Wang *et al. J Mater Inf* 2022;2:1
DOI: 10.20517/jmi.2021.11

Journal of Materials Informatics

Review

Open Access

Check for updates

Big data-assisted digital twins for the smart design and manufacturing of advanced materials: from atoms to products

William-Yi Wang[1,7], Junlei Yin[1], Zaixian Chai[1], Xin Chen[2], Wenping Zhao[3], Jiaqi Lu[1], Feng Sun[4], Qinggong Jia[4,8], Xingyu Gao[2], Bin Tang[1,7], Xidong Hui[5], Haifeng Song[2,*], Fei Xue[1,*], Zi-Kui Liu[6], Jinshan Li[1,7,*]

[1]State Key Laboratory of Solidification Processing, Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China.
[2]CAEP Software Center for High Performance Numerical Simulation & Institute of Applied Physics and Computational Mathematics, Beijing 100088, China.
[3]CRRC Tangshan Co., LTD, Tangshan 063035, Hebei, China.
[4]Western Superconducting Technologies Co., Ltd., Xi'an 710018, Shaanxi, China.
[5]State Key Laboratory for Advanced Metals and Materials, University of Science and Technology Beijing, Beijing 100083, China.
[6]Department of Materials Science and Engineering, The Pennsylvania State University, University Park, PA 16802, USA.
[7]Innovation Center, NPU Chongqing, Chongqing 401135, China.
[8]School of Materials Science and Engineering, Xi'an University of Technology, Xi'an 710048, Shaanxi, China.

***Correspondence to:** Jinshan Li, State Key Laboratory of Solidification Processing, Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China. E-mail: ljsh@nwpu.edu.cn; Haifeng Song, CAEP Software Center for High Performance Numerical Simulation & Institute of Applied Physics and Computational Mathematics, Beijing 100088, China. E-mail: song_haifeng@iapcm.ac.cn; Fei Xue, State Key Laboratory of Solidification Processing, Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China. E-mail: 13913573200@139.com

How to cite this article: Wang WY, Yin J, Chai Z, Chen X, Zhao W, Lu J, Sun F, Jia Q, Gao X, Tang B, Hui X, Song H, Xue F, Liu ZK, Li J. Big data-assisted digital twins for the smart design and manufacturing of advanced materials: from atoms to products. *J Mater Inf* 2022;2:1. https://dx.doi.org/10.20517/jmi.2021.11

Received: 2 Nov 2021 **First Decision:** 27 Dec 2021 **Revised:** 11 Jan 2022 **Accepted:** 10 Feb 2022 **Published:** 23 Feb 2022

Academic Editors: Xingjun Liu, Tong-Yi Zhang **Copy Editor:** Xi-Jun Chen **Production Editor:** Xi-Jun Chen

Abstract

Motivated by the ever-increasing wealth of data boosted by national strategies in terms of data-driven Integrated Computational Materials Engineering (ICME), Materials Genome Engineering, Materials Genome Infrastructures, Industry 4.0, Materials 4.0 and so on, materials informatics represents a unique strategy in revealing the fundamental relationships in the development and manufacturing of advanced materials. Materials developments are becoming ever more integrated with robust data-driven and data-intensive technologies. In the present review, big data-assisted digital twins (DTs) for the smart design and manufacturing of advanced materials are presented from the perspective of the digital thread. In the introduction of the DT design paradigm in the ICME era, the

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

simulation aspects of DT and the data and design infrastructures are discussed. Referring to the simulation and theoretical factors of DTs, high-throughput simulation and automation and artificial intelligence-assisted multiscale atomistic modeling are detailed through several cases studies. With respect to data and data mining technologies, entropy and its application for attribute selection in decision trees are discussed to emphasize knowledge-based modeling, simulation and data analysis in machine learning coherently. Guided by the perspectives and case studies of the digital thread, we present our recent work on the design, manufacturing and product service via big data-assisted DTs for smart design and manufacturing by integrating some of these advanced concepts and technologies. It is believed that big data-assisted DTs for smart design and manufacturing effectively support better products with the application of novel materials by reducing the time and cost of materials design and deployment.

Keywords: ICME, Materials Genome Engineering, high-throughput, automation, workflow, data mining, digital thread

INTRODUCTION

With the dramatic development of advanced technologies, the ever-increasing wealth of data from computations and experiments is considered an essential component in the modern innovation ecosystem of materials[1-7]. As an important output of high-throughput (HT) and high-performance computing, the emergence of "big data" and the "ocean of data" is constructing an ecosystem that represents a significant research opportunity, as well as stimulating new demand for data-intensive capabilities[7,8]. Data-driven research, as well as modeling, simulation and advanced fabrication, has become a smart design/manufacturing paradigm within the so-called "age of design", which generally combines several advanced technologies, including HT computations, data mining, machine/deep learning, artificial intelligence (AI) and additive manufacturing[6,9-16]. Moreover, motivated by worldwide strategies to accelerate materials discovery and deployment, such as Integrated Computational Materials Engineering (ICME) and the Materials Genome Initiative (MGI) in the United States[17,18], Materials Genome Engineering (MGE) and the Human-Cyber-Physical Systems (HCPSs) of China[19], Industry 4.0 in Germany[20] and so on[18], both data and design infrastructures have been boosted substantially. Materials informatics can survey complex and multiscale information in a HT, statistically robust and yet physically meaningful manner[3,21,22], thereby supporting the unique strategy of revealing the composition-processing-structure-property-performance (CPSPP) relationship in the development of advanced materials.

It is noteworthy that materials development must be integrated with manufacturing, quality control and automation, verification and validation, materials synthesis, processing, characterization and property measurements[23]. For instance, digital technologies have been considered as essential aspects in improving the competitiveness of United States manufacturing[9,24]. The target of the Digital Manufacturing and Design Innovation Institute (DMDII) is to be a preeminent worldwide organization for digitizing data across all processes of the lifecycle period and integrating them to yield better solutions and decisions[9,24]. Based on the DMDII's five-year cooperative agreement, some extremely important progress has been made, including (1) improving the operations among organizations through digital manufacturing; (2) accelerating innovations in digital technologies; (3) multi-party collaboration to enable innovative solutions; and (4) solving the "valley of death" problem in digital manufacturing technologies[9,24]. Moreover, it has been reported that "in North America, big manufacturers have spent almost $7 trillion retrofitting old equipment with sensors that allow systems to talk to each other, but that investment only helps them use about 1 percent of their operational data in making business decisions. The European market is leading the way in digital and automated manufacturing"[24].

Accordingly, two perspectives can be found in the new vision for UK manufacturing by 2050[25]. On the one hand, successful firms will be capable of rapidly adapting their physical and intellectual infrastructures to exploit changes in technology as manufacturing becomes faster, more responsive to changing global markets and closer to customers[25]. On the other hand, physical production will play a dominate role in accelerating future developments, creating innovative new revenue streams and increasing the pervasiveness of big data, which will also depend on big data to enhance the competitiveness of the firms and manufacturers[25]. Date-intensive science and technology represent a highly effective combination for addressing upcoming challenges and supporting significant opportunities in innovation and evolution[12,26,27].

In parallel with smart/intelligent manufacturing, the digital thread of manufacturing has been considered and emphasized[5]. In particular, the concept of the digital thread has been utilized for the development of advanced materials, with an emphasis on rapid fielding, the development, employment and integration of digital design tools covering the essential requirements in a lifecycle[28]. The digital thread is the creation and use of a digital surrogate of a material system that allows for a dynamic and real-time assessment of the system's current and future capabilities to inform decisions in the capability planning and analysis, preliminary design, detailed design, manufacturing and sustainment acquisition phases[28]. In contrast, the digital surrogate is a physics-based technical description of the system resulting from the generation, management and application of data, models and information from authoritative sources across the system's lifecycle[28], revealing the physical nature of digital twins through the cyber-physical systems/interactions. By integrating advanced technologies, such as HT and high-performance computations, modeling, data storage and data mining, the digital thread would have the capability to carry out informed decision making at key leverage points in the development process that have the largest impact on acquisition programs[28]. Moreover, an earlier identification and a broader range of feasible solutions can be obtained, including a structured assessment of cost, schedule and performance risk and accelerated analysis, development, testing and operation[28].

Our recent reviews[18,29] have highlighted the dominant roles of HT computation and automated software, benchmarks, databases and platforms, and the principles and standards in the framework of data-driven ICME. Case studies of data-driven ICME for intelligently discovering and fabricating advanced materials and products for high-temperature applications were presented[29]. Perspectives on knowledge-based modeling/simulations, the machine learning knowledge base, platforms and next-generation workforces for a sustainable ecosystem of ICME were highlighted. More efforts are needed to work on the development of advanced structural metal materials and the enhancement of research productivity and collaboration[29].

In the present work, big data-assisted digital twins for the smart design and manufacturing of advanced materials is reviewed in terms of the digital thread. The digital twin design paradigm in the ICME era is first introduced, in which the sub-sections of the simulation aspects of digital twins and the data and design infrastructures are discussed. The data and design infrastructures are then briefly discussed to highlight the rules, principles and standards for the successful achievements of the MGI, MGE and ICME. Next, the simulation aspects of digital twins are highlighted, in which the sub-sections of HT simulation and automation and AI-assisted multiscale atomistic modeling support are studied in detail through the use of several case studies. Data and data mining technologies are discussed in the context of providing tools and aspects to integrate scientific information and theory for materials discovery. Entropy and its application for attribute selection in decision trees are discussed to highlight knowledge-based modeling, simulation and even data analysis in machine learning in one unique term. Perspectives and case studies of the digital thread in Materials 4.0 and Industry 4.0 are then provided, followed by a brief discussion of the outlook in the conclusions.

DIGITAL TWIN DESIGN PARADIGM IN ICME ERA

A digital twin is an integrated multi-physics, multi-scale, probabilistic simulation of a complex system that uses the best available physical models, sensor updates, recorded/reported data and so on to mirror the life of its corresponding twin[28,30]. From their first release in NASA's Apollo program in 2003[31], digital twins have been considered important tools for realizing the real-time interaction and integration between information and the physical world, and represent a key enabling technology for achieving cyber-physical integration for smart assembly/manufacturing[30], i.e., for aircraft, trains and engines[18,24]. Based on the traditional procedure in designing or manufacturing advanced materials, the CPSPP relationship consisting of materials composition, processing, microstructure, properties and performance requires the integration of the aforementioned advanced technologies, such as HT and high-performance computations, data mining, machine/deep learning, artificial intelligence, additive manufacturing and so on[9]. From the perspective of bottom-up design and top-down engineering, as presented in Figure 1, the twin features between experimental and theoretical chains are highlighted by two types of arrows with different background colors. While the MGI highlights the essential role of toolkits consisting of HT experiments and computations and the databases together with their interactions, the HCPs and MGE emphasize the interactions of human-cyber-physical systems[9]. Similarly, Industry 4.0 has highlighted that intelligent manufacturing is enabled through cyber-physical systems[32].

Figure 2 displays the overall hierarchical architecture of the methods, tools, techniques and databases for the applications of ICME methods based on the MGI and MGE[33]. It is understood that the infrastructures integrating HT experimental and computational/theoretical data across material classes are highlighted in the MGI strategic plan to make experimental and computational data accessible, sharable and transformable[4]. Such a materials data infrastructure with the guidance of the MGI and MGE will enable ICME approaches to be deployed with greater success and efficiency and enable the ultimate goals of the MGI to be achieved[4]. In line with the CPSPP relationship, process-structure and structure-property models highlight the significant role of materials knowledge in the transformation from data and informatics to manufacturing, which also indicate the high-value chain toward manufacturing[4]. With increased technology readiness levels, data-driven HT strategies are required to accelerate materials innovations, which are still under development in the field of structural materials[4]. By combining a hierarchical architecture of ICME on the basis of MGI infrastructures, the ICME/ICMD mechanistic design models accelerate innovation, transferring studio ideation into industrial manufacturing[9].

Moreover, the Extensible Self-optimizing Phase Equilibria Infrastructure (ESPEI) software efficiently evaluates the thermodynamic model parameters within the CALculation of PHAse Diagrams (CALPHAD) method[34,35]. It is believed to be an essential part in contributing to the establishment of the "ocean of data" and developing the property database of multi-component materials with multiple defects[7]. Furthermore, computational thermodynamics enable the modeling of the thermodynamics of a state as a function of both external and internal variables and quantitative calculations of a broad range of properties of a multicomponent system in terms of the first and second derivatives of energy at the equilibrium and non-equilibrium states for internal processes[36]. The CALPHAD fundamental database and software system paved the way for the birth of the Materials Genome[37-39]. It is an important methodology that integrates both experimental and theoretical results in a thermodynamic language and is applicable to a much wider context than the original experiments or calculations[37], as well as presenting the digital twin feature by combining the digital thread and surrogate.

As presented in Figure 3, the integration of DFT calculations and CALPHAD has established a robust materials development framework for data-driven ICME[18,29]. HT first-principles calculations can be utilized

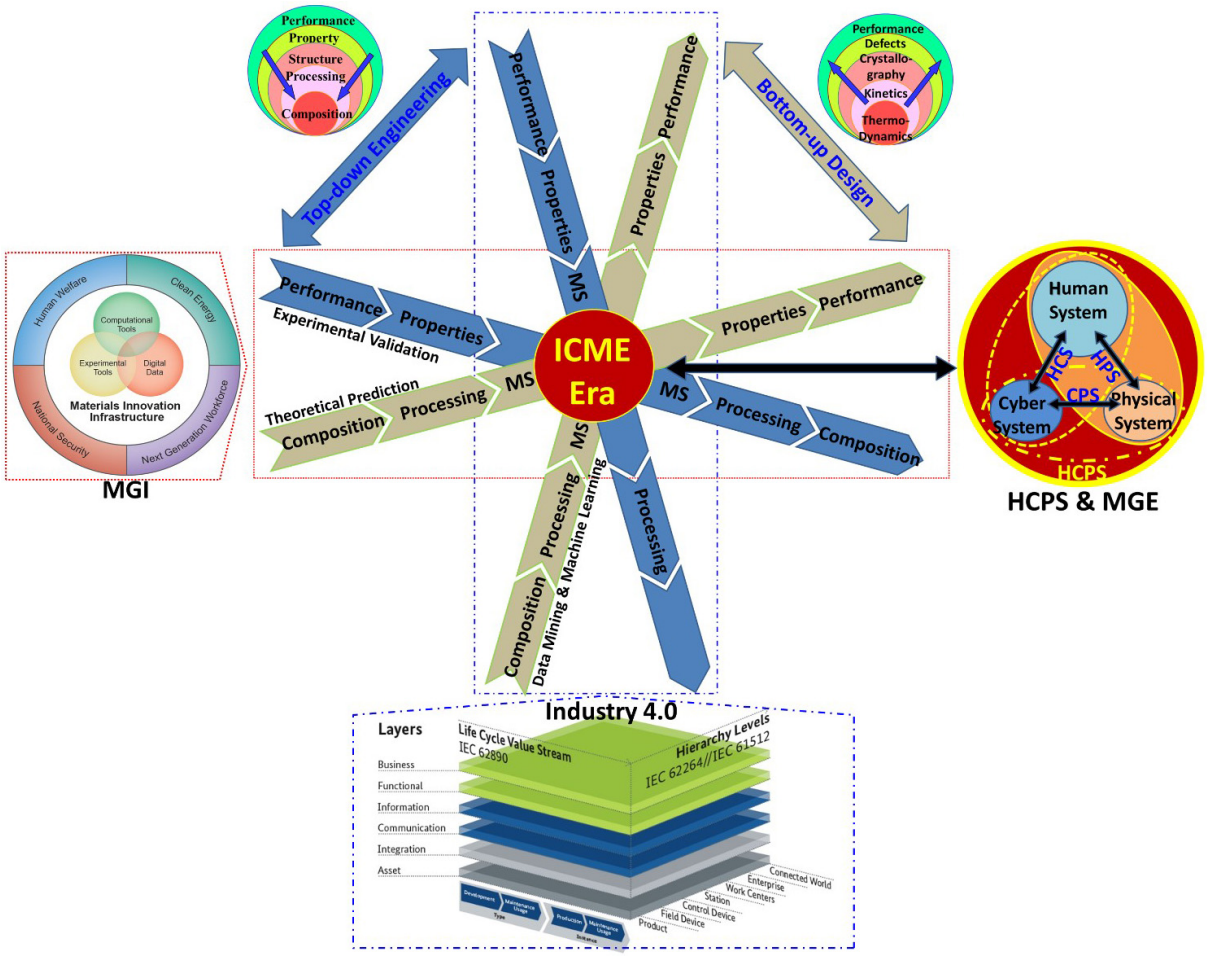

Figure 1. Schematic of digital twin design paradigm in the ICME era, with reference to the MGI, HCPS, MGE and Industry 4.0 national strategies[9]. Reproduced from Ref.[9] with permission from Elsevier. ICME: Integrated Computational Materials Engineering; MGI: Materials Genome Initiative; HCPS: Human-Cyber-Physical System; MGE: Materials Genome Engineering.

to predict thermodynamic properties combined with CALPHAD, kinetic properties, mechanical properties and many other fundamental physical properties[29,40,41]. Consequently, external constraints, such as fixed strain and internal degrees of freedom, including ordering and defects, can be described in a coherent framework and applied to materials design[36]. As shown in Figure 3, these properties also have further contributions to multiscale modeling or computations, together with the corresponding crossover experimental validations, thereby boosting the various kinds of databases and constructing one of the key Material Genome infrastructures. With support from the foundations and milestones of the MGI, MGE and ICME, recent progress in emphasizing the importance of computational materials/platforms/systems for the future has been completed[42-45]. In the investigation of the CPSPP relationship, the flow chain of "data-cyber-knowledge-wisdom" displays the inheritable feature of the data in the frameworks of the Materials Genome[29], calling for more efforts to work on data-driven and data-intensive technologies.

DATA AND DESIGN INFRASTRUCTURES

Data and data infrastructures have been treated as the dominant foundations for the successful achievements of the MGI, MGE, ICME and Inheritable Integrated intelligent Manufacturing (I^3M)[9]. The letter "I" in these strategies indicates "informatics", which is a broad term that encompasses data-driven

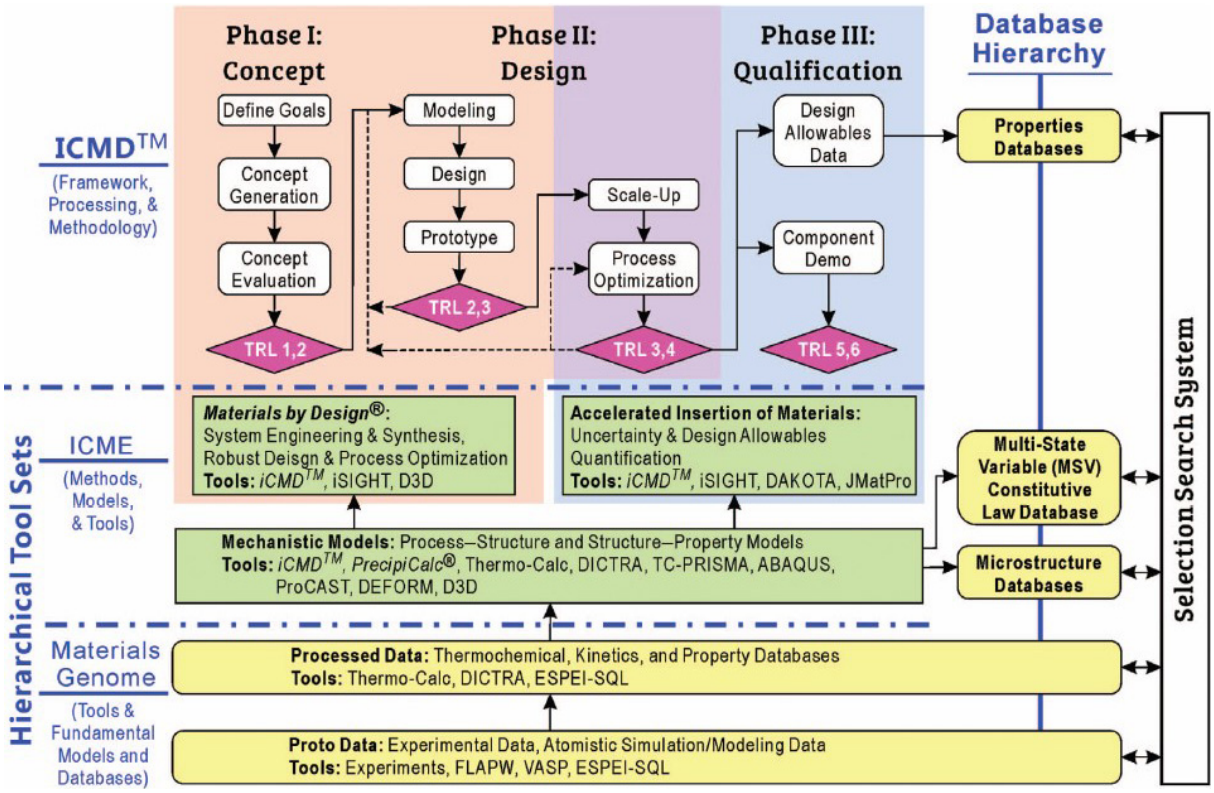

Figure 2. Overall hierarchical architecture of methods, tools, techniques and databases for applications of ICME methods based on the MGI and MGE[33]. ICME: Integrated Computational Materials Engineering; MGI: Materials Genome Initiative; MGE: Materials Genome Engineering.

design stages[9,46]. In order to make integrated decision-making easier and more accessible, the modeling standards for products and processes, the interface standards for the functional and physical connections between components and the mechanisms for flexible system integration should be additional essential aspects of the data and design infrastructure[8]. In particular, a significant amount of manufacturing data and sophisticated design knowledge are required to be effectively used in the innovation of advanced materials, product technology platforms and advanced manufacturing processes[8]. Data-driven or knowledge-based decisions to achieve this integration can reduce the process time and cost for improvements to be instituted and for products to reach customers[8]. Moreover, the manufacturing initiative should include a number of elements that will strengthen the advanced manufacturing data and design infrastructure[8].

As shown in Figure 4, the "digital data" has been considered as an important foundational element in MGI infrastructures for the envisioned acceleration of materials development and deployment when releasing the MGI strategy[5]. In line with its evolutions, a new interdisciplinary field of study known as Materials Data Science and Informatics has been developed, which focuses on all technical and ecosystems of the data- and cyber-infrastructure needed to streamline the efficient extraction of high-value materials knowledge[5]. Based on advanced statistics and computer/computational sciences, modern data science has already played an important role in many research fields, with the aim of developing novel approaches, algorithms, methods and tools, as well as the associated infrastructures, required to organize and streamline the processes and sub-processes involved in extracting high-value (actionable) information from all available data and resources[4].

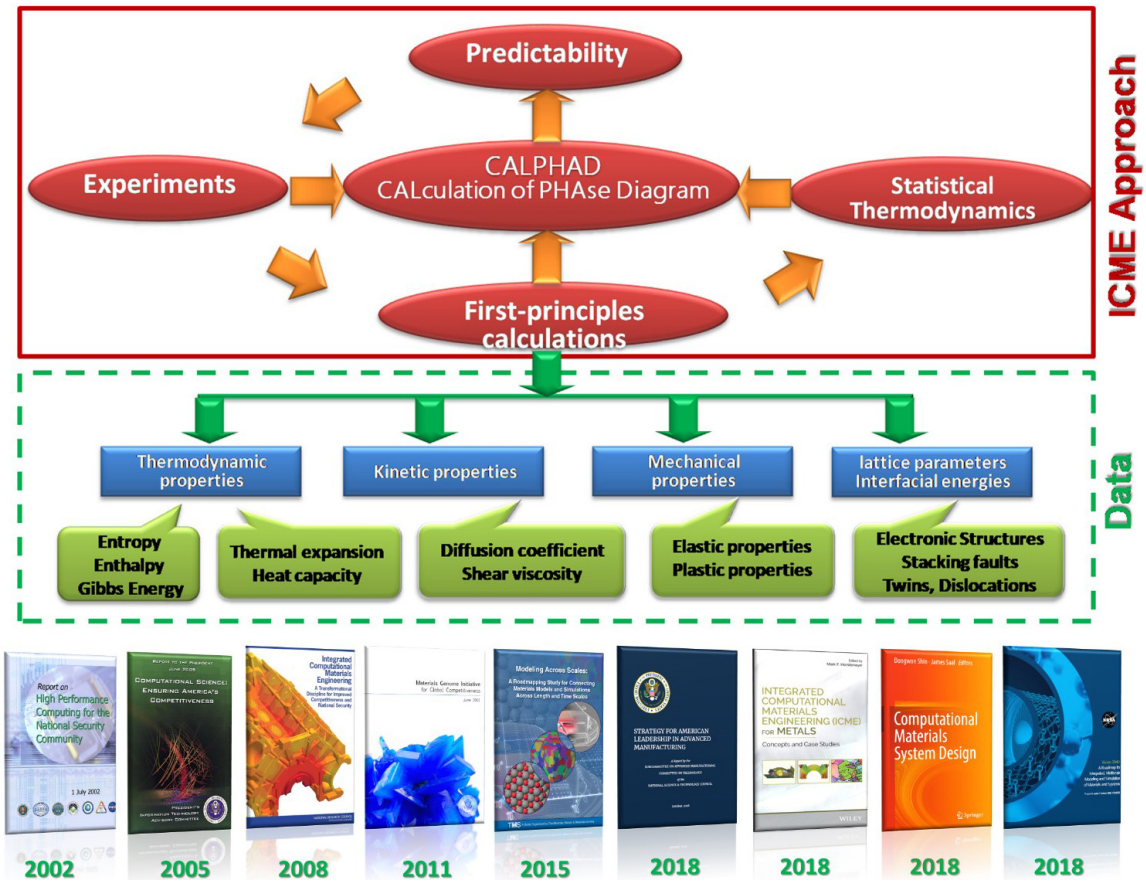

Figure 3. Proposed data-driven ICME approach together with the corresponding foundations and milestones[29]. ICME: Integrated Computational Materials Engineering.

In line with the digital twin design and manufacturing paradigm, the key enabling technologies for digital twin data (DTD) listed in Figure 5 are discussed in detail as follows. The DTD gathering deals with rare-event, extreme environment (i.e., ultra-high temperatures and pressures, high strain rates, deep sea and space, radiation and so on) and multi-physics coupling data, which are difficult to measure or obtain directly. This is why theoretical data generation based on highly efficient simulations plays a crucial role and the combinations of multiple technologies can be used to support data generation[47]. By integrating unified data modeling, database management, data space expansion, data integration, data security and so on, the DTD storage technology dealing with data in different structures, formats, types, encapsulations and interfaces can be conveniently completed[47]. For the data interaction technology, the data collection technology is required to connect data from different parts of DTD by various means, such as sensors and data crawling[47]. The data association technology combines data mining, spatial-temporal data alignment, knowledge reasoning and representation and so on[47]. The data fusion technology is mainly integrated by anomaly detection, granularity transformation, heterogeneous data fusion and fault-tolerant technologies[47]. Data evolution deals with the iterative process of data fusion, including complex network modeling, information measurement, propagation dynamics modeling, rule extraction technology and so on[47]. In particular, the data network and updates to the network in light of emerging data can be constructed by complex network modeling[47]. The data servitization technology mainly consists of resource encapsulation, demand decomposition, multi-objective combinational optimization and data visualization based on virtual reality and augmented reality[47].

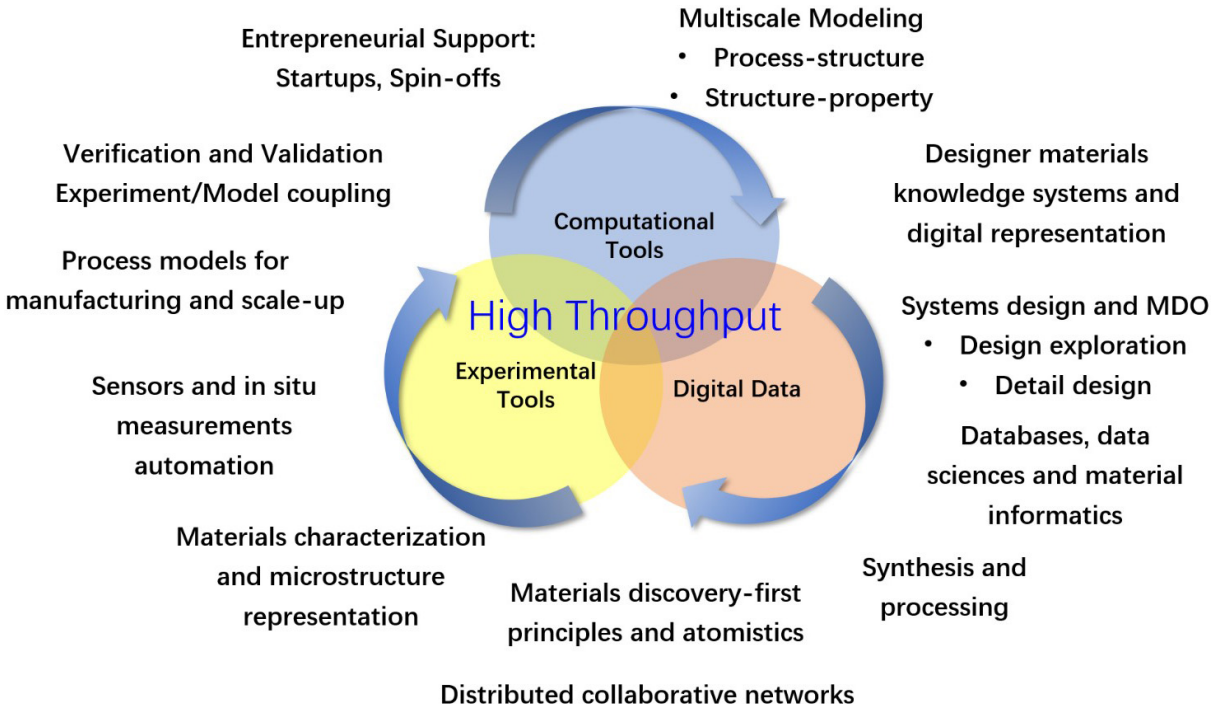

Figure 4. Materials innovation infrastructures to accelerate the discovery, development and deployment of materials with the MGI vision of experimental and computational tools and digital data at the core[5,23]. MGI: Materials Genome Initiative.

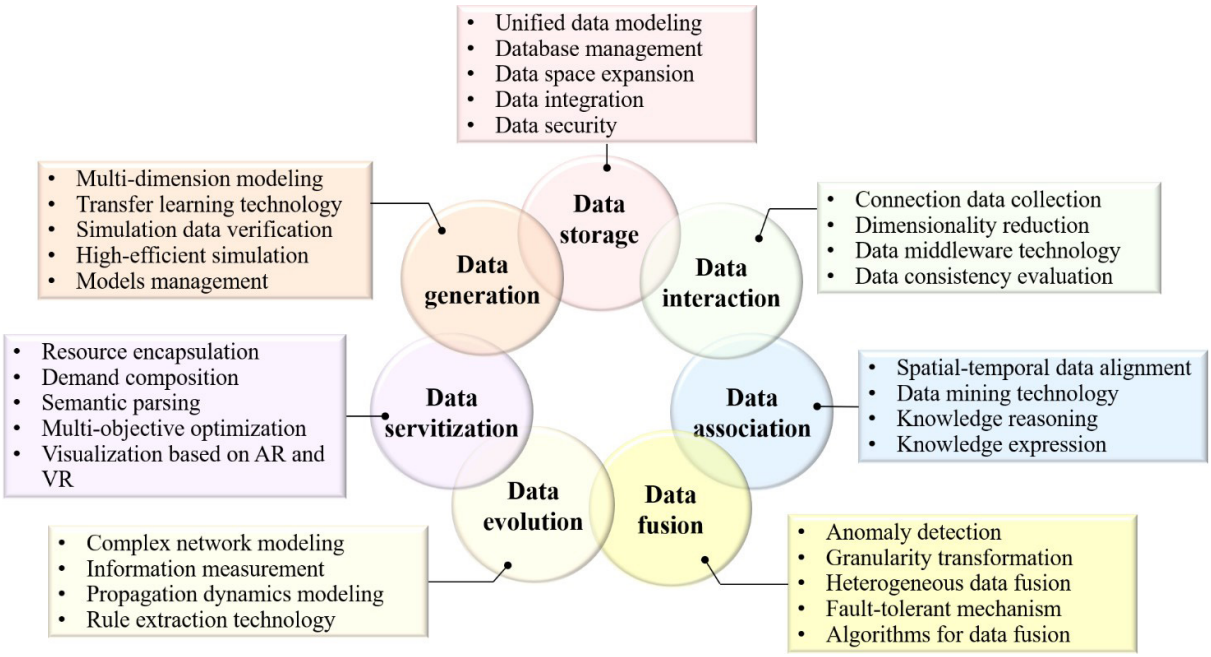

Figure 5. Key enabling technologies for digital twin data[47]. AR: Augmented reality; VR: virtual reality.

In order to fit the FAIR (findability, accessibility, interoperability and reusability) rules, it is essential to address the issues and challenges in data collection, storage and sharing by constructing numerical standards for materials informatics. The National Institute of Science and Technology has proposed or

recommended both performance and interoperability standards to solve issues or challenges during the translation of technologies from laboratory research to industrial engineering applications, thereby accelerating innovation and minimizing the risk involved in the application of novel smart manufacturing technologies[9,25]. The standard draft "General Rules for Materials Genome Engineering Data" released by the Chinese Society of Testing and Materials is the first standard to organize the content of MGE data and will have significant implications in transforming materials science into a data-driven scientific regime[9]. In addition, the framework of proposed systematic standards of big data and the Internet of Things in China is presented in Figure 6 and is the foundation of I^{3}M. The relative numerical standards for material informatics are critical in the development of data-driven and data-intensive technologies, such as databases, data mining, machine learning, artificial intelligence and the acceleration of materials innovation, discovery and design, which are also important in meeting the gaps in innovation and I^{3}M[9,14]. Therefore, based on the data, principles, methodologies and technologies related to digital twins, DTD-based production control can better align the practical process with the simulated plan through efficient real-time data interactions based on the knowledge mining of key information for designers[47].

SIMULATION ASPECTS OF DIGITAL TWINS

With the guidance of the digital thread concept, each hierarchical level of the aforementioned multistep inferences of "data-cyber-knowledge-wisdom" indicates a higher level of refinement of all available data[4], which also indicate the improved/integrated data wealth in the value chain in the design-engineering-operation-service process of I^{3}M and the construction of the modern innovation ecosystem[28]. Smart manufacturing must embrace big data[14]. From the simulation perspective, a digital twin refers to the description of a component, product or system by a set of well aligned executable models with the following characteristics[28]. Firstly, the digital twin is the linked collection of relevant digital artefacts including engineering data, operation data and behavior descriptions via several simulation models[28]. Secondly, the digital twin evolves along with the real system with respect to the entire lifecycle and integrates the currently available knowledge[28]. Thirdly, the digital twin can be utilized to capture the behavior and to derive solutions relevant for the real system, supporting functionalities to assist systems in optimizing operation and service[28]. Therefore, the digital twin extends the concept of model-based systems engineering from engineering and manufacturing to the operation and service phases[28]. Four aspects of the digital twin should be considered comprehensively, including the principle approach and benefit, architecture, lifecycle aspects and value chains[28]. In the following subsections, only the design and engineering phases referring to the digital twin and value chains are discussed to highlight the acceleration of materials discovery and manufacturing in a cost-effective approach.

HT simulation and automation

Powerful techniques for accelerating configurational sampling can overcome the timescale bottlenecks of simulating rare events[12]. HT computational materials design is a straightforward approach in big data-assisted digital twins for smart design and manufacturing[41,48-53]. Based on HT computational materials design, advanced thermodynamic and electronic structure methods are integrated with intelligent data mining and database construction. By exploiting the power of current supercomputer architectures, scientists can generate, manage and analyze enormous data repositories for the discovery of novel materials[15]. In other words, it is based on the integration between computational quantum mechanical-thermodynamic approaches and a multitude of techniques dominated by data-intensive technologies, such as database construction and intelligent data mining[15].

With the guidance of this powerful concept, a large database integrating both the calculated thermodynamic and other physical properties of existing and hypothetical materials can be constructed and intelligently

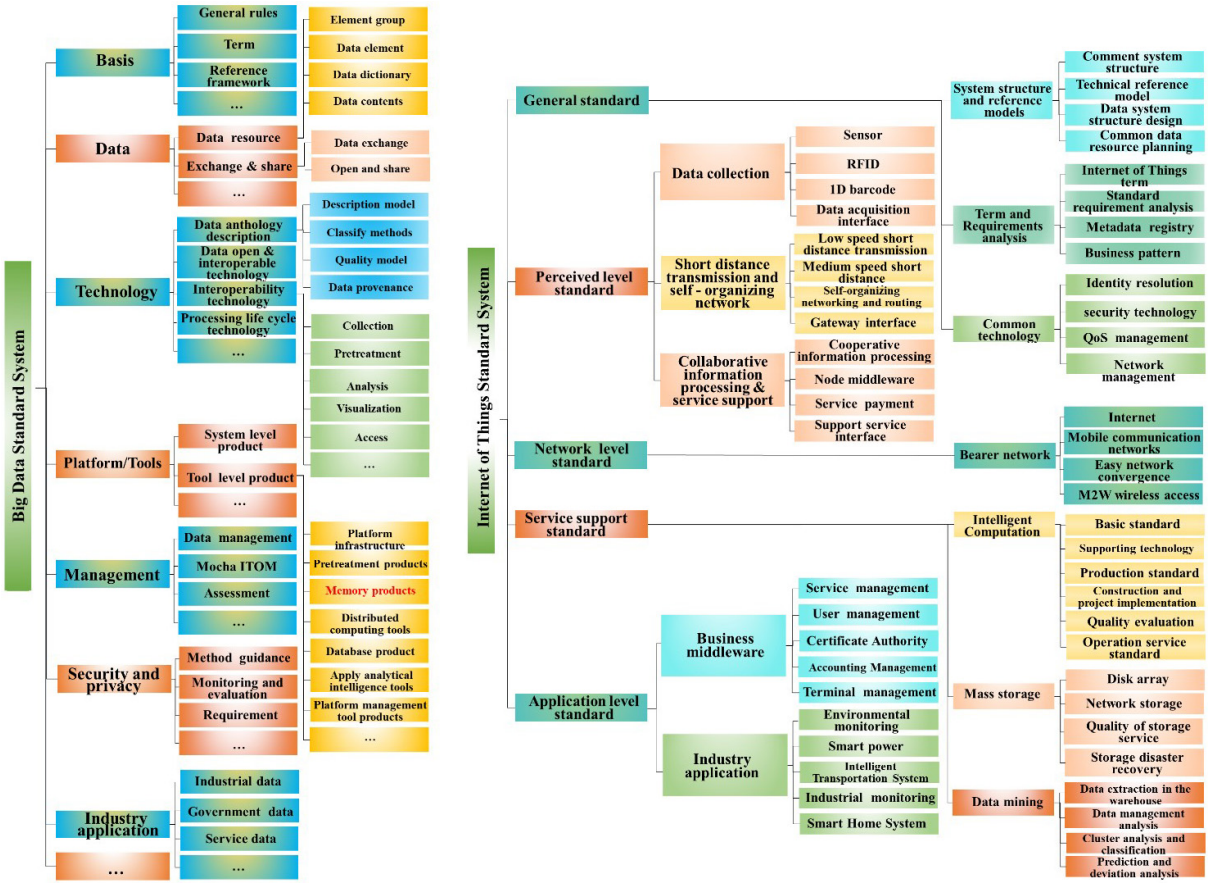

Figure 6. Framework of proposed systematic standards of big data and the Internet of Things in China[9]. ITOM: Information technology operations management; RFID: radio frequency identification; QoS: quality of service.

interrogated in the search for materials with target properties[15]. For instance, by enhancing the accuracy of different kinds of predictions, HT modeling powered by PyCalphad has the capability of expanding CALPHAD modeling into more aspects of ICME[49]. HT density functional theory (HT DFT) has become a powerful tool for accelerating materials design and discovery by its significant contributions to the construction of large databases[54,55]. The applications of HT DFT calculations to search for new materials and conduct fundamental research represent an opportunity for materials science and innovation[41,55]. It has been demonstrated that the method using a machine learning model trained on DFT data from the Open Quantum Materials Database could significantly accelerate materials discovery by predicting the stability of a material based on its crystal structure and chemical composition[56], with the effectiveness of the method illustrated by its application to finding new quaternary Heusler compounds[57].

Moreover, it is believed that automation will be an essential feature/function of calculations and simulations in the future. Robust optimization can be a powerful tool in determining materials and processing tolerance in ICME[49,58]. The capability of automatically handling thousands to hundreds of thousands of calculations represents a novel materials milestone. Correspondingly, systematic database-driven, database-filling protocols can be released to investigate the unknown quarters of materials space, thereby supporting new insights or correlations yielded from data analytics[12]. The systematic efforts of the research community in curating and verifying material properties can be formulated[12]. Since the ESPEI software efficiently evaluates the thermodynamic model parameters within the CALPHAD method, new theoretical and

experimental input data could continuously optimize the associated models by capturing ever more fundamental building blocks of materials and highlighting their function, the procedure of which is similar to that of a knowledge-based one yielded from expert experience[34]. The combination of automatic methods and user online interfaces provides a powerful tool to discover and characterize quantum computational materials efficiently[58].

Furthermore, structure prediction drives materials discovery[59]. Here, our recent works on automated HT DFT calculations are based on the Atomate package[41,60]. FireWorks in Atomate can track both the status of the running job and the progress of user-defined outputs, which can be combined into a built-in database robustly. By interacting with various queue systems, generated reports are provided that summarize the statistics of completion, failure and job start, as well as exploring workflow status[60]. Currently, the PBS, SLURM, Sun Grid Engine and IBM Loadleveler are supported in the code. The job packing organizes multiple calculations executed one after one another within a queue submission, as well as weak parallelization across multiple nodes[60]. The user prioritization of runs will be at both the workflow and individual FireWorks level[60]. The control of reruns will support feedback and be stored in the database[60]. Figure 7 presents two workflow diagrams based on FireWorks, which is a dynamic workflow system designed for robust HT applications[60]. In particular, the workflow for the elastic tensor workflow together with the Atomate dependencies, as shown in Figure 7A. It is noted that the workflow consists of a great number of tasks without complex dependencies. The crystal structure is the initial input to construct the following workflow. Since FireWorks for each of 24 deformations are constructed, these strain states and perturbation stencils may be further customized to yield a minimal workflow and thus to support an optional solution to utilize the fewest possible deformations by symmetry[60].

Based on the Atomate package and integrating expert knowledge and experience in recent decades in the development of theoretical methods and computational software, the Density Functional Theory ToolKit (DFTTK) software in Python has been developed[41]. Its main functions include task submissions on all major operating systems and task executions on high-performance computing environments, which is good at the calculations of thermodynamic properties, including the heat capacity, entropy, phonon properties, enthalpy and free energy of stoichiometric phases[41]. Figure 7B displays the workflow of DFTTK software for the Gibbs energy at finite temperatures, which consists of the structural relaxation completed by the Robust Optimize or CheckRelax module, energy-volume curve and QHA phonon calculations via the EVcheck_QHA module and thermodynamic calculations via the QHAanalysis module. These FireWorks are executed serially.

In particular, the EVcheck_QHA FW contains a series of multilevel judgments and branches[41]. By utilizing the DFTTK, a great number of structures can be constructed using only the required simple input settings. It is noteworthy that the HT post-processing of data will be robustly stored in MongoDB, which can support the original outputs of existing calculations and settings via data mining/searching. Moreover, most common thermodynamic properties and phonon dispersion/DOS with excellent quality can be plotted automatically[41].

Based on the two workflows presented in Figure 7, we designed intelligent HT computing software based on FireWorks, as illustrated in Figure 8. The main code of the software consists of four modules to construct FWs, including the main flow, output analysis (I/O), algorithm and checking modules. The checking module is used to compare the output of each calculation step with the experimental data and knowledge base, with the aim of monitoring common errors and evaluating/estimating the rationality of the output. The key feature of this software is its capability in handling a dynamic workflow by setting up multilayer

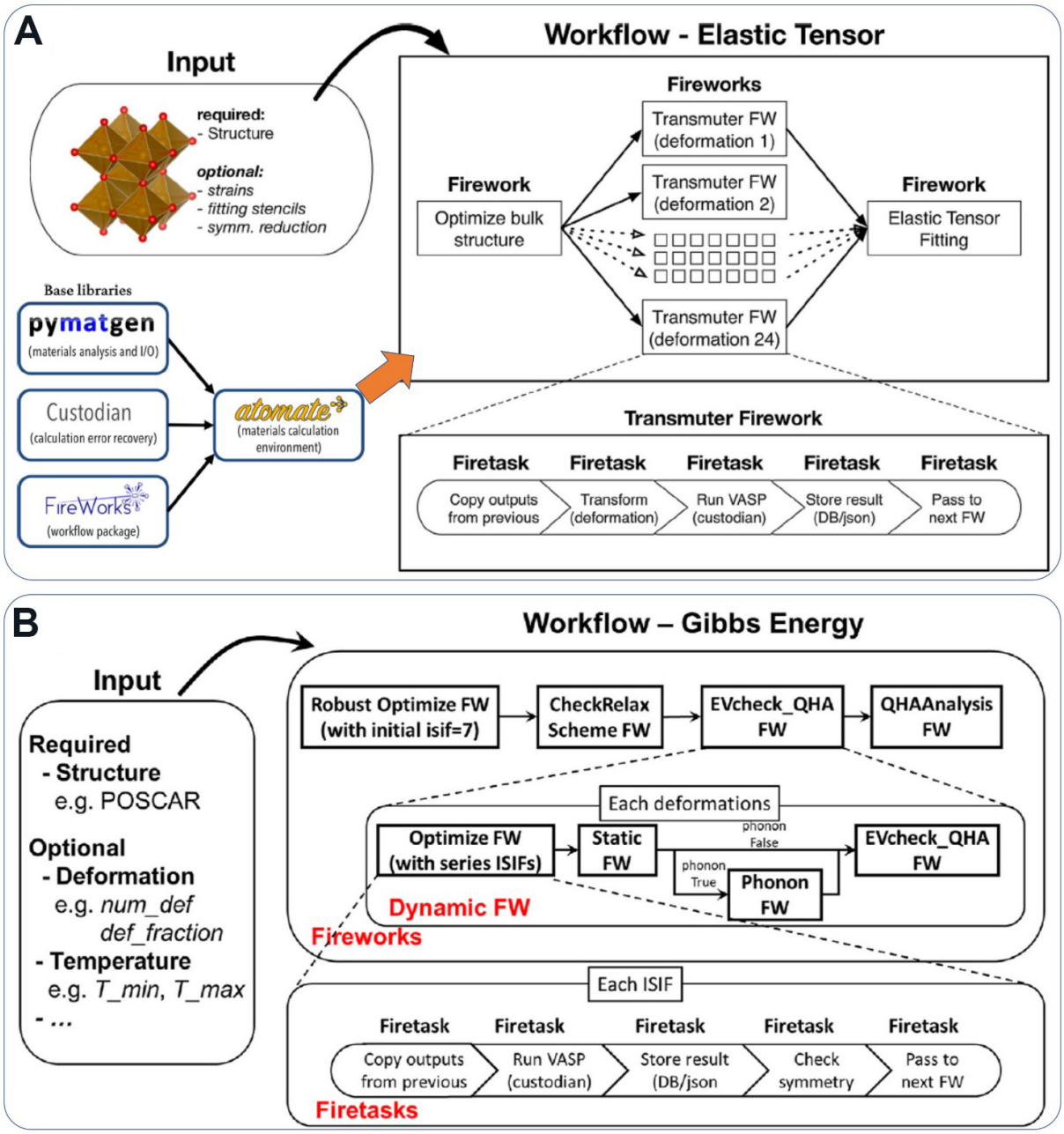

Figure 7. Workflow diagrams based on Atomate. (A) Elastic tensor workflow together with Atomate dependencies[60]. (B) Workflow of DFTTK software for Gibbs energy at finite temperatures[41]. DFTTK: Density functional theory toolkit.

judgment and branch tasks for a more complex and longer workflow. Moreover, with the guidance of the checking module for smart decision making, the function of the automatic correcting process can be completed during/after each calculation step, estimating whether the corrected task to be inserted or not and resulting in the dynamic workflow.

AI-assisted automated multiscale atomistic modeling

Atomistic calculations, for example, *ab initio* calculation or molecular dynamics simulations, have become key approaches for studying the physical and chemical properties of materials. To accurately describe or

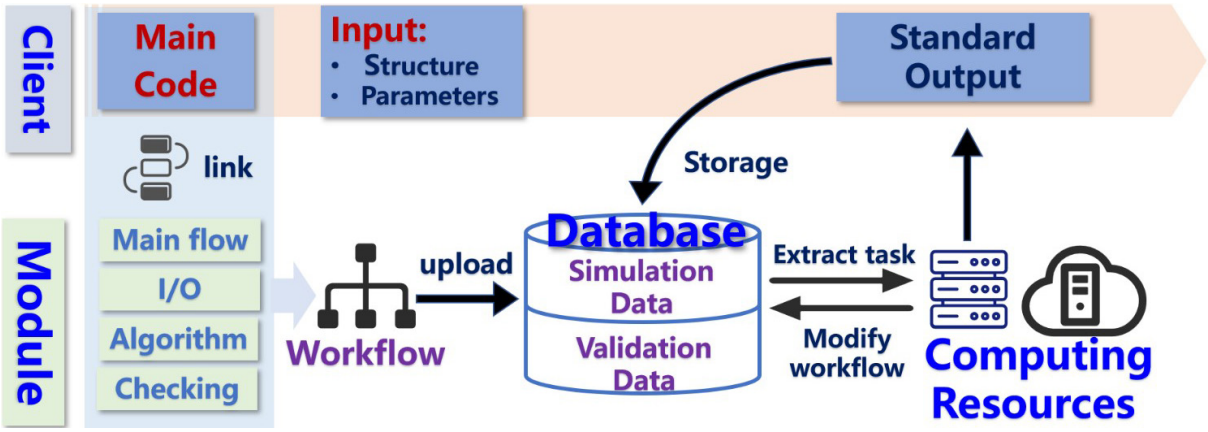

Figure 8. Proposed robust workflow diagram for HT first-principles calculations. HT: High-throughput.

predict materials properties, reliable atomistic modeling is a prerequisite. Over the past few decades, first-principles methods have gained significant improvements and have been widely used in automated materials discovery. Attributed to the computational resource dependence of first-principles calculations, it is challenging to utilize them by solving either a big system or a large scale. On the contrary, physical model-based empirical potentials, for example, the embedded atom method (EAM)[61,62] and angular-dependent potential (ADP)[63,64], are cheap to run but their physical reliability is limited. Hence, a balanced atomistic modeling method is vital for materials discovery.

Recently, machine learning techniques, especially artificial neural networks, have been applied to precisely model atomistic interactions. Neural network potentials (NNPs) can reproduce and predict thermodynamics and kinetic processes in materials[65-67]. The incorporation of NNPs with physical simulation has become ever more common and important, as indicated by the 2020 Gordon Bell Prize[68]. Various NNP codes have so far been published. Our TensorAlloy code[69,70], characteristic of automation and expansibility, is promising in materials modeling and discovery.

Figure 9 demonstrates the workflow of TensorAlloy. TensorAlloy is an automated and expansible code for modeling atomistic interactions and is built upon the virtual atom approach. By smartly introducing virtual atoms, a direct computation graph from atomic positions to the total energy of a structure of arbitrary stoichiometry can be built automatically. Thus, the analytic derivatives of the total energy with respect to atomic positions (atomic forces) or cell tensor (virial stress tensor) can be calculated automatically by the AutoGrad feature of any modern machine learning framework (TensorFlow, PyTorch and so on). Furthermore, complicated physical constraints, such as the equation of state or elastic constants, can be computed and incorporated into the loss function as well. The integration of physical constraints with machine learning techniques makes TensorAlloy highly efficient in optimizing interaction potentials with much lower data requirements.

TensorAlloy is also highly expansible because of the universal descriptor approach. In TensorAlloy, atomic positions are first transformed to universal tensor descriptors (UTDs). Interaction potentials are then constructed based on these UTDs, which act as middleware. The adoption of UTDs has several advantages. Firstly, both NNPs and empirical potentials (EAM, ADP and so on) can be universally implemented in the framework of UTDs, making TensorAlloy a universal platform for modeling various interaction potentials. Secondly, UTDs and UTD-based potentials are built upon tensors. Hence, efficient parallelism and GPU

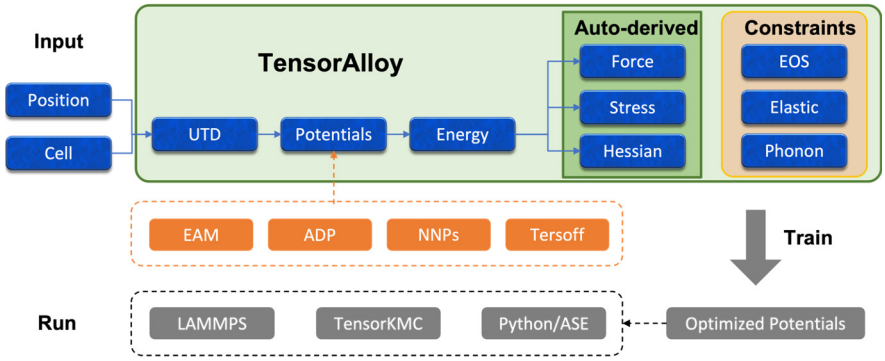

Figure 9. Workflow of TensorAlloy. Atomic positions and cell tensors are transformed to universal tensor descriptors (UTDs) and various interaction potentials can be used to model the total energy. Atomic forces, stress tensors and Hessian matrices can be derived automatically. Physical constraints, like the equation of state or elastic constants, may be utilized in the training stage. Optimized potentials can be called either by C++ codes (LAMMPS for MD and TensorKMC for AKMC) or Python. EAM: Embedded atom method; ADP: angular-dependent potential; NNPs: neural network potentials.

acceleration can be utilized. Thirdly, the basic interaction research workflow consists of three steps: (1) designing or modifying a potential; (2) optimizing the parameters; and (3) running it in a real simulation code. Although step 3 typically requires significant coding effort, all UTD-based potentials share the same input interface and no extra effort is needed to invoke a newly designed potential. These characteristics make TensorAlloy an ideal platform for developing interaction potentials for materials.

Furthermore, TensorAlloy can play a vital role in mesoscale simulations. Very recently, the TensorKMC code was published[71]. TensorKMC is a deep-learning driven code capable of simulating micron second kinetics of 50 trillion atoms on a new-generation of Sunway supercomputer. It is the integration of TensorAlloy and OpenKMC, an atomic kinetic Monte Carlo (AKMC) program. In TensorKMC, highly accurate NNPs are used for real-time energy calculations, which can significantly improve the physical reliability of AKMC simulations. This opens new perspectives for predicting and investigating microstructural evolution at experimental resolutions.

DATA AND DATA MINING

The purpose of data mining is the extraction of knowledge and insight from massive databases, which has been expressed as "data + corrections + theory = knowledge-base"[3]. It is noted that knowledge discovery has been defined as the "non-trivial extraction of implicit, previously unknown and potentially useful information from data", the process of which is in the data mining forms[72]. Knowledge discovery in databases or data mining is an interdisciplinary field that merges ideas from statistics, machine learning, databases and parallel and distributed computing, thereby providing a unique tool to integrate scientific information and theory for materials discovery[3]. Its ultimate target can be achieved through the systematic integration of big data, correlation analysis and theoretical and experimental validation. It is essential to combine the knowledge across scales in the development of advanced technologies/toolkits and materials/products and in the education of the next-generation workforce[18,73]. Such knowledge would generally combine the methods, mechanisms and concepts to execute at the multiscale, which also merge the scientific fields of condensed matter and statistical physics, materials science, applied mechanics, computer science and so on[18,73]. Moreover, it is noted that the sources of data can be varied and numerous, covering HT experiments and computations, combinatorial experiments and huge databases of legacy information[3]. Big data has been considered as the driver for the innovation of databases[27]. Large sets of information can be obtained robustly and efficiently by utilizing advance pioneering data-mining tools,

thereby supporting an informatics driven strategy for materials design[3]. There are several demonstrations of the potential of materials data analytics in extracting high-value materials knowledge from raw data sets in a broad range of materials applications[3,4,33,36,74,75].

Figure 10 presents a logical framework for data-driven ICME for advanced Ti alloys. The discovery and design procedure is composed of four modules. Firstly, the database of Ti-X can be constructed conveniently via HT DFT calculations and experiments. Secondly, data mining and knowledge-based screening are utilized to screen the candidate solutes with those critical features under the guidance of strengthening mechanisms and defect engineering concepts. Thirdly, knowledge-based modeling and machine learning are completed to obtain the self-learned knowledge base and the novel correlations or model in the prediction of the target properties. Finally, with the guidance of various strengthening mechanisms, the best candidate novel high-strength ductile Ti alloy will be discovered by utilizing and validating/optimizing the model, which can be furtherly optimized by repeating this procedure[75].

As shown in Figure 11, based on a database consisting of 44 binary Ti-X systems from HT DFT calculations and 81 reported experimental alloys, data mining is utilized to reveal the correlations among various physical properties at multiple scales. It is noteworthy that the bonding charge density has been considered as the basic building block in the MGI and MGE, revealing the atomic and electronic basis of microstructure-dominated properties and performance at the micro- and macroscales and accelerating the development of advanced metal materials[29,43,73,76,77]. In particular, with the guidance of the concept of defect engineering, the candidate solutes can be efficiently screened out via data mining and knowledge-based screening in terms of stacking fault energy, lattice distortion, segregation energy and dislocation width, which have been considered as key features in screening the best candidate[75]. Through machine learning, the self-learned knowledge base was produced, which is expected to generate plausible explanations and address new hypotheses, particularly, to feedback a novel training model that is superior to the empirical previous ones[75]. For instance, several designed high-strength ductile Ti alloys were fabricated to validate these aforementioned strengthening mechanisms and models, updating the corresponding databases and optimizing the models consistently[75].

In line with the aforementioned digital thread in the CPSPP relationship, data mining is similar to the phenomenological structure-property paradigms when investigating engineering materials, thereby paving the path to derive the correlations, trajectories, clusters, trends and anomalies among disparate data[3]. It presents two primary functions, namely, pattern recognition and prediction, which construct the foundations in revealing material behavior[3].

ENTROPY AND ITS APPLICATION FOR ATTRIBUTE SELECTION IN DECISION TREES

Since knowledge-based modeling/calculation plays an important role in revealing physical properties across scales, knowledge-based simulation, computation, related data-driven and data-intensive research works are strongly recommended for the discovery of advanced materials. With respect to small data sets, decision strategies are developing into data-driven and computation-enabled approaches, the latter of which always combine robust and reliable codes and the availability of computing power to enable the application of pioneering technologies and support an alternative strategy for the discovery of advanced materials[12,75]. Machine learning models, such as neural networks, excel at modeling complex data relationships but generally do not directly provide fundamental scientific insights, thereby motivating more efforts to analyze the models that identify the fundamental composition-property and composition-structure-property relationships. The analysis of machine learning models could accelerate the generation of fundamental materials insights[78]. Although there is a significant amount of manual analysis and materials theories for

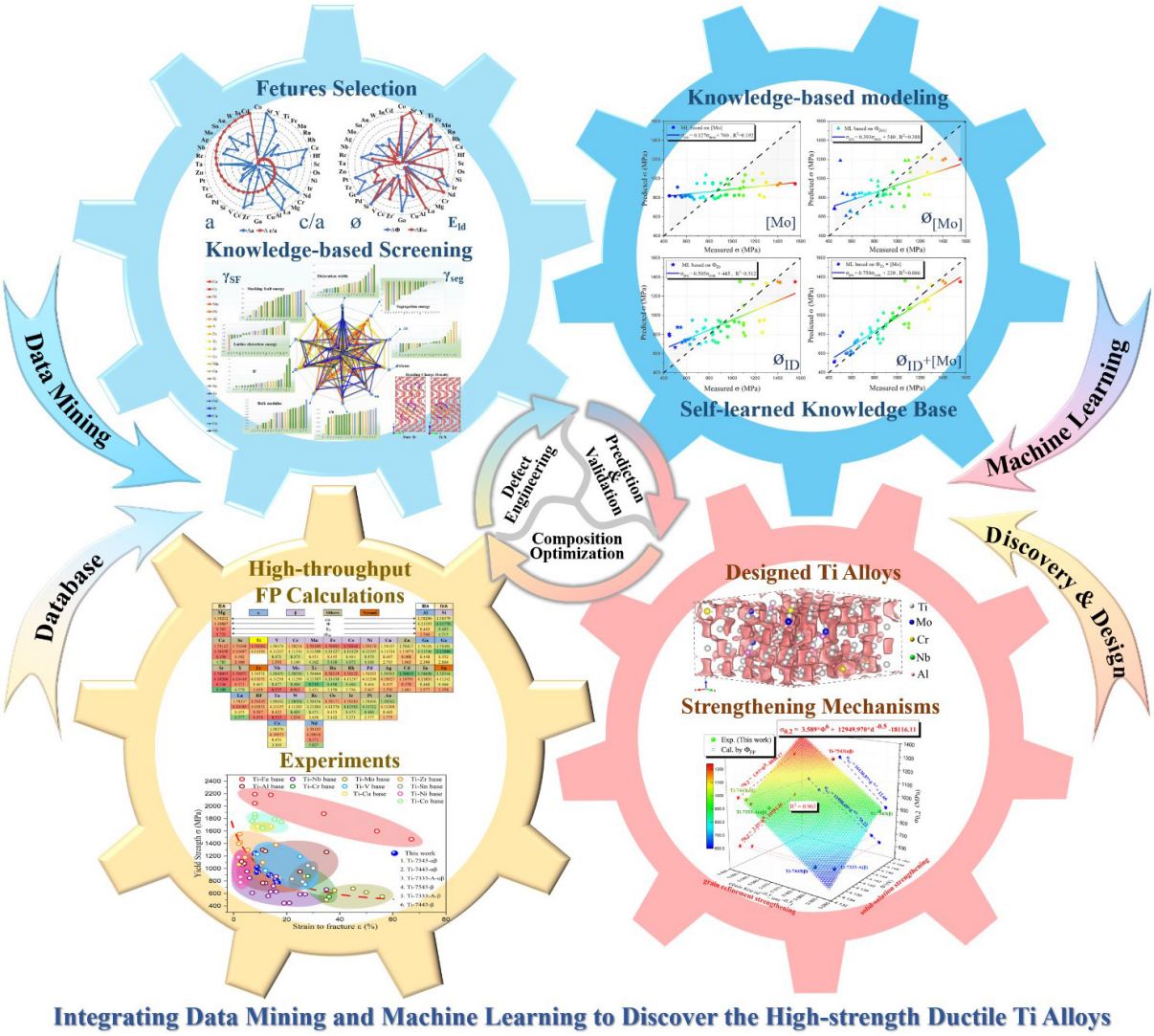

Figure 10. Integrating data mining and machine learning to discover high-strength ductile Ti alloys[75].

constructing the so-called knowledge base, there is still an open question of whether the data-to-insights process can be accelerated via machine learning[78]. Therefore, it is critical to combine this knowledge base and advanced data science and technologies in decision strategies.

Entropy and its application for attribute selection in decision trees are discussed to highlight knowledge-based modeling, simulation and data analysis in machine learning in one unique term. In statistical thermodynamics, entropy is related to the number of microstates (i) and their probabilities of being occupied (p^i)[79]. The microstate configurational entropy (S_{MCE}) is expressed as:

$$S_{MCE} = -k_B \sum_i p^i ln(p^i) \quad (1)$$

where k_B is the Boltzmann constant. Although there are several contributions (i.e., thermal electrons, lattice vibrations and magnetic movements) when predicting the free energies and stability of a system, it is noted that S_{MCE} has only been selected as the key criterion in the discovery of novel high-entropy alloys[80-82].

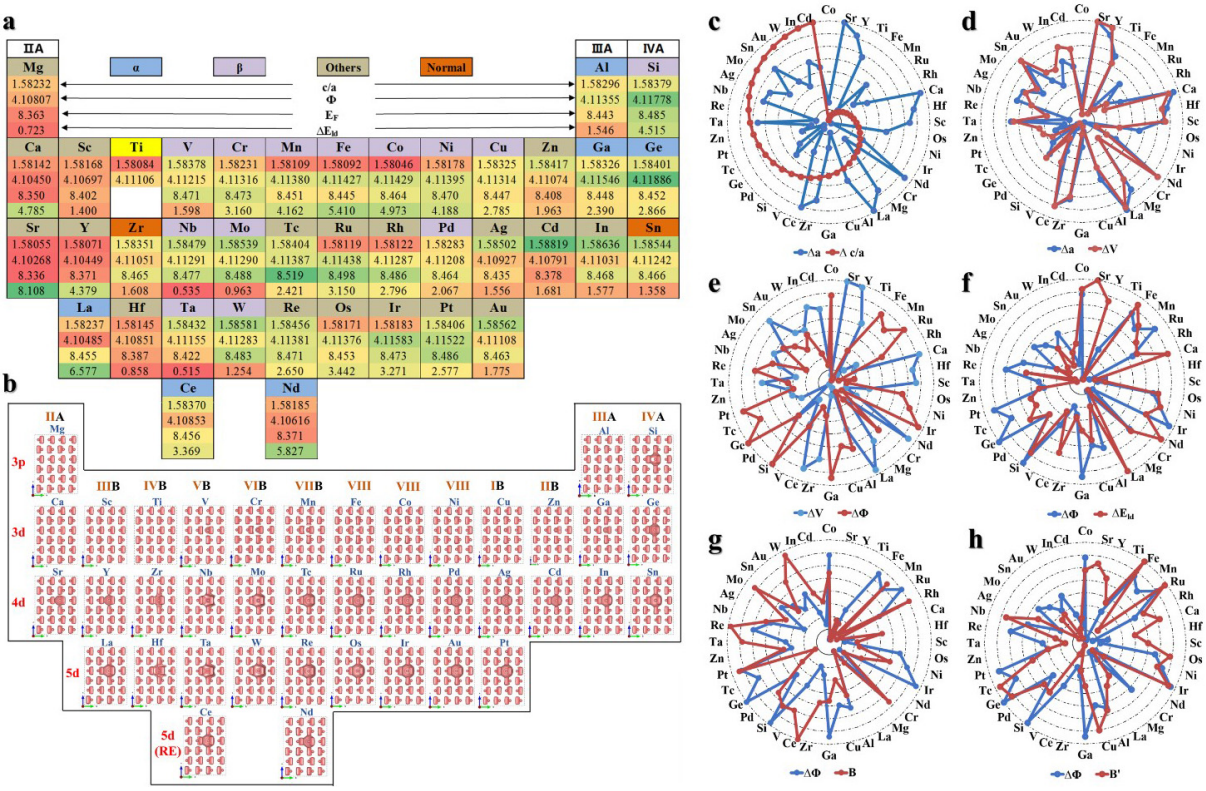

Figure 11. Database and data mining of HCP $Ti_{95}X_1$ from HT first-principles calculations[75]. (a, b) Periodic tables illustrating several dominant properties, including c/a ratio, electron work function (Φ), Fermi energy (E_F), lattice distortion energy (ΔE_{LD}) and bonding charge density isosurface ($\Delta\rho$ = 0.025 e^- $Å^{-3}$). (c-h) Different variation tendencies of c/a ratio change (Δc/a), volume change (ΔV), ΔΦ, bulk modulus (B) and its first derivative (B'), and ΔE_{ld} referred to various reference states.

In contrast, entropy has been set as one principle in the attribute selection of decision trees[83,84]. In data mining variables are often called attributes, while each object is described by a number of variables that correspond to its properties[83,84]. Based on the mathematical function $\log_2 X$, entropy (defined as E_{DT} to present the difference to that used in thermodynamic analysis) is an information-theoretic measure of the "uncertainty" contained in a training set, due to the presence of more than one possible classification[83]. The value of probability (p_i) for i = 1 to K is the number of occurrences of class i divided by the total number of instances within K classes[83]. Accordingly, the entropy of the training set is defined as[83]:

$$E_{DT} = -\sum_i p_i \log_2 p_i \quad (2)$$

Summed over the non-empty classes only. The process of decision tree generation by repeatedly splitting attributes is equivalent to partitioning the initial training set into smaller training sets repeatedly, until the entropy of each of these subsets is zero[83]. The splitting of any attribute has the property that the average entropy of the resulting subsets will be less than (or occasionally equal to) that of the previous training set at any stage of the process[83]. Based on the same algorithm of the logarithmic function and the similar physical meaning in the definition of entropy, it is suggested that machine learning models could be revealed comprehensively to accelerate the generation of fundamental insights in the sequence of "data-cyber-knowledge-wisdom". More efforts in this area are needed in future work.

DIGITAL THREAD IN MATERIALS 4.0 AND INDUSTRY 4.0

Materials 4.0 represents data-driven opportunities for the future of materials manufacturing[32] and has grown from Manufacturing 4.0, Industry 4.0 and the 4th industrial revolution[85] to bridge the gap between multiscale models and experiments, as shown in Figure 12. It represents a new paradigm of materials research, which is excellent at coordinating and analyzing the knowledge related to materials theory, processing and properties in the cyber physical space, as well as reducing the time transforming concept to commercialized products[32]. Moreover, data have been considered as incredibly important inputs for materials innovations by enabling better products with the efficient use of materials and the reducing time and cost of materials design and deployment.

In contrast, Industry 4.0 and Manufacturing 4.0, as additional fundamental innovation and manufacturing paradigms, motivate data-driven and data-intensive technologies at the center of economic and social systems[85]. They integrate the advanced digitalization of factories, the Internet and future-oriented technologies to bring intelligence to devices, machines and systems[86,87], which enable the development of advanced intelligent technologies and tools for digital twins[88]. Moreover, pioneering technologies, including HT/cloud computing, big data analytics, artificial intelligence and so on, have greatly stimulated the development of smart manufacturing[89], which is dominated by cyber-physical integration and is continuously being embraced by manufacturers[89].

In line with the digital thread of HCPs, the aforementioned CPSPP relationship presents the material structural evolution with various process parameters in the cyber-physical integration and expresses the properties and performance as a function of the material structure[5]. In our opinion, the human in the HCPs highlights the significance of the knowledge base while the CPS indicates the important role of data-driven and data-intensive technologies, such as artificial intelligence, machine learning, data mining and so on. As shown in Figure 13, the hierarchical levels of CPS and digital twins in manufacturing are compared[89]. It is found that both CPS and digital twins focus on the achievement of the cyber-physical integration, setting up the foundations of smart manufacturing[89]. In the implementation of functions, sensors and actuators are the main modules in CPS, while the model-based system-engineering approach emphasizing data and models is the foundation of digital twins[87,89]. Both CPS and digital twins are inseparable from new internet technology, which provides their technical basis[89]. It is noteworthy that the complexity of digital twins varies based on the use case, the vertical industry and the business objective[90]. Different levels of digital twin complexity, from simple devices to complex assets, will have differing complex aspects to be considered[90].

Considering the advances in data-intensive, data-driven and even generation information technologies, smart manufacturing is becoming the focus of global manufacturing transformation and upgrading[87]. As presented in Figure 14, digital twins integrate all manufacturing processes and pave the way for the cyber-physical integration of manufacturing, which can achieve the closed loop and optimization of product design, manufacturing and smart maintenance, repair and overhaul[87]. Digital twins drive the business impact of smart design and I^3M by efficiently monitoring and controlling assets and processes, and developing economic and business models with the benefits of development costs[90], with the aim of the long-term targets of MGE and ICME. With regards to the digital thread in the process flow for computation and knowledge representation displayed in Figure 14, it is expected that the knowledge base and graph will drive smart solution generation[91]. With feedback loops in which physical processes affect cyber parts and vice versa, CPS and digital twins can endow manufacturing systems with greater efficiency, resilience and intelligence[89]. CPS and digital twins share the same essential concepts of an intensive cyber-physical connection, real-time interaction, organization integration and in-depth collaboration[89]. By integrating

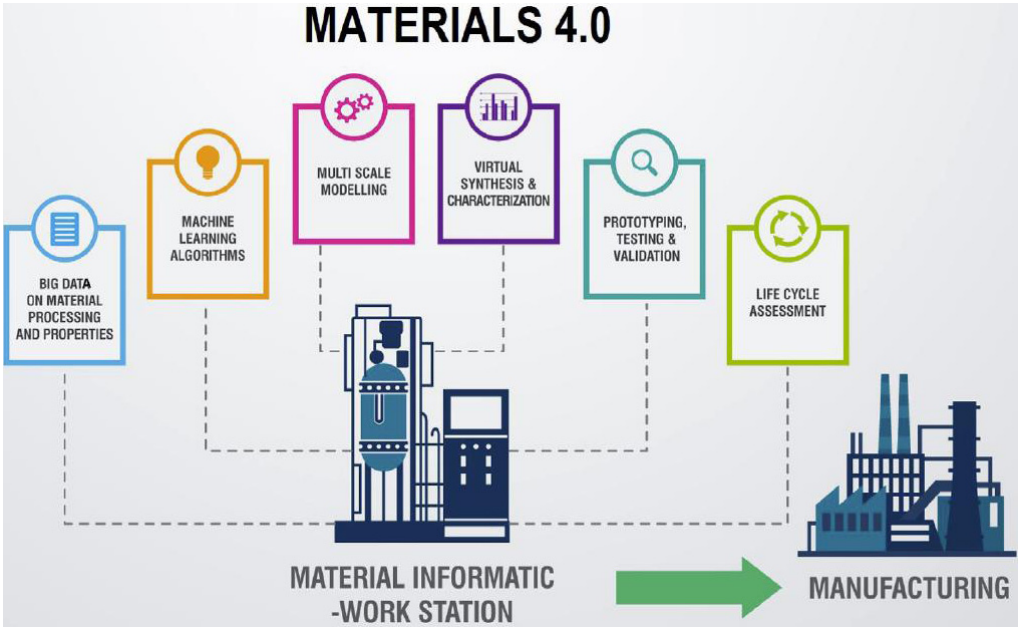

Figure 12. Concept of a web-based materials big data platform, i.e., Materials 4.0[32]. Reproduced with permission from Elsevier.

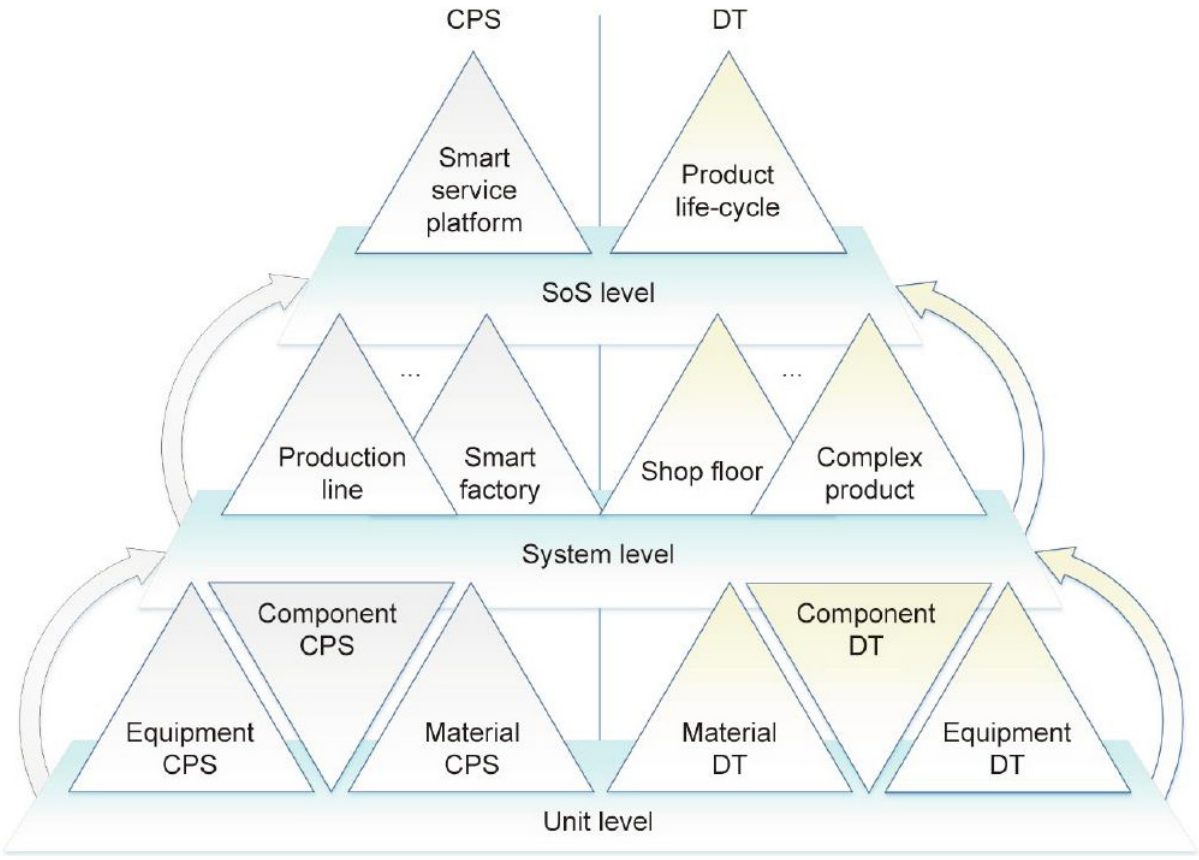

Figure 13. Hierarchical levels of cyber-physical System (CPS) and digital twins (DT) in manufacturing[89].

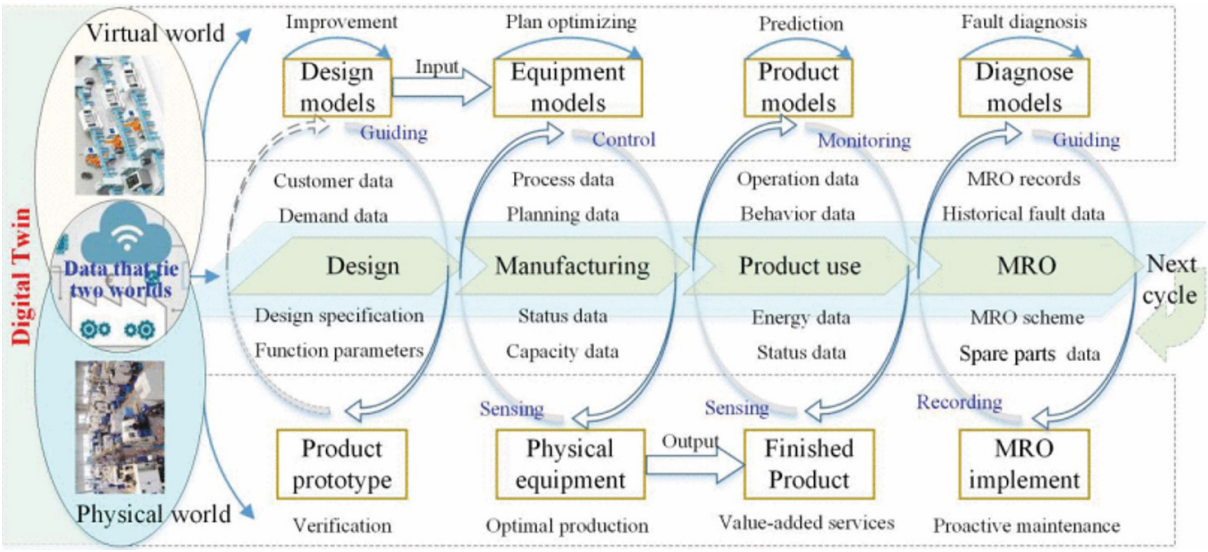

Figure 14. Digital thread in the process flow for computation and knowledge representation[87].

these advanced concepts and technologies, our recent works on design, manufacturing and product use via big data-assisted digital twins for smart design and manufacturing are presented in the following case studies.

Figure 15 displays the HT casting model of standard mechanical samples together with their digital twin products, which has the capability to *in-situ* measure the fluidity of melts presented by the red spiral path. By taking advantage of digital twin technologies, the casting blocks consisting of different types and numbers of standard tensile and fatigue samples and raw plates can be designed, the wax product of which can be fabricated by additive manufacturing conveniently. Correspondingly, the twinning product can be obtained via the investment casting approach. It is noted that both the cost and time to investigate the same number of samples will be decreased significantly since there is a limited machining work for these near net-shape samples. Moreover, the mechanical properties of the near net-shape sample will be better than those yielded from the rod, which also requires more machining works.

Figure 16 presents a case study of the digital twin design and manufacturing approaches for a new high-strength near β Ti7333 landing gear torque arm. This digital twin process was used to systematically explore the microstructural evolution rules, establish constitutive and microstructural models of hot deformation and predict several fundamental properties, including average grain size, phase volume fraction, dislocation density and macro physical fields (i.e., stain and stress fields) during hot deformation[29]. In the finite element numerical simulation of a die forging, the microstructures, geometry, fundamental physical properties and structure-property relationships were integrated to present a knowledge-based and data-driven design approach and yield a precise result. It can be seen that the predicted average grain size and component size of the lower anti-torque arm of the Ti-7333 alloy matched well with the experimental values. Since the process parameters of hot and isothermal die forging can be well optimized in the cyber system[29], the fabrications of lower anti-torque arms in the physical system will be completed using an efficient and cost-effective approach with the power of digital twins.

With the rapid development of product digital design technology, the efficiency and agility of product design and manufacturing are more required by customers. For example, stereolithography (SLA) is widely

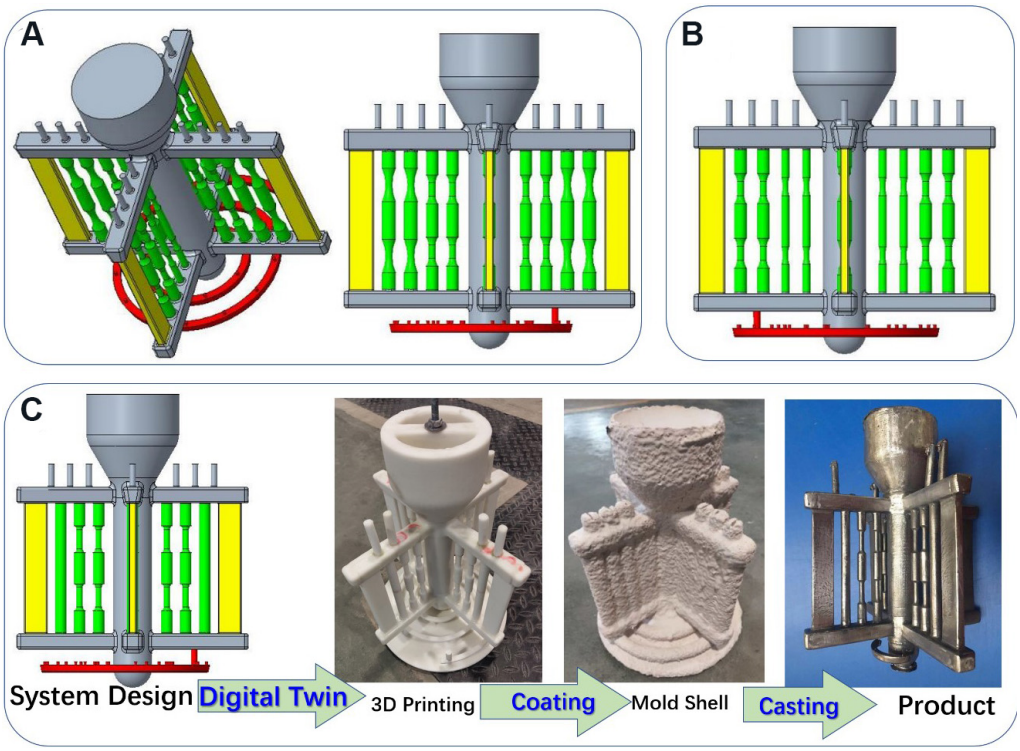

Figure 15. HT casting model of standard mechanical samples together with their digital twin products. (A) 3D and 2D views of the casting blocks consisting of the standard tensile and fatigue samples and the raw plates. (B) Casting blocks consisting of two kinds of standard fatigue samples and raw plates. (C) Experimental fabrication of digital twin products. HT: High-throughput.

applied for structural verification and project display in product design due to its personalized production, high dimensional accuracy (25-100 μm) and surface quality of printed parts[92]. As shown in Figure 17, the size of the tiny structure is 1 mm in diameter in the digital modeling of composite floor slabs (local structure), which is used for case presentation and equipment capability verification. Firstly, classical CAD software is used to build the designed 3D model, which also considers many essential data, including product structure, size, material and so on. The standardized model should be converted into the STL file format (.stl) for SLA 3D printing slice analysis with high precision and molding speed. Secondly, it is essential to optimize the printing parameters (placement angle, layer thickness, filler type, support setting and so on) via slicing software for intelligent design. Here, the ChiTuBox software is utilized in the slicing analysis to compare the influence of different layer thicknesses, including 100, 50 (default value) and 25 μm. It is found that the default value setting takes about 1 h 2 m 29 s and exhausts 15.72 mL of resin. For the 25 and 100 μm ones with the same exhausted material, the processing time is improved and decreased by 50%, respectively. Moreover, the surface accuracy of the 25 μm one is the best. Therefore, it is determined that the thinner layer thickness should be utilized to maintain the quality of the tiny structure and to control the printing time spontaneously.

Since the setting principles of other parameters are similar, the aforementioned simulated processing could balance between precision and time, which also considers of the consumable material (cost) and strength. Finally, the digital twins of the designed part will be implemented via the post-processing processes consisted of cleaning, support removal and drying. This procedure presents digital twin designing and additive manufacturing approaches for product entity. Compared with the traditional technology, the cost and production cycle are effectively reduced and the formability is excellent.

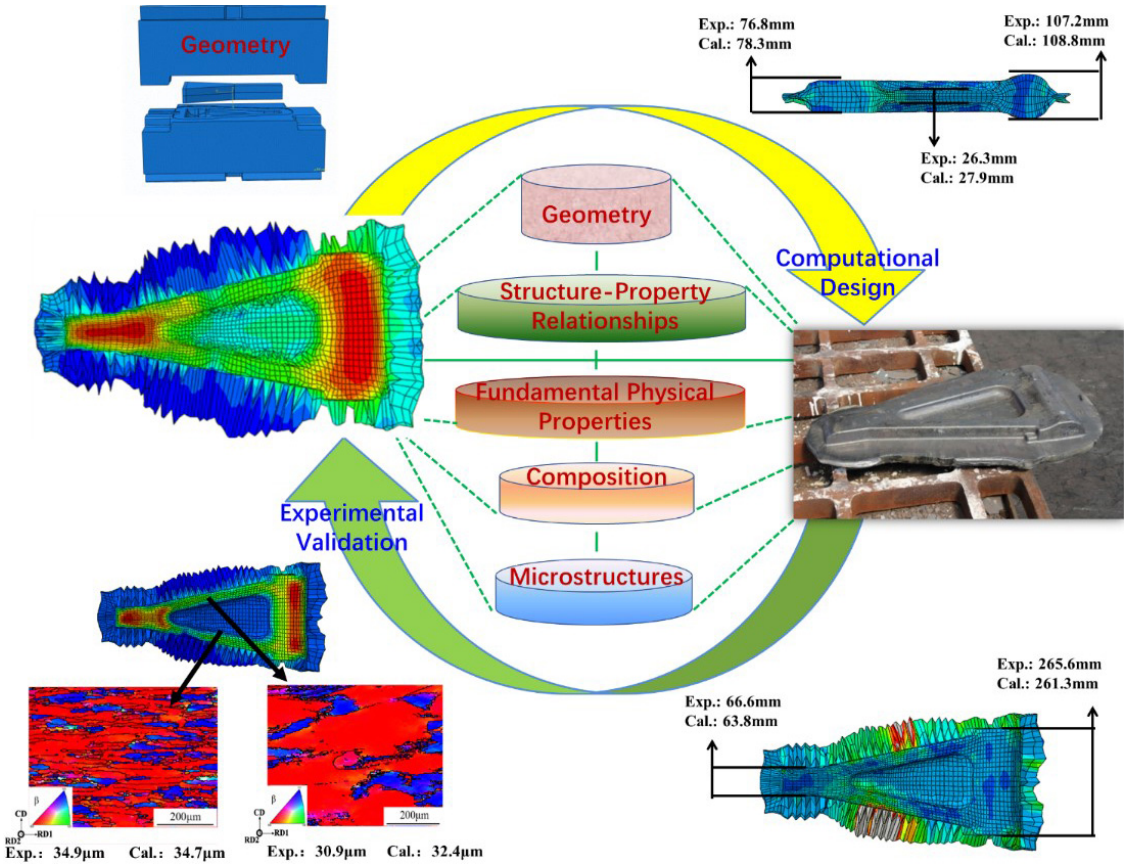

Figure 16. Digital twin design and manufacturing approaches for a Ti7333 landing gear torque arm[29].

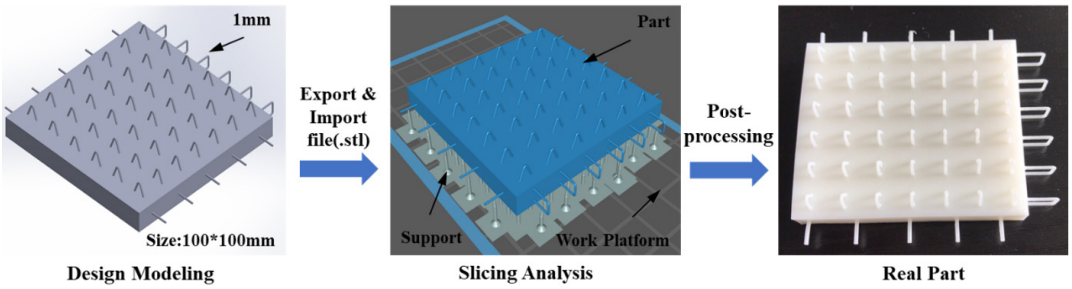

Figure 17. Smart manufacturing of composite floor slabs through the digital twin paradigm.

The implementations of Manufacturing 4.0 and Industry 4.0 come about with the development and full industrial implementation of CPs[24,85,93]. In order to minimize or avoid unintentional and problematic consequences, the knowledge-based design cycle of CPs is recommended, which starts with the requirement of analysis and defines the criteria for the subsequent CP solution in later design/synthesis activities[93]. The expected properties should be compared with the criteria established during the analysis stage to evaluate the provisional CP design solution[93]. At the end of this design cycle, the production system designer decides whether to continue developing the CP design by further elaborating the provisional design solution or whether to try a different type of solution to generate a better CP design proposal[93]. Once the solution of the CP design fits the requirements and criteria, the status will be upgraded to that of final design and the project can move on to implementation planning[93]. For instance, the case study of a big data-assisted digital

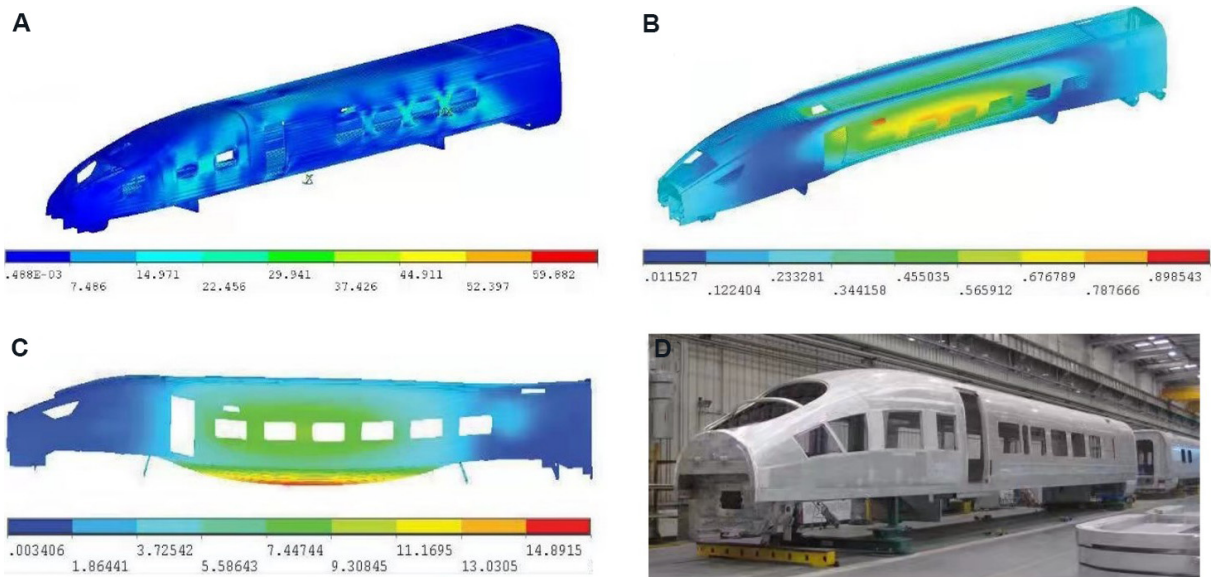

Figure 18. A big data-assisted digital twin for the smart design and manufacturing of the Al body frames of the China Railway High-speed. (A-C) Predicted properties of designed body frame in terms of strength, stiffness and fatigue intensity, respectively. (D) Fabricated body frame as one kind of final solution.

twin for the smart design and manufacturing of the Al body frames of the China Railway High-speed is present in Figure 18. In line with the knowledge-based design cycle of CPPS, the simulated mechanical properties related to the safety and lifetime of the body frame, such as strength, stiffness, fatigue intensity and so on, should be comprehensively investigated, supporting the fundamental information for the design, optimization and final decision. It is noteworthy that both the time and the cost will be significantly reduced in the development of new types of rails from the China Railway High-speed to subways through the big data-assisted digital twin for the smart design and manufacturing routine.

CONCLUSION

A digital twin is an integrated multi-physics, multiscale, probabilistic simulation of a complex product of a system that uses the best available physical models, sensor updates, history data and so on to mirror the life of its corresponding twin. In line with the digital thread of HCPs, the CPSPP relationship captures the foundations of material structural evolution as a function of the process parameters in the cyber-physical integration and expresses the properties and performance as a function of the material structure. In our opinion, the human in the HCPs highlights the significance of the knowledge base while the CPS indicates the important role of data-driven and data-intensive technologies, such as artificial intelligence, machine learning, data mining and so on. Since knowledge-based modeling/calculation is a key feature in connecting physical properties across scales, knowledge-based simulation, computation, relative data-driven and data-intensive research works are strongly recommended in the discovery and development of advanced materials. Knowledge discovery in databases or data mining provides an unique tool to integrate scientific information and theory for materials discovery, which is an interdisciplinary field merging ideas from statistics, databases, machine learning and HT high-performance computations. With respect to data and data mining technologies, the entropy and its application for attribute selection in decision trees are discussed to emphasize knowledge-based modeling, simulation and data analysis in machine learning coherently.

Big data-assisted digital twins integrate all manufacturing processes and pave the way for the cyber-physical integration of manufacturing. In order to minimize or avoid unintentional and problematic consequences, the knowledge-based design cycle of CPPS is recommended, which starts with the requirement of analysis and defines the criteria for the subsequent CPPS solution in later design/synthesis activities. The case study of a big data-assisted digital twin for the smart design and manufacturing of the Al body frames of the standard China Railway High-speed is presented, highlighting the reduced time and the cost in the development of new types of rails from this to subways. It is believed that big data-assisted digital twins for smart design and manufacturing would effectively support better products with the application of novel materials by reducing the time and cost of materials design and deployment.

DECLARATIONS

Authors' contributions

Writing, review and editing: Wang WY, Yin J, Chai Z, Chen X, Zhao W, Lu J, Sun F, Jia Q, Gao X, Tang B, Hui X, Song H, Xue F, Liu ZK, Li J

High-throughput automated calculations: Yin J, Lu J under the supervision of Wang WY; Chen X, Gao X under the supervision of Song HF

Digital design and fabrications: Wang WY, Tang B, Li J, Xue F, Liu ZK, Hui X, Cai Z, Zhao W

Resources, supervision: Wang WY, Li J, Liu Z, Song H, Xue F, Gao X, Hui X

Project administration: Wang WY, Li J, Sun F, Song H, Gao X, Hui X

Availability of data and materials

Not applicable.

Financial support and sponsorship

This work is financially supported by the National Basic Scientific Research project of China (No. JCKY2020607B003), Science Challenge Project (Contract No. TZ2018002), the National Key Research and Development Program of China (2018YFB0703801 and 2018YFB0703802), the Key Project of the Equipment Pre-Research Field Fund of China (No.6140922010302), National Natural Science Foundation of China (No. 51690164), and CRRC Tangshan Co., LTD (Contract No. 201750463031).

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2022.

REFERENCES

1. Zhang TY, Liu XJ. Informatics is fueling new materials discovery. *J Mater Inf* 2021;1:6. DOI
2. Zhang TY. New tool in the box. *J Mater Inf* 2021;1:1. DOI
3. Rajan K. Materials informatics. *Materials Today* 2005;8:38-45. DOI
4. Kalidindi SR, Brough DB, Li S, et al. Role of materials data science and informatics in accelerated materials innovation. *MRS Bull* 2016;41:596-602. DOI
5. Kalidindi SR, Medford AJ, Mcdowell DL. Vision for data and informatics in the future materials innovation ecosystem. *JOM* 2016;68:2126-37. DOI
6. Liu Z. A materials research paradigm driven by computation. *JOM* 2009;61:18-20. DOI

7. Liu Z. Ocean of data: integrating first-principles calculations and CALPHAD modeling with machine learning. *J Phase Equilib Diffus* 2018;39:635-49. DOI
8. A national strategic plan for advanced manufacturing. Available from: https://www.nist.gov/oam/national-strategic-plan-advanced-manufacturing [Last accessed on 23 Feb 2022].
9. Wang WY, Li P, Lin D, et al. DID code: a bridge connecting the materials genome engineering database with inheritable integrated intelligent manufacturing. *Engineering* 2020;6:612-20. DOI
10. Olson GB. Computational design of hierarchically structured materials. *Science* 1997;277:1237-42. DOI
11. Raccuglia P, Elbert KC, Adler PD, et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* 2016;533:73-6. DOI PubMed
12. Marzari N. Materials modelling: the frontiers and the challenges. *Nat Mater* 2016;15:381-2. DOI PubMed
13. Kalinin SV, Sumpter BG, Archibald RK. Big-deep-smart data in imaging for guiding materials design. *Nat Mater* 2015;14:973-80. DOI PubMed
14. Kusiak A. Smart manufacturing must embrace big data. *Nature* 2017;544:23-5. DOI PubMed
15. Curtarolo S, Hart GL, Nardelli MB, Mingo N, Sanvito S, Levy O. The high-throughput highway to computational materials design. *Nat Mater* 2013;12:191-201. DOI PubMed
16. Debnath A, Krajewski AM, Sun H, et al. Generative deep learning as a tool for inverse design of high entropy refractory alloys. *J Mater Inf* 2021;1:3. DOI
17. National Science and Technology Council. Materials genome initiative for global competitiveness. Available from: https://www.researchgate.net/publication/267901251_Materials_Genome_Initiative_for_Global_Competitiveness [Last accessed on 23 Feb 2022].
18. Wang W, Li J, Liu W, Liu Z. Integrated computational materials engineering for advanced materials: a brief review. *Computational Materials Science* 2019;158:42-8. DOI
19. Zhou J, Li P, Zhou Y, Wang B, Zang J, Meng L. Toward new-generation intelligent manufacturing. *Engineering* 2018;4:11-20. DOI
20. Zhong RY, Xu X, Klotz E, Newman ST. Intelligent manufacturing in the context of Industry 4.0: a review. *Engineering* 2017;3:616-30. DOI
21. Rickman J, Lookman T, Kalinin S. Materials informatics: from the atomic-level to the continuum. *Acta Materialia* 2019;168:473-510. DOI
22. Lookman T, Alexander F, Rajan K. Information science for materials discovery and design. Cham: Springer; 2016. DOI
23. Liu Z, Mcdowell DL. The Penn State-Georgia Tech CCMD: ushering in the ICME Era. *Integr Mater Manuf Innov* 2014;3:409-28. DOI
24. National Academies of Sciences, Engineering, and Medicine; Policy and Global Affairs; Government-University-Industry Research Roundtable. The Fourth Industrial Revolution: Proceedings of a Workshop - In Brief. Washington: The National Academies Press; 2017. DOI PubMed
25. The future of manufacturing: a new era of opportunity and challenge for the UK. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/25 [Last accessed on 23 Feb 2022].
26. Tao F, Qi Q. Make more digital twins. *Nature* 2019;573:490-1. DOI PubMed
27. Cui B, Mei H, Ooi BC. Big data: the driver for innovation in databases. *Natl Sci Rev* 2014;1:27-30. DOI
28. Boschert S, Rosen R. Digital twin - the simulation aspect. In: Hehenberger P, Bradley D, editoes. Mechatronic Futures. Cham: Springer; 2016. p. 59-74. DOI
29. Wang WY, Tang B, Lin D, et al. A brief review of data-driven ICME for intelligently discovering advanced structural metal materials: Insight into atomic and electronic building blocks. *J Mater Res* 2020;35:872-89. DOI
30. Yi Y, Yan Y, Liu X, Ni Z, Feng J, Liu J. Digital twin-based smart assembly process design and application framework for complex products and its case study. *J Manuf Syst* 2021;58:94-107. DOI
31. Ferguson S. Apollo 13: the first digital twin. Available from: https://blogs.sw.siemens.com/simcenter/apollo-13-the-first-digital-twin/ [Last accessed on 23 Feb 2022].
32. Jose R, Ramakrishna S. Materials 4.0: materials big data enabled materials discovery. *Applied Materials Today* 2018;10:127-32. DOI
33. Xiong W, Olson GB. Cybermaterials: materials by design and accelerated insertion of materials. *npj Comput Mater* 2016:2. DOI
34. Craig PL. Modeling software for Materials 4.0. Available from: https://news.psu.edu/story/670090/2021/09/21/academics/modeling-software-materials-40 [Last accessed on 23 Feb 2022].
35. Bocklund B, Otis R, Egorov A, Obaied A, Roslyakova I, Liu Z. ESPEI for efficient thermodynamic database development, modification, and uncertainty quantification: application to Cu-Mg. *MRS Communications* 2019;9:618-27. DOI
36. Liu Z. Computational thermodynamics and its applications. *Acta Materialia* 2020;200:745-92. DOI
37. Kaufman L, Ågren J. CALPHAD, first and second generation - Birth of the materials genome. *Scripta Materialia* 2014;70:3-6. DOI
38. Olson G, Kuehmann C. Materials genomics: from CALPHAD to flight. *Scripta Materialia* 2014;70:25-30. DOI
39. Liu Z. Perspective on Materials Genome®. *Chin Sci Bull* 2014;59:1619-23. DOI
40. Liu Z. First-principles calculations and CALPHAD modeling of thermodynamics. *J Phase Equilib Diffus* 2009;30:517-34. DOI
41. Wang Y, Liao M, Bocklund BJ, et al. DFTTK: Density Functional Theory ToolKit for high-throughput lattice dynamics calculations. *Calphad* 2021;75:102355. DOI
42. Shin D, Saal J. Computational materials system design. 1st ed. Cham: Springer; 2018. DOI: 10.1007/978-3-319-68280-8 DOI
43. Zhou BC, Wang WY, Liu ZK, Arroyave R. Electrons to phases of magnesium. In: Horstemeyer MF, editor. Integrated computational materials engineering (ICME) for metals: concepts and case studies. Hoboken: John Wiley & Sons; 2018; p. 237-82. DOI

44. Liu X, Furrer D, Kosters J, Holmes J. Vision 2040: a roadmap for integrated, multiscale modeling and simulation of materials and systems. Available from: https://ntrs.nasa.gov/api/citations/20180002010/downloads/20180002010.pdf [Last accessed on 23 Feb 2022].
45. National Science & Technology Council. Strategy for American Leadership in Advanced Manufacturing. Available from: https://trumpwhitehouse.archives.gov/wp-content/uploads/2018/10/Advanced-Manufacturing-Strategic-Plan-2018.pdf [Last accessed on 23 Feb 2022].
46. Broderick SR, Santhanam GR, Rajan K. Harnessing the big data paradigm for ICME: shifting from materials selection to materials enabled design. *JOM* 2016;68:2109-15. DOI
47. Zhang M, Tao F, Huang B, et al. Digital twin data: methods and key technologies. *digitaltwin* 2021;1:2. DOI
48. Robbins DW, Hartwig JF. A simple, multidimensional approach to high-throughput discovery of catalytic reactions. *Science* 2011;333:1423-7. DOI PubMed PMC
49. Otis RA, Liu Z. High-throughput thermodynamic modeling and uncertainty quantification for ICME. *JOM* 2017;69:886-92. DOI
50. de Walle A, Sun R, Hong Q, Kadkhodaei S. Software tools for high-throughput CALPHAD from first-principles data. *Calphad* 2017;58:70-81. DOI
51. Mounet N, Gibertini M, Schwaller P, et al. Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds. *Nat Nanotechnol* 2018;13:246-52. DOI PubMed
52. Li R, Xie L, Wang WY, Liaw PK, Zhang Y. High-throughput calculations for high-entropy alloys: a brief review. *Front Mater* 2020;7:290. DOI
53. Shang S, Zhou B, Wang WY, et al. A comprehensive first-principles study of pure elements: vacancy formation and migration energies and self-diffusion coefficients. *Acta Materialia* 2016;109:128-41. DOI
54. Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *JOM* 2013;65:1501-9. DOI
55. Jain A, Hautier G, Moore CJ, et al. A high-throughput infrastructure for density functional theory calculations. *Computational Materials Science* 2011;50:2295-310. DOI
56. Krajewski AM, Siegel JW, Xu J, Liu ZK. Extensible structure-informed prediction of formation energy with improved accuracy and usability employing neural networks. Available from: https://arxiv.org/abs/2008.13654. https://arxiv.org/abs/2008.13654 [Last accessed on 23 Feb 2022].
57. Kim K, Ward L, He J, Krishna A, Agrawal A, Wolverton C. Machine-learning-accelerated high-throughput materials screening: discovery of novel quaternary Heusler compounds. *Phys Rev Materials* 2018;2:123801. DOI
58. Curtarolo S, Setyawan W, Hart GL, et al. AFLOW: an automatic framework for high-throughput materials discovery. *Computational Materials Science* 2012;58:218-26. DOI
59. Oganov AR, Pickard CJ, Zhu Q, Needs RJ. Structure prediction drives materials discovery. *Nat Rev Mater* 2019;4:331-48. DOI
60. Mathew K, Montoya JH, Faghaninia A, et al. Atomate: a high-level interface to generate, execute, and analyze computational materials science workflows. *Computational Materials Science* 2017;139:140-52. DOI
61. Baskes MI. Application of the embedded-atom method to covalent materials: a semiempirical potential for silicon. *Phys Rev Lett* 1987;59:2666-9. DOI PubMed
62. Daw MS, Baskes MI. Semiempirical, quantum mechanical calculation of hydrogen embrittlement in metals. *Phys Rev Lett* 1983;50:1285-8. DOI
63. Mishin Y, Lozovoi A. Angular-dependent interatomic potential for tantalum. *Acta Materialia* 2006;54:5013-26. DOI
64. Mishin Y, Mehl M, Papaconstantopoulos D. Phase stability in the Fe-Ni system: investigation by first-principles calculations and atomistic simulations. *Acta Materialia* 2005;53:4029-41. DOI
65. Onat B, Cubuk ED, Malone BD, Kaxiras E. Implanted neural network potentials: application to Li-Si alloys. *Phys Rev B* 2018;97:094106. DOI
66. Zong H, Pilania G, Ding X, Ackland GJ, Lookman T. Developing an interatomic potential for martensitic phase transformations in zirconium by machine learning. *npj Comput Mater* 2018;4:48. DOI
67. Zhang L, Han J, Wang H, Car R, E W. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Phys Rev Lett* 2018;120:143001. DOI PubMed
68. Jia W, Wang H, Chen M, et al. Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning. SC20: International Conference for High Performance Computing, Networking, Storage and Analysis; 2020 Nov 9-19; Atlanta, GA, USA. IEEE; 2020. p. 1-14. DOI
69. Chen X, Gao X, Zhao Y, Lin D, Chu W, Song H. TensorAlloy: an automatic atomistic neural network program for alloys. *Computer Physics Communications* 2020;250:107057. DOI
70. Chen X, Wang L, Gao X, et al. Machine learning enhanced empirical potentials for metals and alloys. *Computer Physics Communications* 2021;269:108132. DOI
71. Shang HH, Chen X, Gao XY, et al. TensorKMC: kinetic monte Carlo simulation of 50 trillion atoms driven by deep learning on a new generation of Sunway supercomputer. The International Conference for High Performance Computing, Networking, Storage and Analysis (SC' 21); St. Louis, Missouri. New York, NY, USA: Association for Computing Machinery; 2021. p. 1-14. DOI
72. Bramer M. Principles of data mining. London: Springer: 2016. DOI
73. Ball P. Four decades of materials developments transform society. *MRS Bull* 2013;38:873-85. DOI
74. de Walle A, Nataraj C, Liu Z. The thermodynamic database database. *Calphad* 2018;61:173-8. DOI
75. Zou C, Li J, Wang WY, et al. Integrating data mining and machine learning to discover high-strength ductile titanium alloys. *Acta*

Materialia 2021;202:211-21. DOI
76. Wang WY, Shang SL, Wang Y, et al. Atomic and electronic basis for the serrations of refractory high-entropy alloys. *npj Comput Mater* 2017:3. DOI
77. Wang WY, Darling KA, Wang Y, et al. Power law scaled hardness of Mn strengthened nanocrystalline AlMn non-equilibrium solid solutions. *Scripta Mater* 2016;120:31-6. DOI
78. Umehara M, Stein HS, Guevarra D, Newhouse PF, Boyd DA, Gregoire JM. Analyzing machine learning models to accelerate generation of fundamental materials insights. *npj Comput Mater* 2019;5:34. DOI
79. Liu ZK, Wang Y, Shang S. Thermal expansion anomaly regulated by entropy. *Sci Rep* 2014;4:7043. DOI PubMed PMC
80. Rickman JM, Chan HM, Harmer MP, et al. Materials informatics for the screening of multi-principal elements and high-entropy alloys. *Nat Commun* 2019;10:2618. DOI PubMed PMC
81. Ye Y, Wang Q, Lu J, Liu C, Yang Y. High-entropy alloy: challenges and prospects. *Materials Today* 2016;19:349-62. DOI
82. Rost CM, Sachet E, Borman T, et al. Entropy-stabilized oxides. *Nat Commun* 2015;6:8485. DOI PubMed PMC
83. Bramer M. Decision tree induction: using entropy for attribute selection. Principles of data mining. London: Springer; 2016. p. 49-62. DOI
84. Bramer M. Decision tree induction: using frequency tables for attribute selection. Principles of data mining. London: Springer; 2016. p. 63-78. DOI
85. Liu ZK. Materials 4.0 and the Materials Genome Initative. *Adv Mater Process* 2020;178:50.
86. Chiarello F, Trivelli L, Bonaccorsi A, Fantoni G. Extracting and mapping industry 4.0 technologies using wikipedia. *Comput Ind* 2018;100:244-57. DOI
87. Qi Q, Tao F. Digital twin and big data towards smart manufacturing and industry 4.0: 360 degree comparison. *IEEE Access* 2018;6:3585-93. DOI
88. Qi Q, Tao F, Hu T, et al. Enabling technologies and tools for digital twin. *J Manuf Syst* 2021;58:3-21. DOI
89. Tao F, Qi Q, Wang L, Nee A. Digital twins and cyber-physical systems toward smart manufacturing and industry 4.0: correlation and comparison. *Engineering* 2019;5:653-61. DOI
90. Pettey C. Prepare for the impact of digital twins. Available from: https://www.gartner.com/smarterwithgartner/prepare-for-the-impact-of-digital-twins [Last accessed on 23 Feb 2022].
91. Lim KYH, Zheng P, Chen C, Huang L. A digital twin-enhanced system for engineering product family design and optimization. *J Manuf Syst* 2020;57:82-93. DOI
92. Mukhtarkhanov M, Perveen A, Talamona D. Application of stereolithography based 3D printing technology in investment casting. *Micromachines (Basel)* 2020;11:946. DOI PubMed PMC
93. Francalanza E, Borg J, Constantinescu C. A knowledge-based tool for designing cyber physical production systems. *Comput Ind* 2017;84:39-58. DOI