scientific data

Check for updates

OPEN

COMMENT

Unleashing the power of AI in science-key considerations for materials data preparation

Yongchao Lu[1], Hong Wang[1] ✉, Lanting Zhang[1] ✉, Ning Yu[1], Siqi Shi[2] & Hang Su[3]

The release of ChatGPT has triggered global attention on artificial intelligence (AI), and AI for science is thus becoming a hot topic in the scientific community. When we think about unleashing the power of AI to accelerate scientific research, the question coming to our mind first is whether there is a continuous supply of highly available data at a sufficiently large scale.

The urgent demand for data from AI for materials science

AI is greatly accelerating the pace of human understanding and changing the world. The release of ChatGPT (https://chat.openai.com/chat), a system based on large language model that can engage in meaningful conversations, generate creative text, and assist in completing various complex tasks, has undoubtedly demonstrated enormous potential of AI in supporting and promoting human social progress. A typical capability of AI is that it can establish the correlation between the features of the research object directly through data without relying on a priori knowledge, which provides a new perspective for scientific research under the condition of lack of usable knowledge models. As such, AI is expected to find many applications in scientific field to study the unknown. For example, AlphaFold[1] (https://www.deepmind.com/research/highlighted-research/alphafold), a representative application of AI for science developed by DeepMind, leverages deep learning and the evolutionary information contained within multiple sequence alignments (MSA) to predict the 3D structure of proteins. It uses the transformer architecture's attention mechanism to understand the spatial relationships between amino acids. After training on a large dataset of about 100,000 known protein sequences and structures from the Protein Data Bank (PDB), AlphaFold is capable of predicting the structures of previously unknown proteins with an accuracy comparable to experimental methods. Moreover, AlphaFold significantly reduces the time required for structure determination, condensing what traditionally takes months to years into a matter of minutes to hours. This case also shows us the disruptive acceleration opportunities that large-scale scientific data can bring to scientific research with the support of AI technology. To fully leverage this advantage, some autonomous experimental platforms have been developed[2–4], which integrate robot technology, databases, and AI technology to integrate the generation, management, mining, and verification of scientific data. As an example, the A-Lab, an autonomous laboratory for the solid-state synthesis of inorganic powders, which uses computations, historical data from the literature, machine learning (ML) and active learning to plan and interpret the outcomes of experiments performed using robotics. The work[4] reported that over 17 days of continuous operation, the A-Lab realized 41 novel compounds from a set of 58 targets including a variety of oxides and phosphates that were identified using large-scale ab initio phase-stability data from the Materials Project and Google DeepMind. Although the novelty of these successfully synthesized compounds remains a topic of debate from the perspective of solid-state chemistry[5], A-Lab continues to demonstrate disruptive acceleration in materials exploration. Further efforts focused on enhancing the reliability of experimental results (such as comprehensive and detailed characterization) and promoting autonomy throughout the entire research process (from synthesis to characterization to analysis) will progressively enhance the reliability and efficiency of this novel research facility, thus fully exploiting the benefits of data-driven approaches.

Sufficient, high-quality domain-specific data is the key to AI's participation in scientific research. GPT-3.5/GPT-4 is a versatile large language model (LLM) trained on extensive internet text data, which includes a certain proportion of materials science information. Based on this model, ChatGPT is capable of addressing

[1]Materials Genome Initiative Center & School of Materials Science and Engineering, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai, 200240, China. [2]Materials Genome Institute, Shanghai University, Shanghai, 200444, China. [3]Material Digital R&D Center, China Iron and Steel Research Institute Group, Beijing, 100081, China. ✉e-mail: hongwang2@sjtu.edu.cn; lantingzh@sjtu.edu.cn

relevant questions in the field of materials science. These include explaining material science concepts, information retrieval and summarization, research proposal design, and writing computational code. To evaluate the performance of LLMs in solving material-related problems, Mohd Zaki[6] and colleagues curated a dataset called MaScQA. This dataset comprises 650 challenging materials science questions from the Graduate Aptitude Test in Engineering(GATE), India's national examination for graduate admission. Correctly answering these questions requires knowledge and skills equivalent to an undergraduate degree in materials science. The results revealed that the top-performing GPT-4 model achieved an accuracy rate of 62%. However, in several core materials science areas, the model exhibited subpar performance. For instance, it struggled with concepts related to atomic and crystalline structures of materials, as well as their electrical, magnetic, and thermodynamic behaviors. Conceptual errors were identified as the primary factor contributing to the decline in performance (approximately 72%). Interestingly, even with prompts to guide its thinking, the accuracy did not significantly improve. Similar to other machine learning models, the output of GPT-4 is consistent with the patterns observed in the network data used to train it[7]. For materials science research, which requires a high degree of rigour and precision, this result suggests that the quality and scale of current materials science data for training GPT-4 is clearly insufficient to support the free and reliable exploration requirements of materials researchers. Compared with Internet data from a wide range of sources, domain-specific data, as exemplified by materials science, is more specialized and more expensive to obtain. It is generally generated, collated and used by personnel with specialized knowledge in specific fields, which to a certain extent leads to the scarcity of the domain data.

Since the announcement of the Materials Genome Initiative (MGI, https://mgi.gov/), the development of high-throughput experiments and computational technologies (such as combinatorial materials chips and first-principles calculations) has deliberately accelerated the rate of materials data generation. However, this centralized and costly effort can only cover a fraction of the materials research field[8]. Although a series of databases have been established in the materials field[9], such as the Materials Project[10] (https://materialsproject.org/), AFlow[11] (http://aflowlib.org/), OQMD[12] (https://www.oqmd.org/), NOMAD Laboratory[13] (https://nomad-lab.eu/), HTEM[14] (https://htem.nrel.gov/), researchers are still unable to easily obtain any required data at will[9]. Due to the limitations of the "cottage industry" research mode and the publication mode centered on scientific research papers, most of the materials data generated by traditional methods are still in a distributed and individualized state. Researchers confront many obstacles in querying, obtaining, integrating, and reusing these data. Consequently, building large, highly applicable materials datasets is very difficult.

The existing material science data ecosystem faces challenges in supporting the widespread and reliable application of AI technology within this field. Urgently, an AI-ready material science data ecosystem must be established, guided by the convenient construction of reliable AI models. This endeavor will fully unleash the disruptive acceleration potential of AI in materials science research. However, the transformation of scientific data governance toward AI involves stakeholders across various stages of the data lifecycle. These stakeholders include researchers, funding institutions, publishing bodies, research equipment suppliers, data management platforms, and standardization organizations. Their sustained attention and collaborative efforts are essential for achieving this transformation.

The FAIR principles[15], from a user perspective, have laid critical groundwork for constructing an AI-ready scientific data ecosystem. These principles provide governance directions and fundamental requirements for data sharing and reuse in scientific endeavors. However, these principles still exhibit partiality when it comes to meeting the specific needs of reliable AI. In light of this, this commentary aims to propose a systematic data governance framework tailored to AI for the materials science research community. Within this framework, we outline the core common issues that need addressing to meet these requirements and offer suggestive solutions. Based on this collaborative working framework that aligns with the common interests of the research community, we call upon all stakeholders involved in materials science data—researchers, institutions, and organizations—to actively participate and accelerate the creation of an AI-ready materials science data ecosystem, thereby fully unlocking the potential of AI to accelerate materials science research.

In terms of content organization, this article follows the materials data lifecycle and divides the discussion into two distinct parts based on the entities responsible for data governance: the individual data sample generation & management, and the datasets construction. Finally, these components are synthesized to form a holistic perspective on an AI-ready materials data governance system. It is essential to emphasize that the focus of this article lies in providing a comprehensive reference for data governance requirements and recommendations tailored to the materials research community, rather than the narrow scope of traditional machine learning data preparation tasks such as data cleaning and preprocessing.

The requirements of data samples for AI for materials science

Accuracy of data generation and quality control. Data serves as the medium through which researchers explore and understand the physical world. The accuracy of scientific data is crucial for ensuring the reliability of research. In the traditional research paradigm, scientists formulate hypotheses based on known or empirical physical models. They then generate scientific data using experimental or computational tools, analyze and interpret the data, and further refine existing knowledge models. However, AI-driven scientific research diverges from this approach. Instead of relying solely on established empirical models, AI has the potential to leverage its advantages in high-dimensional analysis. By exploring the possibility of constructing new knowledge models from the intrinsic correlations within scientific data, AI may potentially accelerate research, bypassing the slower development associated with traditional physical models. Nevertheless, the reliability of AI knowledge models built entirely on scientific data hinges on the accurate representation of that data. Additionally, accurate data provides a crucial entry point for interpretable research in later stages of AI model development.

The accuracy of materials data is closely related to its production stage, including measurement tools, conditions, operator operations, data processing, and evaluation methods. Due to the complexity of scientific data production and influencing factors, it is difficult to develop general data output specifications to effectively guarantee data accuracy. In the current landscape of data governance, researchers' comprehensive control over relevant variables, coupled with domain expertise, assumes paramount importance in evaluating and verifying data reliability and repeatability.

Here we discuss a mechanism for ensuring the accuracy of materials data from a macro perspective. Each piece of materials data, generated under rigorous scientific conditions, is a phenotypic reflection of the material's intrinsic properties and serves as a valuable unit of identification that allows AI models to fit the real physical world. In this data-driven model of research, each piece of accurate data has its own academic value and exists as a coordinate point on a map of the material's intrinsic laws, whether or not it aligns with the researcher's personal research goals. We call for a data-centric academic publishing ecosystem that includes peer-reviewed scientific data accuracy assessment and publishing mechanisms, such as Scientific Data, an innovative journal in this regard. In this mode, each piece of accurate scientific data will be recognized for its academic value and established as an important evaluation factor for scientific contributions. The evaluation will focus on the objectivity and unknowns of the data, motivating researchers to produce and share more high-quality data.

The discourse surrounding data publication has been ongoing within research communities[16–18]. A survey[16] conducted among 250 researchers in scientific and social science fields focused on data publication and peer review revealed several key findings. Researchers expressed a desire for data to be disseminated through databases or repositories to enhance accessibility. Few respondents expected published data to be peer-reviewed, but peer-reviewed data enjoyed much greater trust and prestige. Adequate metadata was recognized as crucial, with nearly all respondents expecting peer review to include an evaluation of data files. Citation and download counts were deemed important indicators of data impact. These survey results can inform publishers in constructing data publication formats that meet community expectations. However, it is essential not to overlook the challenges associated with data peer review. Although most researchers consider peer review critical for ensuring data quality, evaluating data quality extensively remains a complex issue. Challenges related to data peer review include:

- Resource intensiveness: Researchers must invest additional effort to meet standardized data publication requirements, which may yield limited tangible benefits.
- Specialized knowledge: Unlike traditional paper reviews, defining criteria for data review is complex and requires specialized domain knowledge. Additionally, the scale of data may exceed reviewers' capacity.
- Heterogeneous material data: Material data, primarily collected and managed by individual researchers, exhibit varying expression formats and quality, complicating understanding and comparison during review.
- Verification complexity: Considering time and cost, data validation poses difficulties.

As discussed earlier, assessing the quality of material data for AI applications should prioritize both the objectivity and novelty of the data. Furthermore, scalable and maintainable data peer review is closely tied to various aspects of the data ecosystem. Suggested combinations of transformation actions to address these challenges include:

- Establishing consensus-based data standards: Engage stakeholders, including material research equipment and software providers, researchers, publishers, and data repositories, to adopt unified data collection, management, transmission, review, and storage formats. This consensus on data expression formats will facilitate understanding, evaluation, comparison, and integration of data.
- Integrating anti-tampering technologies: Incorporate tamper-proof technologies (e.g., blockchain hash values) into standardized data formats. Equipment or software generating data should create tamper-proof markers added to the original data in standardized formats to support the assessment of data authenticity.
- Maintaining researcher practices while enhancing data publication: Researchers can continue using raw data to support scientific discoveries for traditional manuscript publications. Simultaneously, they could also publish complete, standardized raw data and use metrics like publication volume, citation rates, and download rates as academic impact indicators. This approach encourages researchers to publish data, fostering a data publication community and providing a fairer evaluation of contributions, moving away from extreme winner-takes-all academic norms.
- Using automated programs for peer review: Publishing institutions should use data repositories as the conduit for data reception, review, and publication. Automated data screening techniques should pre-assess data for compliance with format requirements and authenticity markers. Compare existing data in the database to determine whether the data is novel.
- Developing visual data review processes: Integrate automated program-identified data into similar clusters, assessing whether data reflect object characteristics and align with population trends. Specialized data requiring manual review can be efficiently handled. Experienced reviewers can then decide whether to approve publication.

Integrity of data collection and maintenance. The demand for large data sets by AI and the scarcity of scientific data make it increasingly urgent for researchers to use multi-source data. Only by fully capturing information such as the background, conditions, and results of research data into a self-explanatory unit of meaning can data flow freely and independently in the field, and be repeatedly accessed, understood, and correctly used by different researchers. At the same time, these rich details of data production will also provide fine-grained guidance for AI models in the verification and application of the physical world.

However, at the data collection stage, frontline materials researchers often collect only the data fragments that they are interested in, e.g., only images of the results of a particular characterization are collected, without recording the context and conditions under which they were generated. Such data can only be understood by those who generated it based on their previous research, and other users cannot properly understand and reuse it. It also affects the proper understanding of the data selected to build datasets and the accurate reproduction of models in reality.

The evolving demands of data collection and description reflect the expanded scope of data utilization and value pathways in the context of data-driven research. As previously mentioned, each data sample represents a discrete reflection of the target group under study. These samples not only serve the short-term, specific research needs of data producers but can also be repeatedly accessed and integrated by different researchers into various scales of target data clusters. Leveraging AI to explore these data clusters allows for insights into the inherent properties of the research subjects.

However, achieving the convenient and widespread implementation of the latter approach requires a shift in mindset among all data stakeholders. They must adjust data governance practices at their respective stages to facilitate data flow and utilization. For instance, modifying data collection descriptions ensures data completeness, enabling long-term reusability beyond merely burdening researchers with new data governance tasks within the traditional research paradigm.

To address the data integrity requirements of AI-driven research, suggested portfolio reform measures include:

- Funding agencies: When defining project metrics and allocating funds, funding agencies should guide applicants in understanding the long-term, macroscopic value of data within the data-driven context. Establishing assessment criteria for data collection methods and metadata descriptions ensures sustained support for data management across projects.
- Research equipment and software providers: In response to researchers' short-term and long-term data usage patterns, equipment and software providers should adapt their data collection and export services. Presenting detailed data production context (including device, environment, and technical conditions) to users and opening data collection interfaces for automated programmatic[19–23] access can reduce the burden of data integrity descriptions during data processing.
- Researchers: Embed relevant metadata (such as researcher information, study subjects, and research objectives) into collected data. When using data, researchers can selectively analyze areas of interest. During data storage, transmission, and publication, maintaining the original data format with comprehensive metadata descriptions ensures integrity.
- Data repositories and publishing institutions: Establish metadata description workflows that promote data integrity[24]. Encourage the inclusion of original data with complete metadata descriptions when accepting data for publication.

Consistency of data storage and representation. AI can process structured data, unstructured data, and semi-structured data simultaneously. However, in specific model construction, data must conform to certain formats and organizational schemes, so that AI can accurately and efficiently compare and identify the relationships and patterns hidden between data. When constructing data sets from multiple sources, the presentation of data content and organizational structure is often inconsistent. It directly affects the efficiency of data integration and the accuracy of AI models. Recent advancements in large language models (LLMs) have opened new avenues for handling heterogeneous data. Currently, mature applications involve using LLMs to assist in literature reading, summarization, and analysis[25]. Researchers are also exploring LLMs for extracting relevant scientific information from vast amounts of unstructured scientific literature[26,27]. However, given that LLMs generate text, the effectiveness of information extraction depends closely on the composition and distribution of training data. Additional verification against the original text is necessary during information extraction, which remains a significant work burden. Moreover, the data extracted by LLMs from unstructured texts are predictive outputs generated from their training data. Unlike traditional experimental or computational data, these LLM-derived data are characterized by an inherent indirectness and predictive nature. When used to train AI applications for scientific research, they may undermine the reliability and interpretability of the resulting models. To enable seamless integration of diverse data sources in AI-driven research, establishing community-consensus data standards and normalizing data representations are essential. This aligns with the FAIR principles, promoting interoperability across human-machine interactions. These standardized data formats will serve as reference benchmarks for data collection, storage, transmission, and integration within the scientific community, supporting training, validation, and optimization of scientific AI models.

However, material data standardization remains in its infancy within the materials domain. Several prominent challenges hinder its implementation:

- Materials research involves various experimental and computational processes, resulting in a multitude of material data types with varying formats. A unified standard covering all material data remains elusive.
- Insufficient funding and research focus on material data standardization hinder progress.
- Material data standardization spans various lifecycle stages, and clarifying responsibility for driving standardization remains essential.

- The adoption and promotion of data standards pose significant data processing burdens due to the sheer volume of material data.

To address these challenges, recommended reform measures include:

- Modular approach: Given the long-tail nature of material data, incremental standardization using modular approaches is advisable. Establishing standard modules for specific material research aspects (e.g., synthesis, test, or data processing) allows gradual progress. Representative examples include the Crystallographic Information Framework (CIF, https://www.iucr.org/resources/cif/) for crystallography and NeXus (https://www.nexusformat.org/) for experimental data in neutron, X-ray, and muon science. Another noteworthy effort is the universal framework developed by the Materials Genomics Committee of the China Society of Testing Materials (CSTM/FC97), which leverages modular construction methods and adheres to FAIR principles. This framework provides practical guidance and case studies for material data standardization(http://www.cstm.com.cn/article/details/390ce11f-41a2-4d01-8544-04012bb13782/).
- Community collaboration: Engage all stakeholders related to material data in forming domain-specific data standardization communities and leading organizations. Notably, the FAIRmat[8](https://www.fairmat-nfdi.eu/fairmat/) initiative could serve as an reference model for community building efforts.
- Tool development: Develop automated data collection tools, storage platforms, and publication systems aligned with data standards to ease adoption and promote effective dissemination.
- Funding support: Funding agencies should recognize the immense value of data standardization for future research and proactively support material data standardization projects.

Findability of data retrieval and indexing. Building large datasets requires rapid querying and retrieval of the necessary data samples. The FAIR Principles clearly define the minimum requirements for data findability, including assigning persistent unique identifiers to (meta)data, describing (meta)data with a variety of accurate and relevant attributes, and registering (meta)data in a searchable resource. Meeting these requirements requires the power of data-sharing platforms and segmented research communities.

Data platforms can integrate identification, registration, and publication for open access to data. However, it is difficult for individual researchers to afford to build and operate a data platform by themselves. Although the materials research community has established many data-sharing platforms, such as the Novel Materials Discovery Laboratory ((NOMAD, https://nomad-lab.eu/) and the Materials Data Facility (MDF, https://www.materialsdatafacility.org/), researchers still face the problem of not knowing where and how to search for data of interest, especially experimental data. This implies that these data platforms have not been seamlessly integrated into material science research communities to effectively promote discoveries within the field.

The integration of data platforms with research communities can draw inspiration from the operational models of academic conferences and journals. In the materials science research community, a range of specialized materials topics (such as steel materials, nanomaterials, and catalytic materials) naturally gives rise to dedicated academic conferences and journals. These platforms reflect certain research trends within the community and indicate a substantial number of researchers supporting these themes. Researchers rely on these conferences and journals to stay updated on the latest advancements, report their findings, and adjust their future research directions. Consequently, these platforms have become organically integrated into the academic field, ensuring sustainable development.

Similarly, data platforms can be designed and maintained to emulate the functions of academic conferences and journals within research communities. By recognizing the academic value of scientific data, defining the data themes to be accepted, and providing a platform for researchers to report novel scientific data and access existing datasets, data platforms can incentivize and acknowledge researchers. This approach encourages the integration of data platforms into research communities, fostering the discovery and exchange of data of interest. Following certain data publication standards or norms, as the data platform operates, specific material systems' datasets will gradually grow in diversity and quantity. This growth will meet researchers' data discovery needs and support the widespread adoption of AI-driven research in the field.

Combined recommendations to promote the integration of data platforms into research communities and establish mechanisms for data discovery and sharing include:

- Forming scientific data communities: Based on current thematic forums or research associations (e.g., semiconductor materials, nanomaterials, biomaterials), establish scientific data communities within each field to discuss data classification and develop data standards.
- Creating data publication journals: Establish new data publication journals supported by thematic journals in the field, and develop corresponding data publication platforms. The scope of data publications would include unpublished specific material data, stored according to field consensus standards.
- Integrating basic functions: Ensure these thematic data platforms include essential functions such as data identification, registration, uploading, querying, and downloading, thus providing a public digital community for data management and communication among researchers.
- Organizing data publication forums: Concurrently organize sub-forums on data publication at periodic academic forums in the field. These sub-forums would focus on reporting novel scientific datasets and expanding the understanding of field-specific data, while also covering AI applications based on published datasets, to attract broad participation from researchers.

- Establishing sustainable business models: Seek to establish sustainable business models that leverage the high-standard data exploration and AI application support provided by the data platform, ensuring long-term maintenance and operation of the platform.

Accessibility of data access and sharing.

Building and processing large datasets requires easy access to data for researchers, and this is a major challenge to achieving large-scale data sharing. According to the accessibility requirement of the FAIR principle, data should be retrievable by their identifier using a standardized communication protocol that is open, free, and universally implementable, allowing for authentication and authorization if necessary. At a minimum, metadata should be accessible even if the data are unavailable.

The technical requirements of the Accessibility Principle have been maturely applied to the construction of data platforms. The integration of these functions into data platforms in the field of subdivided materials can meet the minimum requirements for data management. However, the willingness of data owners to share is the key to solving the accessibility of data, which is the main obstacle to the large-scale construction of domain data sets. As a new and unique factor of production, data has not yet formed a well-defined mechanism in various fields to coordinate the contradictions of ownership, use rights, and revenue rights, making it difficult for domain data sharing to operate systematically. Blockchain methods are a technical choice to ensure data traceability and transactions[28]. However, there are still many inconveniences for researchers to obtain and integrate domain data due to its technical complexity and privacy characteristics.

From the researcher's perspective, the factors that influence their willingness to share scientific data include the following:

- Intellectual property protection: Researchers worry that sharing data may allow others to publish related findings first, potentially diminishing the commercial value and patent potential of the data, which could negatively impact their own competitive standing.
- Lack of incentives: The scientific community has weak reward mechanisms for data sharing, with no relevant incentives in place.
- Insufficient infrastructure support: Researchers often lack the resources to establish platforms for open data access and face a shortage of user-friendly infrastructure to support data sharing.
- Misuse of data: Researchers are concerned that their data might be misused, which may harm their academic reputation.
- Cultural barriers: The culture of data sharing is not yet established, and researchers may be reluctant to share data due to habitual practices.

To address these challenges and encourage researchers to share scientific data, the following measures are suggested:

- Academic recognition: Publishing institutions should recognize the academic value of data on par with traditional papers. Establishing data ownership through identifiers and citations, and providing academic impact incentives such as impact factors and citation metrics for shared data. Additionally, researchers could be rewarded with data usability credits proportional to the amount of data they share, incentivizing further data publication and sharing.
- Infrastructure development: Academic institutions should build data-sharing platforms within specific thematic material research communities. These platforms should integrate data storage, identification, and controlled access, providing researchers with user-friendly infrastructure for data sharing.
- Quality control: Researchers could share tamper-evident original data generated by objective equipment, in alignment with Data Generation and Quality Control recommendations. This approach minimizes the influence of human factors on the data, preserving its objectivity, and reduces the risk of misinterpretation by other users, thereby protecting the reputation of the data providers.
- Establishment of data sharing communities: A viable approach is to model data-sharing communities after the Human Genome Project. This involves classifying and dividing tasks related to specific material systems, followed by internal sharing and utilization of data within the community. A case of accelerating materials research through division of labor and data sharing can be found in the literature[29].
- Funding agency requirements: Funding agencies could mandate data sharing within a defined timeframe and scope for research projects they support, fostering a culture of data openness.

The requirements of datasets for AI for materials science

AI for materials science research requires a continuous supply of high-quality materials data. To achieve this, targeted adjustments in the production, collection, storage, access, and sharing of materials data samples are needed to create an AI-ready data ecosystem. On this basis, the reliability of the AI model will also be directly affected by how researchers select these data to build a training set. In the following sections, we will briefly discuss the preparation requirements at the dataset level, aiming to provide a concise and systematic reference for materials science researchers at various levels to conduct data-driven studies. Although these principles may seem fundamental to seasoned machine learning researchers, they are essential to the reliability of AI models.

Adequacy of the data volume.

AI technology combines statistical principles and requires sufficient data samples to reflect the significance of potential correlation relationships between data features. Certainly, some AI models can effectively control model accuracy by adjusting hyperparameters. For instance, in long short-term memory networks (LSTM), balancing the dimensions of hidden layers helps mitigate overfitting and underfitting issues. However, abundant data can improve a model's performance by providing it with superior generalization ability and robustness. In practice, it is impossible to generalize how much data is enough for AI model training. Instead, it is determined comprehensively based on the model type, feature set, task complexity, data quality, experimental effects, etc. For example, low-parameter models such as logistic regression require relatively little data, while high-parameter models such as deep neural networks may need more data to support high-dimensional analysis. During model training, these factors need to be considered comprehensively to ensure that the sample size can fully cover different types and features. In addition, it is also necessary to consider the cost of data acquisition and comprehensively determine the optimal data scale.

Comprehensiveness of data features. The features embedded in the data set are essentially human empirical observational perspectives on the research object. The more comprehensive these features are, the closer the AI model's analysis of the data is to that of a human. At the same time, rich features can provide more opportunities for data correlation and prevent AI from being limited by human knowledge. For material science research, material research data is typically centered on various aspects of material composition, process, structure, and performance. For example, steel materials include varying elemental contents such as C, Mn, Cr, Ni, Mo, hot working process, microstructure, tensile strength, yield strength, elongation, and other data features unique to such materials. Collecting as many of these expert knowledge features as possible helps AI models build reliable feature relationships from material data. Especially for deep learning models such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer-based models, they are capable of adapting to complex, high-dimensional data, automatically extracting, combining, and correlating high-level features, fully leveraging the advantages of rich data features. It's worth noting, however, that for some traditional models, such as linear regression and support vector machines (SVMs), an excessive number of features may lead to overfitting, redundancy, and noise, increasing the time required for model training and inference. In such cases, feature engineering methods, such as correlation analysis, can be employed to comprehensively analyze various features and determine the optimal feature set.

Diversity of data samples. Data sets should contain as many sample cases as possible under real-world conditions. When there are few types of samples, AI models are prone to overfitting and have low generalization ability in realistic environments. Current scientific research is often conducted under specific research purposes, collecting only positive data that conform to the research purpose, and not capturing non-positive data. As a result, the accumulated scientific research data is limited. Models trained on this data may perform well on specific data samples but also miss the potential to mine the uncollected parts. For example, in a famous case published in 2016, Raccuglia *et al.*[30] used decision tree methods to predict new metal-organic oxide materials, including both "successful" and "failed" experimental data in the training set. Diverse data samples can provide opportunities to study the gradual changes in scientific phenomena and also reveal valuable rules hidden in unexplored samples.

In practical terms, during data collection concerning the studied object, it is recommended to encompass various scenarios, characteristics, and classifications of the target entity as comprehensively as possible, ensuring an ample number of samples within each classification. Under limited conditions and with small sample sizes, techniques like data augmentation and cross-validation can enhance the representation of intrinsic feature relationships.

Uniformity of sample distribution. The sample distribution in the data set should be uniform. Deliberately or unintentionally, when data samples are collected, the difference in the number of data samples of different categories will cause the model's training results to have a certain tendency and may cause bias in AI. For example, Joshua Schrier[31] *et al.* curated a dataset comprising several hundred synthetic conditions employed in the production of vanadium borates. Subsequently, they developed a machine-learning model using this dataset to forecast the outcomes of reactions, determining their success or failure. Surprisingly, the team observed that a model trained on randomly generated reaction conditions outperformed one trained on a dataset generated by human experts in predicting the success or failure of reactions. To ensure the homogeneity of samples, it is recommended to collect as much data as possible for each category, particularly for a few categories. As an auxiliary measure under limited data conditions, active learning can selectively focus on training samples from minority classes, thereby enhancing the model's ability to recognize those specific categories. Additionally, techniques such as data augmentation can generate additional samples, effectively increasing the quantity of minority data points. Furthermore, incorporating a weighted loss function during training can balance the impact across different categories by assigning varying weights to each class.

Professionalism of data annotation. Machine learning encompasses a variety of methods, including supervised, unsupervised, semi-supervised, reinforcement learning, and more. Of these, supervised learning is particularly dependent on labeling, especially when working with unstructured data, where manual labeling is essential to maintain data quality and model accuracy. In materials science research, text and image data, such as literature and metallographic micrographs, are the primary datasets requiring annotation. As in other domains, the difficulty of annotating materials science data is a professional barrier. Only people with domain knowledge can annotate the required critical information. This is a prerequisite for ensuring the professionalism and reliability of the model. However, data annotation is a time-consuming and tedious process. In the case of small data sets, researchers can annotate the data themselves to ensure that the quality of the data is controllable. For large data sets, the time and energy required may beyond the personal capabilities of researchers. For example[32], the observation of dynamic phenomena at the atomic scale from scanning transmission electron microscopy (STEM) images is a significant challenge. Each STEM image contains tens of thousands of atoms, and the task is further complicated by instrument noise and image artifacts associated with atomic motion. Manually determining the precise morphological characteristics and distribution of individual nanomaterials, identifying all

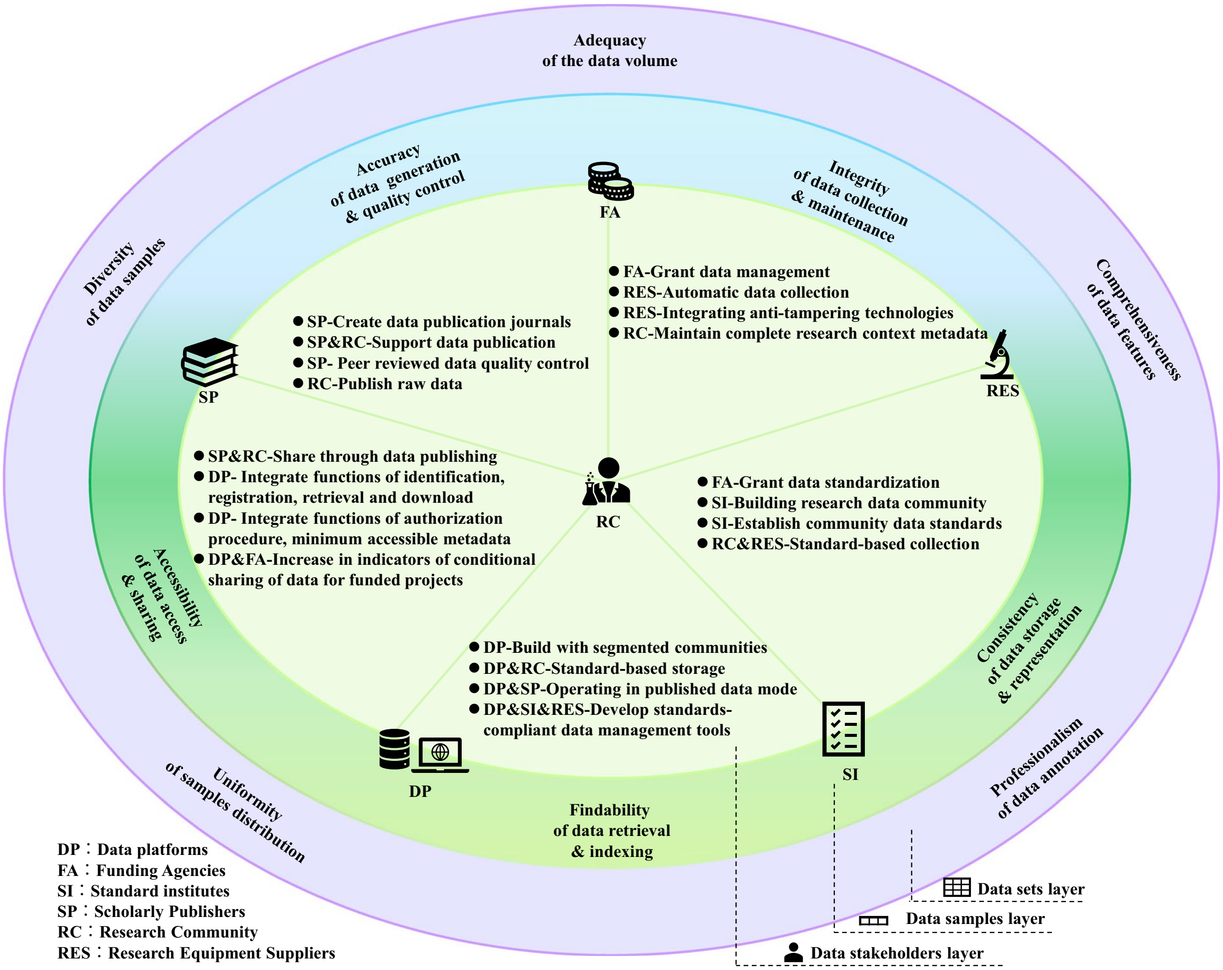
Adequacy
of the data volume
Accuracy
of data generation
& quality control
FA
Integrity
of data collection
& maintenance
Diversity
of data samples
Comprehensiveness
of data features
● FA-Grant data management
● RES-Automatic data collection
● RES-Integrating anti-tampering technologies
● RC-Maintain complete research context metadata
● SP-Create data publication journals
● SP&RC-Support data publication
● SP- Peer reviewed data quality control
● RC-Publish raw data
SP
RES
● SP&RC-Share through data publishing
● DP- Integrate functions of identification, registration, retrieval and download
● DP- Integrate functions of authorization procedure, minimum accessible metadata
● DP&FA-Increase in indicators of conditional sharing of data for funded projects
RC
● FA-Grant data standardization
● SI-Building research data community
● SI-Establish community data standards
● RC&RES-Standard-based collection
Accessibility
of data access
& sharing
Consistency
of data storage
& representation
● DP-Build with segmented communities
● DP&RC-Standard-based storage
● DP&SP-Operating in published data mode
● DP&SI&RES-Develop standards-compliant data management tools
SI
DP
Uniformity
of samples distribution
Findability
of data retrieval
& indexing
Professionalism
of data annotation
DP：Data platforms
FA：Funding Agencies
SI：Standard institutes
SP：Scholarly Publishers
RC：Research Community
RES：Research Equipment Suppliers
Data sets layer
Data samples layer
Data stakeholders layer

Fig. 1 A holistic view of the AI-ready scientific data ecosystem.

particles, and tracking their trajectories for dynamic studies is an exceedingly labor-intensive process. Handling the high-throughput and large-volume datasets exceeds the limits of manual processing.

In such scenarios where data annotation is extensive, particularly for massive image datasets, leveraging convolutional neural networks (CNNs), U-Net, and Mask R-CNN for automatic or semi-automatic segmentation and classification based on a small amount of manually annotated data can significantly enhance annotation efficiency and scale. In addition, professional annotation services provide automated or outsourced annotation. However, these annotators often lack domain-specific knowledge. Therefore, researchers must convey their expertise clearly and precisely to annotation companies to train annotators. Additionally, implementing multiple annotations, periodic sample checks, and other quality control measures is essential to ensure the professionalism of the annotated data and the reliability of the models.

The Overall Perspective of an AI-Ready Materials Data Ecosystem

The AI-ready data governance system from a holistic perspective is illustrated in Fig. 1. The outermost layer of Fig. 1 shows the management requirements of reliable AI models at the data set level, including the adequacy of data volume, comprehensiveness of data features, diversity of data samples, uniformity of data distribution, and professionalism of data annotation. These requirements need to be considered by data users when selecting and organizing data samples. The middle layer of Fig. 1 shows the requirements for individual data samples for AI-driven data set construction, including the accuracy of data generation and quality control, integrity of data collection and maintenance, consistency of data storage and representation, findability of data retrieval and indexing, and accessibility of data access and sharing.

To address the demands at the data sample level, a collaborative effort across the materials research community is essential, as illustrated in the innermost layer of Fig. 1: (1) Funding agencies should enhance financial allocations dedicated to data management and standardization initiatives. Additionally, they should incorporate requirements into the acceptance criteria of funded projects that mandate conditional data sharing within specified timeframes and scopes. (2)Research equipment suppliers should provide automated technological solutions that ensure comprehensive data capture and implement tamper-evident markers to preserve data's intrinsic objectivity. (3)Scholarly publishers are encouraged to acknowledge the academic value of data by establishing dedicated journals for data publication, cultivating an ecosystem that supports data dissemination, and offering robust data platform infrastructure to facilitate these processes. (4)Standard institutes should establish data communities and spearhead the development of consensus-driven data standards within each niche field. This will standardize the presentation of data across collection, storage, publication, and distribution phases. (5) Data platforms, acting as the custodians of data governance, should consider the holistic data usage requirements of specialized research groups. They should be designed with capabilities for data identification, registration, storage, publication, retrieval, and acquisition, and should integrate seamlessly with standardized storage formats and automated systems for data collection, publication, and sharing. (6)Researchers, in their roles as generators, utilizers, and disseminators of data, are urged to embrace a paradigm shift in their approach to data usage. They should proactively manage data with an awareness of its enduring and extensive value within a data-driven research landscape. Active engagement in the evolving landscape of data collection, storage, publication, and standardization is crucial to nurturing an AI-ready data ecosystem that will underpin future scientific advancements.

Conclusion

Artificial Intelligence is showing its disruptive advantages in scientific research and has attracted widespread attention from researchers worldwide. Domain-specific scientific data is the core and scarce research resource for AI to intervene in scientific research in this field. The current materials data ecosystem is small-scale, decentralized, and non-standardized, which cannot meet the needs of AI research for large-scale, high-quality, and sustainable data supply. Hence, the widespread promotion and application of AI in materials science research are limited.

We discussed the core requirements for materials data at various stages of its lifecycle, emphasizing ease and reliability in building artificial intelligence models. We propose some suggestions for the main obstacles in the existing data ecosystem to call on all stakeholders of scientific data to jointly build a new AI-ready scientific data ecosystem to accelerate the application of AI in scientific fields. Based on this, we discuss the preparation requirements that need to be considered at the dataset level to provide a reference for individual researchers to build AI-ready datasets. While this discussion is specific to materials science, it applies to most other natural sciences.

In AI-driven scientific research, scientific exploration can overcome the dimensions and scale of data analysis imposed by human capabilities. The availability, scope, and scale of scientific data will determine the progress of scientific exploration. The importance of data management goes beyond the narrow scope of data availability and directly impacts the scientific exploration process. Faced with this opportunity and challenge, the scientific community must adapt to future common research interests. This effort will be reflected in public basic research facilities such as data management plans, data standards, data sharing platforms, and automated data production equipment based on specific research communities. These measures will make research data in certain fields a public resource that all researchers can produce, collect, store, query, and access fairly and conveniently, thereby promoting scientific innovation in the field broadly and deeply.

Returning to the starting point of the discussion, just as Tony Stark's Jarvis appeared, when we are amazed by the stunning performance of ChatGPT, material researchers are also fantasizing about having an omniscient material version of an AI system. Building an AI-ready material science data ecosystem community may be the first step we should take toward it.

Received: 16 June 2023; Accepted: 27 August 2024;
Published online: 27 September 2024

References

1. Jumper, J. *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
2. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).
3. MacLeod, B. P. *et al*. Self-driving laboratory for accelerated discovery of thin-film materials. *Sci. Adv.* **6**, eaaz8867 (2020).
4. Szymanski, N. J. *et al*. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **624**, 86–91 (2023).
5. Leeman, J. *et al*. Challenges in high-throughput inorganic materials prediction and autonomous synthesis. *PRX Energy* **3**(1), 011002 (2024).
6. Zaki, M., Jayadeva, M. & Krishnan, N. M. MaScQA: investigating materials science knowledge of large language models. *Digital Discovery* **3**(2), 313–327 (2024).
7. White, A. D. The future of chemistry is language. *Nat. Rev. Chem* **7**, 457–458 (2023).
8. Scheffler, M. *et al*. FAIR data enabling new horizons for materials research. *Nature* **604**, 635–642 (2022).
9. Himanen, L., Geurts, A., Foster, A. S. & Rinke, P. Data-driven materials science: status, challenges, and perspectives. *Adv. Sci.* **6**, 1900808 (2019).
10. Jain, A. *et al*. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
11. Curtarolo, S. *et al*. AFLOW: An automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **58**, 218–226 (2012).
12. Saal, J. E., Kirklin, S., Aykol, M. & Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* **65**, 1501–1509 (2013).
13. Draxl, C. & Scheffler, M. The NOMAD laboratory: from data sharing to artificial intelligence. *J. Phys. Mater.* **2**, 036001 (2019).
14. Zakutayev, A. *et al*. An open experimental database for exploring inorganic materials. *Sci Data* **5**, 180053 (2018).
15. Wilkinson, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
16. Kratz, J. E. & Strasser, C. Researcher perspectives on publication and peer review of data. *PloS one* **10**(4), e0117619 (2015).
17. Austin, C. C. *et al*. Key components of data publishing: using current best practices to develop a reference model for data publishing. *Int. J. Digit. Libraries* **18**, 77–92 (2017).
18. Seo, S. & Kim, J. Data journals: types of peer review, review criteria, and editorial committee members' positions. *Sci. Ed.* **7**(2), 130–135 (2020).

19. Taillon, J. A. *et al*. NexusLIMS: A laboratory information management system for shared-use electron microscopy facilities. *Microsc. microanal.* **27**, 511–527 (2021).
20. Jain, A. *et al*. FireWorks: A dynamic workflow system designed for high-throughput applications. *Concurr. Comput.-Pract. Exp.* **27**, 5037–5059 (2015).
21. Supka, A. R. *et al*. AFLOWπ: A minimalist approach to high-throughput ab initio calculations including the generation of tight-binding hamiltonians. *Comput. Mater. Sci.* **136**, 76–84 (2017).
22. Mathew, K. *et al*. Atomate: A high-level interface to generate, execute, and analyze computational materials science workflows. *Comput. Mater. Sci.* **139**, 140–152 (2017).
23. Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N. & Kozinsky, B. AiiDA: automated interactive infrastructure and database for computational science. *Comput. Mater. Sci.* **111**, 218–230 (2016).
24. Duan, Q., Wang, X. & Song, N. Reuse-oriented data publishing: How to make the shared research data friendlier for researchers. *Learn. Publ.* **35**(1), 7–15 (2022).
25. Zheng, Z. *et al*. A GPT-4 Reticular Chemist for Guiding MOF Discovery. *Angew. Chem.-Int. Edit.* **62**(46), e202311983 (2023).
26. Qu, J. *et al*. Leveraging language representation for materials exploration and discovery. *npj Comput. Mater.* **10**, 58 (2024).
27. Choi, J. & Lee, B. Accelerating materials language processing with large language models. *Commun. Mater.* **5**, 13 (2024).
28. Ortega, D. R. *et al*. ETDB-Caltech: a blockchain-based distributed public database for electron tomography. *PLoS One* **14**(4), e0215531 (2019).
29. Strieth-Kalthoff, F. *et al*. Delocalized, asynchronous, closed-loop discovery of organic laser emitters. *Science* **384**, eadk9227 (2024).
30. Raccuglia, P. *et al*. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
31. Jia, X. *et al*. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature* **573**, 251–255 (2019).
32. Cheng, D. *et al*. Computer vision analysis on material characterization images. *Adv. Intell. Syst.* **4**(3), 2100158 (2022).

Acknowledgements

This work is supported by the National Key Research and Development Program of China (No. 2021YFB3702104), the National Natural Science Foundation of China (No. 52042301) and the Technology Standards Project of the Shanghai Municipality (No. 21DZ220600).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.W. or L.Z.

Reprints and permissions information is available at www.nature.com/reprints.

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