SCIENCE CHINA
Information Sciences

• POSITION PAPER •

January 2025, Vol. 68, Iss. 1, 110301:1–110301:19
https://doi.org/10.1007/s11432-024-4257-6

From CAS & CAE Members

When AI meets sustainable 6G

Xiaohu YOU[1,2*], Yongming HUANG[1,2*], Cheng ZHANG[1,2], Jiaheng WANG[1,2], Hao YIN[3] & Hequan WU[4]

[1]*National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China;*
[2]*Purple Mountain Laboratories, Nanjing 211111, China;*
[3]*Chinese Academy of Military Science, Beijing 100091, China;*
[4]*China Information Communication Technologies Group, Beijing 100083, China*

Received 11 November 2024/Accepted 18 December 2024/Published online 24 December 2024

Abstract Sixth-generation (6G) networks are anticipated to achieve transformative advancements, characterized by extreme connectivity, deep integration with artificial intelligence (AI) and sensing, and airground integration. The evolution of 6G exhibits two major trends: ubiquitous intelligence and sustainability. The former aims to embed state-of-the-art AI technology into the 6G network, from the physical layers to applications, while the latter emphasizes reducing energy consumption while enhancing network performance to address environmental concerns. Despite the amazing progress in recent years, AI advancements come with substantial increases in data and computational overhead, posing critical challenges for integrating AI into sustainable 6G networks. First, high energy consumption from large datasets and heavyweight AI models contradicts 6G's green goals. Second, the precise collection of large datasets, message delivery latency, and inference delays in AI models pose challenges for real-time tasks in 6G. Third, the uninterpretability and unpredictability of AI models complicate meeting the stringent requirements for controllable transmission in dynamic wireless environments. Addressing these challenges and achieving sustainable 6G with ubiquitous intelligence calls for a revolutionary design of 6G architecture and AI frameworks. To this end, this paper introduces a novel and practical methodology for green, real-time, and controllable 6G native intelligence, starting with knowledge graph (KG) analysis to extract small but critical datasets, followed by the development of distributed lightweight AI models, and the use of digital twins (DTs) to create precise replicas of physical 6G networks. This leads to a pervasive multi-level (PML)-AI framework supported by a task-centric, three-layer 6G architecture. The AI framework operates through non-real-time and real-time cycles, leveraging three key technologies: wireless data KGs for efficient data management, lightweight AI models for sub-millisecond real-time responsiveness, and DTs for AI pre-validation. A prototype system is built on the proposed 6G architecture and PML-AI framework, and experimental results show that data overhead is significantly reduced and real-time intelligence at the millisecond level can be realized.

Keywords artificial intelligence, sustainable 6G, green, real-time, controllable

Citation You X H, Huang Y M, Zhang C, et al. When AI meets sustainable 6G. Sci China Inf Sci, 2025, 68(1): 110301, https://doi.org/10.1007/s11432-024-4257-6

1 Introduction

Sixth-generation (6G) mobile network represents the next critical milestone in global connectivity, envisioned as the infrastructure for a highly intelligent and extensively interconnected world. Unlike its predecessors, 6G is anticipated to mark a fundamental paradigm shift rather than a merely incremental enhancement, unlocking new frontiers such as holographic communication and autonomous systems at the cutting edge of technological innovation [1, 2]. Compared to the fifth-generation (5G), 6G will bring magnitude increase in terms of peak data rate, spectrum efficiency, latency, and reliability, providing extreme communication connectivity. Other 6G key capabilities include artificial intelligence (AI), enhanced sensing and precise localization, and full coverage and ubiquitous access supported by air-ground integration [3]. These capabilities can effectively support the six key usage scenarios identified by the International Telecommunication Union (ITU), e.g., immersive communication, massive connectivity, and integrated AI/sensing and communication, as well as diverse new applications such as extended reality (XR) and pervasive sensing ecosystems [1].

Ubiquitous intelligence is regarded as a crucial feature of 6G networks, where AI is expected to play the most important role in managing the increasing complexity and dynamism of future communication

* Corresponding author (email: xhyu@seu.edu.cn, huangym@seu.edu.cn)

systems [3, 4]. AI will empower precise adaptation of 6G network resources, facilitating autonomous network management, real-time customized services, and end-to-end high quality of service (QoS) optimization. In turn, 6G networks will also provide flexible and scalable infrastructure to support AI-driven applications, e.g., the deployment of general large AI models through cloud-edge-end collaboration [1]. Recently, AI has made significant advances in various fields such as natural language processing and computer vision. In particular, large language models, e.g., OpenAI's GPT, have revolutionized natural language processing and broadened AI's impact across various domains [5]. However, this progress came at the cost of dramatically escalating computational investments, driven by the growing demand for data and computation-intensive models [6].

Sustainability is another major demand for 6G development, emphasizing the need to reduce energy consumption while enhancing network capabilities. This will align with global sustainability initiatives, such as the Paris Agreement, and address growing environmental, social, and economic concerns. By embedding sustainability throughout the network's lifecycle, 6G supports long-term environmental stewardship and global sustainability goals as network infrastructure evolves and expands [1]. However, the integration of AI into sustainable 6G networks faces many challenges, especially the green, real-time, and controllable requirements. First, the enormous energy required for acquiring, storing, and utilizing large datasets, coupled with the substantial power consumption of heavyweight AI models during both training and inference, conflicts with 6G's green requirements. Second, the precise collection of large datasets, the latency in data and control message transmission, and the inference delays of heavyweight AI models collectively exacerbate the overall latency in real-time processing. Third, the uninterpretability and unpredictability of AI models and the rapidly fluctuating wireless environment make it hard to guarantee controllable performance, thus increasing the risk of performance lapses and inconsistent service quality.

Due to these facts, the concept of green AI has drawn extensive attention to develop more efficient training and inference strategies. Training efforts focus on enhancing initialization techniques, regularization methods, progressive training, and advancements in efficient automated machine learning (AutoML) [6]. Inference strategies, including model pruning, distillation, low-rank factorization, and quantization, intend to reduce computational load while maintaining model accuracy [7, 8]. Meanwhile, the academic and industrial societies also highlight the importance of real-time intelligence and pay increasing efforts on embedding intelligence at the edge and utilizing edge computing for real-time decision-making closer to data sources [9, 10]. These approaches leverage edge devices and base stations for localized inference, thereby reducing latency and network overhead. Note that, however, current research lacks a comprehensive framework to achieve performance-guaranteed, green, and real-time intelligence through a closed-loop system design. To achieve controllable performance through AI models, some researchers try to merge AI with traditional optimization methods and the application of classical statistical learning theories to mitigate the uninterpretability and unpredictability [11–13], while others attempt to combine robustness testing with simulation testing to ensure the controllability of AI systems [14]. Nevertheless, the efficient construction of controllable testing environments for such systems remains an open question.

Addressing the aforementioned limitations and fulfilling the deep integration of AI in sustainable 6G networks require a fundamental rethinking of both network architecture and AI frameworks. As an effort to seek promising solutions, we consider leveraging knowledge graph (KG) data analysis to extract key features to form small feature datasets, developing lightweight AI models based on these feature datasets, and employing digital twins (DTs) to create accurate replicas of physical 6G networks, consequently forming an innovative and practical roadmap towards green, real-time, and controllable 6G native intelligence. Specifically, we introduce a pervasive multi-level (PML)-AI framework supported by a task-centric, three-layer 6G network architecture. The PML-AI framework is built around three key components-data, models, and algorithms. Starting from the data source, it emphasizes the effective use of wireless big data. By extracting a minimal yet highly effective feature dataset closely connected to the network AI performance from massive wireless data, this framework supports subsequent lightweight AI models, thereby reducing computational costs. To ensure the real-time performance and efficiency of lightweight models, the PML-AI framework supports a layered distributed deployment of the models, which is conducive to achieving full-element endogeneity of data, computing power, and control, and simplifies the AI learning challenges in the real-time layer.

Based on the above considerations, the PML-AI framework is specifically built upon a dual-cycle structure. The outer cycle, operating in non-real-time, generates key performance indicator (KPI)-driven feature datasets using the wireless data KGs, enabling efficient use of lightweight AI models. The inner cycle, running in real-time, conducts AI training and inference on these reduced-scale datasets, facilitating

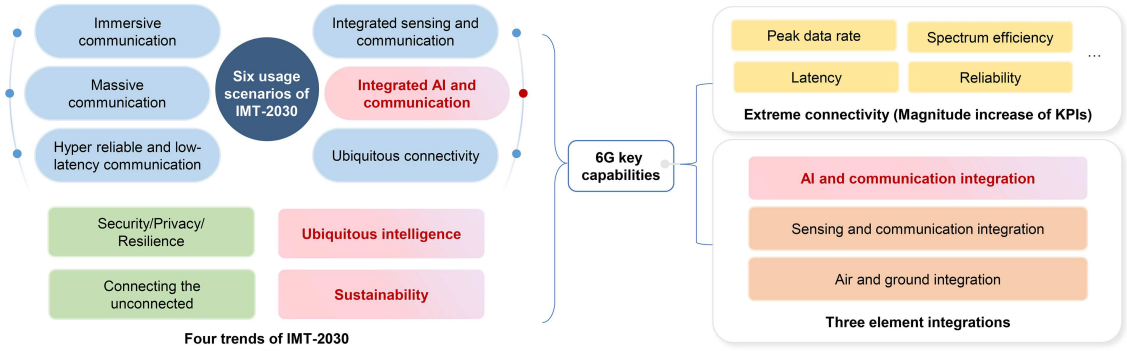

Figure 1 Six usage scenarios, four trends, and key capabilities of IMT-2030 [1].

real-time intelligent network resource allocation at the millisecond level ($\leqslant$ 1 ms). This ensures energy-efficient, low-cost, and agile network intelligence. Additionally, wireless DT based on generative learning is used to create a virtual replica of the 6G network, enabling the pre-validation of AI models and ensuring high-quality QoS. A real-time intelligent prototype verification system is built on the 6G integrated test platform, where a QoS-aware real-time intelligent resource allocation experiment is conducted to validate the proposed technological roadmap. The results demonstrate nearly an order-of-magnitude reduction in data and computational overhead, time-slot-level intelligent control, and significantly boosted system performance, e.g., system throughput and supported user numbers, compared to current 5G benchmarks.

2 6G vision and trends

The advent of 6G represents not just an evolutionary step but a paradigm shift in mobile communications, characterized by six key usage scenarios that empower a wide range of advanced applications. Beyond incremental improvements, 6G brings forth several critical trends, such as sustainability and ubiquitous intelligence, while pushing connectivity performance to extreme levels and integrating various elements including communication, sensing, AI, and space-air-ground networks. These advancements necessitate the development of novel network architectures capable of supporting the complexity and demands of this next-generation ecosystem, enabling seamless and intelligent interactions across diverse domains.

The vision for 6G encompasses a significant expansion of the foundational scenarios introduced by 5G, to meet the needs of an increasingly interconnected and data-driven world. Six key usage scenarios define the 6G landscape [1]: immersive communication, massive communication, hyper reliable and low-latency communication, ubiquitous connectivity, integrated AI and communication, and integrated sensing and communication, as illustrated in Figure 1. These scenarios will facilitate a range of innovative applications, from immersive XR experiences and holographic interactions to autonomous collaboration among devices and pervasive sensing networks. Each scenario reflects the growing demand for more sophisticated and seamless digital experiences, driving the development of new business models and service paradigms that will shape the future of communication.

The ambitious vision of 6G is built upon a set of demanding KPIs that significantly surpass those of previous generations. Central to these KPIs is the TKμ extreme connectivity performance, including Tbps-scale data rates, Kbps/Hz-scale spectral efficiency, and µs-scale latency [2]. These capabilities will enable seamless and high-capacity communication, ensuring superior connectivity and robust performance across diverse advanced applications, including immersive media and complex industrial processes. Additionally, 6G will introduce new capabilities based on element integration including AI-related capabilities, sensing-related capabilities, coverage, positioning accuracy, and interoperability. To meet these objectives, 6G will incorporate cutting-edge technologies such as AI, extremely massive multiple-input multiple-output (MIMO), and Terahertz (THz) communication [1, 3].

Sustainability is recognized as a foundational trend in the development of 6G technology. While the capability of the 6G network will increase by at least one magnitude, the energy consumption per bit should be reduced by at least one magnitude. Ensuring sustainable development is imperative for the long-term success of 6G. This also caters to the increasing demand for environmental, social, and economic sustainability, aligning with global initiatives like the Paris Agreement under the United Nations Framework Convention on Climate Change. Beyond reducing energy consumption, sustainable 6G also

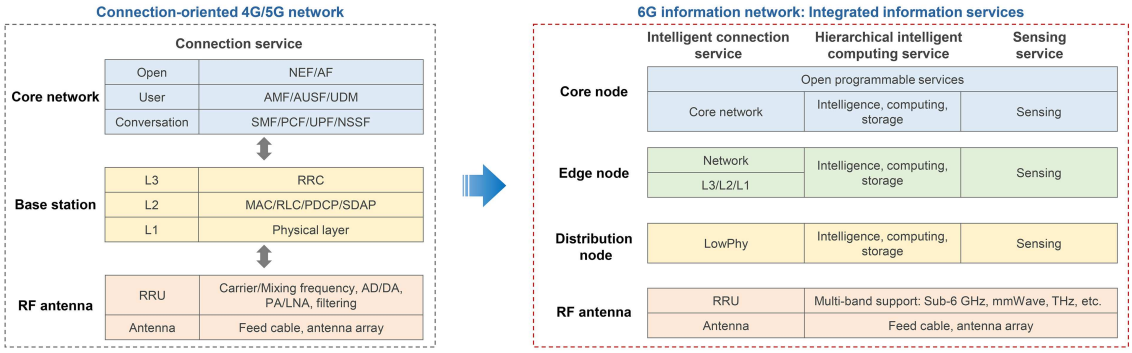

Figure 2 Task-centric three-layer 6G network architecture in [2].

seeks to integrate green technologies and promote digital inclusivity, ensuring that the benefits of advanced connectivity are accessible and environmentally responsible [1]. This includes adopting energy-efficient hardware and network management to optimize resource usage dynamically, thereby minimizing the carbon footprint of 6G deployments.

Ubiquitous intelligence rises as another significant trend of 6G evolution. The integration of AI is not merely an enhancement but a fundamental necessity for 6G to realize its full potential. Traditional network architectures lack the capability to manage the complexity and dynamism of future communication environments, where responsiveness and adaptability are crucial. AI provides the intelligence required for predictive maintenance, dynamic resource allocation, and efficient decision-making, all of which are critical for ensuring seamless user experiences across increasingly diverse and complex applications. 6G will break the boundaries between traditional network architectures and intelligent applications by integrating AI and communication, as well as sensing and communication, to build a novel intelligent communication ecosystem. The integration of AI and 6G networks can be approached from two perspectives: AI for 6G networks and 6G networks for AI [1, 3, 4]. AI for 6G networks facilitates advanced autonomous driving, customized services within the 6G network, and the optimization of end-to-end QoS at the lowest cost for a diverse range of applications. Conversely, 6G networks for AI enable flexible customization of connections, computing, and data for AI applications, potentially supporting the deployment of general large AI models through cloud-edge-end collaboration.

To achieve the extreme connectivity targets of 6G networks and support advanced scenarios related to sustainability and ubiquitous intelligence, a task-centric three-layer network architecture was proposed in [2]. This architecture extends the traditional two-layer cellular architecture, comprising the core network and base stations, into a three-layer architecture, consisting of core nodes (CoNs), central nodes, and distributed nodes, as illustrated in Figure 2. By leveraging distributed massive cell-free MIMO technology and utilizing higher frequency bands, it achieves Tbps-scale data rate and Kbps/Hz-scale spectral efficiency [15–17]. Additionally, the incorporation of spatiotemporal two-dimensional (2D) channel coding reduces latency to the microsecond scale [18]. The coordinated orchestration of resources across these three layers optimizes computing and storage, facilitating the seamless integration of AI and communication, as well as sensing and communication. This architecture offers significant potential for supporting full-service, full-scenario applications, providing valuable direction for future 6G research and development.

3 Ubiquitous AI for sustainable 6G: challenges and solutions

AI has undergone significant technological transformations over the past decades, characterized by both remarkable advancements in technology and exponential growth in data and computational requirements. Starting with early work in symbolic reasoning and expert systems in the mid-20th century, AI research laid the foundation for machine learning (ML) models that would later revolutionize the field. The rise of deep learning, particularly with models like AlexNet in 2012, marked a pivotal shift, driving breakthroughs in image classification, natural language processing, and even text-to-video generation. Later, advanced AI systems like OpenAI's ChatGPT and Sora pushed the boundaries of what AI can achieve. On the other hand, however, the development of such big AI models results in a dramatic increase in computational demands and dataset sizes, as illustrated in Table 1 [19]. For instance, the

Table 1 AI roadmap and trends: from AI to AGI [19].

Year	Model	Number of parameters	Volume of training data	Arithmetic requirement
2014	AlexNet	60M	ImageNet (∼1.5 million images)	$\sim 1.4 \times 10^{10}$ FLOPs
2015	ResNet-50	25.6M	ImageNet (∼1.5 million images)	$\sim 3.8 \times 10^{10}$ FLOPs
2017	Transfomer	65M	WMT2014 (∼1.5 million images)	$\sim 3.8 \times 10^{13}$ FLOPs
2018	BERT (base)	110M	BookCorpus+Wikipedia	$\sim 3.3 \times 10^{16}$ FLOPs
2019	GPT-2	1.5B	40 GB web text	$\sim 1.5 \times 10^{18}$ FLOPs
2020	GPT-3	175B	570 GB web text	$\sim 3.14 \times 10^{23}$ FLOPs
2021	DALL-E	12B	250M text-image pairs	Unpublished (estimated at $\sim 10^{20}$ FLOPs)
2021	Gopher	280B	10 TB of text data	$\sim 2.5 \times 10^{23}$ FLOPs
2022	PaLM	540B	780 GB of text data	$\sim 1.56 \times 10^{24}$ FLOPs
2023	GPT-4	Unpublished	Unpublished (approximately hundreds of GigaBytes or more)	Unpublished (estimated to be over $\sim 10^{24}$ FLOPs)

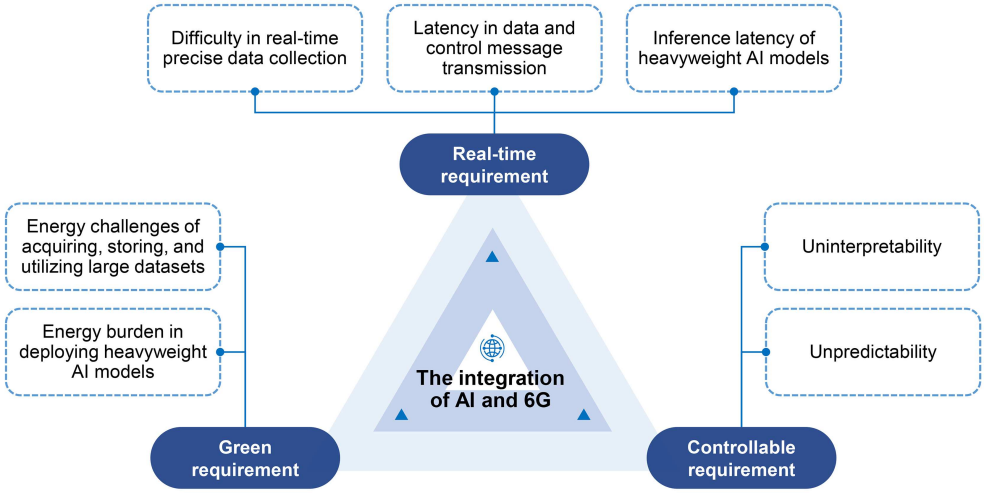

Figure 3 Three challenges faced by the integration of AI and 6G.

cost of training deep learning models increased by 7.14×10^{13} times between 2014 and 2022, highlighting both the progress and the growing operational challenges [5, 20].

Nowadays, AI has become a crucial driver of technological innovation, revolutionizing how we handle complex data and automate processes. The advanced algorithms and models developed through AI research have led to significant breakthroughs across various fields, including healthcare, finance, and autonomous systems. As the era of 6G is coming, integrating AI into this next-generation communication network architecture becomes increasingly critical. In 6G, AI's role extends beyond traditional network management to include real-time decision-making, predictive maintenance, and dynamic resource allocation, all of which are essential for ensuring seamless user experiences in increasingly complex and heterogeneous network environments. However, the "miracle-driven" development paradigm of AI is at odds with the vision of a sustainable 6G future, particularly in meeting green, real-time, and controllable requirements. These challenges, as listed in Figure 3, will be examined in detail in the following.

Challenge 1—Green requirement. As AI continues to evolve, the pursuit of more sophisticated and data-intensive models presents a significant challenge for the integration of AI with 6G. The increased computational power required for advanced AI models conflicts with 6G's green requirements, which calls for reduced energy consumption and improved efficiency. The key lies in balancing the need for advanced AI capabilities with the imperative to minimize energy consumption. High-performance AI systems not only consume substantial power during model training but also have ongoing energy needs throughout their lifecycle. Specifically, the green requirement faces the following two difficulties.

• **Energy challenges of acquiring, storing, and utilizing large datasets.** The traditional AI paradigm, which relies heavily on large datasets, presents significant energy challenges. The process of acquiring large datasets necessitates the continuous operation of numerous sensors and devices, which consumes substantial energy. Once collected, storing these vast amounts of data requires significant

space and energy-intensive data centers. Furthermore, the computational power needed for processing and analyzing this data is considerable, particularly for complex, large-scale computations.

• **Energy burden in deploying heavyweight AI models.** The use of heavyweight AI models also adds to the energy burden, especially during model training and inference. These models demand vast computational resources, including GPUs and TPUs, which consume substantial energy. As AI models become increasingly complex, the ongoing energy costs for their maintenance and updates also rise. Addressing this issue involves developing and implementing more energy-efficient AI techniques, such as model pruning, quantization, and the construction of lightweight models.

Challenge 2—Real-time requirement. Meeting the real-time requirement in 6G networks is a formidable challenge due to the need for ultra-fast processing and decision-making in highly dynamic and latency-sensitive environments. 6G networks are expected to support agile edge-access services with rapid resource scheduling to accommodate differentiated QoS needs, particularly in hyper-reliable low-latency communications scenarios like machine interactions, emergency services, and telemedicine, where even minor delays can have severe consequences. Moreover, real-time applications within radio access networks (RANs), such as dynamic traffic management, autonomous troubleshooting, and adaptive QoS provisioning, all require immediate data processing and decision-making to maintain network performance and user experience. To meet these demands, 6G must achieve slot-level, ms-scale real-time intelligence, which presents the following key challenges.

• **Difficulty in real-time precise data collection.** 6G networks generate vast amounts of heterogeneous, redundant, and dynamic data, including channel state information (CSI), network topology changes, and user behaviors. The precise collection and effective management of this data pose significant challenges, impacting the ability to achieve accurate and timely data acquisition.

• **Latency in data and control message transmission.** Current networks employ external AI frameworks, where data collection, model inference, and control commands are processed across different network elements. This setup introduces transmission delays of tens of milliseconds between nodes. Such latency, along with the sequential execution of tasks, hampers the achievement of slot-level, real-time intelligence, particularly in applications demanding microsecond-scale responsiveness.

• **Inference latency of heavyweight AI models.** The deployment of heavyweight AI models exacerbates the challenge of meeting real-time requirements due to their significant computational demands. These models require substantial processing power for both training and inference, which can lead to delays in real-time applications. The complexity of such models necessitates extensive computations, potentially resulting in slower response times.

Challenge 3—Controllable requirement. The controllable requirement in 6G networks is a significant concern due to the complex and dynamic nature of both the wireless environment and AI-driven systems. The rapid variation of wireless channels and increasingly sophisticated traffic models create an environment filled with fluctuations and uncertainties, making consistent, controllable performance difficult to achieve. This challenge is exacerbated by the inherent unpredictability of AI models, which often make real-time decisions under volatile and non-stationary conditions. These factors increase the risk of performance lapses, potentially leading to significant, even catastrophic, consequences, particularly in mission-critical applications where continuous, high-quality service is essential. Therefore, ensuring reliability in such a dynamic and uncertain environment necessitates addressing several key considerations.

• **Uninterpretability.** Many AI models function as "black boxes", obscuring their internal decision-making processes and complicating the prediction of their performance. This lack of transparency poses challenges for ensuring that AI models consistently meet controllable requirements, particularly in mission-critical applications such as industrial control systems where reliability is paramount.

• **Unpredictability.** The fluctuating nature of wireless channels and various traffic in 6G networks challenges the consistency of AI model performance. This unpredictability, driven by the non-stationary nature of wireless environments, complicates the deployment of AI models that require stable and predictable behavior, particularly in applications demanding strict QoS requirements.

To meet the green, real-time, and controllable requirements, a fundamental transformation in 6G's network architecture and AI frameworks is required. Traditional AI approaches, characterized by large-scale data processing and heavyweight models, are insufficient to address the stringent demands of sustainable 6G networks. To bridge this gap, in the following, we propose a novel PML-AI framework that is specifically tailored to provide a sustainable, responsive, and performance-guaranteed solution for 6G.

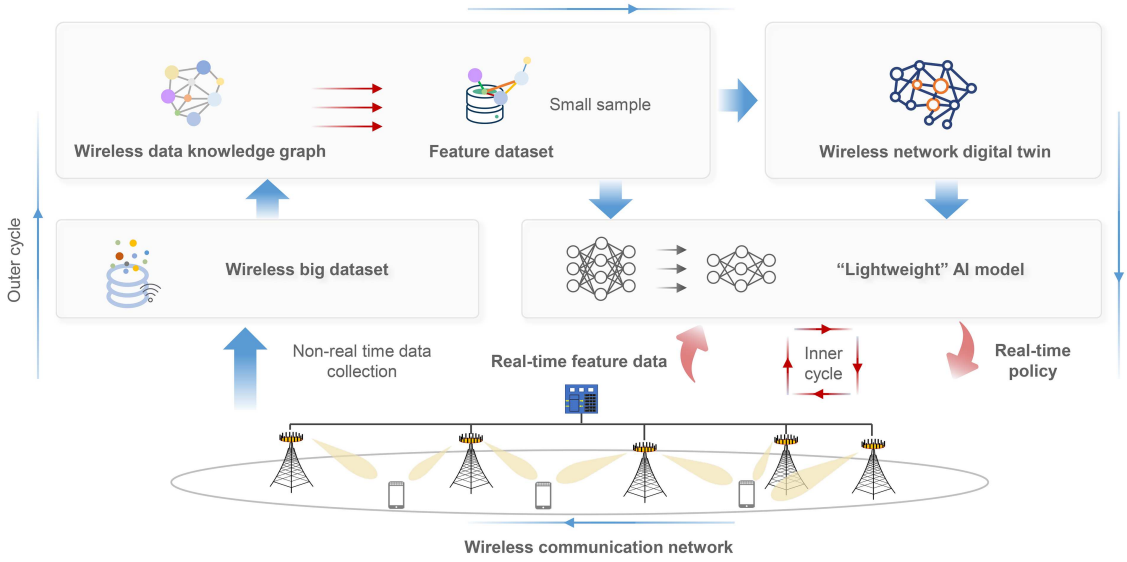

Figure 4 Pervasive multi-level (PML)-AI framework.

4 PML-AI framework for 6G

To address the three primary challenges in integrating AI with 6G-green, real-time, and controllable requirements, we consider leveraging the wireless data KG technology to guide the precise utilization of key data, facilitating AI optimization driven by small but crucial datasets. The construction of lightweight AI models further enhances real-time performance, forming a real-time intelligent provisioning capability of network resources at the slot level ($\leqslant$ 1 ms). Additionally, DTs are embedded within the network to enable AI pre-validation and ensure high-quality QoS through AI-driven optimization. Motivated by this thinking, we propose a dual-cycle architecture, as illustrated in Figure 4. The non-real-time outer cycle focuses on long-term optimization and learning from historical data, while the real-time inner cycle emphasizes immediate response and real-time adjustments.

In the outer cycle, wireless big data, which encompasses a vast array of disorganized, complex, and interrelated data fields, is collected in a non-real-time manner. We construct a semi-dynamic wireless data KG to characterize and analyze the relationships among extensive data fields, thereby enabling the organization and clear interrelation of these data fields. This facilitates the generation of minimal yet indispensable datasets, consisting of significantly reduced feature data that have the greatest impact on target KPIs, thereby forming a KPI-oriented feature dataset. Furthermore, a wireless network DT can be constructed based on the feature dataset, which can serve as a faithful replica of a physical 6G network. This enables the pre-validation of AI through the DT, ensuring the safe deployment of network AI, even when the AI model exhibits unpredictability and unexplainability. The removal of a considerable amount of field redundancy in the feature dataset allows the DT to achieve enhanced generalization performance, specifically in terms of modeling accuracy.

The inner cycle, guided by the outer cycle, performs the real-time collection of a significantly reduced-scale feature dataset and conducts real-time AI training and inference. This allows for the efficient implementation of real-time AI for wireless networks. Specifically, the requirement for real-time data collection is minimized to a small number of key data fields, facilitating the training of lightweight AI models. This strategy reduces the costs associated with data collection and computation, thereby enabling the realization of real-time and green network intelligence. Consequently, a considerably lesser amount of computing power is required, and real-time processing becomes significantly more straightforward. Furthermore, before the online deployment of real-time AI inference results, they must undergo a pre-validation process via the wireless network DT module. Deployment is contingent upon meeting the requisite performance requirements.

To summarize, the new framework of PML possesses three key features. First, the utilization of KG to guide key feature data facilitates accurate data utilization. Second, the employment of critical small data to drive lightweight AI models enables real-time intelligent optimization. Finally, the application of DTs to support AI pre-validation ensures a high QoS guarantee for the 6G network. Three key technologies

encompassed for the implementation of the proposed PML-AI framework are detailed below.

4.1 Wireless data KG and feature dataset generation

To achieve green intelligence, reducing the energy consumption associated with data collection and AI model training is essential. Wireless networks generate a vast amount of data fields and metrics during operation, but only a fraction of these significantly impact the performance of network AI models. Therefore, the green intelligence of communication systems largely depends on a small subset of critical data that has a profound influence on these models. Consequently, effective classification, analysis, and feature extraction from diverse data types, as well as the generation of minimal but effective datasets (referred to as feature datasets) tailored to different application needs, are crucial for driving AI training, inference, and validation. To this end, we draw on the KG technology, combining wireless big data with expert knowledge in wireless communications to innovatively propose a wireless data KG. This technique enables the generation of feature datasets that significantly impact KPIs, paving the way for real-time performance and green intelligence.

Using wireless data KGs can effectively represent the complex relationships between diverse wireless data. This not only helps us gain a deeper understanding of the operating mechanisms of wireless networks, but also provides strong support for in-depth mining of wireless data. Wireless data KGs facilitate this by providing a comprehensive view of network performance across various dimensions, such as network, terminal, user, and service, thus bridging the gap between different network modules like core and radio access networks. This understanding is critical for addressing the challenges associated with green and real-time AI, as it enables detailed analysis and extraction of valuable knowledge, thereby revealing key insights into network efficiency and environmental impact. Despite their potential, current research on wireless data KGs is still in its early stages, with notable deficiencies in construction and representation learning methodologies. These methodologies must accommodate the unique attributes of wireless data, where each node is linked to extensive metrics, necessitating a dual-driven approach that integrates both knowledge and data. Effective categorization, analysis, and feature extraction from these datasets are essential to generate minimal yet crucial feature datasets that support AI training and validation. Our proposed approach seeks to bridge these gaps by improving the construction of wireless data KGs and leveraging feature data mining to optimize feature dataset generation.

A wireless data KG captures the interrelationships between various environmental factors, device properties, and the complete protocol stack within wireless communication networks. It characterizes and analyzes the dynamic relationships among extensive data fields, reflecting wireless communication protocols and principles. However, due to the ever-changing wireless network environment, which varies over time and across locations, these relationships are not static but evolve based on specific temporal, service, and locational characteristics. In light of these observations, we propose the definition of two distinct categories of wireless data KG: the basic wireless data KG and the learned wireless data KG. The basic wireless data KG represents foundational wireless communication protocols and communication principles. In contrast, the learned wireless data KG captures the dynamic and rapidly evolving relationships among wireless data fields in the real physical world. To construct a learned wireless data KG with spatio-temporal dynamics, one should utilize wireless big data to learn and update on the basis of the basic wireless data KG.

Based on the basic wireless data KG, knowledge representation learning methods, e.g., the STREAM proposed in [21], explore the semantic attributes of nodes, relationships, and the underlying graph structure, thus enabling efficient computations of intricate semantic associations among entities and relationships. Specifically, they map the semantics inherent to relationships or entities, along with the real-world data they encapsulate, into compact low-dimensional vectors. Such approaches offer several advantages: (1) They ensure computational efficiency in subsequent operations, as vector-based computations can be expedited through readily accessible acceleration algorithms. (2) They effectively compress verbose information. (3) They simplify the integration of data from multiple sources. As a result, the knowledge representation learning of wireless KGs substantially streamlines downstream tasks. Through the knowledge representation learning, we get the learned wireless data KG.

As shown in Figure 5, to identify a subset of critical nodes from a large volume of wireless data fields that have the most substantial impact on the target KPI, we employ the above-constructed wireless data KG for node selection. In this framework, each node represents a feature related to the KPI node. Initially, the graph structure is used to identify all paths connecting to the KPI. The influence of each

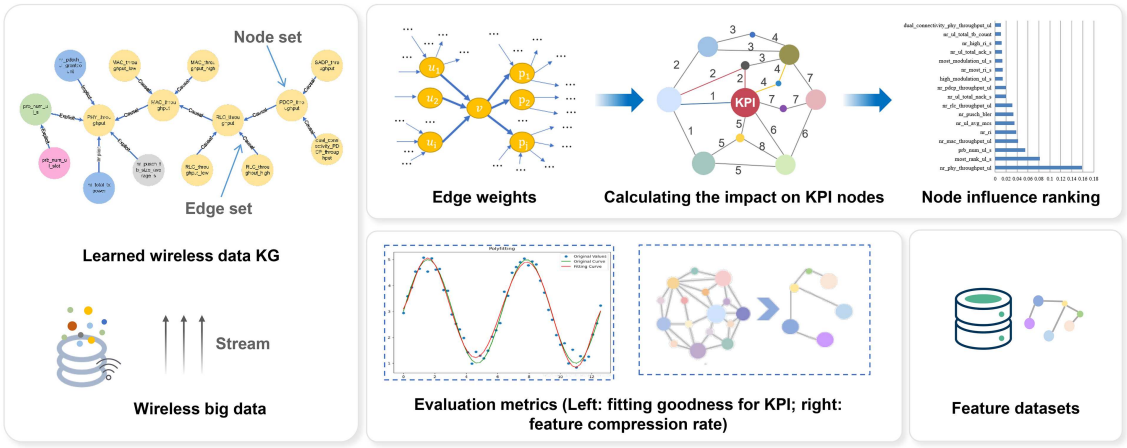

Figure 5 Wireless data KG-based feature dataset generation.

node on the KPI is then determined based on the relationship among neighboring nodes on the paths. The strength of these relationships can be quantified by measuring node similarity, which is accomplished through the calculation of the cosine similarity between the node representation vectors. Subsequently, the nodes are ranked according to their impact on the KPI to guide the selection of nodes.

Once the most significant nodes have been identified based on their impact on the KPI, the feature dataset is then evaluated to ensure that the data features identified are both minimal and crucial. This evaluation involves two key metrics. (1) Goodness of fit. This metric involves utilizing the selected nodes and the corresponding collected data to predict the target KPI and calculating the discrepancy between the predicted values and the actual values. It is crucial to maintain a satisfactory level of goodness of fit to meet the requirements of real-world scenarios. (2) Feature compression ratio. This metric evaluates the effectiveness of feature selection by aiming to maximize the retention of critical information while minimizing redundancy. The goal is to ensure that the feature dataset contains the most essential information with the fewest possible features. The goodness of fit metric ensures the high quality of the selected features, while the feature compression ratio metric focuses on minimizing the number of features. Together, these metrics provide a balanced approach to optimizing both the quantity and quality of features in the dataset.

To validate the advantages of the feature dataset, an experiment based on real collected data [21] is introduced in the following. In the initial stage, we filter 82 data fields from a pool of 201, creating a wireless data knowledge graph focused on uplink throughput. Then, following the aforementioned feature dataset generation process, we obtain a feature dataset for uplink throughput, as shown in Table 2. This feature dataset offers several advantages. Firstly, regarding the uplink throughput KPI, the original dataset comprising 201 data fields is streamlined to only 4 data fields. This drastic reduction eliminates extraneous nodes, allowing subsequent research on uplink throughput in real network environments to concentrate on essential data fields. Secondly, while retaining maximum information transmission, the size of the feature dataset is minimized to facilitate efficient data transmission. According to the experimental results, we achieve a fit of 97.36% with a dataset reduced by about 97.9% in scale. Lastly, achieving real-time intelligence in wireless networks requires minimizing computational costs to avoid latency and energy wastage. During the training of downstream AI algorithms using the feature dataset, the number of parameters is reduced by about 71.87%, and both the floating point operations (FLOPs) and execution time are reduced by nearly an order of magnitude. These results indicate a significant reduction in computational overhead, providing preliminary support for the subsequent implementation of green intelligence.

4.2 Feature dataset powered lightweight AI

Real-time intelligence is essential for achieving ubiquitous AI in 6G. Due to the time-varying nature of wireless channels in the space-time-frequency domain, the RAN performs user scheduling and resource allocation on a millisecond-scale granularity, known as transmission time intervals (TTIs) or time slots. For example, in 5G, a time slot is 0.5 ms when the subcarrier spacing is 30 kHz. To realize the vision

Table 2 Performance and cost comparison of AI models based on raw dataset and feature dataset.

	AI models based on raw dataset	AI models based on feature dataset
Number of feature	188	4
Fitting degree	99.97%	97.36%
Model parameters	8193	2305
FLOPs (G)	1.63×10^{-5}	4.51×10^{-6}
Execution time (s)	465.75	28.33

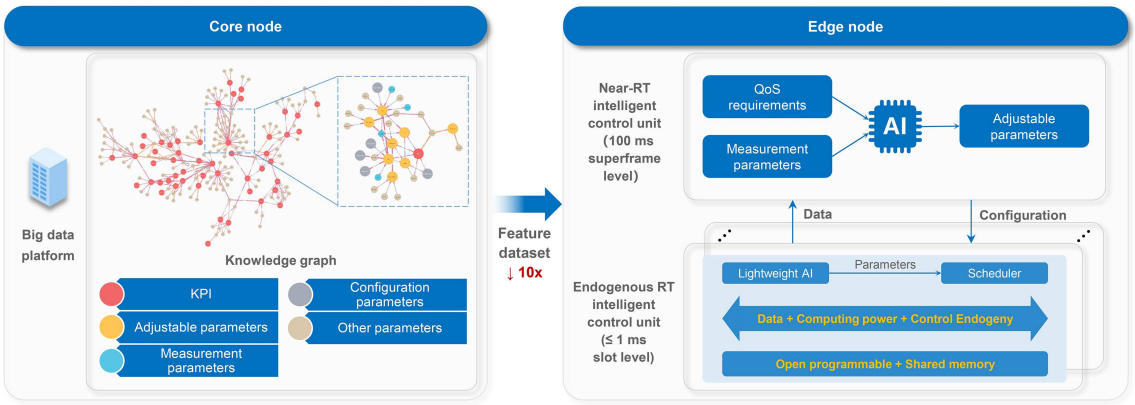

Figure 6 6G hierarchical endogenous intelligent communication system, where RT refers to real-time.

of ubiquitous AI in 6G, slot-level, millisecond-scale real-time intelligence is needed, particularly for user scheduling and resource allocation.

Achieving real-time intelligence requires endogenizing all elements of network intelligence, including data, computing power, and control. Traditional plug-in AI involves transferring data from the RAN to edge servers and sending control strategies back, with delays in data transfer and strategy delivery often reaching tens of milliseconds. To mitigate these delays, it is essential to realize endogenous data, computing, and control within a single network element, such as distributed units (DUs), thereby eliminating communication delays between elements. Lightweight AI models are also critical, as they reduce the requirements on data volume and computing power for model training/inference, minimizing time overhead within network elements.

The accurate use of data underpins lightweight AI models. Identifying critical data from large datasets maintains model performance while reducing data processing time. As illustrated in Figure 6, we propose a feature dataset-driven, hierarchical intelligence architecture based on the PML-AI framework and wireless data KG technology. In this architecture, CoN uses wireless data KGs to construct feature datasets that guide data utilization and enable real-time intelligence at the edge. Edge nodes, leveraging hierarchical and distributed AI, deploy lightweight AI models to achieve real-time intelligence via the endogeny of all network elements.

The CoN is responsible for utilizing the wireless KG to construct feature datasets, guiding accurate data utilization, and driving real-time intelligence at edge nodes. A general KG is constructed at the CoN based on non-real-time data collection and analysis. From this, specialized KGs tailored to specific scenarios and KPIs can be extracted, allowing for the construction of scenario-specific feature datasets. This approach balances generalization and scenario-specific adaptation, with the specialized KGs typically comprising KPIs, adjustable parameters, measurement parameters, configuration parameters, and other data, revealing relationships among these categories.

Edge nodes implement real-time intelligence in a hierarchical and distributed manner, utilizing the feature datasets. The nodes consist of a near-real-time intelligent control unit and multiple endogenous real-time intelligent control units. The near-real-time unit, responsible for network-level AI functions and operating on a superframe-level timescale (e.g., 100 ms), is typically deployed at a dedicated server at the network edge. The endogenous real-time units, responsible for network element-level AI functions, operate on a slot-level timescale (e.g., 1 ms) and are implemented within DUs.

The near-real-time intelligent control unit orchestrates multiple endogenous real-time units, aggregating data and leveraging its more powerful computing resources for tasks such as training lightweight

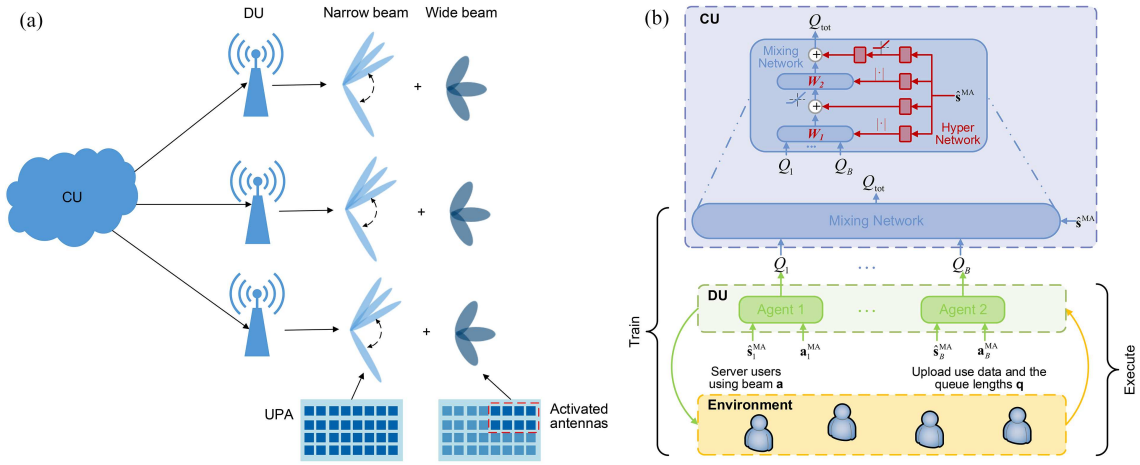

Figure 7 Real-time intelligence: a case study. (a) Wide beam-based narrow beam prediction; (b) proposed partially distributed beam selection.

models or collaborating in distributed AI training for endogenous real-time intelligent control units. For example, the near-real-time unit can host the critical model in a multi-agent reinforcement learning framework, coordinating local AI training across multiple real-time units. Additionally, network optimization tasks can be decomposed via assigning high-computation, latency-tolerant tasks or tasks requiring global information to the near-real-time unit, while latency-sensitive, low-computation tasks are handled by the endogenous real-time units. For instance, the near-real-time unit can use more powerful AI models and longer-term statistical data to pre-configure resources, reducing the complexity of real-time AI tasks at the real-time units.

Empowered by feature datasets and lightweight AI, the endogenous real-time intelligent control units perform real-time closed-loop network optimization via the endogeny of all elements. Feature datasets and lightweight AI reduce data collection and model sizes by an order of magnitude, enabling real-time data collection and inference. With open programmable technologies, data can be efficiently shared, and inference results can be efficiently acted on the scheduler through modular interfaces. This allows endogenous real-time units to efficiently complete the closed loop from data collection to network control, achieving real-time intelligence.

In the following, the delay-aware beam management for cell-free MIMO systems, as a typical use case of real-time resource allocation, is adopted to illustrate the above mentioned hierarchical and distributed intelligence architecture. Beam selection for joint transmission in cell-free MIMO faces challenges like high training overhead, computational complexity, and varying traffic-aware QoS requirements. Our previous work [22] effectively addresses these challenges by employing a hierarchical, distributed, and intelligence-empowered real-time beam selection approach.

The proposed approach employs a hierarchical architecture with distinct timescales to manage the beam selection process, as shown in Figure 7(a). In the long-timescale, the centralized unit (CU) aggregates wide beam responses from multiple DUs and uses a convolutional neural network (CNN) to predict narrow beam power profiles, effectively reducing the candidate beam space. In the short-timescale, DUs utilize real-time local data (e.g., user queue length, current CSI) to adjust beam selection in the pruned beam space dynamically. Specifically, short-timescale beam selection is modeled as a partially observable Markov decision process (POMDP), and a deep Q-network (DQN) is employed to maximize delay satisfaction rates. This hierarchical structure, combining CU-based action space pruning and DU-based real-time optimization, enhances latency performance and beamforming accuracy. To address the issues of high signaling exchange overhead and system delay in the centralized scheme for cell-free MIMO systems, a Q-decomposition multi-agent independent extension (QMIX)-based partially distributed beam selection scheme (QMIX-PDBS) is further proposed. In the training phase, the CU aggregates all DUs' local observations to construct a global state. Compared to the completely independent training of each DU, the QMIX-PDBS facilitates the training of both the global network in the CU and the local networks in the DUs, yielding more favorable training outcomes. The execution of QMIX-PDBS is fully distributed to guarantee real-time local decisions at the DUs.

Simulations are conducted to verify the effectiveness of the proposed scheme, with detailed parameters

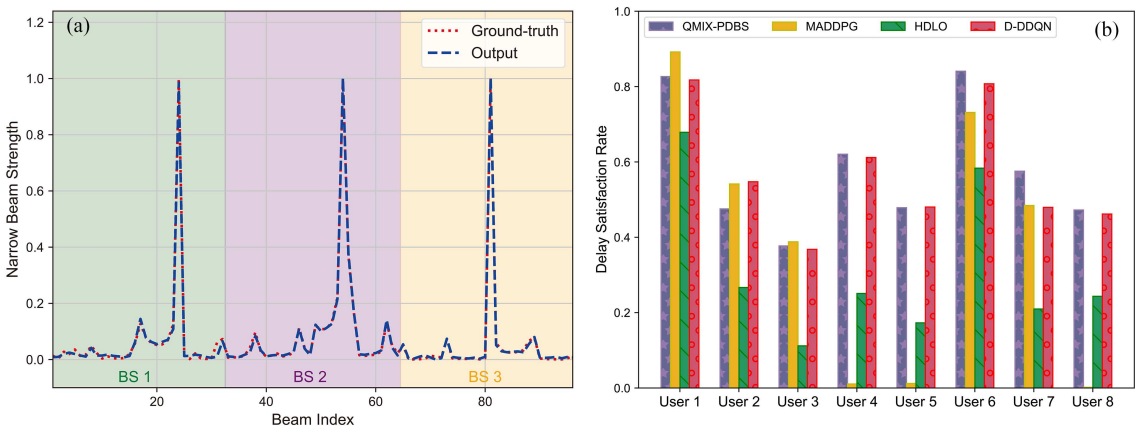

Figure 8 Performance of the proposed hierarchical, distributed, and intelligence-empowered real-time beam selection approach. (a) Wide beam-based narrow beam prediction; (b) delay satisfaction rate versus different users.

provided in [22]. Figure 8(a) compares the predicted narrow beam power profile from the CNN model with the actual narrow beam strength. The results indicate that the CNN can precisely learn the relationship between the wide beam strengths and the narrow beam strengths of multiple DUs well, and it can accurately predict multiple beams with relatively high strength. Figure 8(b) compares the delay satisfaction rate of the proposed QMIX-PDBS scheme to those of three benchmarks, i.e., multi-agent deep deterministic policy gradient (MADDPG), hierarchical distributed Lyapunov optimization (HDLO), fully distributed double DQN (D-DDQN), versus the number of users. The QMIX-PDBS demonstrates superior performance in balancing system QoS, user delay, and minimizing beam training overhead compared to other approaches. Unlike MADDPG, which improves overall system QoS by sacrificing certain users' quality, QMIX-PDBS ensures a more equitable distribution of QoS across all users. In contrast to HDLO, QMIX-PDBS not only achieves a lower average delay per user but also significantly improves the delay satisfaction rate due to its optimized beam training process. Furthermore, the integration of interaction between the DU and CU in QMIX-PDBS outperforms D-DDQN by enhancing coordination across DUs, resulting in a higher delay satisfaction rate.

4.3 Data-driven wireless network DT for pre-validation

A wireless network DT is a virtual digital representation of a physical wireless network. Its primary role is to accurately replicate the physical network, facilitating real-time monitoring, analysis, and optimization. Currently, there are two typical paradigms for wireless network DTs. One relies on models and expert knowledge, using static or historical measurement data to fit model parameters. This method faces significant challenges when solving complex problems due to insufficient model accuracy, and the parameters, derived primarily from historical data, are inadequate for adapting to the dynamically changing states of wireless systems. The other paradigm is based on traditional deep neural networks (DNN) with supervised learning. This approach suffers from inefficient data collection and utilization, as well as a heavy dependence on labeled data. In many complex scenarios, obtaining labels is challenging, and capturing latent complex feature dataset distributions or the stochastic nature of system state transitions proves to be problematic [23].

To address the aforementioned issues, we propose a paradigm shift towards a data-driven approach for network DTs, strategically incorporating KGs, feature datasets, and advanced ML techniques such as generative AI (GAI), Transformers, and long short-term memory (LSTM) models. As illustrated in Figure 9, this architecture leverages both historical and real-time data to dynamically adapt to changes in the network environment, providing a more accurate and up-to-date representation of the physical network. A key aspect of this architecture is the construction and utilization of feature datasets based on the wireless data KG technology. The wireless data KG not only captures complex interdependencies between different network entities, such as user behavior, resource allocation, and traffic patterns, but also enhances the selection of critical features. Furthermore, GAI is integrated to empower the network DTs with the capability to generate synthetic data, thereby augmenting the feature datasets. This augmentation reduces the reliance on labeled data, which is often limited, and improves the model's ability to capture the stochastic nature of system state transitions, leading to more robust training. By

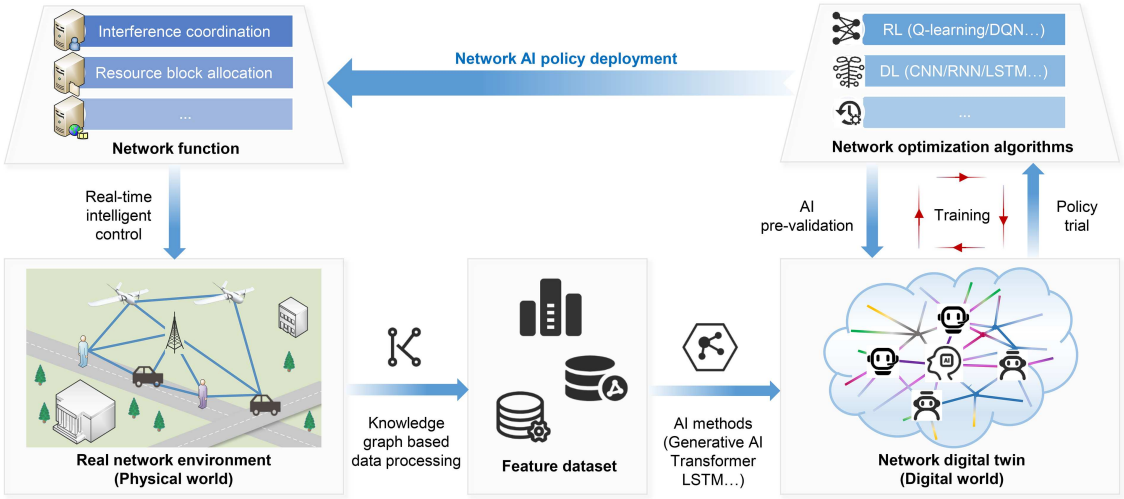

Figure 9 6G network digital twin architecture.

continuously enriching the feature dataset with both real and synthetic data, the DT-based models can better adapt to diverse and evolving network conditions. Advanced ML techniques, e.g., GAI, enable the creation of network DTs customized for various network entities, such as RAN, wireless environments, and network policies, based on specific optimization tasks [23, 24]. Based on these DTs, AI models for network optimization are trained and pre-verified. Upon achieving adequate learning within the DT environment, these models are deployed in the real network for real-time intelligent control. This iterative closed-loop process allows for the continuous refinement of the optimization policy. Network AI models that integrate prior knowledge through DTs exhibit greater stability and faster convergence compared to models initialized randomly, thereby ensuring a safer deployment in real networks. Based on the aforementioned architecture, we propose a GAI-based DT framework for wireless resource management, and further elaborate on a specific design for an intelligent RAN slicing use case.

For the wireless resource management problem, to address challenges such as the classical deep reinforcement learning (DRL) algorithm being unsafe or having performance lower than the default strategy in the initial stage, failing to capture latent complex environment distribution, as well as slow algorithm convergence speed, we propose a GAI-based DT network framework. As illustrated in Figure 10, the DT trains GAI-based state prediction, reward prediction, and policy models to model a virtual interaction environment. It essentially clones the historical wireless resource allocation behaviors and hopefully generates a rich set of high-quality samples. After the pre-training of the DT network, it provides pre-validation services at either the policy or decision level. At the policy level, the real network policy can be pre-verified against interaction trajectories predicted by the DT network, with subsequent alignment to the pre-verified policy. At the decision level, the DT network can predict the performance of policy decisions and facilitate appropriate adjustments to ensure decision safety. Through pre-training via behavioral cloning, virtual interaction environments, pre-validation at the policy or decision level, and the powerful generative capabilities of GAI, the DT network is expected to provide safe intelligent wireless resource management for 6G.

Based on the paradigm of GAI and DT-enhanced intelligent wireless resource management, a data-driven intelligent time-frequency resource management technique is proposed for the physical resource blocks (PRBs) allocation problem in the RAN slicing scenario. The optimization objective is to allocate the PRBs as efficiently as possible while fulfilling the long-term service level agreements (SLAs) for each network slice. The proposed DT network (DTN) consists of a virtual interaction environment, expert analysis for correcting resource allocation strategy, and two types of long-term predictive models, i.e., a behavior cloning model (BCM) based on the conditional generative diffusion model (CGDM) [25] and a KPI prediction model (KPIM). Additionally, there is a synergistic mechanism between the DTN and a DRL agent for RAN slicing, aimed at mimicking the behavior of the default policy and providing pre-validation of policies and decisions before implementing the PRB allocation strategy.

Specifically, the virtual interaction environment is modeled as a Markov decision process (MDP) using CGDM. The DRL algorithm is designed based on a constrained multi-agent MDP, where each agent

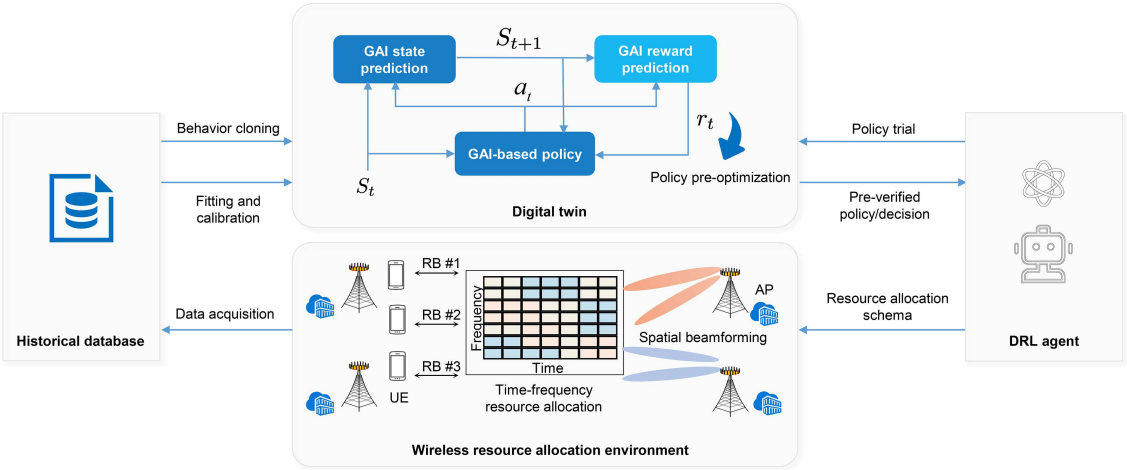

Figure 10 Illustration of GAI and DT enhanced intelligent wireless resource management.

employs a CGDM-based actor to guide its decision-making process. This model-based agent utilizes the conditional aspects of the environment to generate optimal actions within the learning framework. Regarding the synergistic mechanisms between the DTN and the DRL agent, the DTN data warehouse continuously collects system behavior trajectories under the default policy, with each PRB allocation result being refined through expert analysis. Previous trajectory records from the data warehouse are then used to iteratively train the virtual environment, BCM, and KPIM. Once these models are trained and converge, the parameters of the BCM are synchronized with the DRL actor to warm up the DRL agent. Subsequently, the DRL agent is deployed in the actual network environment. The PRB allocation decision generated by the DRL agent requires pre-validation via KPIM before configuration along with a decision pre-adjustment when it cannot fulfill the specific SLA, namely the decision-level pre-validation. Additionally, to update the DRL agent's policy, it is essential to perform policy-level pre-validation using the DTN virtual interaction environment. This process involves interacting with the virtual environment and pre-updating the agent's policy accordingly.

We validate the capability of the proposed PML-AI framework to provide controllable QoS using a DT design tailored for the intelligent RAN slicing use case. The simulations focus on downlink transmission involving three enhanced mobile broadband (eMBB) and two massive machine type communication (mMTC) RAN slices, utilizing 150 PRBs per subframe. Detailed simulation parameters, including the channel model, slice types, SLA requirements, and traffic patterns, can be referenced from [26]. We compare the DRL and DRL-DT against KBRL with 50000 decision stages, where DRL-DT is implemented after DRL is empowered by the DT network, DRL is designed on the basis of the algorithm proposed in [27], and KBRL is a model-based RL approach that exhibits relatively high performance [26]. As shown in Figure 11, DRL, DRL-DT, and KBRL achieve $(0.042, 0.027, 0.114)$ SLA violations and $(105.53, 101.46, 111.79)$ total allocated PRBs in average, respectively. It is not difficult to discern that KBRL experiences non-negligible fluctuations in SLA violations, and DRL exhibits a relatively high rate of SLA violations during the initial training phase. Consequently, DRL-DT superiors in SLA violations, cumulative SLA violations, and PRB conservation compared to other algorithms. Specifically, DRL-DT is safer in the whole training stage and achieves a higher convergence speed than DRL and KBRL.

5 State of the art: 6G prototype and experimental results

The development of comprehensive 6G prototypes is crucial for advancing network evaluation and technology validation. Notable efforts include the European Union's 6G-BRICKS (building reusable testbed infrastructure for validating cloud-to-device breakthrough technologies) project, which focused on creating a pan-European research infrastructure to explore advanced technologies such as AI, cell-free massive MIMO, and reconfigurable intelligence surface (RIS). In China, a 6G prototype platform was developed by China Mobile Research Institute and Beijing University of Posts and Telecommunications, which supports multiple frequency bands and cloud-based technologies [28]. Huawei's 6G research team also posted a prototype for short-range communications at 70 GHz, designed for ultra-low power consumption, ultra-

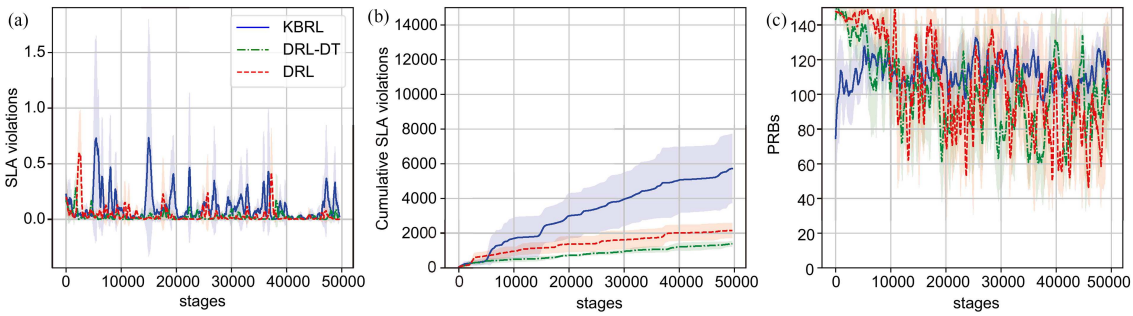

Figure 11 Performance comparison between DRL, DRL-DT, and KBRL. (a) SLA violations; (b) cumulative SLA violations; (c) resource allocation.

high throughput, and ultra-low latency [29]. Additionally, the Zhongguancun Institute of Ubiquitous-X Innovation and Applications, in collaboration with China Mobile and other key partners, designed a cloudified ultra-massive MIMO prototype system, which facilitates the validation of cloudified wireless networks and ultra-massive MIMO [30].

To validate the core advantages of 6G over existing mobile networks, Purple Mountain Laboratories recently developed an intelligent, service-oriented, and open programmable 6G prototype platform [2]. This platform is built on a task-centric, three-layer 6G architecture and supports the validation of key capabilities across six primary 6G use cases, including peak data rates at the Tbps level, spectral efficiency in the Kbps/Hz range, and latency at the microsecond level. It also facilitates key technology verifications such as cell-free access, native intelligence, and THz communication. Furthermore, the platform achieves cloud-network convergence by integrating computation, communication, data, and AI, while enabling service customization through dynamic orchestration and open capabilities. The platform contains a data plane, an intelligence plane, and a converged open service layer, forming a 6G super edge node (SEN). It collects data such as radio resource control (RRC) signaling, slot-level uplink/downlink schedules, measurement reports, and detailed control channel data. Over 400 air interface data distributed across L1, L2, and L3 are open and programmable, including the control parameters of the medium access control (MAC) scheduler, whose real-time performance of intelligent self-optimization attains the millisecond level. The 6G edge network implemented on this platform leverages technologies like baseband unit (BBU) resource pooling, programmability, cloudification, virtualization, and microservices to enable rapid service deployment, flexible scaling, and efficient optimization. Through standardized interfaces, hardware white-boxing, open-source software, and end-to-end programmability, it achieves dynamic resource allocation, flexible capacity expansion, on-demand service customization, and rapid orchestration, creating a flexible, reconfigurable, open programmable network.

Building upon this foundational platform, we have developed a real-time intelligent prototype verification system to evaluate the performance of the proposed PML-AI intelligent framework in addressing green, real-time, and controllable requirements, as shown in Figure 12. This system consists of three layers: CoNs, edge nodes, and end nodes. The CoNs comprise the core network, a big data platform, and the service management and orchestration (SMO). The big data platform focuses on non-real-time data collection and analysis to construct a general KG, which is further refined into specialized KGs for specific scenarios and KPIs to facilitate real-time intelligence at edge nodes. And just as the name means, the SMO is for the service management and orchestration of the whole network. The edge nodes are mainly composed of virtualized CUs (vCUs), virtualized DUs (vDUs), a near-real-time intelligent control unit, the user plane function (UPF), and a wireless data platform. The vDUs are embedded with endogenous real-time intelligent control units, while the near-real-time intelligent control unit is usually implemented in a dedicated edge server. The UPF, which is the user plane function module of the core network, is sunk to the RAN side. The wireless data platform is designed for data collection and management. The intelligent control functions and DTs are implemented in the SMO, the near-real-time intelligent control unit, and the endogenous real-time intelligent control units, constituting the intelligence plane and the DT plane across the CoNs and edge nodes, and operating at different time scales. Each edge node possesses over 10 distributed DUs and provides access and communication services for more than 100 users. The end nodes consist of edge DUs (eDUs) and user equipments (UEs), with eDUs primarily hosting the low-level physical layer (PHY) functions, e.g., modulation and demodulation. This system supports experimental verification and performance evaluation of innovative use cases, such as network

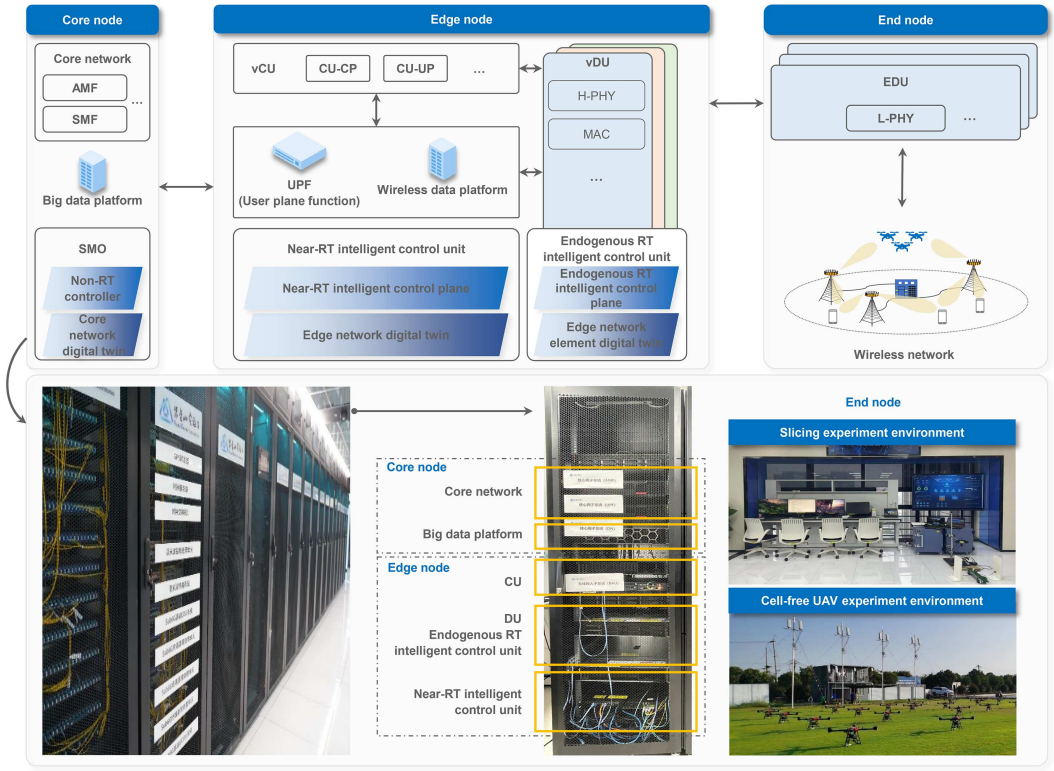

Figure 12 6G real-time intelligent prototype verification system.

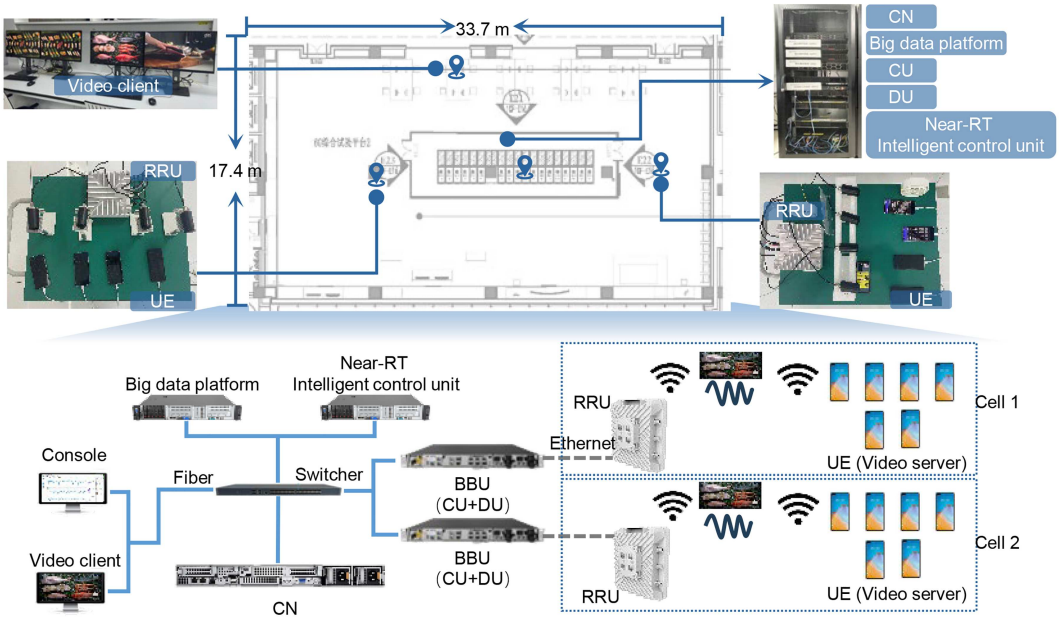

Figure 13 4K video uploading experiment system.

slicing and cell-free unmanned aerial vehicle (UAV) communication.

Based on the above real-time intelligent prototype verification system, we conduct a 4K video uploading experiment, as shown in Figure 13. The edge nodes consist of one CU, one near-real-time intelligent control unit, and two DUs, with one endogenous real-time intelligent control unit being implemented in each DU. The near-real-time intelligent control unit is deployed in a dedicated server at the network edge. The end nodes comprise remote radio units (RRUs) and commercial UEs. Totally two RRUs and twelve commercial UEs are deployed, constituting a canonical two-cell and twelve-user interference scenario.

Figure 14 Experimental results of the uplink 4K video transmission. (a) Screen shots of results; (b) statistical results.

In this scenario, the UEs function as video servers, facilitating the uplink transmission of 4K videos to video clients through communication with the RRUs. To ensure the quality of video transmission for each user, an uplink throughput exceeding 40 Mbps is essential. To this end, the near-real-time intelligent control unit and endogenous real-time intelligent control units collaborate to conduct slot-level on-demand resource allocation in a hierarchical way. In brief, the near-real-time intelligent control unit performs coarse-grained resource pre-allocation for users in each cell based on reference signal received power (RSRP) reported by users, allocating resources to users on demand meanwhile coordinating the interference among users. While the endogenous real-time intelligent control units or DUs predict the channel quality of each resource block (RB) or resource block group (RBG) in real time, to further allocate resources to users in a fine-grained way.

The experimental results, illustrated in Figure 14, reveal that prior to the activation of the real-time AI-powered resource allocation algorithm, both Cell 1 and Cell 2 could only support four users each for 4K video uploads. Notably, by deploying the real-time AI algorithm, both Cell 1 and Cell 2 are capable of supporting all six users for 4K video uploads and thus achieve a 50% increase in the number of supported users. Additionally, the uplink throughputs of Cell 1 and Cell 2 increase from 251.12 and 233.58 Mbps to 275.17 and 287.85 Mbps, respectively, i.e., 9.5% and 23.2% performance improvement. Overall, the total throughput of the two cells increases by 16%. Apparently, the proposed algorithm can remarkably enhance both the overall system uplink throughput and the individual uplink throughput requirements of UEs via the efficient allocation of resources based on user QoS requirements, facilitated by the wireless data KG and its enabled real-time AI capability. It is also worth noting that with the help of KG, a feature dataset composed of 5 data fields is extracted from a total of 44 collected data fields, with a reduction of nearly 90%. The data volume, model size, and model training cost are therefore reduced by a factor of 10.

6 Conclusion

This paper investigated the critical technologies for next generation 6G networks, with a focus on integrating AI into sustainable 6G networks. After providing a brief review of the 6G vision and emerging trends, we identified three key challenges for the integration of AI and sustainable 6G-green, real-time, and con-

trollable requirements. In response, we proposed a novel PML-AI framework built upon a task-centric three-layer architecture. This framework incorporates wireless data KG, lightweight AI driven by feature datasets, and DTs of wireless networks within a dual-cycle architecture. By simplifying data for real-time AI processing, reducing computational overhead, and enhancing network reliability, the proposed AI framework can resolve the core challenges. To validate the proposed PML-AI framework and the corresponding 6G architecture, we developed a real-time intelligent prototype verification system on the 6G integrated test platform and conducted QoS-aware real-time intelligent resource allocation experiments. Experimental results demonstrated time-slot-level intelligent control and a nearly tenfold reduction in data and computational overhead, underscoring the framework's potential to drive 6G advancements while aligning with sustainability objectives.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 62225107), Natural Science Foundation on Frontier Leading Technology Basic Research Project of Jiangsu (Grant No. BK20222001), and Fundamental Research Funds for the Central Universities (Grant No. 2242022k60002).

References

1 International Telecommunication Union. IMT vision-framework and overall objectives of the future development of IMT for 2030 and beyond. Recommendation ITU-R M.2160-0, 2023
2 You X, Huang Y, Liu S, et al. Toward 6G TKμ extreme connectivity: architecture, key technologies and experiments. IEEE Wireless Commun, 2023, 30: 86–95
3 You X H, Wang C-X, Huang J, et al. Towards 6G wireless communication networks: vision, enabling technologies, and new paradigm shifts. Sci China Inf Sci, 2021, 64: 110301
4 Letaief K B, Chen W, Shi Y, et al. The roadmap to 6G: AI empowered wireless networks. IEEE Commun Mag, 2019, 57: 84–90
5 Zhao W X, Zhou K, Li J, et al. A survey of large language models. 2023. ArXiv:2303.18223
6 Schwartz R, Dodge J, Smith N A, et al. Green AI. Commun ACM, 2020, 63: 54–63
7 Xu J, Zhou W, Fu Z, et al. A survey on green deep learning. 2021. ArXiv:2111.05193
8 Yigitcanlar T, Mehmood R, Corchado J M. Green artificial intelligence: towards an efficient, sustainable and equitable technology for smart cities and futures. Sustainability, 2021, 13: 8952
9 Zhou Z, Chen X, Li E, et al. Edge intelligence: paving the last mile of artificial intelligence with edge computing. Proc IEEE, 2019, 107: 1738–1762
10 Shen X, Gao J, Wu W, et al. Holistic network virtualization and pervasive network intelligence for 6G. IEEE Commun Surv Tut, 2022, 24: 1–30
11 Miller T. Explanation in artificial intelligence: insights from the social sciences. Artif Intell, 2019, 267: 1–38
12 Lipton Z C. The mythos of model interpretability. Commun ACM, 2018, 61: 36–43
13 Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell, 2019, 1: 206–215
14 Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proceedings of IEEE Symposium on Security and Privacy (SP), 2017. 38: 39–57
15 Wang D, Zhang C, Du Y, et al. Implementation of a cloud-based cell-free distributed massive MIMO system. IEEE Commun Mag, 2020, 58: 61–67
16 Wang D M, You X H, Huang Y M, et al. Full-spectrum cell-free RAN for 6G systems: system design and experimental results. Sci China Inf Sci, 2023, 66: 130305
17 Zhang J, Zhu M, Hua B, et al. Real-time demonstration of 100 GbE THz-wireless and fiber seamless integration networks. J Lightwave Technol, 2023, 41: 1129–1138
18 You X, Zhang C, Sheng B, et al. Spatiotemporal 2-D channel coding for very low latency reliable MIMO transmission. In: Proceedings of IEEE GLOBECOM Workshops, 2022. 473–479
19 Epoch AI. Large-scale AI models. 2024. https://epochai.org/data/large-scale-ai-models
20 Zhang P, Xiao Y, Li Y, et al. Toward net-zero carbon emissions in network AI for 6G and beyond. IEEE Commun Mag, 2024, 62: 58–64
21 Huang Y, You X, Zhan H, et al. Learning wireless data knowledge graph for green intelligent communications: methodology and experiments. IEEE Trans Mobile Comput, 2024, 23: 12298–12312
22 Wang C, Zhang C, Meng F, et al. Traffic-aware hierarchical beam selection for cell-free massive MIMO. IEEE Trans Commun, 2024, 72: 6490–6504
23 Tao Z, Xu W, Huang Y, et al. Wireless network digital twin for 6G: generative AI as a key enabler. IEEE Wireless Commun, 2024, 31: 24–31
24 Zhang Z, Huang Y, Zhang C, et al. Digital twin-enhanced deep reinforcement learning for resource management in networks slicing. IEEE Trans Commun, 2024, 72: 6209–6224
25 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 6840–6851
26 Alcaraz J J, Losilla F, Zanella A, et al. Model-based reinforcement learning with kernels for resource allocation in RAN slices. IEEE Trans Wireless Commun, 2023, 22: 486–501
27 Huang Y, Xu M, Zhang X, et al. AI-generated network design: a diffusion model-based learning approach. IEEE Netw, 2024, 38: 202–209
28 Beijing University of Posts and Telecommunications. New 6G achievements | Impressive results from the joint innovation center of Beijing University of Posts and Telecommunications and China Mobile Research Institute! 2022. https://news.bupt.edu.cn/info/1012/27648.htm
29 Huawei. Ultra-low power and high-data rate short-range wireless enables fully immersive 6G. 2022. https://www.huawei.com/en/technology-insights/future-technologies/6g-short-range-communications
30 Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park. Release of the 6G cloud-based massive MIMO prototyping and verification system. 2024. https://kw.beijing.gov.cn/art/2024/4/30/art_10680_675554.html

Profile of Xiaohu YOU

Xiaohu YOU is a member of the Chinese Academy of Sciences and the chief professor at Southeast University. He is also the director and chief scientist of Purple Mountain Laboratories, the deputy director of Pengcheng Laboratory, and the director of the National Mobile Communications Research Laboratory. As a leading scientist in mobile communications, he has been the chief expert for the National 863 Programs of 3G, 4G, and 5G research in China during the past 30 years, and currently he serves as the chief expert for the National Key R&D Program on Broadband Communication and New Networks as well as the chief expert for the Ministry of Science and Technology's Program on 6G. He has led the development strategy research on mobile communication, drafted its holistic technical framework, promoted the advance of broadband mobile communication technology in China, and made remarkable contributions to telecommunication technology research and development in China.

Prof. You has achieved a number of crucial accomplishments in communication theory and technologies, and holds over 100 domestic and international invention patents. He has made original innovation contributions to the capacity-approaching theory on broadband wireless transmission and its wide-scope engineering applications, which won the First Class National Technological Invention Award. He has also made pioneering contributions to distributed MIMO and cell-free wireless transmission and received the IET Achievement Medal and the Tan Kah Kee Science Award. He also achieved critical breakthroughs in high-frequency wireless transmission with large-scale integrated phased arrays and promoted their widespread applications in industry, which won the Second Class National Technological Invention Award. His research on multipath channel modeling and corresponding transmission methods has been widely validated and applied in industry, which won the Second Class National Science and Technology Progress Award.

He has published over 500 academic papers in top international journals with around 30000 citations. His research papers on 5G and 6G wireless transmission and system architecture, published in *SCIENCE CHINA Information Sciences*, are among the most highly cited academic papers globally. He authored the world's first academic monograph on distributed MIMO and cell-free mobile communications. He is the recipient of the IEEE Communications Society Fred W. Ellersick Best Paper Award, as well as three Best Paper Awards at top international conferences, such as IEEE GLOBECOM and IEEE WCNC. He also received three Outstanding Paper Awards in the fields of electronics, information, and communication in China.

In recognition of his outstanding contributions to mobile communications in China, he was awarded the National May 1 Labor Medal and the title of the National Outstanding Science and Technology Worker. Renowned in the international academic community, he was elected an IEEE Fellow in 2011 and has served as the general chair of prominent international conferences such as IEEE WCNC 2013 and IEEE ICC 2019. He has always remained dedicated to the forefront of scientific research and teaching, and cultivated a large number of scholars and experts in both academic and industrial societies.

Selected publications

• You X H, Wang C-X, Huang J, et al. Towards 6G wireless communication networks: vision, enabling technologies, and new paradigm shifts. Sci China Inf Sci, 2021, 64: 110301

• You X H, Wang D M, Wang J Z. Distributed MIMO and Cell-Free Mobile Communication. Singapore: Springer, 2020

• You X H, Zhang C, Tan X S, et al. AI for 5G: research directions and paradigms. Sci China Inf Sci, 2019, 62: 021301

• Wu S B, Wang C-X, Alwakeel M, et al. A general 3-D non-stationary 5G wireless channel model. IEEE Trans Commun, 2018, 66: 3065–3078

• You X H, Pan Z W, Gao X Q, et al. The 5G mobile communication: the development trends and its emerging key techniques (in Chinese). Sci Sin Inform, 2014, 44: 551–563

• Wang C-X, Haider F, Gao X Q, et al. Cellular architecture and key technologies for 5G wireless communication networks. IEEE Commun Mag, 2014, 52: 122–130

• Wang D M, Wang J Z, You X H. Spectral efficiency of distributed MIMO systems. IEEE J Sel Areas Commun, 2013, 31: 2112–2127

• You X H, Wang D M, Zhu P C, et al. Cell edge performance of cellular mobile systems. IEEE J Sel Areas Commun, 2011, 29: 1139–1150

• You X H, Wang D M, Sheng B, et al. Cooperative distributed antenna systems for mobile communications. IEEE Wireless Commun, 2010, 17: 35–43

• Yu (You) X H, Chen G-A, Cheng S-X. Dynamic learning rate optimization of the backpropagation algorithm. IEEE Trans Neural Networks, 1995, 6: 669–677

• Wang D M, You X H, Huang Y M, et al. Full-spectrum cell-free RAN for 6G systems: system design and experimental results. Sci China Inf Sci, 2023, 66: 130305

• Wang D M, Wang M H, Zhu P C, et al. Performance of network-assisted full-duplex for cell-free massive MIMO. IEEE Trans Commun, 2020, 68: 1464–1478

• Zhang J J, Zhao D X, You X H. A 20-GHz 1.9-mW LNA using gm-boost and current-reuse techniques in 65-nm CMOS for satellite communications. IEEE J Solid-State Circ, 2020, 55: 2714–2723

• You X H, Zhang C, Sheng B, et al. Spatiotemporal 2-D channel coding for very low latency reliable MIMO transmission. In: Proceedings of IEEE Globecom Workshops (GC Wkshps), 2022. 473–479

• Huang Y M, You X H, Zhan H, et al. Learning wireless data knowledge graph for green intelligent communications: methodology and experiments. IEEE Trans Mobile Comput, 2024, 23: 12298–12312

• You X H, Huang Y M, Liu S H, et al. Toward 6G TKμ extreme connectivity: architecture, key technologies and experiments. IEEE Wireless Commun, 2023, 30: 86–95

• Tao Z Y, Xu W, Huang Y M, et al. Wireless network digital twin for 6G: generative AI as a key enabler. IEEE Wireless Commun, 2024, 31: 24–31

• Zhou W Y, Shen Y F, Li L P, et al. Belief-selective propagation detection for MIMO systems. IEEE Trans Commun, 2023, 71: 7244–7257

• Zhang Y, Zhou W Y, Zhang Y W, et al. BayesBB: a 9.6Gbps 1.61ms configurable all-message-passing baseband-accelerator for B5G/6G cell-free massive-MIMO in 40nm CMOS. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2024. 48–50

• Zhang Z M, Huang Y M, Zhang C, et al. Digital twin-enhanced deep reinforcement learning for resource management in networks slicing. IEEE Trans Commun, 2024, 72: 6209–6224