AI-Ready Materials Data, Standards, and Infrastructure

Authors: Lu Yongchao, Wang Hong, Zhang Lanting, Ning Yu, Wang Hong, Zhang Lanting

Date: 2022-11-17T00:00:00+00:00

Abstract

In recent years, the data-driven research paradigm characterized by "data + artificial intelligence (AI)" has experienced rapid development. The data generation, collection, storage, and application systems established around traditional research paradigms can no longer satisfy the requirements of the new paradigm, necessitating the urgent establishment of an AI-oriented novel data ecosystem to unleash the disruptive potential of data-driven approaches. This paper analyzes the characteristics of AI methods and proposes that material data in the AI context should adhere to the principles of being massive, comprehensive, complete, balanced, and shareable. Among these, data integrity and shareability are attributes of individual data records, which can be guaranteed through data standardization. By contrast, whether data satisfies the conditions of being massive, comprehensive, and balanced depends more on the characteristics of the data ecosystem, requiring support from an entirely new material data infrastructure. As a conceptualized ideal material data infrastructure, the "data factory" will revolutionize existing data production models, bringing about comprehensive improvements in both the quantity and quality of material data, and continuously providing AI-ready data.

Full Text

AI-Ready Material Data, Standards and Infrastructure

LU Yongchao, WANG Hong, ZHANG Lanting, YU Ning
Materials Genome Initiative Center and School of Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Correspondence:
WANG Hong, professor, Tel: 17321042821, E-mail: hongwang2@sjtu.edu.cn
ZHANG Lanting, professor, Tel: (021) 54747471, E-mail: lantingzh@sjtu.edu.cn

Funding: Supported by the National Key Research and Development Program of China (No.2020YFB0704504), National Natural Science Foundation of China (No.52042301) and Shanghai 'Science and Technology Innovation Action Plan' Technical Standards Project (No.21DZ2206000)

Manuscript received: 2022—, **in revised form:** 202*—

Author Contributions:
LU Yongchao, male, born 1993, Ph.D. candidate
WANG Hong (corresponding author): hongwang2@sjtu.edu.cn, research focuses on materials genome engineering theory, data-driven materials innovation infrastructure, and data standardization systems
ZHANG Lanting (corresponding author): lantingzh@sjtu.edu.cn, research focuses on high-throughput experimental techniques and data standardization for materials genome engineering

---

Abstract

In recent years, the data-driven research paradigm characterized by "data + artificial intelligence (AI)" has developed rapidly. The data generation, collection, storage, and application systems built around traditional research paradigms can no longer meet the requirements of the new paradigm. There is an urgent need to establish a new AI-oriented data ecosystem to unleash the disruptive advantages of data-driven research. This paper analyzes the characteristics of AI methods and proposes that material data in the AI context should follow the principles of massive scale, comprehensiveness, integrity, balance, and shareability. Among these, data integrity and shareability are characteristics of individual data records that can be ensured through data standardization. Whether data meets the conditions of massive scale, comprehensiveness, and balance depends more on the characteristics of the data ecosystem and requires support from new material data infrastructure. As a conceptualized ideal material data infrastructure, the "Data Factory" will revolutionize existing data production models, bringing comprehensive improvements in both the quantity and quality of material data, and continuously providing AI-ready data.

KEY WORDS: material genome engineering, data-driven, AI-ready materials data, data standardization, data infrastructure

---

In 2011, the United States launched the Materials Genome Initiative (MGI) [1], aiming to utilize quantitative data and computational codes to discover and predict material behavior, transforming materials R&D from a trial-and-error approach to a predictive paradigm, thereby accelerating the discovery, design, development, and deployment of new materials while reducing costs. Its core content involves developing advanced tools for materials computation, experimentation, testing, and data informatics, and integrating them to build a new

type of materials innovation infrastructure. The European Union, Japan, and other developed nations quickly launched similar government-led research programs [2][3][4]. The Chinese Academy of Sciences and the Chinese Academy of Engineering organized extensive consultations and investigations starting from the year MGI was released. Based on the Chinese Academy of Engineering' s consultation report on China' s version of the Materials Genome Initiative, Wang et al. [5] summarized the concepts of materials genome engineering and proposed recommendations for development strategies, technical routes, and policy measures for implementing China' s Materials Genome Initiative according to China' s actual needs and existing conditions. In 2015, the Ministry of Science and Technology launched the "Key Technologies and Support Platforms for Materials Genome Engineering" key R&D program [6]. Subsequently, Wang et al. [7] further discussed three representative working modes of materials genome engineering, clarifying that the fundamental difference between materials genome engineering methods and traditional methods lies in being data-based. They explicitly proposed that the data-driven model, marked by "data + artificial intelligence," revolves around data generation and data processing, revealing correlations among massive datasets and uncovering hidden parameter relationships through AI analysis. Data-driven approaches open new perspectives for materials research. Benefiting from AI' s efficient data analysis and processing capabilities, the data-driven model significantly increases the dimensionality of research questions and accelerates materials exploration speed, thus promising disruptive effects. In contrast, experiment-driven and computation-driven approaches still rely on traditional fact-based judgment or theoretical deduction, without changing established research thinking. Therefore, data-driven represents the core concept and development direction of materials genome engineering. In recent years, materials research work based on the data+AI methodology has been rapidly increasing (as shown in Figure 1 [Figure 1: see original paper]), and the trend of data-driven materials research has initially taken shape.

The scientific definition of artificial intelligence can be elaborated from multiple perspectives [8][9]. From a data perspective, AI is defined as "the capability of a system to correctly interpret external data, learn from these data, and flexibly adapt and apply the acquired knowledge to achieve specific goals and tasks" [10]. This capability provides materials research with a method to explore patterns through correlations among data, which differs from the causality relied upon by traditional physical models and offers a new perspective for scientific 规律研究 under conditions lacking fundamental physical models [11]. Materials are complex systems composed of vast numbers of atoms, and material properties are often the result of coupling among multiple physical mechanisms, rarely influenced by a single factor alone. Therefore, establishing simple models that correlate properties with only one parameter is difficult to describe clearly. From everyday experience, the human brain can typically only imagine three-dimensional images, and simultaneously processing problems with more than three variables is highly challenging. AI methods can easily investigate the coupling effects

of hundreds or thousands of parameters simultaneously, significantly increasing the dimensionality of understanding. Consequently, AI holds tremendous advantages in solving such problems.

Meanwhile, prior knowledge formed from traditional experimental or computational research is often used in practice to provide fundamental references for feature selection, model optimization, and interpretation when building knowledge models for AI [12][13][14]. Therefore, data-driven is not a simple replacement for experiment-driven and computation-driven modes but rather a supplement and extension based on them.

AI is built upon data. The scale and quality of data are positively correlated with the reliability of AI models. Therefore, data + AI together constitute the core content of the data-driven paradigm. Simply put, data represents the results we obtain through observation, experimentation, or computation [15]. In traditional thinking, data' s primary role is to provide facts as numerical bases for scientific research, technical design, verification, and decision-making, mainly manifesting its apparent value. For a long time, the materials science data ecosystem has been built around traditional research paradigms such as computation and experimentation. Data is often generated and collected as results from experiments or computations conducted by individual researchers to obtain specific information for particular research objectives. Consequently, the overall characteristics are multi-source heterogeneity, small scale, discrete distribution, and lack of standardization. In the AI context, data serves as a carrier of the comprehensive effects of various parameters, providing an information source for data mining. AI methods process and analyze large amounts of data, establishing correlations among data and uncovering the parameters and relationships that constitute these correlations, thereby more reflecting the intrinsic value of data.

Due to the new characteristics exhibited by the data-driven model in data usage, new requirements different from the past have been proposed for data used in AI, both in terms of organizational form and content. The vast majority of existing materials data has been collected, organized, stored, and presented for traditional application forms. In fact, such data encounters certain difficulties in AI-based applications regarding discovery, access, preparation, sharing, reuse, and automated machine processing, which objectively hinders and delays the faster and broader application of the data-driven model in scientific research [16]. Therefore, as materials science strides toward a new data-driven future, it is necessary to achieve a profound understanding and clear recognition of the characteristics, properties, and features that materials data should possess in the AI context, thereby guiding the collection, organization, storage, and use of future-oriented materials data to make it suitable for AI methods and helping them fully unleash their special potential. Data with such characteristics has been appropriately termed AI-ready in the recently released new version of the U.S. Materials Genome Initiative Strategic Plan [17]. Providing a clear explanation of the meaning of AI-ready will propose necessary basic guidelines

for building future-oriented materials science data infrastructure. This is of decisive significance for promoting the application of AI methods in materials science and accelerating the transformation of research paradigms from trial-and-error to data-driven predictive approaches.

This paper starts from the inherent characteristics of AI, combines the current status and latest trends of data governance in the materials field, conducts a comprehensive analysis of the characteristics and requirements that AI-ready materials data must satisfy, summarizes based on this analysis, and discusses measures for achieving AI-ready and related work being carried out in the field.

1.1 Massive Data

AI itself incorporates relevant knowledge from statistics and requires sufficient sample sizes to characterize the significance of potential patterns in the training data [18], and then applies the learned data correlation knowledge to decision-making for new samples. Numerous cases demonstrate that as the volume of training data increases, models become more accurate. For example, Schmidt et al. [19] reported that the prediction error for the formation energy of perovskite compounds shows a monotonic power-law decrease with increasing training set size, with error reductions of approximately 20% when the training set is doubled. Lee et al. [20] studied machine learning models for the band gaps of inorganic compounds, finding that the error of some models stabilizes with increasing data volume, while for support vector machine models, the error still shows a clear downward trend even at the maximum data volume used in their work, indicating that further increases in data volume will continue to promote model optimization. Therefore, massive data is the fundamental guarantee for AI to employ correlation-based strategies for exploration.

The materials research field has long maintained a project group-based work mode, where the research community primarily uses traditional low-throughput experimental or computational methods to characterize or simulate material properties, and then uses the generated results to construct structure-property relationships. The dispersion of work modes and the diversity of characterization and simulation methods result in numerous sources of materials data, and the research community has not established clear and unified data management standards, leading to different research teams collecting data of different types and formats, giving data the characteristic of multi-source heterogeneity [21]. Even when collecting research data centered on a specific material type, such as preparation and characterization data for nickel-based superalloys, the total available data volume remains limited due to differences in data template formats among different teams [22]. The problem of scarce available data is frequently mentioned in machine learning-related research works and review articles [23][24][25][26].

Currently, there are two main approaches to obtaining massive data in the materials field: (1) High-throughput experimentation and computation tech-

nologies are direct means for efficiently generating large amounts of materials data [27][28][29]. For example, high-throughput preparation techniques represented by combinatorial chip technology [30] can rapidly prepare thin-film samples covering complete ternary composition ranges on a 1-inch square substrate. Using synchrotron microbeam X-ray diffraction for characterization, the single-point diffraction measurement time can be shortened to 1-2 seconds, obtaining over 5,000 diffraction patterns on a single combinatorial materials chip sample with a total time consumption of less than 7 hours, enabling the completion of point-by-point structural characterization of 3 combinatorial materials chips per day. High-throughput computation represented by first-principles calculations leverages the powerful computing capabilities of advanced supercomputers, intelligent error-correcting automated computational workflows, and standardized computational parameter settings to rapidly and batch-produce large amounts of computational model data serving materials design [31]. High-throughput experimentation and computation technologies are effective ways to fundamentally accelerate the generation of materials data volume. During China' s 13th Five-Year Plan period, systematic arrangements for high-throughput experimental and computational technologies were made through the Materials Genome Engineering Key R&D Program, achieving numerous advances [6]. Various subdisciplines in materials science are continuously advancing this work [27][32][33].

(2) Extracting data from massive literature [34]. To date, various publicly published scientific literature represents the primary outlet and gathering place for large amounts of important scientific research data, and collecting them is of great significance [34]. Currently, there is no standard format for presenting research results, with most being publicly available in unstructured, heterogeneous forms. The Pauling File project [35] is one of the largest manually collected inorganic crystal materials data projects, extracting crystal structure, physical property, and phase diagram data from scientific literature in materials science, engineering, physics, and inorganic chemistry from 1891 to the present. To date, it contains over 350,000 crystal structures, 150,000 physical properties, and 50,000 phase diagrams, and launched its online version MPDS (Materials Platform for Data Science, https://mpds.io) in 2016. Meanwhile, drawing on experience from the biomedical informatics field, researchers have begun attempting to use computer technologies such as natural language processing and text mining methods to automatically extract data from literature. J. Cole from Cambridge University developed a natural language processing toolkit for chemical texts called ChemDataExtractor [36] and used it to construct large datasets of magnetic material phase transition temperatures [37] and electrochemical properties of battery materials [38]. Extracting data from literature is a compensatory solution for the current unstructured data publication ecosystem. Manual extraction mode requires expert knowledge for annotation, yielding high data accuracy but consuming substantial manual effort with low efficiency. In contrast, using natural language processing and text mining algorithms to automatically

extract literature data is more efficient but has lower accuracy. Extracting data from literature is an indirect data collection method, and the process of unstructured publication and subsequent extraction leads to significant loss of effective information. Therefore, it is necessary to reform knowledge attribution methods and sharing mechanisms to directly publish valuable research data.

1.2 Comprehensive Feature Quantities

The feature quantities contained in materials data determine the possible perspectives through which AI can describe phenomena. If data contains only a single feature quantity, the resulting understanding will inevitably be limited to the relationship between research variables and this feature quantity, unable to extend beyond this feature. A feature set that can completely reflect the materials research process will help AI generate more accurate understanding of correlations among data. In traditional research modes, since humans live in three-dimensional space, the human brain can only directly handle lower-dimensional research problems. During scientific reasoning, idealized forms are often adopted to simplify natural phenomena. For instance, classical physics frequently uses idealized assumptions such as "perfectly smooth planes," "ignoring air resistance," and "ideal gases" to remove complex interfering factors and retain only key factors for research and analysis. In modern materials science research tools, limited by technical conditions, similar approaches are often employed to ensure scientific inquiry can be conducted, such as "vacuum conditions" and "simulated seawater corrosion." These simplifications are reflected in data as low-dimensional descriptions of complex high-dimensional phenomena through dimensionality reduction, to facilitate human processing. Naturally, this inevitably leads to certain deviations between the real world and our understanding.

One of the characteristics of AI methods is their ability to process high-dimensional data, providing a new pathway to explore and understand a more realistic natural world. Correspondingly, AI-ready datasets should include as comprehensive and integrated feature parameters as possible to fully unleash the potential of AI. From a workflow perspective, both experiment-driven and computation-driven approaches first propose possible theories, then collect data, and verify them through characterization or simulation methods. This reliance on prior knowledge facilitates efficient optimization by focusing on known characteristic parameters but may inadvertently exclude many parameters that could be equally meaningful in practical problems, limited by the understanding at the time. This restricts our imagination in practical work [39] and causes some unknown key factors to be missed. In contrast, the data-driven paradigm, theoretically, does not presuppose which parameters are important or unimportant, thus avoiding habits and biases in parameter selection. For example, Ward et al. [40] created a universal feature space containing 145 material parameters around four aspects: stoichiometric proper-

ties, statistical properties, electronic structure properties, and ionic compound properties of chemical elements. This can represent features for inorganic materials composed of any chemical elements. Combined with various machine learning models and training data, it can predict the physical and chemical properties of materials, and its universality and effectiveness were verified in two different aspects: predicting crystal band gap energy, specific volume, and formation energy, and discovering novel amorphous materials. Simultaneously, to quantify the predictive capability of each feature for target properties, quadratic polynomial fitting was sequentially employed to measure the root mean square error of models. It was found that for different materials and properties, the characteristic parameters most influential for optimal modeling may vary significantly. For instance, the formation energy and melting temperature variations of intermetallic compounds are most correlated with the d-electron count among constituent elements, while compounds containing at least one non-metal are most closely related to average ionic characteristics (a quantity based on electronegativity differences among constituent elements). The variation of most relevant features in these examples further supports the necessity of having a large number of available features in machine learning feature sets. Although the 145 features covered in this work cannot completely encompass all features of inorganic materials, it represents a major step toward creating a rich material feature space, demonstrating the important value of comprehensive feature spaces for AI automatic analysis and exploration to acquire unknown patterns.

1.3 Data Record Integrity

From the perspective of the production process of materials research data, these data reveal not only the intrinsic properties of material samples themselves but also contain the influences of related factors such as material preparation, characterization, computational facilities, and processing workflows [41]. When using AI to analyze and process research data, these factors will all be reflected in the intrinsic correlations manifested in the data. In research aimed at process optimization, performance improvement, etc., the prerequisite for researchers to effectively obtain and utilize these hidden relationships is that the dataset contains complete feature dimensions that can reflect the research processes of preparation, characterization, computation, etc. Only then can refined, quantitative reference guidance be obtained on corresponding characteristic parameters and rapidly implemented in computational simulations and experiments.

Meanwhile, any preparation, characterization, or computation process contains numerous detailed parameters. AI-ready data must comprehensively include these parameters, enabling data users to fully understand the conditions, environment, and process of data generation as if they had experienced it themselves, truly ensuring correct and reasonable use of the data. From the perspective of current data collection methods, data producers primarily conduct experimental preparation, characterization, or computational simulation research on materi-

als based on their own research objectives. When recording each data point, they often select only a portion of "key parameters" that meet their research needs while directly ignoring or discarding other parameters generated during the research process. When these "incomplete" data records are published and reused, the lack of details often leads to irreproducible research results, which frequently occurs across various scientific fields [42]. To ensure the reliability of scientific work and the reusability of scientific data, some journals now require users to submit all source data for a achievement simultaneously when submitting pre-published manuscripts. For example, all journals under Nature Publishing Group have this requirement [43] and encourage authors to store all necessary data in public repositories with open access and describe the complete pathways for data acquisition. Some recommended public repositories include Figshare [44], Zenodo [45], and Dryad [46], among others. Considering that the spatial and temporal scope of data usage is continuously expanding under the data-driven paradigm, and different users have increasingly broad perspectives on data utilization, it is necessary to fully consider implementing a "collect all that should be collected" approach for parameters related to data generation actions during original data collection, leaving complete data parameter records to provide as detailed information as possible for data reuse and to offer a comprehensive feature space for AI to efficiently guide materials optimization design.

1.4 Data Distribution Balance

As mentioned earlier, AI reveals intrinsic data patterns by identifying correlations among multiple parameters. However, if the dataset used for training is unevenly distributed in parameter space, it will cause biased judgment results from standard models [47]. This is relatively common in commercially mature AI applications. For instance, Amazon abandoned an AI-powered intelligent recruitment system for scoring job applicants' resumes because the system produced unfair judgment results for female applicants. The reason for this bias was that the training dataset used to develop the algorithm was based on data related to previous applicants (primarily male) [48][49].

Similarly, when AI is applied to materials science inquiry, if the material feature dataset contains human biases, it will cause corresponding biases in the model' s judgment results. For example, in traditional materials science research, researchers often only pay attention to recording so-called "positive data" that aligns with research objectives while directly ignoring or discarding "negative data" that does not match the objectives. When such collected data is used for AI model training, it causes the model to lose partial objectivity in a statistical sense and misses opportunities to mine potential materials patterns. Essentially, the so-called "good or bad" of scientific data is a human qualitative assessment from a narrow perspective; data itself has no inherent quality distinction. From a statistical viewpoint, under rigorously designed scientific conditions, every piece of data generated from materials experiments reflects the objective pat-

terns of materials and should be recorded and preserved. Only then can models comprehensively and objectively reflect materials patterns during subsequent AI data analysis, possess strong robustness and scalability, and fully demonstrate the potential value of each data point. For example, in a famous case published in 2016, Raccuglia et al. [50] included both previous "successful" and "failed" experimental data in the training set when using decision tree methods to predict new metal-organic oxide materials.

1.5 Data Shareability

The reuse of scientific data is a fundamental requirement related to transforming research culture from individual work to collaborative big science models and is also a practical need for the data-driven model in the big data era. To constitute the required massive, multi-parameter, evenly distributed datasets, single-source data is often insufficient, necessitating the integration of discrete data from multiple sources. This requires that discrete individual materials data records have formal expressions that enable participation in large datasets, meeting users' shareability needs for convenient data access and use.

In recent years, data sharing has received widespread attention. China' s 13th Five-Year Plan Materials Genome Engineering Key Special Project also made specific provisions and mandatory requirements for data submission. The "Federal Data Strategy and 2020 Action Plan" released by the United States in 2019 [51], the "European Data Strategy" released by Europe in 2020 [52], and the "Scientific Data Management Measures" released by China in 2018 [53], among others, have all formulated supporting policies and implementation plans from the national strategic level to promote scientific data sharing. In some specific scientific fields, mandatory measures to promote data sharing have also been deployed. For example, the U.S. National Institutes of Health (NIH) stipulates [54] that starting from January 2023, it will require most of its 300,000 funded researchers and 2,500 institutions to include data management plans in their grant applications and ultimately make their research data publicly available. Data sharing has formed a consensus in the scientific community.

However, scientific data has traditionally been stored on local servers and lacks clear and consistent management standards. Data from different sources varies in expression format and completeness, making data neither easily accessible nor easily integrated and utilized, resulting in low sharing benefits. The international scientific community has been discussing for many years how scientific data can be more widely and fully utilized [55][56]. In 2016, Professor Barend Mons from Leiden University in the Netherlands, together with representatives from a series of data stakeholder industries including academia, industry, funding agencies, and academic publishers, jointly designed and endorsed a concise and measurable set of data management principles—the FAIR (Findable, Accessible, Interoperable, Reusable) principles [57]—to enhance data shareability and reusability on a broader scale. The FAIR principles have received widespread recognition in the scientific community, and some new data sharing infrastruc-

ture constructions are being built based on the FAIR principles [58][59].

The basic requirements of the FAIR principles can be summarized as follows:

Findable: The Findable principle addresses the questions of where and how to query target feature research data needed for AI research. It requires data to be identified by unique and persistent identifiers, with the typical representative being the DOI (Digital Object Identifier) system [60], which can provide a resolution access method for data objects through system-assigned unique data identifiers in the public internet space to locate the storage position of target data. Simultaneously, it requires describing data with rich metadata and registering or indexing it in searchable data resource platforms, enabling searchers to precisely retrieve target data through data characteristic attributes, satisfying the basic requirement that target data can be discovered by users.

Accessible: The Accessible principle addresses how to obtain queried target data and specifies minimum implementation requirements for data access methods. It requires retrieving data and its metadata under open, free, and universally implementable standardized communication protocols, enabling data to be transmitted freely and simply through network infrastructure. Regarding intellectual property rights issues in the data acquisition process, it allows data owners to set data access permissions and conduct identity verification and authorization processes for data users when necessary, encouraging data openness while respecting data ownership. Simultaneously, it requires that metadata describing basic data information be persistently accessible, ensuring that the information carried by data can be stably obtained at a minimum level even if the data object becomes inaccessible due to various reasons. Additionally, the Reusable principle requires that clear and accessible data usage license requirements be included during data publication, providing explicit annotation prompts for proper data access and acquisition.

Interoperable: The Interoperable principle requires data to be described using formal, accessible, shareable, and widely applicable languages. The vocabulary involved should be selected from FAIR-compliant vocabularies or existing authoritative terminologies, thereby making its expression form universal within the domain, avoiding semantic and format incompatibilities when integrating data from different sources, and enabling both humans and machines to conveniently process data, establishing a domain-consensus understandable language mechanism for AI applications.

Reusable: For data not generated by oneself, the Reusable principle addresses the issue of how to completely understand and correctly use these data when building AI models. It requires describing data with multiple accurate and relevant metadata that should be related to the detailed provenance of the data and organized in compliance with domain-related standards, enabling users to understand the background and content composition of data as detailed as possible and facilitating their reasonable utilization. The Reusable principle fully considers the content requirements that non-data producers should possess

to completely understand data, providing reliability guarantees for the correct understanding, use, model interpretation, and application of multi-source data needed for AI models.

2.1 Standardized Governance of Materials Data

The AI-ready requirements for materials data—massive scale, comprehensiveness, integrity, balance, and shareability—reflect the new data ecosystem characteristics under the data-driven research paradigm. Among these, data integrity and shareability are characteristics of individual data records that can be ensured through standardization. Standardization is the activity of establishing common and reusable terms as well as preparing, publishing, and applying documents for actual or potential problems to achieve optimal order and promote common benefits within a defined scope [61]. Real-world materials data covers the entire chain of materials R&D, from electronic, atomic, and molecular phenomena, to the effects of multi-scale process conditions on material performance and service behavior, and even to details of application design and manufacturing technologies. Managing such massive and diverse data requires domain-wide coordination to establish common rules, thereby seamlessly achieving data exchange and sharing to realize AI-ready goals. Traditional materials databases generally collect analysis results derived from raw data processing (such as various material performance parameters), while raw data is typically scattered among experimenters, not included in databases, and has diverse formats that are inconvenient for others to reuse. Furthermore, when this data is generated, it often targets specific applications, contains relatively limited material attributes, and lacks comprehensiveness. Consequently, the parameters that can be correlated with the data are quite limited. This is closely related to traditional materials research methods and data generation approaches. Therefore, existing materials databases mostly cannot meet the needs of materials genome engineering. Under the premise of data-driven approaches, it is necessary to propose general rules for establishing materials data structures that meet AI-ready requirements through top-level design to regulate the content composition of AI-ready data.

Data standards provide important safeguards for constructing AI-ready databases (or datasets). Materials data is characterized by large volume, diverse types, various formats, different producing units, and complex intellectual property ownership. Without unified standards to follow, not only does collection and storage become more complex, but usage also becomes inconvenient. Under current conditions where multiple data infrastructures coexist, some form of standardization is essential for practicing the data-driven paradigm [62]. Therefore, establishing unified data standards is a key measure for standardized data governance and lays an important foundation for large-scale adoption of AI methods in the materials field.

2.1.1 Content of AI-Ready Data Standardization

Metadata is a relatively intuitive data organization and management method. Metadata is typically defined as data about data, essentially a form of structured description of data objects from certain perspectives. For example, describing a person can be done through numerous elements such as name, gender, height, age, personality, etc. Reflecting the characteristics possessed by data objects from specific perspectives requires selecting relevant elements to form specific metadata schemas. Standardization is the activity of regulating the content covered in metadata schemas through consensus achieved within a certain social scope. In the data-driven model, metadata is the actual carrier for data retrieval and AI analysis. Data integrity and shareability can be ensured by including corresponding elements in standard metadata schemas. Due to the complexity and diversity of materials systems, developing information-rich, detailed, and adaptable standardized metadata for materials science is a prominent challenge [62]. Currently, in terms of materials metadata standard construction, the international community is still in its infancy. Existing metadata standards are either completely missing or incomplete. Standards organizations (such as the International Organization for Standardization (ISO)) have made many attempts to provide standards related to metadata normalization such as controlled vocabularies, data formats, and data processing, but so far none have been adopted within the domain [63].

Ontology is "a formal, explicit specification of a shared conceptualization" [64]. Ontology can describe specific conceptual systems within a domain and the definite relationships among elements within them. At the practical construction level, ontology itself does not define its representation form and can be expressed through various languages such as OWL, DAML, RDFS, IDEF5, etc. [65], transforming ontology design into computer-processable patterns. Currently, the more commonly used ontology language is OWL (Web Ontology Language). Various ontologies share similarities in expression structure, all employing basic constructive elements such as concepts (also called classes), instances, properties, relationships, and constraints for more specific descriptions [65]. For example, when describing data "properties" such as "tensile strength" and "elongation" for 45# steel material, this conceptual system includes: "Material" is a "class" representing all types of materials; within "Material" , there can be subclasses such as "Metallic Materials" , "Inorganic Materials" , "Polymer Materials" , etc. ("subclass" represents the "relationship" between them), and "Steel Materials" is a subclass of "Metallic Materials" ; "45# steel" is a specific instance within "Steel Materials" ; this instance has multiple "properties" such as "tensile strength" and "elongation" , and "tensile strength 460MPa, elongation 17%" defines two property values for the instance "45# steel" . Furthermore, "ferritic steel" can be defined as a steel containing at least one ferrite structure. We can use the relationships among "steel" , "ferrite" , and "basic microstructure" in the materials ontology to constrain and define the concept of "ferritic steel" [66]. Such constraints make data meaningful to both humans and com-

puters and form the basis for establishing automatic reasoning by computers on conceptual systems. Both ontology and metadata are tools for describing data resources, both representing features contained in objects through concepts or terms. The difference is that metadata organizes these terms in a tree structure, making expression more modular and intuitively concise, while ontology represents them in a network structure, more prominently highlighting the interconnections among these terms and providing semantic background for data resource understanding and utilization. They can be mutually transformed through the terms and relationships they contain.

Standard metadata schemas can be expressed as a universal conceptual system with standardized expressions and interconnections defined within a discipline. The metadata elements within it, based on their logical relationships in the conceptual system, can be regarded as constructive elements of different concepts in the system. Therefore, ontology can be used to describe and reflect the relationships among metadata elements.

In materials science, existing materials ontologies are classification schemes for materials, material properties, units, and constraints and their interrelationships [62]. Establishing standardized materials ontologies will provide researchers in the materials field with a shared standard conceptual system, promoting standardized collaboration among different researchers in the domain for describing and managing similar data, and enhancing data interoperability. Data exchange among multiple heterogeneous databases can be conveniently achieved through intermediate data representations based on materials ontologies. As the adoption scope of ontologies expands, it will also unleash the potential for machine automatic reasoning and mining of implicit knowledge associations among massive materials data. Currently, ontology construction for materials science has just begun and is far from covering a complete knowledge system. Meanwhile, various ontologies and less formal standards compete with each other [62]. For example, NOMAD Meta-info [67], ESCDF [67], and OpenKIM [68] are initial attempts to classify computational results in atomistic materials science; PLINIUS [69] is used for the ceramics field; ONTORULE [70] for the steel industry; SLACKS [71] for laminated composites; and PIF [72], Ashino [73], EMMO [74], MatOnto [75], Premap [76], and MatOWL [77] represent general materials science data, among others. No standardized ontology ensuring complete representation of materials has emerged. Although the development process of materials ontologies has accelerated, they have not yet matured like in other fields (such as biosciences) [66]. In industrial applications, these publicly available ontologies are often insufficient, forcing commercial companies to create their own internal, use-case-specific ontologies [74].

2.1.2 Current Status of Material Data Standards (Domestic and International)

In recent years, the materials informatics field has begun to widely recognize and emphasize the importance of data standardization [6][41][62][63][67][78]. How-

ever, in practice, establishing and promoting standards is a time-consuming and labor-intensive task. Especially in countries with good database foundations, forming consensus among all parties seems like an impossible mission. To address the need for rapidly accumulating large amounts of data, institutions such as the U.S. National Institute of Standards and Technology (NIST) have adopted the data warehouse approach, which does not restrict the format of materials data but stores as much data as possible for future development of tools for organization, analysis, and mining. The data warehouse form is a quick and pragmatic solution for solving data volume bottlenecks and is also a compromise for the current lack of data standards. With subsequent technological and standard advancements, data can certainly be standardized later, but information missing from the original data cannot be remedied afterward. Therefore, standardization should be implemented from the beginning as much as possible.

European and American countries focus on utilizing existing data and data systems, striving to improve the interoperability of multi-source heterogeneous data by establishing complete sets of materials science ontologies. However, this metadata coordination approach still requires developing data converters and shared data schemas. The European Novel Materials Discovery (NOMAD) Laboratory focuses on collecting, storing, and cleaning computational materials science data. For instance, they can directly store raw data generated by more than 10 mainstream ab initio computational codes worldwide and then normalize the raw data into formats meeting certain standards by developing translators [67]. FAIRmat [63] is a data consortium organization supported by the German National Research Data Infrastructure (NFDI, https://nfdi.de) that will build a joint infrastructure for many specific data repositories in the materials field. All participating groups or institutions will use a unified framework to manage their data, i.e., sharing a central metadata repository in computation, management, and storage. Since metadata differs across different subdomains and topics, a bottom-up hierarchical approach is adopted in management, extracting common metadata elements to the upper layer as public attributes, such as material composition and research methods, thereby forming a layer-by-layer progressive data organization and query mode similar to shopping websites. Based on these metadata, an encyclopedia of materials data descriptions is formed, which can simultaneously support both general queries for non-expert users and specific queries for expert users. FAIRmat has already begun establishing metadata and dictionaries for digital translation of vocabularies used in different fields. The next step is to develop ontologies, establish hierarchical and other relationship descriptions among metadata, and then deploy standard metadata and ontologies into Electronic Laboratory Notebooks (ELN) and Laboratory Information Management Systems (LIMS) to achieve interoperability of data collected and stored by different groups. This bottom-up metadata normalization work mode gives FAIRmat high flexibility when connecting new subdomains, but this metadata coordination approach requires developing data converters and shared data schemas. A concrete example of this metadata coordination scheme is the first version API recently released by the Open Databases Integration for Materials

Design (OPTIMADE) [79] consortium, which allows users to access the common subset of metadata schema items from participating databases, achieving unified access to distributed databases.

China recognized the importance of standards early in its exploration and research of materials genome engineering methods. When the China Standards of Testing and Materials (CSTM) was established in 2017, Chinese scientists foresightedly proposed establishing the CSTM Materials Genome Field Standardization Committee, the first standardization committee in the materials genome engineering field internationally, taking the lead in important exploration and demonstration of standards and standardization in the materials genome engineering field. On November 22, 2017, during the First Materials Genome Engineering High-Level Forum, the CSTM Materials Genome Field Standardization Committee (CSTM FC-97) was formally established, with six technical committees under it: General Principles, Computation, Preparation, Characterization, Data, and Application, respectively responsible for group standard system construction work in various fields of materials genome research, development, and application, including materials products, material process methods, material test methods, material test technology evaluation methods, material evaluation methods, material models and software, material computation, material data specifications, and materials field management and work standards. Considering the data-centric characteristics of materials genome engineering, the FC-97 Committee determined to focus materials-related standard development around data. Currently, there are no ready-made materials genome engineering data standards internationally to draw upon. Referencing the actual situation in international materials data standard construction and combining China' s materials R&D field characteristics and institutional advantages, the overall standardization construction policy proposed by FC-97 is: through top-level design, establish a future-oriented data standard system suitable for materials genome engineering data systems. The data standard system will contain a series of standards and rules covering all aspects of technology, processes, and functions involved in the entire lifecycle of materials data, regulating the content that must be collected for data entries and the formats, protocols, and regulations to be followed, ensuring that materials data obtained, stored, and used all meet AI-ready requirements and comply with the data-driven model.

First, FC-97 chose to start with establishing general data rules on the CSTM platform. Based on the fundamental starting point of maximizing satisfaction of the FAIR principles for data, it established principles for the content included in data entries. In August 2019, CSTM released the world' s first group standard on materials genome engineering data—T/CSTM 00120 "General Rule for Materials Genome Engineering Data" (hereinafter referred to as the "General Rule") [6], jointly formulated by more than 30 major materials research institutions in China. The "General Rule" breaks through the limitations imposed by material and division-of-labor diversity on standard development, approaching from the data level and proposing a highly compatible materials data classification framework. As shown in Figure 2 [Figure 2: see original paper], the "General

Rule" divides data into three categories according to the needs of materials science under the data-driven model: sample information, raw data (unprocessed characterization data), and derived data (data obtained through analysis and processing). Here, samples can be physical objects produced by experiments or virtual objects generated by computation. Similarly, raw data can come from characterization or direct measurement, or can be generated through simulation calculations. Note that each data record here takes a single operation (sample preparation/characterization/computation/data processing) as the unit, collecting only content related to that operation. For example, a data record about sample information contains only information about the preparation of that sample and does not include content about characterizing that sample. Each data record is assigned an independent and permanent resource identifier (for example, according to national standard GB/T 32843 or any independently assigned unique and permanent identification system).

Figure 2 The classification of material data categories and their contents in the General Rule for Materials Genome Engineering Data

The design of the "General Rule" focuses on solving three problems: First, raw data (unprocessed data) contains a large amount of information. Its multiple uses, especially for different purposes, is an important guarantee for data reusability. Currently, raw data is mostly scattered among producers, not included in databases, greatly limiting data reuse. This classification ensures from a systemic perspective that raw data is recorded, thereby guaranteeing the possibility of reuse. Second, traditional data currently organizes and presents the relationships among composition-structure-process-performance from the data producer' s perspective in an integrated manner, which formally limits the scope of data application fields and is not conducive to expanding application areas. The "General Rule" defines the content unit of data entries as a single action (preparation/characterization/processing). Under the premise of ensuring rich metadata, individual data records can circulate and be used independently based on their own information, conveniently participating in users' multi-perspective materials investigations and being flexibly and freely combined and reused under different research purposes and contexts.

Third, listing samples as a separate category of data is a practice not found in any previous data systems. The greatest advantage of this approach is that it makes the sample itself a public social resource compliant with FAIR principles, facilitating sample sharing, multiple uses, and reuse in digital proxy form. In addition, there are several other important considerations: 1) Avoiding data processing burdens caused by including excessively large and repetitive sample information in characterization metadata and derived data, which may become unacceptably large especially in derived data; 2) The premise of listing samples as separate items is that each sample is a unique individual. Even for two samples with completely identical apparent parameters, the reproducibility they reflect has statistical significance in materials data science. Traditional databases use one sample as a representative of samples with the same name,

essentially assuming that the listed parameters are characteristic values of the given material, which objectively eliminates differences brought by detailed factors.

Currently, a series of regulatory general standards based on the "General Rule" principles, such as materials genome engineering terminology standards, data identification standards, and data general specifications, are under construction [80][81][82], respectively providing authoritative terminology, identification methods, and standardized processes and method references for the development of data standards for various research methods, to more specifically serve data standardization work construction. For example, sufficient metadata is a basic condition for data reuse and an important component of AI-ready requirements. Currently, metadata included in materials data collections is often very incomplete and fails to meet AI-ready requirements. Therefore, the data general specifications will explicitly stipulate that specific standards must follow the principle of "collect all that should be collected" to gather sufficient metadata. At the current stage, since the software and hardware used in the data/metadata generation/collection process do not consider the requirement of collecting all that should be collected, completing such actions inevitably involves substantial manual recording and entry, causing data management to occupy large amounts of time and energy, making implementers overwhelmed and inevitably leading to slackness or even resistance. The key to resolving this contradiction lies in completing data standardization as soon as possible and implementing standard rules in software and hardware configurations. With the development of high-throughput experimental and computational technologies, the data generation/collection process will inevitably achieve full automation, and this problem will gradually weaken until it disappears. To this end, some workflow control software systems based on the characteristics of experimental equipment or computational software data generation have been developed. For example, the U.S. NIST developed a laboratory information management system for electron microscopes—NexusLIMS [83], which can package all data and metadata generated during a user' s session on a Nexus electron microscope into a structured text document representing an experimental snapshot, achieving automatic backup and archival storage of all raw research data, while also building a web-based portal for users to search and access previous experimental records by date, user, instrument, sample, or any other metadata parameters. The computational materials field, due to its inherent normalization and digital characteristics, has also developed multiple automated data workflow management software systems centered around materials computation, including Fireworks [84], AFLOW Π [85], Atomate [86], AiiDA [87], etc., which can achieve automated collection and storage management of computational data, having relative advantages in complete data collection. Furthermore, unified data templates, i.e., data standards, should be established for general preparation, characterization, computational techniques, methods, and workflows, enabling convenient sharing of such data.

Since the release of the "General Rule" , a series of exemplary data standardization efforts around specific research methods have also been actively carried

out. Supported by the national key special project "Key Technologies and Support Platforms for Materials Genome Engineering" , a standardized metadata template for ion beam sputtering deposition samples has been constructed and deployed on the National Materials Genome Engineering Data Submission and Management Platform (http://nmdms.ustb.edu.cn), which has been used in daily scientific research data management. Supported by the major science and technology special project "Yunnan Rare and Precious Metals Materials Genome Engineering" , data standardization work around the research process of rare and precious metals materials is underway, and based on this, a large professional database for precious metals materials has been constructed (http://ipm-int.matclouds.com). The standard for material thermal property calculation data based on DFT methods has been completed [88]. Additionally, several full-area high-throughput in-situ statistical mapping characterization technology standards for large-size components have been jointly formulated in combination with the national key industrialization project for high-speed train wheel-axle systems and comprehensive fields, providing scientific support for evaluating the quality of related material components using innovative materials genome engineering methods. Currently, 13 items have been applied for project establishment, and more than 30 project proposals have been submitted. Related work covers data generation, collection, storage, sharing, and utilization, and is being actively promoted among member units of the Materials Genome Engineering Field Committee of the China Standards for Testing and Materials Group Standard Committee. The CSTM standard system will ensure that materials genome engineering research activities and their outcomes are leading, standardized, accurate, efficient, and reproducible, and the innovation-driven standardization of materials genome engineering will certainly provide strong support for the high-quality development of the materials industry.

2.1.3 Materials Data Standard System

A complete AI-ready materials data ecosystem needs to be ensured by constructing a complete data standard system. The "General Rule" has established a foundation and pointed out the direction for the standardization of AI-ready materials data and has also been used as a basic principle followed in a broader sense for establishing materials data standards [80][81][82][88]. Materials data is complex and diverse. The data standardization work centered on the "General Rule" adopts a combined top-down and bottom-up work mode. First, starting from top-level design, a standard system framework that comprehensively covers all aspects of materials data-related issues is proposed, providing overall planning for the standards that need to be established. In practice, based on the standard system framework, experts from various fields are mobilized to leverage their respective professional expertise, with the "General Rule" as the core guiding principle, starting from specific problems and gradually establishing various detailed data standards. The framework for the materials genome engineering data standard system is shown in Figure 3 [Figure 3: see original paper] and can be divided into five sections in terms of content.

Figure 3 Schematic diagram of the framework for standardization of materials genome engineering data

• **Basic General Standards** clarify the general requirements for materials data. Among them, the "General Rule" provides overall design and planning for the objectives and content of materials data standardization. Standards such as materials genome engineering terminology standards, data identification standards, and data general specifications concretize the various general requirements for data in the "General Rule" . As mentioned earlier, they respectively provide authoritative terminology, identification methods, and standardized process and method references for the development of data standards for various research methods, to more specifically serve and guide the overall construction of data standardization work. Currently, these three general standards are in the review and revision process [80][81][82].

• **Experimental Data and Computational Data** are two sections related to materials data generation. Corresponding standards specify the content that should be included in data entries generated by various experimental or computational methods from the perspective of materials data producers. In specific implementation, three aspects need to be focused on: data classification, standard construction granularity, and standardization content. First, according to the classification of materials data in the "General Rule" , data is divided into three categories—sample information, raw data, and derived data—based on several data generation processes: experimental preparation/computational (virtual) preparation, experimental characterization/computational characterization, and data analysis. Second, each standard takes an independently existing data generation action (sample preparation/characterization/computation/data processing) as the entry theme and the specific method employed by that action (sample preparation/characterization/computation/data processing) as the carrier. For example, for the thin-film sample preparation process using the "Physical Vapor Deposition (PVD) method" , the corresponding "Physical Vapor Deposition (PVD) Thin-Film Sample Information Metadata Standard" is established; for "X-ray Diffraction Analysis (XRD)" characterization, the "XRD Characterization Metadata Standard" is established; for "XRD Data Phase Analysis" , the corresponding "XRD Phase Analysis Derived Metadata Standard" is established. Computational data standard examples include VASP structure optimization calculation metadata standards (virtual samples), VASP force constant calculation metadata standards (virtual characterization), etc. Furthermore, the content of standards constructs standardized metadata patterns with the data output action process as the description object. Standards for high-throughput experimental and computational data should not only include provisions for the basic technologies of corresponding sample preparation/characterization/computation/data processing but also reflect the characteristics of high-throughput technologies.

• **Data Application Standards** include a series of standardized application dataset metadata patterns established from the perspective of materials data

application in research, based on material properties and parameters of concern in different materials subfields. For example, in low-alloy high-strength steel research, people typically focus on parameters such as key composition, mechanical properties, microstructure, and processing technology. Domain experts, based on years of experience, construct metadata patterns including commonly used characteristics for that material and form domain consensus to make it the "Low-Alloy High-Strength Steel Application Metadata Standard" . Data application standards are granularly divided according to material types, providing users with an expert experience perspective.

• **Data Technology Standards** proceed from computer science to establish consensus protocols, specifications, and standards for materials data standards in aspects such as data storage, interaction, mining, quality control, and data security, providing information technology guarantees for consistent management and interoperability of data at the machine level. Related work is being actively promoted among member units of the Materials Genome Engineering Field Committee of the China Standards for Testing and Materials Group Standard Committee.

2.2 AI-Ready Data Infrastructure

The implementation of data standardization provides governance solutions for constructing complete, reusable, and shareable normalized individual data records, and also lays the foundation for the extensive construction of massive, feature-comprehensive, and evenly distributed materials datasets. Existing materials research infrastructure has been designed and developed based on current needs, and the data produced differs significantly from AI requirements in both quantity and quality. Therefore, obtaining AI-ready data requires support from new materials data infrastructure that matches it. The new materials innovation infrastructure will be data-centric with AI as the keyword, consisting of three components: a data platform, a high-throughput experimental platform, and a high-throughput computational platform. The data platform includes software tool libraries based on AI methods and AI-ready databases; the high-throughput experimental and computational platforms serve as data production sources, providing effective pathways for rapidly acquiring large amounts of data. In this way, the three technical elements of materials genome engineering achieve intrinsic synergy, forming an inseparable deep integration relationship.

The related technologies for building AI-ready new data infrastructure include high-throughput experimental technologies for data, automated data collection and storage technologies, high-throughput computational technologies, data standard systems, standardized storage of data semantics and structures, unified data identification, and network access and acquisition. Data standardization permeates every 环节 of each data record. Through comprehensive application of these technologies, full-chain comprehensive basic capabilities for AI-ready data generation, collection, storage, processing, exchange, sharing, use, analysis, and

network collaboration are realized [16].

Based on the above considerations, Wang et al. [91] proposed the conceptual model of "Data Factory," which posits that under ideal conditions, AI-ready data should be generated from a dedicated facility platform that mass-produces data in a standardized manner like an industrial production line. Figure 4 [Figure 4: see original paper] shows the conceptual diagram of the Data Factory. The center of the conceptual diagram is the data facility of the Data Factory. The right wing of Figure 4 shows the experimental Data Factory, which can be a large-scale, systematic high-throughput integrated preparation and characterization platform facility based on large scientific facilities (such as synchrotron light sources, neutron sources, etc.), integrating a series of in-situ preparation and multi-parameter characterization methods, capable of generating multi-parameter data including mechanical, electrical, optical, thermal, magnetic, and acoustic characteristics and properties. Ideally, all property measurements are conducted in real-time and in-situ on the same sample. The left wing of Figure 4 shows the concept of the computational Data Factory, which is essentially a computing center with various high-throughput computational software and hardware, equipped with high-throughput computational workflows through multiple methods such as density functional theory, molecular dynamics, CALPHAD methods, phase-field simulations, and finite element analysis, capable of generating large batches of comprehensive computational data from atomic to macroscopic scales. The Data Factory can be centrally established at a single location or can be a distributed platform composed of a group of virtually linked sites.

Figure 4 Conceptual diagram of the Data Factory—a dedicated facility capable of mass-producing data in a standardized manner, just like an industrial production line [91]

The Data Factory will directly respond to all aspects of AI-ready requirements for materials data: automated, uninterrupted pipeline-style data collection and storage methods provide guarantees for massive data generation; public data production facilities weaken the strong purposefulness that researchers typically have, making feature parameter distribution more balanced; high-throughput generation methods facilitate obtaining data with better systematicity and consistency; comprehensive observation indicators provide a huge feature space for AI exploration of unknown patterns. Data standards can be conveniently implemented in the Data Factory, making data collection, storage, and management all follow uniform methods, ensuring that FAIR principles are satisfied for any data record. Simultaneously, due to the realization of automation and standardization, collecting large numbers of parameters following the "collect all that should be collected" principle is no longer a burden.

The emergence of the Data Factory will bring a series of major transformations to data production. First, for broader and more long-term goals, comprehensive and balanced materials datasets will be consciously generated on a large scale, rather than being limited to byproducts of dispersed experiments or com-

putations with specific purposes. Second, the comprehensive implementation of data standards ensures parameter integrity and data shareability, significantly improving the usability and applicable scope of each data record. Third, the Data Factory transforms data generation from individual activities to organized social activities. Fourth, this organized effort will change the social attribute of data from private property to public resources. The result will bring comprehensive improvements in both the quantity and quality of materials data, make data sharing much simpler, and reduce overall social costs. This new data generation approach represents a revolutionary change for materials science.

The Data Factory conceptual model reflects the latest development trends in materials innovation infrastructure. The latest "Materials Genome Initiative Strategic Plan" released by the U.S. White House National Science and Technology Council in November 2021 [17] made key deployments for materials innovation infrastructure, proposing to connect, create, and strengthen elements such as computational tools, experimental tools, and data storage and sharing software frameworks, building a national materials data sharing network and integrating it into a unified materials research continuum, thereby expanding MGI coverage and improving the accessibility of research resources. On the basis of this unified data network architecture, with the goal of building AI-ready data, utilizing and strengthening materials innovation infrastructure will greatly accelerate materials R&D through the application of AI methods.

Currently, a series of database platforms based on high-throughput computational platforms or computational "Data Factories" have been developed internationally. The Materials Project [92], jointly developed by MIT and Lawrence Berkeley National Laboratory, relies on the supercomputing cluster of the U.S. National Energy Research Scientific Computing Center (NERSC) and utilizes its developed Fireworks workflow software and Custodian job management software to automatically manage computational and data processing workflows, establishing a large-scale materials first-principles calculation database. To date, it includes computational property data for over 146,000 materials, 24,000 molecules, more than 4,000 battery materials, etc., with computational volume reaching 100 million CPU hours per year, and provides various retrieval and analysis tools to help researchers quickly obtain and analyze data (https://nextgen.materialsproject.org). Other well-known high-throughput computational data platforms include Automatic Flow for Materials Discovery (AFLOW) [93], Open Quantum Materials Database (OQMD) [94], Novel Materials Discovery (NOMAD) [95], and MatCloud [96], among others. It is worth noting that these infrastructures adopt their own unique approaches in data management and storage and do not follow the same standard among each other, creating many inconveniences when integrating multi-source data into AI-ready data [62].

The common API released by the OPTIMADE [79] consortium supports data infrastructures including AFLOW, Materials Project, NOMAD, OQMD, Materials Cloud [97], etc. Through the OPTIMADE API, cross-database retrieval can be achieved across these materials data infrastructures with different physical

locations, reflecting the characteristics of distributed construction and virtual linking of the "Data Factory".

Compared with computation, there are currently fewer large-scale experimental database platforms with "Data Factory" characteristics. The High Throughput Experimental Materials Database (HTEM DB) [98] is one of the few typical representatives. HTEM DB was built by the U.S. National Renewable Energy Laboratory (NREL) based on its high-throughput preparation and characterization experimental data for combinatorial thin-film samples using physical vapor deposition (PVD), and developed the LIMS materials experiment information management system responsible for automatically collecting, indexing, and archiving experimental data. The current public version covers composition (55,000+), structure (65,000+), optical (46,000+), and electrical property data (19,000+) for over 82,000 various thin-film material samples (oxides, nitrides, sulfides, phosphides, intermetallic compounds) synthesized by physical vapor deposition. It also provides a user interface for researchers to query and retrieve data and allows obtaining more data for data mining and analysis through the provided Application Programming Interface (API) (https://htem.nrel.gov).

The data-driven model has brought disruptive development opportunities for materials science research, and the value of data is shifting from a supporting role to a core role. Data formed under the traditional paradigm—discretely distributed, multi-source heterogeneous, small-scale, and non-standardized—cannot effectively interface with AI, constraining the effectiveness of data-driven approaches in the materials field. AI-oriented data governance and new data infrastructure construction have become issues that the materials field must confront.

This paper starts from the principles of AI analysis and systematically proposes the conditions that must be satisfied for constructing AI-ready materials data: massive scale, comprehensiveness, integrity, balance, and shareability, aiming to provide basic reference and direction for data-driven research to construct more and more usable materials data from broader fields.

Standardization is an important foundation for achieving AI-ready materials data and is also a global concern. European and American countries focus on matching existing data, striving to improve the interoperability of multi-source heterogeneous data by establishing complete sets of materials science ontologies, but this metadata coordination approach still requires developing data converters and shared data schemas. China has redefined the construction principles for AI-ready materials data by establishing the "General Rule for Materials Genome Engineering Data". The materials data standardization framework system proposed based on the core concepts of the "General Rule" provides a concrete data governance solution for constructing AI-ready materials data ecosystems. Regardless of the approach taken, materials data standardization is imperative but remains a long and arduous task.

The "Data Factory"new data infrastructure is an ideal venue for comprehensively

constructing AI-ready databases and will continuously provide massive, comprehensive, complete, balanced, and shareable AI-ready standardized data for the materials research field. When the day comes that the "Data Factory" becomes the primary form of data production, the potential of data-driven approaches will hopefully be truly unleashed.

Author Contributions: WANG Hong, ZHANG Lanting: Research proposition formulation; LU Yongchao: Research scheme design, research data collection, paper drafting; YU Ning: Article revision suggestions; WANG Hong, ZHANG Lanting: Article quality control, argumentation; LU Yongchao, WANG Hong: Responsible for final revised version.

[1] White House Office of Science and Technology Policy. Materials genome initiative for global competitiveness [EB/OL]. (2011-06). https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/materials_{{{genome}}}_{{initiativ
final.pdf
[2] Seventh framework programme of the European Community (EU FP7). ACCMET - Accelerated Metallurgy - the accelerated discovery of alloy formulations using combinatorial principles [EB/OL]. (2011-06). https://cordis.europa.eu/project/id/263206/reporting
[3] The Materials Science and Engineering Expert Committee (MatSEEC) of the European Science Foundation(EFS). Metallurgy Europe: a renaissance programme for 2012–2022 [R]. Strasbourg: EFS, 2012
[4] The Center for "Materials research by Information Integration" (CMI2) of MaDIS, NIMS "Materials research by Information Integration" Initiative (MI2I) [EB/OL]. (2015-07). https://www.nims.go.jp/MII-I/en/about/index_m.html
[5] Wang H, Xiang Y, Xiang X D, et al. Materials genome enables research and development revolution [J]. Sci. Technol. Rev., 2015, 33(10): 13 (汪洪, 向勇, 项晓东等. 材料基因组——材料研发新模式 [J]. 科技导报, 2015, 33(10):13)
[6] Su Y J, Fu H D, Bai Y, et al. Progress in Materials Genome Engineering in China [J]. Acta. Metall. Sin., 2020, 56(10): 1313 (宿彦京, 付华栋, 白洋等. 中国材料基因工程研究进展 [J]. 金属学报, 2020, 56(10): 1313)
[7] Wang H, Xiang X D, Zhang L T. Data + AI: The core of materials genomic engineering [J]. Sci. Technol. Rev., 2018, 36(14): (汪洪, 项晓东, 张澜庭. 数据 + 人工智能是材料基因工程的核心 [J]. 科技导报, 2018, 36(14): 15)
[8] Lian S Y. Introduction to artificial intelligence [M]. Beijing: Tsinghua University Press, 2020: 3 (廉师友. 人工智能导论 [M]. 北京: 清华大学出版社, 2020: 3)
[9] Liu F Q. Artificial intelligence [M]. Beijing: China machine Press, 2011: 1 (刘凤岐. 人工智能 [M]. 北京: 机械工业出版社, 2011: 1)
[10] Kaplan A, Haenlein M. Siri, Siri, in my hand: Who' s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence [J]. Bus. Horiz., 2019, 62(1): 15
[11] Warren J A. The materials genome initiative and artificial intelligence [J]. MRS Bull., 2018, 43(6): 452
[12] Murdock R J, Kauwe S K, Wang A Y T, et al. Is domain knowledge necessary for machine learning materials properties? [J]. Integr. Mater. Manuf. Innov., 2020, 9(3): 221-227.

[13] Masood H, Toe C Y, Teoh W Y, et al. Machine learning for accelerated discovery of solar photocatalysts [J]. ACS Catal., 2019, 9(12): 11774-11787.
[14] Childs C M, Washburn N R. Embedding domain knowledge for machine learning of complex material systems [J]. MRS Commun., 2019, 9(3): 806-820.
[15] Liu Z H，Wang T Y. Data governance [M]. Beijing: Party School of the CPC Central Committee Press, 2021: 1 (李振华，王同益. 数据治理 [M]. 北京: 中共中央党校出版社, 2021: 1)
[16] Fagnan K, Nashed Y, Perdue G, et al. Data and models: a framework for advancing AI in science [R]. United States: USDOE Office of Science (SC), 2019
[17] National Science and Technology Council, Committee on Technology and Subcommittee on the MGI Initiative. Materials genome initiative strategic plan [EB/OL]. (2021-11). https://www.mgi.gov/sites/default/files/documents/MGI-2021-Strategic-Plan.pdf
[18] Ghahramani Z. Probabilistic machine learning and artificial intelligence [J]. Nature, 2015, 521(7553): 452
[19] Schmidt J, Shi J, Borlido P, et al. Predicting the thermodynamic stability of solids combining density functional theory and machine learning [J]. Chem. Mat., 2017, 29(12): 5090
[20] Lee J, Seko A, Shitara K, et al. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques [J]. Phys. Rev. B, 2016, 93(11): 115104.
[21] Zhou J, Hong X, Jin P. Information fusion for multi-source material data: progress and challenges [J]. Appl. Sci, 2019, 9(17):
[22] Kalidindi S R, De Graef M. Materials data science: current status and future outlook[J]. Ann. Rev. Mater. Res., 2015, 45: 171
[23] Schmidt J, Marques M R G, Botti S, et al. Recent advances and applications of machine learning in solid-state materials science [J]. npj. Comput. Mater., 2019, 5(1): 1
[24] Schmidt J, Shi J, Borlido P, et al. Predicting the thermodynamic stability of solids combining density functional theory and machine learning [J]. Chem. Mat., 2017, 29(12): 5090
[25] De Jong M, Chen W, Notestine R, et al. A statistical learning framework for materials science: application to elastic moduli of k-nary inorganic polycrystalline compounds [J]. Sci. Rep., 2016, 6(1): 1
[26] Kim J, Kang D, Kim S, et al. Catalyze materials science with machine learning [J]. ACS Mater. Lett., 2021, 3(8): 1151-1171.
[27] Liu Y H, Hu Z H, Suo Z G, et al. High-throughput experiments facilitate materials innovation: a review [J]. Sci. China-Technol. Sci., 2019, 62(4): 521
[28] Curtarolo S, Hart G L W, Nardelli M B, et al. The high-throughput highway to computational materials design [J]. Nat. Mater., 2013, 12(3): 191
[29] Hattrick-Simpers J R, Gregoire J M, Kusne A G. Perspective: composition–structure–property mapping in high-throughput experiments: turning data into knowledge [J]. APL. Mater., 2016, 4(5): 053211
[30] Xiang X, Sun X, Briceño G, Lou Y, WangK, Chang H, et al. A combinatorial approach to materials discovery [J]. Science 1995,268(5218):1738

[31] Ceder G, Persson K. The stuff of dreams [J]. Sci. Am., 2013, 309(6): 36
[32] Zhang X, Chen A, Zhou Z. High-throughput computational screening of layered and two-dimensional materials [J]. Wires Comput. Mol. Sci., 2019, 9(1): e1385.
[33] Brunin G, Ricci F, Ha V A, et al. Transparent conducting materials discovery using high-throughput computing [J]. npj. Comput. Mater., 2019, 5(1): 1
[34] Kononova O, He T, Huo H, et al. Opportunities and challenges of text mining in materials research [J]. Iscience., 2021, 24(3):
[35] Blokhin E, Villars P. The PAULING FILE project and materials platform for data science: From big data toward materials genome [A]. Handbook of Materials Modeling-Methods: Theory and Modeling [C], Cham: Springer, 2020: 1837
[36] Swain M C, Cole J M. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature [J]. J. Chem. Inf. Model., 2016, 56(10): 1894
[37] Court C J, Cole J M. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction [J]. Sci. Data., 2018, 5(1): 1
[38] Huang S, Cole J M. A database of battery materials auto-generated using ChemDataExtractor [J]. Sci. Data., 2020, 7(1): 1
[39] Moosavi S M, Jablonka K M, Smit B. The role of machine learning in the understanding and design of materials [J]. J. Am. Chem. Soc., 2020, 142(48): 20273
[40] Ward L, Agrawal A, Choudhary A, et al. A general-purpose machine learning framework for predicting properties of inorganic materials [J]. npj Comput. Mater., 2016, 2(1): 1
[41] Ghiringhelli L M, Baldauf C, Bereau T, et al. Shared metadata for data-centric materials science [J]. arXiv, 2022: 2205.14774
[42] Baker M. 1,500 scientists lift the lid on reproducibility [J]. Nature, 2016, 533(7604)
[43] Editorial. The importance and challenges of data sharing [J]. Nat. Nanotechnol., 2020(15):83
[44] Thelwall M, Kousha K. Figshare: a universal repository for academic resource sharing? [J]. Online Inf. Rev., 2016
[45] Dillen M, Groom Q, Agosti D, et al. Zenodo, An archive and publishing repository: a tale of two herbarium specimen pilot projects[J]. Biodivers. Inf. Sci. Stand, 2019 (2)
[46] White H, Carrier S, Thompson A, et al. The Dryad Data Repository: A Singapore Framework Metadata Architecture in a DSpace Environment[A]. Proceedings of the International Conference on Dublin Core and Metadata Applications [C]. Göttingen: Universitätsverlag Göttingen, 2008:157
[47] Krawczyk B. Learning from imbalanced data: open challenges and future directions [J]. Prog. Artif. Intell., 2016, 5(4): 221
[48] Weissman J. Amazon created a hiring tool using aI it immediately started discriminating against women [J]. Slate, 2018

[49] Davenport T, Guha A, Grewal D, et al. How artificial intelligence will change the future of marketing [J]. J. Acad. Mark. Sci., 2020, 48(1): 24
[50] Raccuglia P, Elbert K C, Adler P D F, et al. Machine-learning-assisted materials discovery using failed experiments [J]. Nature, 2016, 533(7601): 73
[51] White House Office of Management and Budget. Federal data strategy 2020 action plan [EB/OL]. (2019-12). https://strategy.data.gov/assets/docs/2020-federal-data-strategy-action-plan.pdf
[52] European Commission. A European strategy for data [EB/OL]. (2020-2). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0066&from=EN
[53] General Office of the State Council of the People' s Republic of China. Scientific Data Management Measures [EB/OL]. (2018-3). https://www.cgs.gov.cn/xwl/zfw/201804/W020180403526880358641.pdf (中华人民共和国国务院办公厅. 科学数据管理办法 [EB/OL]. (2018-3). https://www.cgs.gov.cn/xwl/zfw/201804/W020180403
)
[54] Kozlov M. NIH issues a seismic mandate: share data publicly [J]. Nature, 2022,602(7898):558
[55] Mauthner N S, Parry O. Open access digital data sharing: Principles, policies and practices [J]. Soc. Epistemol., 2013, 27(1): 47
[56] Tenopir C, Dalton E D, Allard S, et al. Changes in data sharing and data reuse practices and perceptions among scientists worldwide [J]. PloS one, 2015, 10(8): e0134826
[57] Wilkinson M D, Dumontier M, Aalbersberg I J J, et al. The FAIR Guiding Principles for scientific data management and stewardship [J]. Sci. Data., 2016, 3(1): 1
[58] Berman F. The research data alliance–the first five years [EB/OL]. (2019). https://www.rd-alliance.org/sites/default/files/attachment/RDA%20RETROSPECTIVE%20FINAL%20-%20HDSR.pdf
[59] Mons B, Neylon C, Velterop J, et al. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud [J]. Inf. Serv. Use, 2017, 37(1): 49
[60] Liu J. Digital Object Identifier (DOI) and DOI Services: An Overview[J]. Libri, 2021, 71(4): 349
[61] State General Administration of the People' s Republic of China for Quality Supervision and Inspection and Quarantine, Standardization Administration of China. GB/T 20000.1-2014 Guidelines for standardization—Part 1: Standardization and related actives—General vocabulary[S]. Beijing: Standards Press of China, 2014 (中华人民共和国国家质量监督检验检疫总局、中国国家标准化管理委员会. GB/T20000.1-2014, 标准化工作指南第 1 部分：标准化和相关活动的通用术语 [S]. 北京: 中国标准出版社, 2014)
[62] Himanen L, Geurts A, Foster A S, et al. Data-driven materials science: status, challenges, and perspectives [J]. Adv. Sci., 2019, 6(21): 1900808
[63] Scheffler M, Aeschlimann M, Albrecht M, et al. FAIR data enabling new horizons for materials research [J]. Nature, 2022, 604(7907): 635
[64] Studer R, Benjamins V R, Fensel D. Knowledge engineering: principles and methods [J]. Data Knowl. Eng., 1998, 25(1-2): 161
[65] Yang T. Research on some key technologies of agricultural knowledge

service based on ontology [D]. Shanghai: Fudan University, 2011 （杨涛. 基于本体的农业领域知识服务若干关键技术研究 [D]. 上海: 复旦大学,2011.）
[66] Zhang X, Zhao C, Wang X. A survey on knowledge representation in materials science and engineering: An ontological perspective [J]. Comput. Ind., 2015, 73: 8
[67] Ghiringhelli L M, Carbogno C, Levchenko S, et al. Towards efficient data exchange and sharing for big-data driven materials science: metadata and data formats [J]. npj Comput. Mater., 2017, 3(1): 1
[68] Tadmor E B, Elliott R S, Sethna J P, et al. The potential of atomistic simulations and the knowledgebase of interatomic models [J]. Jom, 2011, 63(7): 17.
[69] van der Vet P E, Speel P H, Mars N J I. The Plinius ontology of ceramic materials [A]. Eleventh European Conference on Artificial Intelligence (ECAI' 94) Workshop on Comparison of Implemented Ontologies [C]. New York: John Wiley & Sons, 1994: 187
[70] Sainte Marie C, Iglesias Escudero M, Rosina P. The ONTORULE project: where ontology meets business rules [A]. International Conference on Web Reasoning and Rule Systems [C]. Berlin: Springer, 2011: 24
[71] Premkumar V, Krishnamurty S, Wileden J C, et al. A semantic knowledge management system for laminated composites [J]. Adv. Eng. Inform., 2014, 28(1): 91
[72] Michel K, Meredig B. Beyond bulk single crystals: a data format for all materials structure–property–processing relationships [J]. Mrs Bull., 2016, 41(8): 617
[73] Ashino T. Materials ontology: An infrastructure for exchanging materials information and knowledge [J]. Data Sci. J., 2010, 9:
[74] European Materials Modelling Council. EMMO: an Ontology for Applied Sciences [EB/OL]. (2017). https://emmc.info/emmo-info/
[75] Cheung K, Drennan J, Hunter J. Towards an Ontology for Data-driven Discovery of New Materials [A]. AAAI Spring Symposium: Semantic Scientific Knowledge Integration [C]. Menlo Park: The AAAI Press, 2008: 9
[76] Bhat M, Shah S, Das P, et al. Premλp: knowledge driven design of materials and engineering process [A]. ICoRD' 13 international conference on research into design [C]. India: Springer, 2013: 1315
[77] Zhang X, Hu C, Li H. Semantic query on materials data based on mapping MATML to an OWL ontology [J]. Data Sci. J., 2009,
[78] Ramakrishna S, Zhang T Y, Lu W C, et al. Materials informatics [J]. J. Intell. Manuf., 2019, 30(6): 2307
[79] Andersen C W, Armiento R, Blokhin E, et al. OPTIMADE, an API for exchanging materials data [J]. Sci. Data, 2021, 8(1): 1
[80] China Standards of Testing and Materials (CSTM). Announcement on the establishment of the CSTM standard "materials genome terminology". [EB/OL]. (2021-11), http://www.cstm.com.cn/article/details/8a276faa-f242-4c3b-8bf1-4fb238af8ef3 (中国材料与试验团体标准委员会. 关于 CSTM 标准《材料基因工程术语》的立项公告. [EB/OL]. (2021-11), http://www.cstm.com.cn/article/details/8a276faa-f242-4c3b-8bf1-4fb238af8ef3)

[81] China Standards of Testing and Materials (CSTM). Announcement on the establishment of the CSTM standard "Data Identifier Naming Method for Materials Genome Engineering". [EB/OL]. (2021-11), http://www.cstm.com.cn/article/details/b356a5f0-a75e-4671-8b7a-f95f95351ade (中国材料与试验团体标准委员会. 关于 CSTM 标准《材料基因工程数据标识符命名方法》的立项公告. [EB/OL]. (2021-11), http://www.cstm.com.cn/article/details/b356a5f0-a75e-4671-8b7a-f95f95351ade)
[82] China Standards of Testing and Materials (CSTM). Announcement on the establishment of the CSTM standard "General Metadata Specification for Materials Genome Engineering Data". [EB/OL]. (2021-11), http://www.cstm.com.cn/article/details/69bfbb17-e88e-481a-bd16-756cdbc969cd (中国材料与试验团体标准委员会. 关于 CSTM 标准《材料基因工程数据通用元数据规范》的立项公告. [EB/OL]. (2021-11), http://www.cstm.com.cn/article/details/69bfbb17-e88e-481a-bd16-756cdbc969cd)
[83] Taillon J A, Bina T F, Plante R L, et al. NexusLIMS: A laboratory information management system for shared-use electron microscopy facilities [J]. Microsc. microanal., 2021, 27(3): 511
[84] Jain A, Ong S P, Chen W, et al. FireWorks: A dynamic workflow system designed for high-throughput applications [J]. Concurr. Comput., 2015, 27(17): 5037
[85] Supka A R, Lyons T E, Liyanage L, et al. AFLOWπ: A minimalist approach to high-throughput ab initio calculations including the generation of tight-binding hamiltonians [J]. Computational Materials Science, 2017, 136: 76
[86] Mathew K, Montoya J H, Faghaninia A, et al. Atomate: A high-level interface to generate, execute, and analyze computational materials science workflows [J]. Comput. Mater. Sci., 2017, 139: 140
[87] Pizzi G, Cepellotti A, Sabatini R, et al. AiiDA: automated interactive infrastructure and database for computational science [J]. Comput. Mater. Sci., 2016, 111: 218
[88] Rao Y C, Lu Y C, Ju S H, et al. Metadata standard for phonon thermal conductivity by first-principles calculation [J]. J. Mater. Inf., 2022, (In-revision)
[89] Chinese Society fot Testing & Materials, Zhongguancun. T/CSTM00120-2019 General rule for materials genome engineering data [S]. Beijing: Metallurgical Industry Press, 2019 (中关村材料试验技术联盟. T/CSTM00120-2019, 材料基因工程数据通则 [S]. 北京: 冶金工业出版社, 2019)
[90] State General Administration of the People' s Republic of China for Quality Supervision and Inspection and Quarantine, Standardization Administration of China. GB/T232843-2016 Science and technology resource identification [S]. Beijing: Standards Press of China, 2016 (中华人民共和国国家质量监督检验检疫总局、中国国家标准化管理委员会. GB/T232843-2016, 科技资源标识 [S]. 北京: 中国标准出版社, 2016)
[91] Wang H, Xiang X D, Zhang L T. On the data-driven materials innovation infrastructure [J]. Engineering., 2020, 6(6): 609
[92] Jain A, Ong S P, Hautier G, et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation [J]. APL. Mater., 2013, 1(1): 011002

[93] Curtarolo S, Setyawan W, Hart G L W, et al. AFLOW: An automatic framework for high-throughput materials discovery [J]. Comput. Mater. Sci., 2012, 58: 218
[94] Kirklin S, Saal J E, Meredig B, et al. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies [J]. npj Comput. Mater., 2015, 1(1): 1
[95] Draxl C, Scheffler M. NOMAD: The FAIR concept for big data-driven materials science [J]. MRS. Bull., 2018, 43(9): 676
[96] Yang X, Wang Z, Zhao X, et al. MatCloud: A high-throughput computational infrastructure for integrated management of materials simulation, data and resources [J]. Comput. Mater. Sci., 2018, 146: 319
[97] Talirz L, Kumbhar S, Passaro E, et al. Materials Cloud, a platform for open computational science [J]. Sci. Data, 2020, 7(1): 1
[98] Zakutayev A, Wunder N, Schwarting M, et al. An open experimental database for exploring inorganic materials [J]. Sci. Data, 2018, 5(1): 1

Note: Figure translations are in progress. See original paper for figures.