# K12 AI-Driven Research Network: A New Paradigm Coupling Data Production and Talent Innovation

<sup>1</sup> Department of Mathematics, Hong Kong Baptist University, Hong Kong

2026-06-11

ChinaRxiv

## Abstract

Against the backdrop of the rapid development of generative artificial intelligence, the depletion of high-quality non-standard data corpora and the discontinuity in top innovative talent have become dual hidden risks constraining the development of China' s new quality productive forces in the intelligent era. This paper proposes a reconstruction of the foundational education-research framework: building a six-step progressive, distributed K12 research network of "AI scientific simulator → real-environment DNA sampling → AI-controlled desktop automated workstation → cloud-based AI laboratory invocation → organoid/traditional Chinese medicine digitization → feature mining of the complex system of Chinese civilization." Its core lies in lowering the threshold for participation through an "AI mentorship system" : AI large models undertake script generation and equipment control, while students issue tasks and verify results through natural-language interaction, obtain non-transferable academic credit credentials by contributing real data, advance along the "innovation funnel," and realize "the capable come first." Taking environmental DNA watershed monitoring as a starting point with zero ethical risk, this paper demonstrates feasibility and compliant governance, and clarifies its structural connectivity with the digitization of traditional Chinese medicine and even life-system networks through tensor isomorphism analysis; the concept of "treating disease before it arises" is extended to the capture of critical slowing down in ecological early warning. It further proposes a "same-frequency convergence" strategy: dividing complex scientific problems into parallelizable micro-topics, allowing tens of millions of nodes to divide and conquer them, and converging the results into national-level data assets and a foundation for innovative talent. Finally, it presents policy pathways for sandbox pilots, interministerial coordination, and legislation on minors' research data.

## Full Text

# K12 AI-Driven Research Network: A New Paradigm Coupling Data Production and Talent Innovation

Dehao Li*1,2,3, Kaijie Liu2, Jianrong Lin7,3, Menghuan Song4, Hao Hu4, Jianzong He†5, Ping Shen6, Guobao Wu‡1

1 Department of Mathematics, Hong Kong Baptist University, Hong Kong
2 Institute for ESG and New Quality Productivity, Guangdong University of Technology, Guangzhou, Guangdong–Hong Kong–Macao Greater Bay Area
3 Guangdong Institute of Digital Industry, Guangzhou
4 Institute of Chinese Medical Sciences, University of Macau; State Key Laboratory of Quality Research in Chinese Medicine, Macau
5 The Education University of Hong Kong, Hong Kong

6 Southern University of Science and Technology, Shenzhen
7 School of Public Policy and Management, Tsinghua University, Beijing

June 2026

## Abstract

Against the backdrop of the rapid development of generative artificial intelligence, the depletion of high-quality data corpora and the discontinuity in the cultivation of top innovative talent have become two major hidden risks constraining the development of China' s "new quality productive forces" in the intelligent era. This paper proposes a reconstruction of the underlying educational framework: a six-step progressive pathway beginning with an "AI scientific simulator," proceeding through "real-environment DNA sampling," an "AI-controlled desktop automated workstation," and "cloud-based AI laboratory invocation," and extending to "organoids/digitalization of traditional Chinese medicine" and "feature mining of complex systems in Chinese civilization." The core of this framework is AI-driven research automation—large AI models serve as "programmers," automatically decomposing research tasks into executable scripts and controlling automated equipment; students issue tasks and verify results through natural-language interaction, cultivating scientific thinking under an "AI apprenticeship system." By contributing real data and participating in distributed research networks, students obtain non-transferable academic credit credentials, thereby unlocking more advanced tasks and realizing the principle that "the capable go first." Taking environmental DNA watershed monitoring as a zero-ethical-risk starting case, this paper demonstrates the feasibility and compliance of the framework; through an analysis of mathematical isomorphism, it points out that the spatiotemporal sparse tensors of eDNA, the hypergraphs of traditional Chinese medicine compatibility, and even the underlying structures of human life-system networks possess an intrinsic unity. The concept of "treating disease before it arises" is extended to the field of ecological monitoring—capturing critical slowing-down signals before irreversible phase transitions occur in a system. This paper proposes the strategy of "frequency-aligned convergence (building a tower by gathering grains of sand)" : complex scientific problems are divided into micro-topics that can be explored in parallel, through unified task specifications, metadata standards, audit chains, and contribution credentials.

* Co-corresponding author; email: alexlihk@qq.com
† Jianzong He: Foreign Academician of the European Academy of Natural Sciences.
‡ Co-corresponding author.

an institutional–technological cavity, enabling the low signal-to-noise contributions of tens of millions of K12 nodes to achieve phase alignment and thereby transform from statistical noise into coherent signals. This strategy takes tensor decomposition and spectral manifold mapping as its mathematical foundation,

allowing heterogeneous and biased distributed observations to be phase-aligned and coherently superposed. After crossing the threshold, they are released directionally, forming an information beam usable for national-scale ecological monitoring and the screening of innovative talent—not simple aggregation, but concentrating light into a blade. Finally, this paper proposes launching synchronized pilots in the Guangdong-Hong Kong-Macao Greater Bay Area (Hetao-Qianhai) and Xiong' an New Area, and, through an interministerial joint-conference mechanism, promoting the inclusion of academic credit credentials in the "Strong Foundation Plan." This paper argues that the complex systemic characteristics in five thousand years of Huaxia civilization that "seem impossible to quantify" (such as holistic perspectives and dynamic balance) are precisely the richest blue ocean for the next generation of data science and artificial intelligence; and that a K12 AI-driven research network is the strategic infrastructure for cultivating data collectors, algorithm validators, and civilizational verifiers for this blue ocean.

**Keywords:** AI-driven research; K12 education; environmental DNA; co-frequency convergence; AI mentorship; treating disease before it arises; distributed digital identity; datafication of Huaxia civilization

# 1 Introduction: The Dual Predicament and Coupling Logic of New Quality Productive Forces in the Intelligent Era

Amid the global technological upheaval triggered by generative artificial intelligence, the construction of computing infrastructure is advancing at an unprecedented pace. However, with the homogenization of algorithmic models, the core barrier determining the level of "new quality productive forces" in the intelligent era is shifting from "computing scale" toward "high-quality data elements" and "top innovative talent." Examining China' s current macro strategy and educational ecosystem, these two underlying foundations are facing hidden dangers that cannot be ignored.

First, there is the depletion crisis of high-order and non-standard data corpora. At present, global training of large models is rapidly consuming the internet' s existing high-quality public text. Studies show that reserves of high-quality language data are expected to face depletion within the next few years [1]. Even more severe is that the "multimodal non-standard data" that can truly empower the real economy and possess local Chinese characteristics—from regional microclimates to functional-state drift in population health, from spatiotemporal dynamics of biodiversity to the formalization of traditional knowledge systems—are extremely scarce. Such data cannot be obtained through web crawlers; they must rely on sustained collection by large-scale, distributed physical nodes in the real world.

Second, "performing incremental work within known paradigms" has led to a fault line in top innovative talent. Although China has a K12 student population of nearly 200 million, the traditional education model still carries strong industrial-era characteristics: it regards students as "containers" of knowledge and places them in homogeneous competition within standardized evaluation systems. This model not only suppresses adolescents' curiosity, but also deprives them of opportunities to cultivate critical thinking through authentic scientific trial and error [2].

The strategic-window judgment is that global competition over AI data is at a critical turning point: high-quality, multimodal, non-standardized data will be the core barrier determining national AI competitiveness over the next 5–10 years. The proposal of this paper is precisely a proactive offensive plan to "seize the commanding heights of data elements."

The central proposition of this paper is: the above two predicaments are not independent of each other, but a mutually causal "double helix." High-quality non-standard data can only be produced through genuine scientific exploration; and true innovative capacity can only be cultivated in the process of producing such data. This means that the production of data elements and the cultivation of top innovative talent are not two independent problems that can be solved separately—they must be solved simultaneously within the same system.

**Data–Talent Double-Helix Coupling Mechanism and "Synchronous Convergence" Strategy**

- **Upward helix: data-production cycle**
  - High-quality non-standard data
  - Train next-generation AI models
  - Enhance AI-assisted research capability (AI mentorship model)
  - Empower more students to participate in authentic research
  - Produce more high-quality non-standard data (synchronous convergence)
- **National distributed knowledge-production network (research by all)**
- **Downward helix: talent-cultivation cycle**
  - Students participate in scientific research
  - Experience the "hypothesis–failure–revision" cycle
  - Cultivate critical thinking and innovation ability
  - Grow into top innovative talent
  - Design higher-quality research tasks (problem decomposition)

Figure 1: Data–Talent Double-Helix Coupling Mechanism and "Synchronous Convergence" Strategy

This judgment leads us to a fundamental strategic question: Is there a systematic solution that can reconstruct China's vast K12 education system from a passive conduit for knowledge transmission into a "dual-base" infrastructure for the synchronized production of data elements and innovative talent? The answer of this paper is: to build a six-step, progressively advancing distributed knowledge-production network—"AI scientific simulator → real-environment DNA sampling → AI-controlled desktop automated workstation → cloud-based AI laboratory invocation → organoids / digitalization of traditional Chinese medicine → feature mining of the complex system of Chinese civilization" —thereby dividing complex scientific problems into miniature topics that can be explored in parallel and addressed by distributed nodes through divide-and-conquer. Ultimately, this realizes "synchronous convergence," a leap from quantitative accumulation to qualitative transformation—namely, the phase alignment of tens of millions of weak contributions within an institutional-technological cavity, their directional amplification after crossing a threshold, and the formation of an information beam usable for decision-making at the national scale.[1]

## 2 Theoretical Transition: From "Recipients of Instruction" to "AI-Collaborative Researchers"

### 2.1 Citizen Science, AI Mentorship, and the New Paradigm of "Research by All"

"Citizen science" refers to a collaborative paradigm in which the public participates in the process of scientific research. Over the past decade or more, international citizen-science projects have demonstrated the enormous value of non-professional groups in processing massive amounts of data [3][4][5]. However, existing citizen science is mostly based on spontaneous participation by adults and lacks a macro-level architecture for systematically integrating it into the national basic education system. Today, with the popularization of large AI models, the technical threshold for this transition has been completely broken through.

This paper proposes the "AI mentorship model" as the core teaching mechanism: large AI models serve as junior mentors and automatic programmers, decomposing complex scientific research tasks into "building-block" steps consistent with the laws of students' cognitive development, and automatically generating scripts for controlling automated equipment; students interact with AI through natural language, understand experimental design, issue task instructions, and observe num

data results; human teachers are responsible for higher-level guidance, ethical supervision, and stimulating innovation. The three work in concert to realize inclusive science education with "no/low thresholds" —students can participate in real scientific research without needing to learn a programming language.

[1] In July 2026, the World Young Scientist Greater Bay Area Frontier Dialogue will officially release a joint R&D project centered on the framework of this paper.

Example: A student says to the AI through a mobile App: "AI teacher, today I found that the water by the river is a bit turbid, and I would like to test the biodiversity here." The AI then guides the student through standardized operations such as positioning, taking photographs, collecting water, and filling in metadata, and automatically uploads the sample information to the cloud. Throughout the process, the student only needs to interact by voice or text, without manually writing any code.

"Research for all" does not require every student to become a scientist; rather, it gives every student the opportunity to participate in some part of real scientific research—perhaps sampling, perhaps data annotation, perhaps directing AI through natural language to complete an experiment, or perhaps verifying a hypothesis. When tens of millions of nodes contribute simultaneously, scientific problems that would originally have required decades to solve may be conquered within a few months. This is precisely the power of "synchronized convergence."

Eliminating thresholds: whether it is an AI-assisted TCM consultation system or an organoid automated workstation (eStation), the original intention of their design is to break the high entry barriers of traditional scientific research. Students do not need to master abstruse theories or costly experimental techniques; by interacting with AI through natural language, they can complete cutting-edge explorations ranging from constitution identification to gene-editing simulation. The essence of the AI mentorship system is to transform "professional thresholds" into "interest-based guidance," giving every young person the chance to become a research participant in which "the capable take the lead."

## 2.2 Six-Step Progressive Path: From Low Thresholds to High Value, from Modern Science to Chinese Wisdom

The six-step progressive path designed in this paper (Fig. 2) follows the principle of "low-threshold entry and high-ceiling extension." Each stage has explicit competency assessment and credit certification based on "the capable take the lead" :

First stage: AI science simulator (zero cost, zero risk). Students understand the scientific inquiry process through the simulator. The AI tutor decomposes complex tasks and demonstrates simulated experiments; students participate through simple operations such as clicking and dragging, thereby developing initial interest.

Second stage: real-environment DNA sampling. Students use standardized kits to collect water samples and record metadata such as GPS and water temperature. Through a mobile App, the AI guides the sampling steps by voice/text to ensure standardized operation. Samples are sent to a regional sequencing center, and data are automatically uploaded.

Third stage: AI-controlled desktop automated workstation (eStation). Drawing on the trend of automation in frontier biomedicine [6][7][8], low-cost desktop-level automated workstations are developed. Students describe experimental needs to the AI in natural language—for example: "Please aspirate 200 microliters of liquid from sample tube 1 and transfer it to position A1 of the 96-well plate." The AI automatically generates the script that controls the robotic arm and executes it. This is the core practice of "AI-driven scientific research" — students do not need to program, yet can command real equipment to operate.

Fourth stage: cloud-based AI laboratory (Cloud Lab). It connects with mature cloud laboratory platforms such as Emerald Cloud Lab and Strateos [9][10]. Students describe experimental protocols in natural language; the AI automatically converts them into API calls, remotely executes the experiment, and returns the data automatically. Because the equipment is expensive, this stage may first be piloted in regional STEM centers.

**Table 1: Six-Step Progressive Pathway for Research Participation**

| Stage | Name | Core Activities | Risk Level | Competency Certification |
|---|---|---|---|---|
| 1 | AI Science Simulator | Conduct virtual experiments under AI guidance; understand scientific procedures | Zero | Basic Operation Certificate |
| 2 | Real eDNA Sampling | Collect water samples outdoors; record metadata; perform standardized operations under AI guidance | Very low | Sampler Qualification |

| Stage | Name | Core Activities | Risk Level | Competency Certification |
|---|---|---|---|---|
| 3 | AI-Controlled Desktop Workstation (eStation) | Through natural-language instructions, AI automatically controls a robotic arm to process samples | Low | AI Collaboration Certification |
| 4 | Cloud AI Laboratory (Cloud Lab) | Describe experimental requirements in natural language; AI automatically calls APIs to execute remote experiments | Medium (controlled) | Advanced AI Collaboration Certification |
| 5 | Organoids / Digitization of Traditional Chinese Medicine | AI-assisted data annotation and knowledge-graph construction (under mentor supervision) | High | Research Assistant Credential |
| 6 | Feature Mining in Complex Systems of Chinese Civilization | AI-assisted analysis of traditional medical data; exploration of high-dimensional emergent features | Extremely high (special ethical approval) | Future Explorer |

Fifth stage: organoids / digitization of traditional Chinese medicine (a channel for top talent). Only students aged 18 and above who have passed the assessments of the first four stages may, under the guidance of mentors in university biosafety laboratories, use AI-assisted tools to participate in auxiliary work such as organoid image annotation and the construction of knowledge graphs for traditional Chinese medicine formulas. This is a bridge connecting "research by all" with "frontier biomedicine."

Sixth stage: feature mining of the complex system of Chinese civilization (future outlook). After distributed sensing networks and AI algorithms have matured, cautious exploration may be carried out on dynamic-response data of the human complex system based on traditional Chinese medicine theory, such as the biophysical properties of acupoints and multimodal data on mind–body interaction. AI will help establish mathematical models and reveal high-dimensional emergent features.

**Innovation Funnel: "Achievement First" in Six Progressive Steps**

*Diagram text, from top to bottom:*

- All K12 students (tens of millions)
- L1: AI science simulator (zero threshold)
- L2: real-world eDNA sampling (extremely low risk)
- L3: AI-controlled desktop workstation
- L4: cloud-based AI laboratory access
- L5: organoids / traditional Chinese medicine digitization (preparatory)
- L6: feature mining of the complex system of Chinese civilization
- Top innovative talent

*Side annotations:*

- Capability assessment; credit credential
- Sustained contribution; achievement first

*Design principle:* "Low threshold, high ceiling" ; capability is not limited by age, and contribution determines advancement.

**Figure 2. Innovation funnel: the "achievement first" six-step progressive research-participation pathway**

## 2.3 Difficult Problem Segmentation, Synchronous Convergence ( "Gathering Sand into a Tower" ), and National Innovation Competitiveness

Frontier problems in modern science—such as ecological early warning, new-drug development, and climate prediction—are often too complex for any single institution to overcome independently. This paper proposes a methodology of "difficult problem segmentation" : breaking down grand scientific problems into miniature topics that can be explored in parallel, and addressing them through a distributed K12 network.

Taking ecological monitoring as an example: traditional methods rely on professional teams conducting seasonal sampling, resulting in extremely low spatial and temporal resolution. A K12 network, however, can divide "nationwide monitoring of river biodiversity" into thousands of subtasks involving "one-kilometer river sections around schools," with students sampling every week. Once these dispersed contributions are aggregated, they form an unprecedented ecological dynamic atlas with high spatial and temporal resolution. This is precisely the power of "synchronous convergence" —when tens of millions of nodes contribute simultaneously, tasks that were originally impossible become feasible.

Standardized management provides the engineering foundation for "synchronous convergence" : drawing on the design philosophy of standardized interfaces in information networks, a K12 distributed research network can define a unified task-description language and data-metadata standards for each node—AI simulators, eDNA samplers, and eStation workstations. As long as the top-level design establishes this set of "research-node interface specifications," schools in different regions can plug in and use them immediately, without redundant development, thereby enabling collaboration among millions of nodes.

The distributed implementation pathway for "AI *Bencao Gangmu* 2030" is as follows: taking the digitalization of traditional Chinese medicine as an example, the grand objective of "building a Chinese-herbal-medicine–organoid response database" can be divided according to regional resource endowments. For example, the Lingnan region could focus on authentic regional medicinal materials such as dried tangerine peel and patchouli; the Southwest could be responsible for *Panax notoginseng* and *Gastrodia elata*; and the Northwest could concentrate on goji berries and licorice. With AI assistance, K12 students in each region would use standardized workstations to conduct organoid experiments on assigned medicinal herbs, with data automatically converging on a central platform. This model of "thematic division, regional division, and synchronous convergence" substantially reduces the cost of single-point research, naturally forms a nationwide functional atlas of medicinal herbs, and provides an actionable technical route for "AI *Bencao Gangmu* 2030."

This is not merely a technical strategy; it is also a source of national competitive advantage in innovation. Whoever can mobilize the broadest socialized research forces in the intelligent era will gain the initiative in data-intensive scientific discovery.

### 2.4 Mathematical Mechanism of Synchronous Convergence: From Random Noise to Coherent Signals

Whether a distributed K12 network can transform the weak contributions of tens of millions of students into trustworthy data assets at the national scale depends on whether it possesses a mathematical foundation capable of "phasealigning" heterogeneous and biased observations to a unified reference system. The research team's early theoretical work—the HoloTSH framework—uses block term decomposition (BTD) to decouple the systematic biases of different devices into device-related modal factors, while compressing the true signals reflecting ecological or biological system states onto a low-dimensional invariant manifold. Spectral manifold mapping then projects observations from each node into the same space, and hypergraph learning further encodes high-order synergistic relationships among multiple species and multiple components [14]. The combination of these three elements is equivalent to providing a "universal phase reference system" for the distributed network. Under this framework, even if the sampling of an individual student contains certain errors or device biases, after decomposition, mapping, and reconstruction, that student' s contribution can still be coherently superposed with information from other nodes, rather than canceling each other out or being submerged as noise. This is the fundamental distinction between this framework and traditional data aggregation: it is not simple "addition," but rather the use of mathematical structure to ensure that "garbage in" does not necessarily lead to "garbage out," thereby allowing every student' s participation to generate a valuable signal contribution.

## 3 Governance Architecture: Institutional Design for Building a "K12 Distributed Knowledge-Production Network"

Incorporating nearly 200 million underage students into a national real-world research system faces its greatest challenge not in technical bottlenecks, but in ethical risks and privacy-compliance issues arising from power asymmetries. To this end, this paper proposes a "dynamic governance architecture for minors' participation in research" based on frontier privacy-preserving computation and distributed technologies.

### 3.1 Task Crowdsourcing and the "Contribution Spectrum"

This architecture establishes a national-level platform for crowdsourcing scientific research tasks and the concept of a "contribution spectrum." Leading research institutions can decompose and release frontier research tasks, enabling students to transition from simulator testing to data collection in real-world environments. More importantly, the system creates standardized metadata descriptions for "negative results," granting them scientific value recognition equivalent to that of "successful discoveries."

### 3.2 Hierarchical Dynamic Authorization Based on Distributed Digital Identity

Data collection involving minors faces extremely high thresholds under the *Personal Information Protection Law* [11]. To address this, distributed digital identity technology [12] is introduced to establish an independent data-sovereignty node for each student. Different authorization thresholds are set according to the sensitivity of the data (Level A: environmental observations; Level B: self-reported bodily perceptions; Level C: biometric features), and students or their guardians may revoke authorization at any time with a single click.

### 3.3 Non-transferable Academic Credit Credentials and the "Preventive Medicine" Data Space

How can students be motivated to participate continuously? This paper proposes "Non-transferable Academic Credit Credentials" [13], technically implemented as soulbound credentials. When students complete valid data collection, submit high-quality negative reports, or unlock advanced tasks, the system issues credit credentials with timestamps and hash-based evidence preservation. These credentials cannot be traded or transferred; they remain permanently bound to the student' s digital identity and become an authentic foundational database for comprehensive quality evaluation.

Anti-fraud, local endowment, and credibility-equity assurance mechanisms: This system ensures the authenticity of the D-Index and its fairness and comparability across schools through three mechanisms. First, dual-witness signatures—each sampling event must be simultaneously confirmed and uploaded by two people, eliminating single-point fabrication. Second, full-process hashing on-chain—all operations from edge devices to data upload generate tamper-proof audit logs, removing room for human intervention. Third, equity through natural geographic resources—in response to the hardware disadvantages of old revolutionary base areas, ethnic minority areas, border areas, impoverished areas, or underdeveloped watersheds, the system introduces an "ecological uniqueness weighted score" at the mechanism level. Although central cities possess advantages in capital and informatization, remote watersheds and forest headwaters naturally possess uncontaminated, extraordinarily rich, exclusive ecological manifolds (a high-dimensional blue ocean of eDNA). On the basis of their distinctive geo-ecological characteristics, localities directly confer on local youth an irreplaceable "sovereignty over raw data." This natural-resource endowment offsets the spatial mismatch of traditional educational hardware, establishing opportunity fairness and educational equity in an ecological sense. At the same time, the system randomly selects 1%–2% of samples and sends them to gold-standard laboratories for independent retesting, keeping the inter-site coefficient of variation within 20% to ensure the scientific rigidity of macro-level selection.

Trusted scientific research certification system: In addition to the three mechanisms described above, the system will also introduce a "random flight inspection" mechanism (AI or experts randomly recheck a subset of samples) and an "anomalous data-pattern recognition" algorithm (AI automatically identifies data patterns indicative of batch fabrication), forming a continuously iterating "trusted scientific research certification system."

Mathematical correction foundation: Beyond the institutional and technical audits described above, this framework can rely on the previously proposed tensor–spectral manifold homomorphic mapping framework (HoloTSH) [14] to complete mathematical correction of systematic biases among heterogeneous devices before data aggregation, thereby from the source

suppressing the risk of "Garbage In, Garbage Out" and ensuring that the mathematical premise of "same-frequency convergence" holds.

Here, the concept of "treating disease before it arises" is extended to the field of ecological monitoring: by capturing geodesic deviations between environmental-state trajectories and the healthy manifold at high frequency, critical slowing-down signals are detected before irreversible phase transitions occur in the system—such as eutrophication collapse or species extinction—thereby realizing "ecological treatment before disease arises."

### 3.4 Underlying Compliance Technical Support: Data Security Packages and Computing-Power Allocation

To realize compliant data circulation among distributed nodes, the framework relies on core RegTech foundational technologies from existing research, such as "distributed big-data edge security protection packages" and "multi-center dynamic computing-power allocation algorithms"[14][15]. The former uses mathematical transformations to make biological-feature data "usable but invisible" ; the latter decomposes complex computational tasks into subtasks that can be executed in different judicial jurisdictions, dynamically schedules computing resources, and avoids compliance risks caused by the physical migration of data.

## 4 Limit Validation: Starting from Environmental DNA Watershed Monitoring

### 4.1 Why Start with eDNA?

Environmental DNA (eDNA) watershed biodiversity monitoring is selected as the first-stage limit-validation scenario for the following reasons: - Zero ethical risk: it only involves water-sample collection and does not touch human or animal living bodies. - High data credibility: central laboratories generate objective and reproducible gene-fragment abundance data through high-throughput sequencing. - High policy alignment: it closely matches strategies such as "Beautiful China" and "data elements × green development" [18]. - Perfect mathematical fit: the spatiotemporal distribution of eDNA data in river networks exhibits high-dimensional sparsity, seasonal periodicity, and nonlinear coupling among multiple species, making it a natural physical field for tensor decomposition and spectral-flow analysis.

### 4.2 Distributed eDNA Observation Network and "Pre-Disease" Ecological Early Warning

Let the ecological-state trajectory of a given water body be $\mathbf{X}(t)$, and let the healthy equilibrium state correspond to a low-dimensional invariant manifold

$\mathcal{M} \subset \mathbb{R}^n$. Using the open-source neuro-symbolic tensor-logic architecture [14], geodesic deviations $\mathrm{dist}(\mathbf{X}(t), \mathcal{M})$ of the state trajectory away from the healthy basin of attraction can be captured at high frequency. When the deviation grows at an accelerating rate and the fluctuation variance increases, it means that the system is approaching a bifurcation boundary. Traditional methods can issue warnings only after a phase transition has occurred, whereas this framework can intervene in advance during the critical slowing-down stage—this is precisely the ecological embodiment of the concept of "treating disease before it arises."

### 4.3 Mathematical Homology: From eDNA Tensors to TCM Medicinal-Compatibility Hypergraphs and Then to Life-System Networks

The methodology proposed in this paper has cross-domain universality. Block Term Decomposition (BTD) of eDNA spatiotemporal tensors can decouple the hydrological features of different watersheds into independent subspaces; the same algorithm can be directly applied to Chinese materia medica

"Jun-Chen-Zuo-Shi" hypergraph learning for multi-drug synergistic networks—each medicinal ingredient is regarded as one mode in a tensor, and herb-pair synergy is regarded as a low-rank approximation. Both are "high-dimensional sparse tensors" : eDNA is represented as (time $\times$ location $\times$ species), while Chinese medicine is represented as (time $\times$ formula $\times$ component).

Mathematical structural homology shows that, let the eDNA data tensor be $\mathcal{X}_{\mathrm{eco}} \in \mathbb{R}^{I\times J\times K}$ and the Chinese-medicine compatibility tensor be $\mathcal{X}_{\mathrm{tcm}} \in \mathbb{R}^{U\times V\times W}$. Through low-rank reconstruction by block-term tensor decomposition (BTD), both can be formally characterized as a modal product mapping of a core tensor and factor matrices [14]: $\mathcal{X} \approx \sum_{r=1}^{R} \mathcal{G}_r \times_1 \mathbf{A}_r \times_2 \mathbf{B}_r \times_3 \mathbf{C}_r$. From the perspective of graph signal processing, the hypergraph Laplacian operators $\mathcal{L}$ of the two exhibit highly consistent spectral-gap distributions. In algebraic topology, this supports an invariant manifold transformation mapping $\mathcal{L}_{\mathrm{eco}} = \Phi(\mathcal{L}_{\mathrm{tcm}})$. Using this homological relationship, the multidimensional interspecies coexistence patterns observed in watershed eDNA topological flows can be directly extrapolated and mapped to the compatibility-antagonism kinetic constants of specific molecules in the complex system of Chinese medicine. Rigorous mathematical confirmation of this low-rank approximation formula and of the convergence boundary of invariant manifolds may be found in existing foundational methodological studies [14][15].

Furthermore, human life-system networks—such as meridians and acupoints—can also be modeled as high-dimensional flows, and their remote effects can likewise be quantified by geodesic-deviation theory [14]. It has been demonstrated through the high-dimensional topological robustness auditing guardrail mechanism [15] that, when medical foundation models face extremely sparse, high-noise, non-standard data, topological constraints can achieve effective convergence, thereby completing the historic leap of non-standard knowledge from "experiential inheritance" to a "deterministic axiomatic model" [14][15][16].

Defining the standard: choosing Chinese medicine as a "limit-test"scenario serves not only to verify the framework' s capacity to handle high ethical risk and non-standardized data, but, more importantly, because its success will open a new track for China in global preventive medicine and complex systems science—one independent of the Western reductionist paradigm and capable of exporting standards. This is precisely the strategic significance of "defining the future" rather than "following the future."

### 4.4 Research Prospects and Verification Path

To move the "Laplacian spectral homology mapping"in this framework from theoretical metaphor toward hard empirical verification, the following three fundable and falsifiable deep-dive research topics are planned for the next stage: 1. An adversarial manifold-transfer experiment to verify the homology mapping Φ: design a dedicated neurosymbolic generative network that projects the shear-layer laminar-flow topological features of river eDNA, as constraints, into the Chinese-medicine compatibility hypergraph, and quantitatively test the reconstruction accuracy of cross-transfer of spectral-domain features. 2. A multicenter dynamic calibration algorithm under an extremely sparse RegTech environment: among high-noise physical nodes on the scale of tens of millions of grids, quantitatively evaluate the coupled interference characteristics of spatial variation in geographic microclimates and endogenous noise in microfluidic workstations on the rate of low-rank tensor completion. 3. Verification of the convergence boundary of the cross-modal tensor spectral gap: using the real-time in vitro organoid responses of classical formulas, such as Ephedra Decoction and Cinnamon Twig Decoction, as the test field, and through high-throughput sequencing, empirically decouple the geodesic approximation limits of multispecies trophic-level hypergraphs and multicomponent pathway hypergraphs.

## 5 Strategic Recommendations and Policy Pathways

### 5.1 Establishing Sandbox Pilots: Synchronized Advancement in the Guangdong-Hong Kong-Macao Greater Bay Area and Xiong' an New Area

It is recommended that pilot programs be launched first in the Guangdong–Hong Kong–Macao Greater Bay Area—specifically the Hetao–Qianhai corridor—and in Xiong' an New Area simultaneously. Relying on the institutional advantages of "One Country, Two Systems" and its foundation for deep Hong Kong–Shenzhen collaborative innovation, the Greater Bay Area can take the lead in exploring implementation models for compliant cross-border data flows and AI-driven scientific research networks. Xiong' an New Area, as a benchmark for digital-twin cities, should focus on validating the integration of standardized interfaces and an all-domain sensing system. Together, the two regions would form an institutional innovation pattern with north–south complementarity, accumulating diverse experience for subsequent nationwide scaling.

## 5.2 Interministerial Joint Conference Mechanism and Collaborative Innovation

It is recommended that the Ministry of Ecology and Environment take the lead, in conjunction with the National Data Administration and the Ministry of Education, in establishing an interministerial coordination group for the "National Civic Science Education Infrastructure for the Intelligent Era." The Ministry of Ecology and Environment would be responsible for issuing ecological monitoring tasks and data standards; the National Data Administration would oversee rules for data-asset rights confirmation and circulation; and the Ministry of Education would be responsible for incorporating academic credit credentials into the Strong Foundation Plan and comprehensive quality evaluation [16][18][20]. To prevent frontier observation networks from giving rise to a new form of digital divide, the interministerial joint conference mechanism must establish a "Special Transfer-Payment Compensation Plan for Science Education in Old Revolutionary Base Areas, Minority Areas, Border Areas, Poor Areas, and Mountainous Areas." Through an open-source computing-power allocation matrix based on sandbox nodes, it should remotely empower schools in western China and fully subsidize the sampling and sequencing costs for students in remote regions, ensuring opportunity inclusion for the D-Index on a national scale.

Recommended charter for collaborative innovation: to resolve potential tensions among departments, it is recommended that the three major departments jointly formulate the following rules:
- The Ministry of Ecology and Environment issues a "National Ecological Monitoring Task List," specifying standards and priorities for data collection;
- The Ministry of Education designs "Standards for Research-Practice Credits," incorporating students' data contributions into comprehensive quality evaluation;
- The National Data Administration formulates "Measures for the Asset-Based Management of Youth Scientific Research Data," clarifying the attribution of data property rights and rules for use.

## 5.3 The Aptitude-Based Education Sandbox: A Special Innovation Pathway for Exceptional Children

To resolve the deep institutional friction between distributed credit networks and China' s current unified college entrance examination evaluation system, this paradigm, in its policy design, by no means seeks to alter or replace the standardized mass-admission foundation. Rather, it serves as a special parallel channel for implementing the national strategies of "teaching according to aptitude"and "teaching according to talent."Through the Level 1 and Level 2 universal science-popularization filters, schools can identify the very small number of "unconventional youths" who display extraordinary talent in data annotation, biochemical programming, or ecological manifold disentanglement. After objective profile audits through school recommendations and the data-customs DID hash chain, the platform unlocks for them advanced operating privileges at Level 3 and above. These gifted students may directly cross traditional age thresholds and, in the form of virtual and physical research assistants, enter university key laboratory sandboxes to receive joint mentorship from AI and national-level scientists.

## 5.4 Promoting Legislation on Research Data Involving Minors

It is recommended that the National Data Administration, together with the relevant legislative bodies, initiate the drafting of regulations

18

to make clear that environmental/non-sensitive data generated by minors within controlled research networks shall be owned by the individual students themselves, with protection provided through distributed digital identity; anonymized datasets shall constitute national strategic assets, and illegal cross-border transfer shall be strictly prohibited

11

.

## 5.5 The National Strategy of "Synchronous Convergence" : From Distributed Nodes to National-Level Data Assets

It is recommended that a "National Research Data Convergence Program for All" be established at the national level, incorporating standardized data produced by K12 networks—such as eDNA, carbon-neutrality data, and observational data —into national scientific data centers as a benchmark case for the "Data Elements ×" strategy

18

. This is not merely an accumulation of data volume, but a revolution in the paradigm of data production—from "passive collection" to "active co-construction," and from "scattered sand forming a tower" to "synchronous convergence."

## 5.6 Limitations of the Proposal and Boundary Statement

It must be stated clearly that, as a macro-level policy proposal targeting the foundational base of new-quality productive forces, the implementation of this work still needs to overcome profound institutional frictions. First, when the interministerial joint conference faces the inherent departmental tensions of "the Data Administration wants circulation, the Ministry of Education wants burden reduction, and the Ministry of Ecology and Environment wants absolute standards," a long-term top-level coordination mechanism will still be required for institutional integration. Second, although the Data Contribution Index (D-Index) is cryptographically protected, in the highly sensitive and highly competitive context of university-entrance selection, subjective and objective disputes over fraud prevention and machine adjudication will still require multiple rounds of closed-door stress testing within a sandbox. Finally, before the theory of high-dimensional tensor homomorphic mapping is supported by ultra-large-scale population physiological data, its cross-domain generalizability still awaits longer-term experimental validation. Taking on and directly facing these governance and theoretical boundaries precisely constitutes the empirical foundation of this study as a rigorous policy-science recommendation.

## 6 Conclusion: Five Thousand Years of Chinese Civilization —A Blue Ocean for the Next Generation of Data Science

National competition in the intelligent era is shifting from a "race in computing-power armament" toward a "complex contest over data assets and innovative talent."The K12 AI-driven research network proposed in this paper is, in essence, not a technical scheme, but a paradigmatic revolution in science education. It attempts to answer the question: when cloud laboratories and autonomous AI scientists become the norm, will our next generation possess the cognitive literacy for "AI-collaborative research" ?

From AI simulators to eDNA sampling, from AI-control desktop workstations to cloud-based AI laboratories, and from organoids/traditional Chinese medicine digitization to mining complex systemic features of Chinese civilization, the six-step progressive pathway provides students with an accessible, gradual, and fulfilling channel for participation in scientific research. Meanwhile, the strategic thinking of "problem segmentation and synchronous convergence" makes large-scale scientific tasks that originally seemed impossible become feasible



.

Strategic implementation dynamics: In July 2026, the Frontier Dialogue of Young Scientists in the Greater Bay Area will formally release the "Greater Bay Area K12 AI-Driven Research Network Joint R&D Project," centered on the framework of this paper, as well as the "Guangdong-Hong Kong-Macao Greater Bay Area Future Edu···

···Education Innovation Cooperation Initiative," marking that the distributed scientific-research network envisioned in this paper has taken the lead in entering the stage of practical validation in the Greater Bay Area.

Looking ahead, the distributed scientific-research network proposed in this paper, together with the methodology of "problem decomposition and same-frequency convergence," will directly serve major national strategic needs. For example, the vision of "AI *Bencao Gangmu* 2030" —to build the world's first large-scale, standardized response database for Chinese herbal medicines and organoids—can be advanced by assigning topics through the K12 network according to regional resource endowments: tens of millions of students would each contribute experimental data on one medicinal herb, ultimately converging into a dynamic and computable modern pharmacological encyclopedia. This would not only democratize scientific discovery, but also constitute the collective creation of Chinese civilization in the digital age.

Those complex-system characteristics in five thousand years of Chinese civilization that "seem not to be quantifiable" —holism, dynamic equilibrium, and the unity of heaven and humanity—are not metaphysics, but high-dimensional emergent features of complex systems. When a distributed sensor network covers the ecological and cultural nerve endings of 9.6 million square kilometers, and when tens of millions of "AI collaborative researchers" learn to use data to "translate" knowledge that once could only be grasped intuitively, we will truly possess the confidence to reinterpret Chinese wisdom in the language of science [14][15].

This is not a technological revolution, but a movement of scientificization at the civilizational level. And its starting point is merely a beaker of river water, a natural-language command, and the curiosity of a K12 student.

## Acknowledgments

This research was jointly supported by the Hang Seng University of Hong Kong, Southern University of Science and Technology, the State Key Laboratory of Quality Research in Chinese Medicine at the University of Macau, and the School of Public Policy and Management at Tsinghua University. Related exploratory work received support during the preliminary research stage from the Hong Kong Education Bureau's "Smart" Start Program.

## References

[1] Epoch AI. Will We Run Out of Data? Analysis of the Future of AI Training Data[R]. 2024.

[2] Five departments including the Ministry of Education. "Artificial Intelligence + Education" Action Plan[Z]. 2026.

[3] Bonney R, et al. Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy[J]. BioScience, 2009, 59(11): 977-984.

[4] Vohland K, et al. The Science of Citizen Science[M]. Springer, 2021.

[5] Sturm U, et al. Citizen science in the age of AI: a review of current practices and future directions[J]. Citizen Science: Theory and Practice, 2024, 9(1): 1-15.

[6] Emerald Cloud Lab. Cloud Labs for Life Sciences[EB/OL]. https://www.emeraldcloudlab.com.

[7] Strateos. The Autonomous Drug Discovery Platform [EB/OL]. https://strateos.com.

[8] LabscriptAI. Multi-agent framework for automated lab control [EB/OL]. GitHub, 2025.

[9] Adaptyv Bio. Cloud-based protein engineering platform [EB/OL]. https://www.adyy.io.

[10] UniteLabs. AI-driven lab automation OS [EB/OL]. https://unitelabs.com.

[11] Standing Committee of the National People' s Congress. Personal Information Protection Law of the People' s Republic of China [Z]. 2021.

[12] W3C. Decentralized Identifiers (DIDs) v1.0 [EB/OL]. W3C Recommendation, 2022.

[13] Buterin V. Soulbound Tokens (SBTs): A New Form of Digital Identity [EB/OL]. 2022.

[14] Liu K, Li A, Hu H, Ho KC. HoloTSH: A Neuro-Symbolic Tensor Logic for TCM Modernization via Mathematical Isomorphism and Theoretical Guarantees [EB/OL]. TechRxiv, 2026.

[15] Liu K, Li A, Hu H, Ho KC. Scaling Laws Hit a Wall: Mitigating Semantic Sorting Collapse in Medical LLMs via the HoloAuditor Topological Guardrail [EB/OL]. SSRN, 2026.

[16] Liu K, Li A, Hu H, Ho KC. Building the "International TCM Data Port" : A Cross-Border Citizen Science Paradigm for the "Data Element ×" Strategy [EB/OL]. SSRN, 2026.

[17] Liu K, Li A, Hu H, Ho KC. The "ShenNong AI" Ecosystem: Building a "Digital Great Wall" for Adolescent Constitution Data [EB/OL]. ResearchGate, 2026. DOI: 10.13140/RG.2.2.26761.81762.

[18] National Data Administration and 17 other departments. "Data Element ×" Three-Year Action Plan (2024–2026) [Z]. 2024.

[19] Standing Committee of the Guangdong Provincial People' s Congress. Regulations on the Shenzhen Park of the Hetao Shenzhen–Hong Kong Science and Technology Innovation Cooperation Zone [Z]. 2025.

[20] CPC Central Committee and the State Council. Outline of the National Informatization Development Strategy [Z]. 2016.

[21] Ministry of Education. Opinions on Strengthening and Improving Experimental Teaching in Primary and Secondary Schools [Z]. Basic Education 〔2019〕 No. 16.

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