, have achieved R^2 values of 0.9863 for Cu–Ni–Si alloys, enhancing predictions of hardness and tensile strength¹⁶. These examples highlight how ML-driven innovations are fostering efficiency across material discovery pipelines.

One particular challenge in deep learning-based material property prediction is extrapolation—the ability to make accurate predictions for materials beyond the model's training distribution. Traditional models tend to be interpolative, limiting their effectiveness to chemical spaces near the data they have seen. Recent analyses suggest that many Out-of-Distribution (OOD) evaluations in materials science may actually reflect interpolation rather than true extrapolation, leading to overestimated generalization capabilities¹⁸³. For instance, many ML models perform well on seemingly OOD tasks simply because the test data remain within the representational domain of the training set rather than exploring truly novel chemical spaces.

However, recent advances in meta-learning and attention-based neural architectures have shown strong potential to tackle this issue. The extrapolative episodic training (E2T) framework¹⁸⁴, for example, employs matching neural networks (MNNs) to develop extrapolative capability through training on artificially generated extrapolative tasks. This approach has proven effective in polymeric and hybrid organic–inorganic perovskite systems, demonstrating superior generalization to materials outside the initial training set. Complementing such frameworks with more rigorous OOD benchmarks¹⁸⁴, would further enhance model robustness by ensuring evaluations account for materials that truly differ from the training data in both chemical composition and structure.

While predictive models enhance our understanding of known materials, generative AI opens a new frontier, allowing algorithms to create entirely novel compounds tailored to performance, sustainability, and application-specific needs.

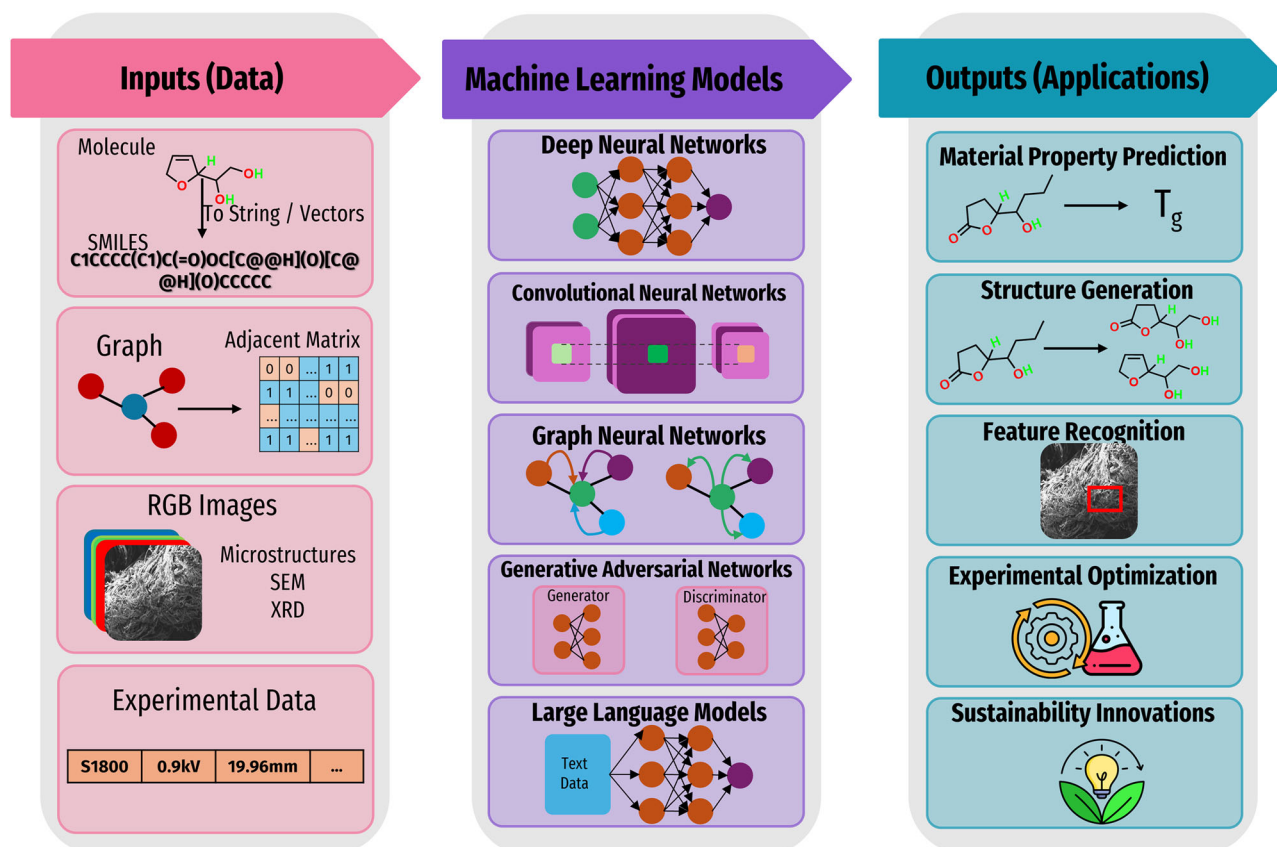


Fig. 5 | Machine learning framework for materials science. Schematic overview of a machine learning workflow in materials science, illustrating representative inputs such as molecular structures, graph-based representations, image data, and experimental datasets, and their processing through models including deep neural networks, convolutional neural networks, graph neural networks, generative

adversarial networks, and large language models. The framework highlights typical application domains enabled by these models, including materials property prediction, structure generation, microstructure analysis, experimental optimization, and sustainability-oriented design tasks.

Federated learning for collaborative materials informatics. Federated learning (FL) is a decentralized training paradigm in which model updates are shared while raw data remain local, enabling collaborative model development under confidentiality, intellectual-property, or regulatory constraints¹⁸⁵. This is particularly relevant to materials science because experimental and computational datasets are routinely fragmented across institutions (“data islands”), and transferring high-volume data can be impractical or undesirable. In materials informatics, parameter-exchange training across multi-source databases has been shown to produce structure–property models (e.g., formation energy prediction) with accuracy close to that obtained from fully centralized datasets, while reducing bandwidth demands and preserving local control over data assets¹⁸⁶. Related work in the physical sciences demonstrates that horizontal FL can mitigate data scarcity and selection bias across heterogeneous contributors, and that federated dimensionality reduction can improve convergence and reduce inter-client statistical heterogeneity in materials-relevant datasets such as metallic nanoparticles¹⁸⁷. Beyond methodological feasibility, secure and scalable implementations have also been proposed for large collaborative materials ecosystems: MatSwarm integrates FL with blockchain-based coordination and trusted execution environments to address non-i.i.d. materials data, governance, and adversarial risks while supporting large-scale multi-institution participation¹⁸⁸.

In smart manufacturing, FL has been actively explored as a privacy-preserving route to improved process monitoring and quality control across distributed production sites, where pooling raw process data across factories is often infeasible. For metal additive manufacturing, an FL-based case study

for melt-pool anomaly (defect) detection shows performance comparable to centralized learning while maintaining confidentiality, and highlights robustness across varying melt-pool conditions that are expected in real deployments¹⁸⁹. More broadly, FL has been demonstrated for part qualification and dimensional prediction across additive-manufacturing factories, where the approach can achieve centralized-learning accuracy under statistically similar production regimes while substantially outperforming isolated local training when each site has limited data¹⁹⁰. However, these studies also emphasize that FL performance can degrade when participating sites exhibit strong distribution shifts (e.g., different defect regimes or quality levels), indicating that federation design and client heterogeneity management are central practical challenges rather than secondary implementation details. Collectively, these findings position FL as a principled mechanism for privacy-preserving collaboration in materials research and manufacturing, while underscoring that practical success depends on explicitly addressing heterogeneity, imbalance, and governance rather than assuming federated training alone resolves these constraints^{188–190}.

Explainable artificial intelligence in materials science. Explainable artificial intelligence (XAI) has become a practical methodological layer in materials science because many high-performing models (e.g., gradient-boosted trees, DNNs, and graph-based architectures) are otherwise difficult to interrogate, limiting scientific trust, debugging, and downstream decision-making. In contemporary materials workflows, explainability is increasingly used for three tightly coupled purposes: (i) validating whether learned relationships are consistent with domain expectations; (ii) identifying influential variables that can be acted upon

in process optimization or materials design; and (iii) translating model behavior into representations that enable human-guided hypothesis generation and constrained exploration of chemical and processing spaces^{191,192}. This positioning is particularly relevant for open-source, modular infrastructures intended to support accelerated discovery and advanced manufacturing, where models must be auditable, reusable across datasets, and interpretable by non-ML stakeholders.

Across materials subdomains, two complementary explainability paradigms are prominent. The first is a post-hoc, model-agnostic explanation, where feature attribution tools are applied after training to quantify how inputs drive predictions locally (per sample) and globally (across a dataset). In porous materials and adsorption problems, SHAP- and LIME-based analyses have been used to expose dominant operational and structural factors underlying predicted uptake trends, thereby connecting predictive performance to interpretable drivers rather than opaque correlations^{193,194}. Similar SHAP-guided interpretation has been integrated into alloy design pipelines to clarify feature influence and to rationally delimit a Bayesian optimization search space for multi-property targeting, linking explainability directly to actionable design decisions rather than retrospective visualization¹⁹⁵. In catalysis, SHAP has likewise been used to extract physically interpretable determinants of reactivity within a theory-infused neural framework, emphasizing explanation as a route to mechanistic insight into electronic contributions rather than a purely statistical ranking¹⁹⁶. In manufacturing-relevant contexts, explainability is also being extended to high-dimensional sensing modalities: deep-learning models applied to in situ process video streams have been coupled with multiple XAI techniques to distill actionable recommendations from process dynamics, while explicitly acknowledging risks such as over-interpretation and confounding in human-in-the-loop inference¹⁹⁷. For signal-based classification tasks, activation mapping approaches such as Score-CAM have enabled identification of the most informative sensor regions and time segments, providing a concrete bridge between learned decision boundaries and physically meaningful sensing responses¹⁹⁸.

The second paradigm focuses on intrinsically interpretable or theory-/physics-infused modeling, where interpretability is embedded in the representation or architecture, reducing reliance on purely post-hoc explanations. Composition-based attention models provide a clear example: interpretability mechanisms incorporated into CrabNet enable inspection of learned element representations and attention patterns, facilitating chemical-space visualization, diagnosis of dataset imbalance effects, and identification of potential data/model issues through attention dynamics¹⁹⁹. Related efforts explicitly target deep network transparency in composition-only predictors, where XAI toolkits are applied to analyze whether inferred element importances align with expected chemical attributes and whether predicted stability landscapes are consistent across binary composition spaces, thereby clarifying both strengths and failure modes of black-box predictors²⁰⁰. In alloy thermodynamics and phase-boundary prediction, graph-based models have been paired with explainability layers to identify influential elemental interactions and to derive empirical equations that improve usability and interpretability for design-oriented workflows²⁰¹. In parallel, theory-infused deep learning has been used to decompose predicted energetic contributions into physically meaningful components, emphasizing interpretability as a structural property of the model rather than an external add-on²⁰². Hybrid frameworks that augment feature-based models with learned graph-derived descriptors and then “decode” latent representations into explicit formulas via surrogate modeling and symbolic regression further illustrate a route to interpretability that yields analyzable, human-readable descriptors without fully sacrificing predictive performance²⁰³. Conceptually, these approaches align with broader calls in materials AI to combine data-driven models with physical knowledge and to use explainability to improve transparency and adoption¹⁹².

A recurring constraint across materials domains is data sparsity and heterogeneity, which shapes how explainability is deployed and what it can credibly support. Small-data perspectives explicitly frame interpretability as

essential for ensuring that limited datasets yield models that remain scientifically informative and not merely predictive, while also motivating constrained learning strategies (e.g., active learning, transfer learning) and physically grounded descriptors¹⁹¹. Physics-informed augmentation and extrapolative modeling strategies provide a concrete example of how interpretability and physical constraints can be coupled under data scarcity: process-variable models for enzymatic fiber refining achieved strong predictive performance while using physics-informed data augmentation and physics-informed regression to support extrapolation and targeted experimental verification, emphasizing a workflow where model-driven guidance remains connected to measurable process variables and mechanistic interpretation²⁰⁴. In parallel, literature-scale synthesis efforts that combine meta-analysis with explainable ML have used XAI to identify the most consequential formulation and processing features in lignin-nanoparticle Pickering emulsions, illustrating how explainability can convert heterogeneous reports into design heuristics while highlighting the importance of curated, structured datasets for robust modeling²⁰⁵. Collectively, these studies reinforce that XAI outputs should be treated as explanations of model behavior under available data, useful for validation, prioritization, and constraint setting rather than direct evidence of causality, particularly when datasets are noisy, biased, or incomplete^{197,206}.

Within an AI-powered, open-source infrastructure for accelerating materials discovery and advanced manufacturing, these patterns motivate an implementation stance where explainability is not an optional visualization step but a standard, versioned artifact of the modeling pipeline. Practically, this means supporting both model-agnostic attribution (e.g., SHAP/LIME for tabular and structured descriptors) and modality-appropriate explainability for deep models (e.g., attention inspection for composition/graph architectures, activation mapping for signals and imagery), while enabling theory-/physics-infused options when domain constraints are known and must be enforced^{194,198,199,201}. Under small-data regimes and heterogeneous experimental records, XAI should be coupled to dataset diagnostics, uncertainty-aware evaluation, and human review to reduce overinterpretation and to maintain traceability from data provenance to predictions and derived design recommendations^{191,197,204,206}. This integration directly supports reproducible decision-making in discovery and manufacturing settings by making the rationale for screening, optimization, and experimental prioritization explicit and inspectable across users, datasets, and model revisions.

Generative AI in materials science. Generative models have become a central component of AI-enabled inverse design, enabling the creation of candidate compositions and structures by learning high-dimensional distributions rather than only predicting properties of known materials. This paradigm complements high-throughput screening by proposing new data points that can expand accessible phase spaces and reduce experimentalist bias in exploration workflows^{207–209}. Across materials classes, the practical impact of generative methods is strongly conditioned by (i) the representation used to encode structure and chemistry, and (ii) the degree to which generated candidates can be validated for stability and synthetic feasibility^{209,210}. Consequently, recent work increasingly frames generative modeling as an integrated design loop, generation, filtering by physics- or ML-based surrogates, and experimental or high-fidelity computational confirmation, rather than as a standalone sampling procedure^{20,207,208}.

Classical generative models. Particularly, VAEs and GANs, have been widely adopted as foundational tools for inverse materials discovery because they enable controlled sampling of latent spaces to propose new compositions, molecules, and microstructures^{209,210}. In materials informatics, these approaches have been applied across multiple design scenarios, including composition design, crystal structure search, microstructure characterization, and defect-related image generation, underscoring their versatility beyond molecule-only settings^{209,211}. Their effectiveness, however, remains tightly coupled to how materials are encoded: low-complexity

representations (e.g., phase-space and composition-only encodings) improve interpretability and breadth, whereas coordinate- or image-based encodings better support downstream coupling to physics-based validation but increase modeling complexity and overfitting risk²⁰⁹.

Despite their demonstrated utility, GAN-based pipelines face well-documented technical constraints that are particularly consequential in materials settings, where data are often limited, and distributions can be multi-modal. Reviews emphasize that robust GAN training often requires large datasets, while common augmentation strategies may increase sample count without adding fundamentally new information, motivating the incorporation of uncertainty-aware or data-efficient strategies when operating in small-data regimes^{211–213}. In addition, generated samples may remain overly similar to the training set, limiting novelty and motivating hybridization strategies (e.g., combining GANs with continuous-latent models or embedding domain priors through tailored objectives) to better balance fidelity with diversity²¹¹. These limitations are also observed in crystal generation tasks, where vanilla GANs can exhibit training instability and mode collapse, and even stabilized variants (e.g., Wasserstein-based objectives) may still struggle to consistently reproduce physically realistic symmetry in generated structures, indicating that model choice and representation jointly govern scientific usefulness²¹⁴. More broadly, survey-level analyses stress that synthetic viability and interpretability remain persistent barriers for mainstream adoption of generative models in materials research, reinforcing the need for validation-aware workflows and transparent error analysis when deploying VAEs and GANs for discovery rather than interpolation²¹⁰.

Diffusion models for materials generation and design. Diffusion models have emerged as a strong alternative to adversarial generative approaches for materials discovery because they generate candidates through iterative denoising, which can improve training stability and yield higher-fidelity structures in representation-sensitive settings²¹⁴. In comparative crystal-generation studies, diffusion models produced more realistic and symmetric structures than vanilla GANs and WGAN variants, indicating an empirical advantage in capturing crystallographic regularities²¹⁴. Recent work further emphasizes scalability: diffusion models coupled to unified crystal representations have been trained on datasets containing millions of materials and evaluated using downstream-relevant metrics such as per-composition formation energy and convex-hull stability from DFT, explicitly addressing the known mismatch between generic generative metrics and discovery objectives²¹⁵. This shift reinforces the broader need for chemically grounded evaluation in generative materials workflows, including hybrid strategies that combine lightweight distributional metrics with physics-based screening for high-confidence candidates²¹⁶.

A second capability that makes diffusion models particularly relevant for inverse design is conditional generation, which allows structures to be generated under compositional, structural, processing, or property constraints while preserving sample diversity^{215,217}. For crystalline materials, point-cloud diffusion approaches demonstrate this principle by generating structures conditioned on elemental composition and validating candidates using DFT-based screening (including stability-related analyses such as hull energy and phonon checks), creating a direct bridge between generation and physics-based verification²¹⁸. For complex non-crystalline regimes, diffusion modeling has also been extended to amorphous materials, where conditional generation enabled sampling across compositions and processing conditions while reproducing short- and medium-range order and multiple macroscopic properties at substantially reduced computational cost relative to conventional MDs, with occasional post-generation relaxation required for rare outlier environments²¹⁷.

From an infrastructure and autonomy perspective, diffusion models map naturally onto the design principles of SDLs and closed-loop materials optimization because they can generate diverse candidates that are subsequently filtered, ranked, and iteratively refined using downstream stability and property models, reducing reliance on exhaustive high-cost simulations and experimental trials²¹⁹.

From LLMs to agentic AI in materials discovery. While LLMs have demonstrated significant value as knowledge assistants for materials science, supporting literature mining, hypothesis generation, representation learning, and property prediction, the key conceptual shift lies in transforming LLMs into agentic systems capable of planning, tool use, memory, and iterative reasoning across complete discovery workflows. Recent perspectives emphasize that the impact of AI in materials research increasingly depends not on isolated model performance, but on the integration of LLMs into autonomous or semi-autonomous pipelines that coordinate generation, evaluation, validation, and human feedback within closed-loop discovery systems^{220,221}. In this context, agentic AI represents a transition from task-level automation toward collaborative, goal-driven scientific workflows.

Several recent frameworks illustrate this evolution by combining LLM reasoning with generative and predictive models. Reinforcement-learning-enhanced language models such as MOFGPT demonstrate how LLM-based generators can be coupled to property predictors and reward functions to enable inverse design under explicit validity, novelty, and performance constraints, outperforming fine-tuning-only approaches in navigating complex chemical spaces²²². More broadly, agent-based architectures such as MAPPS, MatAgent, and MOFGen integrate LLMs with diffusion-based structure generators, physics-informed evaluators, synthesis feasibility checks, and external scientific tools, enabling iterative refinement of candidates and substantially improving stability, novelty, and targeting performance relative to standalone generative models^{221,223,224}. These systems exemplify how agentic AI can orchestrate heterogeneous components, language reasoning, generative modeling, physics-based screening, and expert feedback into coherent discovery pipelines that more closely resemble human scientific workflows.

Despite these advances, several limitations remain that constrain the deployment of agentic AI in real-world materials research. First, reliability and scientific grounding remain central challenges. LLMs trained predominantly on general text data may exhibit hallucinations, weak numerical reasoning, or inconsistent handling of structural and quantitative information, limiting trust in autonomous decision-making^{225,226}. While retrieval-augmented generation, reinforcement learning, and physics-informed tool integration can mitigate these issues, robust self-verification and error-detection mechanisms are still under active development^{221,227}. Second, explainability and transparency remain critical barriers, particularly as agentic systems grow in complexity. Although attention-based transformers and language-centric representations offer partial interpretability, the reasoning chains of multi-agent systems often remain opaque, complicating validation, debugging, and regulatory acceptance^{227,228}.

A further challenge concerns standardization and benchmarking. The diversity of representations (e.g., SMILES, SLICES, MOFid, text descriptions), evaluation metrics, and datasets across studies makes systematic comparison difficult and limits reproducibility^{229–231}. Recent benchmarking efforts highlight that general-purpose LLMs often underperform domain-specific or instruction-tuned models on materials tasks, underscoring the need for standardized benchmarks, task-specific fine-tuning protocols, and shared evaluation frameworks tailored to materials discovery objectives²³¹. In parallel, domain-specific LLMs, trained or adapted using curated materials corpora, structured representations, and multimodal data are increasingly recognized as essential for achieving reliable performance without excessive model scaling^{226,232,233}.

Looking forward, future research is likely to focus on three converging directions. First, developing domain-adapted, multimodal foundation models that integrate text, structure, images, and experimental data will be critical for robust reasoning and generalization across materials classes^{225,234}. Second, advancing agentic architectures with explicit planning, uncertainty awareness, and human-in-the-loop control will be essential to balance autonomy with scientific accountability and trust^{221,227,235}. Third, establishing standardized, open infrastructures including benchmarks, data schemas, tool interfaces, and validation protocols will be necessary to enable reproducible, interoperable, and scalable deployment of agentic AI in both

academic and industrial materials research^{228,230,231}. Together, these directions position agentic AI not as a replacement for human expertise, but as a collaborative layer that augments scientific reasoning, accelerates discovery cycles, and supports more efficient and sustainable materials innovation.

To facilitate comparison across predictive, generative, and agentic AI approaches discussed in “AI modeling”, Table 5 provides a consolidated overview of representative model classes, their typical input representations, application domains, and task-appropriate performance metrics. Because these approaches address fundamentally different objectives, performance metrics are reported as evaluation criteria rather than as directly comparable numerical benchmarks.

AI in cloud-based infrastructure for materials science. The rise of AI has driven leading cloud providers to develop AI solutions tailored to advanced technologies for business and research. Cloud-based AI services offer benefits like scalability, cost-efficiency, and democratization of advanced tools, allowing organizations to access powerful AI resources without significant upfront infrastructure investment. Platforms like TensorFlow and PyTorch thrive with the support of cloud-based compute resources, enabling continuous innovation and accessibility. In materials science, these services accelerate research by analyzing vast datasets, predicting material properties, optimizing manufacturing processes, and simulating experiments. For instance, Google’s DeepMind has leveraged AI to predict protein folding, a breakthrough with profound implications for drug design and material engineering. While generative AI accelerates innovation, it requires substantial computational power, making cloud-based AI critical for scaling these advancements.

The AI cloud services market, led by Amazon SageMaker²³⁶, Google AI Cloud Platform²³⁷, Microsoft Azure AI²³⁸, and IBM Watson, caters to diverse use cases with specialized features. Amazon SageMaker facilitates tasks such as predicting material properties and optimizing experimental parameters. Google AI Cloud Platform excels with frameworks like TensorFlow and AutoML, ideal for recognizing patterns in materials datasets. Microsoft Azure AI enhances collaborative projects through tools like Azure ML and Power BI, while IBM Watson’s expertise in NLP aids in mining insights from scientific literature and patents. Despite their strengths, these platforms require extensive domain-specific customization and multidisciplinary collaboration to fully unlock their potential in material science. Complementing commercial solutions, open-source deployment platforms democratize access, provide code transparency, and enable broader participation in AI-powered materials research.

Open-source deployment

Deploying AI-powered platforms and tools in this domain offers unprecedented opportunities for innovation. AI models can rapidly sift through massive datasets, predict material properties, and simulate experimental outcomes, reducing the time and resources needed for discovery. However, realizing these benefits hinges on robust deployment strategies that ensure accessibility, scalability, and transparency. Open-source initiatives have a critical role in democratizing access to these tools, fostering collaboration, and driving collective progress toward sustainable solutions.

AI-infrastructure platforms and deployment tools

Version control platforms like GitHub²³⁹ have become a cornerstone for deploying AI models in materials science. Their intuitive interface and robust versioning capabilities allow developers to manage codebases, collaborate effectively, and integrate tools like CI/CD pipelines for streamlined deployment. Initiatives like The Materials Project²³ provide comprehensive materials databases and tools like pymatgen for property simulations, while Intel Labs²⁴⁰ develops AI-driven frameworks like matsciml for ML applications. DeepMaterials promotes accessible informatics with tools like SLMat²⁴¹ for serverless materials design, and Materials Virtual Lab²⁴² applies graph-based AI models with repositories like matgl.

Dedicated web platforms offer a user-centric approach to deploying AI models by enabling real-time interaction through graphical interfaces or APIs, making advanced tools accessible to researchers without coding expertise. For instance, GitHub Pages and Docker enhance GitHub deployments with lightweight hosting and consistent runtime environments, while frameworks like Flask and Streamlit provide dynamic dashboards for intuitive interaction. A notable example is a Flask-based battery data analysis platform designed for lithium-ion batteries²⁴³, which integrates ML to estimate the state of charge (SOC) with errors under 2%, alongside interactive visualizations and real-time monitoring capabilities. Similarly, Streamlit-powered applications like LatticeML²⁴⁴ and Stmol²⁴⁵ enable property predictions and 3D molecular visualizations. Additionally, platforms like M2Hub²⁴⁶ and PcMSP²⁴⁷ demonstrate how web-based tools can combine ML with standardized workflows for property simulation. While GitHub excels in collaboration and transparency, web platforms offer broader accessibility but often require additional hosting resources and maintenance.

Best practices for deploying data models involve thoughtful organization and structure to ensure usability and reproducibility. Models should be containerized, using tools like Docker, Kubernetes or Virtualenv to create portable environments that eliminate compatibility issues. Clear documentation, including usage instructions, dependencies, and version history, is essential. For GitHub repositories, modularized codebases and the inclusion of pre-trained model weights or scripts for fine-tuning greatly enhance usability. On web platforms, incorporating tools for visualization, such as Plotly or Dash²⁴⁸, can bridge the gap between raw data and actionable insights. These deployment strategies are particularly valuable in materials discovery, where interdisciplinary teams rely on intuitive and well-documented tools to foster collaboration.

By enabling seamless access to AI tools and data, open-source infrastructures play a pivotal role in addressing sustainability challenges. These platforms empower researchers to share findings and avoid redundant efforts, thereby conserving resources and reducing carbon footprints associated with trial-and-error experimentation. Resources like OpenKIM²⁴⁹, which curate interatomic models and provide tools for integrating them into simulation workflows, exemplify the power of open-access databases in enhancing reproducibility and collaboration in computational materials science. Furthermore, democratizing AI tools accelerates the identification of materials and processes, amplifying the collective impact of global research efforts. Ultimately, the proliferation of open-source AI infrastructures holds immense potential to catalyze materials innovation.

Accessibility and data transparency

Open data, as mentioned before, plays a pivotal role in the development and transparency of AI systems; its availability and use vary significantly across different models. Truly open-source systems often incorporate openly available datasets, ensuring transparency and reproducibility. For example, BigScience’s BLOOM²⁵⁰ stands apart with its full transparency: the model, training data, and documentation are openly accessible. Open data initiatives, such as The Pile²⁵¹ or Common Crawl²⁵², provide vast repositories of text and other materials that support the training of AI models while allowing scrutiny of their content and provenance. However, even open data requires significant labor to clean, curate, and adapt for AI purposes, adding layers of complexity to the notion of openness.

In contrast, semi-open and closed AI systems often obscure their data sources or rely heavily on proprietary datasets, creating significant issues related to transparency and ethics. Semi-open systems, such as Meta’s LLaMA²⁵³ and Google’s Gemini²⁵⁴, exemplify this challenge. For instance, Meta’s LLaMA does not disclose the specific datasets used for training, which makes it difficult for researchers to evaluate biases or assess whether the data complies with IP laws. Similarly, Google’s Gemini, while claiming to build on web data, provides limited details about its training data, leaving its methodology open to speculation. This lack of transparency restricts researchers’ ability to identify potential biases or understand the model’s limitations. In material discovery, such opacity can hinder the development

Table 5 | Comparative overview of AI model classes in materials science

Model class (representative examples)	AI category	Typical inputs representations	Materials domain/tasks	Primary outputs	Performance metrics (evaluation criteria)
Traditional ML models (SVM, RF, ANN)	Predictive	Handcrafted descriptors; process variables; experimental datasets	Synthesis optimization; property screening (polymers, cellulose, alloys)	Scalar property or process outcome	MAE, RMSE, R^2 ; cross-validation error
Bayesian optimization frameworks	Predictive (optimization)	Surrogate-model features + iterative experimental feedback	Experimental minimization; materials/process optimization	Suggested experiments; optimal conditions	Regret; convergence rate; sample efficiency
Graph neural networks (PU-MPNN, SA-GNN, GeoCGNN)	Predictive (deep learning)	Molecular, polymer, or atomic graphs	Property prediction (formation energy, electronic properties)	Predicted material properties	MAE, RMSE; generalization/OOD testing
Multimodal deep learning (MultiMat)	Predictive (multi-modal)	Combined composition, structure, and attribute data	Elastic and tensorial property prediction	Multivariate material properties	MAE/RMSE; performance gain vs. unimodal baselines
ML interatomic potentials (MACE, CHGNet)	Predictive (physics-aware DL)	Atomic environments/local graphs	Energetics and forces (catalysis, nanomaterials)	Energies, forces, stability indicators	RMSE vs. DFT; ranking consistency
Vision-based deep learning (CNNs, Mask R-CNN)	Predictive (imaging)	Microscopy, diffraction, or process images	Microstructure classification; defect detection	Class labels; segmentation masks	Accuracy, F1; IoU (segmentation)
Variational autoencoders (VAEs)	Generative	Latent encodings of composition, structure, or images	Inverse design; candidate generation	New material candidates	Validity; reconstruction loss; novelty/diversity
Generative adversarial networks (GANs)	Generative	Latent noise + structure/image encodings	Crystal and microstructure generation	Synthetic structures/images	Sample validity; diversity; mode-collapse diagnostics
Diffusion models	Generative	Iterative denoising over graphs, point clouds, or voxels; conditional inputs	Conditional crystal and amorphous-material generation	Physically plausible candidate structures	Structural fidelity; conditional accuracy; novelty; DFT-based screening (formation energy, hull stability)
Agentic AI systems (MOFGPT, MAPPS, MatAgent)	Agentic (LLM-orchestrated)	LLM reasoning + generators + physics tools	Closed-loop materials discovery	Ranked, refined candidates and design decisions	Constraint satisfaction; task success rate; relative improvement in stability, novelty, and targeting over non-agentic baselines

Comparison of predictive, generative, and agentic AI model classes used in materials science, highlighting their primary objectives, representative architectures, data requirements, and typical application domains.

of AI-driven tools that rely on high-quality, traceable datasets to generate accurate predictions for material properties or to guide experimental design. Furthermore, semi-open approaches often rely on datasets that aggregate content without clear attribution, raising ethical concerns about the misuse of IP and the perpetuation of harmful stereotypes embedded in unvetted sources.

Closed systems, like OpenAI's GPT-4, take opacity even further by withholding all information about their training datasets, citing competitive advantages and security concerns. While this might protect proprietary interests, it avoids scrutiny and sidesteps accountability. For example, GPT-4 has been criticized for likely relying on vast web-scraped datasets and potentially including sensitive or copyrighted material^{255,256}, but OpenAI's refusal to disclose this information prevents independent verification. Another example is IBM's Watson²⁵⁷, which operates with heavily curated, proprietary datasets that are inaccessible to the public, limiting external evaluation of its models' robustness and fairness. In material discovery, the use of closed systems introduces challenges, as proprietary data sources may not cover the diversity of materials required for comprehensive predictions or could exclude critical information due to IP restrictions. This not only slows progress but also creates barriers for collaboration, which is essential in advancing fields such as energy storage materials, catalysts, or nanotechnology.

To address these pressing challenges, there is a need for not only adopting open data policies but also implementing robust frameworks that enforce ethical and equitable data usage, ensuring respect for IP and fostering greater accountability in AI development. By prioritizing transparency and accessibility, particularly in fields like material discovery, AI can significantly accelerate innovation, enabling the identification of novel materials with unique properties while upholding ethical standards and fostering trust in AI-driven research.

Emerging technologies

Quantum computing

Quantum computing emerges as a transformative tool for materials discovery, offering unprecedented capabilities to model complex quantum systems with higher accuracy and efficiency than classical methods. Google's Willow quantum chip exemplifies this potential, achieving ground-breaking advancements in error correction and stability. This enables computations that would take classical systems thousands of years to complete to be performed in mere seconds, which could accelerate the discovery and optimization of new materials with unprecedented speed and precision²⁵⁸.

Beyond superconducting qubits, molecular qubits offer a chemically tunable alternative with long coherence times and atomic-scale precision. These qubits, based on transition metals, lanthanides, and actinides, provide enhanced control over spin interactions and could be integrated into hybrid quantum-classical systems for materials discovery²⁵⁹. Their tunability allows for precise engineering of qubit interactions, making them promising candidates for quantum simulations in chemistry and material science.

While traditional techniques such as DFT and coupled cluster methods have provided valuable insights into material properties, their computational cost increases rapidly for large, strongly correlated systems. Quantum algorithms take advantage of quantum parallelism and entanglement, overcoming many of the limitations of purely classical methods. Among these algorithms, the Variational Quantum Eigen solver (VQE)²⁶⁰, has proven effective in calculating the ground-state energies of molecules like GeO₂, SiO₂, and LiH, and has facilitated the creation of databases documenting the corresponding Hamiltonians²⁶¹, enabling more accurate material simulations. Recent work underscores the need to further develop classical embedding strategies and sampling techniques to complement quantum methods, since combining high-level descriptions of strongly correlated regions with lower-level treatments of less critical regions can accelerate workflows and guide the search for materials with improved properties²⁵⁹.

Ultimately, quantum computing offers an innovative pathway for addressing the challenges associated with strongly correlated electronic structures, an area where even advanced classical techniques such as DFT and Coupled Cluster expansions struggle to provide accurate predictions. By leveraging superposition and entanglement, quantum computers promise polynomial or exponential speedups in simulating complex material behaviors²⁵⁹. To fully capitalize on these capabilities, robust noise modeling and quantum error mitigation strategies must be developed. This involves understanding how environmental effects such as interactions with phonons, spins, and electromagnetic fluctuations contribute to qubit decoherence. Researchers are actively developing noise-mitigation techniques, such as readout-error correction, zero-noise extrapolation, and randomized compiling, to reduce error rates and extend coherence times in superconducting and spin-based qubits. These approaches provide practical solutions without requiring the extensive overhead of fully fault-tolerant quantum computing²⁵⁹.

An extension of VQE, known as qubit-ADAPT-VQE, optimizes these procedures by reducing both the number of iterations and circuit complexity, allowing for chemically accurate simulations in studies of adsorption in MOFs and CO₂ capture systems²⁶². Grover's search algorithm²⁶³ has also been used to identify Ni-Ti shape-memory alloys with minimized thermal hysteresis, offering a more efficient path than classical optimization. Meanwhile, quantum phase estimation (QPE) has shown great potential in the study of corrosion-resistant materials by simulating high-temperature behaviors in niobium-based alloys and magnesium systems²⁶⁴. As quantum hardware matures, resource estimates suggest that QPE-based approaches will become even more widely accessible.

An important challenge, however, lies in developing deep knowledge of decoherence mechanisms within these platforms, as each qubit architecture, whether superconducting, spin-defect based, or trapped ions, faces materials-based noise that can significantly reduce quantum coherence²⁵⁹. In superconducting circuits, parasitic two-level systems at the interfaces of thin oxide layers are known to limit coherence times²⁵⁹. For spin-defect qubits in semiconductors, atomic-scale disorder or unintended dopants can introduce unwanted charge fluctuations that shorten qubit lifetimes²⁵⁹. Trapped ions and molecules, meanwhile, are vulnerable to surface noise and patch potentials on electrodes, which can introduce anomalous heating. Although targeted experiments and careful materials synthesis can alleviate some of these issues, accurate modeling of decoherence processes for large and heterogeneous qubit systems on classical hardware remains extraordinarily demanding, driving further research into hybrid quantum-classical approaches for error mitigation and noise modeling²⁵⁹.

A key step in leveraging quantum computing frameworks is encoding classical data into quantum-compatible formats. For example, Qiskit Nature²⁶¹ offers amplitude, basis, and angle encoding, which are essential for both quantum ML and quantum chemistry workflows. PennyLane²⁶⁵, a hybrid quantum-classical platform, provides a variety of encoding strategies while allowing data normalization and scaling before forming quantum states. TensorFlow Quantum (TFQ)²⁶⁶ complements this by embedding classical data pipelines into quantum circuits seamlessly.

Data preprocessing requirements vary across different applications. In quantum chemistry, classical tools like ORCA and PySCF²⁶⁷ help prepare molecular orbitals and Hamiltonians, which can then be converted into qubit-compatible formats for integration with Qiskit Nature and similar environments. For graph-based problems, adjacency matrices are constructed with NetworkX²⁶⁸ before being mapped onto quantum registers, while D-Wave's Ocean SDK²⁶⁹ enables transforming classical optimization tasks into Ising or QUBO models for quantum annealing systems. Throughout these processes, data quality remains crucial, with tools like Pandas and NumPy commonly used for cleaning and transforming datasets, while TensorFlow's Data Pipelines offer more robust solutions for larger-scale projects.

Hybrid platforms like AWS Bracket²⁷⁰ and Rigetti's Forest SDK²⁷¹ unify classical and quantum workflows within a single environment, enabling seamless transitions between classical data preprocessing and quantum

algorithm execution. This integration reduces redundant operations and optimizes computational resources, significantly speeding up research in materials science and allowing for more refined control over design parameters.²⁰¹

By integrating quantum algorithms with classical computational techniques, materials science stands on the brink of a new era of discovery. As hardware continues to evolve and hybrid workflows become more accessible, quantum computing is poised to accelerate the identification of new materials while significantly reducing the environmental and economic costs associated with traditional experimental and computational methods. The synergy between quantum and classical computing will be instrumental in achieving groundbreaking advancements in materials design, ultimately transforming how new materials are discovered, optimized, and implemented in real-world applications.

Quantum ML. QML combines quantum mechanics, such as superposition, entanglement, and quantum state mapping, with ML²⁷² and has recently been used to solve complex problems in materials science. By using parameterized quantum circuits and advanced data encoding methods, QML often outperforms classical approaches, especially when dealing with high-dimensional datasets. One of the key algorithms in QML is the Quantum Neural Network (QNN), which embeds classical data into quantum states through layers of quantum gates²⁷³. This process uses rotations, entanglement, and measurements to extract features more efficiently for tasks like hardware Trojan detection, achieving a 97.27% detection accuracy by combining quantum models with techniques like t-SNE for dimensionality reduction²⁷⁴.

Quantum long short-term memory (QLSTM) networks build on classical LSTM architectures by encoding data into quantum states. This allows QLSTMs to explore larger, more complex feature spaces, improving their ability to model sequences. In retrosynthesis tasks on the USPTO-50k dataset²⁷⁵, QLSTMs achieved 80% accuracy by encoding chemical pathways into quantum states, outperforming classical LSTMs, which struggled with complex chemical dependencies²⁷⁴. Variational quantum classifiers (VQCs) and quantum support vector machines (QSVMs) also show the strength of QML in classification tasks. VQCs optimize quantum circuits to create decision boundaries in high-dimensional quantum spaces, while QSVMs use quantum kernels to project classical data into quantum spaces, making it easier to separate complex datasets. For example, QSVMs reached 99% accuracy in classifying transition metal compounds and oxides by using feature mappings like ZZFeatureMap²⁷⁶.

Generative and regression-based algorithms in QML also play a crucial role in materials science. VAEs combine classical and quantum elements, where quantum circuits generate representations of material structures. This has been used to model redox flow battery configurations, helping design more efficient membranes and electrolytes²⁷⁷. Quantum Gaussian Process Regression (QGPR) enhances classical methods by using quantum kernels, leading to more accurate predictions of properties like piezoelectric coefficients and band gaps²⁷⁸. Quantum Kernel Ridge Regression (QKRR) extends this by mapping classical data into quantum spaces using kernels like NPQC (Non-Parametric Quantum Circuit kernels). QKRR has been shown to accurately predict NMR chemical shifts and scale well for atomic-level descriptors from graph neural networks²⁷⁹.

QML's distinct advantage is especially clear in combinatorial optimization tasks in materials engineering. Quadratic unconstrained binary optimization (QUBO) is a framework for encoding design variables as quadratic cost functions over binary variables. These QUBO formulations are suitable for quantum annealing and Ising models, which allow for efficient exploration of large design spaces. Applications include developing complex metamaterials, such as radiative structures and optical diodes, by identifying the best configurations from many possibilities²⁸⁰. QUBO highlights how QML can handle problems that are too difficult for classical hardware, speeding up innovation in materials science.

By utilizing the unique properties of quantum mechanics, QML can tackle complex computational challenges that traditional methods struggle

with, such as high-dimensional data and optimization problems. This makes QML particularly valuable for advancing the design and prediction of sustainable materials. As quantum hardware improves and hybrid quantum-classical workflows become more accessible, QML is set to play a key role in accelerating the discovery of materials while contributing to reducing the environmental impact of the research process. By supporting open-source infrastructures, QML not only helps drive innovation but also aligns with the growing need for sustainability in materials science.

Building on these insights, current research underscores the importance of fostering open data ecosystems where researchers can share validated qubit designs, computational protocols, and experimental benchmarks²⁵⁹. Such infrastructures are critical for verifying quantum algorithms, refining error-mitigation techniques, and correlating theoretical predictions with experimental realizations. As a result, the integration of classical high-performance computing with nascent quantum processing units stands poised to strengthen both the reliability of quantum simulation outputs and the pace at which new, sustainable materials can be identified²⁵⁹.

As we consider transparency and trust in data-driven material discovery, blockchain technology emerges as a potentially powerful complement, offering secure, auditable, and decentralized mechanisms for collaboration and data sharing.

Blockchain for materials discovery

Blockchain technology is increasingly viewed as a transformative tool in materials science, where it can enhance data security, traceability, multi-institution collaboration, and IP protection. By integrating blockchain-based solutions, researchers address many persistent challenges in collecting, storing, and sharing complex materials data, ultimately fostering more transparent, reproducible, and efficient discovery processes.

Blockchain for data organization and storage. The diverse nature of materials-related information, ranging from atomic-scale structures to bulk mechanical properties, often complicates efforts to standardize systems for storage and retrieval. As datasets grow exponentially through high-throughput experiments and advanced simulations, scalability has become a central obstacle. The decentralized architecture of blockchain networks also adds complexity to ensuring consistent interoperability across multiple nodes and platforms.

A persistent challenge is maintaining data provenance, because many existing systems fail to comprehensively trace how materials data are created and modified, thereby undermining reproducibility and trust in scientific outcomes. To address these issues, emerging approaches focus on specialized on-chain data structures and hybrid storage models. Improved Merkle Trees (MMT), for instance, can map on-chain metadata to off-chain data stores, improving retrieval performance and data integrity without requiring changes to the blockchain's overall structure²⁸¹. Systems that combine blockchains with distributed databases or IPFS store essential metadata on-chain while large files remain off-chain, thus striking a balance between security and scalability²⁸². Dynamic data templates, such as those used in the Materials Genome Initiative, further standardize a wide range of property types and facilitate data analysis, though large-scale adoption remains a work in progress²⁸³.

Researchers have also turned their attention to designing indexing strategies specifically tailored to blockchain-based repositories. Efficient retrieval requires indexing methods capable of handling diverse data without imposing excessive computational loads²⁸⁴. Traditional hierarchical indexes can manage detailed searches, but often prove difficult to maintain in dynamic blockchain environments. Adaptive indexing, which modifies its structure based on query patterns, shows promise in enhancing performance but is not yet widely implemented²⁸². Beyond improved retrieval, blockchain-based systems offer robust traceability. Applications that track the temperature-sensitive handling of aerospace-grade prepreg materials underscore how cryptographic transactions secure each step of the chain-of-custody²⁸⁵. Platforms that use enhanced data formats like Physical Information Files (PIF) allow for the hierarchical organization of material

properties, improving data interoperability and reuse²⁸³. Enhanced data formats, including PIF, also foster hierarchical organization of materials properties and promote interoperability.

When optimized, blockchain platforms can reduce computational overhead through adaptive indexing and selective on-chain storage, thus preventing redundant calculations. Realizing these benefits, however, depends on resolving critical issues such as metadata standardization and the heightened computational demands of certain blockchain solutions. A unified framework that meets performance, scalability, and energy-efficiency goals is therefore essential for accelerating materials discoveries while minimizing resource use.

Secure and transparent data sharing. Robust data sharing is vital in high-impact sectors such as advanced materials, energy systems, and industrial manufacturing, where research often extends across multiple institutions and nations. By providing a distributed, immutable ledger, blockchain technology addresses the long-standing risks of unauthorized data alterations, poor traceability, and reliance on centralized intermediaries.

Secure and transparent data sharing is fundamental to progress in sectors that depend on collaborative innovation, particularly in materials discovery and engineering. In this context, blockchain has emerged as a powerful technologies that enhance data integrity and traceability, while also promoting open-source collaboration.

Smart contracts embed governance and permissions directly into the blockchain network, enabling transparent access control and shared accountability. Decentralization also helps dismantle data silos: for instance, recent work has explored how layered indexing in permissioned blockchains can reduce retrieval bottlenecks in industrial big-data environments^{231,286}. Similarly, timestamping or monitoring the half-lives of sensitive materials is critical in nuclear chemistry, where security and safety regulations demand robust traceability²⁸⁷.

Because traceability and auditability are paramount in advanced materials production, combining Hyperledger Fabric with Merkle Patricia Trees can bolster both the reliability and speed of data retrieval²⁸⁸. Nevertheless, the dual need for transparency and confidentiality remains challenging, especially when handling proprietary or sensitive data. Cryptographic protocols can protect confidentiality without compromising the open character of blockchain transactions²⁸⁹, while incentives for high-quality data contributions help maintain trust and safeguard IP²⁹⁰.

Collaborative and open research. Blockchain also fosters a culture of collaborative and open research in materials discovery. Decentralized ledger systems, combined with other informatics innovations, strengthen reproducibility and enable geographically dispersed teams to coordinate in real time or asynchronously. This integrated approach not only expedites scientific progress but also aligns with universal principles of transparency, sustainability, and FAIR data sharing²⁹¹.

Recent implementations illustrate how blockchain-based workflows facilitate secure and reproducible scientific research. BlockFlow, for example, combines blockchain with the ProvONE model to ensure the traceability of genomics workflows²⁹². Meanwhile, Hyperledger Fabric provides fine-grained access control by leveraging attribute-based encryption, allowing researchers to maintain data ownership while preserving transparency²⁹³. These initiatives reflect efforts within the Materials Genome Initiative, wherein decentralized platforms such as the materials data curation system and the NIST Materials Resource Registry accelerate the development of new materials²⁹⁴.

Despite their promise, blockchain solutions still face computational overhead and resource constraints that arise from large material datasets. Off-chain data storage with on-chain references has emerged as a practical remedy, addressing both scalability and network performance concerns^{291,295}. Interoperability challenges persist, prompting researchers to adopt standardized data formats and dynamic containers (e.g., NMDMS) that facilitate cross-platform integration in accordance with FAIR guiding

principles^{292,296}. Cultural and institutional acceptance are equally important, since even the most advanced technical frameworks must be accompanied by policies and incentives that encourage widespread adoption²⁹⁷.

Beyond secure data sharing, a particularly promising direction lies in the coupling of blockchain infrastructures with federated learning, enabling decentralized collaboration without requiring data centralization and thereby reducing privacy and intellectual-property risks^{295,298}. Federated learning allows distributed actors to collaboratively train models while retaining local control over experimental, computational, or process data, whereas blockchain provides immutable provenance, auditability, and rule-based coordination over model updates and material metadata. In this context, social manufacturing platforms such as Makerchain illustrate how blockchain-enabled chemical and material signatures can be leveraged to coordinate production activities, verify material authenticity, and track transformations across the materials lifecycle²⁹⁹. Although Makerchain does not natively implement federated learning, its provenance-centric architecture offers a natural substrate for integrating decentralized AI workflows, supporting collaborative optimization while avoiding centralized data aggregation.

Complementary efforts further demonstrate how these paradigms can be operationalized across different stages of the materials pipeline. Federated materials data infrastructures such as OPTIMADE³⁰⁰ enable decentralized materials discovery by supporting machine-actionable queries across distributed databases without consolidating raw data, while more tightly integrated systems such as MatSwarm show how federated (or swarm) learning, blockchain-based coordination, and trusted execution environments can be co-designed to support large-scale, non-IID materials collaboration¹⁸⁸. Collectively, these approaches point toward self-organizing and resilient socio-technical manufacturing ecosystems, in which data sovereignty, lifecycle traceability, and collaborative intelligence are embedded directly into the infrastructure rather than imposed post hoc.

Across MSE, such blockchain–AI synergies are beginning to transform how researchers manage complex datasets spanning discovery, manufacturing, and deployment. By embedding decentralized intelligence, provenance-aware governance, and lifecycle traceability directly into the infrastructure, these frameworks improve reproducibility, robustness, and trust while reducing single points of failure. More broadly, they align with open-source principles and ethical AI objectives by promoting transparency, privacy preservation, and responsible data stewardship across the full materials innovation lifecycle. As remaining challenges related to scalability, energy efficiency, and institutional adoption are addressed, blockchain-enabled federated ecosystems are positioned as critical enablers of sustainable, globally integrated materials research and advanced manufacturing.

Carbon footprint and environmental impact of computational resources

The integration of AI-powered computational methods in materials discovery presents transformative opportunities to advance global net-zero goals and align with the United Nations' Sustainable Development Goals (SDGs)³⁰¹. These approaches drive critical innovations in renewable energy technologies, carbon capture, and waste minimization^{302–304}. Additionally, they promote circular economies³⁰³ by enabling the design of recyclable and biodegradable materials. These advancements directly support SDGs such as Affordable and Clean Energy and Sustainable Cities and Communities³⁰⁵, making AI an essential tool in global sustainability efforts.

However, the environmental cost of AI's computational demands poses significant challenges. In 2020, the ICT sector was responsible for 700 MtCO₂e, or 1.4% of global greenhouse gas (GHG) emissions, with data centers and transmission networks contributing between 0.1–0.2%³⁰⁶. The energy requirements of ML are increasing rapidly, with training demands growing by 150% annually and inference by 105%³⁰⁶. Large-scale data centers have seen energy consumption rise by 20% annually, even as overall global data center energy use increased by only 6% between 2010 and 2018. This divergence highlights the growing energy intensity of ML operations within the broader ICT ecosystem.

Training large ML models contributes significantly to these emissions. For example, the BLOOM language model emitted between 24.7 and 50.5 metric tons of CO₂, depending on the accounting method³⁰⁷. Other models, such as GPT-3, have emissions reaching 284 metric tons, equivalent to the lifetime emissions of an average car^{308,309}. Embedded hardware emissions, which account for 40–80% of lifecycle emissions in personal devices, further amplify the environmental impact. By contrast, embedded emissions in data centers are typically less than 10%, thanks to optimized hardware utilization and operational efficiency³⁰⁶.

To address these challenges, a consistent framework for evaluating and reporting GHG emissions is crucial^{306,310}. A key metric for measuring computational intensity is Floating Point Operations (FLOPs), which estimates energy use and enables comparisons across models. For instance, ResNet-50 requires 4 billion FLOPs per image, while ResNet-152 demands 11 billion FLOPs, achieving only marginal accuracy improvements at significantly higher energy costs³⁰⁶. Life cycle assessment (LCA) provides a comprehensive approach to evaluating emissions^{311,312}. It considers both operational energy use and embedded emissions from hardware manufacturing and disposal. Embedded emissions can constitute up to 80% of lifecycle emissions for personal devices, but less than 10% for data centers^{306,311,313}.

Tools like CodeCarbon³¹⁴, Carbontracker³¹⁵, and eco2AI³¹⁶ support real-time monitoring of energy consumption and GHG emissions during model training and inference, integrating grid-specific emission factors. CodeCarbon focuses on user-friendly integration with Python applications, offering visualization tools and considering factors like location, infrastructure, and runtime for estimating emissions. Carbontracker is tailored for deep learning environments, providing predictive capabilities to monitor energy consumption and even stop training if emissions exceed predefined thresholds. Eco2AI emphasizes precision, leveraging a detailed and regularly updated database of regional emission coefficients and hardware specifics to provide accurate insights into energy usage and promote energy-efficient AI architectures. Together, these tools enable researchers and developers to track and minimize the environmental impact of their computational workflows.

Using LCA, FLOPs, and tracking tools in combination provides a comprehensive and precise analysis of the environmental impact of ML models. LCA offers a holistic framework that evaluates the entire lifecycle impact, including operational and embedded emissions. Tracking tools contribute real-time, detailed operational data to refine LCA calculations, while FLOPs serve as a standardized metric to compare models' computational intensity and efficiency. Together, these approaches enhance decision-making by identifying inefficiencies, optimizing resource usage, and reducing environmental impact. This integrated approach not only improves sustainability but also establishes a robust foundation for accountable and environmentally responsible AI development.

Reducing the environmental footprint of AI-powered materials discovery also requires the adoption of best practices. Optimizing model architectures to reduce computational demands is vital. Leveraging advanced hardware such as TPUs, which are 30–80 times more energy-efficient than traditional CPUs and GPUs, can further improve efficiency³⁰⁶. Prioritizing energy-efficient data centers powered by renewable energy is another crucial step. Major technology companies such as Google¹⁵³, AWS¹⁵¹, IBM, and Microsoft are leading the way in enhancing data center energy efficiency by integrating renewable energy and adopting innovative sustainability practices. Google has operated on 100% renewable energy since 2017 and aims for 24/7 carbon-free energy by 2030. AWS, the largest corporate purchaser of renewable energy globally, is on track to reach 100% renewable energy by 2025 and has designed highly efficient data centers with a power usage effectiveness (PUE) rating as low as 1.08. IBM focuses on green data centers through advanced cooling systems, compact designs, and renewable energy use while emphasizing IT equipment recycling. Microsoft, aiming to be carbon negative by 2030, is investing in renewable energy projects and developing innovative solutions like grid-interactive UPS systems to enhance energy efficiency and support grid stability.

Additionally, extending hardware lifespans through recycling and reuse practices can significantly reduce embedded emissions, minimizing the overall lifecycle impact of computational resources. By doing so, the materials discovery community can ensure the development of sustainable AI infrastructure that maximizes transformative potential while minimizing environmental harm.

Ethics and responsible innovation in AI-driven materials discovery

As AI becomes deeply embedded in materials discovery workflows, it is paramount to consider the ethical implications that accompany the deployment of these technologies. Large-scale AI models may reflect hidden biases in their training datasets, leading to skewed predictions, asymmetrical verifications, or inequitable material recommendations. Moreover, opaque “black-box” models can potentially reduce trust in scientific outcomes, especially in fields where accuracy and reproducibility are critical.

Responsible innovation demands transparency, accountability, and equitable access to tools and data. Frameworks such as the OECD AI Principles and UNESCO's AI Ethics Guidelines provide valuable direction by advocating for and emphasizing explainability, transparency, accountability, privacy, and human oversight. Adopting these guidelines in materials discovery helps ensure that the AI infrastructure promotes socially beneficial outcomes and avoids reinforcing systemic inequities.

Additionally, reproducibility and data integrity are essential in scientific research. AI systems must be developed with clear audit trails, version control, and data provenance tags and protocols. XAI approaches, lifecycle assessments (LCA), and ethical review practices should be integrated into model development to promote trust and accountability. When implemented within open-source ecosystems, these practices also empower diverse global communities to collaborate effectively while aligning discovery efforts with broader societal values and environmental priorities.

Case study: end-to-end AI infrastructure for fiber-based packaging materials

AI provides a robust framework for advancing materials development by preserving native material properties, optimizing functional performance, and supporting environmental sustainability. Unlike conventional empirical trial-and-error approaches, AI enables a data-driven workflow that integrates heterogeneous experimental datasets, computational simulations, and predictive modeling to systematically refine material compositions and processing parameters. This section presents a representative end-to-end case study illustrating how the proposed AI-enabled infrastructure can be deployed to design, validate, and scale high-performance, fiber-based packaging materials.

Step 1: data acquisition and knowledge aggregation

The case study begins with the definition of target material attributes relevant to high-performance fiber-based packaging, including moisture resistance, mechanical integrity, recyclability, and compatibility with existing manufacturing processes. To support these objectives, data are aggregated from publicly available materials repositories such as the Materials Project, NOMAD, and domain-specific databases maintained by academic and industrial consortia (Fig. 6). These sources, accessed through standardized application programming interfaces (APIs) or structured dataset exports (e.g., CSV, JSON), provide experimental measurements of mechanical properties, thermal stability, and barrier performance metrics such as water vapor transmission rates.

In parallel, automated data scraping frameworks (e.g., Scrapy, Beautiful Soup) extract complementary information from scientific literature and publicly available industry research reports, capturing fiber morphology descriptors, coating chemistries, and reported barrier enhancements. LLMs are employed to process these unstructured sources by extracting methodological details, experimental conditions, and summarized findings, transforming heterogeneous text-based information into machine-readable representations suitable for downstream analysis. Industry collaborations and

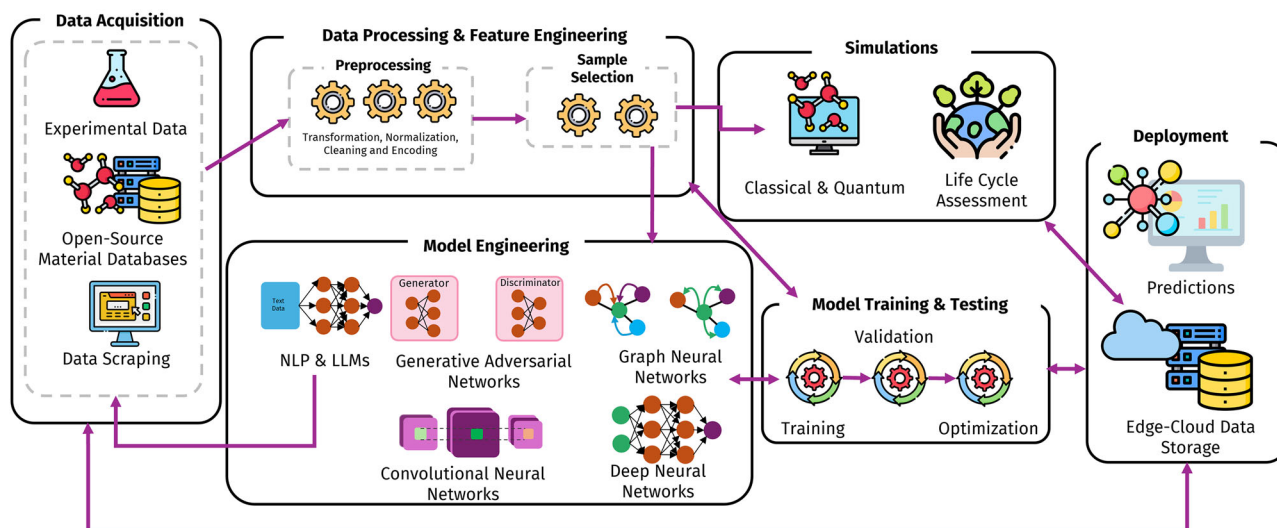


Fig. 6 | AI-driven materials discovery pipeline. Schematic representation of an AI-enabled materials discovery workflow integrating data acquisition from experimental sources, open-source materials databases, and automated data extraction. The pipeline includes data preprocessing and feature engineering, model

development informed by simulation outputs, iterative training and evaluation, and deployment for materials prediction, with feedback loops enabling continuous improvement.

open-data initiatives further contribute pilot-scale manufacturing data and quality-control records, increasing dataset diversity and practical relevance.

Step 2: data standardization and feature engineering

All collected datasets undergo systematic standardization, curation, and metadata enrichment to ensure consistency across sources (Fig. 6). Automated preprocessing scripts clean, normalize, and label entries, while statistical filtering methods implemented using tools such as Pandas and NumPy remove incomplete or anomalous records. The resulting structured dataset enables the identification of correlations among fiber composition, coating formulations, processing conditions, and end-use performance metrics.

Step 3: model development and generative screening

With a curated dataset in place, the AI infrastructure supports multi-model learning and generative exploration. GANs are used to propose candidate polymer–fiber composites, nano-additive dispersions, and bio-based laminate architectures designed to enhance moisture resistance while preserving the porosity, flexibility, and biodegradability of the fiber-based matrix. Bayesian optimization frameworks guide the screening of these candidates by balancing trade-offs among functional performance, raw material availability, cost considerations, and environmental impact indicators, thereby prioritizing formulations that are compatible with sustainable manufacturing constraints.

Step 4: structure–property analysis and multiscale modeling

Candidate formulations are further evaluated using specialized predictive models operating across length scales. CNNs analyze high-resolution scanning electron microscopy (SEM) and atomic force microscopy (AFM) images to detect surface defects, coating non-uniformities, and delamination risks. GNNs model cellulose fibril interactions, crosslinking behavior, and crystalline–amorphous phase distributions, linking molecular-scale organization to predicted oxygen and moisture permeability. These insights are integrated within DNNs that combine thermophysical properties, processing parameters, and degradation models to generate holistic performance predictions.

Step 5: simulation-guided validation and autonomous experimentation

Physics-based simulations serve as a validation layer bridging data-driven predictions and manufacturability. MD simulations estimate water vapor

and oxygen diffusion through composite layers, while finite element analysis (FEA) evaluates mechanical stability under representative stress conditions. These simulations are executed using cloud-based HPC resources, reducing reliance on extensive physical prototyping. Selected candidates are subsequently evaluated using SDLs equipped with robotic synthesis and characterization capabilities, enabling iterative refinement based on AI-generated recommendations.

Step 6: scale-up, monitoring, and deployment

To support translation from laboratory design to industrial production, the infrastructure integrates LCA and techno-economics (TEA) metrics to quantify environmental footprint, energy consumption, end-of-life management, and cost effectiveness. PINNs model pulp refining, coating deposition, and drying kinetics to assess scale-up feasibility. Blockchain-based ledgers secure data provenance, track regulatory compliance, and authenticate sustainability claims across the supply chain.

Pilot-scale manufacturing runs are optimized using digital twin simulations coupled with real-time analytics to anticipate process bottlenecks and refine operating conditions. Edge AI sensors embedded within production lines continuously monitor fiber alignment, coating distribution, and mechanical performance, enabling closed-loop adaptive control of refining, drying, coating, and converting stages. Through this integrated workflow, the case study demonstrates how an AI-powered infrastructure can coordinate data ingestion, modeling, simulation, and manufacturing feedback to support scalable development of sustainable, high-performance fiber-based packaging materials.

Outlook

The integration of AI-powered open-source infrastructures in materials discovery represents a paradigm shift and provides a transformative pathway for developing new materials to address sustainability and carbon footprint challenges. At the heart and foundation of this transformation lies the creation of a unified data architecture that integrates and harmonizes experimental, computational, simulation, and literature-derived datasets. Standardizing protocols for data collection, preprocessing, and storage ensures scalability and interoperability. Advanced AI models and computational tools tailored to materials science challenges, such as predicting material properties, optimizing synthesis processes, and simulating performance, further accelerate innovation. Additionally, collaborative open-source initiatives must

prioritize modularity and accessibility, fostering a global ecosystem where researchers can seamlessly contribute and exchange knowledge. Integrating mechanisms for monitoring environmental impacts within the platform ensures alignment between technological advancements and sustainability objectives.

However, widespread adoption of these technologies brings significant environmental concerns, particularly in the energy consumption and carbon emissions associated with AI operations and cloud-based frameworks. Addressing these issues requires a dual focus on innovation and sustainability. Efforts to integrate renewable energy sources, optimize hardware efficiency, and employ tools for real-time carbon footprint monitoring are crucial. Hybrid computing approaches that balance cloud resources with edge processing can reduce latency and energy consumption while maintaining robust performance. Additionally, leveraging energy-efficient data centers and adopting best practices for resource management can significantly mitigate the environmental impact of AI-driven materials discovery.⁶⁸

Emerging technologies such as SDLs, quantum computing, and blockchain underscore the transformative potential of integrating cutting-edge tools in materials science. SDLs enhance efficiency and reproducibility, enabling real-time optimization of experimental workflows. Quantum computing offers unparalleled accuracy in simulating quantum systems, addressing computational limitations of classical methods. Blockchain technology ensures secure and transparent data sharing while promoting collaboration and maintaining the integrity of complex datasets. Integrating these advancements into open-source frameworks fosters innovation while addressing challenges such as scalability, reproducibility, and interoperability. The convergence of these technologies forms a robust ecosystem for rapid, ethical, and traceable scientific discovery.

Looking forward, the materials science and discovery communities stands at a critical juncture where it must actively shape AI integration, balancing rapid advancements in AI and open-source technology with ethical, sustainable, and collaborative considerations and practices. Adhering to ethical AI principles, fostering open-source collaboration, and employing tools like LCAs and real-time carbon tracking are essential for the development of innovative solutions that align with global environmental goals.

The convergence of AI, open platforms, and sustainable practices represents an opportunity to achieve breakthroughs in materials science while addressing critical global challenges. To this end, future priorities should include:

- Advancing human-centered, explainable AI (XAI) to improve model transparency, trustworthiness, and human interpretability in materials property predictions and synthesis.
- Developing, strengthening, and expanding federated and modular learning infrastructures to enable decentralized AI model training across institutions and entities, while preserving data security, sovereignty, and privacy, and reducing central infrastructure demands.
- Advancing integration with experimental automation, including computer vision, digital twins, robotics, and edge-AI capabilities to close the loop between prediction and validation.
- Driving forward inclusive access to tools, data, and education, ensuring equitable global participation in materials discovery. This, in turn, amplifies materials innovation and incubation.
- Promoting interdisciplinary convergence, combining materials science, AI, ethics, data science, environmental science, and policy to guide responsible innovation. The union of all those disciplines provides results bigger than the sum of the parts.
- Developing hybrid classical-quantum infrastructures to tackle computationally intensive problems in strongly correlated systems and accelerate specific simulations beyond the benchmark limits of classical hardware.
- Implementing FAIR-aligned data practices to ensure material datasets are Findable, Accessible, Interoperable, and Reusable, fostering global collaboration and reproducibility.

- Integrating and embedding sustainability and lifecycle metrics early and by design, into every stage of the AI-driven discovery pipeline, and including real-time carbon footprint tracking and LCAs. This ensures that AI-powered architectures for material and manufacturing align with environmental goals from the outset.

By emphasizing responsible, collaborative, and innovative approaches, the field is positioned to create a future where scientific progress drives sustainability, ensuring that materials discovery contributes meaningfully to both environmental and technological imperatives.

Data availability

The data, tools, and software resources discussed in this study are publicly accessible through a curated GitHub Pages repository, which serves as a supplementary reference to the manuscript. This resource provides organized links to open-source simulation tools, materials databases, and AI frameworks relevant to AI-driven materials discovery. The repository is available at: <https://palarmi.github.io/ai-infrastructure-for-materials-discovery>.

Received: 25 May 2025; Accepted: 3 February 2026;

Published online: 17 February 2026

References

1. Murr, L. E. Examples of Materials Science and Engineering in Antiquity. in *Handbook of Materials Structures, Properties, Processing and Performance* 1–13 (Springer International Publishing, Cham, 2016).
2. Agrawal, A. & Choudhary, A. Perspective: materials informatics and big data: Realization of the “fourth paradigm” of science in materials science. *APL Mater.* **4**, 053208 (2016).
3. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136**, B864–B871 (1964).
4. Gibbs, J. W. On the equilibrium of heterogeneous substances. *Am. J. Sci.* **s3-16**, 441–458 (1878).
5. Alder, B. J. & Wainwright, T. E. Phase transition for a hard sphere system. *J. Chem. Phys.* **27**, 1208–1209 (1957).
6. Rahman, A. Correlations in the motion of atoms in liquid argon. *Phys. Rev.* **136**, A405–A411 (1964).
7. Car, R. & Parrinello, M. Unified approach for molecular dynamics and density-functional theory. *Phys. Rev. Lett.* **55**, 2471–2474 (1985).
8. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–A1138 (1965).
9. Dinic, F. et al. Applied machine learning for developing next-generation functional materials. *Adv. Funct. Mater.* **31**, 2104195 (2021).
10. Guo, K., Yang, Z., Yu, C. H. & Buehler, M. J. Artificial intelligence and machine learning in design of mechanical materials. *Mater. Horiz.* **8**, 1153–1172 (2021).
11. Merchant, A. et al. Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023).
12. Takahashi, K. & Tanaka, Y. Materials informatics: a journey towards material design and synthesis. *Dalton Trans.* **45**, 10497–10499 (2016).
13. Sun, S. et al. Accelerated development of perovskite-inspired materials via high-throughput synthesis and machine-learning diagnosis. *Joule* **3**, 1437–1451 (2019).
14. Cherukara, M. J. & Mannodi-Kanakkithodi, A. Deep learning the properties of inorganic perovskites. *Model. Simul. Mat. Sci. Eng.* **30**, 034005 (2022).
15. Debnath, A. et al. Generative deep learning as a tool for inverse design of high entropy refractory alloys. *J. Mater. Inform.* **1**, 3 (2021).
16. Kolev, M. A novel approach to predict the effect of chemical composition and thermo-mechanical processing parameters on

- Cu–Ni–Si alloys using a hybrid deep learning and ensemble learning model. *Compos. Commun.* **48**, 101903 (2024).
17. Torrents-Barrena, J. & Pellicer, E. Smart cellulose composites: advanced applications and properties prediction using machine learning. In *Encyclopedia of Materials: Composites* Vol. 2 (ed. Hashmi, S.) 527–538 (Elsevier, 2021).
 18. Manujesh, B. J. & Prajna, M. R. Damage detection and classification for sandwich composites using machine learning. *Mater. Today Proc.* **52**, 702–709 (2022).
 19. Ferdosian, F., Pan, Z., Gao, G. & Zhao, B. Bio-based adhesives and evaluation for wood composites application. *Polymers* **9**, 70 (2017).
 20. Pollice, R. et al. Data-driven strategies for accelerated materials design. *Acc. Chem. Res.* **54**, 849–860 (2021).
 21. Almeida, P. de & Bernardino, J. A comprehensive overview of open source big data platforms and frameworks. *Serv. Trans. Big Data* **2**, 15–33 (2015).
 22. Nosek, B. A. et al. Promoting an open research culture. *Science* **348**, 1422–1425 (2015).
 23. Jain, A. et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
 24. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *JOM* **65**, 1501–1509 (2013).
 25. Scheidgen, M. et al. NOMAD: a distributed web-based platform for managing materials science research data. *J. Open Source Softw* **8**, 5388 (2023).
 26. Elliott, K. C. & Resnik, D. B. Making open science work for science and society. *Environ. Health Perspect.* **127**, 75002 (2019).
 27. Oellermann, M. et al. Harnessing the benefits of open electronics in science. *Integr. Comp. Biol.* **62**, 1061–1075 (2021).
 28. Samuel, G. & Lucivero, F. Responsible open science: moving towards an ethics of environmental sustainability. *Publications* **8**, 54 (2020).
 29. Pearce, J. M. Open source research in sustainability. *Sustainability* **5**, 238–243 (2012).
 30. Rao, R. et al. AI-powered municipal solid waste management: a comprehensive review from generation to utilization. *Front. Energy Res.* **13**, 1670679 (2025).
 31. Nair, A., Pillai, S. V. & Senthil Kumar, S. A. Towards emerging Industry 5.0 – a review-based framework. *J. Strategy Manag.* <https://doi.org/10.1108/J SMA-12-2022-0217> (2024).
 32. European Commission. *Industry 5.0: Towards a Sustainable, Human-Centric and Resilient European Industry* (Directorate-General for Research and Innovation, 2021).
 33. Kaur, M. J., Mishra, V. P. & Maheshwari, P. The convergence of Digital Twin, IoT, and machine learning: transforming data into action. In *Internet of Things* (eds. Mahmood, Z. & Ramachandran, M.) 3–17 (Springer, 2020).
 34. Bai, J. et al. A dynamic knowledge graph approach to distributed self-driving laboratories. *Nat. Commun.* **15**, 462 (2024).
 35. Adel, A. & Alani, H. S. N. Human-centric collaboration and industry 5.0 framework in smart cities and communities: fostering sustainable development goals 3, 4, 9, and 11 in society 5.0. *Smart Cities* **7**, 1723–1775 (2024).
 36. Stahl, B. C. Addressing ethical issues in AI. In *Artificial Intelligence for a Better Future* 55–79 (Springer, 2021).
 37. Gabsi, A. E. H. Integrating artificial intelligence in industry 4.0: insights, challenges, and future prospects—a literature review. *Ann. Oper. Res.* 1–28, <https://doi.org/10.1007/s10479-024-06012-6> (2024).
 38. Greenpeace. *Oil in the Cloud: How Advanced Computing is Being Used throughout Oil Production* (Greenpeace USA, 2020).
 39. Heikkilä, M. AI's Carbon Footprint Is Bigger Than You Think. *MIT Technology Review* <https://www.technologyreview.com/2023/12/05/1084417/ais-carbon-footprint-is-bigger-than-you-think/> (2023).
 40. Pan, I., Mason, L. R. & Matar, O. K. Data-centric engineering: integrating simulation, machine learning and statistics. Challenges and opportunities. *Chem. Eng. Sci.* **249**, 117271 (2022).
 41. Himanen, L., Geurts, A., Foster, A. S. & Rinke, P. Data-driven materials science: status, challenges, and perspectives. *Adv. Sci.* **6**, 1900808 (2019).
 42. Heid, E., McGill, C. J., Vermeire, F. H. & Green, W. H. Characterizing uncertainty in machine learning for chemistry. *J. Chem. Inf. Model.* **63**, 4012–4029 (2023).
 43. Ghiringhelli, L. M. et al. Shared metadata for data-centric materials science. *Sci. Data* **10**, 626 (2023).
 44. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
 45. Orio, M., Pantazis, D. A. & Neese, F. Density functional theory. *Photosynth. Res.* **102**, 443–453 (2009).
 46. Vasudevan, R. K. et al. Materials science in the artificial intelligence age: high-throughput library generation, machine learning, and a pathway from correlations to the underpinning physics. *MRS Commun.* **9**, 821–838 (2019).
 47. Cohen, A. J., Mori-Sánchez, P. & Yang, W. Challenges for density functional theory. *Chem. Rev.* **112**, 289–320 (2011).
 48. Mao, Q. et al. Classical and reactive molecular dynamics: principles and applications in combustion and energy systems. *Prog. Energy Combust. Sci.* **97**, 101084 (2023).
 49. Verma, P. & Truhlar, D. G. Status and challenges of density functional theory. *Trends Chem* **2**, 302–318 (2020).
 50. Valsson, O., Tiwary, P. & Parrinello, M. Enhancing important fluctuations: rare events and metadynamics from a conceptual viewpoint. *Annu. Rev. Phys. Chem.* **67**, 159–184 (2016).
 51. Elber, R. Long-timescale simulation methods. *Curr. Opin. Struct. Biol.* **15**, 151–156 (2005).
 52. Reible, B. M., Hille, J. F., Hartmann, C. & Delle Site, L. Finite-size effects and thermodynamic accuracy in many-particle systems. *Phys. Rev. Res.* **5**, 023156 (2023).
 53. Lopes, P. E. M., Guvench, O. & MacKerell, A. D. Current status of protein force fields for molecular dynamics simulations. In *Methods in Molecular Biology* Vol. 1215 (eds. Monticelli, L. & Salonen, E.) 47–71 (Humana Press, 2015).
 54. Davies, D. W. et al. Computational screening of all stoichiometric inorganic materials. *Chem. Mater.* **1**, 617–627 (2016).
 55. Axelrod, S. et al. Learning matter: materials design with machine learning and atomistic simulations. *Acc. Mater. Res.* **3**, 343–357 (2022).
 56. Barrios, N., Parra, J. G., Venditti, R. A. & Pal, L. Elucidation of temperature-induced water structuring on cellulose surfaces for environmental and energy sustainability. *Carbohydr. Polym.* **329**, 121799 (2024).
 57. Jain, A. Machine learning in materials research: developments over the last decade and challenges for the future. *Curr. Opin. Solid State Mater. Sci.* **33**, 101189 (2024).
 58. Shi, L., Pan, B., Chen, W. & Wang, Z. Deep learning-based multifidelity surrogate modeling for high-dimensional reliability prediction. *ASCE-ASME J. Risk Uncertain. Eng. Syst. Part B Mech. Eng.* **10**, 031106 (2024).
 59. Park, H., Li, Z. & Walsh, A. Has generative artificial intelligence solved inverse materials design?. *Matter* **7**, 2355–2367 (2024).
 60. Zeni, C. et al. A generative model for inorganic materials design. *Nature* **639**, 624–632 (2025).
 61. Lu, S., Zhou, Q., Chen, X., Song, Z. & Wang, J. Inverse design with deep generative models: next step in materials discovery. *Natl. Sci. Rev.* **9**, nwac111 (2022).

62. Jena, P. & Sun, Q. Theory-guided discovery of novel materials. *J. Phys. Chem. Lett.* **12**, 6499–6513 (2021).
63. Salas, M. et al. Hyperspectral imaging for real-time waste materials characterization and recovery using endmember extraction and abundance detection. *Matter* **8**, 102365 (2025).
64. Medina, J. et al. Accelerating the adoption of research data management strategies. *Matter* **5**, 3614–3642 (2022).
65. National Center for Biotechnology Information. PubChem 2025 update. *Nucleic Acids Res.* **53**, D1516–D1525 (2024).
66. European Bioinformatics Institute. ChEMBL: a large-scale bioactivity database for drug discovery. <https://www.ebi.ac.uk/chembl/#> (2024).
67. Crystallography Open Database. *Nucleic Acids Research/COD Project* <https://www.crystallography.net/cod/> (Accessed 25 September 2024).
68. Tom, G. et al. Self-driving laboratories for chemistry and materials science. *Chem. Rev.* **124**, 9633–9732 (2024).
69. Tamura, R., Tsuda, K. & Matsuda, S. NIMS-OS: an automation software to implement a closed loop between artificial intelligence and robotic experiments in materials science. *SSci. Technol. Adv. Mater. Methods* **3**, 2232297 (2023).
70. Volk, A. A. et al. AlphaFlow: autonomous discovery and optimization of multi-step chemistry using a self-driven fluidic lab guided by reinforcement learning. *Nat. Commun.* **14**, 1403 (2023).
71. Bayley, O., Savino, E., Slattery, A. & Noël, T. Autonomous chemistry: navigating self-driving labs in chemical and material sciences. *Matter* **7**, 2382–2398 (2024).
72. Seifrid, M. et al. Chemspeed: an open-source python interface for Chemspeed robotic chemistry and materials platforms. *Digit. Discov.* **3**, 1319–1326 (2024).
73. Du, X. et al. Elucidating the full potential of OPV materials utilizing a high-throughput robot-based platform and machine learning. *Joule* **5**, 495–506 (2021).
74. Volk, A. A., Campbell, Z. S., Ibrahim, M. Y. S., Bennett, J. A. & Abolhasani, M. Flow chemistry: a sustainable voyage through the chemical universe en route to smart manufacturing. *Annu. Rev. Chem. Biomol. Eng.* **13**, 45–72 (2022).
75. Tabor, D. P. et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.* **3**, 5–20 (2018).
76. MacLeod, B. P. et al. Self-driving laboratory for accelerated discovery of thin-film materials. *Sci. Adv.* **6**, eaaz8867 (2020).
77. Leng, J. et al. Physics-informed machine learning in intelligent manufacturing: a review. *J. Intell. Manuf.* 1–43, <https://doi.org/10.1007/s10845-025-02641-1> (2025).
78. Ra, D. et al. Physics-informed machine learning across manufacturing processes: recent advances, challenges, and directions. *J. Manuf. Syst.* **85**, 72–95 (2026).
79. Zhu, S.-P. et al. Physics-informed machine learning and its structural integrity applications: state of the art. *Phil. Trans. R. Soc. A* **381**, 20220406 (2023).
80. Raissi, M., Perdikaris, P. & Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378**, 686–707 (2019).
81. Karniadakis, G. E. et al. Physics-informed machine learning. *Nat. Rev. Phys.* **3**, 422–440 (2021).
82. Liu, D., Lu, Y. & Wang, Y. Physics-informed machine learning for metal additive manufacturing. in *Machine Learning for Powder-Based Metal Additive Manufacturing* 77–106 (Elsevier, 2025).
83. Tian, M. et al. Physics-informed machine learning-based real-time long-horizon temperature fields prediction in metallic additive manufacturing. *Commun. Eng.* **4**, 168 (2025).
84. Sajadi, P., Dehaghani, M. R., Tang, Y. & Wang, G. G. Two-dimensional temperature field prediction with in-situ data in metal additive manufacturing using physics-informed neural networks. *Eng. Appl. Artif. Intell.* **150**, 110636 (2025).
85. Chew, A. K. et al. Advancing material property prediction: using physics-informed machine learning models for viscosity. *J. Cheminform.* **16**, 31 (2024).
86. Moon, H., Lee, J., Yu, J. & Ryu, S. Physics-informed machine learning for material characterization: a perspective on data-efficient discovery through physics-informed neural networks. *Int. J. AI Mater. Des.* **2**, 1–9 (2025).
87. Liu, S. et al. Physics-informed machine learning for composition–process–property design: Shape memory alloy demonstration. *Applied Materials Today* **22**, 100898 (2021).
88. Hong, Y., Hou, B., Jiang, H. & Zhang, J. Machine learning and artificial neural network accelerated computational discoveries in materials science. *WIREs Comput. Mol. Sci.* **10**, e1450 (2020).
89. Yao, Z. et al. Machine learning for a sustainable energy future. *Nat. Rev. Mater.* **8**, 202–215 (2023).
90. Vasudevan, R., Pilania, G. & Balachandran, P. V. Machine learning for materials design and discovery. *J. Appl. Phys.* **129**, 070401 (2021).
91. Thompson, A. P. et al. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* **271**, 108171 (2022).
92. Abraham, M. J. et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
93. Phillips, J. C. et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* **153**, 044130 (2020).
94. Giannozzi, P. et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Matter* **21**, 395502 (2009).
95. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B Condens. Matter Mater. Phys.* **54**, 11169–11186 (1996).
96. Mabrouk, Y. et al. Reactive molecular dynamics simulations of lithium-ion battery electrolyte degradation. *Sci. Rep.* **14**, 10281 (2024).
97. Aldabesh, A. D. A numerical study of CsSn_xBr_{3-x} perovskite material as an electron transport layer (ETL), in the perovskite solar cell of a photovoltaic system by molecular dynamics method with LAMMPS software: the effects of external convective heat transfer. *Eng. Anal. Bound. Elem.* **147**, 51–58 (2023).
98. Fernandes, S. Q. & Madhuranthakam, C. M. R. Molecular dynamics simulation of a superhydrophobic cellulose derivative targeted for eco-friendly packaging material. *Macromol. Theory Simul.* **30**, 2000056 (2021).
99. Mahfud, R. Molecular dynamics computational study of sustainable green surfactant for application in chemical enhanced oil recovery. *ACS Omega* **9**, 27177–27191 (2024).
100. Mariano Colombari, F., Marcos Nascimento, V., Liu, Y. L., De Moraes Rocha, G. J. & Driemeier, C. Density functional theory with implicit solvents for accurate estimation of aqueous and organic solvation free energies of lignin fragments. *ACS Sustain. Chem. Eng.* **10**, 10870–10878 (2022).
101. Batchelor, C. et al. Computationally-accelerated prediction of polyester-melamine coatings degradation to design sustainable organically-coated steels for outdoor applications. *RSC Adv.* **14**, 18343–18354 (2024).
102. Singh, M., Shukla, A. & Chakraborty, B. Highly efficient hydrogen storage of a Sc decorated biphenylene monolayer near ambient temperature: ab initio simulations. *Sustain. Energy Fuels* **7**, 996–1010 (2022).
103. Kistanov, A. A. et al. First-principles prediction of two-dimensional B₃C₂P₃ and B₂C₄P₂: structural stability, fundamental properties, and renewable energy applications. *J. Phys. Chem. Lett.* **12**, 3436–3442 (2021).

104. Itas, Y. S., Suleiman, A. B., Yamusa, A. S., Razali, R. & Danmadami, A. M. Ab initio studies of the structural and electronic properties for single-walled armchair MgONT, SiCNTs and ZnONTs for next generations' optoelectronics. *Gadua J. Pure Allied Sci.* **1**, 160–165 (2022).
105. Curtarolo, S. et al. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **58**, 227–235 (2012).
106. Fiedler, L., Shah, K., Bussmann, M. & Cangi, A. Deep dive into machine learning density functional theory for materials science and chemistry. *Phys. Rev. Mater.* **6**, 040301 (2022).
107. Pederson, R., Kalita, B. & Burke, K. Machine learning and density functional theory. *Nat. Rev. Phys.* **4**, 357–358 (2022).
108. Dral, P. O. Quantum chemistry in the age of machine learning. *J. Phys. Chem. Lett.* **11**, 2336–2347 (2020).
109. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
110. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
111. Batzner, S. et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **13**, 2453 (2022).
112. Lo, S., Seifrid, M., Gaudin, T. & Aspuru-Guzik, A. Augmenting polymer datasets by iterative rearrangement. *J. Chem. Inf. Model.* **63**, 4266–4276 (2023).
113. Tempke, R. & Musho, T. Autonomous generation of single photon emitting materials. *Nanoscale* **16**, 10239–10249 (2024).
114. Allen, C. et al. Deep learning strategies for addressing issues with small datasets in 2D materials research: microbial corrosion. *Front. Microbiol.* **13**, 1059123 (2022).
115. Khder, M. Web scraping or web crawling: state of art, techniques, approaches and application. *Int. J. Adv. Soft Comput. Appl.* **13**, 145–168 (2021).
116. Leonard Richardson. Beautiful Soup 4.4.0 documentation. *Beautiful Soup Project* <https://beautiful-soup-4.readthedocs.io/en/latest/> [Accessed 7 November 2024].
117. The Selenium Browser Automation Project. <https://www.selenium.dev/documentation/> [Accessed 7 November 2024].
118. Puppeteer. <https://pptr.dev/> [Accessed 7 November 2024].
119. Octoparse. Web Scraping Tool and Free Web Crawlers: Octoparse. <https://www.octoparse.com/> [Accessed 7 November 2024].
120. ParseHub Free web scraping - The most powerful web scraper. <https://www.parsehub.com/> [Accessed 7 November 2024].
121. GitHub - scrapinghub/portia: Visual scraping for Scrapy. <https://github.com/scrapinghub/portia> [Accessed 7 November 2024].
122. Sequentum Low Code Web Data Pipelines at Enterprise Scale. <https://www.contentgrabber.com/> [Accessed 7 November 2024].
123. Kalhor, P. et al. Functional material systems enabled by automated data extraction and machine learning. *Adv. Funct. Mater.* **34**, 2302630 (2024).
124. Polak, M. P. & Morgan, D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nat. Commun.* **15**, 1569 (2024).
125. Huang, S. & Cole, J. M. BatteryDataExtractor: battery-aware text-mining software embedded with BERT models. *Chem. Sci.* **13**, 11487–11495 (2022).
126. Itani, S., Zhang, Y. & Zang, J. The northeast materials database for magnetic materials. *Nat. Commun.* **16**, 9415 (2025).
127. Shetty, P. et al. A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. *NPJ Comput. Mater.* **9**, 52 (2023).
128. Gupta, S., Mahmood, A., Shetty, P., Adeboye, A. & Ramprasad, R. Data extraction from polymer literature using large language models. *Commun. Mater.* **5**, 269 (2024).
129. Shetty, P. & Ramprasad, R. Automated knowledge extraction from polymer literature using natural language processing. *iScience* **24**, 101922 (2021).
130. Zheng, Z. et al. Shaping the water-harvesting behavior of metal-organic frameworks aided by fine-tuned GPT models. *J. Am. Chem. Soc.* **145**, 28284–28295 (2023).
131. Makino, K., Kuniyoshi, F., Ozawa, J. & Miwa, M. Extracting and analyzing inorganic material synthesis procedures in the literature. *IEEE Access* **10**, 31524–31537 (2022).
132. Zheng, Z. et al. Image and data mining in reticular chemistry powered by GPT-4V. *Digit. Discov.* **3**, 491–501 (2024).
133. Shetty, P., Adeboye, A., Gupta, S., Zhang, C. & Ramprasad, R. Accelerating materials discovery for polymer solar cells: data-driven insights enabled by natural language processing. *Chem. Mater.* **36**, 7676–7689 (2024).
134. Dong, Q. & Cole, J. M. Snowball 2.0: generic material data parser for ChemDataExtractor. *J. Chem. Inf. Model.* **63**, 7045–7055 (2023).
135. Mukherjee, A., Giri, D. & Rajan, K. Uncertainty-informed screening for safer solvents used in the synthesis of perovskites via language models. *J. Chem. Inf. Model.* **65**, 7901–7918 (2025).
136. Microsoft Corporation. Microsoft Excel. <https://office.microsoft.com/excel>.
137. pandas - Python Data Analysis Library. <https://pandas.pydata.org/> [Accessed 11 November 2024].
138. Welcome to Python.org. <https://www.python.org/> [Accessed 11 November 2024].
139. Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. dplyr: A Grammar of Data Manipulation. <https://dplyr.tidyverse.org> (2023).
140. Ham, K. OpenRefine (version 2.5). Free, open-source tool for cleaning and transforming data. *J. Med. Libr. Assoc.* **101**, 233 (2013).
141. R Core Team. R: A Language and Environment for Statistical Computing. <https://www.R-project.org/> (2021).
142. Shaikh, E., Mohiuddin, I., Alufaisan, Y. & Nahvi, I. Apache Spark: A Big Data Processing Engine. in *2019 2nd IEEE Middle East and North Africa COMMunications Conference (MENACOMM)* 1–6 (IEEE, 2019).
143. Talend A Complete, Scalable Data Management Solution. <https://www.talend.com/> [Accessed 11 November 2024].
144. rdkit. <https://doi.org/10.5281/ZENODO.13990314>.
145. KNIME Analytics Platform. <https://www.knime.com> [Accessed 11 November 2024].
146. AI Analytics Platform - Alteryx. <https://www.alteryx.com/> [Accessed 11 November 2024].
147. Sakr, S., Liu, A., Batista, D. M. & Alomari, M. A survey of large scale data management approaches in cloud environments. *IEEE Commun. Surv. Tut.* **13**, 311–336 (2011).
148. Trinks, S. & Felden, C. Edge computing architecture to support real time analytic applications: a state-of-the-art within the application area of smart factory and industry 4.0. in *2018 IEEE International Conference on Big Data (Big Data)* 2930–2939 (IEEE, 2018).
149. Chen, C. et al. Accelerating computational materials discovery with machine learning and cloud high-performance computing: from large-scale screening to experimental validation. *J. Am. Chem. Soc.* **146**, 20009–20018 (2024).
150. Strieth-Kalthoff, F. et al. Delocalized, asynchronous, closed-loop discovery of organic laser emitters. *Science* **384**, eadk9227 (2024).
151. Amazon Sustainability Report 2023. <https://sustainability.aboutamazon.com/2023-report> [Accessed 11 October 2024].
152. 2023 ESG REPORT. <https://newsroom.ibm.com/2023-04-11-IBM-Releases-2023-Impact-Report> [Accessed 9 December 2024].
153. Google 2024 Environmental Report - Google Sustainability. <https://sustainability.google/reports/google-2024-environmental-report> (2024) [Accessed 9 December 2024].
154. Singh, S. Optimize cloud computations using edge computing. in *2017 International Conference on Big Data, IoT and Data Science (BID)* vols 2018–January 49–53 (IEEE, 2017).

155. Bangui, H., Rakrak, S., Raghay, S. & Buhnova, B. Moving to the edge-cloud-of-things: recent advances and future research directions. *Electronics* **7**, 309 (2018).
156. GitHub - emmo-repo/EMMO: Elementary Multiperspective Material Ontology (EMMO). <https://github.com/emmo-repo/EMMO> [Accessed 17 November 2024].
157. Huber, S. P. et al. AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. *Sci. Data* **7**, 300 (2020).
158. Gupta, A. et al. Bridging eResearch infrastructure and experimental materials science process in the quantum data hub. Preprint at <http://arxiv.org/abs/2405.19706> (2024).
159. Wang, Z. et al. AlphaMat: a material informatics hub connecting data, features, models and applications. *NPJ Comput. Mater.* **9**, 130 (2023).
160. Borysov, S. S., Geilhufe, R. M. & Balatsky, A. V. Organic materials database: an open-access online database for data mining. *PLoS One* **12**, e0171501 (2017).
161. Ong, S. P. et al. Python Materials Genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
162. Ward, L. et al. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60–69 (2018).
163. Pedregosa F et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
164. Hjorth Larsen, A. et al. The atomic simulation environment—a Python library for working with atoms. *J. Phys. Condens. Matter* **29**, 273002 (2017).
165. Allison, J., Collins, P. & Spanos, G. *Proceedings of the 3rd World Congress on Integrated Computational Materials Engineering (ICME 2015). Proceedings of the 1st World Congress on Integrated Computational Materials Engineering, ICME* (Springer International Publishing, Cham, 2016).
166. Abadi, M. et al. TensorFlow: a system for large-scale machine learning. <https://arxiv.org/abs/1603.04467> (2016).
167. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, 8026–8037 (2019).
168. Cao, Y., Taghvaie Nakhjiri, A. & Ghadiri, M. Different applications of machine learning approaches in materials science and engineering: comprehensive review. *Eng. Appl. Artif. Intell.* **135**, 108783 (2024).
169. Agrawal, A. & Choudhary, A. Deep materials informatics: applications of deep learning in materials science. *MRS Commun.* **9**, 779–792 (2019).
170. Adetunla, A., Akinlabi, E., Jen, T. C. & Ajibade, S.-S. Harnessing the power of artificial intelligence in materials science: an overview. in *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)* 1–6 (IEEE, 2024).
171. Zhang, X. et al. Polymer-unit graph: advancing interpretability in graph neural network machine learning for organic polymer semiconductor materials. *J. Chem. Theory Comput.* **20**, 2908–2920 (2024).
172. Focassio, B., M. Freitas, L. P. & Schleder, G. R. Performance assessment of universal machine learning interatomic potentials: challenges and directions for materials’ surfaces. *ACS Appl. Mater. Interfaces* **17**, 13111–13121 (2025).
173. Jha, D. et al. Enabling deeper learning on big data for materials informatics applications. *Sci. Rep.* **11**, 4244 (2021).
174. Dong, R. et al. DeepXRD, a deep learning model for predicting XRD spectrum from material composition. *ACS Appl. Mater. Interfaces* **14**, 40102–40115 (2022).
175. Mishra, S. P. & Rahul, M. R. A comparative study and development of a novel deep learning architecture for accelerated identification of microstructure in materials science. *Comput. Mater. Sci.* **200**, 110815 (2021).
176. Moro, V. et al. Multimodal foundation models for material property prediction and discovery. *Newton* **1**, 100016 (2025).
177. Cui, Y. et al. SA-GNN: Prediction of material properties using graph neural network based on multi-head self-attention optimization. *AIP Adv.* **14**, 055033 (2024).
178. Ronchetti, C. et al. Machine learning techniques for data analysis in materials science. in *2022 AEIT International Annual Conference (AEIT)* 1–6 <https://doi.org/10.23919/AEIT56783.2022.9951839> (IEEE, 2022).
179. Liz, M. F., Nartova, A. V., Matveev, A. V. & Okunev, A. G. Using computer vision and deep learning for nanoparticle recognition on scanning probe microscopy images: modified U-net approach. in *2020 Science and Artificial Intelligence conference (S.A.I.ence)* 13–16 (IEEE, 2020). <https://doi.org/10.1109/S.A.I.ence50533.2020.9303184>.
180. Shi, M. et al. Visualization of deep convolutional neural networks to investigate porous nanocomposites for electromagnetic interference shielding. *ACS Appl. Mater. Interfaces* **15**, 22602–22615 (2023).
181. Wankhade, N. P. et al. Metallurgical microstructure classification using CNN: a comprehensive study on heat treatment analysis for steel. *Mater. Today Proc.* <https://doi.org/10.1016/j.matpr.2024.05.066> (2024).
182. Devadig, R., Shrestha Gurung, B. D., Gnimpieba, E., Jasthi, B. & Gadhamshetty, V. Computational methods for biofouling and corrosion-resistant graphene nanocomposites. A transdisciplinary approach. in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 4494–4496 (IEEE, 2023). <https://doi.org/10.1109/BIBM58861.2023.10385810>.
183. Li, K. et al. Probing out-of-distribution generalization in machine learning for materials. *Commun. Mater.* **6**, 9 (2025).
184. Noda, K., Wakiuchi, A., Hayashi, Y. & Yoshida, R. Advancing extrapolative predictions of material properties through learning to learn using extrapolative episodic training. *Commun. Mater.* **6**, 36 (2025).
185. McMahan, B. et al. Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proc. Mach. Learn. Res.* **54**, 1273–1282 (2017).
186. Guo, J. et al. Neural network training method for materials science based on multi-source databases. *Sci. Rep.* **12**, 15326 (2022).
187. Huang, W. & Barnard, A. S. Federated data processing and learning for collaboration in the physical sciences. *Mach. Learn. Sci. Technol.* **3**, 045023 (2022).
188. Wang, R. et al. MatSwarm: trusted swarm transfer learning driven materials computation for secure big data sharing. *Nat. Commun.* **15**, 9290 (2024).
189. Zhou, S. & Tian, W. Privacy-preserving process-defect modelling for metal-based additive manufacturing processes: a federated learning-based case study. *Manuf. Lett.* **44**, 1016–1025 (2025).
190. Mehta, M., Bimrose, M. V., McGregor, D. J., King, W. P. & Shao, C. Federated learning enables privacy-preserving and data-efficient dimension prediction and part qualification across additive manufacturing factories. *J. Manuf. Syst.* **74**, 752–761 (2024).
191. Zhou, Q., Chen, X. & Wang, J. Machine learning assisted material discovery: a small data approach. *Acc. Mater. Res.* **6**, 685–694 (2025).
192. Otyepka, M., Pykal, M. & Otyepka, M. Advancing materials discovery through artificial intelligence. *Appl. Mater. Today* **47**, 102981 (2025).
193. Thanh, H. V., Dai, Z. & Rahimi, M. Data-driven explainable machine learning approaches for predicting hydrogen adsorption in porous crystalline materials. *J. Alloys Compd.* **1028**, 180709 (2025).
194. Nait Amar, M., Alatefi, S., Alkhouh, A. & Agwu, O. E. Modeling CO₂ adsorption capacity of diverse porous adsorbent materials: robust machine learning frameworks and insights from local and global explainable artificial intelligence. *Adsorption* **31**, 106 (2025).

195. Padhy, S. P. et al. Integrated design framework for titanium aluminides through interpretable machine learning. *J. Alloys Compd.* **1047**, 184937 (2025).
196. Omidvar, N. et al. Explainable AI for optimizing oxygen reduction on Pt monolayer core-shell catalysts. *Electrochem. Sci. Adv.* **4**, e202300028 (2024).
197. Klein, L. et al. Discovering process dynamics for scalable perovskite solar cell manufacturing with explainable AI. *Adv. Mater.* **36**, 2307160 (2024).
198. Fukui, Y., Minami, K., Yoshikawa, G., Tsuda, K. & Tamura, R. Harnessing explainable AI to explore structure-activity relationships in artificial olfaction. *ACS Appl. Mater. Interfaces* **17**, 52728–52737 (2025).
199. Wang, A. Y. T., Mahmoud, M. S., Czasny, M. & Gurlo, A. CrabNet for explainable deep learning in materials science: bridging the gap between academia and industry. *Integr. Mater. Manuf. Innov.* **11**, 41–56 (2022).
200. Wang, K. et al. XElemNet: towards explainable AI for deep neural networks in materials science. *Sci. Rep.* **14**, 25178 (2024).
201. Wang, J., Kwon, H., Oh, S.-H., Kim, H. S. & Lee, B.-J. Interpretable and physics-informed modeling of solidification in alloy systems: a generalized framework for multi-component prediction. *Acta Mater.* **286**, 120716 (2025).
202. Huang, Y., Wang, S.-H., Cao, S., Achenie, L. E. K. & Xin, H. Decoding the stability of transition-metal alloys with theory-infused deep learning. Preprint at <http://arxiv.org/abs/2506.03031> (2025).
203. Gouvêa, R. A. et al. Combining feature-based approaches with graph neural networks and symbolic regression for synergistic performance and interpretability. *npj Comput. Mater.* **12**, 67 (2026).
204. Signori-lamin, G. et al. Enzyme-assisted high-consistency fiber refining: enhancing cellulose materials performance in the paper industry through process and physics-informed machine learning modeling. *Cellulose* (2026) <https://doi.org/10.1007/s10570-025-06921-0> (2026).
205. Marquez, R. et al. Lignin nanoparticles as pickering stabilizers: emulsion engineering through physicochemical design and meta-analysis. *Adv. Colloid Interface Sci.* **349**, 103752 (2026).
206. Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S. & Stodtmann, S. Practical guide to <scp>SHAP </scp> analysis: explaining supervised machine learning model predictions in drug development. *Clin. Transl. Sci.* **17**, e70056 (2024).
207. Wilfong, B. et al. Ternary materials discovery using human-in-the-loop generative machine learning. *RSC Adv.* **15**, 19126–19131 (2025).
208. Han, X.-Q. et al. AI-driven inverse design of materials: past, present and future. *Chinese Physics Letters* **42**, 027403 (2025).
209. Fuhr, A. S. & Sumpter, B. G. Deep generative models for materials discovery and machine learning-accelerated innovation. *Front. Mater.* **9**, 865270 (2022).
210. Menon, D. & Ranganathan, R. A generative approach to materials discovery, design, and optimization. *ACS Omega* **7**, 25958–25973 (2022).
211. Jiang, Y., Li, J., Yang, X. & Yuan, R. Applications of generative adversarial networks in materials science. *Mater. Genome Eng. Adv.* **2**, e30 (2024).
212. Marquez, R. et al. Surfactants produced from carbohydrate derivatives: part 2. A review on the value chain, synthesis, and the potential role of artificial intelligence within the biorefinery concept. *J. Surfactants Deterg.* **28**, 25–76 (2025).
213. Hong, H., Kim, S., Lee, J. & Kim, S. S. A comprehensive review of artificial intelligence applications in composite materials: Predictive, generative, and automation approaches. *Int. J. AI Mater. Des.* **2**, 1 (2025).
214. Alverson, M. et al. Generative adversarial networks and diffusion models in material discovery. *Digit. Discov.* **3**, 62–80 (2024).
215. Yang, S. et al. Scalable Diffusion for Materials Generation. In *Proc. Int. Conf. Learn. Represent.* (ICLR, 2024).
216. Veillon, E., Klipfel, A., Sayede, A. & Bouraoui, Z. A chemically grounded evaluation framework for generative models in materials discovery. Preprint at <http://arxiv.org/abs/2601.00886> (2025).
217. Yang, K. & Schwalbe-Koda, D. A generative diffusion model for amorphous materials. *NPJ Comput. Mater.* **12**, 29 (2025).
218. Li, Z. et al. Generative design of crystal structures by point cloud representations and diffusion model. *iScience* **28**, 111659 (2025).
219. Leng, J. et al. Diffusion model-driven smart design and manufacturing: prospects and challenges. *J. Manuf. Syst.* **82**, 561–577 (2025).
220. Priyakumar, U. D. Towards self-driving/autonomous material discovery lab. *J. Chem. Sci.* **137**, 82 (2025).
221. Zhou, L. et al. Toward greater autonomy in materials discovery agents: unifying planning, physics, and scientists. Preprint at <http://arxiv.org/abs/2506.05616> (2025).
222. Badrinarayanan, S., Magar, R., Antony, A., Meda, R. S. & Barati Farimani, A. MOFGPT: generative design of metal-organic frameworks using language models. *J. Chem. Inf. Model.* **65**, 9049–9060 (2025).
223. Takahara, I., Mizoguchi, T. & Liu, B. Accelerated inorganic materials design with generative AI agents. *Cell Rep. Phys. Sci.* **6**, 103019 (2025).
224. Inizan, T. J. et al. System of agentic AI for the discovery of metal-organic frameworks. Preprint at <http://arxiv.org/abs/2504.14110> (2025).
225. Miret, S. & Krishnan, N. M. A. Enabling large language models for real-world materials discovery. *Nat. Mach. Intell.* **7**, 991–998 (2025).
226. Jiang, X. et al. Applications of natural language processing and large language models in materials discovery. *NPJ Comput. Mater.* **11**, 79 (2025).
227. Zhang, J., Chen, X., Ye, X., Yang, Y. & Ai, B. Large language model in materials science: roles, challenges, and strategic outlook. *Adv. Intelligent Discov.* **2025**, 202500085 (2025).
228. Liu, Y. et al. Generative artificial intelligence and its applications in materials science: current situation and future perspectives. *Journal of Materiomics* **9**, 798–816 (2023).
229. Chen, Y. et al. MatterGPT: a generative transformer for multi-property inverse design of solid-state materials. Preprint at <http://arxiv.org/abs/2408.07608> (2024).
230. Hu, J., Li, Q. & Fu, N. Generative AI for materials discovery: design without understanding. *Engineering* **39**, 13–17 (2024).
231. Niyongabo Rubungo, A., Li, K., Hatrick-Simpers, J. & Bousso Dieng, A. LLM4Mat-bench: benchmarking large language models for materials property prediction. *Mach. Learn. Sci. Technol.* **6**, 020501 (2025).
232. Huang, H., Magar, R., Xu, C. & Farimani, A. B. materials informatics transformer: a language model for interpretable materials properties prediction. Preprint at <http://arxiv.org/abs/2308.16259> (2023).
233. Wang, Y. et al. Exploring lightweight language models for materials informatics with AlchemBERT. *Cell Rep. Phys. Sci.* **6**, 102724 (2025).
234. Niyongabo Rubungo, A., Arnold, C., Rand, B. P. & Dieng, A. B. LLM-Prop: predicting the properties of crystalline materials using large language models. *NPJ Comput. Mater.* **11**, 186 (2025).
235. Ferreira, J. J. et al. Human-AI co-creation approach to find forever chemicals replacements. Preprint at <http://arxiv.org/abs/2304.05389> (2023).
236. Amazon SageMaker. <https://aws.amazon.com/sagemaker/> [Accessed 3 December 2024].
237. AI Infrastructure ML and DL Model Training Google Cloud. <https://cloud.google.com/ai-infrastructure> [Accessed 3 December 2024].
238. Azure AI Platform Cloud AI Platform Microsoft Azure. <https://azure.microsoft.com/en-us/solutions/ai> [Accessed 3 December 2024].

239. GitHub · Build and ship software on a single, collaborative platform · GitHub. <https://github.com/> [Accessed 4 December 2024].
240. Intel Labs. <https://www.intel.com/content/www/us/en/research/overview.html> [Accessed 4 December 2024].
241. Choudhary, K. SLMat: a comprehensive serverless toolkit for advanced materials design. Preprint at <https://chemrxiv.org/doi/full/10.26434/chemrxiv-2024-fqq27> (2024).
242. Materials Virtual Lab – advancing materials science through AI. <https://materialsvirtuallab.org/> [Accessed 4 December 2024].
243. Liang, Z., Liang, Z., Zheng, Y., Liang, B. & Zheng, L. Data analysis and visualization platform design for batteries using flask-based python web service. *World Electric Vehicle J.* **12**, 187 (2021).
244. Mishra, A. LatticeML: a data-driven application for predicting the effective Young Modulus of high temperature graph based architected materials. *Int. J. Interact. Des. Manuf.* **19**, 3899–3920 (2025).
245. Nápoles-Duarte, J. M. et al. Stmol: a component for building interactive molecular visualizations within streamlit web-applications. *Front. Mol. Biosci.* **9**, 2022 (2022).
246. Du, Y. et al. M²Hub: unlocking the potential of machine learning for materials discovery. In *Proc. Neural Inf. Process. Syst.* Vol. 36 (2023).
247. Yang, X., Wilson, S. & Petzold, L. MatKB: semantic search for polycrystalline materials synthesis procedures. Preprint at <http://arxiv.org/abs/2302.05597> (2023).
248. Data Apps for Production Plotly. <https://plotly.com/> [Accessed 4 December 2024].
249. Tadmor, E. B., Elliott, R. S., Sethna, J. P., Miller, R. E. & Becker, C. A. The potential of atomistic simulations and the knowledgebase of interatomic models. *JOM* **63**, 17–17 (2011).
250. BigScience Workshop. BLOOM: a 176B-parameter open-access multilingual language model. Preprint at <https://huggingface.co/bigscience/bloom> (2022).
251. Gao, L. et al. The pile: an 800GB dataset of diverse text for language modeling. Preprint at <http://arxiv.org/abs/2101.00027> (2020).
252. Common Crawl – Open Repository of Web Crawl Data. <https://commoncrawl.org/> [Accessed 8 January 2025].
253. Touvron, H. et al. LLaMA: open and efficient foundation language models. Preprint at <http://arxiv.org/abs/2302.13971> (2023).
254. Gemini Developer API Gemma open models Google AI for Developers. <https://ai.google.dev> [Accessed 13 January 2025].
255. New York Times Sues OpenAI and Microsoft Over Use of Copyrighted Work - The New York Times. <https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html> [Accessed 13 January 2025].
256. Samuelson, P. Generative AI meets copyright: Ongoing lawsuits could affect everyone who uses generative AI. *Science* **381**, 158–161 (2023).
257. IBM Watson. <https://www.ibm.com/watson> [Accessed 13 January 2025].
258. Meet Willow, our state-of-the-art quantum chip. <https://blog.google/technology/research/google-willow-quantum-chip/> [Accessed 14 January 2025].
259. National Academies of Sciences, Engineering, and Medicine. *Advancing Chemistry and Quantum Information Science: An Assessment of Research Opportunities at the Interface of Chemistry and Quantum Information Science in the United States* (National Academies Press, 2023).
260. Peruzzo, A. et al. A variational eigenvalue solver on a photonic quantum processor. *Nat. Commun.* **5**, 4213 (2014).
261. Kenson Wesley, R. & Reena Monica, P. A quantum computing-driven aid for new material design. Preprint at <http://arxiv.org/abs/2304.07879> (2023).
262. Barroca, M. A., Neumann Barros Ferreira, R. & Steiner, M. Exploring qubit-ADAPT-VQE for materials discovery in direct air capture. *APL Quantum* **1**, 046103 (2024).
263. Grover, L. K. A fast quantum mechanical algorithm for database search. In *Proc. 28th Annual ACM Symposium on Theory of Computing* 212–219 (ACM, 1996).
264. Nguyen, N. et al. Quantum computing for corrosion simulation: workflow and resource analysis. *npj Quantum Inf.* **12**, 27 (2026).
265. Bergholm, V. et al. PennyLane: automatic differentiation of hybrid quantum-classical computations. Preprint at <http://arxiv.org/abs/1811.04968> (2022).
266. TensorFlow Quantum. <https://www.tensorflow.org/quantum> [Accessed 22 December 2024].
267. Sun, Q. et al. PySCF: the Python-based simulations of chemistry framework. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **8**, e1340 (2018).
268. NetworkX — NetworkX documentation. <https://networkx.org/> [Accessed 22 December 2024].
269. D-Wave Ocean Software Documentation — Ocean Documentation 8.1.0 documentation. <https://docs.ocean.dwavesys.com/en/stable/> [Accessed 22 December 2024].
270. Cloud Quantum Computing Service - Amazon Braket - AWS. <https://aws.amazon.com/braket/> [Accessed 22 December 2024].
271. Rigetti QCS. <https://qcs.rigetti.com/sdk-downloads> [Accessed 22 December 2024].
272. Biamonte, J. et al. Quantum machine learning. *Nature* **549**, 195–202 (2017).
273. Altaisky, M. V., Kaputkina, N. E. & Krylov, V. A. Quantum neural networks: current status and prospects for development. *Phys. Part. Nuclei* **45**, 1013–1032 (2014).
274. Beaudoin, C., Kundu, S., Topaloglu, R. O. & Ghosh, S. Quantum machine learning for material synthesis and hardware security. In *Proc. 41st IEEE/ACM Int. Conf. Comput.-Aided Des. (ICCAD, 2022)*.
275. Lowe, D. M. Extraction of chemical structures and reactions from the literature. (Apollo - University of Cambridge Repository, 2012). <https://doi.org/10.17863/CAM.16293>.
276. Vedavyasa, K. V & Kumar, A. Classification analysis of transition metal compounds using quantum machine learning. *Adv. Quantum Technol.* **8**, 2400192 (2025).
277. Tang, L., Leung, P., Xu, Q. & Flox, C. Machine learning orchestrating the materials discovery and performance optimization of redox flow battery. *ChemElectroChem* **11**, e202400024 (2024).
278. Lourenço, M. P. et al. Exploring quantum active learning for materials design and discovery. *Adv. Quantum Technol.* **7**, 2300268 (2024).
279. Shiota, T., Ishihara, K. & Mizukami, W. Universal neural network potentials as descriptors: towards scalable chemical property prediction using quantum and classical computers. *Digit. Discov.* **3**, 1714–1728 (2024).
280. Guo, Z., Li, R., He, X., Guo, J. & Ju, S. Harnessing quantum power: revolutionizing materials design through advanced quantum computation. *M.* **2**, e73 (2024).
281. Wang, R., Xu, C. & Zhang, X. Toward materials genome big-data: a blockchain-based secure storage and efficient retrieval method. *IEEE Trans. Parallel Distrib. Syst.* **35**, 1630–1643 (2024).
282. Muzammal, M., Qu, Q., Nasrulin, B. & Skovsgaard, A. A blockchain database application platform. Preprint at <http://arxiv.org/abs/1808.05199> (2019).
283. O'Mara, J., Meredig, B. & Michel, K. Materials data infrastructure: a case study of the citrination platform to examine data import, storage, and access. *JOM* **68**, 2031–2034 (2016).
284. Patel, O. Blockchain indexing strategies: enhancing performance in software applications. *J. Eng. Appl. Sci. Technol.* **2**, 1–10 (2020).
285. Coronado Mondragon, A., Coronado Mondragon, C. E. & Coronado, E. S. Exploring the applicability of blockchain technology to enhance manufacturing supply chains in the composite materials industry. In *Proc. 2018 IEEE International Conference on Applied System Innovation (ICASI)* 1300–1303 (IEEE, 2018).

286. Yue, L., Junqin, H., Shengzhi, Q. & Ruijin, W. Big data model of security sharing based on blockchain. In *Proc. 3rd International Conference on Big Data Computing and Communications (BIGCOM)* 117–121 (IEEE, 2017).
287. Ecemis, I. N. et al. Exploring blockchain for nuclear material tracking: a scoping review and innovative model proposal. *Energies* **17**, 3028 (2024).
288. Wang, R. et al. A secured big-data sharing platform for materials genome engineering: state-of-the-art, challenges and architecture. *Future Gener. Comput. Syst.* **142**, 59–74 (2023).
289. Wilson, S. et al. Blockchain-enabled provenance tracking for sustainable material reuse in construction supply chains. *Future Internet* **16**, 135 (2024).
290. Wang, C., Su, H., Duan, L. & Li, H. InterMat: a blockchain-based materials data discovery and sharing infrastructure. *Processes* **11**, 3168 (2023).
291. Karastoyanova, D. & Stage, L. Towards collaborative and reproducible scientific experiments on blockchain. In *Lecture Notes in Business Information Processing*. **316**, 144–149 (Springer-Verlag, 2018).
292. Coelho, R. et al. A blockchain-based architecture for trust in collaborative scientific experimentation. *J. Grid Comput.* **20**, 19 (2022).
293. Alniamy, A. M. & Liu, H. Blockchain-based secure collaboration platform for sharing and accessing scientific research data. In *2020 3rd International Conference on Hot Information-Centric Networking (HotICN)* 34–40 (IEEE, 2020). <https://doi.org/10.1109/HotICN50779.2020.9350856>.
294. Dima, A. et al. Informatics infrastructure for the materials genome initiative. *JOM* **68**, 2053–2064 (2016).
295. Rajan, S. N. et al. A blockchain based framework for opening up big data in collaborative edges with less resource consumption. In *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* 2818–2822 (IEEE, 2023).
296. Gong, H. et al. A repository for the publication and sharing of heterogeneous materials data. *Sci. Data* **9**, 787 (2022).
297. LLeible, S., Schlager, S., Schubotz, M. & Gipp, B. A review on blockchain technology and blockchain projects fostering open science. *Front. Blockchain* **2**, 16 (2019).
298. Salah, K., Rehman, M. H. U., Nizamuddin, N. & Al-Fuqaha, A. Blockchain for AI: review and open research challenges. *IEEE Access* **7**, 10127–10149 (2019).
299. Leng, J. et al. Makerchain: a blockchain with chemical signature for self-organizing process in social manufacturing. *J. Clean. Prod.* **234**, 767–778 (2019).
300. Evans, M. L. et al. ‘optimade-python-tools’: a Python library for serving and consuming materials data via OPTIMADE APIs. *J. Open Source Softw.* **6**, 3458 (2021).
301. THE 17 GOALS | Sustainable development. <https://sdgs.un.org/goals> (2024).
302. Motadayen, M., Devabharathi, N. & Agarwala, S. Advancing sustainability: biodegradable electronics and materials discovery through artificial intelligence. *Int. J. AI Mater. Des.* **1**, 1 (2024).
303. Ali, A. H. & Green, A. AI for sustainability: leveraging machine learning to drive a circular economy. *Babylonian J. Artif. Intell.* **2023**, 15–16 (2023).
304. Chen, H., Zheng, Y., Li, J., Li, L. & Wang, X. AI for nanomaterials development in clean energy and carbon capture, utilization and storage (CCUS). *ACS Nano* **17**, 9763–9792 (2023).
305. Salem, K. S. et al. A critical review of existing and emerging technologies and systems to optimize solid waste management for feedstocks and energy conversion. *Matter* **6**, 3348–3377 (2023).
306. Kaack, L. H. et al. Aligning artificial intelligence with climate change mitigation. *Nat. Clim. Chang.* **12**, 518–527 (2022).
307. Luccioni, A. S., Viguier, S. & Ligozat, A.-L. Estimating the carbon footprint of BLOOM, a 176B parameter language model. *J. Mach. Learn. Res.* **24**, 11990–12004 (2023).
308. Everman, B. et al. Evaluating the carbon impact of large language models at the inference stage. In *2023 IEEE Int. Perform. Comput. Commun. Conf. (IPCCC)* 150–157 (IEEE, 2023).
309. Strubell, E., Ganesh, A. & McCallum, A. Energy and policy considerations for deep learning in NLP. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics* 3645–3650 (2019).
310. Schwartz, R., Dodge, J., Smith, N. A. & Etzioni, O. Green AI. *Commun. ACM* **63**, 54–63 (2020).
311. Pernici, B. et al. A methodology for assessing the environmental effects induced by ICT services. In *Proc. 7th International Conference on ICT for Sustainability* 36–45 (ACM, 2020).
312. Itten, R. et al. Digital transformation—life cycle assessment of digital services, multifunctional devices and cloud computing. *Int. J. Life Cycle Assess.* **25**, 2093–2098 (2020).
313. Malmodin, J. & Lundén, D. The energy and carbon footprint of the global ICT and E&M Sectors 2010–2015. *Sustainability* **10**, 3027 (2018).
314. CodeCarbon.io. <https://codecarbon.io/#howitwork> (2024).
315. Anthony, L. F. W., Kanding, B. & Selvan, R. Carbontracker: tracking and predicting the carbon footprint of training deep learning models. Preprint at <http://arxiv.org/abs/2007.03051> (2020).
316. Budenny, S. A. et al. eco2AI: carbon emissions tracking of machine learning models as the first step towards sustainable AI. *Dokl. Math.* **106**, S118–S128 (2022).
317. Irwin, J. J. et al. ZINC20 - a free ultralarge-scale chemical database for ligand discovery. *J. Chem. Inf. Model.* **60**, 6065–6073 (2020).
318. ChemSpider <https://www.chemspider.com/> (2024).
319. Cambridge Crystallographic Data Centre. *The Cambridge Structural Database*. <https://www.ccdc.cam.ac.uk/solutions/software/csd/> (2024).
320. Inorganic Crystal Structure Database (ICSD) <https://icsd.products.fiz-karlsruhe.de/> (2024).
321. RCSB Protein Data Bank <https://www.rcsb.org/> (2024).
322. Case, D. A. et al. AmberTools. *J. Chem. Inf. Model.* **63**, 6183–6191 (2023).
323. Gonze, X. et al. The Abinit project: impact, environment and recent developments. *Comput. Phys. Commun.* **248**, 107042 (2020).
324. Blaha, P. et al. *WIEN2k: an augmented plane wave plus local orbitals program for calculating crystal properties. calculating crystal properties*, vol. 2 (Techn. Universität, 2019).
325. Frisch, M. J. et al. *Gaussian 16 Revision C.01*. (Gaussian, Inc., 2016).
326. Neese, F. & Wiley, J. The ORCA program system. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2**, 73–78 (2012).
327. Shao, Y. et al. Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Mol. Phys.* **113**, 184–215 (2015).
328. Barca, G. M. J. et al. Recent developments in the general atomic and molecular electronic structure system. *J. Chem. Phys.* **152**, 154102 (2020).
329. Aprà, E. et al. NWChem: past, present, and future. *J. Chem. Phys.* **152**, 42 (2020).
330. Kühne, T. D. et al. CP2K: an electronic structure and molecular dynamics software package—quickstep: efficient and accurate electronic structure calculations. *J. Chem. Phys.* **152**, 34 (2020).
331. *Scrapy 2.11 Documentation*. <https://docs.scrapy.org/en/latest/> (2024).
332. *WebHarvy Web Scraping Software*. <https://www.webharvy.com/index.html> (2024).
333. Hermann, M. *Helium: Lighter web automation with Python*. <https://github.com/mherrmann/helium> (2024).
334. *MechanicalSoup 1.3.0 documentation*. <https://mechanicalsoup.readthedocs.io/en/stable/> (2024).

Acknowledgements

This material is based upon work supported by the U.S. Department of Energy's Office of Critical Minerals and Energy Innovation (CMEI) under the Bioenergy Technologies Office (BETO) Award Number DE-EE0009669. Disclaimer: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States government nor any agency thereof, nor any of its employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness. We also gratefully acknowledge IBM for the Smart Sustainability Award that supported this work.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Lokendra Pal.

Peer review information *Communications Materials* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026