学术展望

Academic Perspectives

美国华裔教授学者协会学术期刊

ISSN 1936-1246

第18期

2026年

USA CHINESE SCHOLARS ASSOCIATION

Los Angeles, California, USA, October 10, 2026

美国华裔教授学者协会
USA CHINESE SCHOLARS ASSOCIATION

- 美国华裔教授学者协会是一个非盈利、非政治性的社区公益组织；
- 自1998年成立到现在已经走过了28个年头，是南加地区成立最早，历史最悠久的以学术交流为基础的专业人员团体之一；
- 云集了美国院士、国际知名奖项获得者、大学和研究机构的教授和学者，现有207位终身会员。

协会理念

CORPORATE PHILOSOPHY

促进中国学者在世界各地的知识交流和友谊

协会宗旨

ENTERPRISE TENET

促进中美两国科学、技术和高等教育的发展

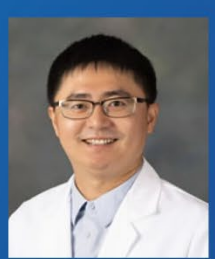
协会使命

ENTERPRISE SPIRIT

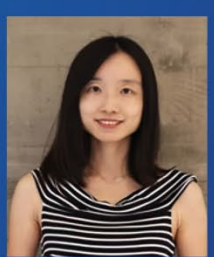
增进中美友谊和东西方文化的相互了解

美国华裔教授学者协会现任理事

会长
李永田



副会长
李明爽



秘书长
张佳辰

财务长
宋敏

社会科学理事
何绍义

自然科学理事
郦永刚

外联理事
蒋文森

- 美国华裔教授学者协会网站：http://usacsa.org/
- 美国华裔教授学者协会活动：专题讲座，学术刊物，年会报告

Academic Perspectives

学术展望

USA Chinese Scholars Association

美国华裔教授学者协会

Volume 18, 2026
ISBN 1936-1246

目录

Table of Contents

会长寄语

President's Message

李永田 / Yongtian Li

美国华裔教授学者协会

USA Chinese Scholars Association

值此美国华裔教授学者协会会刊《学术展望》2026年第18卷出版之际，我谨代表协会，向长期关心和支持协会发展的广大会员、专家学者及各界朋友，致以诚挚问候和衷心感谢。

《学术展望》作为协会重要的学术交流平台，始终致力于展示会员成果、促进思想交流、推动跨学科合作。本卷继续同步推出纸质版与电子版，体现了数字时代知识传播的开放与便利。

当前，科技创新日新月异，国际交流不断深化，广大华裔教授学者在教学、科研和社会服务中持续贡献智慧与力量。《学术展望》汇聚学术成果，推动交流合作，展现华裔学者的专业风采与责任担当。

本卷收录综述、评述及原创性文章，涵盖医学与生命科学、地球科学、社会科学、人文语言学、人工智能与城市治理等多个领域，体现出跨学科、多视角的鲜明特色。内容兼具学术探索性、理论思辨性与现实关怀，既关注自然科学与工程技术前沿，也回应人文社会科学的重要议题，是一期具有专业深度与应用价值的学术文集。

展望未来，协会将继续秉持服务会员、促进交流、支持创新、回馈社会的宗旨，加强平台建设，凝聚优秀人才，共同推动华裔学术共同体不断发展。

衷心祝愿《学术展望》越办越好，祝愿各位同仁学术精进、事业顺遂。

李永田

美国华裔教授学者协会会长

2026年9月

President's Message

On the occasion of the publication of Volume 18 (2026) of Academic Perspectives, the journal of the USA Chinese Scholars Association, I would like, on behalf of the Association, to extend my sincere greetings and heartfelt thanks to all members, scholars, and friends from various sectors who have long supported our development.

As an important platform for academic exchange, Academic Perspectives is dedicated to showcasing members' achievements, promoting intellectual dialogue, and encouraging interdisciplinary collaboration. This volume is published in both print and digital formats, reflecting the openness and convenience of knowledge dissemination in the digital age.

This volume includes review articles, commentaries, and original research papers covering medicine and life sciences, earth sciences, social sciences, humanities and linguistics, artificial intelligence, and urban governance. It highlights interdisciplinary perspectives and combines academic exploration, theoretical reflection, and practical relevance.

Looking ahead, the Association will continue to uphold its mission of serving members, promoting exchange, supporting innovation, and giving back to society. We will further strengthen our platform, bring together outstanding talent, and advance the development of the Chinese American academic community.

I sincerely wish Academic Perspectives continued success and all colleagues great achievement and fulfillment in their academic careers.

Yongtian Li
President, USA Chinese Scholars Association
September 2026

阳光诱导产生的气体信号分子沿人体丰富微循环路径参与经络系统的发育、生物分子构成、功能及生物物理现象

Sunlight-Induced Gasotransmitters along Body's Rich Microcirculation Pathways Involved in the Development, Biomolecules, Functions, and Biophysical Phenomena of the Meridian System

马胜兴 / Sheng-Xing Ma

Lundquist Institute for Biomedical Innovation at Harbor-University of California at Los Angeles (UCLA) Medical Center and Department of Obstetrics and Gynecology, David Geffen School of Medicine at UCLA and Harbor-UCLA Medical Center, Torrance, CA 90502

ABSTRACT

The meridian system flows "meridian Chi," and connects the surface of the body to the internal organs as the central theory of traditional Chinese medicine (TCM) and guiding principle for therapies. Although the essential pathways of the meridians have been described for thousands of years, the development, biomolecules, structure, and functions of the meridian system are unclear. Anatomical studies have shown that most acupuncture points (acupoints) are located intimately in the soft tissues with rich blood vessels, nerve trunks, and hair follicles between the muscle/tendon/bone. Biophysical studies have demonstrated that acupoints possess characteristics of low impedance, and the sensations propagated along acupuncture channel pathways (PSCP), and the tracers along linear pathways resembling meridians. As described in an ancient classical medical book: "There is Yang energy in the human body like there is sun in the sky." The sun primarily radiates energy as visible light, ultraviolet, and infrared radiation. Exposure of human skin to sunlight-induced thermal effects and ultraviolet radiation causes the release of nitric oxide (NO) and calcitonin gene-related peptide (CGRP). Sunlight also increases endogenous hydrogen sulfide (H_2S) and carbon monoxide （CO). Numerous studies have demonstrated that these gasotransmitters play very important roles in physiological regulations and pathological changes involved in various diseases. The endothelial surface glycocalyx (ESG) at the luminal side of the microvessel wall plays a crucial role in mechanosensing, transduction, and initiating the production of NO and H_2S for regulating blood flow and vascular health. Near-infrared light and labors/exercises induce the generation of gas transmitters is predominantly enhanced in the acupoints and/or meridian regions between muscles/tendons along somatic pathways, which mediate chemical-electrical activity through the gracile nucleus-thalamic-cortex pathways for PSCP generation. In this review, we summarize the current understanding of simulated sunlight (near-infrared light or heat) on

the generation of gasotransmitters in ESG at the luminal side of the microvessel wall and their impact on chemical-electrical activity of low skin resistance, PSCP, and the tracers along linear pathways resembling meridians. These studies should advance our understanding of the roles of thermal stimuli and near-infrared light from sunlight and labors/exercises-induced gasotransmitters over ESG at the luminal side of the microvessel wall in the development of the meridian pathways, and yield new insights regarding the biophysical characteristics of gasotransmitters in the meridians, which opens a light to explore the scientific basis including the biomolecules, structure, functions, and development of the meridian system in TCM.

Keywards: Gasotransmitters; sunlight; endothelial surface glycocalyx; propagated sensation along the meridian; tracers; axon reflex; nitric oxide; biophysical approaches; acupuncture points; meridian system

INTRODUCTION

The Meridian System (channels and collaterals, jingluo) is an essential pathway system that unifies all parts of the body and deals with physiological regulation and pathological changes of the human body according to the description of Traditional Chinese Medicine (TCM) [1,2]. Acupuncture points (acupoints) are located along the 12 main meridian pathways, and acupoints/meridians have been used in many unconventional medical practices such as acupuncture, electrical acupuncture (EA), transcutaneous electrical nerve stimulation (TENS), moxibustion (heating acupoints), Qi Gong, meditation, and Kungfu [1-3]. There has been a widespread and increasing interest in scientific studies explore the Meridian System and the use of acupuncture for the treatment of disorders over the world. However, the most logical and direct approach to identify a specialized anatomic entity of acupoints and meridians is still lacking [2-5]. One of the 125 major exploration and discovery questions is: “Is there a scientific basis to the Meridian System in traditional Chinese medicine?” as a frontier, global common need and foresight gathering released in 2021 by Shanghai Jiao Tong University and "Science" magazine [5]. There have been several hypotheses about the structure of the meridians, i.e., autonomic nerve, body fluid channel, electromagnetic field, and connective tissue interstitial [6-9]. Not all of these hypotheses are mutually exclusive. The scientific basis of the biomolecules, structure, functions, and development of the meridian system in the body is unclear.

Biochemical properties of sunlight-induced gas transmitters with rich distributions along skin acupoints/meridians.

The meridian system's flow, “meridian qi,” evokes dynamic processes such as energy exchange, movement, and connection from the surface of the body to the internal organs, as well as communication with the universe, environment, and our body as a whole. Various diseases are associated with dysfunctions of the meridians, and acupuncture stimulates acupoints on the body, releasing and easing the “meridian qi” for treating disorders[1-3]. "This is the great channel tunnel of Chi" - Pulse Degree 17 of "Huangdi Neijing". It is believed that the Meridians are neither nerve networks nor blood vessels, but channels through which Chi circulates [6-8]. Chi is an invisible and extremely fine substance that is constantly moving in the universe, and it is the origin from heaven (the sun) [10]. As described in Chapter 3 of "Huangdi Neijing": “There is Yang energy in the human body like there is sun in the sky. ……and the bodily health of a man depends on the clear and floating of Yang energy which guards against outside.” [11] The ancient descriptions clearly pointed out that the meridian Chi, particularly the Yang energy in the human body, is from the sun in the sky. Interestingly, growing observational and experimental evidence has shown that the sun is the most important energy source for life on Earth and

mainly radiates energy in the form of visible light, ultraviolet, and infrared rays [13-14].

Various studies have demonstrated that long-wavelength light, especially near-infrared light, has good tissue penetration depth [13-18]. When infrared rays irradiate the human body, it produces two major effects: thermal effect and non-thermal effect [13]. Thermal effect is the effect caused by the body's molecules transferring energy through collision or radiation under the action of electromagnetic waves. For every 1°C increase in body tissue temperature, tissue metabolism increases by 10%, thereby promoting local blood circulation and increasing cell and enzyme activity [13-14]. The thermal effect and ultraviolet exposure of human skin release nitric oxide (NO) and transient receptor potential vanilloid (TRPV), which induces calcitonin gene-related peptide (CGRP) [15-18]. TRPV3 is a receptor mainly responding to thermal stimuli of around 32 – 40 ºC, and there is 40% homology in protein sequence between the TRPV1 receptor and TRPV3 receptor [19-21]. Sunlight [heat and near-infrared light (NIL)] increases endogenous gasotransmitters, including NO, hydrogen sulfide (H_2S), carbon monoxide（CO）, and carbon dioxide (CO2) [13,14,22,23]. Gasotransmitters play very important roles in physiological regulation and pathological changes involved in various diseases, which allow us to understand why dysfunctions of the meridians are associated with many pathological changes [25-27].

Anatomical and morphological studies of acupoints/meridians along the soft tissues between the muscle/tendon/bone with rich endothelial surface glycocalyx structure/pathways and microcirculation are highly exposure to sunlight

Anatomical and morphological studies have identified that the most acupuncture points (acupoints) are located intimately at the soft tissues (connective tissue) with rich blood vessels, nerve trunks, and hair follicles between the muscle/tendon/bone in both humans and rats [9, 28-30]. The endothelial surface glycocalyx (ESG) at the luminal side of the microvessel wall is now recognized as a multifunctional and dynamic structure that participates in many vascular processes, including vascular permeability, mechanosensing, transduction, inflammation, thrombosis, and cytokine signaling [31-34]. The ESG consists of NO synthases, proteoglycans (PGs), glycosaminoglycans (GAGs), and glycoproteins, while GAGs and PGs are abundant structural components of the extracellular matrix, which have been evaluated as important contributors to skin aging [35-38]. Thus, ESG plays a crucial role in sensing blood flow and initiating the production of NO, while H_2S plays a protective role for ESG. PGs are present in all basement membranes, including skin and its appendages such as the epithelial compartment of hair follicles [31-34]. Consistent with the micro-CT/histological studies reported that long-distance fibrous matrices from the thumb to the right atrium for interfacial fluid transport [39], the fibrous connective tissue is composed of three basic components: fibers, including collagen fibers and elastic fibers, and a gelatinous matrix, which includes PGs and GAGs, and is consistent with ESG.

It is described in TCM that the meridian systems flow “meridian qi” evoking dynamic processes such as energy exchange, movement, and connection from the surface of the body to the internal organs, as well as communication with the universe, environment, and our body as a whole [1-3]. The “meridian qi”, which differs from external qi such as air, oxygen, hydrogen, clouds, and earth qi, is the essence of life remaining after a living being absorbs external "qi" such as area, oxygen into the body and excretes waste gases generated as "internal qi" which possesses a self-regulating mechanism for new metabolic processes, possessing both qi and energy, and circulates through the acupoints, meridians, and collaterals within the body, forming a life field [40,41]. Numerous studies have demonstrated that gasotransmitters, which include NO generated from L-arginine and oxygen, and H_2S generated from hydrogen and cysteine/homocysteine, play very important roles in physiological regulations and pathological changes involved in various diseases [25-27]. The

ESG at the luminal side of the microvessel wall are predominantly located in the vertical distribution of the fascia networkand soft tissues between the muscle/tendon/bone along somatic meridians [30-34], which play a crucial role in mechanosensing, transduction, and initiating the production of gasotransmitters for regulating blood flow and vascular health [35-38]. Consistently, these skin regions with rich blood vessels, nerve trunks, and hair follicles are highly exposed to sunlight. Most acupoints and meridian lines are located intimately at the soft tissues between the muscle/tendon/bone with ESG structure and rich microcirculation [28-30].

Several studies have demonstrated that NO and cGMP levels are higher over acupoints and meridian skin surfaces [43-45] and in the subcutaneous tissue by dermal microdialysis [46]. The investigation of NO levels at higher levels over the skin acupoints is confirmed by a different method from other studies [47], which consistently suggest that higher levels of NO exist on the skin surface of acupoints involved in physiological distribution and therapeutic manipulation of the skin microvasculature. Meridian channels/lines in humans and mammals have been developed through long-term biological evolution under the influence of labor, exercise, heartbeats, respiratory movements, and the stimulation, impact, and regulation from environmental temperatures, daily and monthly cycles, Earth's and celestial light, gravity, and cosmic radiation, especially sunlight exposure. The twelve primary meridians are distributed between the muscles and muscle tendons of the limbs, which are the areas with the greatest muscle contraction and tension deformation, and which are frequently exposed to sunlight. Hair in these areas is relatively dense, as are blood vessels and nerve fibers, which produce much more gasotransmitters in the regions. It appears that the interaction and distribution of sunlight-induced gasotransmitters along the pathways of the human body’s fascia network and soft tissues with rich microvessel walls and nerve adventitia contribute to the development of the meridian pathways with rich production of gasotransmitters. The biochemical and functional of the gasotransmitters are like the described characteristics of the "internal qi". Research studies of sunlight on induction of gasotransmitters and anatomical and morphological studies of accuoints/meridians have advanced considerably; however, the most logical and direct approach to identify a specialized anatomic entity of the meridian system and biomolecules of meridian Chi needs to be established in the future. A more sophisticated approach would be required to address this issue. Despite these limitations, increasing numbers of the publications would be consistent with sunlight-induced elevations of gasotransmitters over all skin regions at the soft tissues between the muscle/tendon/bone with ESG structure and rich microcirculation, and demonstrate a consistent response to high NO release at the most acupoints/meridians, which are located intimately at the connective tissues with rich blood vessels, nerve trunks, and hair follicles between the muscle/tendon/bone. The results from both anatomical and biochemical studies consistently suggest that gasotransmitter signaling molecules induced/generated by labor, exercise, heartbeats, respiratory movements, and environmental temperatures, gravity, and cosmic radiation, especially sunlight exposure, are located with a high level along acupoints/meridian at the soft tissues between the muscle/tendon/bone with rich ESG structure and microcirculation.

Gas transmitters play important roles in biophysical approaches of meridian phenomenon

Over the last few decades, there has been increasing evidence from biophysical, physiological, and biochemical approaches implicating some specificity for the acupoints in the body around the world. An abundance of studies from different international groups have demonstrated that acupoints possess characteristics of low electrical resistance and high electrical conductance, and the propagated sensation along channels (PSC) and tracers along linear pathways resembling meridians are the most commonly observed and identified meridian phenomena [2-4,6]. Previous anatomical studies have demonstrated that acupoints/meridians exist higher number of nerve fibers/trunks, blood vessels, hair follicles, and sweat glands [28-30]. We have shown that NOergic signaling molecules and neuropeptides play important roles in mediating the skin conductance responses to electrical stimulation

and contribute to low electrical resistance and high electrical conductance of acupoints/meridians [48]. 3H-NE synthesis/release was enhanced in acupoints and facilitated by the presence of an exogenous NO donor and inhibited by an inhibitor of NO synthesis [49]. A recent review summarized that NO concentrations are enhanced in skin acupoints/meridians, and L-arginine-derived NO synthesis and noradrenergic transmission modify skin electric conductance, which contributes to low resistance characteristics of acupoints and meridians (one of the meridian phenomena) [50].

Several studies demonstrated that low force/rate of several studies demonstrated that low force/rate of transcutaneous electrical nerve stimulation (TENS), manual acupuncture (MA), electroacupuncture (EA), and electrical heat produce an elevation of NO release in all skin areas, including non-acupoints and non-meridian skin regions but with higher levels at acupoints [51-55]. Consistently, these stimuli, TENS, electrical heat, and EA have been used to elicit axonal reflexes, which are characterized by local release of NO and CGRP mediated by activation of TRPV1 [56-58]. TENS, MA, EA, and electrical heat produce an elevation of NO release in all skin areas, including non-acupoints and non-meridian skin regions but with higher levels at acupoints. In addition, the stimuli-induced release of NOergic molecules (NO synthases, NO, and cGMP) with high levels in the acupoints containing rich neuronal components and blood vessels mediated axon reflexes-evoked elevation of cutaneous blood flow may spread over acupoints one after another along the median lines, differing from nerve pathways following the stimuli to induce PSCP [59]. The same driver also elicits NO release in the gracile nucleus, which contributes to the somatosensory signal transduction of PSCP perceptions through the dorsal medulla-thalamic pathways. The stimuli-evoked biomolecular changes in the acupoints and gracile nucleus play important roles in transductions of the sensory signals induced by PSCP through the acupoints-dorsal medulla-thalamic pathways [60-62]. The PSCP perceptions are similar to the “de qi” sensation during acupuncture treatment. That is the reason why stronger “de qi” sensation usually possesses better therapeutic effects following MA, EA, electrical stimulation, and heat therapies. NO-mediated axon reflex over acupoints along the median lines [56-58] and the sensory signals mediated by NO release in the gracile nucleus through the acupoints-dorsal medulla-thalamic pathways may also participate in the processes of “de qi” sensation as well as produce therapeutic effects through central regulation induced by the stimuli [59]. When acupoints rich in blood vessels and nerve fibers are stimulated, they release high concentrations of NO-related molecules (NO synthase and cGMP), which mediate axon reflexes that increase skin blood flow. This can spread from point to point along the meridian pathways (different from the stimulated nerve pathways), inducing sensory propagation (numbness, pressure, heaviness, and/or warmth perception, PSCP). The same stimulus also causes NO release in the thalamic nuclei, helping convey somatosensory signals for PSCP perception through the dorsal column-thalamic pathway [59-62].

Moreover, an abundance of studies from different international groups have demonstrated that tracers along linear pathways resembling meridians over the body surface of humans [63-66]. All experiments of the studies have been conducted by injecting a radiotracer solution or tracer dyes in a volume of solution into acupoints [63-66]. Solutions injected into acupoints produce much stronger mechanical stimuli than acupuncture, which causes an axon reflex [67]. Anatomical studies have demonstrated that acupoints/meridians have a higher number of small nerve fibers and blood vessels with rich NO, CGRP, and neuropeptides in the cutaneous tissues as structures for the biomolecule-mediated axon reflexes. Recent advances have determined that NO, CGRP, and neuropeptides play crucial roles in the comprehension of the axon reflex [56-58]. The stimuli-evoked axon reflex and NOergic biomolecules/neuropeptides increase local blood flow with higher levels in acupoints/meridians, which move radioactive substances or tracer dyes in the skin and subcutaneous tissue along a linear path resembling acupoints and meridians, an important phenomenon of meridians induced by the stimuli [63-66]. The evidence and understanding of the biomolecular processes of the tracers along linear pathways resembling meridians have been summarized in a review paper with an emphasis on recent developments of NO, CGRP, and neuropeptides mediating stimuli-evoked axon reflexes to increase local blood flow with higher levels in acupoints/meridians, which move radioactive substances or tracer dyes in the skin and subcutaneous tissue contributing to tracers along linear pathways resembling meridians [67].

NO signaling molecules and neuropeptides (CGRP, SP, and NPY) are involved in the stimuli-evoked axon reflexes, and the same driver-elicited biomolecules-mediated axon reflex may also participate in the somatosensory signal transduction and processes of low electrical resistance of acupoints/meridians and PSC/LPSC. The stimuli-evoked NOergic biomolecules-mediated axon reflex increases local blood flow, which moves radioactive substances or tracer dyes in the skin and subcutaneous tissue along a linear path resembling acupoints and meridians, another important phenomenon of meridians induced by the stimuli. In this paper, the evidence and understanding of the biophysical approaches of acupoints/meridians have been discussed and summarized with an emphasis on recent development of stimuli-evoked release of NOergic biomolecules and neuropeptides through the axon reflex and mediating meridian (Jingluo) phenomenon. Although more sophisticated approaches are required to address other gasotransmitters, particularly H_2S, CO, and CO_2 contributing to biochemical-physiological interactions in these processes in this fields, the results from animal and human studies consistently suggest that the stimuli-evoked NOergic biomolecules mediated axon reflex increase local blood flow, which are involved in the processes of low electrical resistance of acupoints/meridians and PSC/LPSC as well as moving radioactive substances or tracer dyes in the skin and subcutaneous tissue under a linear path resembling acupoints and meridians, the all current phenomena of meridians induced by the stimuli.

CONCLUSION

The meridian system flows “meridian Chi,” and connects the surface of the body to the internal organs as the central theory of TCM and guiding principle for therapies. Although the essential meridian pathways have been described for thousands of years, the meridian development, biomolecules, structure, and functions of the system are unclear. There are various studies from biophysical approaches dealing with the phenomena of meridians and have demonstrated that acupoints/meridians possess characteristics of low electrical resistance, the PSCP, and the tracers along linear pathways resembling meridian pathways during the last few decades. Previous anatomical studies have demonstrated that acupoints/meridians exist higher number of nerve fibers/trunks, blood vessels, hair follicles, and sweat glands. Recent studies have demonstrated that NO and cGMP levels are higher over acupoints and meridian skin surfaces and in the subcutaneous tissue. L-arginine-derived NO synthesis modifies skin NE synthesis /release in acupoints/meridians, and NO-NE activations play an important role in mediating the skin conductance responses to electrical stimulation. Acupoints have higher levels of nervous components, and NOergic signaling molecules (NO-NE, and NO-gap junction/TRPV1) and neuropeptides are also involved in the skin's low resistance at acupoints.

The PSCP and the latent PSCP (LPSCP) over the body surface, which are consistent with the courses of the classical meridians, have been internationally elicited by TENS, EA, or electrical heat stimulation of acupoints in many human subjects. Animal studies showed that NO levels are elevated by EA in the acupoints/meridians in rats, associated with an enhanced expression of NO synthase endowed with TRPV1. Dermal microdialysis in humans reported that NO-cGMP releases are increased by EA in the subcutaneous tissue of acupoints. Several studies demonstrated that TENS, MA, and electrical heat produce an elevation of NO release predominantly over acupoints. Consistently, these stimuli, TENS, electrical heat, and EA, have been used to elicit axonal reflexes, which are characterized by local release of NO and CGRP mediated by activation of TRPV1. Interestingly, the same stimuli-, heat, mechanical, and electrical stimuli-evoked axon reflexes, and electrical heat, TENS, and EA-evoked PSCP/LPSCP, induce NO signaling molecules that are involved in the biochemical-physiological processes of axon reflexes and PSCP/LPSCP. These lead us to suggest that similar driver-induced NO-mediated axon reflexes participate in the somatosensory signal transduction of PSCP/LPSCP, the important phenomena of meridians induced by the stimuli. Similarly, all experiments of the tracers along linear pathways resembling meridians have been conducted by injection of a radiotracer solution or tracer dyes in a volume of solution into acupoints. Solutions injected intoacupoints produce much stronger mechanical stimuli, which cause axon reflex, and NOergic biomolecules/neuropeptides increase local blood flow with higher levels in acupoints/meridians,

thus moving radioactive substances or tracer dyes in the skin and subcutaneous tissue along a linear path resembling acupoints and meridians, the important phenomena of meridians induced by the stimuli.

The sun mainly radiates energy in the form of visible light, ultraviolet, and infrared rays, and exposure to sunlight contributes to thermal and ultraviolet effects on human skin, inducing the release of NO, H_2S, and CGRP. It is described in TCM that the meridian systems flow “meridian qi” evoking dynamic processes such as energy exchange, movement, and connection from the surface of the body to the internal organs, as well as communication with the universe, environment, and our body as a whole. The “meridian qi”, which differs from external qi such as air, oxygen, hydrogen, clouds, and earth qi, is the essence of life remaining after a living being absorbs external "qi" into the body and excretes waste gases generated as "internal qi" which possesses a self-regulating mechanism for new metabolic processes, possessing both qi and energy, and circulates through the acupoints, meridians, and collaterals within the body, forming a life field. Although more sophisticated approaches are requested to address gasotransmitter-physiological interactions in the processes in this field, the results from animal and human studies consistently suggest that NO, one of the gasotransmitters, is involved in the stimuli-evoked axon reflexes, and the same driver-elicited NO-mediated axon reflex may also participate in the processes of PSC/LPSC and tracers along linear pathways resembling meridians, the important phenomena of meridians are induced by the same stimuli. The evidence supports the hypotheses that sunlight (heat and NIL) contributes to enhanced gasotransmitters (NO, CO, H_2S) and CGRP levels through the pathways of acupoint/meridians with rich ESG at the luminal side of the microvessel wall; and meridian practices, NIL or heat or stimulation induce gasotransmitters, which mediate chemical-electrical activity for the biophysical characteristics of acupoints/meridians. Together, the stimuli-induced NO signaling molecules and axon reflexes participate in the biochemical-physiological processes of all meridian phenomena, low electrical resistance of acupoints/meridians, the PSC/LPSC, and tracers along linear pathways resembling meridians, leading to the suggestion that NO signaling molecules are important meridian biomolecules and contribute to structure, functions, and development pathways of the meridian system. Anatomical studies have demonstrated that acupoints/meridians exist higher number of nerve fibers/trunks, blood vessels, hair follicles, and sweat glands. The higher number of small nerve fibers, blood vessels with rich ESG at the luminal side of the microvessel wall containing rich endothelia synthases for NO, CO, H_2S, TRPV1, and neuropeptides, and sweat glands in the cutaneous tissues serve as structures for gasotransmitters and neuropeptide-mediated axon reflexes, resulting in low electrical resistance, PSC, and tracers along linear pathways resembling meridians. The evidence supports the hypotheses for future studies to systematically investigate the distribution of gasotransmitters along the pathways of the human body’s fascia network and soft tissues with rich microvessel wall and nerve adventitia, known as the meridian system in TCM.

ACKNOWLEDGMENTS

This project was made possible by NIH Grants (AT002478, AT004620, and AT004504) from the National Center for Complementary and Integrative Medicine (to SXM).

Conflict of Interest

The author has stated explicitly that there are no conflicts of interest in connection with this article.

REFERENCES

1. Beijing College of Traditional Chinese Medicine, Shanghai College of Traditional Chinese Medicine, Nanjing College of Traditional Chinese Medicine, The Acupuncture Institute of the Academy of Traditional Chinese Medicine (eds). Essentials of Chinese Acupuncture. Beijing: Foreign Languages Press; 1980.
2. Wang GJ, Ayati MH, Zhang WB. Meridian studies in China: a systematic review. J Acupunct Meridian Stud 2010; 3:1-9.
3. Zhu ZX, Hao JK. Acupuncture Meridian Biophysics-Scientific Verification of the First Great Invention of China. Beijing: Beijing Press; 1989.
4. Chan SH. What is being stimulated in acupuncture: evaluation of the existence of a specific substrate? Neurosci Biobehav Rev 1984; 8:25-33.
5. Levine AG. 125 questions: exploration and discovery. In: Sanders S (Eds.). Washington, DC: Science/AAAS Custom Publishing; 2021. p. 10.
6. https://www.science.org/do/10.1126/resource.2499249/full/sjtu-booklet-1714066892333.pdf.
7. Shang C: The past, present and future of meridian system research. Clinical Acupuncture and Oriental Medicine 2000; 1: 115-124.
8. Hua P, Lü H, Yuan L, et al. Four main schools of thought and analysis in studies of channels and collaterals. Chin Acupunct Moxibustion. 2006;26:407–13.
9. Benias PC, Wells RG, Sackey-Aboagye B, et al. Structure and Distribution of an Unrecognized Interstitium in Human Tissues. Sci Rep. 2018;8:4947.
10. Langevin HM, Yandow JA. Relationship of acupuncture points and meridians to connective tissue planes. Anat Rec. 2002;269:257–65.
11. Wang B (Englished by Wu NL, Wu AQ): Yellow Empero's canon internal medicine. Beijing: China Science & Technology Press; 1997. Chapter 17 (The Length of the Channels), p. 597.
12. Wang B (Englished by Wu NL, Wu AQ): Yellow Empero's canon internal medicine. Beijing: China Science & Technology Press; 1997. Chapter 3 (on the human vital energy connecting with nature), p. 18-19.
13. van der Rhee HJ, de Vries E, Coebergh JW: Regular sun exposure benefits health. Med Hypotheses. 2016; 97:34-37.
14. Wenfei Liang, Chengming Nie, Jun Du et al: Near-infrared photon upconversion and solar synthesis using lead-free nanocrystals. 2023; Nature Photonics 2023; 17;346–353.
15. Wang R. Gasotransmitters: growing pains and joys. Trends Biochem Sci. 2014; 39:227–232. doi: 10.1016/j.tibs.2014.03.003.
16. Barolet AC, Litvinov IV, Barolet D: Light-induced nitric oxide release in the skin beyond UVA and blue light: Red & near-infrared wavelengths. Nitric Oxide. 2021;117:16-25.
17. Wong BJ, Fieger SM. Transient receptor potential vanilloid type-1 (TRPV1) channels contribute to cutaneous thermal hyperaemia in humans. J Physiol 2010; 588(Pt 21): 4317–26.
18. Hazell G, Khazova M, Cohen H, Felton S, Raj K: Post-exposure persistence of nitric oxide upregulation in skin cells irradiated by UV-A. Sci Rep. 2022;12(1):9465.
19. Hazell G, Khazova M, O'Mahoney P: Low-dose daylight exposure induces nitric oxide release and maintains cell viability in vitro. Sci Rep. 2023;13(1):16306. doi: 10.1038/s41598-023-43653-2.

20. Huang SM, Lee H, Chung MK, Park U, Yu YY, Bradshaw HB, Coulombe PA, Walker JM, Caterina MJ. Overexpressed transient receptor potential vanilloid 3 ion channels in skin keratinocytes modulate pain sensitivity via prostaglandin E2. J Neurosci. 2008;28:13727–13737. doi: 10.1523/JNEUROSCI.5741-07.2008.
21. Cioffi DL. The skinny on TRPV1. Circ Res. 2007;100:934–936. doi: 10.1161/01.RES.0000265139.10277.62.
22. Hu J, Ma SX: Infrared heat treatment reduces food intake and modifies expressions of TRPV3-POMC in the dorsal medulla of Obese Prone rats. Int J Hyperthermia 2011; 27:708-716.
23. Aslami H, Schultz MJ, Juffermans NP: Hydrogen sulfide: a hot molecule. Anesthesiology. 2011;115:921-922.
24. Liu J, Liu Y, Bu W, Bu J, Sun Y, Du J, Shi J. Ultrasensitive nanosensors based on upconversion nanoparticles for selective hypoxia imaging in vivo upon near-infrared excitation. J Am Chem Soc. 2014;136:9701–9709.
25. Paul BD, Snyder SH. H_2S: a novel gasotransmitter that signals by sulfhydration. Trends Biochem Sci. 2015;40(11):687–700. doi: 10.1016/j.tibs.2015.08.007.
26. Wang R. Physiological implications of hydrogen sulfide: a whiff exploration that blossomed. Physiol Rev. 2012;92(2):791–896. doi: 10.1152/physrev.00017.2011
27. Hu L-F, Lu M, Tiong CX, Dawe GS, Hu G, Bian J-S. Neuroprotective effects of hydrogen sulfide on Parkinson's disease rat models. Aging Cell. 2010;9(2):135–146.
28. Wallace JL, Wang R. Hydrogen sulfide-based therapeutics: exploiting a unique but ubiquitous gasotransmitter. Nat Rev Drug Discov. 2015;14(5):329–345. doi: 10.1038/nrd4433.
29. Anonymous (Shanghai College of Medicine), Acupuncture, ShangWu Publisher, Hong Kong, 1975.
30. Luciani RJ: Direct observation and photography of electroconductive points on human skin. Am J Acup 1978; 6:311-317.
31. Wang ZT, Wu SL, Cao YC, Zhu ZX, Xu RM: Morphological study on the low impedance line along channel. Acup Res 1987; 1:82-85.
32. Yen W, Cai B, Yang J, Zhang L, Zeng M, Tarbell JM, Fu BM: Endothelial surface glycocalyx can regulate flow-induced nitric oxide production in microvessels in vivo. PLoS One. 2015 Jan 9;10(1):e0117133. doi: 10.1371/journal.pone.0117133. eCollection 2015.
33. Li J, Qi ZJ, Sun CG et al: Circulating glycocalyx shedding products as biomarkers for evaluating prognosis of patients with out-of-hospital cardiac arrest after return of spontaneous circulation. Sci Rep. Article number: 17582 (2024)
34. Sugimoto K, Fujii S, Takemasa T, Yamashita K: Detection of intracellular nitric oxide using a combination of aldehyde fixatives with 4,5-diaminofluorescein diacetate. Histochem Cell Biol. 2000 May;113(5):341-7. doi: 10.1007/s004180000151.
35. Allen CL, Wolanska K, Malhi NK, Andrew V Benest AV: Hydrogen sulfide is a novel protector of the retinal glycocalyx and endothelial permeability barrier. Front Cell Dev Biol 2021;9:724905. doi: 10.3389/fcell.2021.724905.
36. Yen W, Cai B, Yang J, Zhang L, Zeng M, Tarbell JM, Fu BM: Endothelial surface glycocalyx can regulate flow-induced nitric oxide production in microvessels in vivo. PLoS One. 2015 Jan 9;10(1):e0117133. doi: 10.1371/journal.pone.0117133. eCollection 2015.
37. Li J, Qi ZJ, Sun CG et al: Circulating glycocalyx shedding products as biomarkers for evaluating prognosis of patients with out-of-hospital cardiac arrest after return of spontaneous circulation. Scientific Reports volume 14, Article number: 17582 (2024).
38. Sugimoto K, Fujii S, Takemasa T, Yamashita K: Detection of intracellular nitric oxide using a combination of aldehyde fixatives with 4,5-diaminofluorescein diacetate. Histochem Cell Biol. 2000 May;113(5):341-7. doi: 10.1007/s004180000151.

39. Allen CL, Wolanska K, Malhi NK, Andrew V Benest AV: Hydrogen sulfide is a novel protector of the retinal glycocalyx and endothelial permeability barrier. Front Cell Dev Biol 2021;9:724905. doi: 10.3389/fcell.2021.724905
40. Li H, Yang C, Yin Y, Wang F, Chen M, Xu L, Wang N, Zhang D, Wang X, Kong Y, Li Q, Su S, Cao Y, Liu W, Ao Z, Dai L, Ma C, Shang L, Han D, Ji F, Li H: An extravascular fluid transport system based on structural framework of fibrous connective tissues in human body.Cell Prolif. 2019 Sep;52(5):e12667. doi: 10.1111/cpr.12667.
41. Wang Dayou (Shaohua Zi): The Holographic Law of the Universe, China Medical Science and Technology Press. 2002 (in Chinese).
42. Yang Li: Zhou Yi and Traditional Chinese Medicine, pp. 618-667, Beijing Science and Technology Press. Third edition, 2006 (in Chinese).
43. Ma SX. Enhanced nitric oxide concentrations and expression of nitric oxide synthase in acupuncture points/meridians. J Altern Complement Med 2003; 9: 207-215.
44. Ma S.X., Li X.Y., Sakurai T, Pandjaitan M. Evidence of enhanced non-enzymatic nitric oxide generation on the skin surface of acupuncture points: An innovative approach in humans. Nitric Oxide: Biol Chem 2007; 17:60-68.
45. Ma SX, Lee P, Li XY, Jiang I, Ma E, Hu J. Influence of age, gender, and race on nitric oxide release over acupuncture points-meridians. Sci Rep 2015 Dec 1:5:17547. doi: 10.1038/srep17547.
46. Lim N, Ma SX. Responses of nitric oxide-cGMP releases in acupuncture point to electroacupuncture in human skin in vivo using dermal microdialysis. Microcirculation 2009; 16: 434-443.
47. Ha Y, Kim M, Nah J, Suh M, Lee Y. Measurements of location-dependent nitric oxide levels on skin surface in relation to acupuncture point. Evid Based Complement. Alternat Med 2012:781460. doi:10.1155/2012/781460.
48. Chen, JX, Ma SX. Effects of nitric oxide and noradrenergic function on skin electric resistance of acupoints and meridians. J. Altern. Complement. Med. 2005, 11, 423-431.
49. Chen JX, Ibe BO, Ma SX. Nitric oxide on modulation of norepinephrine production in acupuncture points. Life Sci. 2006, 79, 2157-2164.
50. Ma SX. Low electrical resistance properties of acupoints: Roles of NOergic signaling molecules and neuropeptides in skin electrical conductance. Chin J Integr Med 2021; 27:563-569.
51. Ma SX, Mayer E, Lee P, Li XY, Gao EZ. Transcutaneous electrical nerve stimulation increased nitric oxide-cyclic cGMP release biocaptured over skin surface of the pericardium meridian and acupuncture points in humans. Acup & Electro-Therapeutics Res INT J 2015; 40:73-86.
52. Ma SX, Lee P, Anderson TL, Li XY, Jiang I. Response of local nitric oxide release to manual acupuncture and electrical heat in humans: Effects of reinforcement methods. Evid Based Complement Alternat Med 2017:2017:4694238.doi: 10.1155/2017/4694238.
53. Dippel E, Mayer B, Schonfelder G, et al. Distribution of constitutive nitric oxide synthase immunoreactivity and NADPH-diaphorase activity in murine telogen and anagen skin. J Invest Dermatology 1994； 103：112-115.
54. Ji B, Hu J, Ma SX. Effects of electroacupuncture Zusanli (ST36) on food intake and expression of POMC and TRPV1 through afferents-medulla. pathway in Obese Prone Rats. Peptides 2013； 39:188-194.
55. Xiong K, Li H, Wang T. Origin of nitric oxide synthase positive nerve fibers at zusanli area in rats. Chin J Integ Tradit West Med (Chin) 1998; 18: 230-232.
56. Hodges GJ, Kosiba WA, Zhao K, Johnson JM. The involvement of norepinephrine, neuropeptide Y, and nitric oxide in the cutaneous vasodilator response to local heating in humans. J Appl Physiol 2008; 105: 233–240.
57. Houghton BL, Meendering JR, Wong BJ, Minson CT. Nitric oxide and noradrenaline contribute to the temperature threshold of the axon reflex response to gradual local heating in human skin. J Physiol 2006; 572(Pt 3): 811–20.

58. Shinbara H, Okubo M, Kimura K, Mizunuma K, Sumiya E, Shinbara H, et al. Participation of calcitonin gene related peptide released via axon reflex in the local increase in muscle blood flow following manual acupuncture. Acupunct Med 2013; 31(1): 81–7.
59. Ma, S.X.: Stimuli-evoked NOergic molecules and neuropeptides at acupuncture points and gracile nucleus contribute to signal transduction of propagated sensation along the meridian through the dorsal medulla-thalamic pathways. J Integ Med. 2024. https://doi.org/10.1016/j.joim.2024.07.001
60. Ibrahim TS, Chen ML, Ma SX. TRPV1 expression in acupuncture points: Response to electroacupuncture stimulation. J Chem Neuroanat 2011, 41:129-136.
61. Rong PJ, Ma SX: Electroacupuncture Zusanli (ST36) on release of nitric oxide in the gracile nucleus and improvement of sensory neuropathies in Zucker Diabetic Fatty rats. Evid. Based Complement. Alternat Med 2011; 2011:134545.
62. Ma SX, Ma J, Moise G, Li XY. Responses of neuronal nitric oxide synthase expression in the brainstem to electroacupuncture Zusanli (ST 36) in rats. Brain Res 2005; 10:70-77.
63. Kovacs F.M., Gotzens V., García A., García, F., Mufraggi, N., Prandi, D., Setoain, J., Román, F.S. Radioactive pathways of hypodermically injected technetium-99m, Am. J. Chin. Med. 1996, 24,101-102.
64. Kovacs FM, Garcia A, Mufraggi N, et al.: Migration pathways of hypodermically injected technetium-99m in dogs. Eur. Radiol. 2000, 10, 1019-1025.
65. Wu CC, Chen MF, Lin CC: Absorption of subcutaneous injection of Tc-99m pertechnetate via acupuncture points and non-acupuncture points. Am J Chin Med. 1994; 22:111-118.
66. Li T, Tang BQ, Zhang WB, Zha MY, Hu QC, Ahn A: In-vivo visualization of pericardium meridian with fluorescent dyes. Evid. Based Complement Alternat. Med. 2021; doi: 10.1155/2021/5581227. eCollection 2021.
67. Ma SX: Stimuli-induced NOergic molecules and neuropeptides mediated axon reflexes contribute to tracers along meridian pathways. Curr Top Med Chem. 24: 393–400, 2024.

Correspondence Author

Sheng-Xing Ma, M.D., Ph.D.
Professor, Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center
1124 W. Carson Street, E4-room 28C
Torrance, CA 90502
Tel: (310) 974-9573; E-mail: sxma@lundquist.org

传染病博弈论模型

Game Theory Models for Infectious Diseases

黄胜和，周文生，阿姆勃劳斯.琼 / Sheng-He Huang[1]，Wensheng Zhou[2]，Ambrose Jong[1]

1. *Childrens Hospital Los Angeles, University of Southern California, Los Angeles, CA,*
2. *HRL Laboratories, Malibu, CA, Corresponding author: shhuang@hsc.usc.edu*

ABSTRACT

Extending along the dynamic continuum from conflict to cooperation, microbial infections always involve symbiosis (Sym) and pathogenesis (Pat). There exists a dynamic Sym-Pat duality (DSPD) in microbial infection that is the most fundamental problem in infectomics. DSPD is encoded by the genomes of both the microbes and their hosts. Three focal point (FP) theory-based game models (pure cooperative, dilemma and pure conflict) are proposed for resolving those problems. Our health is associated with the dynamic interactions of three microbial communities [nonpathogenic microbiota (NP) (Cooperation), conditional pathogens (CP) (Dilemma), and unconditional pathogens (UP) (Conflict)] with the hosts at different health statuses. Sym and Pat can be quantitated by measuring symbiotic index (SI), which is quantitative fitness for the symbiotic partnership, and pathogenic index (PI), which is quantitative damage to the symbiotic partnership, respectively. Symbiotic point (SP), which bears analogy to FP, is a function of SI and PI. SP-converting and specific pathogen-targeting strategies can be used for the rational control of microbial infections. (This paper was published in J Biomed Biotechnol. 2008 Mar 2; 2008:856314; PMID: 18350122).

INTRODUCTION

Infectious diseases caused by bacterial, viral, fungal or parasitic pathogens continue to be the leading cause of morbidity and mortality worldwide despite the availability of effective anti-microbial agents and vaccines over the last fifty years [1]. The continual emergence of previously undescribed new pathogens, reemergence of old pathogens, and the rising crisis of antibiotics resistance will certainly heighten the global impact of microbial infections in the 21st century. These problems are mainly due to inadequate knowledge of the dynamic duality relationships between symbiosis (Sym) and pathogenesis (Pat) in microbial infections [2]. The term symbiosis, which may have many variations on its definition, in this paper refers to living together through a close and prolonged association between two or more organisms of different species [3, 4]. Duality is defined as different ways of looking at the same thing [5]. There are two major limitations inherent in the conventional theories of microbial infection. On the one hand, in the past century biology and medicine including infectious diseases have been dominated by the reductionistic approaches. Focusing research on individual virulence genes and the important pathogens has been the traditional approach to human infectious diseases. On the other hand, as Joshua Lederberg pointed out [6, 7], medical science is imbued with the Manichaean view of the microbe–human host relationship: “we good; they evil.” Almost all broad-spectrumantimicrobial agents, which are in

the best interest of pharmaceutical industries, kill both the good microbes as well as the bad germs. Even though narrow-spectrum antiinfective agents are not “narrow” for pathogens, they also target both the good and bad microorganisms with a limited range of species.

Animals and plants are continually infected by an extensive diversity of symbiotic or invading organisms including bacteria, virus, fungus or parasites. Infection of bacteria by phages started long before the emergence of animals and plants [8]. Microbial infection is an evolutionary paradigm which is associated with co-evolution between hosts and microbes [6, 7, 9]. This co-evolution can be defined as the process of reciprocal and dynamic genetic changes in two or more species [2]. The conventional wisdom in medicine holds that microbial infection is a pathogenic process in which a pathogen enters, establishes itself and multiplies in the host [10]. The emphasis is on the antagonism or conflict, not the mutualism. This represents “zero-sum thinking” – the belief that if one player gains, other player must inevitably lose. Methods and concepts of the zero-sum game theory have proved successful in studying the strategy of pure conflict. The most challenging issue in infectious diseases is how to dissect the dynamic Sym-Pat duality in microbial infections using infectomics and mathematics such as focal point-based game theory. Game theory, defined in the broadest sense, is the study of the strategies of conflict, cooperation and mixed situations in which both coexist. Generally there are multiple equilibria in a game. Thomas Schelling’s concept of focal point, which is an equilibrium usually standing out from the others, addressed the crucial question of how to interpret the multiplicity of equilibria [11, 12]. This article attempts to enlarge the scope and application of focal point game theory in microbial infections, extending from the zero-sum games to the nonzero-sum games.

DEFINITIONS AND METHODS

Three Community Principle of Microbial Infections

Our health is associated with the dynamic interactions of three microbial communities [2] [nonpathogenic microbiota (NP), conditional pathogens (CP), and unconditional pathogens (UP)] with the hosts at three different health statuses [nonsusceptibility (NS), conditional susceptibility (CS), and unconditional susceptibility (US)] (Fig.1). NP is the major microbial community which forms a healthy symbiotic ‘superorganism’ with the hosts. The ecology and evolution of NP-NS interaction is essential and fundamental for health. From birth to death, we share a benign coexistence with a vast, complex, and dynamic consortium of microbes. Most of our microbial commensals reside in our gastrointestinal (GI) track packed with up to 100 trillion (1014) microbes [1, 13]. The GI tract harbors a rich microbiota of >600 different bacterial species. Some of these microorganisms have important health functions. These include stimulating the immune system, protecting the host from invading bacteria and viruses, and aiding digestion. The gut microbiota, which is essential for human homeostasis, is established rapidly after birth and remains relatively stable throughout the life [1]. The GI mucosa provides a protective interface between the internal environment and the constant external challenge from food-derived antigens and microbes. CP and UP are minor microbial communities that mainly contribute to the pathogenesis of microbial diseases. The distinction between the commensal and the pathogen in the CP community can be blurred because they may cause diseases under certain sub-health conditions of the hosts, or in immunocompromised hosts. For example, pneumococcus, meningococcus and Haemophilus bacteria regularly exist as part of the normal microbiota of the host respiratory track and are mostly carried asymptomatically despite the fact that they can cause well-defined diseases [14, 15]. Microbes in the CP community dynamically evolve in two opposite directions, which are toward either the NP (more cooperative or mutualistic) or UP (more pathogenic) microbial community. Microbes with

high pathogenicity belong to the UP microbial community. The three microbial communities and three statuses of the hosts are subjected to dynamic reciprocal changes driven by transfer of genetic materials.

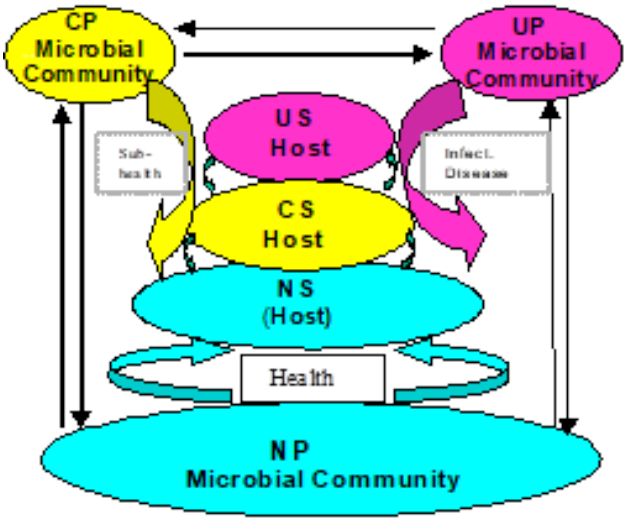


Figure.1 Schematic representation of interactions of three microbial communities [non-pathogenic (NP), conditional pathogenic (CP) and unconditional pathogenic (UP)] with with the hosts at three different health statuses [non-susceptibility (NS), conditional susceptibility (CS), and unconditional susceptibility (US)].

Dynamic Duality Relationships between Sym and Pat in Microbial Infections

Extending along the dynamic continuum from conflict to cooperation, microbial infections always involve symbiosis and pathogenesis, which are two fundamental components of the host-microbe interactions (Fig.2). There exists a dynamic Sym-Pat duality (DSPD) in microbial infection, which is the most fundamental issue of infectomics [2]. DSPD is reflected in the genotypic and phenotypic infectomes, which are encoded by the genomes of both the microbes and their hosts [2]. The opposition and unity of Sym and Pat are indispensable, and the academic viewpoint that the unity of opposites of Sym and Pat gives impetus to the development of microbial infection is considered as the core idea and radical principle of the duality representations of microbial infections. In certain circumstances and at a certain stage of the development of microbial infection, each of the two aspects of Sym and Pat will transform from antagonism into mutualism or from mutualism into antagonism. Sym and Pat can be quantitated by measuring symbiotic index (SI), which is quantitative fitness for the symbiotic partnership, and pathogenic index (PI), which is quantitative damage to the symbiotic partnership, respectively. The most crucial studies are to identify infectomic signatures specific for SI and PI. The set of symbiotic or pathogenic parameters is defined as a function SI(x) or PI(x*). SI(x) and PI(x*) are continuous functions ranging from 0 to 1 to admit different degrees of Sym and Pat, respectively. SI(x)=0 and PI(x*)=0 indicate that x and x* are perceived to be zero-symbiotic and zero-pathogenic, respectively. SI(x)=1 and PI(x*) =1 indicate that x and x* are perceived to be completely symbiotic and completely pathogenic, respectively. Intermediate values of SI(x) and PI(x*) indicate that x and x* are perceived to be partially symbiotic and partially pathogenic, respectively. Symbiotic points are used to determine the dynamic duality between Sym and Pat. SI and PI are interdependent parameters. The symbiotic point (SP) is a function of SI and PI.

$$SP= f (SI, PI)$$

The focus of the dynamic duality research is to examine the ability of SP to transform situations of potential conflict (UP-US & CP-CS) into situations of cooperation (NP-NS). SP bears analogy to Schelling's focal point, which is any feature of such a game that provides a focus of convergence [16]. In the games with multiple Nash equilibria, one equilibrium usually stands out from the others (salient). Such an equilibrium is a focalpoint which can be easily recognized by all the players [12]. Thomas Schelling's Strategy of Conflict (1960) has been recognized as one of the most important works of game theory [11, 17]. There is no doubt that focal points play a central role in Schelling's game theory. Schelling has made a significant contribution to a reorientation of game theory. Understanding focal points is not only a key to improving game theory but also a key to dissecting SPs.

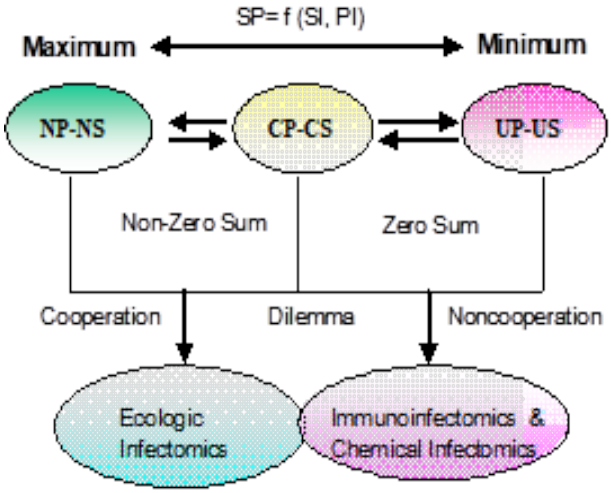


Figure 2. A continuum model of host-microbe interactions coupling with infectomic approaches to dissect the problems in microbial infections.

Game Theoretical Models (GTMs) of Microbial Infections

In this paper, three types of GTMs are proposed for studies on NP-NS interactions (cooperative game), UP-US interactions (noncooperative games), and CP-CS interactions (dilemma or bargaining game). First, the NS-NS interactions are dissected with pure cooperative games in which each player chooses the strategy corresponding with the focal point in the expectation that the others will do the same. The significance of focal points can be shown most clearly in the pure cooperative games. As there is no conflict of interests in these games, all the players merely want to cooperate and they do not choose the alternative ways. Analysis of the cooperative game issues is to focus on coalition formation and distribution of the gains through cooperation. The SP in the NP-NS games tends to be maximal (Fig.2). Secondly, noncooperative GTMs are used for analysis of the UP-US interactions. In contrast to cooperative games which focus on collective rationality and common interest, noncooperative games emphasize individual rationality and individual optimal strategy. The SP in the UP-US games tends to be minimal (Fig.2). In games of pure conflict, defection is the equilibrium strategy and the total benefit to all players in the game, for every combination of strategies, always adds to zero (zero-sum). In the antagonistic UP-US interaction model, the surviving strategies of the UP community conflict with that of the US host. The UP evolves to exploit the host as much as possible, and the host adapts to exclude or limit the damage caused by the UP. Thirdly, we consider the strategic use of focal point theory in mixed situations to analyze the CP-CS interactions in which there is both conflict and mutual dependence. The most well-known example is the Prisoner's Dilemma game (a two player game) in which each player chooses between a cooperating and defecting strategy. In this game, each player receives a higher playoff by defecting than by cooperating. However, a higher playoff is received if both

cooperate than both defect. The two player game can be extended to the N-player Prisoner's Dilemma game with arbitrary numbers of players.

RESULTS

Three communities in Escherichia coli species

E. coli is one of the best understood and most thoroughly studied organisms and is advantageous as a model microorganism for the current studies. This bacterium is genotypically and phenotypically a highly diverse species, which is present in the three microbial communities (Table 1). Most E. coli strains are commensals of higher vertebrates belonging to NP, but some are pathogenic (CP and UP). Uropathogenic E. coli (UPEC) in the CP group are the most common cause of community-acquired urinary tract infection (UTI). UPEC are responsible for about 80% of the estimated 150 million UTIs diagnosed annually [18]. E. coli O157, which belong to the UP group, is a major food pathogen. Shigella species, the cause of dysentery, are now known to be multiple distinct lineages of E. coli. Genomes of those E. coli strains have been sequenced (Table 1). Recently, 'better' E. coli strains (MDS41, 42, 43) have been engineered in which about 15% of the genome has been removed with the use of synthetic biology [19]. Coliphage, a virus which infects E. coli, is a major contributor responsible for diversification of E. coli [20, 21]. From a population-dynamic view, the interactions between coliphage and E. coli are analogous to those of a predator and a prey.

Table 1. *E. coli* in the three microbial communities

	E. coli strains	characteristics	Genome (Mb)	Putative	
				SI	PI
NP	MG1655 (K12)	Commensal	4.6	> 0.75	< 0.25
	Nissle 1917	Probiotics	5.1	> 0.75	< 0.25
	A0 34/86	Probiotics	?	> 0.75	< 0.25
	MDS41, 42, 43	K12 strains	3.9	> 0.75	< 0.25
CP	RS218	Low pathogenicity	5.1	0.5±0.25	0.5±0.25
	CFT073	Uropathogenicity	5.2	0.5±0.25	0.5±0.25
UP	O157 RIMD	High pathogenicity	5.5 5.5	< 0.25	> 0.75
	O157 EDL	High pathogenicity	4.4	< 0.25	> 0.75
	Shigella Sd197	High pathogenicity		< 0.25	> 0.75

Sym-Pat duality is encoded by the genomes of both microbes and their hosts

The development of microbial infections depends on the dynamic Sym-Pat duality, which is governed by the genomes of both the microbes and their hosts [2]. There are three types of genetic/genomic determinants (GGDs) that may contribute to the Sym-Pat duality of microbial infections (Fig.3). The first type is the Sym GGD pool, which contributes to symbiosis. The Sym GGDs from the microbial partner include symbiosis-related genomic islands (SYI), plasmids, transposons, and microbiome, which is a collective genome of microbiota [2, 22, 23]. The gene pool contributing to immune tolerance belongs to the host Sym GGDs [24]. The symbiotic homeostasis of the superorganism formed by the microbiota and its host is governed by hologenome, a complex of the host genome and microbiome [22, 25]. The second pool of GGDs contributes to the pathogenesis of microbial infections. These include pathogenicity islands (PAI), virulence plasmids, pathogen-associated molecular patterns (PAMPs) and endogenous alarmins [26]. PAMPs are a diverse group of microbial molecules, which are recognized by the host innate and adaptive immune system, primarily through toll-like receptors (TLRs) [26]. Alarmins are endogenous molecules within the host that signal tissue and cell damage. Effector cells of the innate and adaptive immunity can release alarmins when they are activated by PAMPs. Endogenous alarmins and exogenous PAMPs elicit similar responses by conveying a similar message. Therefore,

they constitute a larger family of damage-associated molecular patterns (DAMPs). The third type of GGD pool has dual functions that depend on external and internal environments. These include ecological islands (ECI), certain GGDs from conditionalpathogens [such as CP factors (CPFs) and CP islands (CPIs)], and the host GGDs with dual effects on microbial infections. A comparative infectmic study suggests that GimA, a 20-kb genomic island, is a typical CPI [27]. The dual biological functions of GimA depend on the genomic environments in E. coli strains. GimA present in meningitic E. coli K1 genome (O18:K1:H7) is essential for bacterial crossing the blood-brain barrier to cause meningitis [28]. In contrast, GimA is required for the probiotic function of E. coli K24 strain A0 34/86 (O83:K24:H31), which has been safely and effectively used in Czech pediatric clinics since 1967 [29]. The dual Sym-Pat properties of microbial determinants were also observed in photorhabdus, which is a genus of Gram-negative bacteria mutualistically associated with entomophagous nematodes of the family heterorhabditiae [30]. A hexA homologous gene from photorhabdus is able to regulate both symbiosis and pathogenesis [30]. Some microbes exhibit dual behavior as symbionts and pathogens in a manner dependent on the hosts. Sooty mangabey (SM) monkeys infected with simian immunodeficiency virus (SIV) do not develop acquired immunodeficiency syndrome (AIDS) [31]. In contrast, SIV infection of non-natural host monkeys, such as rhesus macaques (RM), causes AIDS that closely resembles the human disease [31]. Similarly, polydnaviridae, a family of double-stranded DNA viruses, have evolved complex life cycles in which they interact as symbionts with one host and pathogens with another [32].

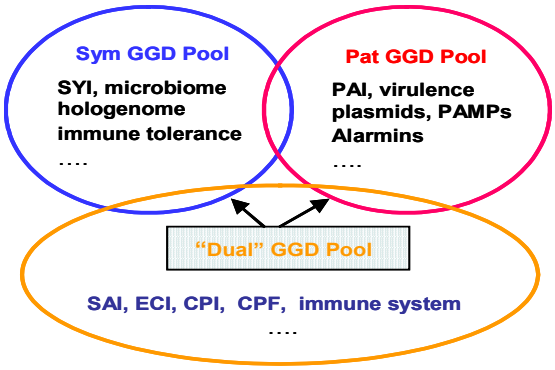


Figure 3. Gene pools contributing to the Sym-Pat duality. CPF, CP factors; CPI, CP islands; ECI, ecological islands; HPI, high pathogenicity islands; PAI, pathogenicity islands; SYI, symbiosis island.

Duality relationship between Sym and Pat

Sym and Pat, the fundamental components of microbial infection, can be defined as a function SI(x) or PI(x*). SI(x) and PI(x*) are continuous functions ranging from 0 to 1 to admit different degrees of Sym and Pat, respectively (Fig.4). SI(x) = 0 and PI(x*) = 0 indicate that x and x* are perceived to be zero-symbiotic and zero-pathogenic, respectively. SI(x)=1 and PI(x*) =1 indicate that x and x* are perceived to be completely symbiotic and completely pathogenic, respectively. If SI is close to or equal to 1, microbial infection is a physiological process. It is now well accepted that mitochondria were derived from an endosymbiotic relationship with internalized proteobacteria, via a progressive transfer of genetic material [33]. This long symbiotic relationship reaches the maximum Sym value. The symbiotic nitrogen fixation process for converting atmospheric dinitrogen (N2) to ammonia (NH3) is essentially dependent on two partners, the host legume plant and bacteria belonging to the family Rhizobiaceae. [34]. This type of microbial infection is more typically associated with a physiological process.

If the PI value approaches or reaches the maximum limit, microbial infection is more completely associated with a pathogenic process. Most of serious infectious diseases fall into this category. Intermediate values of SI(x) and PI(x*) indicate that x and x* are perceived to be partially symbiotic and partially pathogenic, respectively. Microbial infections induced by conditional pathogens represent a competitive relationship (II or III in Fig.4). The hosts have large influences on SI and PI. For

example, polydnaviruses have evolved complex life cycles in which they are able to adapt to a mutualistic partnership with one host and become pathogens with another. Their genomes reflect the dual roles as mutualists and pathogens [32].

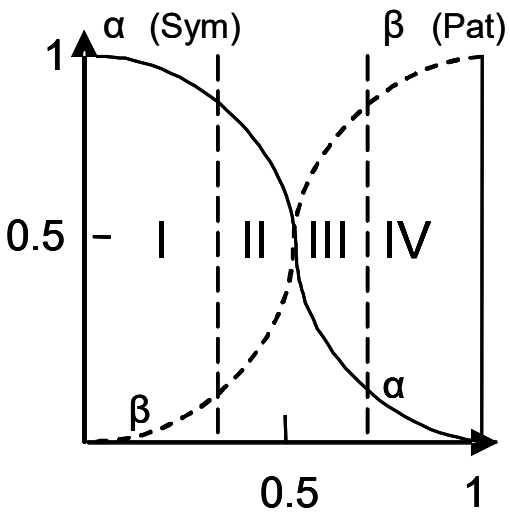


Figure 4. The duality relationship between Sym (αi) and Pat (βi). A cooperative relationship (NP-NS: health and mutualism) (I) occurs between the host and the NP microbial community. A competitive relationship (CP-CS) (II and III) exists between the host and the CP microbial community. There are two types of competitions, better (II) and worse (III). An antagonistic relationship (UP-US) (IV) between the host and the UP microbial community.

Outcomes of three types of games in microbial infections

The outcome of a game is not only determined by one individual's choices, but also depends on the strategies used by all the others. The dynamic Sym-Pat duality influenced by both microbes and their hosts is the key factor that determines the outcome of a game associated with microbial infection. One of the most fundamental issues in game theoretical solution concepts is that strategies used by individual players are based on the differences in payoff perceived by them [35]. This issue can be solved by Schelling's focal point (FP) theory. FPs constitute shared expectations that coordinate the activities of diverse players collectively or independently seeking their goals [36]. By harmonizing anticipated behaviors or responses despite the presence of imperfect information, individuals are able to coordinate their activities towards their ends. In 1960 Schelling classified games into three major categories: pure cooperative game on one side, pure conflict games on the other and combinations of partial cooperation/partial conflict games in between [11]. Recently, a similar classification was designated in Gao's book, "Principles of Systemics" [37]. The pure cooperative game models are used for dissecting NP-NS interaction problems. The payoff matrix for a pure NP-NS problem would resemble something like that in Fig.5a. In contrast, the pure UP-US interaction, a situation of pure antagonism, is characterized by completely opposing interests, where the pathogenesis of microbial infection is the most predominant event. The payoff matrix for the pure UP-US game would look something like that in Fig.5b. The state of depiction comes from the Manichaean view of the microbe–human host relationship. This situation is depicted in Fig.5b as a two-player game. In the pure conflict game, the relationships between the microbes and their hosts end up in a "war of all against all" in which the payoffs of the outcomes add to zero (zero-sum). The three microbial community principle and the dynamic Sym-Pat duality concept would help establish a holistic view of microbial infections. The CP-CS interaction is a situation where cooperation and conflict coexist. As illustrated in Fig.5c, thepayoffs for the CP-CS games would lie midway between the pure cooperative and pure conflict games. The CP-CS problems are microbial dilemmas, in which there is a mixture of mutual

dependence and conflict, of partnerships and competition. The underlying idea arises naturally from the well-known games for the social dilemmas and the Prisoner's Dilemma (PD), in which each player chooses between a cooperating and a defecting strategy [38]. As shown in Fig.5c, each player receives a higher playoff by defecting than by cooperating, no matter what the other player chooses. However, they receive a higher playoff if both cooperate than both defect. The CP-CS interactions can co-evolve toward two different directions, increasing (more cooperation) or decreasing (more antagonism) the SP (Fig.2).

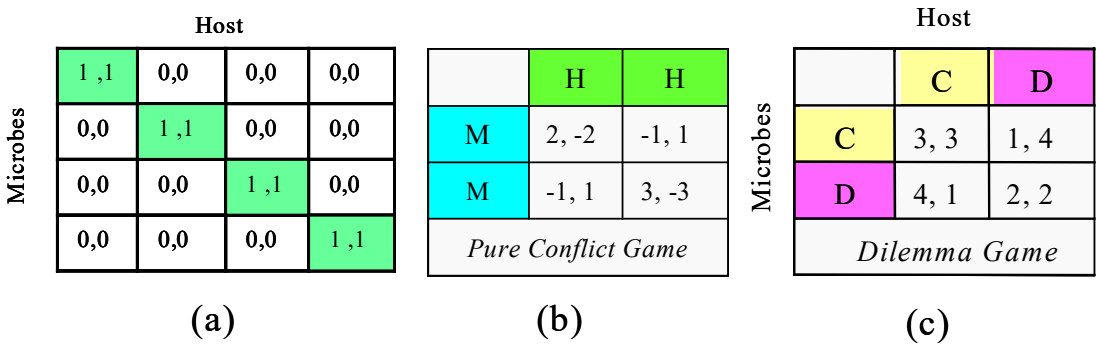


Figure 5. (a) Pure cooperative game. (b) Pure conflict game. H: host; M: microbes. (c) Dilemma game. The number in the left of each pair indicates the payoff for Microbes; the right, Host. Higher numbers represent greater payoff for the individual. Two strategies [Cooperation (C) and Defection (D)] are used.

DISCUSSION

In this paper, focal point theory-based game models are proposed for analysis of the dynamic Sym-Pat duality (DSPD) in microbial infections. DSPD is the most fundamental problem in infectomics, which is the integration of omics and mathematical/computational approaches. There are three types of infectomic approaches that can be used for the control of microbial infections: ecological infectoimcs, immunoinfectomics and chemoinfectomics [2]. Ecological infectomics will explore symbiotic solutions to microbial infections. Developing novel immunological intervention strategies for the prevention and treatment of microbial infections using infectomic signatures and immunomic approaches falls within the field of immunoinfectomics. Chemoinfectomics represents the most powerful approach to the development of a new generation of drugs for antimicrobial chemotherapy.

Symbiosis Point Converting (SPC): Ecological infectomics-based approaches for rational control of microbial infections.

As microbial infection is an ecological and evolutionary paradigm which is associated with co-evolution between hosts and microbes (such as human host and microorganisms, phages and bacteria) in dynamic ecosystems, two ecological infectomics-based SPC approaches (increasing and decreasing SP) can be used for rational control of infectious diseases [2]. The focus in SP increasing approaches is how to transform situations of potential conflict (pathogenesis) into cooperation (symbiosis) by dissecting the dynamic duality relationships between Sym and Pat in microbial infections and developing symbiotic agents (symbiotics) that favor a healthy symbiosis [2]. Symbiotics are defined as products that are beneficial to symbiotic ecology of the super-organisms consisting of microbes and their human hosts. These include microbial (e.g., probiotic bacteria) and nonmicrobial agents (e.g., prebiotics) [2]. The introduction of beneficial symbiotics with higher SP in our body should be a very attractive rationale for modulating the microbiota, improving the symbiotic homeostasis of the superorganism and providing a microbial stimulus to the host immune system against pathogens. Decreasing SP is another rational strategy for control of microbial infections. As phages, which specifically kill bacteria, play an important role in the ecology, evolution and virulence

of a number of pathogens, there is a rational use of phages for treatment and prevention of bacterial infections. The use of phages to treat bacterial infections has a long history dating back to mid 1910's [2]. Due to the availability of effective broad spectrum antibiotics in the early 1940's, phage therapy was discarded in Western medicine at that time. The rising crisis of antibiotic resistance has recently increased great interest in phages and their use as natural antimicrobial agents to fight microbial infections [2]. Compared with commonly used antibiotics, a great advantage of phages is their narrow host range. Recent studies showed that co-infection with GB virus C (GBV-C) is associated with a decreased mortality in HIV-infected patients [39]. Therefore, reducing SP between microbial agents (such as phages and GBV-C) and targeted pathogens is another excellent ecological approach for the development of novel antimicrobial agents.

Specific Pathogen-Targeting (SPT): Immunoinfectomics- and chemoinfectomics-based approaches for prevention and treatment of infectious diseases

In contrast to the ecological infectomics-based SPC approaches that focus on the symbiotic relationships (such as NP-NS and CP-CS interactions) between the hosts and microbial communities, immunoinfectomics- and chemoinfectomics-based SPT approaches emphasize the use of antagonistic relationships (such as UP-US interactions) between the hosts and microorganisms. It is important to point out that the SPT approaches are intrinsically different from the conventional pathogen-targeting antimicrobial agents, which kill both pathogens and nonpathogens [2]. The availability of the genomic information from both microbes and their hosts has resulted in exciting new progress in the field of immunoinfectomics. Nanobody-based technologies and immune epitope mapping have emerged as the very powerful tools for the discovery and development of novel antimicrobial agents. Concurrent advances in both high-throughput chemistry and infectomics have given rise to the field of chemoinfectomics for elucidating and validating drug targets, and generating novel therapeutics. Chemoinfectomics refers to the use of small synthetic molecules that are highly specific for defined infectomic targets, for biological function analysis and to discover new drug leads. The progress towards understanding the dynamic Sym-Pat duality in microbial infections using focal point theory-based game models will greatly facilitate the use of ecological infectomics, immunoinfectomics and chemoinfectomics for the rational control of infectious diseases.

REFERENCES

[1]. S.H. Huang, T. Triche and A.Y. Jong, "Infectomics: genomics and proteomics of microbial infections," Functional & Integrative Genomics. Vol.1, pp.331-44, 2002.

[2]. S.H. Huang, X. Wang and A. Jong, "The evolving role of infectomics in drug discovery," Expert Opinion on Drug Discovery, Vol. 2, pp.961-975, 2007.

[3]. M.J. Roossinck, "Symbiosis versus competition in plant virus evolution," Nature Review Microbiology, Vol. 3, pp.917-24, 2005.

[4]. G.G. Dimijian, "Evolving together: the biology of symbiosis," Baylor University Medical Center Proceedings, Vol. 13, pp.217-26, 2000.

[5]. S. Lee, "Three roads to Quantum Gravity," Basic Books, 2001, page 113.

[6]. J. Lederberg, "Infectious history," Science, Vol. 288, pp.287-93, 2000.

[7]. J. Lederberg. "The future of infectious diseases," Journal of Urban Health. Vol.75, pp.463-70, 1998.

[8]. H. Brussow, C. Canchaya, W.D. Hardt, "Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion," Microbiology and Molecular Biology Reviews, Vol. 68, pp.560-602, 2004.

[9]. M.E. Woolhouse, J.P. Webster, E. Domingo, B. Charlesworth, B.R. Levin, “Biological and biomedicalimplications of the co-evolution of pathogens and their hosts,” Nature Genetics, Vol.32, pp.569-77, 2002.

[10]. A. Casadevall and L.A. Pirofski, “Host-pathogen interactions: basic concepts of microbial commensalism, colonization, infection, and disease,” Infection and Immunity, Vol. 68, pp.6511-8, 2000.

[11]. T. C. Schelling, The Strategy of Conflict, Harvard University, Cambridge, Massachusetts, pp. 83-118, 1960.

[12]. J. Metha, C. Starmer, R. Sugden, ”Focal points in pure coordination games: an experimental investigation,” Theory and Decision, Vol. 36, pp.163-185, 1994.

[13]. R.E. Ley, D.A. Peterson, J.I. Gordon, “Ecological and evolutionary forces shaping microbial diversity in the human intestine,” Cell. Vol.124, pp.837-48, 2006.

[14]. S. Falkow, “Is persistent bacterial infection good for your health?” Cell, Vol.124, pp.699-702, 2006.

[15]. D. Kuklinska, M. Kilian, “Relative proportions of Haemophilus species in the throat of healthy children and adults,” European Journal of Clinical Microbiology, Vol.3, pp.249-52, 1984.

[16]. K. Benmore, L. Samuelson, “The evolution of focal points,” Games and Economic Behavior, Vol.55, pp.21-42, 2006

[17]. R. Sugden, I.E. Zamarron, “Finding the key: the riddle of focal points,” Journal of Economic Psychology, Vol. 27, pp.609-621, 2006.

[18]. R.A. Welch, V. Burland , G. Plunkett 3rd, et. al. “Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli,” Proceedings of the National Academy of Sciences of the United States of America, Vol.17020-4, 2002.

[19]. G. Pósfai, G. Plunkett 3rd, T. Fehér T, et. al., Emergent properties of reduced-genome Escherichia coli. Science. Vol.312, pp.1044-6, 2006.

[20]. Brussow H. “Phage therapy: the Escherichia coli experience”. Microbiology. Vol.151:2133-40, 2005.

[21]. M. Ohnishi, K. Kurokawa, T. Hayashi, “Diversification of Escherichia coli genomes: are bacteriophages the major contributors?” Trends Microbiology, Vol.9, pp.481-5, 2001.

[22]. D.A. Relman, “New technologies, human-microbe interactions, and the search for previously unrecognized pathogens.” Journal of Infectious Diseases. Vol.186, pp.S254-8, 2002.

[23]. J. Hacker, E. Carniel. “Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes.” EMBO Reports, Vol.2, pp.376-81, 2001.

[24]. P.J. Sansonetti, “War and peace at mucosal surfaces.” Nature Review Immunology. Vol4, pp.953-64, 2004.

[25]. E. Rosenberg, O. Koren, L. Reshef, R. Efrony, I. Zilber-Rosenberg. “The role of microorganisms in coral health, disease and evolution.” Nature Review Microbiology. Vol.5, pp.355-62, 2007.

[26]. M.E. Bianchi, “DAMPs, PAMPs and alarmins: all we need to know about danger.” Journal of Leukocyte Biology, Vol.81, pp.1-5, 2007.

[27]. S. H. Huang, “Infectomics-based new theory for dissecting the pathogenesis, prevention and therapeutics of microbial infections”. Teda-Watson International forum on Biotechnology and Biomedicine, Vol.4, pp.73-74, 2005.

[28]. S.H. Huang, Y.H. Chen, G. Kong, et. al., “A novel genetic island of meningitic Escherichia coli K1 containing the ibeA invasion gene (GimA): functional annotation and carbon-source-regulated invasion of human brain microvascular endothelial cells. Functional and Integrative Genomics. Vol.1, pp.312-22, 2001.

[29]. J. Hejnova, U. Dobrindt, R. Nemcova, et. al., “Characterization of the flexible genome complement of the commensal Escherichia coli strain A0 34/86 (O83 : K24 : H31).” Microbiology. Vol.151, pp.385-98, 2005.

[30]. S.A. Joyce, D.J. Clarke, “A hexA homologue from Photorhabdus regulates pathogenicity, symbiosis and phenotypic variation.” Molecular Microbiology, Vol.47, pp.1445-57, 2003.

[31]. G. Silvestri, “Naturally SIV-infected sooty mangabeys: are we closer to understanding why they do not develop AIDS?” Journal of Medical Primatology, Vol.34, pp.243-52, 2005.

[32]. B.A. Webb, M.R. Strand, S.E. Dickey, et. al., “Polydnavirus genomes reflect their dual roles as mutualists and pathogens,” Virology, Vol. 347, pp.160-174, 2006.

[33]. R. Lucattini, V.A. Likic, T. Lithgow, Bacterial proteins predisposed for targeting to mitochondria. Molecular Biology and Evolution. Vol.21, pp.652-8, 2004.

[34]. T. Uchiumi, T. Ohwada, M. Itakura, et. al., Expression islands clustered on the symbiosis island of the Mesorhizobium loti genome. Journal of Bacteriology. Vol. 186, pp.2439-48, 2004.

[35]. M.C. Janssen, “Rationalizing focal point.” Theory and Decision, Vol.50, pp.119-48, 2001.

[36]. P.T. Leeson, C.J. Coyne, P.J. Boettke, “Converting social conflict: Focal points and the evolution of cooperation.” Review Austrian Economics, Vol.19, pp.137-47, 2006.

[37]. L.C. Gao, “Principles of Systemics”, Science Press, Beijing, China, pp.126-132, 2005.

[38]. A.M. Colman, “Thomas C. Schelling’s psychological decision theory: Introduction to a special issue.” Journal of Economic Psychology, Vol.27, pp.603-608, 2006.

[39]. D.E Yirrell, E. Wright, L.A. Shafer, et. al., “Association between active GB virus-C (hepatitis G) infection and HIV-1 disease in Uganda,” International Journal of STD & AIDS, Vol. 18, pp.244-9, 2007.

Correspondence Author

Sheng-He Huang, M.D.
Childrens Hospital Los Angeles, University of Southern California, Los Angeles, CA
E-mail: shhuang@usc.edu

GLP-1四十余年的发展与未来：从肠促胰素到多器官保护

Four Decades of GLP-1: From an Incretin Hormone to Multi-Organ Protection

刘西平 / Xiping Liu

NantWorks, Culver City, California, USA

摘要

过去四十余年，围绕胰高糖素样肽-1（glucagon-like peptide-1，GLP-1）的基础研究、临床转化和创新药物开发取得了历史性突破，成为现代转化医学最具代表性的成功范例之一。GLP-1受体激动剂（GLP-1 receptor agonists，GLP-1RAs）不仅显著改善2型糖尿病患者的血糖控制，还在肥胖、动脉粥样硬化性心血管疾病、糖尿病相关慢性肾脏疾病、代谢功能障碍相关脂肪性肝炎及其他肥胖相关疾病中展现出明确的临床获益，推动代谢性疾病治疗由单纯降糖向体重管理、多器官保护及长期临床结局改善转变。近年来，多靶点受体激动剂、口服小分子药物、偏向性信号调控、结构生物学及人工智能辅助药物设计等新技术不断推动GLP-1药物的优化与创新，进一步拓展了其在神经退行性疾病、健康老龄化及其他慢性疾病中的应用前景。本文系统回顾GLP-1领域四十余年来的重要科学发现、关键临床证据及创新药物研发进展，重点总结其基础生物学特征、受体信号转导机制、临床应用及未来发展方向，并结合精准医学和人工智能驱动药物研发的发展趋势，探讨GLP-1治疗平台未来的发展方向及其在下一代代谢性疾病治疗中的应用潜力，以期为基础研究、临床实践及创新药物研发提供参考。

关键词：胰高糖素样肽-1；GLP-1受体激动剂；2型糖尿病;肥胖;心-肾-代谢疾病;精准医学;人工智能

ABSTRACT

Over the past four decades, fundamental research, clinical translation, and innovative drug development centered on glucagon-like peptide-1 (GLP-1) have achieved remarkable breakthroughs, making the GLP-1 field one of the most representative success stories in modern translational medicine. GLP-1 receptor agonists (GLP-1RAs) not only substantially improve glycemic control in patients with type 2 diabetes but also provide well-established clinical benefits in obesity, atherosclerotic cardiovascular disease, diabetes-associated chronic kidney disease, metabolic dysfunction-associated steatohepatitis, and other obesity-related conditions. These advances have driven a major shift in the treatment of metabolic diseases from glucose lowering alone toward comprehensive weight management, multiorgan protection,

and improved long-term clinical outcomes.

In recent years, emerging technologies and therapeutic strategies, including multi-receptor agonists, oral small-molecule drugs, biased signaling, structural biology, and artificial intelligence-assisted drug design, have continued to advance the optimization and development of GLP-1-based therapeutics while further expanding their potential applications in neurodegenerative diseases, healthy aging, and other chronic disorders.

This review systematically examines more than four decades of major scientific discoveries, pivotal clinical evidence, and advances in drug development in the GLP-1 field, with particular emphasis on its fundamental biology, receptor signaling mechanisms, clinical applications, and future directions. In the context of emerging trends in precision medicine and artificial intelligence-driven drug discovery, we further discuss the evolution of the GLP-1 therapeutic platform and its potential role in next-generation treatment strategies for metabolic diseases, with the aim of providing a useful reference for basic research, clinical practice, and innovative drug development.

Keywards: Glucagon-like peptide-1 (GLP-1); GLP-1 receptor agonists (GLP-1RAs); type 2 diabetes; obesity; cardio-kidney-metabolic disease; precision medicine; artificial intelligence.

1. 引言

20世纪80年代，胰高糖素样肽-1(glucagon-like peptide-1，GLP-1）作为由胰高糖素原(proglucagon）裂解产生的重要肽类激素被发现，并被证实具有葡萄糖依赖性促进胰岛素分泌的肠促胰素（incretin）作用。早期研究主要聚焦于其在餐后血糖调节和葡萄糖稳态维持中的生理功能。随着分子生物学、药理学、结构生物学及临床医学的快速发展，人们逐渐认识到，GLP-1不仅是连接肠道与胰腺的重要内分泌信号分子，还广泛参与食欲调控、能量代谢、胃肠运动、神经内分泌调节及多器官功能稳态，在机体代谢调控网络中发挥着核心作用。

以GLP-1受体激动剂（GLP-1 receptor agonists，GLP-1RAs）为代表的新一代代谢治疗药物，深刻改变了2型糖尿病和肥胖的治疗策略。除显著改善血糖控制外，多项大型随机对照临床试验进一步证实，部分GLP-1RAs可显著降低体重，减少主要不良心血管事件（major adverse cardiovascular events，MACE）风险，延缓糖尿病相关慢性肾脏疾病（chronic kidney disease，CKD）进展，并在代谢功能障碍相关脂肪性肝炎（metabolic dysfunction-associated steatohepatitis，MASH）、肥胖相关心力衰竭及阻塞性睡眠呼吸暂停等疾病中展现出明确的临床获益。GLP-1RAs由此突破了传统降糖药物的治疗定位，推动代谢性疾病管理由单纯控制血糖逐步转向兼顾体重管理、多器官保护、生活质量改善及长期临床结局优化的综合治疗模式。

GLP-1广泛的临床获益源于其多层次、多通路的生物学作用。除促进葡萄糖依赖性胰岛素分泌、抑制胰高糖素释放、延缓胃排空及增强饱腹感等经典效应外，GLP-1还可通过减轻体重、改善血压和脂质代谢、抑制炎症反应与氧化应激、改善血管内皮功能等多种机制发挥系统性保护作用。近年来，大量基础和临床研究进一步揭示，GLP-1信号通路在神经系统、心血管系统及免疫炎症调控中同样具有重要生物学意义。然而，目前GLP-1受体在部分外周组织中的表达及其直接作用机制仍存在一定争议，其组织特异性功能及分子机制仍有待结合遗传学、空间组学、单细胞测序及功能学研究进一步阐明。

近十年来，GLP-1研究进入快速发展的新阶段。心血管结局试验（cardiovascular outcome trials，CVOTs）以及肥胖、慢性肾脏疾病、MASH等领域的一系列高质量临床研究不断积累循证医学证据，推动国际临床指南和治疗策略持续更新。与此同时，冷冻电镜结构解析、受体结构生物学、偏向性信号转导（biased signaling）、多肽工程、长效化修饰、口服递送技术及人工智能辅助药物设计等关键技术快速发展，推动

GLP-1药物研发由经验性优化逐步迈向结构指导、机制驱动的理性设计。GIP/GLP-1双受体激动剂、GLP-1/GCGR双受体激动剂以及GIP/GLP-1/GCGR三受体激动剂等新一代药物不断涌现，标志着代谢性疾病治疗进入多靶点协同调控的新阶段。

本文系统回顾GLP-1领域四十余年来的重要科学发现、关键临床证据及创新药物研发进展，重点总结其基础生物学、受体信号转导机制、临床应用及未来发展趋势，并结合精准医学和人工智能驱动药物研发的发展方向，探讨GLP-1治疗平台未来的创新机遇及其在下一代代谢性疾病治疗中的应用前景。

2. GLP-1研究四十余年的发展历程：从科学发现到临床转化

图 1. GLP-1 研究四十余年的发展历程。 图示概括了 GLP-1 领域的重要科学发现、关键药物研发及临床转化里程碑，展示了 GLP-1 由经典肠促胰素发展为覆盖 2 型糖尿病、肥胖及心—肾—肝—代谢疾病等多器官保护治疗平台的演进过程。注： GIP，葡萄糖依赖性促胰岛素多肽（glucose-dependent insulinotropic polypeptide）；GLP-1，胰高糖素样肽-1（glucagon-like peptide-1）；GLP-1RA，GLP-1 受体激动剂（GLP-1 receptor agonist）；T2D，2 型糖尿病（type 2 diabetes）；CVOTs，心血管结局试验（cardiovascular outcome trials）。

四十余年来，GLP-1研究伴随基础生物学、药物化学、结构生物学和临床医学的持续进步，实现了由基础发现向临床转化、由单纯降糖向多器官保护、由单靶点药物向多靶点治疗平台的不断演进。GLP-1也由一种主要参与餐后血糖调节的肠促胰素，逐步发展成为覆盖2型糖尿病、肥胖、心血管疾病、慢性肾脏疾病（chronic kidney disease，CKD）及代谢功能障碍相关脂肪性肝病（metabolic dysfunction-associated steatotic liver disease，MASLD）等多种慢性疾病的重要治疗平台。根据关键科学发现、代表性药物研发及临床证据的演进过程，其发展历程大致可分为生物学发现与机制奠基、临床转化、临床突破和创新拓展四个阶段（图1）。

2.1 生物学发现与机制奠基阶段（1983—1995 年）

20 世纪 80 年代初，胰高糖素原（proglucagon）基因被成功克隆并完成测序。随后研究发现，胰高糖素原可在不同组织中经组织特异性加工，产生胰高糖素、GLP-1 和 GLP-2 等多种生物活性肽。其中，具有完整生物活性的 GLP-1（7–37）和 GLP-1（7–36）酰胺相继得到鉴定，标志着现代 GLP-1 研究的正式开启 [1-4]。

功能研究进一步证实，GLP-1 可通过葡萄糖依赖性方式促进胰岛素分泌，抑制高血糖状态下不适当的胰高糖素释放，并通过延缓胃排空、增强饱腹感等作用共同参与餐后血糖调节。由于其促胰岛素作用随血糖水平下降而减弱，单独使用时发生严重低血糖的风险相对较低。这一独特的生理和药理学特征，为后续 GLP- 1 类药物的研发奠定了重要理论基础[1-3,5,6]。

1992 年，大鼠 GLP-1 受体（GLP-1 receptor，GLP-1R）首先被克隆，随后人 GLP-1R 也得到鉴定。GLP-1R 属于 B 类 G 蛋白偶联受体（class B G protein-coupled receptor，class B GPCR）。GLP-1R 的发现不仅阐明了GLP-1 发挥生物学作用的分子基础，也为受体药理学研究、构效关系分析及新型激动剂设计提供了关键靶点 [2,3,7]。

2.2 临床转化阶段：从天然激素到治疗药物（1995—2010 年）

天然 GLP-1 进入循环后极易被二肽基肽酶-4（dipeptidyl peptidase-4，DPP-4）降解，并经肾脏迅速清除，血浆半衰期仅约 1～2 min，严重限制了其直接临床应用。围绕这一关键瓶颈，研究逐步形成两条具有里程碑意义的药物开发路径：一是开发耐受 DPP-4 降解、具有更长作用时间的 GLP-1RAs；二是开发 DPP-4 抑制剂，以延长内源性 GLP-1 及其他肠促胰素的作用时间[3-5]。

基于从希拉毒蜥唾液中发现的 Exendin-4，研究人员开发出艾塞那肽（exenatide），并于 2005 年获批上市，成为全球首个用于临床的 GLP-1RA，标志着 GLP-1 完成了由内源性激素向治疗药物的历史性跨越。随后，脂肪酸侧链修饰、白蛋白结合和分子工程等长效化策略不断发展，利拉鲁肽（liraglutide）等长效制剂相继问世，并于 2010 年获批用于 2 型糖尿病治疗。GLP-1RAs 与 DPP-4 抑制剂的成功研发，共同开启了以肠促胰素为基础的代谢疾病治疗新时代[4,5]。

2.3 临床突破阶段：从降糖到心肾保护（2010—2020 年）

2008 年，美国食品药品监督管理局 (FDA) 要求新型降糖药物开展大型心血管结局试验（cardiovascular outcome trials，CVOTs），推动糖尿病药物评价体系由主要关注降糖疗效，逐步转向同时重视长期心血管安全性和临床结局 [6]。

随后，LEADER、SUSTAIN-6 和 REWIND 等关键临床试验相继证实，利拉鲁肽、司美格鲁肽和度拉糖肽不仅能够有效改善血糖控制，还可降低主要不良心血管事件（major adverse cardiovascular events，MACE）的发生风险。与此同时，多项研究显示，GLP-1RAs 还可减少白蛋白尿、延缓肾功能下降， 并带来持续减重、血压改善及其他综合代谢获益[8-10]。

这些高质量循证医学证据推动 GLP-1RAs 逐步突破传统降糖药物的治疗定位，成为兼具降糖、减重和心肾保护作用的重要治疗策略，也促使糖尿病管理理念由单纯控制糖化血红蛋白，转向以降低心血管和肾脏风险、改善长期临床结局为核心的综合管理模式。

2.4 创新拓展阶段：肥胖治疗与多靶点时代（2020 年至今）

进入 2020 年代，GLP-1 研究步入快速发展的新阶段，适应证拓展、药物创新和作用机制研究同步推进。STEP 系列研究证实，每周一次司美格鲁肽 2.4 mg 可显著降低超重或肥胖患者的体重，使 GLP-1Ras 正式成为肥胖长期药物治疗的重要选择。随后，SELECT 研究进一步证实，司美格鲁肽可降低无糖尿病但伴超重或肥胖患者的 MACE 风险，提示其心血管获益并非完全依赖降糖作用。FLOW 研究则首次通过专门设计的肾脏结局试验证实，

GLP-1RA 可降低 2 型糖尿病合并 CKD 患者的主要肾脏结局事件风险，进一步确立了其在心—肾—代谢综合管理中的重要地位[11-13]。

与此同时，多靶点药物开发成为 GLP-1 领域新的研究热点。GIP/GLP-1 双受体激动剂替尔泊肽(tirzepatide) 在降糖和减重方面展现出显著疗效；GIP/GLP-1/胰高糖素受体（glucagon receptor，GCGR） 三重激动剂 retatrutide 则进一步整合了抑制食欲、改善胰岛素敏感性、促进能量消耗和脂质动员等多重作用机制，代表了下一代代谢治疗的重要发展方向[14-16]。

另一方面，口服多肽制剂、非肽类小分子 GLP-1R 激动剂、超长效制剂和偏向性激动剂不断涌现。高分辨率冷冻电镜、人工智能辅助药物设计、蛋白质结构预测及多组学分析等新技术，正在推动 GLP-1 药物研发由经验性优化迈向结构指导、机制驱动和数据赋能的新阶段。与此同时，GLP-1 相关治疗正积极拓展至MASLD、阻塞性睡眠呼吸暂停（obstructive sleep apnea，OSA）、神经退行性疾病及其他慢性疾病。然而，不同适应证的循证医学证据成熟度仍存在明显差异，其长期疗效、安全性及适用人群尚需更多高质量随机对照试验进一步验证[15,17-22]。

综观 GLP-1 四十余年的发展历程，其成功充分体现了基础科学、药物工程、临床试验和产业创新协同推进的转化医学规律。从活性肽及其受体的发现，到突破天然 GLP-1 半衰期短的药学瓶颈；从首个 GLP-1RA 获批上市，到心肾保护、肥胖治疗、多靶点协同调控及人工智能辅助药物研发，GLP-1 已由经典肠促胰素发展成为连接代谢调控、多器官疾病治疗与现代转化医学的重要平台，也为创新药物研发提供了具有代表性的实践范式。

3. 基础研究驱动 GLP-1 创新发展：从分子机制到理性药物设计

GLP-1 由经典肠促胰素发展成为现代代谢医学的重要治疗平台，其持续创新始终建立在基础科学不断突破的基础之上。四十余年来，分子生物学、受体药理学、结构生物学、多肽工程、系统药理学及人工智能等领域的持续发展，不断深化了人们对 GLP-1 生物学功能和作用机制的认识，并推动 GLP-1 药物研发由经验性优化逐步迈向机制驱动、结构指导和数据赋能的理性设计。总体而言，GLP-1 领域的创新遵循了“生物学发现—机制解析—结构阐释—药物工程—临床验证”的经典转化医学路径 [2-6]。

3.1 GLP-1 生物学功能的持续拓展

GLP-1 主要由回肠和结肠 L 细胞分泌，脑干孤束核（nucleus tractus solitarius，NTS）中的部分神经元亦可合成 GLP-1。进食后，葡萄糖、脂肪酸和氨基酸等营养物质可刺激肠道 L 细胞释放 GLP-1，并通过连接肠道、脑和胰腺的“肠—脑—胰轴”（gut–brain–pancreas axis），共同调控葡萄糖稳态、胃肠运动、食欲及能量平衡[2,3,5,6]。

早期研究主要关注 GLP-1 促进葡萄糖依赖性胰岛素分泌、抑制高血糖状态下不适当的胰高糖素释放以及延缓胃排空等经典肠促胰素效应。近年来，大量基础和临床研究进一步表明，GLP-1 还广泛参与饱腹感形成、食物奖赏调节、脂肪组织代谢、能量平衡和神经内分泌调控，其生物学功能已远超传统肠促胰素的范畴。

目前，GLP-1R 在胰岛 β 细胞、胃肠神经系统和中枢神经系统中的表达及功能已得到较为充分的证实，而其在心脏、肾脏、肝脏及免疫细胞等外周组织中的表达水平和直接作用仍存在一定争议。现有证据提示, GLP-1相关治疗的多器官获益既可能部分来源于局部 GLP-1R介导的直接作用，也与体重下降、胰岛素抵抗改善、血压和脂质代谢优化、血管功能改善及全身炎症减轻等系统性效应密切相关。然而, 不同组织中 GLP-1R的真实表达谱、细胞特异性分布及直接作用机制，仍有待借助遗传学、单细胞测序、空间组学及功能学研究进一步阐明 [3,6,21,22]。

3.2 GLP-1 受体信号转导及偏向性调控

GLP-1R 属于 B 类 G 蛋白偶联受体。配体结合后，受体主要与 Gs 蛋白偶联，激活腺苷酸环化酶（adenylyl cyclase），促进环磷酸腺苷（cyclic adenosine monophosphate，cAMP）生成，并进一步激活蛋白激酶 A（protein kinase A，PKA）、cAMP 直接激活的交换蛋白（exchange protein directly activated by cAMP， Epac）及其下游多条信号通路。在胰岛 β 细胞中，这些信号协同调节离子通道活性、促进钙离子内流，并增强胰岛素颗粒募集、对接及胞吐，从而放大葡萄糖依赖性胰岛素分泌 [2,3,5,6]。

近年来，人们逐渐认识到，GLP-1R 并非简单的“开关型”受体，而是一个具有显著时间和空间特征的动态信号平台。受体激活后可募集 β-arrestin，参与受体脱敏、内化、胞内转运、内体信号转导及膜表面再循环。部分内化后的 GLP-1R 仍可在内体中持续产生 cAMP 信号，表明其信号输出不仅取决于受体是否被激活， 还受到亚细胞定位、转运路径和信号持续时间的精细调控。

不同 GLP-1R 激动剂在 G 蛋白激活、β -arrestin 募集、受体内化及胞内转运等方面可表现出不同程度的功能选择性，即偏向性激动（biased agonism）。理论上，通过优化不同信号通路的相对输出，有望在维持或增强治疗效应的同时，减少受体脱敏、延长药效并改善耐受性。然而，体外实验中观察到的信号偏向能否稳定转化为动物模型和临床中的疗效或安全性优势，仍有待建立标准化药理学评价体系，并通过适宜动物模型和前瞻性临床研究进一步验证[20-22]。

3.3 结构生物学推动理性药物设计

近年来，高分辨率冷冻电镜（cryo-electron microscopy，cryo-EM）的快速发展，使 GLP-1R 与天然配体、多肽药物、非肽类小分子激动剂及 G 蛋白复合物的三维结构相继得到解析，为阐明 B 类 GPCR 的配体识别、受体激活及信号选择机制提供了重要结构基础[19]。

研究表明，肽类配体通常首先通过其 C 端与受体胞外结构域结合，随后 N 端深入跨膜结构域并诱导受体构象改变，最终实现受体活化。不同配体在结合构象、关键氨基酸相互作用及受体微环境上的细微差异，可进一步影响受体活化效率、信号偏向及胞内转运行为。与此同时，部分非肽类小分子可结合于 GLP-1R 跨膜结构域内的别构位点，以不同于天然多肽的方式稳定受体活化构象。这些发现为口服小分子 GLP-1R 激动剂及新型别构调节剂的研发提供了重要理论依据 [19,20,22]。

通过整合冷冻电镜、分子动力学模拟、构效关系（structure–activity relationship，SAR）分析、计算化学及人工智能辅助的蛋白质结构预测，研究人员能够更加精准地优化配体亲和力、受体选择性、药代动力学特征和信号输出模式。结构指导药物设计（structure-based drug design）由此逐渐成为新一代 GLP-1 药物研发的重要策略 [19,20,22]。

3.4 多肽工程与多靶点协同调控

天然 GLP-1 极易受到 DPP-4 降解，血浆半衰期仅约 1～2 min。针对这一药代动力学瓶颈，研究人员通过DPP-4 切割位点修饰、脂肪酸侧链连接、白蛋白结合、Fc 融合及缓释递送等多种工程策略，成功推动GLP- 1 药物由短效制剂向每日一次、每周一次乃至更长给药周期持续发展。口服司美格鲁肽与吸收促进剂 SNAC〔sodium N-[8-(2-hydroxybenzoyl) amino] caprylate〕联合制剂的成功开发，则实现了多肽类 GLP-1 药物口服递送的重要突破 [4,5]。

随着对代谢性疾病系统性和异质性特征认识的不断深化，药物研发理念也由单一受体激动逐步转向多靶点协同调控。GIP/GLP-1双受体激动剂替尔泊肽（tirzepatide）已证实其在降糖和减重方面具有显著临床优势；GIP/GLP-1/胰高糖素受体（glucagon receptor，GCGR）三重激动剂 retatrutide 则进一步整合了抑制食欲、改善胰岛素敏感性、促进能量消耗和脂质动员等多重作用机制，展现出更大的减重及代谢改善潜力[14- 16,22]。

现代多靶点药物设计强调系统药理学理念。不同受体活性并非简单叠加，而需要综合优化各靶点的相对活性、内在效能、信号强度、药代动力学特征和组织暴露，以实现血糖控制、体重下降、能量代谢改善和安全耐受性之间的最佳平衡。如何确定不同受体作用的理想比例，并在不同患者和疾病阶段实现精准匹配， 已成为多靶点药物研发中的关键科学问题[14-16,22]。GLP-1 通过经典内分泌作用及肠—脑—胰轴调控，并结合体重下降、炎症减轻及代谢改善等系统性效应，共同实现多器官保护（图 2）。

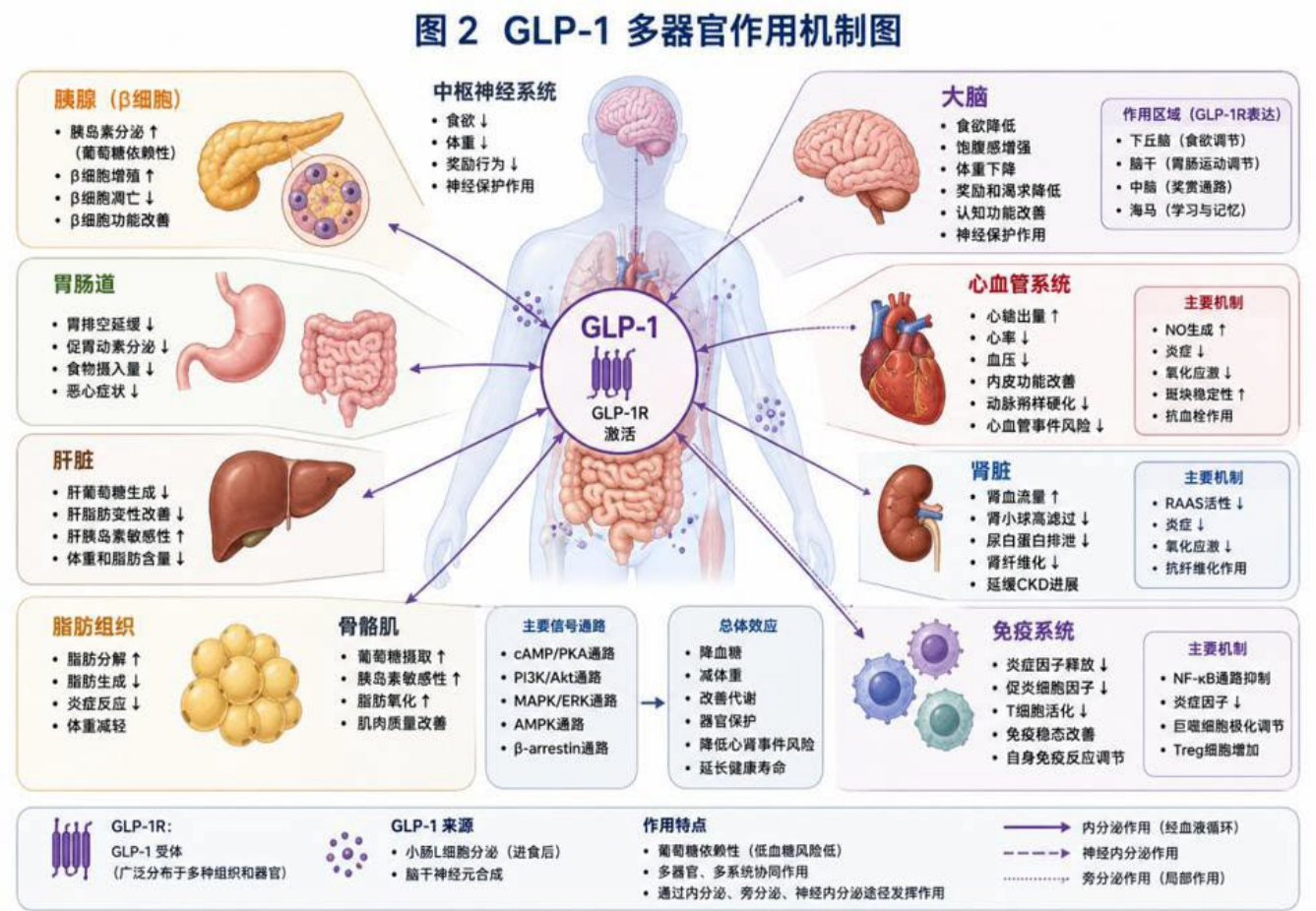


图 2. GLP-1 多器官作用机制示意图 GLP-1 通过经典肠促胰素作用及肠—脑—胰轴调控，并结合体重下降、炎症减轻及代谢改善等系统性效应，协同发挥对胰腺、脑、心脏、肾脏、肝脏、脂肪组织及胃肠道等多个器官的保护作用。

综上所述，基础研究始终是推动 GLP-1 持续创新的核心动力。从活性肽发现、受体克隆及信号网络解析， 到冷冻电镜结构解析、多肽工程、多受体协同调控和人工智能辅助药物设计，每一次科学发现与技术突破都推动 GLP-1 药物研发进入新的阶段。未来，随着单细胞组学、空间组学、系统药理学、计算生物学及人工智能等前沿技术的深度融合，GLP-1 药物研发将进一步由经验性优化迈向机制驱动、结构指导、数据赋能和精准设计的新阶段，为代谢性疾病及相关多器官疾病的精准治疗提供更加坚实的理论基础和创新路径[19-22]。

4. GLP-1 受体激动剂的临床应用：从降糖到多器官保护

GLP-1RAs 的临床应用推动了现代代谢医学治疗理念的深刻变革。自 2005 年首个 GLP-1RA 艾塞那肽获批用于 2 型糖尿病治疗以来，其临床应用已由单纯降糖逐步拓展至肥胖、动脉粥样硬化性心血管疾病（atherosclerotic

cardiovascular disease，ASCVD）、糖尿病相关慢性肾脏疾病（chronic kidney disease， CKD）、代谢功能障碍相关脂肪性肝炎（metabolic dysfunction-associated steatohepatitis，MASH）及其他肥胖相关疾病。GLP-1 相关治疗的核心价值并不局限于单一器官，而在于通过改善肥胖、胰岛素抵抗、能量代谢异常和慢性低度炎症等共同病理基础，促进多个器官系统临床结局的协同改善，推动心—肾—肝—代谢一体化管理理念的发展 [5,6,21,22]。

4.1 糖尿病治疗：从控制血糖到综合代谢管理

GLP-1RAs 主要通过增强葡萄糖依赖性胰岛素分泌、抑制高血糖状态下不适当的胰高糖素释放、延缓胃排空及减少能量摄入等机制发挥降糖作用。由于其促胰岛素效应具有葡萄糖依赖性，单药治疗时发生严重低血糖的风险较低；但与胰岛素或磺脲类药物联合应用时，仍应根据患者的血糖水平和低血糖风险，适当调整相关药物剂量 [2,3,6,21]。

除显著降低糖化血红蛋白（glycated hemoglobin，HbA1c）外，GLP-1RAs 还可减轻体重，并改善血压、血脂等多种代谢危险因素。LEADER、SUSTAIN-6 和 REWIND 等大型随机对照试验进一步证实，部分 GLP- 1RAs 不仅能够改善血糖控制，还可降低主要不良心血管事件（major adverse cardiovascular events， MACE）的发生风险，并对肾脏结局产生积极影响[8-10]。

随着循证医学证据不断积累，2 型糖尿病治疗理念已由传统的“以血糖为中心”，逐步转向综合评估 ASCVD、CKD、心力衰竭、肥胖、低血糖风险及患者个体需求。近年来的糖尿病诊疗指南和专家共识进一 步强调，应根据患者的临床特征、合并症和治疗目标实施个体化治疗，而非仅依据 HbA1c 水平选择降糖药物。对于合并 ASCVD 或心血管高风险的患者，应优先选择具有明确心血管结局获益证据的 GLP-1RA；对于需要显著减重的患者，司美格鲁肽及 GIP/GLP-1 双受体激动剂替尔泊肽显示出较大的临床优势 [6,8- 10,14,17,21,22]。因此，GLP-1RAs 已成为现代 2 型糖尿病综合管理的重要支柱，也是心—肾—代谢综合管理策略的重要组成部分。

4.2 肥胖治疗：开启高效药物减重新时代

肥胖是一种由遗传、神经内分泌、环境和行为等多种因素共同作用的慢性、复发性疾病，也是 2 型糖尿 病、心血管疾病、代谢功能障碍相关脂肪性肝病（metabolic dysfunction-associated steatotic liver disease， MASLD）、阻塞性睡眠呼吸暂停（obstructive sleep apnea，OSA）及多种恶性肿瘤的重要危险因素。生活方式干预始终是肥胖治疗的基础，但单纯依靠饮食控制和运动通常难以长期维持具有临床意义的体重下降[6,17,21,22]。This project was made possible by NIH Grants (AT002478, AT004620, and AT004504) from the National Center for Complementary and Integrative Medicine (to SXM).

STEP 1 研究显示，在生活方式干预基础上，每周一次皮下注射司美格鲁肽 2.4 mg，可使无糖尿病的超重或肥胖成年人在 68 周时平均体重下降 14.9%，而安慰剂组仅下降 2.4% [11]。SURMOUNT-1 研究进一步证实，替尔泊肽 15 mg 治疗 72 周可使患者平均体重下降 20.9%，将药物减重效果提升至新的水平 [14]。

GLP-1 相关药物带来的获益不仅体现在体重下降，还包括腰围缩小、胰岛素敏感性提高，以及血压、血糖、血脂和炎症指标的综合改善。然而，肥胖具有显著的慢性复发特征，停药后食欲增强和体重回升较为常见。因此，肥胖治疗应定位于长期慢病管理，而非短期减重。GLP-1 相关药物的应用，标志着肥胖治疗由主要依赖生活方式干预逐步进入高效、长期药物管理时代，其治疗目标也由单纯降低体重扩展至改善代谢健康、身体功能、生活质量及长期临床结局 [11,14,17,21,22]。

4.3 心血管保护：超越降糖的临床价值

LEADER、SUSTAIN-6、Harmony Outcomes、REWIND、AMPLITUDE-O 及 SOUL 等大型心血管结局试验

（cardiovascular outcome trials，CVOTs）表明，部分具有明确循证医学证据的 GLP-1RAs 能够降低 MACE 发生风险。由于不同药物的分子结构、受体药理学特征、研究人群及试验设计存在差异，各项研究显示的获益程度并不完全一致，因此，目前尚不能将程度相同的心血管保护作用简单视为所有 GLP-1RAs 的共同类效应 [8-10,23-25]。

SELECT 研究首次将 GLP-1RA 的心血管获益拓展至无糖尿病的超重或肥胖人群。该研究纳入 17 604 例既往存在心血管疾病、体质指数≥27 kg/m²且无糖尿病的成年人，结果显示，司美格鲁肽 2.4 mg 可使 MACE 发生风险降低 20%[13]。基于该研究结果，美国 FDA 于 2024 年扩展 Wegovy 的适应证，用于降低伴超重或肥胖且已有心血管疾病的成人发生心血管死亡、非致死性心肌梗死及非致死性脑卒中的风险 [13,21,22]。

目前认为，GLP-1RAs 的心血管保护可能来源于体重下降、血压和血脂改善、炎症反应减轻、血管内皮功能改善，以及其他直接或间接的代谢效应，而并非仅依赖血糖下降[3,21,22]。这些研究不仅确立了 GLP-1RAs 在 ASCVD 患者及高危人群中的重要地位，也推动糖尿病药物评价体系由主要关注降糖疗效，逐步转向重视心血管安全性和长期临床结局。

在肥胖相关射血分数保留型心力衰竭（heart failure with preserved ejection fraction，HFpEF）患者中，STEP-HFpEF 及 STEP-HFpEF DM 研究显示，司美格鲁肽 2.4 mg 可显著改善心力衰竭相关症状和身体活动受限程度，提高运动耐量和生活质量，并伴随明显的体重下降[26,27]。这些结果提示，肥胖和脂肪组织功能异常可能是部分 HFpEF 患者的重要可干预因素。不过，截至 2026 年 7 月，HFpEF 尚未成为司美格鲁肽在美国获批的独立适应证，其对心力衰竭住院、心血管死亡等硬临床终点的长期影响，仍需更大规模、以临床事件为主要终点的研究进一步证实[26,27]。

4.4 肾脏保护：从探索性证据到专门结局试验证实

早期 CVOTs 的次要终点及探索性分析提示，GLP-1RAs 能够减少新发大量白蛋白尿，并可能延缓估算肾小球滤过率（estimated glomerular filtration rate，eGFR）下降，但早期肾脏复合终点的获益主要由白蛋白尿改善所驱动 [8-10,23,24]。

FLOW 是首项以主要肾脏事件为核心终点的 GLP-1RA 专门结局试验，标志着其肾脏保护证据由探索性分析进入专门结局试验证实的新阶段。该研究纳入 3 533 例 2 型糖尿病合并 CKD 患者，结果显示，每周一次司美格鲁肽 1.0 mg 可使主要肾脏复合终点事件风险降低 24%；同时，MACE 和全因死亡风险分别降低 18%和20%[12]。

GLP-1RAs 的肾脏保护可能与改善血糖、体重、血压和白蛋白尿，以及减轻炎症反应、氧化应激和肾脏脂毒性等多重机制有关 [3,12,21,22]。在临床实践中，GLP-1RAs 并不替代肾素—血管紧张素系统抑制剂、钠— 葡萄糖协同转运蛋白 2 抑制剂（sodium–glucose cotransporter 2 inhibitors，SGLT2is）或非甾体盐皮质激素受体拮抗剂，而应根据患者的代谢状态、心血管风险、白蛋白尿程度及肾功能状况，与上述药物形成相互补充的综合治疗策略 [6,12]。

4.5 MASH 及其他肥胖相关适应证

MASLD 是肥胖和 2 型糖尿病最重要的代谢相关并发症之一，其伴有肝细胞损伤、炎症及不同程度纤维化的进展性表型称为 MASH。GLP-1RAs 可通过减轻体重、改善胰岛素抵抗、降低脂毒性及优化全身代谢状态，减少肝脏脂肪蓄积，并改善部分患者的肝脏组织学表现 [3,18,21,22,28]。

III期 ESSENCE 研究纳入经肝活检确诊、伴 F2～F3 期肝纤维化的 MASH 患者。第 72 周中期分析显示，司美格鲁肽 2.4 mg 组达到 “MASH 缓解且肝纤维化未恶化” 的患者比例为 62.9%，安慰剂组为 34.3%；达到 “肝纤维化改善且 MASH 未恶化” 的患者比例分别为 36.8%和 22.4%[28]。基于上述组织学替代终点，美国FDA 于 2025 年

8 月加速批准 Wegovy 用于治疗成人非肝硬化 MASH 伴中度至重度肝纤维化，即 F2～F3 期纤维化。ESSENCE 研究计划继续随访，以进一步评价其对肝硬化进展、肝功能失代偿肝移植及死亡等长期临床结局的影响[21,22,28]。

在 OSA 领域，两项III期 SURMOUNT-OSA 研究显示，替尔泊肽治疗 52 周可显著降低呼吸暂停—低通气指数，减轻低氧负荷，并改善睡眠相关症状和体重[29]。基于上述结果，美国 FDA 于 2024 年 12 月批准Zepbound 联合低热量饮食和增加体力活动，用于治疗伴肥胖的成人中重度 OSA，使其成为美国首个获批用于 OSA 治疗的药物。

此外，GLP-1　相关药物在肥胖相关膝骨关节炎、多囊卵巢综合征及其他代谢异常中的应用仍在持续探索。STEP 9 研究显示，在伴肥胖和中重度疼痛的膝骨关节炎患者中，司美格鲁肽 2.4 mg 除显著降低体重外，还可减轻膝关节疼痛并改善身体功能[30]；但截至 2026 年 7 月，其尚未获得独立的骨关节炎适应证。GLP-1 相关药物在神经退行性疾病、肿瘤代谢、成瘾行为及健康老龄化等领域的应用，目前仍处于基础研究或临床探索阶段，其疗效、适用人群及长期安全性尚待进一步验证[21,22]。

总体而言，GLP-1 相关治疗已由传统降糖药物发展成为覆盖 2 型糖尿病、肥胖、ASCVD、CKD、MASH 及其他肥胖相关疾病的重要治疗平台。代表性 GLP-1 相关药物的发展历程、关键临床研究及主要临床应用见表 1。

表1 代表性GLP-1相关药物、关键临床研究及主要临床应用

药物	作用靶点	首次获批*	代表性临床研究	已证实的主要临床价值
艾塞那肽（Exenatide）	GLP-1R	2005	AMIGO	首个GLP-1RA，改善血糖控制
利拉鲁肽（Liraglutide）	GLP-1R	2010	LEADER	降糖、减重、降低MACE风险
度拉糖肽（Dulaglutide）	GLP-1R	2014	REWIND	降低MACE风险，改善肾脏结局
司美格鲁肽（Semaglutide）	GLP-1R	2017	SUSTAIN-6、STEP、SELECT、FLOW、ESSENCE	降糖、减重、ASCVD、CKD、MASH
替尔泊肽（Tirzepatide）	GIPR/GLP-1R	2022	SURPASS、SURMOUNT、SURMOUNT-OSA	强效降糖、减重、OSA
Retatrutide	GIPR/GLP-1R/GCGR	III期	Phase 2 Obesity、Phase 2 T2D	超高幅度减重、多靶点治疗
CagriSema	GLP-1RA + 胰淀素类似物	III期	REDEFINE	增强减重作用
Zenagamtide（Amycretin）	GLP-1R + 胰淀素受体	II期	Phase 1	新型单分子双激动剂
MariTide	GLP-1R激动 + GIPR拮抗	III期	Phase 2	超长效（月制剂）
Orforglipron	小分子GLP-1R激动剂	2026	ACHIEVE、ATTAIN	首个口服非肽GLP-1R激动剂

注：GLP-1R，GLP-1受体；GIPR，葡萄糖依赖性促胰岛素多肽受体；GCGR，胰高糖素受体；MACE，主要不良心血管事件；CKD，慢性肾脏疾病；MASH，代谢功能障碍相关脂肪性肝炎；OSA，阻塞性睡眠呼吸暂停。*首次获批时间主要指美国FDA批准时间。

5. GLP-1 治疗的新兴研究领域：从代谢疾病到更多治疗领域

随着 GLP-1RAs 在糖尿病、肥胖及心—肾—代谢疾病中的临床价值不断确立，其研究范围正逐步拓展至神经退行性疾病、健康老龄化及其他潜在治疗领域。基础研究表明，GLP-1　信号除参与葡萄糖稳态和能量代谢调控外，还可能影响神经炎症、氧化应激、线粒体稳态及细胞代谢网络[3,21,22]。然而，与糖尿病、肥胖及心—肾—代谢疾病已积累较充分的循证医学证据不同，上述新兴领域多数仍处于机制研究、早期临床探索或概念验证阶段，其真实临床价值尚需更多高质量随机对照试验进一步确认。因此，在评价这些潜在适应证时，应严格区分

基础研究发现、观察性关联、早期临床信号与随机对照试验证据，避免将生物学合理性或生物标志物变化过早等同于临床治疗有效性。

5.1 神经退行性疾病：机制前景与临床挑战

中枢神经系统是 GLP-1 新兴研究中最受关注的领域之一。GLP-1R 表达于多个参与学习记忆、自主神经调节、食欲控制和能量平衡的脑区。动物研究提示，GLP-1RAs 可能通过改善脑内胰岛素信号、减轻神经炎症和氧化应激、维持线粒体稳态及保护突触功能等途径发挥神经保护作用[3,21,22]。然而，不同药物在血脑屏障通透性、中枢神经系统暴露、受体激活强度及药代动力学特征等方面存在明显差异；与此同时，体重、血糖、血压及心脑血管危险因素的改善也可能间接影响神经系统结局。因此，在临床研究中，外周代谢改善所带来的间接获益与中枢 GLP-1R 介导的直接作用往往难以完全区分。

阿尔茨海默病（Alzheimer's disease，AD）是研究较为深入的潜在适应证之一。动物实验及部分观察性研究提示，GLP-1RAs 可能影响 β-淀粉样蛋白沉积、Tau 蛋白异常磷酸化、神经炎症及脑能量代谢。然而， EVOKE 和 EVOKE Plus 两项III期随机对照研究显示，口服司美格鲁肽治疗早期症状性 AD 未能显著改善临床痴呆评定量表箱总分等主要临床终点，也未证实其能够稳定延缓认知功能和日常生活能力下降[31]。即使部分炎症指标或疾病相关生物标志物出现变化，也不能据此推断患者必然获得具有临床意义的认知或功能获益。

帕金森病（Parkinson's disease，PD）的研究同样呈现基础机制积极而临床结果不一致的特点。动物模型显示，GLP-1RAs 可能减轻神经炎症、改善线粒体功能并保护多巴胺能神经元。LIXIPARK II期研究显示，利西拉肽治疗 12 个月后运动功能恶化程度低于安慰剂组，但多数次要终点未显示一致获益，且胃肠道不良反应较为常见[32]。随后开展的每周一次艾塞那肽III期研究未发现其能够改善运动症状或延缓 PD 进展，也未显示具有一致性的疾病修饰效应 [33]。总体而言，现有证据尚不足以支持 GLP-1RAs 作为 AD 或 PD 的常规治疗药物。

未来研究需要进一步明确适宜的疾病阶段、治疗时间窗、中枢有效暴露水平及潜在获益人群，并采用认知功能、运动功能、日常生活能力、生活质量和疾病进展等具有患者临床意义的终点进行验证，而不能仅依赖影像学或生物标志物变化[31-33]。

5.2 健康老龄化：改善健康寿命而非简单“抗衰老”

衰老通常伴随内脏脂肪增加、骨骼肌量和肌力下降、胰岛素敏感性降低、慢性低度炎症及多器官储备功能减退。GLP-1 相关药物能够改善肥胖、糖尿病、心血管疾病、CKD、OSA 和 MASH 等多种年龄相关疾病， 因而可能通过降低疾病负担、减少重大临床事件并维持身体功能，间接促进健康老龄化[12-14,21,22,28,29]。

基础研究提示，GLP-1 信号可能参与线粒体稳态、自噬、氧化应激及细胞代谢适应等过程[21,22]。但目前尚无充分临床证据证明 GLP-1RAs 能够直接延缓人类生物学衰老或延长寿命，也不存在经监管机构批准的“抗衰老” 适应证。因此，将其简单描述为延寿药物或抗衰老药物并不严谨。

老年患者应用 GLP-1 相关药物时，还需特别关注身体成分和功能状态的变化。显著减重虽然有助于改善代谢健康、活动能力和肥胖相关并发症，但减轻的体重中通常包含一定比例的瘦体重，可能增加肌肉量下降的风险，尤其需要关注高龄、衰弱、营养不良或原有肌少症患者[17,21,22]。同时，影像学测得的瘦体重下降不能简单等同于临床肌少症；是否同时伴有肌力、步速、身体活动能力和独立生活能力下降，更具有临床意义。

因此，老年患者的治疗目标不应仅聚焦于体重数字，还应综合评估肌肉量和肌力、蛋白质与能量摄入、身体功能、衰弱程度、跌倒风险及独立生活能力。必要时应结合抗阻运动、充足蛋白质摄入和个体化营养干预，在改善代谢风险的同时尽可能维持肌肉与骨骼健康。健康老龄化的核心并非单纯降低体重，而是在减少疾病负担的同时维持功能储备、独立生活能力和生活质量，促进健康寿命（healthspan）的延长，而非简单追求寿命延

长 [17,21,22]。

5.3 未来研究方向

GLP-1 研究向神经科学和老年医学领域延伸，反映了代谢异常与神经功能、慢性炎症及器官衰老之间的密切联系。然而，GLP-1 药物在代谢疾病领域取得的显著成功，也可能引发“适应证外溢”，即依据其广泛的生物学作用，推断其能够治疗多种尚缺乏充分临床证据的疾病。

未来研究应重点解决三个方面的问题。第一，应利用单细胞测序、空间组学、遗传学工具及高特异性受体检测技术，进一步明确 GLP-1R 在不同组织、细胞类型及疾病状态下的真实表达、空间分布和生物学功能[21,22]。第二，应结合药代动力学、组织分布、血脑屏障通透性及受体占有率研究，确定候选药物能否在目标器官达到具有生物学意义的有效暴露，并明确外周代谢效应与局部直接作用之间的相对贡献。第三， 应整合多组学、人工智能及精准分层策略，识别真正可能获益的患者亚群，并采用疾病特异性、患者中心的临床终点验证疗效，而不能仅依赖生物标志物或影像学变化。

总体而言，新兴研究领域再次表明，机制合理、生物标志物改善或动物模型中的积极结果，并不等同于临床有效。GLP-1 新适应证的开发仍应遵循“靶点验证—机制研究—早期临床探索—随机对照试验—长期获益评估”的循证路径。只有将可靠的机制证据最终转化为可重复、具有临床意义且获益—风险比可接受的患者结局，GLP-1 相关药物才能真正由代谢治疗平台进一步拓展为覆盖更多系统性疾病的治疗策略[21,22,31- 33]。

6. GLP-1 药物开发创新：从多靶点协同到结构指导与人工智能驱动

GLP-1 药物开发已进入多种技术路线并行推进、快速迭代的新阶段。早期研发主要致力于克服天然 GLP-1 半衰期短、易被二肽基肽酶-4（dipeptidyl peptidase-4，DPP-4）降解及需要频繁注射等局限；当前研发目标则进一步转向提高减重和代谢改善幅度、拓展心—肾—肝等多器官获益、改善胃肠道耐受性、延长给药间隔，并满足不同患者对口服、注射及个体化治疗的多样化需求 [4,5,21,22]。

多靶点调节剂、口服多肽、非肽类小分子、超长效制剂、组织选择性和偏向性激动剂，以及人工智能辅助药物设计等技术，正在共同重塑 GLP-1 药物研发格局。与此同时，药物评价标准也由单纯比较糖化血红蛋白水平或体重降幅，逐步转向综合评价长期疗效、多器官保护、安全性、身体功能、治疗依从性、可制造性、供应能力及真实世界可及性。

6.1 多靶点协同治疗

肥胖和 2 型糖尿病涉及食欲调控、胰岛功能、胰岛素敏感性、脂质代谢及能量消耗等多条相互关联的生理通路。基于复杂代谢性疾病需要多机制协同干预的认识，药物研发策略已由单一 GLP-1 受体激动，逐步拓展至与葡萄糖依赖性促胰岛素多肽受体（glucose-dependent insulinotropic polypeptide receptor，GIPR）、胰高糖素受体（glucagon receptor，GCGR）及胰淀素相关受体通路的联合调控[5,14-16,21,22]。

GIP/GLP-1 双受体激动剂替尔泊肽的成功，验证了多靶点策略在降糖和减重方面的重要临床价值。GIP/GLP-1/GCGR 三重激动剂 retatrutide 进一步引入 GCGR 介导的脂质动员和能量消耗作用。II期研究显示，retatrutide 治疗 48 周可使肥胖患者体重显著下降，最高剂量组平均降幅达 24.2%，展现出较大的减重潜力 [15,16]。目前，III期研究正在进一步评价其长期减重效果及对肥胖相关并发症的影响，其最终临床定位仍需结合完整研究结果、长期安全性及心血管和其他多器官结局综合判断。

GLP-1 与胰淀素通路的协同调控也是当前的重要研发方向。CagriSema 由长效胰淀素类似物 cagrilintide 与司

美格鲁肽组成，旨在通过两条互补的食欲调控通路增强饱腹感、减少能量摄入并提高减重效果；相关III期研究已分别在无糖尿病的肥胖或超重人群及伴 2 型糖尿病人群中报告结果 [34,35]。Zenagamtide（原研发代号amycretin）是一种兼具 GLP-1R 和胰淀素受体激动活性的单分子候选药物，正在推进后续临床开发[36]。

Maridebart cafraglutide（MariTide）采用 GLP-1R 激动与 GIPR 拮抗相结合的抗体—多肽偶联设计，并探索每月一次或更低频率的给药方案 [37]。上述进展表明，多靶点药物并非不同受体活性的简单叠加，而需要系统优化各靶点的相对活性、内在效能、信号持续时间、组织暴露及药代动力学特征，以在血糖控制、体重管理、多器官获益及长期安全性之间实现最佳平衡 [15,16,21,22,34-37]。

6.2 口服多肽与非肽类小分子

注射给药是影响部分患者治疗接受度和长期依从性的重要因素。口服司美格鲁肽通过与吸收促进剂SNAC〔sodium N-[8-(2-hydroxybenzoyl) amino] caprylate〕共同制剂化，促进多肽药物在胃内吸收，是 GLP-1 多肽口服递送的重要突破 [4,5]。但其口服生物利用度仍较低，对服药时间、饮水量及进食间隔具有特定要求， 剂量利用效率和用药便利性仍有进一步优化空间。

非肽类小分子 GLP-1R 激动剂通常可采用传统化学合成工艺生产，具有口服方便、常温储运及潜在规模化制造等优势。Orforglipron 是一种每日一次口服的非肽类小分子 GLP-1R 激动剂。临床研究显示，其可使肥胖或超重人群获得具有临床意义的体重下降，并可显著改善 2 型糖尿病患者的糖化血红蛋白水平[38,39]。2026 年 4 月 1 日，美国 FDA 批准 orforglipron（商品名 Foundayo）联合低热量饮食和增加体力活动，用于成人肥胖，或伴至少一种体重相关合并症的超重患者减轻并长期维持体重，标志着口服非肽类小分子 GLP- 1R 激动剂正式进入临床应用阶段[40]。

与天然多肽相比，小分子配体通常以不同方式结合 GLP-1R 跨膜结构域，并可能稳定特定的受体构象，从而产生不同的内在效能、信号偏向及受体转运特征[19-22]。但口服小分子并非在所有方面均优于多肽药物，其肝脏代谢、药物相互作用、活性代谢产物、脱靶风险以及长期心血管和肾脏结局仍需持续评价[21,22,38,39]。

6.3 长效化与组织选择性

通过脂肪酸侧链修饰、白蛋白结合、抗体或 Fc 融合以及缓释递送等技术，GLP-1 药物已由每日给药发展至每周给药，并开始探索每月一次乃至更低频率的治疗方案[4,5,37]。低频给药有望提高用药便利性和长期依从性，但超长半衰期也可能降低剂量调整和不良反应处置的灵活性。因此，长效化设计需要在给药便利 性、持续疗效、药效可逆性和个体化滴定之间取得平衡，而不能单纯追求更长的半衰期。组织选择性策略试图通过调节分子大小、亲脂性、电荷、蛋白结合及组织穿透能力，改变药物在中枢神经系统、胰岛和其他外周器官中的分布。理论上，增强特定中枢区域的有效暴露可能提高食欲抑制和食物奖赏调节作用；外周选择性设计则可能在保留部分代谢效应的同时，减少某些中枢相关不良反应[21,22]。

然而，组织选择性不能仅依据分子结构或体外活性加以推断，仍需通过定量组织分布、游离药物浓度、靶点占有率、组织特异性信号及药效学研究进行验证。尤其对于中枢作用，血脑屏障通透性、脑内游离药物暴露、局部受体表达及神经环路可及性均可能影响最终疗效。因此，“能够进入脑组织”并不必然等同于能够在相关神经环路产生充分且具有临床意义的药理作用。

6.4 偏向性激动与受体转运

GLP-1R 激活后可引发 Gs/cAMP 信号、β-arrestin 募集、受体内化、内体信号转导及膜表面再循环。不同配体在这些过程中的相对效能、持续时间和亚细胞空间分布存在差异，由此产生偏向性激动或功能选择性 [20-

22]。

部分 G 蛋白偏向性配体可在维持 cAMP 信号的同时减少 β-arrestin 募集和受体内化，理论上可能有助于保留膜表面受体、减轻脱敏并延长功能反应。然而，受体内化并不等同于信号终止，内化后的 GLP-1R 仍可能在内体中持续产生 cAMP 信号，而受体再循环效率也会影响长期药效。因此，不能简单地将“减少受体内化”视为更优的药理学特征。

偏向性评价还容易受到细胞类型、受体表达水平、参考配体、检测时间、信号放大程度及分析模型等因素影响。单一体外指标不能直接预测临床减重幅度、胃肠道耐受性或多器官获益。未来仍需结合标准化体外评价、定量系统药理学、适宜动物模型和前瞻性临床研究，进一步验证这些药理学特征能否稳定转化为减重增强、耐受性改善及多器官保护等真实临床优势[20-22]。

6.5 结构生物学与人工智能辅助设计

冷冻电镜研究已逐步揭示 GLP-1R、GIPR 和 GCGR 与天然多肽、工程化配体、非肽类小分子及信号蛋白相互作用的结构基础[19]。结合分子动力学模拟、构效关系分析和自由能计算，研究人员能够预测氨基酸替换、侧链修饰及小分子取代基对配体亲和力、受体效能、靶点选择性和构象稳定性的影响，从而提高候选分子的设计效率。

人工智能可进一步整合氨基酸序列、三维结构、受体活性、药代动力学、毒理学、免疫原性及可制造性等多维数据，建立序列—结构—功能关系模型，并在预设受体活性比例、稳定性、溶解性、聚集倾向、免疫原性及生产工艺等约束条件下，生成或筛选候选分子。其潜在应用还包括虚拟筛选、脱靶预测、毒性和免疫原性评估、制剂优化、患者分层，以及真实世界不良反应信号识别[21,22]。

然而，人工智能模型的可靠性高度依赖训练数据的质量、规模、标准化程度和代表性。数据偏倚、实验条件不一致及阴性结果缺失，均可能限制模型的外推能力。因此，人工智能不能替代实验验证，而应与结构生物学、定量药理学、自动化合成、高通量筛选及临床转化研究深度融合，形成“计算设计—实验验证—数据学习—迭代优化”的闭环研发体系。只有经过生化、细胞、动物及临床多层次验证，人工智能生成的候选 分子才能真正转化为具有开发价值的创新药物[21,22]。

6.6 从减重竞争走向综合临床价值

随着部分新型 GLP-1 相关药物在临床研究中的平均体重降幅达到或超过 20%，单纯比较体重下降百分比已不足以全面反映其整体临床价值[14,16,34,37]。未来评价体系还应涵盖减重维持、停药后体重反弹、心—肾—肝临床结局、身体成分变化、骨骼肌和骨量保护、胃肠道耐受性、长期依从性、患者报告结局及生活质量等度。

药物生产规模、供应稳定性、储存运输条件、治疗成本及医疗保障覆盖，同样决定其真实世界可及性。口服小分子、低频注射制剂和高效多靶点药物可能分别适用于不同患者，而不会简单地相互取代。未来更可能形成依据疾病特征、治疗目标、合并症、耐受性、给药偏好及经济条件进行精准分层选择的多元化治疗格局。

总体而言，新一代 GLP-1 药物研发已由单纯追求更强的降糖和减重作用，逐步转向兼顾长期疗效、多器官保护、安全性、身体功能、治疗依从性及真实世界可及性的综合优化。未来最具临床价值的治疗方案，未必是靶点最多、半衰期最长或平均减重幅度最大的药物，而应是在疗效、耐受性、安全性、可制造性、供应能力及可负担性之间实现最佳平衡的治疗策略。

多靶点协同、口服小分子、超长效制剂、组织选择性、偏向性激动及人工智能辅助设计，共同代表了 GLP-1 药物研发的重要创新方向。随着结构生物学、系统药理学和人工智能的持续融合，GLP-1 药物研发正由经验性分子优化迈向结构指导、数据驱动、实验验证与精准分层相结合的新阶段。然而，新技术和新策略能否最终转

化为疗效稳定、安全可控且具有长期临床获益的治疗方案，仍需高质量临床研究和真实世界证据持续验证[19-22]。

7. GLP-1 挑战与展望：从循证证据到精准医学

过去四十余年，GLP-1 研究完成了由肠促胰素基础生物学发现向现代代谢医学重要治疗平台的跨越。GLP-1RAs 不仅显著改善 2 型糖尿病和肥胖患者的血糖控制与体重管理，还在动脉粥样硬化性心血管疾病、糖尿病相关慢性肾脏疾病、代谢功能障碍相关脂肪性肝炎及其他肥胖相关并发症中展现出重要临床价值[5,11-14,21,22,28,29]。这一发展历程充分体现了基础研究、受体药理学、结构生物学、药物工程与大型临床试验协同推进的转化医学模式。

尽管 GLP-1 相关治疗不断取得突破，长期获益与安全性仍是未来发展的首要挑战。胃肠道不良反应是最常见的耐受性问题，胆囊及胆道相关事件、胰腺安全性、营养摄入不足，以及瘦体重和骨骼肌减少等潜在风险，仍需通过长期随访和真实世界研究持续评价[17,21,22]。与此同时，肥胖具有显著的慢性复发特征，停药后食欲恢复和体重反弹较为常见，提示 GLP-1 相关治疗更适合作为长期慢病管理策略，而非短期减重手段。对于儿童和青少年、妊娠期或备孕女性、高龄患者、衰弱人群及多病共存患者等特殊人群，其长期疗效、安全性及获益—风险平衡仍有待更多专门研究进一步明确。

糖尿病和肥胖具有显著的生物学与临床异质性，不同患者在体重下降、血糖改善、胃肠道耐受性、身体成分变化以及心—肾—肝获益等方面存在明显差异。未来应综合临床表型、遗传背景、代谢特征、生物标志物、身体成分、器官功能及伴随疾病等多维信息，建立更加精准的患者分层、药物选择、剂量优化和疗效监测体系。治疗目标也应由单纯关注糖化血红蛋白和体重下降，逐步转向兼顾心血管结局、肾功能保护、肝脏获益、肌肉与骨骼健康、身体功能及生活质量的综合管理，从而推动 GLP-1 相关治疗由群体平均获益迈向真正意义上的精准医学[6,21,22]。

未来，新技术和新策略能否真正转化为患者获益，仍需更加坚实的循证医学证据支持。长期随机对照试 验、真实世界研究及标准化药理学评价，将继续构成判断 GLP-1 相关治疗临床价值的重要基础。无论是多靶点协同调控、受体信号优化、组织选择性设计，还是人工智能辅助药物研发，其最终价值均应以长期临床结局、安全性、身体功能、生活质量及真实世界疗效为评价标准，而不能仅依据体外实验、动物研究、生物标志物变化或短期减重幅度作出判断。

人工智能、结构生物学及多组学技术正在推动 GLP-1 研究进入更加精准和高效的发展阶段。未来，人工智能有望通过整合分子结构、受体信号、药代动力学、遗传信息、电子健康记录及真实世界数据，提高创新药物研发效率，并辅助患者分层、疗效预测和个体化治疗决策[21,22]。然而，人工智能模型的可靠性仍高度依赖训练数据的质量、完整性、标准化程度和代表性。实验平台差异、阴性数据缺失、数据共享不足、算法偏倚及可解释性有限，均可能影响模型的泛化能力。因此，人工智能不能替代实验验证和临床判断， 而应与结构生物学、定量药理学、高通量实验及前瞻性临床研究深度融合，共同构建“设计—验证—学习— 优化”的闭环研发体系。

GLP-1 药物的评价标准也将进一步由关注降糖和减重效果，拓展至长期多器官临床结局、减重维持、停药后体重反弹、身体成分、肌肉与骨骼健康、身体功能、患者报告结局、生活质量及治疗依从性等多维指标。与此同时，药物生产能力、供应稳定性、储存运输条件、治疗成本及医疗保障覆盖，也将直接影响创新成果能否转化为持续而广泛的公共健康获益。如何在创新速度、临床价值、患者可及性和医疗资源承受能力之间取得平衡，将成为 GLP-1 领域未来发展的重要课题 [21,22]。

展望未来，GLP-1 药物研发的目标将不再是寻找适用于所有患者的“最强药物”，而是建立安全、有效、可持续、可负担，并适用于不同疾病阶段、不同并发症谱及不同个体需求的精准治疗体系。随着结构生物学、系统药理学、人工智能、长期临床结局研究及真实世界证据的持续发展，GLP-1 相关治疗有望进一步实现由单靶点干预迈向多通路系统调控，由群体平均治疗迈向精准分层治疗，由单一代谢指标控制迈向心—肾—肝—代谢疾病的一体化管理。

GLP-1 四十余年的发展历程表明，基础科学发现只有经过机制解析、理性药物设计、严格临床验证和产业化转化，才能最终形成持续而广泛的患者获益与公共健康价值。未来，随着精准医学、系统药理学、结构生物学、人工智能及真实世界证据的进一步融合，GLP-1 研究将不断拓展 “从肠促胰素到多器官保护” 的科学内涵，并由代谢疾病治疗进一步迈向多器官综合管理，为复杂代谢性疾病及其他慢性疾病创新治疗模式的建立提供新的理论基础与实践范式。

作者简介

刘西平博士长期从事医学研究与创新药物开发，现任美国一家生物制药公司研发总监，在肿瘤、免疫及代谢性疾病等领域具有丰富的创新药研发经验。曾任美国华裔教授学者协会（USACSA）2024—2025年度理事会会长，现任协会顾问委员会主席。工作之余，热爱运动、旅行、园艺与写作，致力于以文字记录科学、生活与人生。

参考文献

1. Mojsov S, Weir GC, Habener JF. Insulinotropin: glucagon-like peptide I (7-37) co-encoded in the glucagon gene is a potent stimulator of insulin release in the perfused rat pancreas. J Clin Invest. 1987;79(2):616- 619. doi:10.1172/JCI112855.
2. Drucker DJ. The biology of incretin hormones. Cell Metab. 2006;3(3):153-165. doi:10.1016/j.cmet.2006.01.004.
3. Drucker DJ, Nauck MA. The incretin system: glucagon-like peptide-1 receptor agonists and dipeptidyl peptidase-4 inhibitors in type 2 diabetes. Lancet. 2006;368(9548):1696-1705. doi:10.1016/S0140- 6736(06)69705-5.
4. Drucker DJ, Habener JF, Holst JJ. Discovery, characterization, and clinical development of the glucagon- like peptides. J Clin Invest. 2017;127(12):4217-4227. doi:10.1172/JCI97233.
5. Drucker DJ. Mechanisms of action and therapeutic application of glucagon-like peptide-1. Cell Metab. 2018;27(4):740-756. doi:10.1016/j.cmet.2018.03.001.
6. Müller TD, Finan B, Bloom SR, D’Alessio D, Drucker DJ, Flatt PR, et al. Glucagon-like peptide 1 (GLP- 1). Mol Metab. 2019;30:72-130. doi:10.1016/j.molmet.2019.09.010.
7. Thorens B. Expression cloning of the pancreatic beta cell receptor for the gluco-incretin hormone glucagon-like peptide 1. Proc Natl Acad Sci U S A. 1992;89(18):8641-8645. doi:10.1073/pnas.89.18.8641.
8. Marso SP, Daniels GH, Brown-Frandsen K, Kristensen P, Mann JFE, Nauck MA, et al. Liraglutide and cardiovascular outcomes in type 2 diabetes. N Engl J Med. 2016;375(4):311-322. doi:10.1056/NEJMoa1603827.
9. Marso SP, Bain SC, Consoli A, Eliaschewitz FG, Jódar E, Leiter LA, et al. Semaglutide and cardiovascular outcomes in patients with type 2 diabetes. N Engl J Med. 2016;375(19):1834-1844. doi:10.1056/NEJMoa1607141.
10. Gerstein HC, Colhoun HM, Dagenais GR, Diaz R, Lakshmanan M, Pais P, et al. Dulaglutide and cardiovascular outcomes in type 2 diabetes: a double-blind, randomised placebo-controlled trial. Lancet. 2019;394(10193):121-130. doi:10.1016/S0140-6736(19)31149-3.
11. Wilding JPH, Batterham RL, Calanna S, Davies M, Van Gaal LF, Lingvay I, et al. Once-weekly semaglutide in adults with overweight or obesity. N Engl J Med. 2021;384(11):989-1002. doi:10.1056/NEJMoa2032183.

12. Perkovic V, Tuttle KR, Rossing P, Mahaffey KW, Mann JFE, Bakris G, et al. Effects of semaglutide on chronic kidney disease in patients with type 2 diabetes. N Engl J Med. 2024;391(2):109-121. doi:10.1056/NEJMoa2403347.
13. Lincoff AM, Brown-Frandsen K, Colhoun HM, Deanfield J, Emerson SS, Esbjerg S, et al. Semaglutide and cardiovascular outcomes in obesity without diabetes. N Engl J Med. 2023;389(24):2221-2232. doi:10.1056/NEJMoa2307563.
14. Jastreboff AM, Aronne LJ, Ahmad NN, Wharton S, Connery L, Alves B, et al. Tirzepatide once weekly for the treatment of obesity. N Engl J Med. 2022;387(3):205-216. doi:10.1056/NEJMoa2206038.
15. Rosenstock J, Frias J, Jastreboff AM, Du Y, Lou J, Gurbuz S, et al. Retatrutide, a GIP, GLP-1 and glucagon receptor agonist, for people with type 2 diabetes: a randomised, double-blind, placebo- and active-controlled phase 2 trial. Lancet. 2023;402(10401):529-544. doi:10.1016/S0140-6736(23)01053-X.
16. Jastreboff AM, Kaplan LM, Frías JP, Wu Q, Du Y, Gurbuz S, et al. Triple-hormone-receptor agonist retatrutide for obesity: a phase 2 trial. N Engl J Med. 2023;389(6):514-526. doi:10.1056/NEJMoa2301972.
17. Drucker DJ. GLP-1 physiology informs the pharmacotherapy of obesity. Mol Metab. 2022;57:101351. doi:10.1016/j.molmet.2021.101351.
18. Loomba R, Hartman ML, Lawitz EJ, Vuppalanchi R, Boursier J, Bugianesi E, et al. Tirzepatide for metabolic dysfunction-associated steatohepatitis with liver fibrosis. N Engl J Med. 2024;391(4):299-310. doi:10.1056/NEJMoa2401943.
19. Zhang Y, Sun B, Feng D, Hu H, Chu M, Qu Q, et al. Cryo-EM structure of the activated GLP-1 receptor in complex with a G protein. Nature. 2017;546(7657):248-253. doi:10.1038/nature22394.
20. Jones B, Buenaventura T, Kanda N, Chabosseau P, Owen BM, Scott R, et al. Targeting GLP-1 receptor trafficking to improve agonist efficacy. Nat Commun. 2018;9(1):1602. doi:10.1038/s41467-018-03941-2.
21. Drucker DJ. Efficacy and safety of GLP-1 medicines for type 2 diabetes and obesity. Diabetes Care. 2024;47(11):1873-1888. doi:10.2337/dci24-0003.
22. Drucker DJ. GLP-1-based therapies for diabetes, obesity and beyond. Nat Rev Drug Discov. 2025;24(8):631-650. doi:10.1038/s41573-025-01183-8.
23. Hernandez AF, Green JB, Janmohamed S, D'Agostino RB Sr, Granger CB, Jones NP, et al. Albiglutide and cardiovascular outcomes in patients with type 2 diabetes and cardiovascular disease: a double-blind, randomised placebo-controlled trial. Lancet. 2018;392(10157):1519-1529. doi:10.1016/S0140- 6736(18)32261-X.
24. Gerstein HC, Sattar N, Rosenstock J, Ramasundarahettige C, Pratley R, Lopes RD, et al. Cardiovascular and renal outcomes with efpeglenatide in type 2 diabetes. N Engl J Med. 2021;385(10):896-907. doi:10.1056/NEJMoa2108269.
25. McGuire DK, Marx N, Mulvagh SL, Deanfield JE, Inzucchi SE, Pop-Busui R, et al. Oral semaglutide and cardiovascular outcomes in high-risk type 2 diabetes. N Engl J Med. 2025;392(20):2001-2012. doi:10.1056/NEJMoa2501006.
26. Kosiborod MN, Abildstrøm SZ, Borlaug BA, Butler J, Rasmussen S, Davies M, et al. Semaglutide in patients with heart failure with preserved ejection fraction and obesity. N Engl J Med. 2023;389(12):1069- 1084. doi:10.1056/NEJMoa2306963.
27. Kosiborod MN, Petrie MC, Borlaug BA, Butler J, Rasmussen S, Verma S, et al. Semaglutide in patients with obesity-related heart failure and type 2 diabetes. N Engl J Med. 2024;390(15):1394-1407. doi:10.1056/NEJMoa2313917.
28. Sanyal AJ, Newsome PN, Kliers I, Østergaard LH, Long MT, Kjær MS, et al. Phase 3 trial of semaglutide in metabolic dysfunction-associated steatohepatitis. N Engl J Med. 2025;392(21):2089-2099. doi:10.1056/NEJMoa2413258.
29. Malhotra A, Grunstein RR, Fietze I, Hedner J, McEvoy RD, Pack AI, et al. Tirzepatide for the treatment of obstructive sleep apnea and obesity. N Engl J Med. 2024;391(13):1193-1205. doi:10.1056/NEJMoa2404881.
30. Bliddal H, Bays H, Czernichow S, Uddén Hemmingsson J, Hjelmesæth J, Hoffmann Morville T, et al. Once-weekly semaglutide in persons with obesity and knee osteoarthritis. N Engl J Med. 2024;391(17):1573-1583. doi:10.1056/NEJMoa2403664.
31. Cummings JL, Atri A, Sano M, Zetterberg H, Scheltens P, Knop FK, et al. Efficacy and safety of oral semaglutide 14 mg (flexible dose) in early-stage symptomatic Alzheimer's disease (EVOKE and EVOKE Plus): two phase 3, randomised, placebo-controlled trials. Lancet. 2026;407(10544):2167-2179. doi:10.1016/S0140-6736(26)00459-9.
32. Meissner WG, Remy P, Giordana C, Maltête D, Derkinderen P, Houéto JL, et al. Trial of lixisenatide in early Parkinson's disease. N Engl J Med. 2024;390(13):1176-1185. doi:10.1056/NEJMoa2312323.

33. Vijiaratnam N, Girges C, Auld G, McComish R, King A, Skene SS, et al. Exenatide once a week versus placebo as a potential disease-modifying treatment for people with Parkinson's disease in the UK: a phase 3, multicentre, double-blind, parallel-group, randomised, placebo-controlled trial. Lancet. 2025;405(10479):627-636. doi:10.1016/S0140-6736(24)02808-3.
34. Garvey WT, Blüher M, Osorto Contreras CK, Davies MJ, Lehmann EW, Pietiläinen KH, et al. Coadministered cagrilintide and semaglutide in adults with overweight or obesity. N Engl J Med. 2025;393(7):635-647. doi:10.1056/NEJMoa2502081.
35. Davies MJ, Bajaj HS, Broholm C, Eliasen A, Garvey WT, le Roux CW, et al. Cagrilintide-semaglutide in adults with overweight or obesity and type 2 diabetes. N Engl J Med. 2025;393(7):648-659. doi:10.1056/NEJMoa2502082.
36. Gasiorek A, Heydorn A, Gabery S, Hjerpsted JB, Kirkeby K, Kruse T, et al. Safety, tolerability, pharmacokinetics, and pharmacodynamics of the first-in-class GLP-1 and amylin receptor agonist, amycretin: a first-in-human, phase 1, double-blind, randomised, placebo-controlled trial. Lancet. 2025;406(10499):135-148. doi:10.1016/S0140-6736(25)01176-6.
37. Jastreboff AM, Ryan DH, Bays HE, Ebeling PR, Mackowski MG, Philipose N, et al. Once-monthly maridebart cafraglutide for the treatment of obesity: a phase 2 trial. N Engl J Med. 2025;393(9):843-857. doi:10.1056/NEJMoa2504214.
38. Wharton S, Aronne LJ, Stefanski A, Alfaris NF, Ciudin A, Yokote K, et al. Orforglipron, an oral small- molecule GLP-1 receptor agonist for obesity treatment. N Engl J Med. 2025;393(18):1796-1806. doi:10.1056/NEJMoa2511774.
39. Rosenstock J, Hsia S, Nevarez Ruiz L, Eyde S, Cox D, Wu WS, et al. Orforglipron, an oral small-molecule GLP-1 receptor agonist, in early type 2 diabetes. N Engl J Med. 2025;393(11):1065-1076. doi:10.1056/NEJMoa2505669.
40. US Food and Drug Administration. Foundayo (orforglipron) tablets, for oral use: prescribing information. Silver Spring (MD): US Food and Drug Administration; 2026.

Correspondence Author

Xiping Liu, Ph.D.
NantWorks, Culver City, California, USA
E-mail: xiping@yahoo.com

组织工程技术与再生医学：回顾、进展与未来展望

Tissue Engineering Technology and Regenerative Medicine: Review, Progress, and Future Prospects

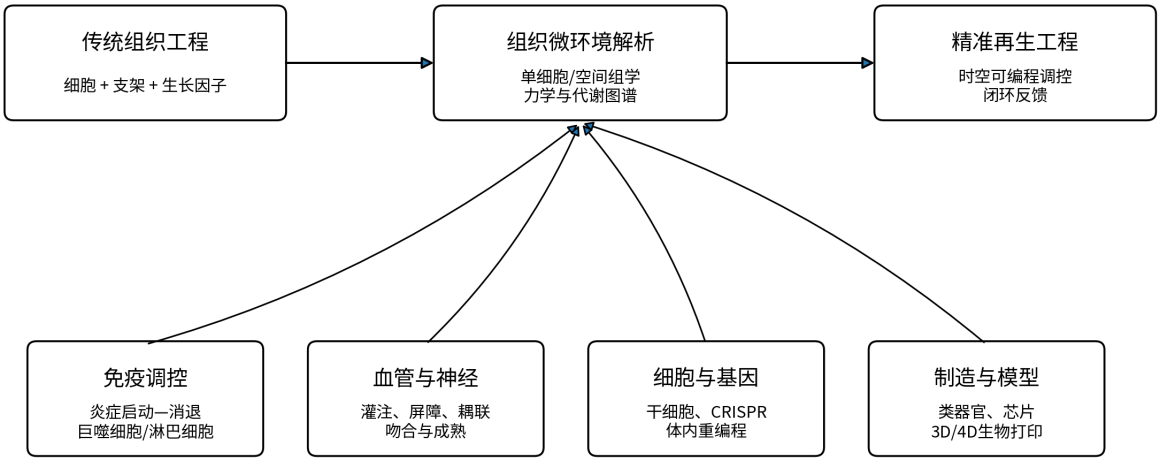
蒋文森 / Wensen Jiang

摘要

组织工程与再生医学旨在恢复、维持或改善受损组织和器官的结构与功能。过去三十余年，该领域以“细胞—支架—生物活性因子”三要素为核心，形成了皮肤、软骨、骨、角膜及血管替代物等一批临床产品。然而，单纯将细胞与材料混合并不能自动重建天然组织复杂的免疫、血管、神经、力学和代谢网络。近年的技术主线正由静态构建转向对再生过程的时空精准调控：利用免疫调控材料塑造炎症消退环境，以可灌注血管网络突破厚组织尺度限制，结合干细胞、基因编辑和体内基因递送修复细胞内在缺陷，并通过类器官、器官芯片、3D/4D生物打印及空间多组学实现患者特异性设计和闭环评价。本文首先梳理组织工程、组织修复与组织再生的概念及历史脉络，继而总结经典材料体系，重点评述免疫调控生物材料、血管化组织工程、干细胞与基因疗法、类器官及3D生物打印的最新进展、关键瓶颈和临床转化路径。总体而言，未来竞争焦点不再是“能否制造一个组织样结构”，而是能否以可制造、可质控、可监管的方式，在正确时间和空间协调多细胞生态与宿主反应，实现可预测、可重复和长期稳定的功能再生。

关键词：组织工程；再生医学；免疫调控生物材料；血管化；干细胞；基因编辑；类器官.
3D生物打印

图1 组织工程从“构建替代物”向“精准调控再生过程”演进的概念框架



目标：在正确的时间、空间和剂量下，协调免疫—血管—细胞—基质，实现结构与功能同步恢复

图 1. 组织工程由传统三要素构建向多尺度、时空精准调控演进。

1. 背景介绍

1.1 组织工程、组织修复与组织再生

组织工程通常被定义为综合运用工程学与生命科学原理，开发能够恢复、维持或改善组织功能的生物替代物[1]。其典型范式包括可扩增的细胞来源、具有适宜结构和降解特性的支架，以及调控细胞命运的生化或生物物理信号。组织修复强调损伤后的结构封闭和功能代偿，常伴随瘢痕形成，未必恢复原有组织的细胞组成和空间架构；组织再生则要求重建接近生理状态的组织结构、细胞多样性、血管神经连接及长期功能。临床上，快速修复可挽救生命，而高质量再生决定远期功能、复发率和生活质量。

再生医学的范围比传统组织工程更广，除支架与细胞外，还包括细胞治疗、基因治疗、外泌体与分泌组、体内重编程以及调动宿主内源性修复能力的材料和药物。近年来，“再生”被逐步理解为一个受免疫系统、血管系统、细胞外基质、神经支配、力学负荷和代谢状态共同控制的动态过程。因此，评价终点也从细胞存活和组织填充，转向灌注、神经连接、屏障功能、力学整合、免疫稳态及长期安全性。

1.2 关键奠基者及其科学贡献

Joseph P. Vacanti与Robert Langer在20世纪80—90年代系统提出并推广“细胞种植于可降解支架形成新组织”的工程化框架，1993年发表于Science的综述被普遍视为现代组织工程学科形成的标志[1]。Langer进一步推动可控药物释放、可降解聚合物和材料界面调控，使支架从被动承载结构转变为可递送信号并指导细胞行为的平台。Vacanti团队则通过软骨、骨及器官样结构研究验证了细胞—支架策略的可行性。与此同时，Anthony Atala等推动组织工程膀胱及泌尿组织的临床探索[2]；Shinya Yamanaka通过诱导多能干细胞技术解决患者特异性细胞来源问题[3]；Jennifer Doudna、Emmanuelle Charpentier及Feng Zhang等推动CRISPR基因编辑，使精准修复遗传缺陷和定向调控细胞命运成为可能[4,5]。这些贡献共同构成了从材料、细胞到基因层

面的技术基础。

其临床意义在于：第一，替代供体组织不足，尤其适用于大面积皮肤、软骨、骨和角膜等结构相对明确的组织；第二，降低长期免疫抑制依赖；第三，为罕见病、遗传病及个体化药物评价提供患者来源模型；第四，将治疗目标从“植入一个永久性器械”转向“诱导宿主形成自己的功能组织”。但目前获批产品仍主要集中于薄层、无血管或力学要求相对可控的组织，复杂实质器官的完整替代尚未实现[6]。

1.3 范式转变：从细胞与支架混合到精准调控

传统组织工程的基本假设是：只要将适宜细胞置于仿生支架中，并提供若干生长因子，细胞便会自组织形成目标组织。实践表明，这一策略在毫米尺度或简单组织中有效，但在厚组织和慢性病背景下常受到细胞异质性、中心缺氧、炎症持续、材料异物反应、血管不成熟和功能整合不足的限制。未来范式更接近“再生控制工程”：先解析损伤组织中不同细胞亚群和信号通路的时空变化，再通过响应性材料、可编程细胞、局部基因递送和生物制造技术，在特定窗口依次启动炎症、血管生成、基质重塑和组织成熟。空间转录组、单细胞多组学、机器学习与计算模型将帮助识别关键调控节点，并用于患者分层和产品放行标准设计[7]。

2. 组织工程材料及其应用

支架材料的核心功能包括提供细胞黏附位点和三维空间、传递力学信号、允许营养物质交换，并以可控速率降解和被新生基质取代。经典材料可分为天然高分子、合成聚合物、无机材料及脱细胞细胞外基质。天然材料具有优良生物活性，但批间差异、力学强度和病原风险需严格控制；合成材料易于规模化和参数调节，但常缺乏特异性细胞信号；无机材料适合骨和牙组织；脱细胞基质保留组织特异性成分和微结构，但去细胞充分性、残留DNA、灭菌和来源一致性是关键质量属性。

材料类别	代表材料	主要优势	主要局限	典型应用
天然高分子	胶原，明胶，纤维蛋白，透明质酸，海藻酸盐，壳聚糖	细胞相容性和生物活性高；可被细胞重塑	力学弱，降解和批间差异较大	皮肤，软骨，神经，创面水凝胶，生物墨水
合成可降解聚合物	PLA、PGA、PLGA、PCL、PEG	组成、孔隙、力学和降解可设计；易规模化	缺少天然配体；酸性降解产物或疏水性可能引发炎症	骨、软骨、血管、神经导管、药物控释
无机与复合材料	骨传导性、矿化能力和压缩性能较好		脆性、降解慢，复杂形状加工困难	骨缺损、牙槽骨、骨软骨界面
脱细胞基质	皮肤、心肌、肝、软骨、脂肪dECM	保留组织特异性ECM和生长因子结合位点	来源和去细胞工艺差异；残留免疫原性	补片、注射水凝胶、器官支架、生物墨水
微/纳米结构材料	电纺纤维、微球、微凝胶、纳米颗粒	高比表面积；可构建梯度和时序递送	制造与表征复杂；潜在颗粒迁移	肌腱、神经、免疫调控、局部递送

表 1 经典组织工程材料的性能特征与代表性应用。

材料设计已由单一化学组成扩展到多尺度结构控制。孔径、孔连通性和纤维取向影响细胞迁移与血管侵入；基质刚度、黏弹性和应力松弛决定干细胞分化和组织形态；可降解交联键则调节细胞介导的重塑速度。模块化微凝胶支架、可注射多孔水凝胶和原位交联材料可降低手术创伤，并改善不规则缺损填充[8]。临床转化时应避免“参数越仿生越好”的简单逻辑，而需在生物活性、可制造性、灭菌稳定性和质量控制之间取得平衡。

3. 免疫调控生物材料

3.1 概念与作用机制

任何植入材料都会触发蛋白吸附、补体激活、凝血、先天免疫细胞募集以及异物巨细胞形成。过去的目标是让材料“免疫惰性”，而现代免疫调控生物材料强调主动塑造有利于再生的免疫反应。其关键不是简单抑制炎症，而是重建炎症启动、病原清除、消退和组织重塑的正确时序。巨噬细胞M1/M2二分法虽便于描述，但体内表型呈连续谱；中性粒细胞、树突状细胞、调节性T细胞、B细胞及组织驻留免疫细胞也共同决定再生结局[9]。

免疫调控策略主要包括：通过材料化学和拓扑降低过度异物反应；递送IL-4、IL-10、TGF-β、促炎症消退脂质介质或免疫代谢调节剂；利用颗粒尺寸、表面电荷和配体实现髓系细胞靶向；通过可降解或刺激响应网络在炎症高峰期释放药物；以及递送抗原、mRNA或基因编辑工具，对免疫细胞进行局部编程。材料本身的刚度、形状、孔径及降解产物也可通过整合素、机械敏感离子通道、炎症小体和代谢通路改变免疫表型。

3.2 当前挑战与最新进展

首要挑战是免疫反应具有组织和疾病特异性。急性创伤需要短暂炎症以清除坏死组织，而糖尿病、衰老和自身免疫病中炎症长期不消退，统一的“抗炎材料”可能削弱宿主防御或延迟重塑。第二，动物模型与人类免疫系统差异显著，尤其是巨噬细胞标志物、补体和纤维化反应。第三，多成分材料中的活性因子释放曲线、空间分布和降解产物难以建立明确的剂量—效应关系。第四，长期免疫抑制、肿瘤风险和感染风险必须在大动物及真实疾病背景下验证。

近年进展集中于“动态和闭环”调控。例如，酶响应、活性氧响应或pH响应水凝胶可在炎症微环境中加速释放抗炎或促血管分子；可注射微凝胶组装体通过微米级孔隙促进宿主细胞快速浸润，并减少致密纤维包裹；具有特定糖基、肽配体或纳米结构的材料可偏向招募修复型髓系细胞或调节性T细胞。另一方面，单细胞和空间组学揭示了植入界面中此前被忽略的成纤维细胞和免疫细胞亚群，使材料评价从少数标志物转向细胞生态网络[7,10]。

3.3 临床应用与未来趋势

近期最可能进入临床的方向包括慢性创面、骨缺损、肌腱修复、神经导管、植入器械抗纤维化及细胞移植物保护。未来产品可能采用分阶段设计：早期释放抗菌和促清除信号，中期促进血管生成与免疫消退，后期增强基质成熟和力学整合。另一个趋势是将免疫调控材料与细胞疗法结合，使局部材料既保护细胞免受免疫攻击，又允许营养交换和功能分泌。监管上需将免疫学终点纳入关键质量属性，建立人体相关的体外免疫模型、补体试验和长期纤维化评价。

4. 血管化组织工程技术

4.1 概念与核心瓶颈

厚组织工程构建体若缺乏血管，氧和营养物质只能通过有限距离扩散，中心区域易发生缺氧、坏死和功能衰退。血管化组织工程不仅要求形成内皮细胞管腔，还要建立可灌注、抗渗漏、具有周细胞和平滑肌支持、能够与宿主循环快速吻合并长期维持的多尺度血管网络。不同组织还需要特异性内皮表型，如肝窦、肾小球、血脑屏障和骨髓窦状血管。

4.2 技术路线与最新进展

现有路线可分为体内血管化和体外预血管化。体内方法通过释放VEGF、FGF、PDGF等因子或设计促血管孔隙，引导宿主血管长入；优点是工艺相对简单，但速度慢且空间可控性有限。体外预血管化则将内皮细胞、周细胞、间充质细胞或器官特异性基质细胞共同培养，形成微血管网络后再移植。微流控芯片可施加流动剪切力促进内皮成熟；牺牲墨水打印、同轴打印和嵌入式打印可制造可灌注通道；去细胞器官支架则保留天然血管树，但再内皮化和微循环完整覆盖仍困难[11,12]。

近期的重要方向是多尺度整合。大直径通道负责即时灌流，微血管床负责细胞交换，二者需通过可吻合接口连接。内皮祖细胞与组织特异性基质细胞共分化可提高器官特异性；血管类器官和自组装微血管可作为“血管单元”嵌入组织；预制动静脉环、带蒂组织工程或外科显微吻合则为大体积移植物提供即时血供。血管化类器官、器官芯片和打印组织正在相互融合，以解决长期培养和移植后的灌注问题[13]。

4.3 临床前景与评价标准

近期临床应用更可能是血管化皮肤、骨、脂肪、肌肉和心肌补片，而非完整器官。评价应包括灌注量、血管密度之外的功能指标，例如血管通透性、血栓形成、屏障蛋白、周细胞覆盖、血流反应性和宿主吻合时间。未来还需处理血液相容性、内皮来源、肿瘤样异常血管生成及大规模制造问题。可用现成异体内皮细胞、诱导多能干细胞来源内皮细胞及低免疫原性细胞库，可能降低患者特异性制造成本。

5. 干细胞、基因编辑与基因疗法及其组织工程应用

5.1 干细胞来源与功能定位

间充质基质细胞因易获得、具有免疫调节和旁分泌作用，长期以来是组织工程中最常用的细胞之一。但其体内长期植入和直接分化贡献往往有限，疗效更多来自分泌因子、细胞外囊泡和免疫调控。胚胎干细胞与诱导多能干细胞可提供理论上无限的细胞来源，并分化为心肌、神经、视网膜、胰岛、肝细胞等多种谱系。iPSC还可建立患者特异性细胞库和疾病模型，但需严格控制残余未分化细胞、基因组稳定性、分化纯度和成熟度[3,14]。组织特异性成体干细胞，如表皮、肠道、角膜缘和造血干细胞，因谱系稳定和临床经验丰富，是目前成功转化较多的来源。

5.2 基因编辑和基因递送的耦合

CRISPR-Cas系统允许敲除致病基因、纠正点突变、插入保护性基因或调控细胞命运。碱基编辑和先导编辑可减少双链断裂，理论上提高精确性；CRISPRa/i则无需永久改变DNA序列即可调节转录。组织工程中的应用包

括：在植入前编辑自体干细胞以纠正遗传缺陷；敲除HLA或增强免疫逃逸以建立通用细胞；插入安全开关降低肿瘤风险；以及通过支架局部递送质粒、mRNA、病毒载体或核糖核蛋白，在体内重编程宿主细胞。材料载体可提高局部浓度、延长表达并减少系统暴露。

最新趋势是“材料辅助原位细胞工程”。例如，水凝胶或纳米颗粒在损伤区递送mRNA、siRNA或CRISPR组件，使成纤维细胞、免疫细胞或内皮细胞暂时获得促再生表型；电刺激、超声和光响应材料也可作为非侵入触发手段。与体外扩增细胞相比，原位策略可能简化制造，但面临递送效率、细胞特异性、脱靶、免疫原性和剂量控制等挑战。

5.3 临床转化风险与路径

临床上应区分“细胞产品”“基因治疗产品”和“组合产品”的监管路径。关键风险包括细胞异质性、克隆选择、脱靶编辑、染色体异常、插入突变、长期表达失控和生殖系暴露。对基因编辑细胞，需建立全基因组层面的脱靶与结构变异分析、长期肿瘤监测及可追溯的细胞库。近期更可行的场景包括表皮、角膜、造血系统和局部可切除组织；复杂实质器官则需要同时解决细胞成熟、血管化和功能整合。

6. 类器官技术

6.1 概念与价值

类器官是由干细胞或组织祖细胞在三维环境中自组织形成、能够重现部分器官细胞组成、空间结构和功能的微型组织。与二维培养相比，类器官更能保留细胞—细胞和细胞—基质相互作用；与动物模型相比，人源类器官在遗传背景和物种特异性上更接近患者。其应用包括发育研究、疾病建模、药物筛选、毒性评价、感染研究、精准医疗及潜在移植。

6.2 当前挑战与工程化进展

类器官的主要问题包括批间差异大、尺寸和形态不可控、缺乏成熟血管和免疫系统、长期培养后出现中心坏死、成熟度偏低，以及常依赖来源不明确的动物基质。工程化方法正在逐步解决这些问题：定义明确的合成水凝胶可独立调节黏附配体、刚度和降解；微流控器官芯片可建立流体剪切、浓度梯度和多器官连接；生物反应器提高传质和规模化；3D生物打印控制初始细胞组成和几何结构；血管共培养、血管类器官融合及宿主移植促进灌注和成熟[13,15]。

近年的突出趋势是从“自发形成的类器官”走向“工程化、可量化的类器官系统”。空间组学和活体成像用于比较类器官与真实组织的细胞谱系；机器学习从形态和多组学数据预测成熟度和药物反应；标准化冻存、自动化培养和多孔板流程提高通量。类器官与免疫细胞、神经元和微生物群共培养，使其更接近疾病微环境。

6.3 临床应用展望

类器官在近期最成熟的用途是伴随诊断和药物敏感性测试，例如肿瘤、囊性纤维化和遗传代谢病。作为移植物时，需要解决规模、血管化、功能成熟、残余多能细胞、基质来源和植入后可控性。肠道、肝、胆管、视网膜和胰岛类器官具有较强转化潜力，但真正进入常规治疗前仍需建立GMP培养、明确的放行指标和长期随访体系。

7. 3D生物打印技术

7.1 技术原理与主要模式

3D生物打印是按照数字模型，将细胞、生物材料和活性因子以空间可控方式逐层构建组织结构。挤出式打印设备简单、材料范围广，适合高细胞密度和大尺度构建，但分辨率有限且剪切力可能损伤细胞；喷墨打印速度快、成本低，但受黏度和液滴稳定性限制；激光辅助打印分辨率和细胞活性较高，但设备昂贵；光固化和数字光处理可快速形成高分辨率结构，但要求光敏材料和细胞相容的引发体系。嵌入式打印通过支撑浴打印柔软水凝胶，显著扩展了复杂形状和悬空结构的制造能力。

7.2 生物墨水与血管化难题

理想生物墨水需要兼顾可打印性、形状保持、细胞存活、组织特异性和可降解性，但这些指标常相互冲突。高黏度和高交联有利于打印精度，却限制细胞迁移和基质重塑；天然基质生物活性高，但机械性能和批次稳定性不足。当前策略包括双网络水凝胶、纳米复合材料、微凝胶颗粒墨水、脱细胞基质墨水和细胞高密度无支架打印。血管化仍是厚组织打印的核心瓶颈，常通过牺牲通道、同轴喷头、多材料打印和内皮自组装相结合来解决[11,16]。

图2 下一代再生医学的技术汇聚与闭环转化路径

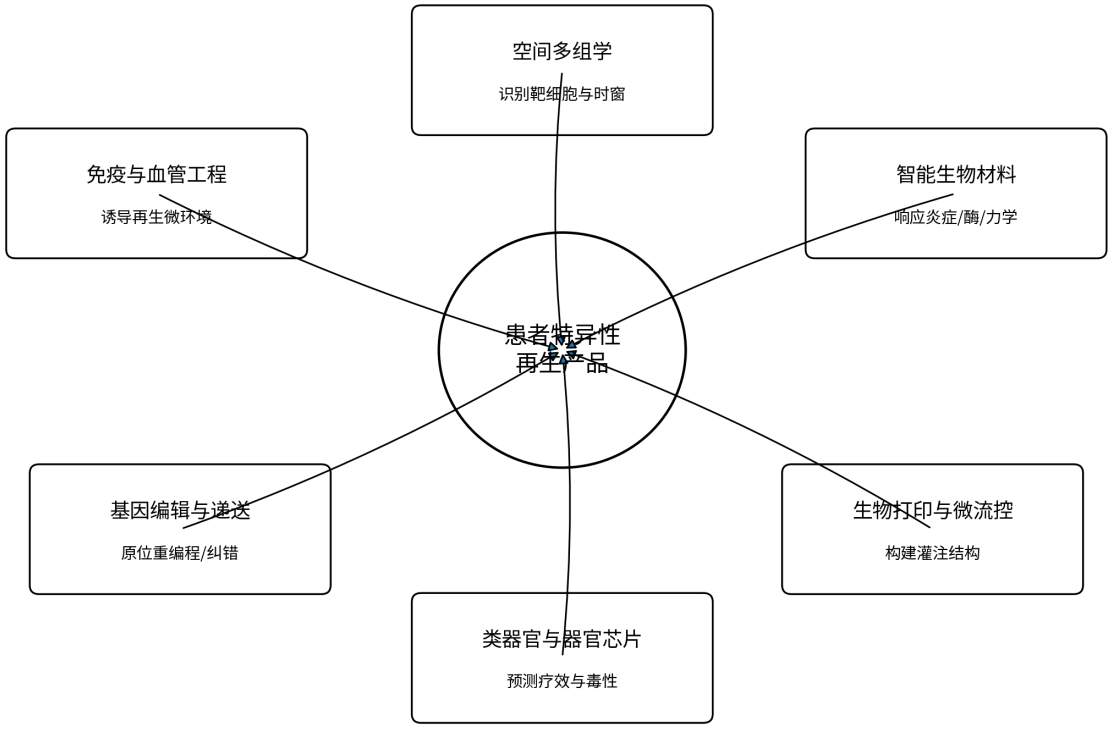


图 2. 多组学、材料、细胞与制造技术汇聚形成患者特异性再生医学闭环。

7.3 4D生物打印、原位打印与临床前景

4D生物打印在三维结构基础上引入时间维度，使构建体在温度、pH、光、磁场、酶或细胞牵引作用下发生形状和功能变化，可模拟组织生长、折叠和动态力学环境[17]。原位生物打印则直接在创面或手术床上沉积材料和细胞，有望用于皮肤、软骨和骨缺损，并减少体外培养和搬运。实时成像、机器人定位和闭环打印控制将提高其精度。

临床转化近期更可能聚焦于皮肤替代物、骨软骨植入物、药物筛选组织、肿瘤模型和个体化补片，而不是完整心、肝或肾。主要障碍包括打印速度、细胞来源、无菌封闭制造、在线质量监测、复杂产品放行以及植入后的长期稳定性。2026年在微重力环境开展多组织生物打印，提示重力可能影响细胞分布和柔软组织制造，但其临床价值仍需严格验证[18]。

8 综合总结与未来趋势

组织工程正在从“制造一个与组织外形相似的构建体”转向“控制损伤后复杂生物过程”。经典材料仍是基础，但下一代产品必须同时处理免疫反应、血管灌注、细胞命运、基质重塑和功能整合。免疫调控材料使炎症从需要回避的副作用变为可利用的再生驱动力；血管工程决定厚组织的尺度上限；干细胞和基因技术提高细胞来源的可获得性与可编程性；类器官和器官芯片提供人体相关的机制研究与患者分层平台；3D/4D生物打印则把这些生物学单元按照可设计的空间结构组装。

未来五至十年的关键趋势可概括为五点。第一，时空精准递送取代持续高剂量刺激，通过响应性材料和多阶段释放匹配炎症、血管生成和成熟窗口。第二，患者特异性与“现货型”产品并行发展：前者用于高价值个体化治疗，后者依赖通用细胞、标准化材料和自动制造以降低成本。第三，空间多组学、人工智能和数字孪生将用于靶点发现、工艺优化和疗效预测。第四，质量控制将从单一成分检测升级为功能性放行，包括灌注、屏障、收缩、传导、免疫调节和基因稳定性。第五，临床路径将由完整器官替代转向分阶段应用，优先发展可局部植入、可监测、可切除和风险可控的补片、微组织及原位再生产品。

真正决定临床成功的并非技术复杂度，而是能否形成清晰的作用机制、稳定的制造工艺、可量化的关键质量属性以及与临床需求一致的终点。只有当材料科学、发育生物学、免疫学、血管生物学、基因工程、制造科学和监管科学协同推进时，组织工程才可能从概念验证走向大规模、可重复的功能再生。

表2 主要前沿方向的关键挑战、近期突破与临床落点

方向	关键挑战	近期突破	较现实的临床落点
免疫调控材料	人群和疾病异质性；感染与纤维化风险	刺激响应释放、微凝胶多孔支架、空间免疫组学	慢性创面、骨修复、植入物抗纤维化
血管化工程	灌注、成熟、吻合和血栓	牺牲通道、微流控、血管类器官、多尺度打印	皮肤、骨、脂肪和心肌补片
干细胞/基因编辑	异质性、肿瘤和脱靶	iPSC标准化分化、碱基/先导编辑、原位mRNA递送	角膜、皮肤、血液及局部遗传病
类器官	成熟度、批间差异、缺乏免疫血管	定义基质、器官芯片、血管化与自动化	药敏测试、疾病模型、局部组织移植
3D/4D生物打印	生物墨水矛盾、速度与质控	嵌入式打印、微凝胶墨水、原位及动态打印	皮肤、骨软骨、药筛组织和个体化补片

表2 主要技术方向的转化瓶颈与近期可实现应用

参考文献

[1] Langer R, Vacanti JP. Tissue engineering. Science. 1993;260:920-926.

[2] Atala A, Bauer SB, Soker S, Yoo JJ, Retik AB. Tissue-engineered autologous bladders for patients needing cystoplasty. Lancet. 2006;367:1241-1246.

[3] Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell. 2006;126:663-676.

[4] Jinek M, Chylinski K, Fonfara I, et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337:816-821.

[5] Cong L, Ran FA, Cox D, et al. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013;339:819-823.

[6] Hoang VT, et al. Tissue Engineering and Regenerative Medicine. Adv Healthcare Mater. 2025.

[7] Chu LX, et al. Exploring molecular landscapes in aging and regenerative medicine with spatiotemporal multi-omics. 2024.

[8] Liu Z, et al. 3D biofabrication of microporous hydrogels for tissue engineering. 2025.

[9] Sadtler K, Singh A, Wolf MT, et al. Design, clinical translation and immunological response of biomaterials in regenerative medicine. Nat Rev Mater. 2016;1:16040.

[10] Zhang W, et al. Smart biomaterials as active immune modulators to shape pro-regenerative microenvironments. 2025.

[11] Chen EP, Toksoy Z, Davis BA, Geibel JP. 3D bioprinting of vascularized tissues for in vitro and in vivo applications. Front Bioeng Biotechnol. 2021;9:664188.

[12] Novosel EC, Kleinhans C, Kluger PJ. Vascularization is the key challenge in tissue engineering. Adv Drug Deliv Rev. 2011;63:300-311.

[13] Zhou R, et al. Vascularised organoids: recent advances and applications in tissue engineering and regenerative medicine. 2025.

[14] Mandai M, Watanabe A, Kurimoto Y, et al. Autologous induced stem-cell-derived retinal cells for macular degeneration. N Engl J Med. 2017;376:1038-1046.

[15] Nwokoye PN, et al. Bioengineering methods for vascularizing organoids. 2024.

[16] Li Z, et al. A review of 3D bioprinting for organoids. 2025.

[17] Lai J, et al. 4D bioprinting of programmed dynamic tissues. Trends Biotechnol. 2024.

[18] Reuters. Space station bioprinting experiment advances quest for lab-grown tissues. 2026-07-09.

[19] Place ES, Evans ND, Stevens MM. Complexity in biomaterials for tissue engineering. Nat Mater. 2009;8:457-470.

[20] Lutolf MP, Hubbell JA. Synthetic biomaterials as instructive extracellular microenvironments for morphogenesis in tissue engineering. Nat Biotechnol. 2005;23:47-55.

[21] Discher DE, Mooney DJ, Zandstra PW. Growth factors, matrices, and forces combine and control stem cells. Science. 2009;324:1673-1677.

[22] Murphy WL, McDevitt TC, Engler AJ. Materials as stem cell regulators. Nat Mater. 2014;13:547-557.
[23] Badylak SF, Freytes DO, Gilbert TW. Extracellular matrix as a biological scaffold material. Acta Biomater. 2009;5:1-13.
[24] Silva AKA, Richard C, Bessodes M, Scherman D, Merten OW. Growth factor delivery approaches in hydrogels. Biomacromolecules. 2009;10:9-18.
[25] Vegas AJ, Veiseh O, Gürtler M, et al. Long-term glycemic control using polymer-encapsulated human stem-cell-derived beta cells in immune-competent mice. Nat Med. 2016;22:306-311.
[26] Takebe T, Sekine K, Enomura M, et al. Vascularized and functional human liver from an iPSC-derived organ bud transplant. Nature. 2013;499:481-484.
[27] Huh D, Matthews BD, Mammoto A, et al. Reconstituting organ-level lung functions on a chip. Science. 2010;328:1662-1668.
[28] Lancaster MA, Knoblich JA. Organogenesis in a dish: modeling development and disease using organoid technologies. Science. 2014;345:1247125.
[29] Clevers H. Modeling development and disease with organoids. Cell. 2016;165:1586-1597.
[30] Murphy SV, Atala A. 3D bioprinting of tissues and organs. Nat Biotechnol. 2014;32:773-785.
[31] Daly AC, Prendergast ME, Hughes AJ, Burdick JA. Bioprinting for the biologist. Cell. 2021;184:18-32.
[32] Skylar-Scott MA, Uzel SGM, Nam LL, et al. Biomanufacturing of organ-specific tissues with high cellular density and embedded vascular channels. Sci Adv. 2019;5:eaaw2459.
[33] Grigoryan B, Paulsen SJ, Corbett DC, et al. Multivascular networks and functional intravascular topologies within biocompatible hydrogels. Science. 2019;364:458-464.
[34] Babb LM, et al. Immunomodulatory biomaterials for treating chronic mucosal inflammatory diseases. Nat Rev Bioeng. 2025.
[35] Li Y, et al. Immunomodulatory biomaterials for bone and soft tissue regeneration in chronic inflammation. 2025.
[36] Liao L, et al. Engineering strategies for the clinical translation of organoids. 2026.
[37] Ju Y, et al. Advancing multiscale vascular engineering for tissue regeneration. Chem Eng J. 2026.

Correspondence Author

Wensen Jiang, Ph.D.
E-mail: wjian005@ucr.edu

中医理论体系中关于“神”的概念

The concept of "Shen" within the theoretical system of Traditional Chinese Medicine.

杨常青 / Changqing Yang

精神医学博士，MD (Psychiatry)，加州执照中医师 *(California Licensed Acupuncturist)*

摘要

本文简述了中医神志理论中的最为重要概念“神”。本文将“神”还原于中国文化的源流之中，从文化和医理的角度极为简要地阐述“神”的几个基本内涵。中医文化源于中华文化，又有丰富的临床实践和几千年的历史积淀，因此，本文仅就显而易见的八个内涵进行一个简要地总结。

引言：

本文也是承接上次在贵刊发表的拙作“中医神志理论中神气的概念探讨”，再进一步进行一个关于中医“神”的讨论的拓展。和中医本身受到的中国文化影响一样，中医有关“神”的概念也是基于古老而又传统的中国文化为背景而产生的。理解神志首先要理解“神”，神本身不仅仅是一个中医的神志概念，而且，更重要的是，神是一个与文化相关的概念，所以，要先将神还原于中国文化背景之下来探讨一下神的几个概念，这样才能进一步对理解神志，以及与其相关的理论体系与实践有一个总体上的把握，以及可以在中国文化总体背景上形成较为全面的认知，这对从事神志病事业在方法论和认识论上都是有所裨益的。

讨论：

一，神是天神和宇宙的主宰

正如《说文》中所说的那样，神的第一个含义就是天神，即宇宙的主宰。在商代的甲骨文里，有极大量向天神问卜的卜辞。也就是《说文》中所言的“神，天神引出万物者也。”在《周礼·大司乐》中就有“冬日至，於地上之圜丘奏之，若乐六变，则天神皆降，可得而礼矣。”都说明了在中国古代，人们非常重视祭祀天神，认为神是宇宙的主宰。这也就是为什么在传统文化中，人们在谈及神的时候，就会产生敬畏之心。由此可知传统文化中的图腾文化对神的概念产生了重要影响。同时，天是至高无上，同于宇宙时空的概念。因此，

以天为背景产生万物，包括意识思维的主体和有形之物的主体。如《系辞》中有云：“天生神物。”《荀子·天论》中也有：“天职既立，天功既成，形具而神生”也表达了“天生神”的观点，因此，天神合用就是上天，上帝之意，这是神的首要概念。

二，神是自然天地及其规律

神也是用来表示自然本身和其规律的概念。如《天元纪大论》中言：“夫变化之为用也，在天为玄，在人为道，在地为化，化生五味，道生智，玄生神。”天是幽深玄远而为生神之所，这个天生之神就是宇宙自然及其变化规律。接着文中又曰：“神在天为风，在地为木”等五行天地与人体的对应关系，要注意文中是以神为首开始论述，就是以天地自然为起始逐步揭示宇宙自然规律，包括阴阳，子午流注，以及天六地五的天地周纪的天元纪概念，是宇宙自然的生命之机，所以，所以从文中不难看出“神”即是“道”本身。

在《荀子·礼论》中说：“列星随旋，明暗递炤，四时代御，阴阳大化，风雨博施，万物各得其和以生，各得其养以成，不见其事，而见其功，夫是之谓神。”在《礼记·祭法》中明言：“山陵川谷丘陵能出云为风雨，皆曰神”，说明古人认为神也是自然的代名词，甚至认为神寓于自然万物之中。

三，神是超乎自然能量和能力

正因为中医认为“在天为玄，玄生神”，这个神就有了超出人类想象的能力和能量。诸如天崩地裂，海啸风暴都是神的能力而同时又有着巨大的能量。再如《山海经》中关于四方神的论述就是“南方祝融，兽身人面，乘两龙；西方蓐收，左耳有蛇，乘两龙；东方有句芒，身鸟人面，乘两龙；北方禺疆，黑身手足，乘两龙。”还有图腾后的五方神以应五行的理论规律，所以认为是东为青龙，西为白虎，南为朱雀，北为玄武居於四方，能够司天运，定天命，能力无边。当然也有恶神，如民间传说中的五通神就有邪恶的能量和能力，也需要给予祭祀来求得平安的。

实际上，在中国传统文化中，神的能力和能量一直是受到尊崇的，从《封神榜》，《西游记》这样的古代小说中就可以看出人们对神的能力和能量的敬拜，而且，希望借助于这样的神力打败对手，或脱离险境达到目的。这种对神的能力和能量的尊崇，使得对一些历史上著名人物进行了神化，如医家扁鹊，哲学家老子，还有《三国演义》里面的武将关公，以及传统文化中门神等等，都是对一些古代和近代的名人进行了神化。一些名人的文化内涵已经不再是其肉身之人的本身，甚至已经完全成为神的属性，所以，传统上对神的敬拜，以及对神的能力和能量的尊崇就会产生这样的文化特点。

四，神是未认识的奥秘

在《易经》中有“一阴一阳之谓道”，“阴阳不测为之神”的系辞，其表明了宇宙自然的奥秘无穷，然而阴阳为其道，是其总的规律与认识原则。《孟子》也有“圣而不可知之谓神”的说法，都表明了神一词有玄奥难解和高深莫测的含义。中国古文化中以《易经》为据进行宇宙自然，以及社会人文的推理演绎，就是一种认识论上的揭秘方法，而《内经》将古文化的自然哲学体系应用于人体而建立的医学体系，也是对人类生命有意义的揭秘和探索。

西方人经常说的“只有上帝知道”就是对自己或人们无法理解的事物所用的一种托词，带有自身的宗教色彩，而中国人对不可理解的人事物认为“很神”就是有中国文化色彩的一种托词。孔子在《论语》中的“樊迟问知，子曰：务民之义，敬鬼神而远之，可谓知矣。”孔子的回答有着较为深奥的哲学内涵，服务民众，对神秘奥妙的鬼神是“敬”而“远”之。因此，不难理解一些人认为对宇宙和生命等奥秘的探索，实际上也是对神的奥秘的贴近。科学追求实证，而在文化观念上留下来的影响依旧还是根深蒂固的。

五，神是最高的赞美和赞颂

神还有最高的赞美和赞颂之意。如在中医历史上出现的名医扁鹊，淳于意等都被尊为神医。伟大的工匠被称为神工，极佳的写作被称作神笔。其他如神机妙算，神武雄才等都是赞美之词。在中医自身的医疗活动中，好的针法可以称之为神针，好的处方用药也可以称之为神药神方而有神效，就是赞美好的方药可以出神入化而效果立现。再如《张衡传》中对其发明的地动仪认为是“驗之以事，合契若神”，这是最高的赞美之词。

“神”一词在传统文化中的赞美之意是非常普遍的，在一些诊所和医疗设施内的匾额以及荣誉表彰的内容里经常出现神的字样，如“神医在世”，神方神药，神来之笔等等，这是最高的赞美之词,
这也是神赞美之意文化的延续。实际上，神的最高赞美含义的根源仍然是传统文化中对神的崇拜，认为神是至高无上的宇宙主宰而具有超乎想象的能力。

六，神是内在的神志与形相对

神在中医学里的重要含义是指人的意识本身，即人的所有心理功能活动。如五脏藏神的概念，而且，中医特别强调了心是藏神之臟而为“君主之官”，这些都是中医神志学说中本神理论的核心内容。中医学认为神的功能包括了所有今天心理学所认识到的感知觉，意识，认知，思维，记忆，意志以及情感等心理过程。

在以《内经》为指导的中医理论体系中，精神是“精”与“神”这两个方面的含义，是人神志与肉体的对应，故也有”形神”之论。人的神，即内里的各种神志心理功能活动，可以是一个协调的统一体而表现出个人的某种类型，即受先天的影响，也有后天养生的干预，因此，才会有通天性格的阴阳属性和五行人格的特征。《八正神明论》中认为神是“若风吹云故曰神”就是对神的存在属性的精确描述，是与肉体相对应的。要注意中医认为精为形体有形之物，而神为灵魂是无形的存在，然而皆来自于自然。男女媾和，两神相感，两精相合，元神入心，五神来居，生命诞生，所以，精是神之居所，故《内经》称之为“五脏藏神”。所以《天年》篇在论述人的生命诞生的理论时认为：“血气已和，荣卫已通，五藏已成，神气舍心，魂魄毕具，乃成为人”。

七，神是生命本身及其活动

中医学认为，神是人的生命。正如《移精变气论》中指出的那样：“得神者昌，失神者亡。”这里的得神与失神是中医生命观中的核心概念，是能够维持生命活动的关键。当人的健康状态良好或由恶转好的情况都可以称之为得神，反之，健康状态不佳或是由好转恶的情况都可以称之为失神。

在《病传》篇中所说的“生神之理”就是生养人生命的道理，这个神就是指的生命，所以才有下文疾病传变的死生之理。同时，神也是生命力，因为神的能量是与气结合在一起的，故神气的活动就可以使神显露于外，就可以通过望闻问切诊断出来。所以，人的得神与失神状态就可以通过人的外表气色和生命活动表现出来，从而，对人的生命力及其活动有一个正确的判断而有助于养生和治疗。

八，神是鬼神迷信而为医家所排斥

在《五脏别论》中认为在诊治疾病时“拘于鬼神者，不可与言至德”正是说明了医术与迷信鬼神的区别。《内经》中的医学理论体系是客观的和实际的，从医巫不分到医巫各行其术，各有其专，甚至还有相互的争执。尽管《内经》中有象鬼臾区这样的巫师大家，但是，《内经》的理论显然是主张信医不信巫的。这也是司马迁在《史记》书里的《扁鹊传》中概述的扁鹊与巫师对立的观点，认为：“病有六不治，骄恣不论于理，一不治也；轻身重财，二不治也；衣食不能适，三不治也；阴阳并，脏气不定，四不治也；形羸不能服药，五不治也；信巫不信医，六不治也。”中医秉承了这个思想体系，使中医的临床治疗能够客观而且实际。

在《贼风》篇中有言：“黄帝曰：今夫子之所言者，皆病人之所自知也。其毋所遇邪气，又毋怵惕之所志，卒然而病者，其故何也？唯有因鬼神之事乎？岐伯曰：此亦有故邪留而未发，因而志有所恶，及有所慕，血气内乱，两气相搏。其所从来者微，视之不见，听而不闻，故似鬼神。”所以，岐伯的回答是基于患者临床实际的神志心理反应，即志有所恶，是人的性情好恶所为，而并非是鬼神所导致的。由此可以看出，中医是以古代的自然观，生命观，疾病观等理论和哲学思想为基础，是合乎天地之道的，即自然规律的医学体系。

Correspondence Author

Changqing Yang, MD
E-mail: yangcq2002@yahoo.com

利用断层带陷波与远震接收函数揭示深部断层损伤带

Fault Damage Zone at Depth Depicted by FZ Trapped Waves and Receiver Functions from Teleseismograms

郦永刚 / Yong-Gang Li

南加州大学、加州地震中心
University of Southern California, California State-Wide Earthquake Center

ABSTRACT

We analyzed the teleseismograms recorded at a seismic nodal-array consisting of 100 stations deployed atop the Calico Fault (CF) in Mojave Desert, California, and identified the fault-zone trapped waves (FZTWs) characterized by large amplitudes at 0.5-2 Hz in ~3.5-4 s and ~6-8 s durations, respectively, following the transmitting P' wave and P-to-S converted S' wave from the Moho discontinuity at stations within the ~1-km-wide CF compliant zone. When the P' and S' waves enter the bottom of the low-velocity waveguide (LVWG) formed by the CF compliant zone at seismogentic depths, FZTWs are produced by constrictive interference of reflective waves within the LVWG. The receiver functions computed by deconvolution between the radial- and vertical-component teleseismograms confirm these newly identified FZTWs arising from the P' and S' waves. This new type of FZTWs has not been used before, but they appear to have great promise for diagnosing the depth extension of fault damage zone at seismogenic depths where the coverage of seismic waves from explosions and local earthquakes is poor. Our 3D finite-difference simulations of these FZTWs show that the CF compliant zone with the maximum ~40-50% velocity reduction in the fault core likely extends to ~8-km depth. Accurately extracting the fault damage zone structure at seismograms helps us in evaluating earthquake physics and predicting hazards during earthquakes. This article provides the details of Author's presentation at the Seismological Society of America (SSA) 2026 annual meeting.

Keywards: teleseismogram; fault compliant zone; low-velocity waveguide; fault-zone trapped wave; Moho discontinuity; P-to-S converted wave, receiver function.

INTRODUCTION

The Calico fault (CF) is located midway between the 1992 Mw7.3 Landers and 1999 Mw7.1 Hector Mine ruptures (Fig. 1a) on which fault-zone trapped waves (FZTWs) generated by explosions and aftershocks were recorded

(Li et al.,1994a, 2002; Cochran et al, 2009). The FZTWs could be produced by coherent interference of waves reflected at the boundary between the low-velocity fault zone and high-velocity surrounding rocks (Li et al,1990; Li and Leary, 1990). Coseismic interferograms for these earthquakes also showed strain localized on the CF and other nearby faults (Fialko et al., 2002; Fialko, 2004). Line-of-sight displacements are clearly associated with the CF (Fig. 1b). Geodetic estimates suggest that the CF has as much as 7 mm/yr dextral slip rates within the eastern California shear zone (Peltzer et al., 2001). The trench data indicate that the last earthquake occurred on the CF at least several hundred years ago. A team of academic researchers from University of Southern California, University (USC), Uinersity of California at Los Angels (UCLA) and University of California at San Diego (UCSD) installed a square array of 100 seismographs atop the CF (Fig. 1c) in Mojave Desert in 2006. In the previous study, a 2-km-wide low-velocity compliant zone along the CF was imaged by FZTWs generated by explosions and local earthquakes, traveltimes inverse for local and teleseismic earthquakes, and InSAR observations (Cochran, et al., 2009). Seismic velocities are reduced to the maximum of ~40-50% at the center of a ~1-km-wide fault core damage zone, which extends to 5-6 km depth, and becomes narrower with a lower velocity contrast at greater depths. It indicates that faults can affect rock properties at substantial distances from the primary fault slip surfaces and throughout much of the seismogenic depth, implying that the portion of energy expended during rupture to drive cracking and the development of fault systems.

The 3-D structure and material properties of the fault damage zone at seismogenic depth will influence rupture propagation and may offer a basis for such elusive source parameters as the "asperities", "barriers" and the upper bound frequency “fmax” of seismic waves that can be radiated from a fault zone in sake for the strong ground motion prediction (Aki, 1979, 1984; Papageorgiou, 2003). Brune (1970) introduced a cutoff frequency of about 10 Hz in the acceleration spectrum in accordance with the general appearance of observed strong motion records. Ida (1973) interpreted the cutoff frequency as the source effect and estimated the critical slip in his slip-weakening model to be of the order of 10 cm. Papageorgiou and Aki (1983) attributed the cutoff frequency to the size of cohesive zone (breakdown zone) acting as a spatial smoothing operator on fault slip. This range of the “fmax” may correspond to the range of cohesive zone size from 50 to >200 m. Ferry et al. (2025) found in their numerical modeling that the co-seismic off-fault damage zone, while narrowing with depth, becomes denser to act as an energy sink, significantly influencing rupture dynamics by stabilizing slip rates. The damage formation markedly reduces rupture velocity and delays, or prevents the transition to super-shear speeds even for a narrow damage zone. Thus, permanent damage zones at seismic depth plays a critical role in development and dynamics of faults, intensity of ground motion, and earthquake hazards along the fault zone due to its waveguide effect. However, the resolution to image the deep portion of fault damage zone is usually less constrained at seismogenic depths where the ray-path coverage of seismic waves from artificial sources and local earthquakes is poor,

In order to accurately extracting the high-resolution image of the CF damage structure at seismogenic depths we innovatively use the new type of FZTWs identified in teleseismograms recorded at this seismic array atop the CF compliant zone. These FZTWs have not been used before, but appear to have great promise for providing unprecedented constrains of the depth extension of fault damage zone because they arise from seismic waves that unprecedented constrains of the depth extension of fault damage zone because they arise from seismic waves that enter the bottom of fault damage zone at deep level. These FZTWs are further confirmed by receiver-functions computed from teleseismograms. We interpret the large amplitudes closely following the transmitting P-wave and converted S-wave in receiver functions owing to FZTWs which are produced within the LVWG of the CF compliant zone when body waves coming from the Moho enter its bottom. Combined with the previous velocity model of the CF (Cochran et al., 2009), he results from observations and simulations of these FZTWs reveal that the compliant zone with velocity reduction of 40-50% along the CF likely extends to ~8-km depth.

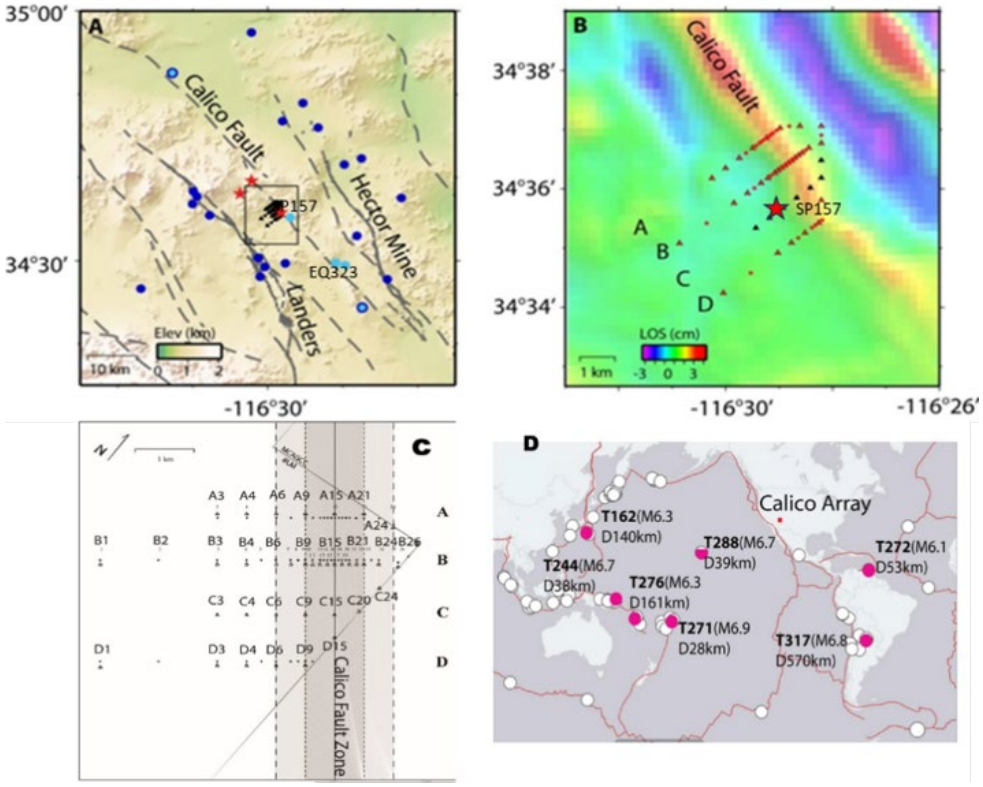


Figure 1. (a) Map of Mojave Desert in S. California shows locations of earthquakes (blue circles), explosions (red stars) and seismic array stations at the CF in the square (after Cochran et al., 2009) which outlines region in (b) Colors denote variations in line of sight (LOS) displacements caused by Hecter Mine M7.1 earthquake on 10/16/1999 that spans 1/13-10/20/1999 (after Fialko et al., 2002). Black triangles and red circles are seismic stations with 40T and L22 sensors of lines A, B, C and D in (c) across the 2-km-wide CF compliant zone (grey color) with the 1-km-wide fault core zone (dark grey). (d) Map shows locations of M≥6 teleseismic earthquakes recorded at the CF array. Red circles labeled with T followed by Julian date, M by magnitude and D by depth, denote teleseismic earthquakes are used in.

THE DATA AND WAVEFORM ANALYSIS

A dense array of 40 intermediate-period (40T 0.025 Hz) stations and 60 short-period (L22 2Hz) stations was installed in a 1.5 km × 5.5 km square adjacent to the Calico fault (CF) for six months starting from June in 2006 to record seismic waves from shots and earthquakes. Station spacings in Line A, Line B, Line C and Line D of this square array are 62.5, 125, 250, 500 m or 1 km unevenly. Each station has a L22 and/or a 40T sensor buried in a 0.5-1-m deep dig hole with three components in the vertical, parallel and perpendicular to the fault strike. The seismometers worked in continuous mode at 100 samples per second, and were synchronized by internal GPS clocks. The data recorded for 3 explosions and 5 local on-fault earthquakes have been used for FZTWs analysis, and 20 local and 8 teleseismic earthquakes used for travel time tomography, showing a ~2-km-wide with ~1-km-wide core damage zone along the CF extending to 5-6 km depth, within which seismic velocities are reduced by up to ~40-50% and shear moduli reduced by -55-70% from wall-rock (Cochran et al., 2009).

In the present study, we examined the teleseismograms from 26 M≥6 teleseismic earthquakes occurring at distances of ~40º-80º great circles to the CF array site during 2006 (Fig. 1d). We identify this type of FZTWs with large amplitudes and long durations of ~3.5-4 s and ~6-8 s, respectively, following the transmitting P-wave and P-to-S converted wave

from the Moho discontinuity arriving at stations within the ~1-km-wide CF core damage zone. Table 1 shows magnitudes and locations of seven teleseismic earthquakes named by TE followed by the Julian date of the event, from which the teleseismograms are exhibited and analyzed in this paper.

Table 1. Teleseismic earthquakes recorded at the seismic array atop the Calico fault

Event ID	Type	Magnitude	Latitude	Longitude	Depth (km)	Distance
TE162	Teleseism	6.3	33.130	131.150	140.0	72.2°
TE244	Teleseism	6.7	-6.780	155.540	38.0	69.5°
TE271	Teleseism	6.9	-16.613	-172.035	28.0	73.5°
TE272	Teleseism	6.1	10.910	-61.650	53.0	55.0°
TE276	Teleseism	6.3	-18.860	169.090	161.0	65.7°
TE288	Teleseism	6.7	19.878	-155.935	38.9	37.8°
TE317	Teleseism	6.8	-26.038	-63.243	547.5	78.9°

Teleseismograms from Teleseismic Earthquakes West of the Array Site

Fig. 2a shows teleseismograms recorded at five 40T stations of seismic Line D across the CF for a M6.7 teleseismic earthquake TE288 (see Table 1 and Fig. 1d) occurring at 39-km depth beneath Hawaii Islands and ~4200 km (37.8^{o}) west from the CF seismic array site in Southern California. The primary P-wave and S-wave are visible in teleseismograms while strong surface waves appear after 1000-s from the event origin time. In Fig. 2b to 2d, we apply a band-pass (0.5-2.0 Hz) filter on teleseismograms between 435 s and 460 s from the event origin time recorded at four cross-fault lines (Lines A, B, C and D) of the square array. The transmitted first-arrival P-wave (called P') closely followed by the P-type fault-zone trapped wave (called P'FZTW) and the P-to-S converted wave (called S') from the Moho interface closely followed by the S-type fault-zone trapped wave (called S'FZTW) appear clearly in vertical and horizontal components, respectively, at stations within the ~1-km-wide CF compliant zone. The time difference between P' and S' is ~4 s when they travel at the average speed of 6 km/s and 3.3 km/s respectively between the Moho discontinuity at 30-km depth beneath Mojave Desert and the seismic array at the ground surface (Fig. 2e). When the P' and S' waves enter the bottom of LVWG formed by the CF compliant zone at certain depth, P'FZTW and S'FZTW are produced within it due to constructive interference of multiple reflected waves at the boundary between the fault zone and high-velocity surrounding rocks (Fig. 2f). The P'FZTW and S'FZTW show large amplitudes and long durations of ~3.5-4 s and ~7.5 s durations following P' and S', respectively, within which the FZTW amplitudes are twice above the background level. The PSSP ratios of these FZTWs are ~1.9 (refer to Li et al., 2016), indicating a prominent LVWG formed by the sub-surface CF compliant zone to trap seismic wave energy from the teleseismic earthquake. Multiple amplitude peaks of FZTWs are noticeable, corresponding to multiple layers of the CF LVWG with depths (refer to Li, 2025). In contrast, such FZTWs do not appear at stations located out of the CF complaint zone. The multiple-phase P" and S" waves (named PPmP and PSmS in Fig. 2e) traveling between the ground surface and the Moho interface, arrive at ~10 s after P' and S', respectively. When they enter the bottom of the CF compliant zone, the secondary fault zone trapped waves P"FZTW and S"FZTW are produced within the LVWG with smaller amplitudes due to twice longer distance and the incomplete reflection coefficient at the Moho discontinuity. Although the later portion of S'FZTWs might be overlapped by the early portion of P"FZTWS, we could measure their durations in horizontal and vertical components of teleseimograms, respectively, to avoid the interference between them.

In practice using receiver function technique, the fundamental assumptions include: (1) plane wave approximation, (2) the PS is primarily recorded on the radial, (3) the receiver function only describes P-S converted energy at the Moho discontinuity. We compute receiver functions by waveform deconvolution between the radial-component and vertical-component teleseisms for isolating the local response. Fig. 2g shows receiver functions computed for teleseismograms from teleseismic earthquake

TE288 recorded at stations of Lines A, B and C, which were located within and out of the CF compliant zone. The peakamplitudes of P' and P-to-S converted wave S' from the Moho in receiver functions (color curves) appear at ~0 s and ~4 s. The multiple-phase P" and S" arrives at ~10 s and ~14 s. Receiver functions (curves in blue and yellow) at stations A15, A24, B15, C09 and C20 located within the CF compliant zone show multiple large amplitude peaks in 7-s-long time window after S' (in red box and marked by red bar). They are likelyassociated with S-type trapped waves S'FZTW produced within the multi-layer LVWG of the CF compliant zone as S' wave enters its bottom (refer to Fig. 2f). In contrast, much lower amplitudes following S' in receiver functions (curves in green color) at stations A03, B01, B26, A04 and C03 located outside the compliant zone as shown by theoretical waveforms of the receiver function in Fig. 2e. Similarly, the large amplitudes in ~4-s time window after P' (in blue box and marked by blue bar) likely associated with P-type trapped waves P'FZTWs produced within the CF compliant zone as P' wave enters its bottom. In the same sense, the large amplitudes after multiple-phase P" and S" likely owning to trapped waves P"FZTW and S"FZTW produced within the CF compliant zone as P" and S" enter its bottom at depth. Thus, we consider that identified FZTWs in teleseismograms are produced by a low-velocity CF compliant zone at depth rather than the site effect or shallow anomaly structure.

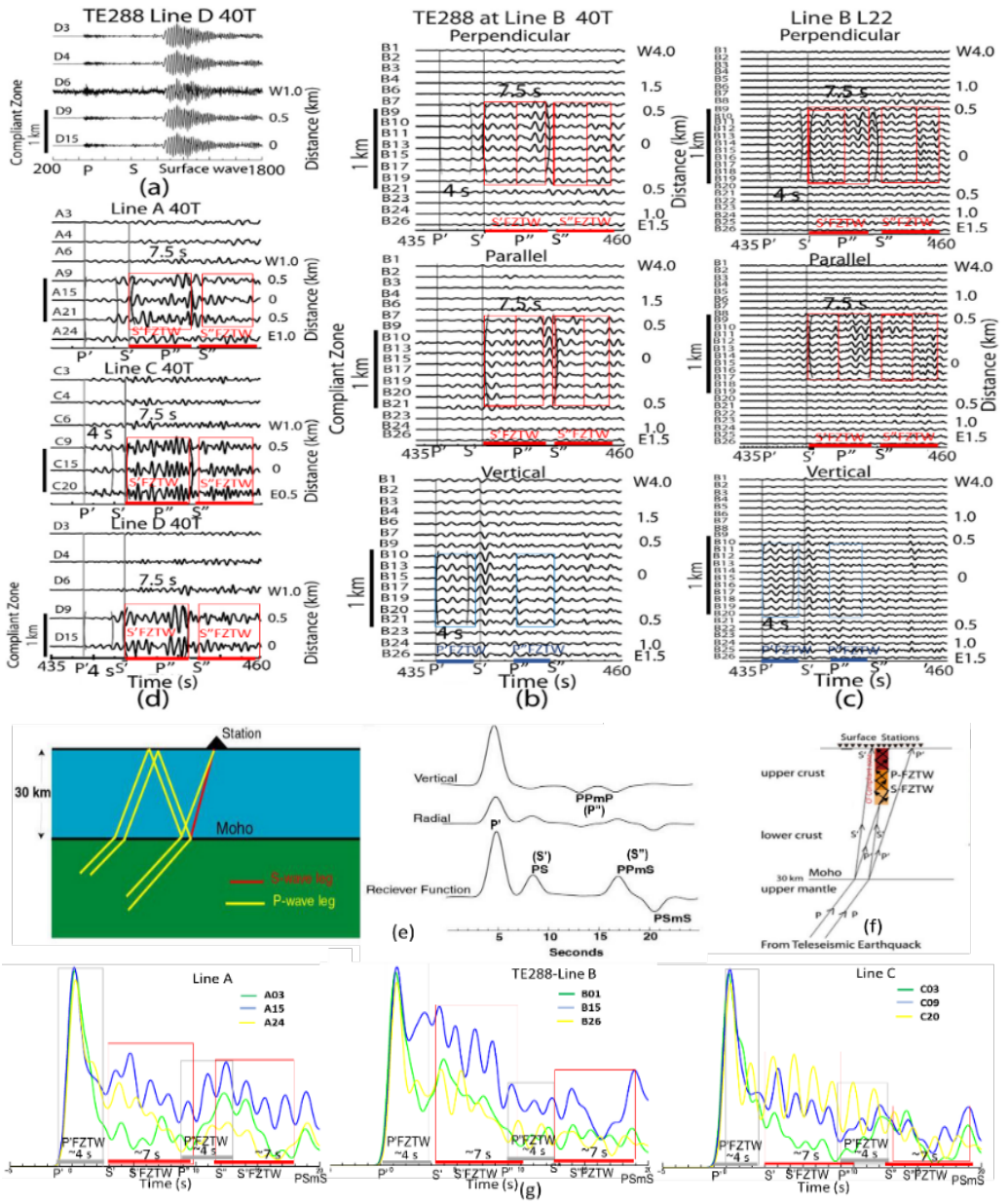


Figure 2. (a) Vertical-component teleseismograms recorded at 40T stations of Line D for teleseismic earthquake TE288; x-axis is 1600-s time. (b) and (c) There-component teleseismograms recorded at Line B with 40T and L22 sensors show S'FZTWs and P'FZTW with large amplitudes (in red and blue boxes) and long-durations (denoted by red and blue bars) at stations within the ~1-km-wide CF compliant zone. Multiple P" and S" followed by secondary FZTWs are visible. X-axis is 25-s time. (d) Teleseismgrams in horizontal-component at Lines A, C and D. (e) Sketch shows ray paths of the P-wave from teleseismic earthquake, the transmitting P-wave and P-S converted wave at the Moho and their multiples to station at the surface. Theoretical waveforms of the receiver function by modeling the amplitude and timing of those waves (PS, PPmP, PPmS, PSmS). (f) Sketch illustrates the P-wave from teleseismic earthquake, P' and S' waves entering the bottom of the CF compliant zone to produce trapped waves within the LVWG. (g) Computed receiver functions (curves in colors) of teleseismograms at Lines A, B and C. Large amplitudes with long durations following P' and S' (in blue and red boxes) at stations within the compliant zone owing to P'FZTW and S'FZTW). Multiple-phase P'' and S'' are followed by P''FZTW and S''FZTW.

In the next example, we examine teleseisms recorded for a M6.3 earthquake TE162 occurring at 140-km depth in Honshu Prefecture of Japan, approximately 9500-km (72.2°) west of the CF array site in California. Fig. 3a, 3b show three-component seismograms recorded at Line B. P-type and S-type fault-zone trapped waves P'FZTW and S'FZTW with large amplitudes and ~4-s post-P' and ~7-s post-S' durations appear at stations located within the ~1-km-wide CF compliant zone. Similarly, the trapped waves P"FZTW and S"FZTW following the multiple-phase P" and S" appear at these stations. P-type and S-type trapped waves are prominent in vertical and horizontal components of teleseismograms, respectively. However, P-type trapped waves are observable in perpendicular-component because waves come from the teleseismic earthquake TE162 to the CF LVWG sub-vertically and transverse to the CF strike. In contrast, brief body waves with weak amplitudes are recorded at stations out of the CF compliant zone, indicating that the CF compliant zone traps abundance of seismic energy. We computed spectral amplitudes of teleseismograms recorded within and out of the surface-rupture zones of the CF compliant zone. Spectral amplitudes are computed using the Seismic Analysis Code program (spectrogram) and normalized with respect to the maximum amplitude among all traces. Fig. 3c illuminates normalized spectral contours, showing large amplitude (in red color) of P'FZTW and S'FZTWs at ~1-2 Hz with ~4-s and ~7-s durations following P' and S' waves at on-fault station B15. The PSSP ratios of these FZTWs are ~1.7. Much lower amplitudes and shorter durations (<2-s) and smaller PSSP ratios (<0.8) are registered at stations B1 and B26 outside the compliant zone. The spectral amplitudes of trapped wave S''FZTW following the multiple-phase S'' waves at station B15 are also higher than those at stations B1 and B26.Both of P-type and S-type fault-zone trapped waves are observed in perpendicular-component teleseismograms recorded at stations of Lines A and C within the CF compliant zone (Fig. 3d). We notice that the late portion of S'FZTW is likely overlapped by the early portion of P"FZTW in the perpendicular-component teleseismograms. Therefore, we measure the duration time of the S'FZTW is mainly in the parallel-component teleseismograms to avoid this overlapping issue.

Fig. 3e shows computed receiver functions of teleseismograms at partial stations of Lines A, B and D. The P' and S' appear at 0 s and ~4 s, and the multiple-phase P" and S'' at ~10 s and ~14 s. Large amplitudes show in time windows ~3.5-4-s and ~7-s following P' and S', respectively, corresponding to the trapped waves P'FZTWs and S'FZTW with PSSP ratios of ~1.5 produced by the fault LVWG and recorded at stations (A15, A24, B13, B14, B15, D9 and D15) within the CF compliant zone. In contrast, much lower amplitudes following P' and S' in receiver functions at stations (B03 and B24) outside the compliant zone. Similarly, large amplitudes of P''FZTW and S''FZTW appear after multiple-phase P'' and S''. Therefore. receiver functions show the existence of low-velocity structure along the CF compliant zone at depth.

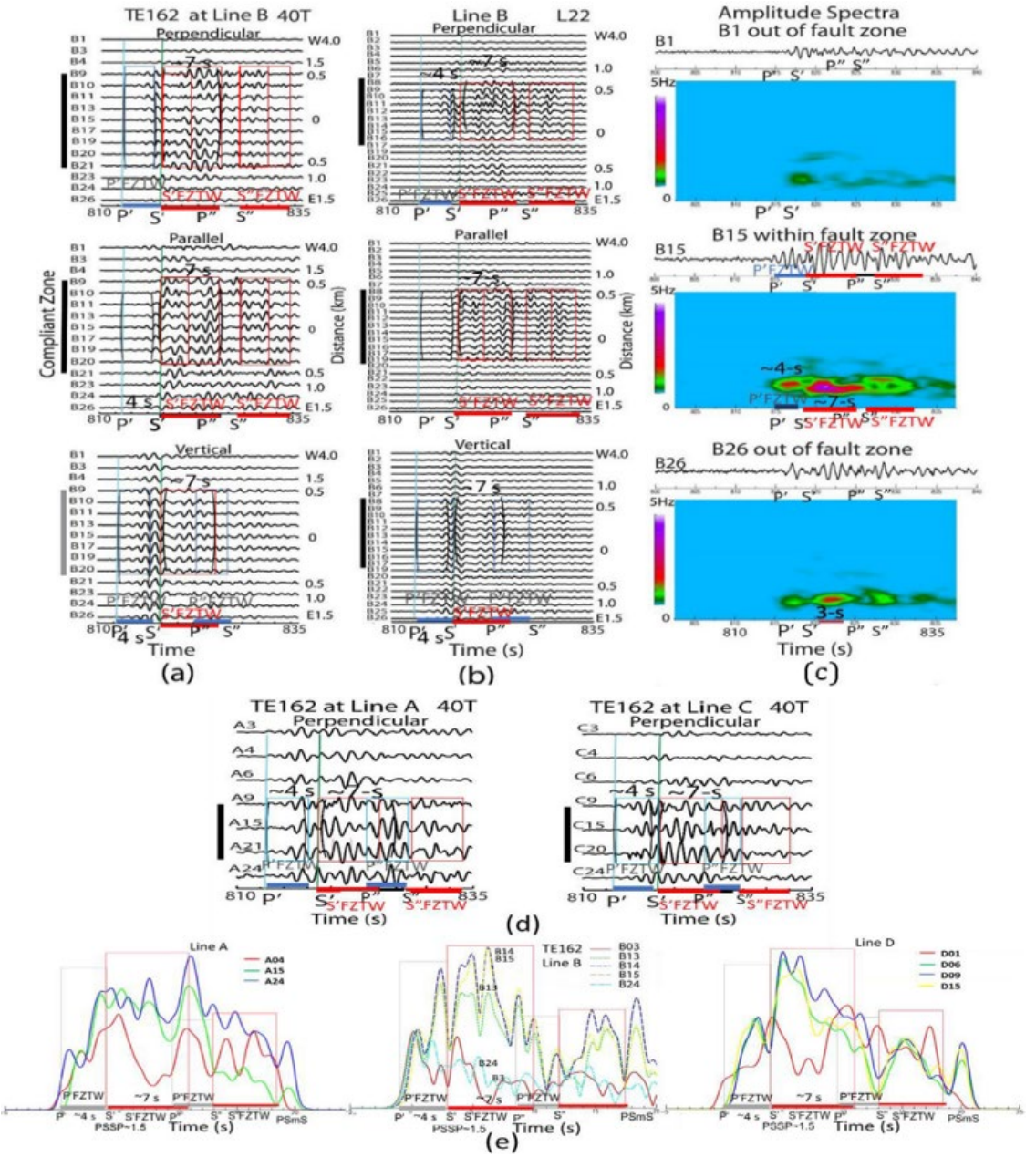


Figure 3. (a) and (b) Three-component teleseismograms (0.5-2 Hz filtered) at Line B for teleseismic earthquake TE162 show S'FZTW and S"FZTW with ~7-s duration as well as P'FZTW and P"FZTW with ~4-s post-P duration iat stations within the CF compliant zone. X-axis is 25-s time. (c) Normalized spectral amplitude contours of perpendicular-component teleseismograms at B1, B15 and B20 show large amplitudes (red color) of FZTWs with ~7-s duration after S' and S" as well as ~4-s duration after P' at station B15 within the CF compliant zone. (d) Same as in (a) but teleseismograms recorded at Lines A and C. (e) Receiver functions of teleseismograms at stations of Lines A, B and D. P' and S' waves are followed by large amplitudes owning to P'FZTWs and S'FZTW at stations within the CF compliant zone, but much lower amplitudes after P' and S' at stations outside the compliant zone. P" and S" arrive at ~10s and ~14s, followed by large amplitudes of P"FZTW and S"FZTW.

Teleseismograms from Teleseismic Earthquakes Southwest of the Array Site

We further examine the data recorded at Lines A, B, C and D across the CF for a M6.7 teleseismic earthquake TE244 occurring at 38-km depth beneath Papua New Guinea Islands ~11,000 km (69.5º) southwest from the square array site in California. P' and S' arrive at ~848 s and 852 s from the event origin time. Figs. 4a to 4d exhibit the recorded teleseismograms and spectral amplitude contours, showing P'FZTW and S'FZTW with large-amplitude and long wavetrains with ~4-s and 7-

8-s durations following P' and S' at stations located within the CF compliant zone but not at stations out of the zone. P'FZTWs are visible in the vertical- and perpendicular-component seismograms from this shallow teleseismic earthquake with the ray-path nearly perpendicular to the fault strike. The PSSP ratios of these FZTWs are 1.7-2. We also observe P"FZTW and S"FZTW following the multiple-phases P" and S" as well as PSmS in the later coda in teleseismograms. Fig. 4e shows receiver functions computed for teleseismograms recorded at stations of Lines A, B and C for this event. Large-amplitudes in time windows of ~4-s and ~7.5-s after P' and S' as well as their multiple phases in receiver functions at stations within the CF compliant zone are interpreted owning to fault-zone trapped waves produced within the CF compliant zone when body waves from the Moho enter its bottom. In contrast, lower amplitudes are in the same time windows after P' and S' at stations outside the compliant zone. Similar observations for the multiple-phase P" and S".

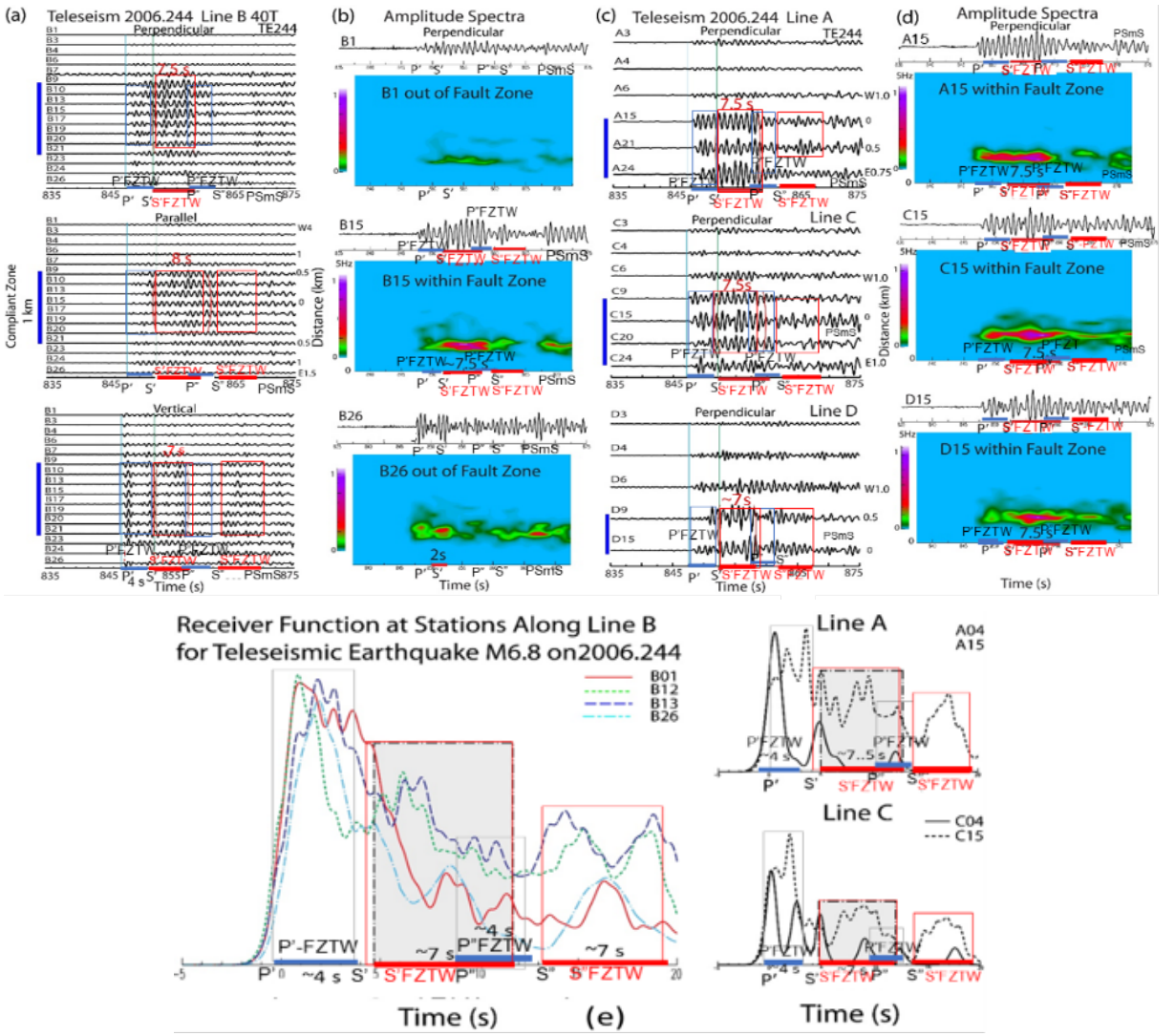


Figure 4. (a) Three-component teleseismograms recorded at Line B for a M6.7 teleseismic earthquake TE244, show P'FZTW and S'FZTW with ~4-s and ~7.5-s durations following P' and S' on seismograms recorded at stations B9–B21 within the CF compliant zone. The multiple-phase P" and S" are followed by P"FZTW and S"FZTW. The PSmS is notable in the later coda of teleseismograms. X-axis is 40-s time. (b) Spectral amplitude contours of teleseismograms show large-amplitude FZTWs at 1-2 Hz with ~7.5-s duration at on-fault station B15, but lower amplitudes with much shorter duration (<2-s) following S' at stations B1 and B26 outside the compliant zone. (c) and (d) Similar to (a) and (b), but at Lines A, C and D. (e) Receiver functions of teleseismograms at stations of Lines A, B and C for TE244.

In the next example, Fig. 5 exhibits teleseismograms and their spectral amplitude contours for a M6.3 deep earthquake TE276 occurring at 161-km depth beneath Solomon Islands, approximately 10,000 km (65.7°) southwest of the seismic array. P-type trapped waves P'FZTW and P"FZTW with ~4-s durations following P' and P' are more visible in vertical component, while S-type trapped waves S'FZTW and S"FZTW with 7-7.5-s durations are more visible in horizontal components ofteleseismograms, because the waves from this deep event arrive at stations nearly vertically. The PSSP ratios of these trapped waves are 1.7-1.9. Receiver functions computed from teleseimograms recorded at partial stations of four lines for this deep event show larger amplitudes in the time windows of ~3.5-s and ~7-s following P' and S' at stations within the compliant zone, owning to trapped waves P'FZTW and S'FZTW produced by the fault LVWG when P' and S' waves from the Moho enter the bottom of CF compliant zone.

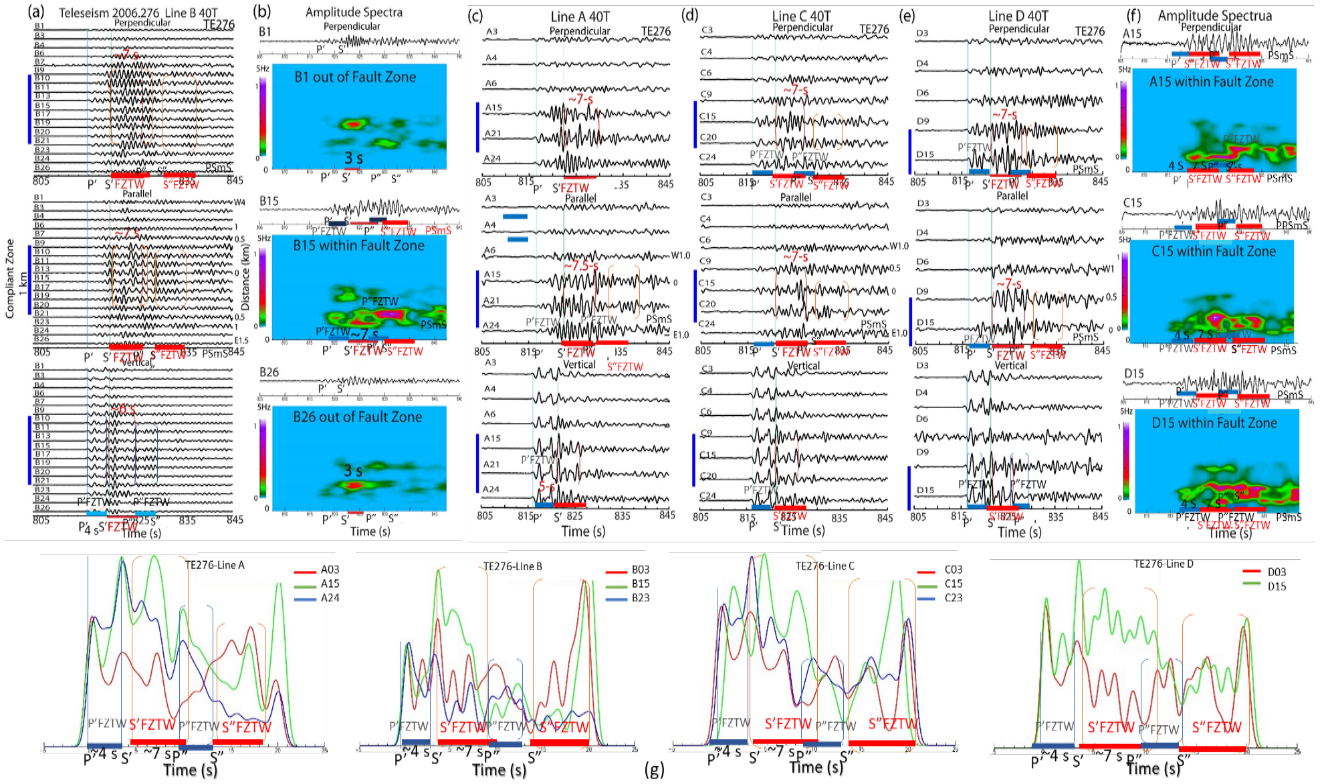


Fig. 5 Similar as in Fig. 5, but teleseismograms recorded at Line B for teleseismic earthquake TE276. FZTWs with ~7-s post-S duration in horizontal seismograms recorded at stations located within the compliant zone. Spectral amplitude contours of teleseismograms at on-fault stations A15, C15 and D15 show large amplitudes of FZTWs. Receiver functions of teleseismograms at stations within the CF compliant zone show large amplitudes in ~4-s and ~7.5-s windows following P' and S', owning to P'FZTW and S'FZTW. In contrast, lower amplitudes in the same time windows at stations out of the compliant zone.

Seismograms from Teleseismic Earthquakes Southeast of the Array Site

In the following two examples, we examine the teleseismograms recorded for teleseisnic earthquakes located south of the array site with the ray-path parallel to the CF strike. Fig. 6 shows teleseismograms recorded at four lines of the square array stop the CF for a M6.1 teleseismic earthquake TE272 with occurring at 53-km depth in South America, ~6,200-km (55°) southeast of the CF array. Large-amplitude S-type trapped waves S'FZTW and S"FZTW with ~7-s durations following S' and S" appear in teleseismograms recorded at stations within the ~1-km-wide CF compliant zone, but not at stations out of the zone. P-type trapped waves P'FZTW and P"FZTW with ~3.5-s durations are more visible in vertical- and parallel-component teleseismograms because P-wave radiated from the source propagates along the fault line. Receiver functions computed from the vertical- and parallel-component teleseismograms recorded at stations within the CF compliant zone of

four lines show large amplitudes in ~3.5-s and ~7-s time windows following the P' and S' waves, owning to P- and S-type FZTWs produced within the fault LVWG. On the other hand, the lower amplitudes after the P- and S-arrivals in receiver functions at stations outside the compliant zone because of no FZTWs approach these stations.

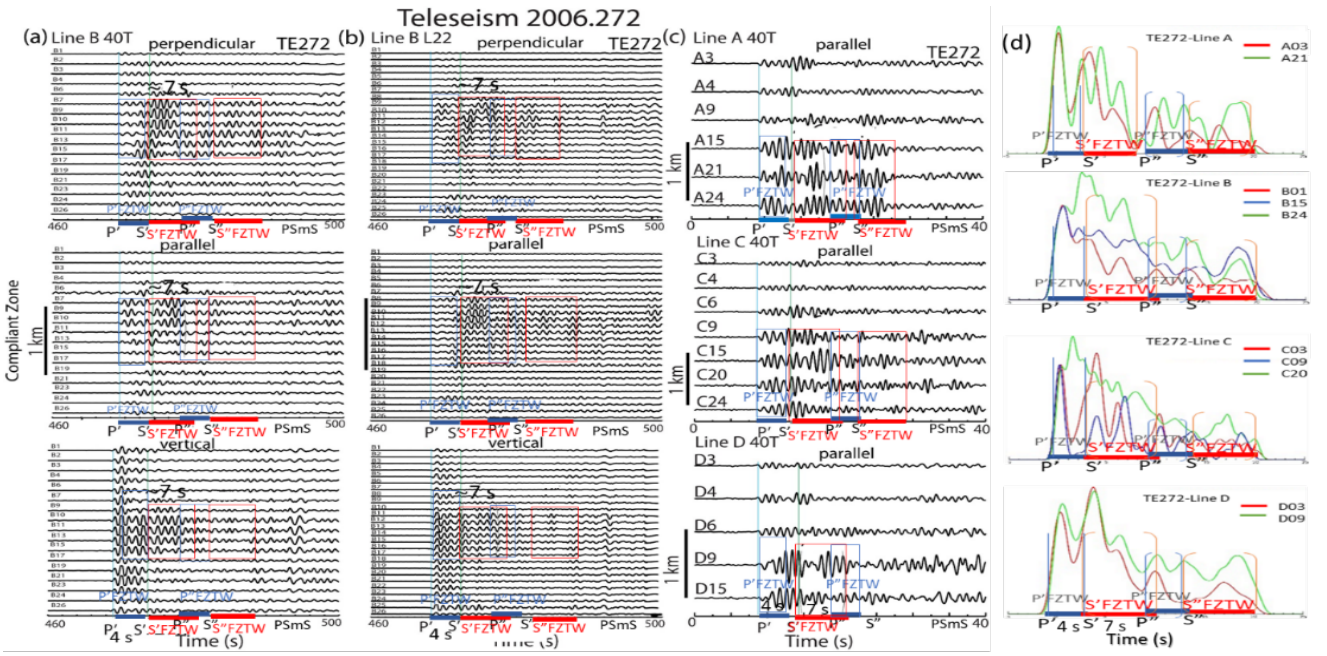


Figure 6. (a) and (b) Parallel-component teleseismograms recorded at Line B for teleseismic earthquake TE272 show P-type and S-type FZTWs with large amplitudes and ~4-s and ~7-s durations arriving at stations located within the ~1-km-wide CF compliant zone. (c) Parallel-component teleseismograms recorded at Lines A, C and D for this teleseismic event show P'FZTWs and S'FZTW at stations within the CF compliant zone. (d) Computed receiver functions of teleseismograms at Lines A, B, C and D show large amplitudes in time windows of ~4-s and ~7-s after P' and S' and multiple-phase P" and S" at stations within the CF compliant zone, but small amplitudes at stations outside the zone.

We finally examine teleseismograms recorded for a M6.8 teleseismic earthquake TE317 occurring at 548-km depth in Bolivia, approximate 7900-km (78.9°) southeast of the array site in California, from which seismic waves travel along the CF strike to the stations. Fault-zone trapped wave S'FZTW and S"FZTW with ~7-s duration following the P-S converted S' and its mutiple-phase S" in horizontal-component teleseismograms recorded at stations within the CF compliant zone (Figs. 7a, b). While, only P'FZTW with ~4-s duration appears in the vertical-component seismograms because the ray path from this deep teleseismic earthquake to the CF array is nearly vertical. Normalized spectral amplitude contours of teleseismograms recorded at on-fault station B15 show large-amplitude P-type and S-type FZTWs at 1-2 Hz with ~4-s and ~7-s durations, but weaker and much shorter (2-s) wavetrains at stations B1 and B26 outside the compliant zone (Fig. 7c). Figs. 7d, 7e show similar features of teleseismograms and speectral amplitudes at Lines A, C and D. Fig. 9f shows receiver functions of teleseismograms recorded at partial stations of Lines A, B, C and D. Large-amplitudes appear in time windows of ~3.5 s and ~7 s, respectively, following P' and S' as well as multiple-phase P" and S" at stations within the CF compliant zone, owning to P-type and S-type fault-zone trapped waves produced within the CF compliant zone. In contrast, much lower amplitudes are shown in the same time windows in the receiver functions of teleseismograms recorded at stations outside the compliant zone.

In Figs. 2 to 7, we exhibit teleseismograms recorded at the square array installed at the Calico Fault at Mojave Desert and their receiver functions for seven teleseismic earthquakes. show the consistent wavetrains with large amplitudes and approximate 3.5-4 s and 7-8 s durations closely following the transmitting P' and P-to-S converted S' waves from the

Moho discontinuity as well as their multiple-phase P" and S". We interpret these large-amplitude long-duration wavetrains to be a new-type of fault-zone trapped waves which arise from cohesive interference of reflected waves within the LVWG formed by the CF compliant zone when P' and S' waves enter its bottom at certain depth in the upper crust. Therefore, it is plausible to use these type FZTWs identified in teleseismograms to accurately depict the depth extension of the CF compliant.

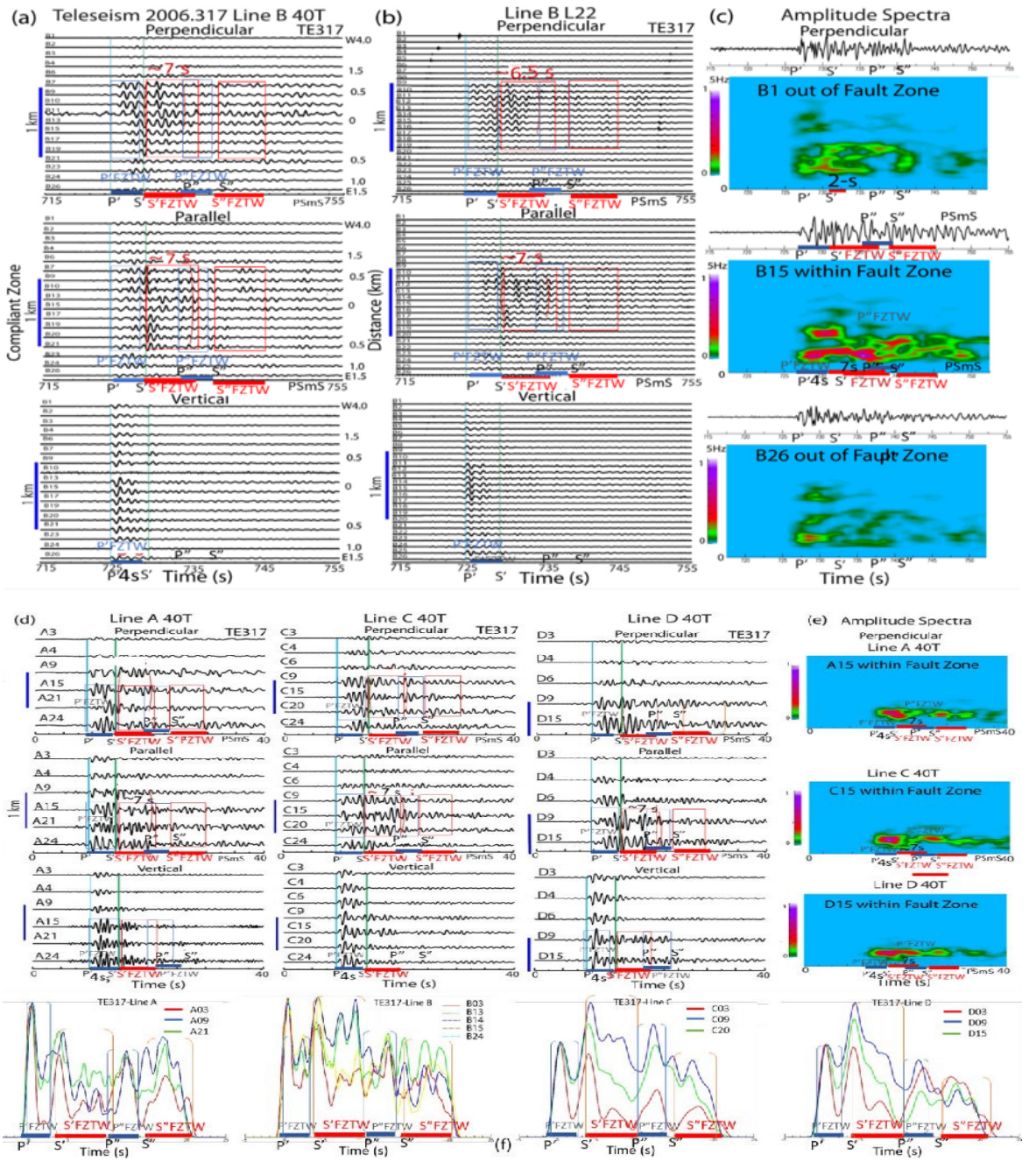


Figure 7. (a) and (b) Three-component seismograms recorded at stations along Line B for teleseismic earthquake TE317. Teleseismograms at stations located within CF compliant zone show FZTWs with long durarions following the P' and S' as well as multiple-phase P' and S". (c) Normalized spectral amplitude contours of teleseismograms show large-amplitude FZTWs at 1-2 Hz with ~3.5.-4-s and ~7-s durations at on-fault station B15, but lower amplitudes with shorter durations (<2-s) at stations B1 and B26 out of the compliant zone. (d) and (e) Same as in (a)-(c) but at Lines A, C and D. (f) Receiver functions of teleseismograms show large amplitudes appearing in time windows following P'- and S'-arrivals at stations within the CF compliant zone, but much lower amplitudes at stations outside the zone.

However, we need carefulness in measurement of the duration time of fault-zone trapped wave S'FZTW following the S'-wave because its late portion might be overlapped by the early portion of P"FZTW following the multiple P" wave in teleseismograms, although the amplitudes of SFZTWs are much larger than the amplitudes of P"FZTW. The time span between arrivals of the S' and P" is only ~5 s while the duration of the S'FZTW is usually longer than ~6-7 s in our study, Therefore, we use three-component teleseismograms in analysis. In practice, we focus on the horizontal-component transverse to the ray-path between the teleseismic earthquake and the seismic array to measure the duration time of the S'FZTW because the multiple trapped wave P"FZTW is weakest in this component.

3-D FINITE-DIFFERENCE SIMULATIONS of FZTWs IDENTIFIED IN in TELESEISMOGRAMS

We use a 3-D finite-difference code to compute synthetic teleseismograms. It is second order in time and fourth order in space, and propagates the complete wave-field through an elastic media with a free surface boundary and spatially variable anelastic damping Q (Graves, 1996; Vidale et al., 1985). The 160-160-640 element grid in x–y–z coordinates with a grid spacing of 62.5 m to simulate a volume of 10 km in width, 10 km in length, and 40 km in depth. The CF compliant zone is represented by a multi-layer low-velocity waveguide (LVWG) embedded at the center of this volume in half space. Fig. 8 shows the model of P-wave velocities across the multi-layer LVWG, within which P and S velocities are reduced by maximum 40-50% from surrounding-rock velocities with ~5% difference on two sides of the LVWG. The LVWG is 1.5-km wide at the free surface, and becomes narrower with depth. The receiver line crosses perpendicularly the LVWG at the free surface. Plane waves with P-motion coming from teleseismic earthquakes at various azimuth directions and incident angles to the Moho interface between the crust and mantle. For 62.5-m grid spacing in the model and the minimum wave velocity 0.75 km/s within the LVWG, the maximum frequency of synthetic seismograms is 3-Hz.

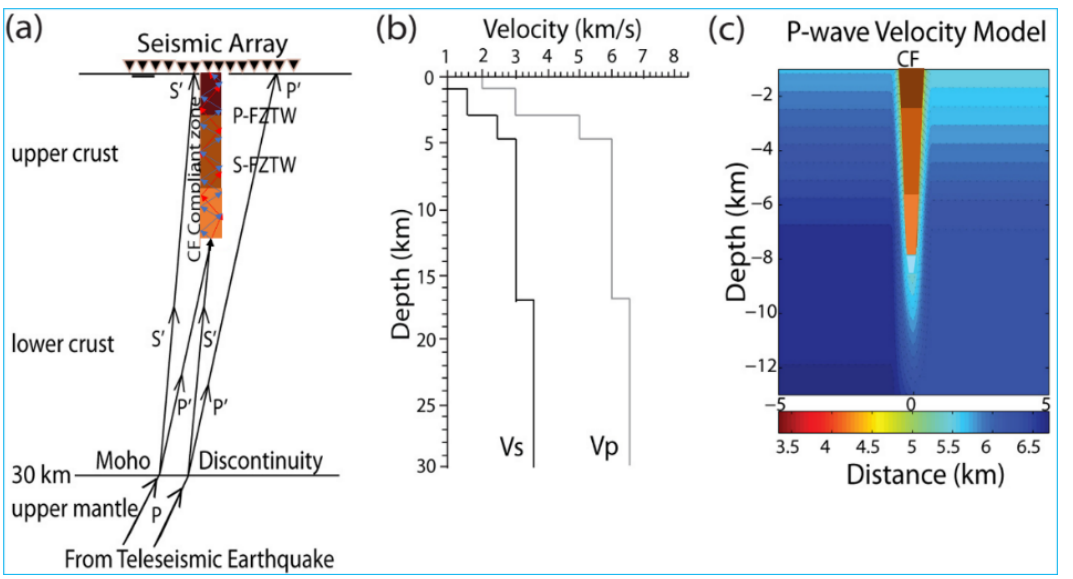


Figure 8. (a) Schematic diagram illustrates ray-paths of the first-arrival P wave from a teleseismic earthquake and transmitting P' and converted S' at the Moho discontinuity at 30-km depth. P' and S' waves enter the bottom of a multi-layer LVWG (brawn color rectangular) to produce trapped waves P-FZTW and S-FZTW within it, and recorded at stations (triangles) atop the LVWG. (b) 1-D P- and S-velocities beneath Mojave Desert represent surrounding-rock velocities of the CF compliant zone. (c) The model of P-wave velocities across the CF compliant zone represented by a multi-layer LVWG within which velocities are reduced by maximum 40-50% from wall-rock velocities.

Fig. 9 exhibits 3-D finite-difference synthetic teleseismograms using the basic model of Fig. 10c, in which the LVWG is truncated at 8 km and 16 km in the upper crust, respectively. The P-type and S-type FZTWs with large amplitudes and long durations (waveform amplitudes are above twice the background noise level) arrive at stations within the 1.5-km wide LVWG. The P-type FZTWs are observable in vertical- and radial-component seismograms. Duration times of P-type and S-type FZTWs produced by an 8-km-deep LVWG are 4 s and 7 s, but are 7 s and 13-14 s produced by a 16-km-deep LVWG. The duration times of synthetic P- and S-type FZTWs produced by the 8-km-deep LVWG are most likely consistent with observations.

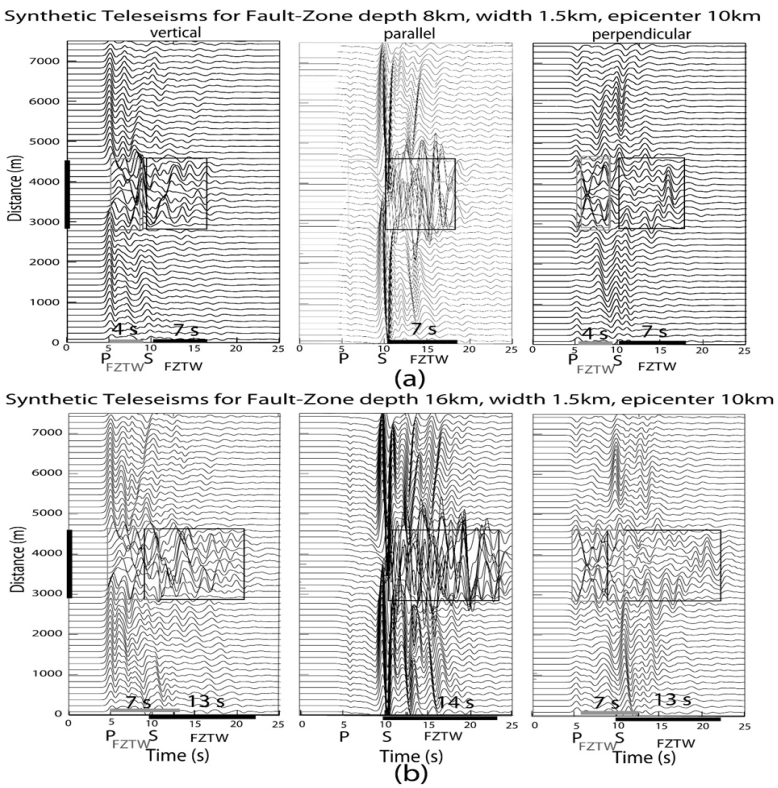


Figure 9. 3D Finite-difference synthetic profiles of three-component teleseismograms across the fault compliant zone using the model in Fig. 10c with a LVWG extending to (a) 8-km and (b) 16-km depths for plane waves hitting the Moho in the direction sub-perpendicular to the fault strike and at 60° incident angle to the Moho. Seismograms have been <3 Hz filtered. First-P and P-to-S converted waves arrive at ~5 s and ~9 s. P- and S-type FZTWs with large amplitudes (in boxes) and long durations (marked by black bars and measured times) appear at stations within the 1.5-km wide LVWG (marked by the vertical black bar).

We further tested the basic model in Fig. 8c with a 1-km-wide LVWG extending to various depths of 4 km, 8 km and 12 km, respectively. Fig. 10a illustrate synthetic three-component teleseismograms. FZTWs with large amplitudes and long wavetrains follow P and S waves at stations within the LVWG, but do not appear at stations out of the waveguide. S-type FZTWs are dominant in three-component teleseisms. P-type FZTWs are observable in vertical- and radial-components, but not clearly seen in the component transecting the wave propagation direction. The durations of P-type and S-type FZTWs increase from 3.5 s though 4 s to 4.5 s and from 6 s through 7.5 s to 9 s, respectively, as the LWVG depth extension increases from 4 km through 8 km to 12 km, because FZTWs travel a longer distance within the deeper LVWG. Fig. 10b shows spectral amplitudes of synthetic teleseismograms at station located at the center of the LVWG which extends to 4 km, 8 km and 12 km, respectively, show large-amplitudes and 3.5-4.5 s durations of P-type FZTW

and 6-9.5 s durations of S-type FZTW. P-type FZTWs appear only in the vertical-component seismograms. The durations of trapped wavetrains increase with the depth extension of the LVWG. The duration times of synthetic P-FZTWs (~4 s) and S-FZTWs (~7.5 s) produced by the 8-km-deep LVWG are mostly consistent with those observed in Figs. 2 to 7, suggesting that the Calico fault compliant zone likely to extend to a depth of approximately 8 km.

Finally, we compute 3D finite-difference synthetic seismograms using the velocity model having an 8-km-deep multi-layer LVWG shown in Fig. 10c to match teleseismograms recorded for three teleseismic earthquakes TE288 and TE244 occurring west of the CF array site, and TE317 south of the site. In modeling for TE288 and TE244, we assume plane waves with P-motion hit the Moho at 30-km depth with the azimuth angle 75^{o} (sub-perpendicular to the fault strike) and incidence angle 60^{o}. For TE371, we assume P-motion with the azimuth angle 15^{o} (sub-parallel to the fault strike) and sub-vertical incidence angle 75^{o}. Fig. 10c shows P-type and S-type fault-zone trapped waves P'FZTW and S'FZTW with ~4-s and ~7.5-s durations closely following P' and S', respectively, in both observed and synthetic teleseismograms arriving at stations between B9 and B20 located within the ~1-km-wide CF compliant zone for these teleseismic earthquakes. In contrast, no such large-amplitude and long-duration wavetrains recorded at stations out of the CF compliant zone. We interpret that when the P' and S' waves from the Moho discontinuity sub-vertically enter the bottom of compliant zone at 8-km depth, FZTWs arise from constructive interference of multiple-reflection within the CF compliant zone, and recorded at the stations located within the CF compliant zone at the ground surface. Synthetic multiple-phase P" and S" arrive at ~10 s after P' and S', respectively, and are closely followed by secondary trapped waves P"FZTW and S"FZTW, which show smaller amplitudes than P'FZTW and S'FZTW due to wave attenuation caused by low Q fault rocks and the reflection coefficient at the ground surface and the Moho discontinuity.

Fig. 10d exhibits receiver functions computed from deconverlation of the radial and vertical components of synthetic teleseismograms at partial stations of Line B located within and out of the CF compliant zone for teleseismic earthquake TE272 shown in Fig. 12c. The P' and P-to-S converted wave S' from the Moho in receiver functions (color curves) arrive at 0 s and ~4 s on x-axis. The multiple-phase P" and S" arrives at ~10 s and ~14 s. Receiver functions at stations located within the CF compliant zone show large amplitudes in ~4-s and 7.5-s-long windows after the P' and S' waves (in blue and red boxes), respectively. They are interpreted to be owing to P-type and S-type trapped waves P'FZTW and S'FZTW produced within the multi-layer LVWG when P' and S' waves enter its bottom at ~8-km. Similarly, large amplitudes appear after multiple-phase P" and S" arrivals in receiver function curves at stations within the CF compliant zone, owning to secondary fault-zone trapped waves P"FZTW and S"FZTW recorded at these stations although their amplitudes smaller than those following P' and S' due to wave attenuation over the longer distance within the low-Q fault zone. In contrast, receiver functions at stations out of the LVWG show low amplitudes after the P' and S' because FZTWs do not appear at them, as shown in theoretical waveforms of the receiver function by modeling the amplitude and timing of P' and S' waves and their multiples waves (Fig. 3e). The receiver functions computed from synthetic teleseismograms are consistent with those from teleseismograms recorded at stations within the CF compliant zone, providing the further constraint to its depth extension to ~8-km.

Synthetic waveforms developed from our preferred velocity model in Fig. 10c are in general agreeable with observations for this teleseismic earthquake, suggesting that the depth extension of the CF compliant core zone with velocity reduction of 40-50% is approximately 8 km. The value is likely ~2 km greater than the depth extension (~5-6 km) of the Calico fault compliant zone imaged by Cochran et al. (2009). Nevertheless, the finite-difference simulations of FZTWs udentified in teleseismograms recorded for seven teleseismic earthquakes in Fig. 1d and Table 1 provide a primary constraint on the depth extension of the CF compliant zone beneath Mojave Desert. Since the depth extension and the velocity reduction within the compliant zone are trad-off in the trial-and-error forward modeling, a systematic simulation of teleseismograms recorded at the CF for more teleseismic events with different incident angles to the CF seismic array will provide further constraints on the true depth extension of the CF compliant zone.

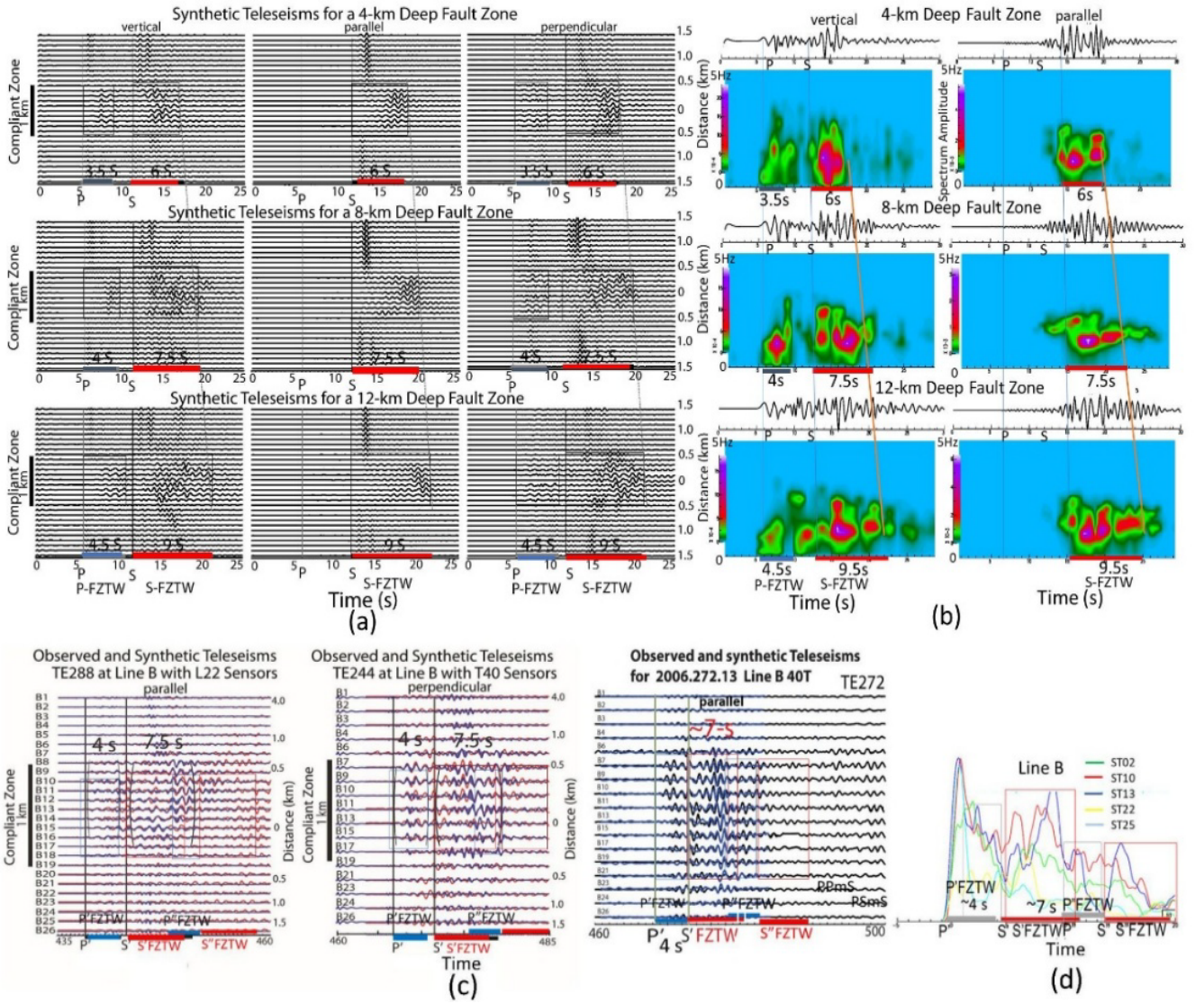


Figure 10. (a) 3D Finite-difference synthetic teleseismograms (1-3 Hz) along the cross-fault line, using the model in Fig. 10c with a LVWG extending to various depths of 4 km, 8 km and 12 km. Duration times of synthetic P-FZTWs (4 s) and S-FZTWs (7.5 s) produced by the 8-km-deep LVWG are consistent with observations. (b) Spectral amplitude contours of synthetic teleseismograms (see left) at stations within the CF compliant zone show that FZTW duration times increase with the LVWG depth extension. (c) Observed (red) and synthetic (blue) teleseismograms (<3 Hz) at Line B for teleseismic earthquakes TE288, TE244 and TE272. P'FZTWs with ~4-s duration following the P' and S'FZTWs with ~7.5-s duration follow the S' wave at stations within the 1-km-wide CF compliant zone. (d) Receiver functions (color curves) of synthetic teleseismograms at stations ST10 and ST13 (in red and blue colors) within, and ST03, ST22 and ST26 (in green, yellow and cryn colors) out of the CF fault damage zone, respectively, for teleseismic earthquake TE272.

CONCLUSION AND DISCUSSION

Finite-difference methods have been employed to synthesize FZTWs arising from near-surface explosions and local earthquakes for fault models (e.g., Li and Vidale, 1996; Li and Malin, 2008; Li et al., 1994b, 2002, 2003, 2004, 2021; Igel et al., 2002; Ellsworth and Malin, 2021). In this article, we introduce a new-type of FZTWs. When the primary P-wave from teleseismic earthquakes strikes the Moho discontinuity, the transmitting P-wave and P-to-S converted wave may enter the bottom of CF compliant zone sub-vertically at certain depth, and are reflected at the boundaries of the low-velocity compliant zone and high-velocity surrounding rocks to produce the FZTWs due to cohesive interference. These

waves are recorded at the dense seismic array atop the Calico fault zone and capable to provide more unprecedented constraints on the depth extension of fault damage structure at seismogenic depth where there is lack of coverage of waves generated by explosions and local earthquakes at shallow depths.

Recently, 3-D finite-difference full-waveform simulations in terms of a depth-dependent multi-layer fault zone low-velocity waveguide (LVWG) show the complexity in analysis and interpretation of FZTWs recorded at realistic fault zones (Li, 2025). The synthetic FZTWs show multiple amplitude peaks in their long wavetrains. The earlier portion of these FZTWs with the larger amplitude peak at lower frequency is most likely produced by cohesive interference of reflected waves within the upper layer of the fault damage zone with slower velocities, while the later portion of FZTWs with smaller amplitude peaks at higher frequencies might arise from the lower layers with relatively faster velocities. Even if a source is located out of the fault damage zone, large-amplitude FZTWs could be produced within its upper layers as long as the source depth is below the bottom of the upper layers. these FZTWs show shorter post-S durations than those for the source located at the same depth within the deep portion of the fault damage zone. However, such type of FZTWs with large amplitudes but short post-S durations are usable to characterize the upper portion of the fault damage zone, but it might lead to underestimate the depth extension of fault damage zone.

The data recorded at the dense array of 100 three-component seismographs deployed atop the Calico fault in the central Mojave Desert, Southern California, provide us a good opportunity to detect the seismic phases closely following the first-arrival P wave from teleseismic earthquakes because the seismic array worked in continuous recording mode at an extremely quiet site for six months. In this paper, we show teleseismograms recorded at four cross-fault seismic lines A, B, C and of the square array for seven teleseismic earthquakes with magnitudes M≥6 and occurring at distances of ~40o-80o great circles (approximate 5,000-11,000 km) to the Calico seismic array site. Both P-type and S-type FZTWs closely following the transmitting P-wave and P-to-S converted wave at the Moho discontinuity as well as their multiple-phases between the ground surface and Moho are identified in teleseismograms recorded at stations located ~1-km-wide CF compliant zone which forms a low-velocity (by 40-50% reduction) multi-layer waveguide (LVWG) to trapped seismic waves within it. In the 3-D finite-difference simulations of this type of FZTWs, we assume the P and converted S waves as plane-waves entering the bottom of the LVWG formed by CF compliant zone that extends from the ground surface to various depths. The synthetic FZTWs with large amplitudes and long-durations in terms of the model having an 8-km-deep LVWG are mostly agreeable with observations.

Our observations and numerical modeling of the FZTWs identified in teleseismograms are further confirmed by receiver-functions of teleseismograms recorded at stations located within the CF compliant zone. Large amplitudes closely following the transmitting P-wave and converted S- in receiver-functions are interpretated to be owing to FZTWs which are produced within the LVWG of the CF compliant zone when the body waves coming from the Moho discontinuity enter it. These FZTWs are capable to provide more unprecedented
constraints on the depth extension of fault damage zone, beyond approach of waves generated by explosions and local earthquakes at shallower depths. A permanent damage zones on the Calico fault at seismic depth may play a critical role in development and dynamics of faults, intensity of ground motion, and earthquake hazards along the low-velocity fault zone due to its waveguide effect.

ACKNOWLEDGEMENTS

The research project was supported by National Science Foundation (grant EAR-0439947) and Southern California Earthquake Center’s award 15195, for which Yong-Gang Li was the Principe Investigator. The present study is benefited by the previous results. Thank Elizabeth Cochran for her work on archive of the data recorded at the seismic array deployed at the Calico fault in 2006.

REFERENCES

Aki, K., Characterization of barriers on an earthquake fault, J. Geophys., Res. 84, 6140-6148, 1979.

Aki, K., Asperities, barriers, characteristic earthquakes, and strong motion prediction, J. Geophys. Res., 89, 5867-5872, 1984.

Cochran, E. S., Y. G Li, P. M. Shearer, S. Barbot, Y. Fialko, and J. E. Vidale, Seismic and geodetic evidence for extensive, long-lived fault damage zones, Geology, v9.37; no. 4, 315–318, doi:10.1130/G25306A.1; Data Repository item 2009082, 2009.

Ellsworth W. L. and P. E. Malin, Deep rock damage in the San Andreas fault revealed by P- and S-type fault zone guided waves, Sibson's volume, Geological Society Special Publication, New Zealand, V359, 39-53, 2011.

Ferry, R., Thomas, M. Y., Bhat, H. S., and Dubernet, P., Depth dependence of coseismic off-fault damage and its effects on rupture dynamics. Journal of Geophysical Research: Solid Earth,130, e2024JB029787. https://doi.org/10.1029/2024JB029787, 2025.

Fialko, Y., Probing the mechanical properties of seismically active crust with space geodesy: Study of the coseismic deformation due to the 1992 Mw7.3 Landers (southern California) earthquake: J. Geophys. Res., v. 109, B03307, doi: 10.1029/2003JB002756, 2004.

Fialko, Y., Sandwell, D., Agnew, D., Simons, M., Shearer, P., and Minster, B., Deformation on nearby faults induced by the 1999 Hector Mine earthquake: Science, v. 297, p. 1858–1862, doi: 10.1126/science.1074671, 2002.

Graves, R.W., Simulating seismic wave propagation in 3D elastic media using staggered-grid finite differences, Bull. Seis. Soc. Am., 86, 1091-1106, 1996.

Igel, H, G. Jahnke, and Y. Ben-Zion, Numerical simulation of fault zone guided waves: accuracy and 3-D effects, Pure Appl. Geophys., 159, 2067-2083, 2002.

Li, Y.G., P.C. Leary, K. Aki, and P.E. Malin, Seismic trapped modes in the Oroville and San Andreas fault zones, Science 249, 763–766, 1990.

Li, Y.G., and P.C. Leary, Fault-zone trapped seismic waves, Bull. Seis. Soc. Am., 80, 1245-1271, 1990.

Li, Y.G. and J.E. Vidale, Low-velocity fault zone guided waves; numerical investigations of trapping efficiency, Bull. Seis. Soc. Am., 86, 371-378, 1996.

Li, Y. G. and P. E. Malin, San Andreas Fault damage at SAFOD viewed with fault-guided waves, Geophys. Res. Lett., 35, L08304, doi:10.1029/2007GL032924, 2008.

Li, Y. G., K. Aki, D. Adams, A. Hasemi, and W. H. K. Lee, Seismic guided waves trapped in the fault zone of the Landers, California, earthquake of 1992, J. Geophys. Res., 99, 11,705-11,722, 1994a.

Li, Y. G., J. E. Vidale, K. Aki, C. J. Marone, W. H. K. Lee, Fine structure of the Landers fault zone; segmentation and the rupture process, Science, 256, 367-371, 1994b.

Li, Y. G., J. E. Vidale, K. Aki, and F. Xu, Depth-dependent structure of the Landers fault zone using fault zone trapped waves generated by aftershocks, J. Geophys. Res. 105, 6237–6254, 2000.

Li, Y. G., J. E. Vidale, S. M. Day and D. Oglesby, Study of the M7.1 Hector Mine, California, earthquake fault plan by fault-zone trapped waves, Hector Mine Earthquake Special Issue, Bull. Seism. Soc. Am., 92, 1318-1332, 2002.

Li, Y. G., J. E. Vidale, D. Oglesby, S. M. Day and E. Cochran, Multiple-fault rupture of the M7.1 Hector Mine, California, earthquake from fault-zone trapped waves, J. Geophys. Res., 108, ESE 11, 1-25, 2003.

Li, Y. G., J. E., Vidale, and S. E. Cochran, Low-velocity damaged structure of the San Andreas fault at Parkfield from fault-zone trapped waves, Geophy. Res. Lett., 31, L12S06, 2004.

Li, Y.G., R.D. Catchings, and M.R. Goldman, Subsurface fault damage zone of the 2014 Mw6.0 South Napa, California, earthquake viewed from fault-zone trapped waves, Bull. Seismo. Soc. Am. 106, 2747-2763, doi:10.1785/01201600310, 2016.

Li, Y. G., Spatial-temporal characterization of the San Andreas Fault by fault-zone trapped waves at seismic experiment site, Parkfield, California, Earthquake Science, 34(3):261-285, doi:10.29382/eqs-2021-0014, 2021.

Li, Y.G., Numerical investigations and observations of waveguide effects for multi-layer fault damage zones at seismogenic depths. In: Li, YG., et al. (eds) Scientific Investigation of Continental Earthquakes and Relevant Studies. Springer, Singapore. https://doi.org/10.1007/978-981-96-2826-1_10, 2025.

Papageorgiou, A. S., The barrier model and strong ground motion, Pure appl. Geophys., 160, 603-634 0033 - 4553/03/040603 – 32, 2003.

Papageorgiou, A. S., and K. Aki, A specific barrier model for the quantitative description of inhomogeneous faulting and the prediction of strong motion, I, Description of the model, Bull. Seismol. Soc. Am., 73, 693-722, 1983.

Vidale, J.E. (1989). Finite-Difference seismograms for SH waves, Bull. Seismo. Soc. Am., 78, 1765-1782.

Correspondence Author

Yong-Gang Li, Ph.D.
E-mail: ygli@usc.edu

建设韧性社区：洛杉矶市的原则、策略与经验

Building Resilient Communities: Principles, Strategies, and Lessons from the City of Los Angeles

颜利平 / Liping Yan

Los Angeles Department of Water and Power

ABSTRACT

Natural disasters continue to impose severe human, economic, and infrastructural losses worldwide. Among these hazards, earthquakes remain uniquely destructive due to their sudden onset and potential to trigger cascading failures across transportation networks, water and power systems, telecommunications, and healthcare services. In response, the concept of community resilience has emerged as a foundational principle for disaster risk reduction and sustainable urban development.

This paper provides a comprehensive examination of community resilience, including its theoretical underpinnings, multidimensional characteristics, and relationship to the disaster management cycle. Particular emphasis is placed on engineering resilience, lifeline protection, emergency preparedness, and community engagement. The City of Los Angeles is presented as a case study illustrating how resilience principles can be operationalized through public policy and engineering initiatives, notably the Resilience by Design program. The paper concludes with recommendations for governments, engineers, planners, and citizens seeking to strengthen resilience in hazard-prone regions.

INTRODUCTION

The frequency and severity of natural disasters have increased markedly in recent decades, driven by rapid urbanization, aging infrastructure, population growth, and climate-related hazards. Earthquakes, in particular, pose significant risks to modern cities due to their unpredictability and capacity to disrupt interconnected lifeline systems.

California, situated along the Pacific Ring of Fire, is among the world's most seismically active regions. Major faults such as the San Andreas Fault are capable of producing catastrophic earthquakes that threaten millions of residents and critical infrastructure.

Although earthquakes cannot be prevented, advances in engineering, earth sciences, urban planning, and emergency management have substantially improved society's ability to reduce disaster impacts. Contemporary disaster risk reduction emphasizes not only response but also the development of communities capable of resisting, adapting to, and rapidly recovering from disruptive events. Community resilience has therefore become a central objective of modern disaster management.

UNDERSTANDING COMMUNITY RESILIENCE

Definition

The United Nations International Strategy for Disaster Reduction (UNISDR, 2009) defines resilience as:

"The ability of a system, community or society exposed to hazards to resist, absorb, accommodate to, and recover from the effects of a hazard in a timely and efficient manner, including through the preservation and restoration of its essential basic structures and functions."

This definition underscores several core elements: resistance, absorption, adaptation, rapid recovery, continuity of essential services, and post-event improvement. Resilience is thus a dynamic, iterative process rather than a static condition.

Dimensions of Resilience

Scholars commonly conceptualize resilience as comprising multiple interdependent dimensions:

- Physical resilience – robustness of buildings and infrastructure
- Social resilience – community cohesion, education, and health
- Economic resilience – diversity, stability, and continuity of economic activity
- Environmental resilience – natural systems that mitigate hazards
- Institutional resilience – governance, policy frameworks, and organizational capacity

Together, these dimensions determine a community's ability to withstand and recover from disasters.

IMPORTANCE OF COMMUNITY RESILIENCE

Earthquakes and other disasters produce impacts that extend far beyond physical damage. Consequences may include:

- Loss of life
- Structural collapse
- Lifeline disruption
- Business interruption
- Long-term economic decline
- Psychological trauma

Communities that invest in resilience consistently demonstrate:

- Reduced fatalities
- Lower reconstruction costs

- Faster economic recovery
- Increased public trust
- More sustainable development trajectories

Resilience should therefore be understood as a long-term investment in societal stability and prosperity.

THE DISASTER MANAGEMENT CYCLE

The disaster management cycle comprises four interconnected phases.

Mitigation

Mitigation aims to reduce future losses before hazards occur. Key measures include:

- Seismic building codes
- Structural and non-structural retrofits
- Land-use regulation
- Slope stabilization
- Flood protection
- Lifeline strengthening

Engineering-based mitigation often yields the greatest long-term economic benefits.

Preparedness

Preparedness enhances readiness for emergency conditions through:

- Emergency planning
- Public education
- Earthquake drills
- Emergency supply storage
- First-responder training
- Early warning systems

Prepared communities respond more effectively under stress.

Emergency Response

Immediately following a disaster, priorities include:

- Search and rescue
- Medical care
- Fire suppression
- Emergency shelter
- Damage assessment
- Utility restoration

Effective interagency coordination is essential for minimizing losses.

Recovery

Recovery encompasses both short-term and long-term activities:

- Infrastructure reconstruction
- Housing recovery
- Economic revitalization
- Mental health support
- Updating codes and standards
- Incorporating lessons learned

Recovery should aim to "build back better," enhancing resilience rather than restoring pre-disaster vulnerabilities.

CHARACTERISTICS OF RESILIENCE COMMUNITIES

According to ARUP (2011), resilient communities exhibit several defining characteristics.

Knowledgeable and Healthy Citizens

Public awareness and participation strengthen:

- Risk perception
- Household preparedness
- Emergency response capability
- Recovery effectiveness

Effective Organization

Resilient communities benefit from:

- Strong leadership
- Clear roles and responsibilities
- Coordinated emergency plans
- Efficient decision-making structures

Social Connectivity

Social cohesion facilitates:

- Resource sharing
- Volunteer mobilization
- Rapid communication
- Mutual aid

Reliable Infrastructure

Critical infrastructure systems include:

- Transportation networks
- Water supply systems

- Electric power systems
- Telecommunications
- Hospitals and emergency facilities

Infrastructure resilience ensures continuity of essential services.

Economic Strength

Economic resilience is supported by:

- Diverse employment sectors
- Financial reserves
- Insurance coverage
- Business continuity planning

Environmental Stewardship

Healthy natural systems contribute to:

- Flood mitigation
- Slope stability
- Water resource protection
- Ecosystem resilience

ENGINEERING CONTRIBUTIONS TO RESILIENCE

Engineering plays a foundational role in reducing disaster risk. Key responsibilities include:

- Earthquake-resistant structural design
- Geotechnical hazard evaluation
- Foundation engineering
- Lifeline system protection
- Seismic retrofitting
- Performance-based design
- Quantitative risk assessment

Geotechnical engineers, in particular, assess hazards such as liquefaction, landslides, fault rupture, settlement, and foundation performance—critical factors in reducing seismic risk before construction.

Modern engineering increasingly emphasizes system-level resilience, recognizing that infrastructure networks must function collectively during and after disasters.

THE CITY OF LOS ANGELES RESILIENCE BY DESIGN PROGRAM

Los Angeles has emerged as a global leader in seismic resilience. Guided by the Mayoral Seismic Safety Task Force, the City launched the Resilience by Design program (2014) to:

- Protect lives
- Strengthen emergency response
- Accelerate recovery
- Safeguard regional economic stability

The Task Force identified four major vulnerabilities:

- Older concrete buildings
- Soft-story buildings
- Water infrastructure
- Telecommunications systems

Building Strengthening

Building collapse is the primary cause of earthquake fatalities. Los Angeles enacted mandatory retrofit ordinances targeting:

- Soft-story residential buildings
- Non-ductile reinforced concrete buildings

Complementary initiatives include seismic safety ratings, business continuity programs, and mandatory reconstruction following severe damage.

Water System Resilience

Water systems are essential for firefighting, public health, and recovery. The City's strategy includes:

- Alternative firefighting water systems
- Aqueduct reinforcement
- Expanded storage capacity
- Local water development
- Seismic-resilient pipe networks
- Long-term planning for system reliability

Telecommunications Resilience

Reliable communication is critical for emergency coordination. Los Angeles has prioritized:

- Internet continuity
- Cellular tower protection
- Backup communication systems
- Earthquake early warning technologies

LADWP Water System Resilience Program

LADWP has implemented a comprehensive seismic resilience strategy involving:

- A dedicated resilience task force
- System-wide seismic evaluations

- Continuous monitoring of critical assets
- Resilient pipe network development
- Enhanced water storage
- Expanded emergency water supplies
- Improved emergency response capability

Representative projects include installation of earthquake-resistant pipelines and retrofitting tunnels that cross active faults.

EMERGING TECHNOLOGIES

Future resilience efforts will increasingly rely on advanced technologies such as:

- Artificial Intelligence (AI)
- Internet of Things (IoT)
- Structural Health Monitoring
- Remote Sensing
- Geographic Information Systems (GIS)
- Digital Twins
- Unmanned Aerial Vehicles
- Satellite Monitoring

These tools enhance real-time monitoring, hazard assessment, and decision-making across all phases of disaster management.

FUTURE CHALLENGES

Despite progress, significant challenges remain:

- Aging infrastructure
- Climate change
- Population growth
- Limited funding
- Increasing infrastructure interdependence
- Cybersecurity risks
- Social inequality

Addressing these challenges requires sustained investment, scientific innovation, and coordinated public policy.

RECOMMENDATIONS

To strengthen community resilience, stakeholders should:

- Continue improving seismic building codes

- Accelerate infrastructure retrofit programs
- Expand public education and preparedness initiatives
- Enhance regional emergency coordination
- Develop resilient lifeline systems
- Integrate advanced monitoring technologies
- Promote interdisciplinary collaboration
- Increase investment in resilience research
- Encourage private-sector continuity planning
- Foster a culture of resilience across society

The Chengdu Declaration's five-point action strategy (UNISDR, 2011) remains a valuable framework for ongoing implementation.

CONCLUSION

Community resilience represents one of the most critical challenges in contemporary disaster management. Achieving resilience requires coordinated efforts among governments, engineers, scientists, businesses, and citizens.

The City of Los Angeles demonstrates that proactive planning, mandatory seismic retrofits, lifeline strengthening, and community engagement can significantly reduce earthquake risk. As natural hazards continue to threaten rapidly urbanizing regions, resilience must become an integral component of sustainable development.

Ultimately, a resilient community is one that not only withstands disasters but also adapts, learns, and emerges stronger. Investing in resilience today protects lives, preserves economies, and ensures a safer future for generations to come.

REFERENCES

ARUP International Development (ARUP, 2011). Characteristics of a Disaster Resilient Community.
City of Los Angeles (2014). Resilience by Design Report.
Los Angeles Department of Water and Power (LADWP, 2014). Water System Seismic Resilience Program.
United Nations International Strategy for Disaster Reduction (UNISDR, 2009). Terminology on Disaster Risk Reduction.
United Nations International Strategy for Disaster Reduction (UNISDR, 2011). Chengdu Declaration for Action.

Correspondence Author

Liping Yan, Ph.D.
E-mail: lipingyan.ge@gmail.com

美国棕地修复方法可持续性探讨

A Discussion of the Sustainability of Brownfield Site Remediation Methods in the US

容躍 / Yue Rong

California Regional Water Quality Control Board
Los Angeles Region (Retired)

ABSTRACT

Sustainability in environmental site remediation has become a new focus of discussion lately. There are 3 components in the definition of sustainability: (1) to consider environmental, economic, and social factors, (2) to have clear benefits when doing site remediation, and (3) to make balanced decisions between environmental protection and economic growth. The concept of sustainability may mean energy and resource conservation, less emission to the environment, use of green technology, maintaining human health and safety, reducing environmental risk, climate change consideration, cost-effectiveness, and public acceptance. This paper conducts a sustainability analysis of each common site remediation method, such as soil vapor extraction and in-situ treatment. Each remediation method's advantages and disadvantages in terms of sustainability are discussed. The paper also presents a cost analysis based on 179 closed underground storage tank cleanup sites in the Los Angeles area. Finally, the paper presents a case study of a soil reuse site where petroleum-impacted soil was reused after remediation as a mix of road pavement material.

Keywards: Sustainability; environmental remediation; brownfield; site cleanup methods; soil vapor extraction; cost analysis; environment; economic; social factors

DEFINITION OF SUSTAINABILITY

Sustainability in environmental site remediation has become a new focus of discussion lately (e.g., Hadley et.al. 2015). According to the definition from Wikipedia, the sustainability of environmental site remediation is defined as "the practice of demonstrating, in terms of environmental, economic and social indicators, that the benefit of undertaking remediation is greater than its impact, and that the optimum remediation solution is selected through the use of a balanced decision-making process."

Another definition can be seen in Ellis and Hadley (2009), as ". . . sustainable remediation is broadly defined as a remedy or combination of remedies whose net benefit on human health and the environment is maximized through the judicious use of limited resources."

These two definitions show commonality and similarity. To summarize, there are 3 components in the definition: (1) to consider environmental, economic, and social factors, (2) to have clear benefits when doing site remediation, and (3) to make a balanced decision between environmental protection and economic growth.

In site remediation industry terms, the concept of sustainability may mean energy and resource conservation, less emission to the environment, use of green technology, maintaining human health and safety, reducing environmental risk, climate change consideration, cost-effectiveness, and public acceptance.

Figure 1 shows the relationship among the environmental, economic, and social elements. As seen in the figure, in order to be sustainable, these three elements need to overlap. Overemphasizing one factor over others will only cause problems down the road. For example, overemphasizing environmental remediation will negatively impact economic growth. Our way of living in society will be impacted by environmental pollution and also low economic growth. Therefore, a balanced decision in setting the cleanup goal is necessary between environmental cleanup and economic development. For example, how to set up a remedial goal is a method involving these 3 factors. To do so, new technologies and limited resources should be considered. To clean up to a non-detect level or background level is certainly an ideal goal for site remediation. However, to remove low concentrations from the environmental media like soil or groundwater is technically challenging and not very cost-effective. Therefore, the goal of non-detect may not be sustainable. If the cleanup site is located in a very populated residential area, then the remedial goal should be more stringent than otherwise in a remote area. A balanced decision in setting up environmental site remediation goals is the key to sustainability.

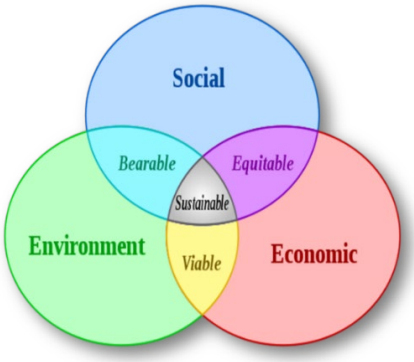


Figure 1. Relations in sustainability (Arulanantham 2018

In summary, balancing the environmental, economic, and social factors is like a string to hold all factors, as demonstrated in Figure 2 below. Case-specific consideration will determine the percentage of each factor, e.g., the economic factor may need 50% of the fraction for a site.

Figure 2: Balancing the three factors

Figure 2. Balancing the three factors

BROWNFIELD OVERVIEW AND DEFINITION

The term "brownfield" in the US is used to describe abandoned land where no grass grows (compared to a "green field"). The definition of brownfield is that a property, its expansion, redevelopment, or reuse is hindered by environmental pollution or impact. Any program that is created and associated with investigating and remediating brownfield sites is called a "Brownfield program." In the US, approximately 450,000 sites that used to be commercial and industrial sites were abandoned due to environmental impact (USEPA 2016). To revitalize the economy, those brownfield sites need to be remediated to get environmental clearance and closure, and then brought back to productive use of the land. The remediation of brownfield or any other kind of contaminated sites should also be considered under the regional surface watershed management. This is because some cleanup methods or activities may or may not be positive to entire watershed management, especially in urbanized areas where impacted sites are concentrated, and remediation methods may demand more resources and energy. To remediate those brownfield sites, various remediation methods must be evaluated. During the evaluation, the sustainability of each method should also be analyzed.

REMEDIAL METHODS SUSTAINABLE ANALYSIS

Table 1 below shows site remediation method mentality over time. In the 1960s (about 60 years ago), when talking about site cleanup, actions were to discard any waste to get out of sight and get out of mind. In the 1990s (about 30 years ago), the regular site cleanup methods were either to dig out the contaminants, pump them out, bury them, or send them to incinerators to burn them out. In this period, many kinds of treatment methods were developed. Now, new technologies of treatment methods are still developing and more and more remediation methods are developed with considerations of recycling, reusing, and using natural attenuation and bioremediation. The new consideration, or more "green" mentality in site remediation, is more or less to pursue energy conservation, less emissions, and cost-effectiveness, which is more sustainable.

Whether a remedial method is sustainable depends upon the analysis if it is effective and efficient. This section will analyze common environmental site remediation methods in terms of each method's advantages and disadvantages in the perspective of sustainability. More detailed discussions of these remedial methods can be seen in Rong (ed.) (2018).

Table 1: Site Remediation Method Evolvement

60 years ago	30 years ago	Now
Discard	Dig Pump Bury Burn	Recycle Reuse Natural Attenuation Biodegradation

The following methods are discussed:

- Soil excavation
- Groundwater pump and treat
- Soil vapor extraction
- Thermal enhancement
- Dual-phase extraction and air sparging

- In-situ treatment: chemical and biological
- Monitoring natural attenuation
- Phytoremediation
- Permeable Reactive Barriers (PRBs)
- Vapor barrier installation

Soil excavation

Soil excavation is to remove contaminated soil and dispose of it (Figure 3). It is good for a quick solution, particularly for sites to be re-developed soon. The advantages of soil excavation also include completeness of the removal of a pollution source area, and good for fine-grained materials where soil vapor extraction or in-situ injection does not work effectively. The disadvantages of soil excavation include cost, interruption to the site business, limited to the open area for operation, and usually not used for removing deep soil contamination. Ultimately, excavation does not destroy the contaminants; in some cases without treatment, it simply transfers them from one place to another.

Figure 3. Soil excavation site.

In summary, soil excavation is

- to be used as a quick solution for redevelopment
- a better choice for fine-material soils
- remove contamination sources
- usually costly
- Interruptive

Groundwater pump and treat

Groundwater pump and treat (P&T) involves pumping underlying groundwater to the surface for treatment (Figure 4). After groundwater is pumped to the surface, many physical, chemical, and/or biological methods can be used to treat and

remove pollutants. It is more effective to remove soluble contaminants, such as hexavalent chromium and perchlorate. As existing dissolved contaminants are removed by groundwater pumping, more contaminants will enter the groundwater by desorbing, dissolving, precipitating, and/or diffusing in aquifer materials. As the remediation progresses, contaminant concentration gets lower and lower, and the massrecovery reaches an asymptotic level. If the treatment goal is not achieved before this asymptotic level, this treatment method becomes less cost-effective. The post-treated water is generally either discharged to the sewer system or re-injected back to the aquifer to conserve water resources.

The advantages of P&T method include the choice of many cleanup methods to treat particular contaminants of interest, and it can be less disruptive to the onsite operation and business. The disadvantages are costly, and it may not conserve water resources if discharging to sewer or surface water. It also may not be a good choice to treat low concentrations of contaminants.

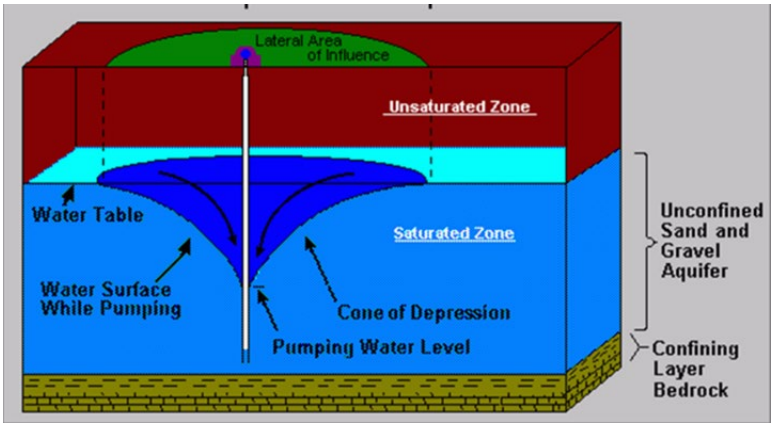


Figure 4. Schematic of groundwater pump and treat

In summary, groundwater pump and treat (P&T) is

- to allow more choice of treatment methods
- less interruptive
- usually costly
- waste water resources if not reinjected back to aquifer
- not cost-effective for low concentration site
- need large area for treatment system construction

Soil vapor extraction

Soil vapor extraction (SVE) uses vacuum power to extract the vapor phase of volatile contaminants in soil (Figure 5). It is effective in coarse-media soil and volatile contaminants, and can be operated continuously over time until it is complete. The cost is moderate. Like groundwater P&T, the contaminant mass removal rate of SVE diminishes as treatment progresses, and the vapor concentration will also show rebound after it stops. Unlike groundwater P&T, treatment options for the SVE process are limited to those for volatile organic constituents. Commonly used vapor treatment options include granular activated carbon (GAC), catalytic oxidation and thermal oxidation. Extracted soil vapor needs to be treated to remove contaminants before being discharged to the atmosphere, which may require an extensive air quality permitting process, which can be a disadvantage of SVE technology. In addition, other disadvantages of SVE include that it may be ineffective in fine-grained soil, it usually requires a sizeable compound for the system and equipment, and it can be very noisy in a residential neighborhood.

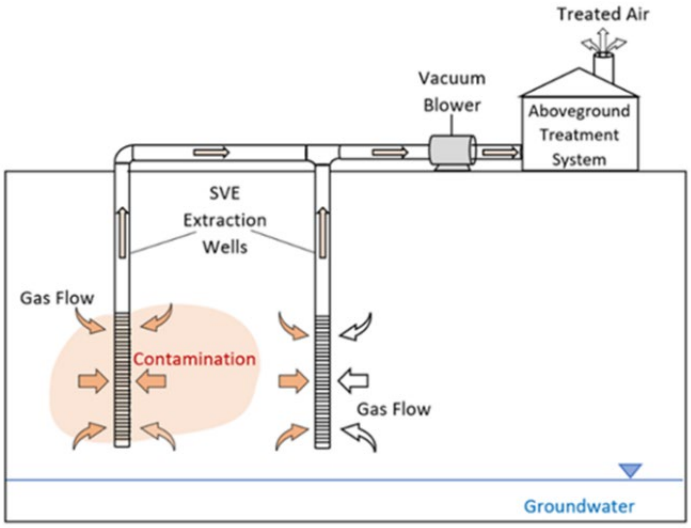


Figure 5. Schematic of soil vapor extraction.

In summary, Soil vapor extraction (SVE) is

- effective in coarse materials
- moderate cost
- rebound concentrations
- ineffective in fine-grained soil
- needs a large area for constructing equipment and system
- noisy

Thermal enhancement

Thermal enhancement uses heat to encourage more volatilization in the subsurface for volatile compound removal (Figure 6). This method is usually used in conjunction with soil vapor extraction (SVE). In addition to the vadose zone, thermal enhancement can also be applied in saturated zone contamination. The advantages of thermal-enhanced SVE are that it is effective for the heating method used for tight soils where SVE does not work effectively, and it is also effective for semi-volatile contaminants, which may be difficult for soil vapor extraction to work effectively under ambient temperature conditions.

Figure 6. Soil heating probe

Another advantage of thermal treatment is to accelerate the cleanup in soil or groundwater. The disadvantage of the heating method is energy consuming, and therefore costly. Thermal treatment is normally the preferred method for sites with expedited redevelopment plans and ample financial resources. Due to heat transport in the soil media, thermal enhancement may have only a limited impact area.

In summary, thermal enhancement method is

- good for fine materials
- treating semi-volatile contaminants
- accelerating cleanup (shorten time frame)
- energy consuming
- usually costly
- interruptive

Dual-phase extraction and air sparging

Dual-phase extraction and air sparging are used to extract vapor and dissolved phases of contaminants, or sometimes NAPL, and to use air injection to the subsurface to encourage volatilization (Figure 7). The advantage of multi-phase extraction is to clean up multiple phases of pollutants at one location. It works well for pollutants distributed in the capillary fringe, which is a difficult area to remediate. Like soil vapor extraction, it can run continuously over time until it is complete. Its disadvantage is to have a too complicated system in piping and engineering units and to take a lot of surface space for the operating units. Sometimes it may not work well without a clear understanding of the different phases of pollutant distribution in the subsurface.

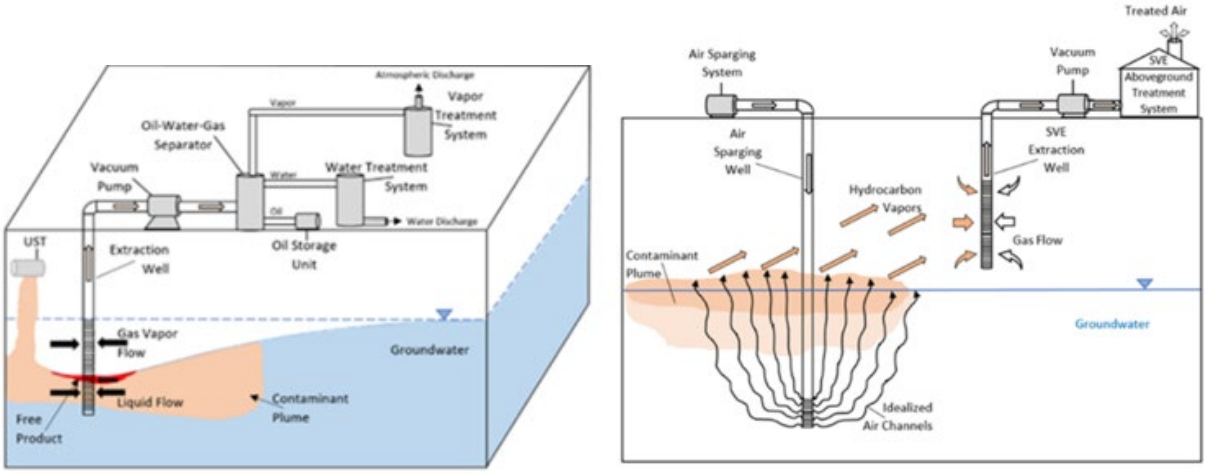


Figure 7. Schematic of dual-phase extraction (left) and air sparging (right).

In summary,

- effective in coarse materials, and multi-phase in the capillary fringe
- moderate cost
- complicated system in engineering construction
- noisy

Chemical and biological In-situ treatment

Chemical and biological In-situ treatment involves injecting chemical and biological agents into the impacted area or zone to allow chemical or biological processes to remediate contamination (Figure 8). The advantage of in-situ injection is to conserve water resources. It is not necessary to pump the water out of the aquifer and then put it back in. Another advantage for in-situ injection is low cost. The disadvantage of in-situ injection is that it is difficult to monitor the effect of the in-situ injection, which may only have very limited impact around the injection points.

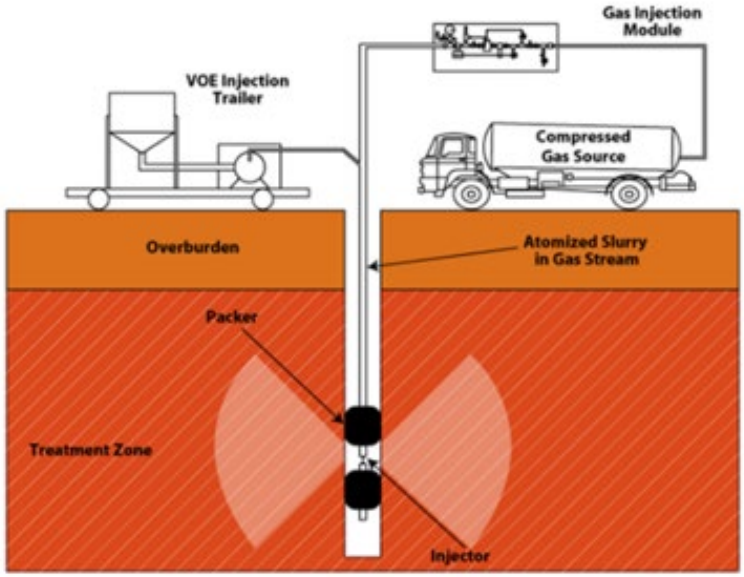


Figure 8. schematic of In-situ injection.

In-situ chemical injection method may work very effectively to reduce the residual concentration as a polishing step after a major physical removal method. However, this method may not be effective to treat very high concentration pollution plumes, such as those in NAPL sites. The biggest concern for the biological injection method is adding bacteria or other microbial species to the water resources. Bacteria injection is not like chemical injection. Chemicals will eventually be diluted in the environment after stopping injection. However, bacteria can grow exponentially even after stopping injection. It is a challenge to control the unwanted growth and just have the bacteria consume and destroy pollutants. It is hoped to have the right species that is not harmful to human beings, but can be effective to remediate the contamination. In the meantime, controlling the scale and quantity of the biological materials to be injected can be done to minimize potential impact on humans.

In summary, chemical and biological In-situ treatment is

- low cost
- less interruptive
- water conservation
- good for the polishing stage in remediation
- impact area of injection may be small
- may not be effective to treat high concentrations
- hard to verify results
- biological agent concern

Monitoring natural attenuation

Monitoring natural attenuation (MNA) is to monitor reduction of contaminant concentrations by natural means. This is not an active remedial method. As seen in Figure 9, the natural attenuation parameter monitoring is not always that straightforward; sometimes it is not very clearly indicative. Therefore, one should use multiple lines of evidence to make a judgement to see if the natural attenuation is indeed a possibility in the subsurface environment. The multiple lines of evidence may include the trend of concentration and mass, the geochemistry parameters (e.g., Fe2+, nitrate), and new technologies like isotope and DNA techniques. The type of contaminants, the subsurface environment, the maximum concentration and distribution, the contamination history, and the estimate of total discharge to the subsurface should also be evaluated in the framework of monitoring natural attenuation. The advantages of the natural attenuation method are economical, good for the polishing stage after active remediation,

desirable for project cost-prohibitive, at a polluted site that is not completely accessible, or other remedial methods may not be feasible. The disadvantages are it is slow, may not work in site-specific conditions, and monitoring may become costly over time.

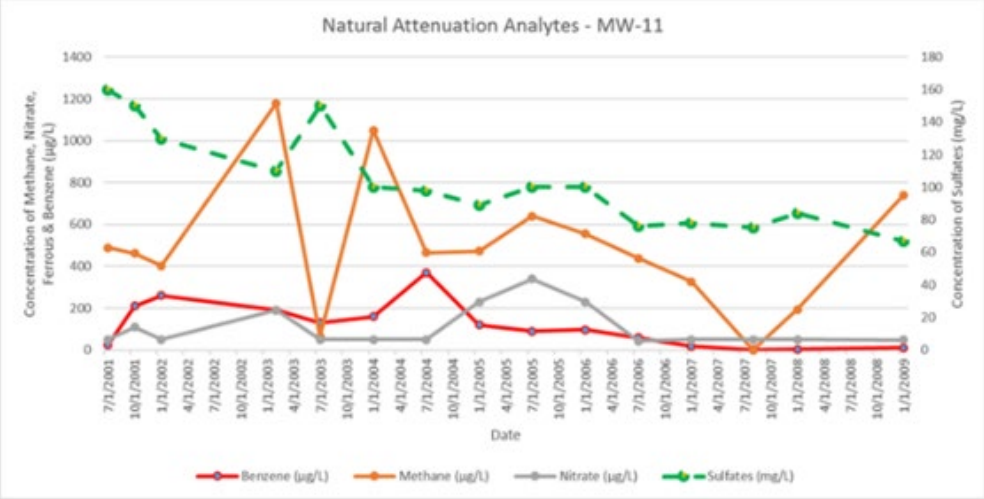


Figure 9. Monitoring natural attenuation geochemistry parameter trend.

In summary, monitoring natural attenuation (MNA) is

- low cost
- good for physically constrained and not accessible sites
- good for the polishing stage of remediation
- less interruptive
- may be slow to see the result
- difficult to verify results
- monitoring for a long time may cause increasing cost

Phytoremediation

Phytoremediation uses natural living green plants to remove, degrade, or break down the soil and/or groundwater contaminants in-situ (Figure 10). This method depends more upon species-specific characteristics, soil medium, and type of pollutants. Green plants' intake may or may not be discriminatory to different types of pollutants, for example, metals or organics. The advantage of phytoremediation is the low cost in most situations. The disadvantages are that

this method may be very slow, the cleanup rate is dependent on seasonality, difficult to remediate deeper zones of contamination beyond the plant root zone, and difficult to monitor the results of the method. In addition, the high pollutant concentrations may cause harm to the plants. at stations within the CF compliant zone (Figs. 7a, b). While, only P'FZTW with ~4-s duration appears in the vertical-component seismograms because the ray path from this deep teleseismic earthquake to the CF array is nearly vertical. Normalized spectral amplitude contours of teleseismograms recorded at on-fault station B15 show large-amplitude P-type and S-type FZTWs at 1-2 Hz with ~4-s and ~7-s durations, but weaker and much shorter (2-s) wavetrains at stations B1 and B26 outside the compliant zone (Fig. 7c). Figs. 7d, 7e show similar features of teleseismograms and speectral amplitudes at Lines A, C and D. Fig. 9f shows receiver functions of teleseismograms recorded at partial stations of Lines A, B, C and D. Large-amplitudes appear in time windows of ~3.5 s and ~7 s, respectively, following P' and S' as well as multiple-phase P" and S" at stations within the CF compliant zone, owning to P-type and S-type fault-zone trapped waves produced within the CF compliant zone. In contrast, much lower amplitudes are shown in the same time windows in the receiver functions of teleseismograms recorded at stations outside the compliant zone.

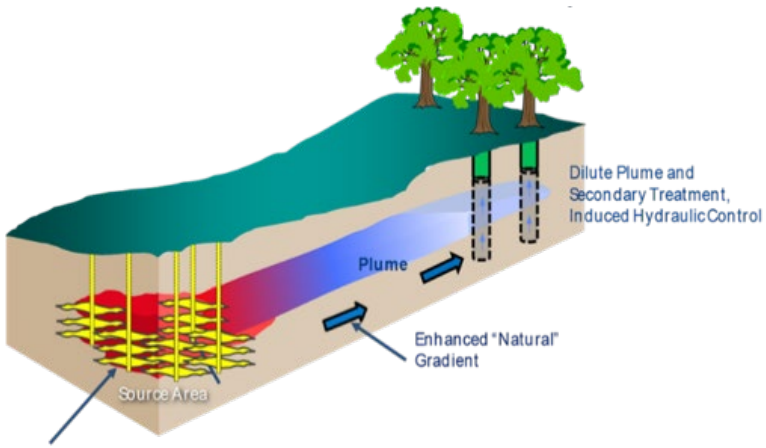


Figure 10. Schematic of phytoremediation (Edward Gatliff & Doug Riddle, 2018).

In summary, phytoremediation is

- a green technology
- low cost
- may work very slowly
- interruptive
- not good for deep zone contamination
- difficult to verify results

Permeable Reactive Barriers (PRBs)

Permeable Reactive Barriers (PRBs) use a man-made infiltration material barrier to treat passing groundwater (Figure 11). The materials stuffed in the barrier usually include sorption materials like GAC, or in-situ chemical or biological treatment materials like oxidation compounds. The advantages of PRBs include in-situ treatment to conserve water resources, and effective treatment of migrating plumes and dissolved contaminants, particularly in high concentrations. Unlike in-situ treatment, this method may be costly and more disruptive. The most significant disadvantage of all is that the PRBs could hardly be replaced and refilled in place. When the barrier materials are saturated with the contaminants, the remedial life could not be sustained.

Figure 11. Schematic of Permeable Reactive Barriers (PRBs).

In summary, Permeable Reactive Barriers (PRBs) is

- water conservation
- effective to treat high concentration and migrating plumes
- costly
- interruptive
- could not be replaced in-situ

Vapor barrier installation

The vapor barrier installation method uses a barrier to prevent a risk exposure pathway to humans (Figure 12). This method is not a treatment method, rather a mitigation method. For volatile pollutants, one of their potential impacts on human health is to transport through the building floor into people's living spaces. It is called "vapor intrusion." In addition to the soil vapor extraction method, one method to remediate the vapor intrusion situation is to install a vapor barrier between the contaminated soil or groundwater and the floor of the building. The vapor barrier usually contains a solid surface and sandwiches a porous layer of sorption material like activated carbon. It is effective and low cost. But it is not an active cleanup method and cannot clean up the contamination.

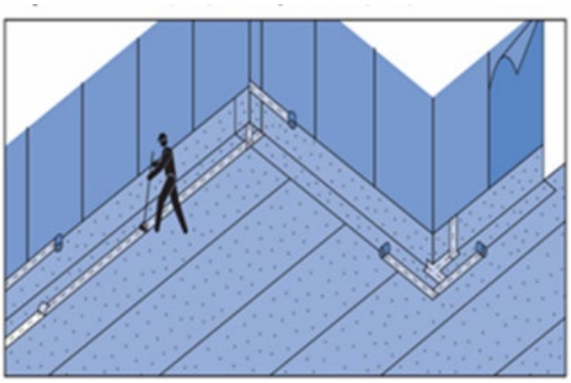


Figure 12. Vapor barrier installation.

COST ANALYSIS FOR SITE REMEDIATION

If a remedial method also depends upon the analysis of cost-effectiveness. Economic consideration is always important technically and engineeringly to site remedial projects. This section presents a site remediation method cost analysis based on a dataset in the California underground storage tank (UST) cleanup fund in Los Angeles, California, USA. The cost analysis used 179 closed UST cases in the Los Angeles area since 2012, where active remediation was conducted. The remedial methods for cost analysis include (1) soil vapor extraction, (2) dual phase extraction, (3) soil excavation, (4) groundwater pump and treat, and (5) free product removal. If a cleanup site had more than one method applied (e.g., both soil vapor extraction and soil excavation were used), the most costly method at the site will be counted into the group with the same method. For example, if a cleanup site applied both soil vapor extraction and soil excavation, and soilexcavation is more expensive as a result, then this site will be classified to the soil excavation group, not the soil vapor extraction group for analysis. The average cost associated with different active remedial methods per site is summarized and presented in Table 2.

Table 2: Average cost by individual method per site in $

Method	Cost per site
Soil vapor extraction	$926,173 (n=68)
Dual phase extraction	$907,440 (n=21)
Soil excavation	$870,803 (n=80)
Groundwater pump and treat	$639,786 (n=6)
Free product removal	$481,953 (n=1)
Total average	$887,438 (n=179)

It needs to be noted that the cleanup at UST-impacted sites deals primarily with petroleum hydrocarbons as contaminants. The cost may be different from treating other types of contaminants. For example, conventional soil excavation is considered a very expensive method for cleanup. However, data in Table 2 suggests its cost ($870K) is below the total average cost ($887K). This may be because the excavation is only conducted in the small source area (e.g., UST area). In addition, groundwater pump and treat is traditionally considered a high-cost method. However, the data in Table 2 show its cost below the total average. This may be because the sample size (n) is too small (only 6 sites used the pump-and-treat method), and therefore the result might be biased. Soil vapor extraction is considered to be a cost-effective method. However, the data in Table 2 suggests this method is the most expensive method compared with other methods. This may be because it lasted too long, and maybe a mobile soil vapor extraction unit, which is more expensive, was used. Again, this dataset in the Los Angeles area may be too small and therefore biased and the multiple remedial methods used at the site may cause uncertainties as well.

CASE STUDY: REUSE OF REMEDIATED SOIL

One of the sustainable remedial methods is to reuse the soil after remediation. The section presents a case study where soil after remediation is used as material for road pavement. The US Environmental Protection Agency (USEPA)

published a guideline for "The use of soil amendments for remediation, revitalization, and reuse" (2007). The guideline discussed the type of soil contamination (e.g., petroleum, metals, volatiles, etc.), exposure pathway evaluation based on the contaminated soil, soil type and ecosystem function, reuse proposal and application, operation of soil reuse, permitting from government agencies, and environmental monitoring.

The case example presented is a soil reuse case where remediated soil is used to mix with asphalt to pave a road onsite at an oil and gas production field. Soil was impacted by petroleum products and treated with natural degradation. From an environmental perspective, this proposal is to reuse the material at the site, to reduce waste, and not to cause risk to humans and the surrounding ecosystem. From an economic perspective, it certainly reduces some cost by using onsite material. From a social perspective, since this site is located in a less populated and remote area of an oil and gas production field, the plan to reuse remediated soil will not cause too many neighborhood concerns. With that, the oil and gas field operator submitted a proposal to the lead government agency for soil reuse. The government agency issued a permit to reuse the treated soil. This is considered a sustainable solution for the situation. The permit specified the following conditions for reuse of the soil.

Reuse of impacted soil with oily sand for road pavement materials must conform to the following conditions.

(1) limit 1001 yard3 oily sand (quantity limitation)

(2) soil is from the oil storage pond bottom (location specification)

(3) the soil will be mixed with asphalt

(4) Concentration must be less than 6% Total Petroleum Hydrocarbons

(5) Only use for pavement of a road (usage specification)

(6) Asphalt materials for the road must be at least 90%

(7) Any actions must be approved by the permitting agency.

In addition, treated soil should be monitored and sampled. The composite soil samples should be obtained from the soil stockpiles, and the number of soil samples is taken per the guideline shown in the following table.

Table 3: Number of soil samples guideline for soil stockpile (DTSC 2001)

Volume of stockpile	Samples per volume
Up to 1000 yard^3	4
1000 – 5000 yard^3	4 + 1 every 500 yard^3
> 5000 yard^3	12 + 1 every 1000 yard^3

REFERENCES

Arulanantham, R. (2018). Chapter 3, Fundamentals of Environmental Site Assessment and Remediation (Rong ed). CRC Press Taylor & Francis Group (ISBN-13: 978-1138105157), Boca Raton, FL 33487, USA.

DTSC (2001). California Department of Toxic Substances Control. Information Advisory – Clean Imported Fill Material (October). https://www.dtsc.ca.gov/Schools/upload/SMP_FS_Cleanfill-Schools.pdf

Ellis, D. E., and Hadley, P.W. (2009). Sustainable Remediation White Paper – Integrating Sustainable Principles, Practices, and Metrics into Remediation Projects. REMEDIATION in Wiley Online Library (wileyonlinelibrary.com) http://www.sustainableremediation.org/library/issue-papers/

Gatliff, Edward and Riddle, Doug (2018). Technical presentation of phytoremediation at Los Angeles Regional Water Quality Control Board (April).

Hadley, P.W., Arulanantham, R, Gandhi, D. (2015). Yardsticks to integrate risk assessment, risk management, and groundwater remediation. REMEDIATION in Wiley Online Library (wileyonlinelibrary.com) http://www.sustainableremediation.org/library/issue-papers/

Rong, Y. (editor) (2018). Fundamentals of Environmental Site Assessment and Remediation. CRC Press Taylor & Francis Group (ISBN-13: 978-1138105157), Boca Raton, FL 33487, USA.

USEPA (2007). US Environmental Protection Agency. The use of soil amendments for remediation, revitalization, and reuse. (December) EPA 542-R-07-013. https://nepis.epa.gov/Exe/ZyPDF.cgi/60000LQ7.PDF?Dockey=60000LQ7.PDF.

USEPA (2016). US Environmental Protection Agency. Brownfield Overview and Definition. (August) https://19january2017snapshot.epa.gov/brownfields/brownfield-overview-and-definition_.html.

Correspondence Author

Yue Rong, Ph.D.
E-mail: yrong1983@gmail.com

连接传统信号处理与边缘人工智能：感知与计算约束下的助听设备语音增强

Bridging Conventional Signal Processing and Edge AI: Speech Enhancement for Hearing Devices Under Perceptual and Computational Constraints

李明爽 / Mingshuang Li

mingshuang.li@csun.edu
Department of Communication Disorders and Sciences, California State University, Northridge, 18111 Nordhoff Street, Northridge, CA 91330, USA

Abstract

Prescriptive amplification restores audibility, yet many hearing-aid users continue to struggle in noise because amplification does not resolve target-to-masker competition. Conventional noise reduction improves comfort and effort more reliably than intelligibility, whereas spectral and formant enhancement approaches strengthen selected speech cues. Deep learning can produce larger laboratory gains, but translation to wearable devices requires causal, real-time processing within tight memory, computation, and battery budgets. This review connects evidence from conventional signal processing, cue-directed enhancement, human-listener studies, and hardware implementations. It argues that hearing-device algorithms should be judged by perceptual benefit within a stated device budget. Targeted formant enhancement illustrates interpretable auditory-informed processing, while compact neural models may support target identification, scene analysis, and control. A likely near-term direction is a hybrid architecture in which compact AI guides efficient, controlled signal processing.

Keywords: Speech enhancement; conventional signal processing; artificial intelligence; hearing aids.

1. Introduction

Modern hearing aids can restore audibility with considerable precision, yet speech in noise remains difficult for many users. Amplification cannot fully compensate for reduced frequency selectivity, temporal and binaural deficits, or the cognitive demands of following one talker among competing voices. Reverberation, moving sound sources, changing communication goals, and source-related degradation such as speech muffling add to the difficulty in everyday listening.

Prescriptive amplification, typically implemented through frequency- and level-dependent gain and WDRC, provides the foundation of hearing-aid processing by restoring audibility within the listener's reduced dynamic range (Keidser et al., 2011). Its purpose, however, is not to separate target speech from competing sound. When speech and noise overlap, amplification may make both more audible without inherently improving the target-to-masker ratio, and restored audibility alone may be insufficient at moderate-to-high noise levels. Speech enhancement addresses this complementary problem by improving access to the target under adverse acoustic conditions. It is best understood as an adaptive processing layer within the broader hearing-aid signal chain, operating alongside rather than replacing prescriptive amplification (Kayser et al., 2022).

Speech enhancement has approached this problem in several ways. Digital noise reduction (DNR) attenuates components judged to be noise, whereas spectral enhancement strengthens selected features of the target. More recently, deep neural networks have learned mappings from noisy mixtures to masks, filters, spectra, or waveforms. The perceptual consequences differ across these approaches. Noise reduction may improve comfort without improving recognition, and cue enhancement may increase the salience of selected speech information without changing the overall signal-to-noise ratio (SNR). Either approach can reduce naturalness or environmental awareness when applied too aggressively.

Deep learning has expanded what is technically possible, but offline performance says little by itself about suitability for a hearing aid. A wearable device must process sound continuously and with little delay, usually without relying on a remote server, while sharing limited computation, memory, and battery capacity with the rest of the hearing-aid system. It must also preserve loudness comfort, spatial cues, and access to relevant background sounds. In this review, Edge AI refers to causal or streaming inference performed on the device or closely coupled local hardware rather than in the cloud. The practical issue is whether AI can deliver reliable listener benefit throughout the day, in unfamiliar acoustic scenes, within the limits of the device that the listener wears.

Previous reviews have examined DNR, deep-learning speech enhancement, and AI in hearing aids (Bentler and Chiou, 2006; Andersen et al., 2021; Healy et al., 2023). The present review takes a narrower translational view. It considers what conventional signal processing has taught us about listener outcomes, what cue-targeted enhancement adds, how far neural systems have progressed, and where device constraints still limit translation. The final section considers how conventional signal processing and Edge AI may be integrated, using targeted formant enhancement as one interpretable test case rather than a complete solution.

2. Scope and Evaluation Framework

This integrative narrative review examines representative work on conventional DNR, spectral and formant enhancement, and AI-based speech enhancement for hearing devices. It is selective rather than exhaustive, with emphasis on foundational studies that established important perceptual principles and recent work providing human-listener evidence, ecological evaluation, causal or streaming operation, or measurements on resource-constrained hardware. A comprehensive taxonomy of neural architecture, microphone-array processing, dereverberation, and target-speaker extraction is beyond its scope.

The evidence is considered along two dimensions: perceptual value and device feasibility. Perceptual outcomes include speech recognition, listening effort, sound quality, preference, and environmental awareness, which should not be treated as interchangeable. Device feasibility includes causal and real-time operation, execution on or near the device, latency, memory, computation, energy use, and integration with the broader hearing-aid signal chain. Objective measures such as STOI, HASPI, PESQ, and SI-SDR remain useful during development, but they do not substitute for listener testing. For hearing devices, an improvement is meaningful only when it produces reliable listener benefit within a clearly stated device budget.

3. Conventional Signal Processing Approaches to Speech Enhancement

Conventional speech-enhancement algorithms have generally followed two broad strategies: attenuating components attributed to competing noise and strengthening perceptually relevant features of the target speech. These strategies address different failure modes, and their perceptual record provides a useful baseline for evaluating AI-based methods.

3.1 Digital noise reduction

Conventional single-channel DNR estimates whether time-frequency regions are dominated by speech or noise and reduces gain where noise is judged to dominate. Implementations have used modulation analysis, spectral subtraction, Wiener filtering, statistical noise tracking, and related gain-control methods (Bentler and Chiou, 2006). Because speech and noise overlap, attenuation can remove low-level speech components as well as the masker, while rapid gain changes can introduce audible artifacts.

The perceptual record is mixed. Conventional DNR has improved comfort and acceptable noise level more reliably than intelligibility. Laboratory and field studies found limited recognition benefit, and a meta-analysis of adult hearing-aid studies reached a similar conclusion (Bentler et al., 2008; Lakshmi et al., 2021). Even without an intelligibility gain, noise reduction may reduce listening effort or release cognitive resources (Sarampalis et al., 2009). The recurring pattern of improved comfort or reduced effort without better recognition remains an important benchmark for AI-based systems.

3.2 Spectral and formant enhancement: From broad contrast modification to targeted cues

Spectral enhancement was motivated by the loss of frequency selectivity associated with sensorineural hearing loss. Early methods narrowed formant bandwidths, amplified spectral peaks, attenuated valleys, or expanded contrast across multiple channels. Their effects were usually small or condition dependent. Some studies reported modest improvements in consonant or sentence recognition, whereas others found no intelligibility benefit or poorer sound quality (Bunnell, 1990; Simpson et al., 1990; Stone and Moore, 1992; Baer et al., 1993; Franck et al., 1999). Dynamic enhancement and individualized parameters produced larger gains in selected noise conditions, but the effect did not generalize broadly (Chen et al., 2012, 2013). Spectral enhancement has also failed when tested with spectrally sharpened sentences, underscoring that greater contrast does not automatically create more usable speech information (Nittrouer et al., 2015).

Formant-directed processing has a longer history than is sometimes recognized. Boers (1980) described formant enhancement for listeners with sensorineural hearing loss, and Summerfield et al. (1985) examined formant-bandwidth narrowing. Narrowing did not improve consonant place identification in quiet, although broadening was harmful. A later approach, contrast-enhancing frequency shaping (CEFS), increased high-frequency emphasis as a function of hearing loss and the speech spectrum. CEFS improved the auditory-nerve representation of F2 in animals with noise-induced hearing loss (Miller et al., 1999). Bruce (2004) extended the scheme to running speech and examined its compatibility with multiband compression using an auditory-periphery model. This work supplied a physiological rationale for formant-directed enhancement, but not evidence of improved perception in hearing-aid users.

A related series of behavioral studies focused specifically on F2 and, more recently, combined F2-F3 enhancement. F2 enhancement improved vowel-formant discrimination for listeners with hearing loss and produced condition-dependent benefits for older listeners and for vowel and consonant identification in noise (Woodall and Liu, 2013; Guan and Liu, 2019; Li, 2021). Extending this work to sentences, Li and Liu (2026) compared unmodified, F2-enhanced, and F2-F3-enhanced natural and N95-muffled speech. Enhancement provided no measurable benefit in quiet, whereas both approaches improved recognition overall in four-talker babble. F2-F3 enhancement was more consistent across speaking

conditions and SNRs, particularly when speech muffling and external masking occurred together. This pattern suggests a benefit from restoring broader mid- to high-frequency spectral structure, although the specific contribution of F3 could not be isolated.

The studies also point to an important boundary condition: benefit is most likely when the relevant cues are degraded, because favorable conditions leave little room for improvement. Enhancement can alter spectral tilt, formant-amplitude relations, or sound quality, and most published formant studies processed the target before it was mixed with noise even though a hearing device receives the mixture. The evidence therefore demonstrates perceptual efficacy under controlled signal access rather than readiness for real-time deployment. Formant enhancement remains useful as an interpretable test bed for selective cue modification, but its computational demand, latency, memory, and energy use have yet to be measured on target hardware.

4. AI-Based Approaches to Speech Enhancement

In practical devices, AI-based enhancement enters an established hearing-aid signal chain in which prescriptive amplification, WDRC, directionality, feedback control, and output limiting remain essential. Learned models can add estimation, separation, and adaptive control under conditions that are difficult for fixed-rule processing. The relevant question is therefore not whether a neural model performs well offline, but what useful role it can sustain within the complete device.

4.1 Neural methods and human-listener evidence

Machine-learning enhancement differs from classical estimation because the mapping from noisy input to desired output is learned from examples. Early supervised systems estimated ideal binary or ratio masks; later networks estimated continuous masks, complex spectra, adaptive filters, or time-domain waveforms (Wang and Chen, 2018). The first human-listener demonstrations produced unusually large intelligibility gains, but training and test conditions were closely matched and future frames were often available (Healy et al., 2013).

Subsequent work addressed increasingly realistic constraints. Training variation expanded across talkers, noises, corpora, recording channels, reverberation, and languages. Causal speaker-separation and dereverberation systems retained perceptual benefit while avoiding future information, although conversion to real-time-capable operation carried a cost (Healy et al., 2021). Cross-corpus studies also showed that generalization to one kind of novelty does not guarantee performance in another (Pandey and Wang, 2020). By 2023, a causal attentive recurrent network trained with large numbers of talkers and noises produced substantial average intelligibility improvements for listeners with hearing loss under varied test conditions (Healy et al., 2023). Other work has demonstrated benefit for hearing-aid users rather than relying only on normal-hearing simulations (Diehl et al., 2023).

Human-listener studies now show that neural enhancement can improve communication, although benefit remains tied to sound quality, failure modes, and the relation between training data and everyday use. Causality alone does not ensure deployability: a model may still require a long analysis window, excessive memory, or more computation than a wearable device can sustain.

4.2 Commercial devices and everyday outcomes

Recent studies have begun to evaluate commercially implemented systems. Christensen et al. (2024) compared premium hearing aids during everyday wear and found that average satisfaction was similar across devices, although

satisfaction with the DNN-based system was less dependent on ambient SNR. Fitzgerald et al. (2025) combined electroacoustic measurements, clinical speech-in-noise tests, and ecological ratings; benefits were evident, but their magnitude depended on the test and masker. Overby et al. (2025) showed that listeners' preference for DNN noise reduction also depends on how it is combined with wide dynamic-range compression (WDRC).

Folkeard et al. (2026) reached a related conclusion using commercial hearing aids configured with different combinations of directionality and noise reduction. The best overall results occurred when DNN processing was combined with beamforming, and benefits varied with masker type and spatial configuration. These interactions make it difficult to evaluate an AI module in isolation: beamforming changes the signal presented to the neural processor, and WDRC may preserve or magnify residual artifacts.

For a hearing aid, the desired output is not necessarily the closest approximation to an anechoic clean signal. It is the signal that best supports the listener's current goal while remaining natural, comfortable, spatially coherent, and informative about the surroundings. This is a different objective from maximum noise suppression.

5. Why AI for Hearing Aids Is a Distinct Engineering Problem

5.1 From offline models to real-time operation under resource constraints

Many neural enhancement systems are evaluated offline on desktop or server-class hardware, whereas a hearing aid must process a continuous audio stream on the device and without dependable access to remote computation. Offline evaluation, causal processing, real-time execution, and untethered operation are not equivalent. A causal model uses no future input, yet analysis windows, buffering, memory, or computation may still produce unacceptable delay or exceed the available processor. End-to-end delay also includes analysis and synthesis, look-ahead, and data transfer; open fittings can be particularly unforgiving (Goehring et al., 2018). Faster-than-real-time execution on an external processor is therefore insufficient evidence of hearing-aid feasibility.

Computation and energy are coupled because enhancement runs for hours from a small battery while sharing resources with the rest of the signal chain. Parameter count alone is a poor proxy for deployability: working memory, numerical precision, data movement, duty cycle, and front- and back-end processing also affect runtime and energy use. TinyLSTMs illustrated how pruning, quantization, and state-update skipping can reduce computation on a microcontroller (Fedorov et al., 2020). Recent hardware studies reported a 28-nm processor consuming 740 microwatts with 39.96 ms total latency and streaming enhancement on a programmable behind-the-ear hearable (Park et al., 2025; Itani et al., 2025). These results establish feasibility, but power and latency must still be interpreted in relation to form factor, battery capacity, expected wear time, duty cycle, and measurement boundaries.

5.2 System integration and device-level evaluation

Enhancement must operate alongside directionality, feedback control, frequency shaping, WDRC, output limiting, and binaural processing. Device-level comparisons should add candidate enhancement strategies to the same appropriately fitted amplification baseline, so that observed differences reflect enhancement rather than audibility or compression. Recognition, effort, and quality should then be considered together with latency, memory, computation, and energy. The relevant endpoint is listener benefit under the constraints of the intended device; Table 1 summarizes the information needed to assess it.

Table 1. Key information for evaluating the perceptual benefit and device feasibility of speech-enhancement systems.

Domain	Information to report
Signal processing and latency	Sample rate; microphone channels; frame and hop; look-ahead; causal status; analysis-synthesis and end-to-end delay
Model and computation	Architecture; parameter count; numerical precision; operations per frame or second; front-end and reconstruction costs
Memory, hardware, and energy	Model storage; peak working memory; target processor; runtime; measured power or energy; included and excluded stages
Perceptual evaluation	Population and audiograms; speech, masker, SNR, and reverberation; recognition; effort; quality; artifacts; spatial awareness
System integration	Beamforming; WDRC; fitting formula; feedback control; binaural processing; cloud, smartphone, or on-device execution
Generalization and ecological validity	Unseen noise, talker, corpus, room, language, device, and recording channel; field or ecological-momentary evidence

6. Future Directions: Integrating Conventional Signal Processing and Edge AI

Future progress is likely to depend less on choosing between conventional signal processing and learned processing than on assigning each component the role it performs best. Prescriptive amplification will remain the foundation of hearing-loss compensation, while speech enhancement may combine efficient conventional signal processing with compact learned models for scene analysis, target estimation, and adaptive control. Within this hybrid framework, four developments appear especially important.

Near-term systems are likely to remain modular. Prescriptive frequency shaping and WDRC will continue to provide hearing-loss compensation, while beamforming, feedback control, and other conventional signal-processing components solve well-defined problems efficiently. Learned systems can complement these components without displacing them with a single end-to-end model. Hardware-software-audiology co-design will matter as much as model architecture: an algorithm that improves an offline score but consumes resources needed by other functions or shortens usable battery life is not necessarily the better hearing-aid solution.

Learned processing may be most useful upstream of acoustic modification. A compact causal model could identify the current target, classify the scene, estimate relevant cue trajectories, and judge its own confidence. A bounded processor could then apply gain, filtering, or cue enhancement. The learned output would remain low dimensional and the acoustic change interpretable. When confidence is low, the device could reduce processing strength or return to a conservative mode rather than fail abruptly.

Evaluation will also need to move beyond a single leaderboard score. Objective measures remain useful for screening, but hearing-device studies need human outcomes and device measurements from the same implementation. A smaller intelligibility gain may be preferable if it preserves quality and spatial awareness, avoids objectionable delay, and can run throughout the day. Comparisons should include the complete processing chain rather than attributing the result to one module.

Personalization is likely to become more contextual and more closely tied to user intent. Audiogram-conditioned processing is only a starting point. Users differ in listening goals, tolerance for suppression, language experience, and the

value they place on environmental sound. Continuous on-device learning may be costly and unstable; a more practical near-term approach is a compact pretrained model with a small number of listener-specific controls and occasional preference updates. The wanted talker may change as the user turns the head, and in traffic the safest output may not be the one with the highest SNR.

Targeted formant enhancement provides an interpretable test case because both the manipulated cues and their perceptual consequences can be specified. The next step is mixture-based F2/F3 estimation with bounded enhancement and individualized WDRC, evaluated in terms of tracking accuracy, listener outcomes, computation, memory, latency, and energy. Its value should be assessed on the same hardware against conventional and full-band neural systems, progressing from unfamiliar acoustic conditions to listeners with clinically representative hearing loss and field use. Such comparisons would clarify the balance between perceptual benefit and practical feasibility and could also be extended to other auditory-informed cue modifications.

7. Conclusions

Prescriptive amplification will remain the foundation of hearing-aid processing, but it does not by itself resolve target-to-masker competition. Speech enhancement complements rather than replaces prescriptive amplification, and future systems are likely to integrate conventional signal processing with Edge AI. Evidence from conventional DNR distinguishes improvements in comfort from gains in intelligibility, while spectral and formant enhancement approaches show that selective cue modification can help under specific forms of degradation. Deep learning offers broader estimation and separation capabilities, although its practical value still depends on generalization, sound quality, latency, memory, computation, energy, and integration with the complete hearing-aid system. A hybrid architecture is a credible near-term direction: compact AI can identify the target, characterize the scene, and guide processing, while bounded auditory-informed methods implement the acoustic modification. Progress should therefore be judged by whether an algorithm provides reliable listener benefit while meeting clearly specified constraints on computation, memory, latency, and energy use.

Acknowledgments

This study was supported by the 2022-2023 Research, Scholarship, and Creative Activity Award, 2024-2025 Pathways to Tenure Grant, Faculty Research Project Development Initiative (FRPDI) from California State University, Northridge to Mingshuang Li.

Author Declarations

Conflict of Interest

The author declares no conflicts of interest related to this manuscript.

Ethics Approval

Not applicable.

References

Andersen, A. H., Santurette, S., Pedersen, M. S., Alickovic, E., Fiedler, L., Jensen, J., & Behrens, T. (2021). Creating clarity in noisy environments by using deep learning in hearing aids. Seminars in Hearing, 42(3), 260–281.

Baer, T., Moore, B. C. J., & Gatehouse, S. (1993). Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: Effects on intelligibility, quality, and response times. Journal of Rehabilitation Research and Development, 30(1), 49–72.

Bentler, R. A., & Chiou, L.-K. (2006). Digital noise reduction: An overview. Trends in Amplification, 10(2), 67–82.

Bentler, R. A., Wu, Y.-H., Kettel, J., & Hurtig, R. (2008). Digital noise reduction: Outcomes from laboratory and field studies. International Journal of Audiology, 47(8), 447–460.

Boers, P. M. (1980). Formant enhancement of speech for listeners with sensorineural hearing loss. IPO Annual Progress Report, 15, 21–28.

Bruce, I. C. (2004). Physiological assessment of contrast-enhancing frequency shaping and multiband compression in hearing aids. Physiological Measurement, 25(4), 945–956.

Bunnell, H. T. (1990). On enhancement of spectral contrast in speech for hearing-impaired listeners. The Journal of the Acoustical Society of America, 88(6), 2546–2556.

Chen, J., Baer, T., & Moore, B. C. J. (2012). Effect of enhancement of spectral changes on speech intelligibility and clarity preferences for the hearing impaired. The Journal of the Acoustical Society of America, 131(4), 2987–2998.

Chen, J., Baer, T., & Moore, B. C. J. (2013). Effect of spectral change enhancement for the hearing impaired using parameter values selected with a genetic algorithm. The Journal of the Acoustical Society of America, 133(5), 2910–2920.

Christensen, J. H., Whiston, H., Lough, M., Gil-Carvajal, J. C., Rumley, J., & Saunders, G. H. (2024). Evaluating real-world benefits of hearing aids with deep neural network-based noise reduction: An ecological momentary assessment study. American Journal of Audiology, 33(1), 242–253.

Diehl, P. U., Singer, Y., Zilly, H., Schönfeld, U., Meyer-Rachner, P., Berry, M., Sprekeler, H., Sprengel, E., Pudszuhn, A., & Hofmann, V. M. (2023). Restoring speech intelligibility for hearing aid users with deep learning. Scientific Reports, 13, Article 2719.

Fedorov, I., Stamenovic, M., Jensen, C., Yang, L.-C., Mandell, A., Gan, Y., Mattina, M., & Whatmough, P. N. (2020). TinyLSTMs: Efficient neural speech enhancement for hearing aids. In Proceedings of Interspeech 2020 (pp. 4054–4058). International Speech Communication Association.

Fitzgerald, M. B., Mysore Athreya, V., Srour, M., Rejimon, J. P., Venkitakrishnan, S., Bhowmik, A. K., Jackler, R. K., Steenerson, K. K., & Fabry, D. A. (2025). Effectiveness of deep neural networks in hearing aids for improving signal-to-noise ratio, speech recognition, and listener preference in background noise. Frontiers in Audiology and Otology, 3, Article 1677482.

Folkeard, P., Ashkanichenarlogh, V., Rahme, M., Parsa, V., Kühnel, V., Sheikh, B., Qian, J., Sung, Y.-Y., & Scollie, S. (2026). The effect of deep neural network implementation on speech recognition, listening effort, and sound quality in older adults with mild to moderately severe hearing loss. Trends in Hearing, 30, Article 23312165261449983.

Franck, B. A. M., van Kreveld-Bos, C. S. G., Dreschler, W. A., & Verschuure, H. (1999). Evaluation of spectral enhancement in hearing aids, combined with phonemic compression. The Journal of the Acoustical Society of America, 106(3), 1452–1464.

Goehring, T., Chapman, J. L., Bleeck, S., & Monaghan, J. J. M. (2018). Tolerable delay for speech production and perception: Effects of hearing ability and experience with hearing aids. International Journal of Audiology, 57(1), 61–68.

Guan, J., & Liu, C. (2019). Speech perception in noise with formant enhancement for older listeners. Journal of Speech, Language, and Hearing Research, 62(9), 3290–3301.

Healy, E. W., Taherian, H., Johnson, E. M., & Wang, D. L. (2021). A causal and talker-independent speaker-separation/dereverberation deep learning algorithm: Cost associated with conversion to real-time capable operation. The Journal of the Acoustical Society of America, 150(5), 3976–3986.

Healy, E. W., Yoho, S. E., Wang, Y., & Wang, D. L. (2013). An algorithm to improve speech recognition in noise for hearing-impaired listeners. The Journal of the Acoustical Society of America, 134(4), 3029–3038.

Healy, E. W., Johnson, E. M., Pandey, A., & Wang, D. L. (2023). Progress made in the efficacy and viability of deep-learning-based noise reduction. The Journal of the Acoustical Society of America, 153(5), 2751–2768.

Itani, M., Chen, T., Raghavan, A., Kohlberg, G., & Gollakota, S. (2025). Wireless hearables with programmable speech AI accelerators. In Proceedings of the 31st Annual International Conference on Mobile Computing and Networking. Association for Computing Machinery.

Kayser, H., Herzke, T., Maanen, P., Zimmermann, M., Grimm, G., & Hohmann, V. (2022). Open community platform for hearing aid algorithm research: Open Master Hearing Aid (openMHA). SoftwareX, 17, Article 100953.

Keidser, G., Dillon, H., Flax, M., Ching, T., & Brewer, S. (2011). The NAL-NL2 prescription procedure. Audiology Research, 1(1), Article e24.

Lakshmi, M. S. K., Rout, A., & O'Donoghue, C. R. (2021). A systematic review and meta-analysis of digital noise reduction hearing aids in adults. Disability and Rehabilitation: Assistive Technology, 16(2), 120–129.

Li, M. (2021). Perceptual effects of formant enhancement with the factors of phonetic type, listening conditions and language experience of listeners [Doctoral dissertation, The University of Texas at Austin].

Li, M., & Liu, C. (2026). Sentence recognition and enhancement in noise: Effects of speech muffling, signal-to-noise ratio, and formant enhancement. JASA Express Letters, 6(7), Article 075204.

Miller, R. L., Calhoun, B. M., & Young, E. D. (1999). Contrast enhancement improves the representation of /ɛ/-like vowels in the hearing-impaired auditory nerve. The Journal of the Acoustical Society of America, 106(5), 2693–2708.

Nittrouer, S., Tarr, E., Wucinich, T., Moberly, A. C., & Lowenstein, J. H. (2015). Measuring the effects of spectral smearing and enhancement on speech recognition in noise for adults and children. The Journal of the Acoustical Society of America, 137(4), 2004–2014.

Overby, N., Dau, T., Zahorik, P., Strelcyk, O., & May, T. (2025). Preference and similarity of combined noise reduction and dynamic range compression in hearing-impaired listeners. The Journal of the Acoustical Society of America, 158(2), 1334–1343.

Pandey, A., & Wang, D. L. (2020). On cross-corpus generalization of deep learning–based speech enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 2489–2499.

Park, S., Lee, S., Park, J., Choi, H. S., Lee, K., & Jeon, D. (2025). A real-time speech enhancement processor for hearing aids in 28-nm CMOS. IEEE Journal of Solid-State Circuits, 60(5), 1830–1843.

Sarampalis, A., Kalluri, S., Edwards, B., & Hafter, E. (2009). Objective measures of listening effort: Effects of background noise and noise reduction. Journal of Speech, Language, and Hearing Research, 52(5), 1230–1240.

Simpson, A. M., Moore, B. C. J., & Glasberg, B. R. (1990). Spectral enhancement to improve the intelligibility of speech in noise for hearing-impaired listeners. Acta Oto-Laryngologica, 110(sup469), 101–107.

Stone, M. A., & Moore, B. C. J. (1992). Spectral feature enhancement for people with sensorineural hearing impairment: Effects on speech intelligibility and quality. Journal of Rehabilitation Research and Development, 29(2), 39–56.

Summerfield, Q., Foster, J., Tyler, R. S., & Bailey, P. J. (1985). Influences of formant narrowing and auditory frequency selectivity on identification of place of articulation in stop consonants. Speech Communication, 4(3), 213–229.

Wang, D. L., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(10), 1702–1726.

Woodall, A. J., & Liu, C. (2013). Effects of signal level and spectral contrast on vowel formant discrimination for normal-hearing and hearing-impaired listeners. American Journal of Audiology, 22(1), 94–104.

Correspondence Author

Mingshuang Li, Ph.D.
E-mail: mingshuang.li@csun.edu

试论中国式现代化

A Discussion on Chinese-Style Modernization

陈岳云 / Yueyun Chen

美国洛杉矶西来大学工商管理系

Department of Business Administration,
University of the West, Los Angeles, USA

摘要

本文探讨了中国式现代化的历史逻辑， 现代化社会的基本要素。现代化社会就是以知识经济为主导，依靠计算机、互联网、机器人、人工智能等现代技术, 以数据为关键生产要素, 进行柔性化、定制化为主要特征的生产,追求可持续发展,精神生活丰富和个人有较充分自由的新型发达社会。文章例举了近五十年改革开放, 为中国式现代化奠定坚实基础, 在经济等领域所取得的成就。讨论了中国式现代化的主要经济指标, 在本世纪中叶, 名义人均国民收入达到当前发达经济体人均国民收入的平均值水平。分析了中国式现代化的主要特征， 並进而强调中国应怎么做, 才能在本世纪中叶, 实现中国式的全面现代化。

I. 引言

中国经济和社会发展已进入了一个新的阶段。其新的目标是实现中国式的现代化。这是一个历史性的宏伟蓝图。一个国家，其经济和社会的变迁由其主观愿望和努力，内部条件，及其外部环境等所决定的。而且这种发展有其客观规律性，特别是先前较发达国家所走过的道理，其经验和教训值得借鉴。

中华民族有几千年悠久的历史。中华文明源远流长,博大精深。中国人民，已从站起來，富起來，到现在的强起來。中华民族的伟大复兴，是无数代中华儿女的共同梦想，一致追求，精神动力。实现中国式现代化，建设一个高度民主高度自由高度文明的强国，是广大中国人民，中华儿女的共同心愿。

建设现代化的强国是近代以來中国人的梦想和矢志追求。孙中山1894年提出的“振兴中华”口号,是近代中国探求现代化的早期精神标志。孙中山1918年至1920年代初的《建国方略》 中的《实业计划》, 系统规划了工业以及交通等建设， 是中国现代化的第一份宏伟蓝图，现代化思想的实质性体现。新中国成立不久,提出了实现“现代化的工业、现代化的农业、现代化的交通运输业和现代化的国防”的设想。1964年第三届全国人民代表大会第一次会议， 向全国人民宣布实现“四个现代化”的任务：建设成为一个具有现代农业、现代工业、现代国防和现代科学技术的社会主义强国, 赶上和超过世界先进水平。1982年提出了工农业总产值到二十世纪末翻二

番，人民生活水平达到小康水平的目标，並于1995年提前五年圆满实现。2017年提出了建立现代化强国的目标,以及2022年揭示了中国式现代化宏伟规划。综上所述，建设现代化的强国，是近代以來中国人民的巨大梦想，矢志不移追求的宏伟目标。无数先辈，高瞻远瞩,承前启后,继往开来,身先士卒，鞠躬尽瘁；广大人民吃若耐劳,不折不挠,前赴后继,团结一致，顽强拚搏。正是这些无数人的牺牲和奋斗，造就了中国的独立和解放，造就了改革开放近五十年的巨大变化，造就了中国现代化事业的显著进步。

什么是中国式现代化，能否实现及如何才能实现，这些都是值得並应该探讨的大问题。本文首先展示人类文明进步的轨迹，三个不同的阶段，並分析现代化社会的基本要素。然后举证中国近五十年改革开放，经济社会发展所取得的成就，为实现全面代化所奠定的坚实基础。重点讨论了中国式现代化的主要经济指标及主要特征，並进而分析中国应怎么做，才能在本世纪中叶，实现中国式的全面现代化。

II. 文明的演变与社会现代化

现代化社会、工业社会、农业社会的差异

要素	现代化社会	工业社会	农业社会
经济与物质水平	发达经济体;物质丰富多样;总体上个人高收入、富裕	中等收入经济体；物质较丰富多样；总体上个人收入较高，较富裕	落后、欠发达经济体；物质匮乏，单一;总体上个人收入低，解决自給自足、温饱问题
主导产业与 经济结构	第三产业（服务业）,特别是信息服务业占主导，知识经济兴起，无形资产（如专利、数据）价值超越有形资产	第二产业（工业）崛起并主导经济，资本与机器成为核心资源，市场交换高度发达	第一产业（农业）占绝对主导，土地是最主要的生产资料，经济循环相对封闭
精神生活水平与个人自由	精神生活丰富;个人有较充分时间和自由,个性化发展	多样化精神生活;个人有足夠的自由	精神生活匮乏，单调；个人缺乏足夠的自由
技术核心与 生产方式	生产呈现柔性化、定制化特征，数据成为关键生产要素； 以智能化与数字化为主，依靠计算机、互联网、机器人、人工智能等现代技术	技术强调标准化、生产效率、规模化与流水线作业;依靠蒸汽、电力等无生命能源， 以机器大生产	核心技术为农耕历法与简单农具，生产具有季节性且自给自足；以手工劳动为主，依靠人力、畜力和自然力
人际关系与社会结构	组织边界模糊，网络化社会，个体联系更紧密但关系碎片化，知识阶层成为核心力量，治理趋向扁平化与去中心化	人口向城市集中，城市陌生人社会，业缘关系取代血缘，形成科层制（官僚制）管理，法治取代人治	以家庭为基本单位，乡村熟人社会，血缘与地缘关系紧密，社会流动性低，管理多靠“人治”与礼俗
人地关系与环境观	追求可持续发展，注重生态修复与集约利用，通过技术调和经济发展与环境保护	征服自然，高强度开发资源导致污染与生态破坏，人地矛盾尖锐	顺应自然，对环境影响小但抗风险能力弱，追求生态平衡
竞争逻辑与 时间观念	竞争转向人与人之间的信息与智力交互;时间倾向未来，注重预测与创新	竞争在于人与人造自然的博弈;时间倾向现在，强调效率与即时目标	竞争主要在人与自然之间;时间倾向过去，依赖经验传承

农业、工业和现代化社会,代表了人类文明的三个主要进化阶段。它们共同依赖人类集体的合作来克服资源稀缺，有相似之处,包括:

1. 适应稀缺性： 这三种社会类型本质上都围绕生产食物和商品来克服环境限制而组织，尽管他们采用越来越先进的方法来实现这一点。
2. 复杂的劳动分工：每一次远离基本生存的转变,都导致了更深的社会分层、专业化分工和相互依存的经济。
3. 集体意识：尽管结构性发生变化，每个系统/社会仍依赖共享的文化价值观和社会团结来维持秩序和社区功能。

但农业社会、工业社会与现代化社会存在巨大差异。它们的核心差异,在于经济与物质水平、主导产业与经济结构、精神生活水平与个人自由、技术核心与生产方式、人际关系与社会结构、人地关系与环境观、及竞争逻辑与时间观念等的区别。

基于上述分析， 现代化社会就是以知识经济为主导，依靠计算机、互联网、机器人、人工智能等现代技术， 以数据为关键生产要素， 进行柔性化、定制化为主要特征的生产,追求可持续发展， 精神生活丰富和个人有较充分自由的新型发达社会。 而中国式现代化，就是根据中国的国情、实情，经济规模，经济发展程度，技术开发水平，市场需求差异，组织管理能力，社会文化，环境等因素，來确定上述现代化要素的界限，要达到的程度和水平。

III. 近50年的快速发展为中国经济社会现代化奠定了坚实基础

中国自1978年实行改革开放政策以来，经济和社会各方面取得了史无前例的突破性进步，人均GDP从1978年的156美元，提高到2024年的13,303美元。产业結构不断变化，从一个农业就业为主导，农业产值占比高的偏农业型经济体，逐步向工业型，然后偏服务型的经济体。下表是三大产业就业和产值占比的变化。与工业化进程相一致，城市化也稳步提高。城镇化率从1978年17.9%，上升到2024年的67%。

1978年与2024年中国经济的比较

		1978	2024
人均GDP		$156	$13,303
城镇化率		17.90%	67%
农业	就业	70.50%	22.2%
	GDP	27.70%	6.8%
工业	就业	17.30%	29%
	GDP	47.70%	36.5%
服务业	就业	12.20%	48.8%
	GDP	24.60%	56.7%

数据来源：世界银行

1978年与2024年中国贸易的比较

		1978	2024
出口	**总价值（十亿美元）**	6.89	3,543
	% GDP	4.6%	18.9%
进口	**总价值（十亿美元）**	7.64	3,219
	% GDP	5.1%	17.2%
总贸易	**总价值（十亿美元）**	14.53	6,762
	% GDP	9.7%	36.1%

数据来源：中国国家统计局、世界银行

得益于积极对开放政策和全球化的良好国际环境，中国在过去近五十年对外贸易长足进步。进出口年度总贸易，从1978年145 亿美元，增长到2024年的6 万7千多亿美元； 进出口贸易占GDP比率从1978年的 9.7% 到2024年的 36.1%。

1978年与2024年中国其他的比较

		1978	2024
在校大学生(万)		85.6	4,846
科研投入(R&D)(亿美元)		11.3	5,090
科研 (R&D) 从业人员(折合全时) (万人)		50	795
专利申请量(万件)		0	180
人均居住面积(平方米)	城镇	3.6	41.8
	农村	8.1	47
人均可支配收入(元)	城镇	343	54,188
	农村	134	23,119
交通运输(万公里)	铁路	5.2	16.2
	公路通车里程	89	549.04

数据来源：中国国家统计局

教育、科技事业也取得了巨大的进步。在校大学生从1978年的85.6万人，猛增到2024年的4,846万人。科研 (R&D) 投入 从1978年的11.3亿美元增加到2024年的5,090亿美元。科研 (R&D) 从业人员折合全时当量从1978年的不足50万人，增加到2024年的 795万人。1978 年中国仍无专利法，因此无专利申请量， 1985年 1.4万件， 而2024 年中国专利申请量约 180 万件，占全球近半数，连续 6年居全球首位 。美国位居第二：同期美国专利申请量约为 50.2 万件，中国数量是美国的 3 倍多。

民众的居住条件也历经显著变化。1978年，城镇人均居住面积3.6平方米，农村人均8.1平方米；2024年，城镇人均41.8平方米，农村人均47平方米。人均收入不断提高，1978 年城镇居民人均可支配收入为343 元,农村居民人均可支配收入为134 元，2024 年城镇居民人均可支配收入为54,188 元,农村居民人均可支配收入为23,119 元。1978年全国约有2.5亿农民处于绝对贫困状态，2021 年初，中国正式宣告消除了绝对贫穷。 1978年全国铁路营业里程5.2万公里，公路通车里程89.0万公里。2024年，铁路营业里程16.2万公里，其中高速铁路营业里程4.8万公里；公路总里程 549.04万公里，其中高速公路里程19.07万公里。

1978年与2024年中国其他的比较

		1978	2024
在校大学生(万)		85.6	4,846
科研投入(R&D)(亿美元)		11.3	5,090
科研 (R&D) 从业人员(折合全时)(万人)		50	795
专利申请量(万件)		0	180
人均居住面积(平方米)	城镇	3.6	41.8
	农村	8.1	47
人均可支配收入(元)	城镇	343	54,188
	农村	134	23,119
交通运输(万公里)	铁路	5.2	16.2
	公路通车里程	89	549.04

数据来源：中国国家统计局

尽管中国过去近五十年取得了如上所述的突破性、历史性、全方位的快速发展和进步，但与其他发达经济体相比，仍有巨大的差距。下表是与美日的比较。从人均GDP及城镇化率方面，中国仍大大落后于这些发达国家。从三大产业就业和GDP分布上比较，我们的农业比重仍过高，也说明农业规模经济弱，效率低。服务业在较快增长，已成主要产业，但与美日相较，比例仍较低。这些比较，一方面說明了我们仍落后；而另一方面，预示出中国发展的巨大潛力，更说明推动现代化建设的必要性，及努力的方向。

尽管中国过去近五十年取得了如上所述的突破性、历史性、全方位的快速发展和进步，但与其他发达经济体相比，仍有巨大的差距。下表是与美日的比较。从人均GDP及城镇化率方面，中国仍大大落后于这些发达国家。从三大产业就业和GDP分布上比较，我们的农业比重仍过高，也说明农业规模经济弱，效率低。服务业在较快增长，已成主要产业，但与美日相较，比例仍较低。这些比较，一方面說明了我们仍落后；而另一方面，预示出中国发展的巨大潛力，更说明推动现代化建设的必要性，及努力的方向。

中国、美国和日本经济的比较 (2024年)

		中国	美国	日本
人均GDP		$13,303	$85,810	$32,476
城镇化率		67%	83%	92%
农业	就业	22.2%	1.57%	3%
	GDP	6.8%	0.9%	1%
工业	就业	29%	19.34%	24%
	GDP	36.5%	17.3%	24.4%
服务业	就业	48.8%	79.09%	73%
	GDP	56.7%	79.7%	74.7%

数据来源：世界银行

这里需要指出的是，中国式现代化，是在过去近百年，建国后七十多年，特别是改革开放近五十年，所取得的经济和社会发展的基础上展开的。历史的进程有其连续性，经济和社会的进步更是如此。特别是 “四个现代化” 的宏伟目标，经过六十多年的努力奋斗，大部分已基本或接近达到。相较于其他三个现代化，农业

的现代化尽管显著改善，农业生产率得到大的提高，务农人数，特别是占总就业的比率大幅降低，但其总体现代化的程度仍很低，农业的机械化、自动化有待加強，科技兴农仍需突破。工业的现代化成效显著，成果丰硕。中国能成为世界工厂，工业生产总值超过美日德总和，产品出口稳居世界第一，均得益于制造业的规模化、现代化发展。但工业发展、现代化提高不平衡，许多关健产业，关健设备、关健产品，仍无法自制自給。许多产品总量大，总出口多，但科技含量低，利润少。国防的现代化，应该是最有成效的。现代化的完整的满足国家防御和保障经济发展的国防体系已基本完成並在不断改进、加強。“上天揽月”，“下洋捉鳖”，中国已成为航天、航海、深海大国、強国。科技现代化也与日俱进。正是科技现代化的成果及广泛应用，促进和保障了 农业、工业和国防现代化的持续推进。当然科技发展永无止境，现代化进步亦如此。中国四个现代化虽成效辉煌，但仍需不断努力，持续前进。中国式的现代化，必然包含这四个领域的进一步现代化。限于篇幅，本文就不深入探讨，农业、工业和国防这三个现代化了。

IV.中国式现代化发展的主要经济指标

经过近五十年的快速发展，中国已从一个落后贫穷的国家，然后成为小康社会，到现在巳实现高中等收入经济体，並向高收入经济体稳步迈进。工业化，城镇化及现代化，是世界各国、各经济体发展中经历的必然之路。中国的工业化之路，借鉴各国的经验，充分利用世界经济开放、产业转移、全球化的浪潮，取得了持续且快速的开发。城镇化进步也载入史冊，近五十年中，年均城镇化率提高1.04%。尽管中国的工业化道路並未完成，特别是高科技、高质量及高精密的产品及行业仍需大力发展，城镇化也仍有很大的发展潜力，特别是许多没有戶籍的农民工的真正的城镇化仍有许多问题需要解决，但今后中国发展的主旋律，主要目标，是加速推进並全面实现经济和社会各个方面的现代化。

经济是基础，要成为一个全面现代化的国家，必须有扎实、雄厚的经济为条件，首先必须成为有实力的发达经济体。按照世界银行(WB)当前的标准，人均国民总收入 (GNI)超过13,935美元是划分成为发达经济体的必要标准。中国2025年GNI 应该巳非常接近这个水平。

世界银行条件下的发达经济体共有87个，名义人均GNI平均值是50,302美元。中国若能在本世中叶达到这个数，將是在经济上取得的巨大成就。另外，经济合作和发展组织（OECD）是发达经济国家的棋乐部，目前共有38 成员国。 OECD只邀请符合其条件的国家申请，並必须获得所有成员国的同意，方可正式入会。其成员国除了是一个发达国家外，还必须符合包括完全实行市场经济等条件。目前所有会员国的名义人均GNI 均值为47,119美元。中国若干年后，即使成为一个发达国家，想成为一个OECD会员国，仍会非常困难，因为实行完全的市场经济等要求。但如前面世界银行平均值所讨论的，OECD 成员国经济标准仍可作为衡量中国现代化程度的一种参照，即到本世纪中叶，中国要达到名义人均GNI 47,119 美元，OECD 成员国当现均值的水平。

如下表所例，到2050年，中国必须保持4.69%的年度名义人均GNI增长率，才能达到OECD 38国2024年的人均GNI平均值水平。 如果想达到2024年世界银行高收入经济体的平均水平，年增长必须达到4.95%。然而，如果其他经济体年增长率为2%，那么相关中国的年增长率应为6.79%和7.05%，才能得配届时相关发达经济群体人均GNI的平均值。

中国的上述名义人均GNI年度增长率能否实现？约5%的名义人均GNI年增长率应可实现。 但要达到6.79%到7.05%的名义人均GNI 年增长率,会有挑战性，但仍然可行。正如 Chen, et al (2017) 分析的那样，除了中等速度年均经济增长率的潜力外，通胀和人民币升值将帮助中国提高名义年度国民总收入和人均国民总收入。此外，从2024年起，中国人口减少，这将帮助中国未来人均GNI更高。

	中国 人均GNI	OECD 人均 GNI 平均值 (38国家)	WB 高收入经济体 人均 GNI 平均值 (87经济体)
2024	$13,660	$47,119	$50,302
By 2050			
年增长率—中国4.69% & OECD 0%	$47,119	$47,119	
年增长率—中国6.79% & OECD 2%	$80,427	$80,427	
年增长率—中国4.95% & WB 0%	$50,302		$50,302
年增长率—中国7.05% & WB 2%	$85,859		$85,859

V. 中国式现代化的主要特征

如本文前面所讨论的，不同社会和国家的现代化，必然带有若干共性。同时，二十一世纪的中叶，中国式的现代化，必然带有中国特色与时代特征。

首先，一个现代化的社会，必须是高度民主，高度自由的社会。不同国家不同民族在社会发展的不同时期，对民主和自由的要求不同。但民主和自由有其一定的共性。特别是作为一个发达的国家，一个现代化的国家，这种共性应该是无可争辩的。当然民主和自由是不断培育和进步的。这里最重要的，是看是否真心实意要借鉴其他国家的经验，逐步扩大民主和自由。更要建立一套机制，确定一个明确的目标，有具体的衡量标准，不断改进民主和自由。要像报告每年的GDP一样，每年公告全体人民，民主和自由的綜合程度是否提高、改善了。建设民主自由富強的国家，是当初中国革命事业的根本宗旨，是对人民的庄严承诺。这也是无数先辈终身奋斗、牺牲的动力和目标。

改革开放的思想，就是民主自由的思想。改革开放是民主自由的一种表现和结果。国家以人民利益为出发点和宗旨，全心全意为人民服务，人民至上， 就应相信和依靠广大人民，在经济和社会各方面已成熟的条件下，充分实行高度的民主和自由。

第二,要成为一个现代化的社会并不断进步，必须是市场经济高度发达、成熟，物质生活高度充裕的社会。市场经济是个好东西，尽管不够完善。建立、健全统一有效有序的市场体系，使广大企业更公平充分自由竞争，优胜劣汰;使各种产品、服务的流通,市场交换更有效有序;使各种资源得到优化合理的分配，增加社会效益，更好满足大众的多样化需求。要建立统一有效有序的市场体系， 必须建立、健全、完善並切实执行各种满足经济、社会稳定和发展所需的法律、法规，依法行政，依法规范市场运行、企业经营，有效保护个人、企业、社会和国家财产，切实有效保护个人、企业等权益。有效保护知识产权。有效保护民（私）营企业及其企业家。有效保护外资及各类其他投资人的权利和利益。

第三, 中国式现代化社会, 必须是高度文明，精神生活高度丰富、多样的社会。必须是和谐、友爱、平安、团结、互助的社会。城乡一体化发展, 各阶层、各民族求同存异, 平等包容,互相尊重,友好相处。中国式现代化的建设，就是要更好传承发扬优秀中华文化，更好树立培育中华民族精神，更好传播弘扬中华文明。

第四, 社会治理现代化是中国式现代化的一个重要方面。社会治理现代化的核心是共建共治共享, 就是广大民众一起,共同建设美好和谐的社区、家园，共同参与社区发展的决策和治理，並共享安全、美丽、互助、团结，富有

活力的社区环境和成果。广义上讲， 社会治理的现代化，也包括国家管理的现代化。各级政府及其不同职位人公责任感,不断提高大众的文明程度，道德水准和素质修养。

第五， 二十一世紀的现代化,必须是人与自然和谐共生的现代化。必须保持绿色发展，像保护眼睛一样保护生的官员和工作人员，其基本责职是履行其职位所赋予的责任，全心全意为大众服务。维护、执行宪法及各类法律、法规，维护公平正义，是政府的本质任务。社会治理的现代化，也包括提高全民的民族意识，爱国热情，主态环境，实现永续发展。人民大众对美好生活的追求，也包括优美干净的环境，清洁的水，蓝色的天空，青秀的山河。环境严重污染，主要是由工业化粗放型发展，及城镇化提高、生活水平提高、汽车使用猛增等引起的。也由不重視森林、土地保护，破坏生态平衡导致的。而大众的文明化程度及环保意识同样重要。特别是各级政府，在招商引资、扩展产业时，是否有长远观念、环保意识，是关健。科技的发展和广泛应用，对減少、解决环境污染问题更是起决定性的作用。

最后， 中国式现代化必须是走和平发展道路的现代化。“各美其美、美美与共”， 是中国儒家共生发展的智慧。只有尊重文明多元性,互鉴互学，才能超越零和博弈思维，实现全球现代化的健康发展。中国式的现代化,是全人类努力构建命运共同体,使世界各国各民族情相联,利相统,共命运,和衷共济,共享成果,共谋发展。

VI.中国应该怎么做，才能在本世纪中叶成为一个高度现代化的国家

要在本世纪中叶成为一个高度现代化的国家， 中国必须明确目标，科学规划，凝聚人心，精心组织， 不断创新，埋头苦干，并采取如下措施：

首先， 必须进一步充分解放思想，实事求是，坚持更广更深入的改革开放政策，坚持以经济建设为中心，永不动摇。 现代化的目标已经明确，方向也很清楚。 但前进道路仍很复杂、困难，仍会碰到前所未有的挑战和危机。 近五十年经济和社会高速全面发展的最基本经验，就是要不断解放思想，

创造性的解决各种新老困难和问题。 改革不是权宜之计，改革应永远在路上，在实践中。 开放不仅是本身继续稳定进步的要求，也是世界发展的主流，时代的主旋律。 即使中国成为了一个发达经济体，经济总量接近，甚至超过了美国，我們的人均GDP仍会不高，与极大部分发达国家相比，仍有很大的差异。 所以，仍必须坚持以经济建设为中心，永不动摇。 不断提高人民大众的物质和精神生活水平，是政府的最高目标。 “发展是硬道理”， 只有经济持续增长，才能达成这样的目标。 实践是检验真理的唯一标准。 任何一种思想或理论，只有能指导总体经济和社会持续稳定、健康发展的，使广大人民的满足感、幸福度不断稳定提升的，才是真正的符合国情的真理。

第二， 要实现中国式现代化，就必须不断调整、优化经济结构，建立一个能满足人民日益增长各种需求的可持续发展的现代化经济体系 。如前面所分析的，中国的结济或产业结构，仍有很大的调整空间。特别是农业就业比率仍占总劳力的22.2%， 与美日等高度发达国家约1-3%相比，差距极大。提高农业生产率，改进农业规模生产和集约化经营，增加科技投入，更好应用现代科技于农牧渔业的生产、深加工、经营和销售，从而实现农业的现代化。随之而來，大量农务工，就能更多地转移到工业，特别是服务业。如Chen等 (2015, 2022, 2026) 反复强调的，稳定、巩固工业，特别是制造业，是中国经济行稳致远，平衡持续增长的关健。随着经济水平的提高，

服务业必然更快增长，在总体经济占比日益扩大。但服务业同样必须提高生产率，增加效率。服务业更要提高国际竞争力，从而扩大服务产品的更多出口及更好维护国内服务市场，避免与外资服务产品竞争中的不利地位。

第三，现代化经济体系的建立， 离不开有效与有为的政府，有效地制定与实施相关法律、法规，有效地解决企业与市场经济活动中出现的各种新问题。政府应是有为的，在若干新兴产业，高科技初创投入，公共设施、公共产品领域，需要各级政府的规划、投资及具体参与。政府也必须是有限的，权力的有限性及功能的限制性。搞市场经济，就必须限制政府的不必要权力，权力放在”笼子中” 、监督下， 并且加于法制化、规范化。政府的基本职能应是监管市场和经济及提供服务。

第四， 要实现中国式现代化，就必须不断提高生产率，特别是全要素生产率。生产率决定一个产品和企业的竞争力,也关系到经营的效率和效益。生产率也是一个经济体总体竞争力的主要标志，体现了一个国家总体科技发展，特别是科技应用的程度和水平。中国过去几十年经济快速发展，出口猛增，但整体生产率並未显著提升，也有研究表明有所下降; 特别是与美国这样的发达经济体相比，人均劳力所创产值比率，差距巨大。尤其是全要素生产率，其差异有扩大的趋势。所以推进中国式现代化，必须了解阻碍生产率提升的主要因素， 以大力提高生产力为重点。生产率不高的一个主要原因，应该是我们的经济结构，具有高附加价值的产品、企业和产业低。美国上市公司市值近一万亿或以上的目前有约十六家，最高市值近五万亿。中国上市公司中，市值近千亿的目前有约十一家，最高市值略超五千亿。所以中国最有价值的企业群体，其市值只为美国同类企业的10%。而中国2025 GDP约占美国的64%。为什么我们顶尖企业总价值相对低？最重要原因应是技术独创、垄断性低，随之而來的市占率不高，从而缺乏足夠的定价权。另一个因素，应是行业內卷普遍、严重。除了中央的反垄断，各地政府大力扶持各自的相关企业和行业，造成总体行业投资过度，竞争过度。特别是在行业发展达到一定程度、技术稳定、市场已形成相当规模后，若不能充分按照经济规律，市场竞争规律，优胜劣汰，內卷继续甚至更剧，就会造成社会资源的极大浪费，破坏行业正常运营和发展，亦会严重阻碍整个行业的技术进步、创新。

第五， 要实现中国式现代化，就必须不断完善、发展金融市场，使其能更有效充分地支持整个经济的营运和增长。金融产业及其发展对个人、企业、社会及整个经济都至关重要。中国金融业附加值占当年整个GDP 约8%。但金融对经济、社会有其他间接及综合的影响。 据估计，金融业创造一个就职岗位，会产生3.6个其他行业的附带岗位。在过去几十年里，中国金融业经历了快速发展。中国是第二大经济体，中国的保险业及股票市场也是世界第二。

下表是中美在2000，2020及2025年股票总市值， GDP及相应比率的比较。这些数字表明，中国股市总价值过去几十年有突破性增长，但股票总市值与当年 GDP 比率仍不高，不足80%。特别是与美国超过100%，甚至超过 200%的比率相对照。这也意味着，中國发展股票市场仍有极大的潜力。

中国与美国股市市值及GDP (万亿美元 Trillion) 比较

	中国			美国		
年份	**股市市值**	**GDP**	**%**	**股市市值**	**GDP**	**%**
2000	0.6	1.21	49.6%	11	10.25	107.3%
2020	11.52	15	76.8%	40	22.09	181.1%
2025	15.51	19.5	79.5%	70	31.42	222.8%

数据来源：https://www.macrotrends.net/ & https://www.statista.com/

对投资人及上市企业來说，最相关的是单个股价及股指的变化。下表是中国上证指数，深证成指及美国三大指数在这三个年份的情况,然后是相关年份总增长率及年均增長率的数据。从2000年到2020年，上证指数总增长67.5%, 年均增长3.4%; 从2020年到2025年，上证指数总增长 14.28%, 年均增长 2.9% 。深证成指从2000年到2020年，其表现比上证指数好，但从2020-2025年, 是负增长。 而美国三大指数，从2000年到2020年的年均增長率是上证指数至少2.7倍, 从2000年到2020年, 至少3.9倍。

按照世界银行的数据，中国经济年均增长率从2000年到2020年是 8.70%, 从2020年到2025年 4.90%; 而美国同期相应的经济年均增長率分别为 1.96% and 2.41%。这里就产生俩个尖锐问题，中国经济年均增长率这么高，为什么股市变化如此少？中国的年经济增长率比美国高許多，甚至是美国的2-4倍，但为什么我们的股市指数增长率与美国差这么多？

造成中国股市过去几十年整体表现平平，特别是不能与经济增长速度相匹配，而与美国相比较，更显不足，原因很多且复杂。在金融及整个经济改革中，在中国式现代化的建设中，若能加以解决、处理以下几个主要因素，相信中国股市会有根本性改变，更正常发展。这里需要強调，投资股市有风险，任何一个国家的股市，都有衰退的时候，关健之处，是看多久会恢复。美国股市在“9.11”及“COVID-19”时，也曾暴跌，出现“熔断”，但都能较快回复。而中国股市曾有近十年，回到原点的情况。

中国股市问题的第一个主要原因，应是发展历史短，整个市场不成熟。上世纪九十年初，才开始恢复股票交流，最初只有八个股票。尽管上世纪早期曾有公开的股票交易所，但几十年中断，恢复后的交易市场与之前的无关连性。由此造成证券交易人才短缺，监管缺乏经验，证券交易法规及管理不完善、不到位、不及时，公信度低。第二个原因是投资主体缺失、不足。股价及整个股市，由供需平衡所决定。进行股票买卖的机构、个人及总资金不足，自然无法带动股市兴旺。第三，优质上市企业少，缺少吸引力。第四，上市企业可持续发展能力不足，上市后，企业增长缺乏后劲，不稳定。这部份是由行业内卷，过度竞争造成的。第五，整体经济规范性不足，法律法规不完善，有其他众多风险低或较可控，但回报较高较稳定的投资机会，这样股票投资就不作考虑了。第六，散戶、个人投资占许多股票总值比例高，交易频繁，波动大，从而风险高。第七，股票投资基金少，且其投资回报不稳定、低，从而缺乏吸引力。第八，对股票投资，及对股票基金的投资限制多，包括各种退休基金及保险公司，如人寿保险公司，能否直接或间接地投资股市，最高比例多少等。

中国与美国股市指数的 比较

	中国		美国		
	上证指数	深证成指	道琼斯工业指数	标普500指数	纳斯达克指数
2000年	2073.47	4730.14	10786.85	1,320.28	2,470.52
2020年	3473	14470.68	30,606.48	3,756.07	12,890.03
2025年	3968.84	13525.02	48,063.29	6,845.50	23,241.99
2000-2020总增长	67.50%	205.92%	183.74%	184.49%	421.75%
2020-2025总增长	14.28%	-6.54%	57.04%	82.25%	80.31%
2000-2020年均增长	3.4%	10.3%	9.2%	9.2%	21.1%
2020-2025年均增长	2.9%	-1.3%	11.4%	16.5%	16.1%

数据来源: 中国与美国相关证券交易所

中国金融市场发展的另一个严重问题是融资难， 尤其是民营企业， 包括民营中小型及微型企业融资难。造成这一问题久未解决，主要是三个原因。一是中国企业直接融资（发行股票及债券）的比例太低，虽在上升，但仍约为20-25%， 而美国企业直接融资比例约为80%。 其次， 民营中小型及微型企业直接融资的机会非常有限。第三， 银行的贷款，大部分给了各级政府和国营企业，民营中小型及微型企业 非常难以获得银行贷款。民营企业占中国GDP 超 60%，占就业超 80% 及占出口超 60%，但在金融市场是孤儿，弱势群体。要根本改变这一久远的问题，需要从根本上解放思想,大胆改革、创新。除了提高民营企业直接融资比例， 银行的商业贷款要更多地流向中小企业特别是民营中小企业。 中央和省市等地方政府，应主要依靠发行各种债券等来直接融资，逐步减少直至基本消除银行贷款。即使是大的建设项目，也应依靠市场，靠发行债券等来筹集资金。国有企业特别是国有大中型企业，也应借助发行股票及债券直接融资。银行特别是国有银行，通过购买各种债券来支持政府及国有企业。这样的改革，使金融市场发挥了其应用的作用，使金融交易更公平，并能减少公共项目盲目过度重复投资的现象。更直接的结果是中小企业特别是民营企业 能够获得更多银行的商业贷款,从而从根本上解决他们的融资难问题。

第六 ,要实现中国式现代化，就必须积极培育、支持众多拥有高质量、卓越声誉的产品/服务与品牌的世界一流跨国公司。国际经济的竞争，本质上是企业之间的竞争;企业间最核心的竞争，是跨国公司间的竞争。根据<<Fortune>> 2025年的数据，在全球100家最大销售收入企业中，美国有42家，中国有30家。但美国最大的公司都是跨国公司，其销售额相当高的比例来自国外，且许多超过50%。在中国最大的公司中，绝大多数是国有企业，以国内销售为主。鼓励、支持民（私）营企业的国际化拓展，是达到拥有更多世界一流跨国公司的唯一途径。企业强则国强，民营跨国公司强则中国公司强。中国未来发展， 需要更多民营跨国公司， 尤其需要更强、更大、更多的民营跨国企业。

第七， 科学技术的现代化是四个现代化的基础,也是中国式现代化的关健。必须建立、健全、实施现代化的科技管理体系，增加科研投入，建立、完善产学研一体化协作、发展的体系，更多更好地应用科研成果。发挥群体合作优势，集中攻关，更有效地解决影响国家战略地位，阻碍关健产业发展的硬核技术。发展科技事业，就必须充分尊重知识，尊重人才，人尽其才，人尽其用。发展科技事业，就要按科技规律办事,专家治所，专家治校，鼓励、支持公平竞争，根据不同科技岗位的特点，借鉴国外行之有效的实践，制订实施有差别的考核评鉴标准，以更好地调动广大科研人员的主动性、积极性。必须大力鼓励创新，出产更多原创性突破性的科研成果，不断提高我们科研的世界影响力和地位。现代科技发展日新月异，颠覆性发明和技术不断涌现，不进则退。我们必须以加倍的勇气、加倍的努力，既要鼓励支持团队合作精神，发挥集体合作优势， 也要理解、支持个人独立科研的自主性，尊重他们的创造性、自由性。

第八， 教育的现代化是实现中国式现代化的必经之路。必须建立健全並有效实施能更好教育青少年、培育出国家和社会所需的各种人才的现代化的教育体系，包括现代化的中小学、职业教育体系及高等教育体系。教书育人是各类教育的本质工作。教育带有強烈的公共产品属性，所以不能用全盘市场化的思路和方法来进行改革，这是世界各国的共识，也是我们在尝试教育改革中学到的教训。教育改革，特别是高等教育的改革是一项复杂的系統工程， 要继续解放思想,深化改革,改变教风学风,保证学术自由，鼓励独立思考精神， 提高学生创造性思维的能力，减少对教授不必要的评鉴和考核。

像任何一个学科或事业一样，教育事业有其独特性。教育的改革,政策的制定了解需要掌握其规律和特点，客观科学地分析存在的各种主要问题，解剖其起因及原因，充分认识解决问题的条件和要求，探索解决问

题的可能的不同方案，及具体实施的步骤和方法。行有行规，业有专术，　应充分征求各级专家学者的独立意见，同时广泛听取大众的意见和建议，集思广益，发扬民主集中制，使各项政策更精准，更合实际，真正能解决迫切的问题，并能获得大众的支持。各地教育水平，经济发展水平不同，社会文化和习惯等也有异，在统一整个国家基本教育制度和政策的前提下，应允许及鼓励各地根据当地实情，采取有别的具体措施，试验解决教育热点问题的不同方案。

中国的国情不同，但各国的教育也有共性-满足民众需求,培养造就各种人才，探索了解未知的世界。所以中国也应更好更多地学习了解各国在教育改革和发展中的经验和教训，少走弯路。全球化是时代趋势，教育的国际化发展也是不可逆转的潮流。中国教育要继续走国际化全球化的道路。要继续对外开放，扩大国际交流与合作。不仅要继续鼓励支持更多的青年人，到国外留学深造，支持鼓励更多的外国人到中国留学攻读学位，也要吸引并留住更多的世界各国优秀人才，到中国来从事科研和教育工作。中国高校中，师资的国际化程度极低，有待加强提高。

教育是崇高的事业，“为天地立心，为生民立命，为往圣继绝学，为万世开太平”。崇高的事业需要崇高的思想和并有愿为之奉献的崇高的人，包括师生和各种职员等。而这一切，在市场经济中，在功利化盛行的社会中，往往时有矛盾和冲突。这也是当代教育面临困境的主要原因。人无完人，金无足赤。教育制度，教育事业也必然在异化，进化中逐步改变，与时代同行，伴历史齐进。

第九，决策的民主化、科学化是确保经济和社会持续、稳定、健康发展的基本条件，是实现中国式全面现代化的制度保证。现代社会错综复杂，国内各种矛盾到了深水区，国际关系纵错交杂，不确定性，各种危机，日益加深，所以重大决策，必须讲究民主化、科学化。集思广益，普纳贤见，广泛参与，科学论证，系統思维,以真实客观事实为依据,人民至上，以民为本，避免、減少错误的决策及其帶來的危害和不利影响，特别是要建立和践行有效的纠错机制，重大项目和决策的监督、评价体系，及时了解、掌握相关结果，为科学决策保驾护航。决策的权威性，來自于决策的科学性，决策的正确性；或者说，來自于决策的结果对社会和广大民众的影响。权威需要並必须得到维护、支持，但权威更需人民的检验，历史的检验。

最后，中国式现代化的实现，需要一个总体和平、稳定、友好的国际环境。世界需要中国，中国也需要世界。和平与发展，应是今后世界的主流，也是世界上绝大部分国家和人民的共同愿望。世界上各种战争时有发生。中国近代历史上，曾饱受不断战争的严重灾难，更懂得和平的珍贵和价值。
我们要与世界各国人民一起，反对战争，支持、声援和平。中国的快速发展前所未有，国际上难免有所不适。我们特别要避免历史上大国崛起的“修昔底德陷阱”。以史为鉴，以史为戒，遵循历史发展规律，顺应历史潮流，走和平发展的中国式现代化道路。二十世纪八十年代，中国提出並基本贯彻了在国际上“韬光养晦，不争出头”的基本战略，获得了广泛理解、认同与支持，对中国改革开放、引进外资、成功加入 WTO 等起到了促进的作用。随着总体国力的提升，中国在世界上的话语权、影响力显著提高,中国已经自然地从过去的“韬光养晦”，转变为“有所作为” 。但我们应充分认识到，中国人均 GDP 仍不高，科技、创新与世界先进国家仍有较大的差距，制造业总体仍大而不强。中国既不能 妄自菲薄，更不能盲目乐观。“有所不为” 才能“有所为”，这是我们自身的实力与条件所决定的。

VII.结论

中国式现代化的经济目标，是要在本世纪中叶，建国一百周年时，名义人均国民收入达到当前发达经济体

人均国民收入的平均值水平。更高的目标，是名义人均国民收入达到本世纪中叶时发达经济体人均国民收入的平均值。为了达到较低的目标，中国必须保持大致5%的名义人均GNI年增长率。为了达到更高的目标，相关年增长率应为7%。如前所述，低目标应能达到；考虑到人民币升值潜力，通货膨胀，及人口总数下降等因素，中国式现代化的高经济目前，也仍有望达到。

如上面所讨论，中国必须采取一系列行之有效的措施，才能确保其现代化真正实现，特别是保证其现代化走在正确的轨道上。需要強调的是，中国式的现代化本质上是人的现代化，人的思想的现代化，思維的现代化，观念的现代化。由此，人们的思想必须更加解放，思维必须更加科学，观念必须更加开放，不忘历史的责任，时代的重托，以更加扎实的行动，更加坚定的意志，共同为实现中国式现代化，建设一个高度民主高度自由高度文明高度富裕，更加平等更加美好的強国而奋斗，实现中华民族的伟大复兴。

“两岸猿声啼不住，轻舟已过万重山”。中国在过去近五十年的发展中,尽管也有挫折,走过弯路,但总体上坚持改革开放,坚持以经济建设为中心,充分调动发挥广大人民的积极性、创造牲，取得了经济持续快速增长，人民生活水平极大提高，整体国力突破性增強的举世公认的成就，为建设现代化的強国奠定了坚固的基础。世界在变，中国在变。国际政治、国际关系，战爭还是和平,全球化合作还是孤立脱钩,都正经历着前所未有的不确定性、存在着极大的风险。现代科技的发展，特别是许多颠覆性技术的发明和应用，带來了极好的机会与潜在的巨大危机。世界正处于其历史发展的特殊关健时期，中国的发展，中国梦的实现也处于关健时刻。“长风破浪会有时，直挂云帆济沧海”。相信世界的未來是美好的，中国的未來是美好的。历史总是在曲折中前进的。中华民族是伟大的民族，中国人民有理想有梦想，富有勇气和胆量，愚公移山,众志成城。中国式的现代化也必將攻坚克难，顺利完成。未來可期，因为我们善于总结经验教训，勇于承认、修正错误。未來更美好，因为我们善于学习，勤于实干。未來光辉灿烂，因为我们的事业是正义的，符合时代潮流、历史发展的趋势。

作者介绍

陈岳云,美国洛杉矶西来大学工商管理系;邮箱：billchen@uwest.edu.

参考文献

Chen, Yueyun (Bill), “China’s Path to the Sustainable, Stable and Rapid Economic Development: From the Largest to the Strongest Manufacturing Country”; Journal of World Economic Research, August, 2015.

Chen, Yueyun(Bill), et al. ”The Productivity, Economic Structure and Middle-Income Trap—Can China Avoid this Trap? Journal of Applied Business and Economics, Volume 19(11), 2017.

陈岳云,“中国下一步怎么走?”, <<人民论坛>>, 2017年11月.

陈岳云, “中国外交的新思维新特点,” <<人民论坛>>, 2018 年 3月.

陈岳云, “中国民营企业要敢担当争创世界一流的跨国企业,” <<人民论坛>>, 2018年12月.

陈岳云，“中国能否避免大国崛起的‘修昔底德陷阱’?” <<G20发展报告（2018-2019）>>，北京外国语大学G20集团研究中心，2019年9月.

陈岳云，“民族精神 ：‘多难兴邦’的强大精神支柱,” <<人民论坛>>，2020年3月.

陈岳云，“中华民族强大凝聚力从何而来？?” <<人民论坛>>，2020年9月.

陈岳云,“浅议‘金融为实体经济服务’,” << 学术展望>>,2021年9月,美国华裔教授学者协会.

Chen, Yueyun(Bill), and Nika Qiao,“An Analysis of China’s Economic Growth-the Effects of the WTO and the RCEP,” Int’l Journal of Business and Economics, Vol. 7, No. 2, 2022.

Chen, Yueyun(Bill), “China’s Economic Prospects: Can China Achieve Stable and High Economic Growth in Coming Decades?” Journal of Global Economics and Business, Vol. 3, No. 11, 2022.

陈岳云,“中国教育：成就、问题与展望,”<< 学术展望>>，2023年9月，美国华裔教授学者协会.

陈岳云，“金融文化、伦理和道德风险—美国的实践与经验教训,”<< 学术展望>>，2025年9月，美国华裔教授学者协会.

魏礼庆(主编)，<<美国高等教育>>，中国科学技术出版社，2025年9月.

Qiao, Nika and Yueyun Chen, “A Value-added Analysis of China’s Economy,” International Journal of Business and Systems Research, Volume 19, No. 2, 2026.

Correspondence Author

Yueyun Chen, Ph.D.
E-mail: billchen@uwest.edu.

神话造物与文明延伸：人工智能作为人类能力外化的社会技术分析

From Mythical Automata to Civilizational Extension:

A Sociotechnical Analysis of AI as Human Capability Externalization

何绍义 / Shaoyi He

School of Cyber & Decision Sciences
California State University, San Bernardino

摘要

本文提出"人工智能文明延伸框架"（Artificial Intelligence Civilizational Extension Framework, AICEF），用以解释人工智能如何嵌入人类长期的能力外化过程，并把常见的"神话—自动机—人工智能"线性进化叙事，重构为一个可被证伪的中层理论。本文区分三条相互关联却不可混同的谱系——文化想象谱系、技术实现谱系与制度嵌入谱系——并把人工智能的影响分解为认知任务、组织社会与基础设施三个层级。全文核心主张是"条件性认知延伸"命题：人工智能并不自动增强人类智能，唯有当系统被可靠地接入行动者的目标、反馈、校验与责任结构时，它才构成稳定的认知延伸；否则，同一系统可能同时产生替代、控制、依赖与权力集中等后果。本文以医疗、教育、公共治理与传媒四个领域的真实案例——Epic 脓毒症预测模型的外部验证失败、生成式人工智能引发的"元认知惰性"、荷兰儿童福利算法丑闻、以及 Arup 公司深度伪造诈骗案——检验并例示此命题，并据此提出生命周期化的"分布式责任矩阵"。本文的贡献不在于宣告一种不可逆的"文明终局"，而在于提供一个可被反例修正的分析框架，继而追问：人工智能何时、以何种机制、对哪些主体构成能力延伸，以及这种延伸如何被权力关系与制度条件所塑造。

关键词：人工智能；能力外化；延伸心智；分布式认知；社会技术系统；基础设施化；大语言模型；条件性延伸；责任治理

Abstract

This article develops the Artificial Intelligence Civilizational Extension Framework (AICEF) to explain how AI participates in the long-term externalization of human capabilities, recasting the familiar linear narrative of "myth–automaton–AI" as a falsifiable middle-range theory. It distinguishes three related but non-identical genealogies—cultural imagination, technical implementation, and institutional embedding—and analyzes AI's effects at three levels: cognitive tasks, organizational-social routines, and infrastructure. Its central claim is a conditional-extension thesis: AI does not automatically augment human agency; it becomes a stable cognitive extension only when reliably coupled to human goals, feedback, verification, and structures of accountability. Otherwise the same system may simultaneously produce substitution, control, dependency, and the concentration of power. The thesis is examined through four documented cases—the failed external validation of the Epic Sepsis Model, "metacognitive laziness" induced by generative AI, the Dutch childcare-benefits algorithm scandal, and the Arup deepfake fraud—spanning health care, education, public administration, and media. On this basis the article proposes a lifecycle-based distributed responsibility matrix. Its contribution is not a deterministic account of a civilizational endpoint, but a revisable framework for analyzing when, how, and for whom AI extends capabilities, and how such extension is shaped by institutions and power.

Keywords: artificial intelligence; capability externalization; extended mind; distributed cognition; sociotechnical systems; infrastructuralization; large language models; conditional extension; accountability.

一、引言：从宏大叙事转向可质询的理论问题

人工智能史通常以现代计算科学为中心展开：数理逻辑、可计算性理论、控制论、达特茅斯会议、符号主义、连接主义、深度学习与基础模型依次构成主要节点。这种叙述准确呈现了技术学科的形成，却较少回答一个更具广度的问题：为何不同社会会反复把记忆、计算、判断、分类、预测乃至语言生产，交由外部符号与人工装置承担？

另一类叙述则把塔洛斯、偃师造人、自动机与现代机器人连成一条漫长的"人工智能前史"。此类叙述富有启发，却容易把文化想象、文学母题、机械装置与计算系统之间的结构相似，误写为直接的历史继承。因此，本文不主张古代神话"预见"了现代人工智能，也不把自动机视为大语言模型的技术祖先。本文关心的是一种更有限、也更可检验的连续性：人类社会如何反复借助外部媒介与人工物件，重新配置其自身的能力。

这一视角的转换，看似只是措辞之别，实则关乎论证的可质询性。把人工智能看作"机器智能的自我进化"，问题会滑向"机器何时超越人类"这类无从检验的思辨；而把它理解为"人类能力的外化"，问题便落到可以经验追问的基础——它外化了谁的能力、它在什么条件下成立、代价由谁承担。本文正是要把"延伸"这个常被浪漫化的隐喻，锻造成一组可以拿去比对案例、也可以被反例推翻的分析判断。

据此，本文提出以下四个研究问题（RQ）：Sustainability

1. RQ1：神话造物、自动机与现代人工智能之间，可以建立何种类型的谱系关系，哪些关系不能成立？
2. RQ2：在何种条件下，人工智能可以被称为个体、组织或社会的"认知延伸"？
3. RQ3：模型规模、统计学习与工具化系统如何改变延伸的范围，又带来哪些认识论限制？
4. RQ4：当能力分布于数据、模型、平台、组织与使用者之间时，责任应如何配置？

本文的核心论点是：人工智能可以被理解为人类能力外化的一种当代形式，但这种"延伸"不是自动发生的。它是由技术性能、组织安排、权力关系、反馈机制与责任制度共同塑造的条件性结果。全文以下述次序展开：第二节交代研究设计与证据边界；第三节梳理理论基础；第四节建构 AICEF 框架并提出可证伪命题与延伸判断；第五、六节分别处理历史谱系与规模智能的认识论边界；第七节以四个真实案例检验框架；第八节提出治理矩阵；第九节自陈反论证与适用边界；第十节提出一个开放而非目的性的结论。

二、研究设计与证据边界

2.1 研究类型与方法

本文是一项理论建构研究，采用三种方法，它们相辅相成。其一，概念分析，用于区分能力外化、认知延伸、自动化、替代、基础设施化与治理责任等易被混用的范畴。其二，比较历史解释，用于考察不同文明中的人工造物想象与自动机实践，但不据此推断直接的技术传承。其三，理论指导下的案例检验：本文不追求行业覆盖的完备，而选择医疗、教育、公共治理与传媒四个差异显著的领域，每一领域锚定一个有据可查的真实案例，作为检验框架解释力的经验参照。

文献与案例的取舍遵循三条可复核的标准：其一，在相关理论传统中具有奠基性；其二，能够直接支持或挑战被讨论的具体主张；其三，尽量采用原始论文、权威专著、同行评议研究或正式政策框架，而非二手转述。凡涉及具体数据者，均以可追溯的权威来源为准。这一标准使本文在扩展论据时，能够抵御"以修辞充实证据"的诱惑——一个理论的说服力，不在于例证之多，而在于每一例证是否经得起质疑与追问。

2.2 证据层级

理论建构最易犯的错误，是用一类材料去证明它根本无力证明的结论——譬如用神话叙事证明技术事实，或用算法论文证明宏观社会后果。为避免这一弊端，本文预先设定一张证据层级表，明确每类材料能支持什么、不能支持什么。

材料类型	可支持的结论	不能支持的结论
神话与文学叙事	某一文化如何想象人工生命、自动行为或人造行动者	现代人工智能的直接技术起源或历史影响关系
自动机与工程文献	特定时期的机械模拟、控制结构与社会象征	与现代学习系统具有相同的智能机制
现代技术论文	模型架构、训练目标、性能与限制	社会后果必然发生，或模型具有意识
组织与行业案例	具体情境中的工作重组、风险与治理实践	所有地区与行业均出现同一趋势
规范与政策文件	责任、风险管理与制度目标	制度已被有效实施，或风险已被消除

以上分层不是形式上的谨慎，而是本文论证诚实性的基础。它意味着：第五节的历史材料只用于建立"文化母题的有限连续性"，而非技术因果；第七节的四个案例只用于例示"条件性延伸"的机制，而非证明某种普遍规律必然发生。证据的边界，即是结论的边界。

三、理论基础：从延伸心智到社会技术基础设施

3.1 延伸心智与分布式认知

本文的第一块理论基石，来自认知哲学中的"延伸心智"命题。Clark 与 Chalmers（1998）提出，某些稳定的可被使用并被行动者持续依赖的外部资源，可以成为认知过程的组成部分。关键在于，这一命题从一开始就是有条件的：并非任何工具都自动成为心智的一部分，成立与否，取决于外部资源与行动者之间是否形成可靠的耦合——可随时取用、被自动信赖、且在需要时确实进入了推理与行动。Clark（2008）在其后的系统阐述中进一步强调了这种耦合的动态性。这一"条件性"，正是本文"条件性认知延伸"命题的哲学先声：延伸从来不是工具的属性，而是耦合关系的属性。

Hutchins（1995）的分布式认知研究，则把分析单位从个体扩展到由人员、符号、仪器与程序共同组成的任务系统。在他对海军舰船导航的经典田野调查中，"计算航向"这一认知行为并不发生在某个人的头脑里，而分布于航海图、量具、口令与船员分工构成的整体之中。由此来看人工智能，其理论意义便不只是"机器替人思考"，而是它可能重新安排认知任务在整个系统中的分布——某些任务被移交给模型，某些被留给人类，而分布方式的改变，会连带改变谁掌握判断、谁承担错误。

3.2 技术中介与社会技术系统

第二块理论基石是"技术中介"。Verbeek（2005）提醒我们，技术不只是被动地听人使唤，它还在暗中塑造人怎么看世界、怎么做选择。一个常见的例子是产房里的B超：它本是一件检查工具，却改变了准父母与胎儿的关系——原本要到出生才"见面"的孩子，如今在屏幕上提前成了一个可观察、可测量、甚至可讨论去留的对象。工具一介入，人所面对的处境就变了。这对人工智能尤为重要：同一份病人数据，系统若以刺眼的红色警报呈现，医生倾向于立刻行动；若只列成一栏不起眼的概率，同样的信息就可能被轻轻略过。呈现方式本身，已经在替人做一半的决定。

3.3 基础设施化

第三块基石是基础设施研究。当一种技术从单一工具，转变为多个活动共同依赖、通常不可见、却一旦失效便产生广泛连带影响的底层系统时，可以说它获得了基础设施属性（Star & Ruhleder, 1996）。某样东西是不是"基础设施"，并没有一个固定答案，而是因人、因情境而定——同一套电力系统，对普通用户是隐形的背景，对电工却是每天打交道的前景。判断人工智能是否已成为"文明基础设施"，不能凭一纸行业清单，而应考察其依赖范围、互操作性、制度嵌入、替代成本、公共影响与系统性风险。正因如此，本文使用"基础设施化"（infrastructuralization）这一进行时的概念，而不说"AI 已成为基础设施"——它是一个正在发生、尚可干预、结局未定的过程，而非既成事实。这一措辞上的克制，同样服务于本文的反目的论立场：把"文明的下一阶段"留作一个开放的问题，交由公共选择去回答，而非一个已被技术预先写定的结局。

四、AICEF：人工智能文明延伸框架

4.1 三条谱系：为何"连续"必须先被限定

把神话、自动机与人工智能放在同一叙事里，最大的风险是把"主题的相似"误认为"技术的传承"。AICEF 的第一步，因此是把这条看似单一的谱系，拆成三条并行而不可互相替代的线索。

谱系	研究对象	连接机制	主要风险
文化想象谱系	神话造物、寓言、文学与拟人化叙事	母题复现、概念类比、文化记忆	把相似性误认为技术传承
技术实现谱系	自动机、控制论、计算机、机器学习与基础模型	工程积累、数学形式化、制度化研究	线性进步史与技术决定论
制度嵌入谱系	企业、专业组织、政府、市场与公共基础设施	部署、标准、采购、监管与劳动重组	忽视权力、利益与地区差异

三条谱系可以相互影响，却不能相互替代：神话为技术提供意义资源，技术为制度提供新的行动可能，制度又决定哪些技术得以规模化部署。这一拆分的作用，是把"人工智能古已有之"这类动人却含混的断言，替换为三个各有边界、各可检验的具体主张。它也直接回应了 RQ1——可以成立的是文化母题层面的有限连续，不能成立的是技术机制层面的直接继承。

4.2 三个分析层级

在这三条谱系之外，AICEF 沿着另一维度把人工智能的当代影响分解为三个层级，用以定位"延伸"究竟发生在何处。

层级	被延伸的能力	主要机制	可能结果
认知任务层	检索、分类、生成、识别、预测与辅助判断	人机耦合、外部记忆、模型推断、工具调用	增强、误导、依赖、技能变化
组织社会层	流程协调、劳动分工、专业决策与资源配置	工作流整合、标准化、平台化、绩效评价	效率提升、岗位重组、监控、权力集中
基础设施层	知识生产、公共服务、社会协调与风险治理	跨系统集成、数据与算力网络、标准与制度	公共能力扩展、系统脆弱性、数字鸿沟

三个层级的关键在于，每一层的"可能结果"都是双向的——既列出增强结果，也列出其反面结果。这一刻意的对称，是本文区别于乐观主义技术叙事的关键所在：延伸与损害并非先后发生的两个阶段，而是同一过程在不同条件下的两种走向。

4.3 五个可修正命题

综合上述三个谱系与三个层级，AICEF 提出五个核心命题。它们不是不容置疑的断言，而是被刻意表述为可被经验反例修正的形式。

命题一（社会技术性）：人工智能是一种社会技术构造，而非脱离人类制度独立演化的主体。它的能力、边界与后果，都在人—技术—制度的耦合中被决定。

命题二（条件性延伸）：人工智能只有在可靠性、可访问性、反馈与责任条件同时满足时，才构成稳定的

认知延伸；否则，它可能只是偶发工具，甚至是风险来源。这是全文的核心命题，第七节四个案例即为其经验之检验。

命题三（规模的有限性）：当代基础模型的能力具有显著的规模依赖性，但规模不是充分解释；架构、数据质量、训练目标、工具调用与推理时计算同样非常重要。因此，本文只保留"规模依赖能力"，而放弃"能力只来自规模"的强命题。

命题四（中介性）：人工智能的社会影响以任务重组与权力重配为中介，不应直接从技术性能推导出"职业消失"或"制度进步"。技术性能是原因链的起点，而非终点。

命题五（基础设施的双刃性）：基础设施化在扩大人工智能公共价值的同时，也放大其失误、垄断、依赖与不平等的系统性后果。公共价值与系统性风险，是同一枚硬币的两面。

4.4 条件性认知延伸判断

命题二若要可用，"条件"就必须被操作化为可逐项核查的判断。下表把抽象的"稳定延伸"，拆解为六个可对案例逐条打分的问题。这六条判断，把"延伸"从一个修辞性比喻，转化为可用于案例比较的分析工具。一个系统满足的判断越多，它构成稳定认知延伸的可能性越高；反之，一个系统缺失的判断越多，它越可能滑向替代、依赖或失控。第七节将用这把尺子，逐一丈量四个真实案例——我们会看到，每一桩广为人知的人工智能失败，几乎都可以精确地定位到某几条判断的缺失。

判断	核心问题
功能耦合	系统输出是否真正进入任务并改变了行动？
可靠可得	系统是否在需要时稳定可用，其性能边界是否可知？
反馈校验	使用者能否识别、纠正并记录错误？
目标一致	系统的优化目标是否与专业目标和公共目标相容？
责任可追溯	数据、开发、部署与使用各环节的责任是否可被追踪？
退出可能	行动者是否保有不用系统、复核系统、恢复人工流程的能力？

五、从神话造物到认知机器：有限连续性与历史断裂

5.1 文化原型而非技术祖先

希腊神话中的塔洛斯、皮格马利翁与中国古代《列子·汤问》中的偃师造人，反映了不同文化对人造生命、服从、情感投射与身体模拟的想象（Mayor, 2018；袁珂, 2006）。希腊神话中的青铜巨人塔洛斯绕岛巡行、执行指令；《列子·汤问》所载偃师为周穆王所造的木偶，"内则肝胆心肺、外则筋骨支节"，剖之而见其"皆假物"，去其心则口不能言、去其肝则目不能视（杨伯峻, 1979）——这是对生命内部机制的一次机械摹写。它们的重要性，在于揭示"人工行动者"并非现代社会才出现的文化问题：早在任何计算装置之前，人类已在想象中反复演练"制造心智"这件事，并预先思考了它带来的敬畏与不安。

然而，依照第二节的证据层级，这些材料所能支持的，仅止于"某种文化如何想象人工生命"，而不能证明古代文明已形成机器智能的理论，更不能证明它们是现代人工智能的技术源头。它们既不含现代计算意义上的算法，也不构成可训练的系统。将其称为现代人工智能的"文化前史"，比称为"技术起源"远为准确。文化想象谱系的价值是真实的，但它的边界同样也是真实的。

5.2 自动机：机械模拟与社会秩序

神话之后，文明开始把想象具象化为装置。早期现代欧洲的自动机——从水力驱动的发声人偶，到沃康松能进食、振翅、模拟消化的机械鸭——把运动分解、反馈控制与机械因果转化为可观察的对象（Riskin, 2016；Mayr, 1986）。思想史研究指出，这些装置远不只是精巧的玩物：它们同时是工程对象、宫廷权力的象征、宗教仪式的道具，以及关于"生命是否即机械"的哲学实验（Kang, 2011；Wood, 2002）。一只具有"消化"功能的机械鸭，逼迫当时的哲学家去追问：若生命的种种功能都可被机械复现，"活着"是否也不过是一套足够复杂的机械过程？值得强调的是，这种以机械模拟生命与运动的传统并非欧洲所独有：中国古代的指南车、记里鼓车、水运仪象台与各类"机关"装置，同样把畜力、水力与精密齿轮组织为可自动运转的系统，其机械工程成就在李约瑟的系统研究中得到详尽考证，相关装置的实物复原研究也进一步佐证了这一技术脉络（李约瑟, 1999）。东西方不约而同地出现了这些现象，旨在提醒我们，"以人工装置模拟生命与秩序"是一个跨文明的共同冲动，而非某一文化的专利。

5.3 计算的断裂性

现代人工智能的形成，依赖数理逻辑、可计算性理论、信息论、控制论、数字化，组织化与计算机技术的科研成果。这些成果不是古代自动机的自然延伸，而是一组全新的历史前提。Turing（1950）把"机器能否思考"这一易陷入定义之争的形而上学问题，转化为一个可操作的行为判断问题；1955 年起草、2006年重刊的由McCarthy、Minsky、Rochester 与 Shannon 等人工智能先驱提交的达特茅斯研究计划书，则把"智能的形式化与机器模拟"确立为一门新兴学科Artificial Intelligence (人工智能)的明确纲领（McCarthy et al., 1955/2006）。这里发生的是一种决定性的断裂：智能不再只通过机械运动被表现，而开始通过符号、概率与可训练的计算过程被操作化。

于是，谱系的图景变得清晰起来：在文化想象与技术实现之间，存在着主题与问题意识的连续——"能否制造会思考的东西"这个问题，从塔洛斯一直问到图灵；而在技术机制上，则横亘着决定性的断裂——回答这个问题的手段，从青铜与齿轮，变成了逻辑、概率与计算。连续与断裂并存，正是 AICEF 对 RQ1 的完整回答。

六、深度学习与基础模型：规模、能力与理解的边界

6.1 从规则编码到表示学习

深度学习带来的关键变化，不是简单地以数据取代知识，而是通过多层参数化模型，从训练任务中学习对预测有用的表示。与依赖显式规则的系统相比，表示学习降低了部分特征工程的成本，并在图像、语音与语言任务上取得显著进展（LeCun et al., 2015）。这一路径的深层意味，恰好在与延伸心智命题遥相呼应：符号主义试图把人类"已能明言的知识"搬进机器，而深度学习外化的，是人类那种"说不清却做得到"的能力，比如，识别一张脸、听懂一句话。但是，需要补充说明的是：这种能力仍取决于训练分布、目标函数与评估标准，不能自动推广

为常识、因果理解或可靠判断。一个在某分布上表现优异的模型，未必在另一分布上仍然可靠——这一限制，将在第七节的医疗案例中以较为突出的形式显现。

6.2 Transformer 与基础模型

Transformer 通过自注意力与并行计算，大幅度地提高了序列建模的可扩展性（Vaswani et al., 2017）。预训练模型随后经由任务提示、微调、检索、工具调用或人类反馈，适配于多种用途；GPT 系列所展示的"少样本学习"，正是这种适配能力的体现（Brown et al., 2020）。把这一过程概括为"把文明知识压缩进参数"，虽有修辞上的吸引力，却容易遮蔽训练数据的选择、版权、表示损失、更新困难与知识时效等一系列问题。更准确的表述是：模型参数编码了对训练数据分布有预测价值的统计结构，并可在一定条件下重组这些结构以便完成新的任务。此处对措辞的克制处理并非咬文嚼字——正是"压缩进文明知识"这类过度承诺的修辞，最容易滋生对模型能力的系统性高估。

6.3 规模不是充分理论

神经缩放研究显示，模型性能常随参数、数据与计算投入呈现经验性规律（Kaplan et al., 2020）；但下游表现还受模型形状、数据质量、训练策略与任务性质的影响。因此，本文保留"规模依赖能力"，而放弃"能力只来自规模"的强命题。规模解释的是能力增长的一个重要维度，而非智能的全部机制。这一区分对本文的整体论证至关重要：若能力只来自规模，则"延伸"几乎是自动的、只待算力堆积；而一旦承认能力还依赖数据、目标与耦合方式，"条件性延伸"命题便有了立足之地——延伸的成败，取决于那些规模之外、且大多由人与制度掌控的条件。本节对规模与能力关系的讨论，仅限于技术性能层面，从而不会自动推导出任何关于社会后果的结论。

6.4 "理解"争论的限定

Bender 与 Koller（2020）指出，仅从语言形式训练的系统，不能据此获得与外部世界和交际意图相联系的意义。这一论证有力地反对了把流畅输出直接等同于语义理解的倾向，其思想与 Searle（1980）著名的"中文房间"论证一脉相承：一套仅仅立足于操作符号的系统，无论表现得多么像懂中文，也未必真的理解中文。然而，本文不把"非理解"当作对所有模型内部状态的最终判决，而将其限定为一种认识论警示：行为上的成功，不足以单独证明模型具有与人类相同的理解能力。这一限定同样服务于条件性命题——正因为模型可能"能干却不懂"，它的输出才必须被接入人类的校验与责任结构，否则，流畅而自信的错误，将比笨拙的错误更难被察觉、也更加危险。

七、四个案例：延伸、替代与控制的分野

本节以四个来自不同领域、有据可查的真实案例，检验第四节提出的条件性延伸判断。四个案例的选择遵循"最大差异"原则——高风险临床、日常学习、公共行政、商业传媒——以考察同一组判断能否跨领域地区分"稳定延伸"与"延伸失败"。每一案例之后，均以六条判断作结构化诊断。

7.1 医疗：当"高准确率"遇上分布漂移

医疗常被视为人工智能延伸认知的典范领域——影像识别、风险筛查、文献检索与病历生成，无一不在扩展临

床能力。但一个真实案例揭示了延伸的脆弱性。Epic 公司开发的脓毒症预测模型（Epic Sepsis Model）曾被美国数百家医院部署，用以自动预警脓毒症的发生。2021 年，密歇根大学的研究团队对其进行外部验证：在近 38,000 次住院、约 27,000 名患者的真实数据上，该模型的敏感度仅为 33%，曲线下面积（AUC）仅 0.63，并因大量误报而引发"警报疲劳"（Wong et al., 2021）。也就是说，这个被广泛信赖、宣称高性能的系统，在真实临床环境中漏掉了约三分之二本应预警的脓毒症患者。

基于六条判断的分析：该系统的功能耦合是充分的（输出确会触发临床警报并影响处置），但可靠可得存在根本缺陷（在部署医院发生了分布漂移，性能边界未被使用者充分知晓）；反馈校验薄弱（临床医生难以识别模型为何误报、亦难以将纠正回馈系统）；目标一致亦有隐患（供应商的模型阈值未必对齐各医院的临床定义）。判断的多重缺失，恰好解释了为何一个"高准确率"的模型，并不构成稳定的认知延伸。这个医疗案例的启示是：延伸关系必须以外部验证、专业复核与持续监测为条件；医生的角色也并非简单地从"信息处理者"升格为"决策者"，而是承担起新的校验、解释与责任协调之责。

7.2 教育：认知外包与"元认知惰性"

生成式系统可提供即时反馈、语言支持与练习材料，因而可能降低个性化辅导的成本。但同一能力也可能反噬学习本身。一项发表于《英国教育技术学刊》的随机实验研究提供了直接证据：研究者让 117 名中国学生（以英语这一非母语完成读写任务）分别在 ChatGPT、人类专家、写作分析工具与无辅助四种条件下学习，发现依赖生成式人工智能的学生虽在当次任务的产出上有所提升，却表现出"元认知惰性"（metacognitive laziness）——他们把本应由自己承担的规划、监控与反思等元认知负荷，外包给了模型，长此以往可能削弱其自我调节学习的能力（Fan et al., 2025）。

基于六条判断的判断：此处的功能耦合与可靠可得都不是问题，真正缺失的是反馈校验与目标一致——当系统直接给出成品答案，学生失去了识别与纠正自身错误的过程，而"完成任务"这一即时目标，与"形成能力"这一教育目标发生了背离。这恰是命题二的一个精确注脚：延伸是否成立，不取决于系统是否好用，而取决于耦合方式是否保全了人类一侧必须亲历的认知过程。教育中的关键，因此不在于教师是否从"知识提供者"转为"学习设计者"，而在于哪些认知任务可以外包、哪些能力必须由学生亲身实践方能形成，以及师生是否保有审查与重构系统输出的自主权（对应"退出可能"判断）。

7.3 公共治理：被算法放大的不平等

在公共部门，人工智能既能延伸国家的行政能力，也可能延伸国家的监控与分类能力——荷兰儿童福利算法丑闻（toeslagenaffaire）是迄今最惨痛的例证之一。自 2013 年起，荷兰税务部门部署了一套风险分类模型，用以筛查儿童保育津贴中的疑似欺诈。该系统将"是否拥有非荷兰国籍"等变量纳入风险评分，系统性地把外国背景的家庭标记为高风险；一旦被标记，津贴便被中止并要求全额追缴，往往不经进一步的人工评估。至丑闻曝光时，约 26,000 个家庭被错误指控，许多家庭被迫偿还数万欧元、陷入财务崩溃，逾千名儿童一度被从家庭中带走。2021 年 1 月，荷兰内阁为此集体辞职；荷兰数据保护局其后认定该系统违法处理国籍数据并处以罚款。

这一案例几乎是条件性延伸判断的反面全景：功能耦合过强而人工缓冲缺失（风险分数被直接当作中止津贴的依据）；反馈校验形同虚设（受影响者难以知情、解释与申诉）；目标一致严重背离（"打击欺诈"的效率目标压倒了平等保护与程序正义）；责任可追溯断裂（"自学习"黑箱使责任难以落实到具体主体）；退出可能被剥夺（家庭无从摆脱系统的裁决）。它印证了 Eubanks（2018）的核心发现——自动化系统往往在福利、治安

等领域，把历史形成的不平等嵌入决策，并集中地伤害最脆弱的群体。此案的教训因此不是"提高算法精度"，而是必须追问一个 AICEF 反复强调的问题：谁的能力被延伸了，谁又承担了错误的成本？

7.4 传媒：生产扩展与信任基础设施的动摇

生成系统降低了文本、图像、音频与视频的生产门槛，也同步降低了伪造与规模化操纵的成本。2024 年初，跨国工程公司 Arup 的香港办公室遭遇了一起震动业界的深度伪造诈骗：一名财务人员先收到一封声称来自英国总部首席财务官的邮件，要求处理一笔机密转账；他起初心存怀疑，直到被邀请加入一场视频会议——会上"出现"了财务官与数名同事，音容笑貌与真人无异。在多位"熟悉面孔"的实时确认下，他打消了疑虑，分 15 笔转出约 2 亿港元（约 2,570 万美元）。事后向总部核实时，方知会议中除他之外无一真人，所有与会者皆为由公开视频资料合成的深度伪造（CNN, 2024）。

此案的判断诊断别具意味：问题不在某个单一模型的性能，而在支撑社会信任的整套基础设施被瓦解。反馈校验失效（受害者无从当场辨伪，因为伪造恰恰针对了"视频即可信"这一直觉）；可靠可得的判断被反向利用（"多位熟人同时在场"这一通常增强可信度的线索，成了施骗的杠杆）。深度伪造的真正威胁，正在于它侵蚀文明赖以运转的信任层——视觉的信任、听觉的信任、在场的信任。这也把第七节引向一个超越个案的判断：当人

工智能进入传媒与身份验证的基础设施层，真正的核心不再是单一模型，而是模型、平台分发、身份认证与新闻规范共同组成的生态。因此，治理的对象，也必须从"模型"上移到"生态"。

7.5 案例比较

四个案例横跨高风险临床、日常学习、公共行政与商业传媒，却指向同一结构性结论。

领域	主要延伸	关键风险	缺失的核心判断
医疗（Epic 脓毒症模型）	识别、筛查、记录与决策支持	分布漂移、误诊、警报疲劳	可靠可得、反馈校验
教育（元认知惰性）	反馈、生成、翻译与个别化支持	认知外包、能力退化、评价不公	反馈校验、目标一致
公共治理（荷兰福利丑闻）	分类、预测、资源配置与行政处理	歧视、不可申诉、监控扩张	责任可追溯、退出可能
传媒（Arup 深度伪造诈骗）	内容生产、检索、身份呈现	深度伪造、信任瓦解、规模操纵	反馈校验、可靠可得

经过比较以后的结论是清晰的：人工智能既不是单纯的工具，也不是单一的行动者；它通过嵌入不同的制度，获得不同的社会能力，也在不同判断的缺失处，滑向不同形态的失败。四桩广为人知的失败，无一是"模型不够大"所致，而无一不是某几条延伸判断的缺失所致。这正是命题二最有力的经验支持——延伸的成败，几乎从不取决于规模，而总是取决于耦合、反馈、责任与退出这些"规模之外"的条件。

八、治理：从"系统责任"到分布式责任矩阵

8.1 责任不能因分布而消失

本文第七节的四个案例，尤其是荷兰福利丑闻，暴露出人工智能治理中最隐蔽的陷阱：当一个系统的行为由数据、模型、接口、组织流程与使用情境共同产生，"分布式因果"极易导致"无人负责"。责任在链条中层层稀释，最终落不到任何一个具体的主体之上。本文由此而主张，治理的首要任务，正是抵抗这种稀释——把不同主体各自可控制的风险，与相应的义务精确对应起来。美国国家标准与技术研究院的人工智能风险管理框架（NIST AI RMF）所采用的"治理—映射—测量—管理"思路表明，风险治理应贯穿系统的整个生命周期，而非在产品上线后附加一场伦理审查（NIST, 2023）。

8.2 分布式责任矩阵

据此，本文提出一个生命周期化的分布式责任矩阵。它把系统从数据到监督的五个环节，与各环节的关键主体、典型风险及最低可审计义务一一对应。其设计原则是：每一环节都存在一个可被追责的主体，且其义务是"可审计的"——即能被外部独立地检查，而非仅凭自我声明。

环节	主要主体	典型风险	最低可审计义务
数据	数据提供者、采购方、管理机构	非法收集、代表性不足、标签偏差	来源记录、许可依据、质量与偏差评估
模型开发	开发者、基础模型供应商	能力误报、安全缺陷、不可复现	模型文档、测试边界、版本与事故记录
系统部署	采购方、集成商、机构管理者	场景错配、权限过度、工作流失控	影响评估、访问控制、人工复核与退出方案
具体使用	专业人员、普通用户、自动代理操作者	误用、过度依赖、未经核实传播	用途限制、核验记录、重大决定的人工确认
制度监督	监管者、法院、行业组织、审计机构	监管空白、利益冲突、救济不足	独立审计、申诉机制、报告与纠正权

当我们把这张矩阵回置于四个案例时，其解释力便立马呈现：Epic脓毒症模型的失败，主要落在"模型开发"环节的测试边界未明与"系统部署"环节的复核缺位；而荷兰福利丑闻的灾难，则是"数据"环节纳入国籍变量、"具体使用"环节以分数径行裁决、"制度监督"环节申诉机制失灵的三重叠加。矩阵的价值，正在于它能把一场笼统的"算法灾难"，分解为若干个各有责任主体、各可审计问责的具体失职。

8.3 治理原则

矩阵之上，本文提炼五条贯穿性原则。比例原则：风险越高，验证、文档、人工复核与外部监督越应严格。可追溯原则：关键输入、模型版本、输出使用与责任决定，均应保留审计链。可争议原则：受影响者应享有知情、解释、申诉与纠正的渠道——这一条恰是荷兰案例中最致命的缺失。可退出原则：组织应保留人工替代流程，避免关键公共功能被单一系统锁定。能力对称原则：在部署人工智能扩大机构能力的同时，必须同步扩大个人与公众的监督与救济能力——否则，"能力延伸"便只是单向地流向强者，而这正是下一节要正面处理的反论证。

九、反论证、适用边界与未来研究

9.1 "延伸"可能掩盖控制

对本文框架最有力的反论证是：把人工智能称为"人类能力延伸"，容易把"人类"当作利益一致的整体，从而掩盖了延伸背后的权力不对称。现实中，企业、政府、专业人员、劳动者与公众的能力，可能朝完全不同的方向变化——平台的预测能力扩大，劳动者的自主权却可能萎缩；行政效率提高，申请人的申诉空间却可能收窄。本文承认这一反论证的分量，并以之为界定框架的约束：AICEF 必须始终回答延伸的四个要素——主体（谁的能力）、对象（延伸了什么）、收益（谁获益）与成本（谁承担）。任何抽掉了这四问的"延伸"论述，都可能沦为为控制辩护的修辞。第七节的荷兰案例，正是这一警觉的经验教训。

9.2 认知外包可能导致能力退化

第二重边界是：外部工具既可能释放认知资源，也可能削弱练习、记忆与判断能力。是否构成净值增强，取决于任务设计、反馈质量与使用频率——第七节的教育案例（Fan et al., 2025）已给出直接证据。这就给我们指出了一个重要的未来研究方向：需要通过纵向实验与组织研究，检验哪些人机协作方式能提高长期能力，哪些只是短期提高产出、却以能力退化为代价。本文的判断提供了假设，而这些假设的实际检验，还需通过专门的实证工作才能完成。

9.3 跨文明比较的限制

第三重边界关乎方法。"跨文明"不应只意味着在希腊与中国神话中各举其例。严谨的比较，需要处理文本年代、译本差异、社会语境与概念的不可通约性。本文的历史部分只建立了有限的文化母题比较，并已在第二节的证据层级中明确其边界——它不能替代专业的区域技术史研究。承认这一限制，不是削弱论证，而是使论证更加夯实。

9.4 AICEF 的可证伪性

最后，当一个理论尚不能被反例修正时，充其量只是修辞而不能被称为理论。因此，本文明确列出足以挑战自身的反例情形：若在某组织中，人工智能系统能长期稳定地完成任务，却不需要任何人类目标、制度支持、数据供给、维护或责任结构，则命题一（社会技术性）将受到挑战；若系统在缺乏可靠性、反馈与可追溯性的情况下，仍持续增强行动者的能力，则条件性延伸判断需要修改；若基础设施化并未放大系统性风险，则命题五需要重新评估。清楚地知道"什么情况下我会错"，是把 AICEF 从文明修辞转化为可修正理论的最后一步，也是本文与它所批评的那类基于目的的叙事之间，最根本的区别。

十、结论：一个开放的问题，而非一个预定的终局

本文对"人工智能是人类自我延伸"这一命题，进行了收缩与重构。人工智能既不是从古代神话线性进化而来的技术，也不是脱离社会独立成长的智能主体。更准确地说，它是由文化想象、工程知识与制度部署共同形成的社会技术系统。神话与自动机提供的，是文化与机械模拟的前史；现代人工智能则建立在可计算性、统计学习、数字基础设施与组织化科研之上。二者之间，既有主题与问题意识的连续，也有决定性的技术与制度断裂。

AICEF 的核心贡献，是把"延伸"从一个动人的隐喻，转化为条件性、分层次且可质询的分析。人工智能在个体层面可以扩展检索、识别与生成，在组织层面可以重组工作流与专业分工，在社会层面可能进入知识、治理与资源配置的基础设施。但同一过程，也可能造成替代、监控、依赖、不平等与系统性风险。第七节的四个案例——脓毒症模型漏诊三分之二的患者、学生把思考外包给聊天机器人、荷兰两万余个家庭被算法误判、Arup 因几可乱真的深度伪造而损失两千余万美元——从四个方向共同证明：延伸的成败，几乎从不取决于模型有多大，而总是取决于耦合、反馈、责任与退出这些由人与制度掌控的条件。

于是，我们得以回到本文开篇的神话与造物——倒在克里特岛上的青铜巨人塔洛斯，与被肢解于周穆王面前的木偶。数千年来，人类反复想象、制造能够替自己行动与思考的造物；而每一次，真正的问题都不在造物本身有多精巧，而在人类为它安排了怎样的位置——它听谁的指令、由谁来校验、出了错谁来负责、以及谁知道它的命门所在。塔洛斯之死，恰恰系于膝盖处那枚遮掩着命门的铜钮。这个古老的意象旨在提醒我们：一个不留退出可能、无人可追责的强大系统，才是真正危险的系统。

因此，人工智能是否会成为文明的"下一阶段"，不是一个可以由模型规模单独回答的技术问题，而是一个取决于公共选择的开放问题。它的答案，不是写在参数里，而是写在我们为这些造物设计的目标、反馈、责任与退出机制里——一言以蔽之，写在我们是否愿意并能够随时拔掉那枚铜钮。文明的下一步走向何方，最终将不取决于机器会变得有多么聪明，而将取决于人类是否足够清醒。

参考文献 / REFERENCES

1. Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5185–5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463.
2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.
3. Chen, H., & Magramo, K. (2024, May 16). Finance worker pays out $25 million after video call with deepfake ‘chief financial officer.’ CNN Business. https://www.cnn.com/2024/05/16/tech/arup-deepfake-scam-loss-hong-kong-intl-hnk.
4. Clark, A. (2008). Supersizing the mind: Embodiment, action, and cognitive extension. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195333213.001.0001.
5. Clark, A. & Chalmers, D. (1998). The extended mind. Analysis, 58(1), 7–19. https://doi.org/10.1093/analys/58.1.7.
6. Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin’s Press.
7. Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. British Journal of Educational Technology, 56(2), 489–530. https://doi.org/10.1111/bjet.13544.
8. Kang, M. (2011). Sublime dreams of living machines: The automaton in the European imagination. Harvard University Press. https://doi.org/10.4159/9780674059412.
9. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., &

Amodei, D. (2020). Scaling laws for neural language models. arXiv. https://doi.org/10.48550/arXiv.2001.08361.

10. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539.
11. Mayor, A. (2018). Gods and robots: Myths, machines, and ancient dreams of technology. Princeton University Press.
12. Mayr, O. (1986). Authority, liberty, and automatic machinery in early modern Europe. Johns Hopkins University Press.
13. McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. AI Magazine, 27(4), 12–14. https://doi.org/10.1609/aimag.v27i4.1904 (Original work written in 1955).
14. National Institute of Standards and Technology (NIST). (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1.
15. Riskin, J. (2016). The restless clock: A history of the centuries-long argument over what makes living things tick. University of Chicago Press.
16. Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417–457. https://doi.org/10.1017/S0140525X00005756.
17. Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. Information Systems Research, 7(1), 111–134. https://doi.org/10.1287/isre.7.1.111.
18. Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433.
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008. https://proceedings.neurips.cc/paper/7181-attention-is-all-you-need.
20. Verbeek, P.-P. (2005). What things do: Philosophical reflections on technology, agency, and design (R. P. Crease, Trans.). The Pennsylvania State University Press. https://doi.org/10.5325/j.ctv14gp4w7.
21. Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penoza, C., Ghous, M., & Singh, K. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Internal Medicine, 181(8), 1065–1070. https://doi.org/10.1001/jamainternmed.2021.2626.
22. Wood, G. (2002). Living dolls: A magical history of the quest for mechanical life. Faber & Faber.
23. 李约瑟.（1999）． 中国科学技术史：第四卷・物理学及相关技术，第二分册・机械工程（鲍国宝、张景泰 译；王铃 协助）．科学出版社；上海古籍出版社.（原著出版于1965年）.
24. 杨伯峻.（2012）．列子集释（中华国学文库）．中华书局.
25. 袁珂.（2006）．中国古代神话．华夏出版社.

Correspondence Author

Shaoyi He, Ph.D.
E-mail: SHe@csusb.edu.

从学界到侨界，从大学到社会

From Academia to the Overseas Chinese Community, from University to Society

林连连 / Lianlian Lin

California State Polytechnic University, Pomona

编者注：这篇文章原题目为“深耕学界育桃李，投身侨界惠社区”，被收集和发表在《我们的美国故事》一书中，由美中和平友好协会和美国洛杉矶华文作家协会联合征文、主编和评审，由世界华人周刊出版影视集团于2026年7月出版，并获得征文大赛优秀提名奖。经本书主编同意转载。

时代的变迁改写着个人的命运。中国的改革开放，使我得以走出国门，远赴美国求学深造；而充满机会的美国，则成为我安居立业的第二故乡，给予我施展理想与抱负的广阔舞台。一段跨越中美的人生轨迹，一段跨越学界与侨界的人生征程，一条未曾设想过的人生路径，最终在这片土地上成为现实。时值美国建国250周年之际，回望来路，从学界的坚守，到侨界的奉献，思绪万千，谨以此文表达我的感激和敬意。

在学校：跨越语言与学习的门槛

我赴美留学，在宾夕法尼亚大学法学院攻读法学硕士学位。初到美国，首先面对的是语言和学业上的双重挑战。出国前，我的托福成绩为583分，在当时已属高分，主要得益于语法和阅读部分的优势，但听力是我英文能力中最薄弱的环节。法律课程涉及大量案例分析，听不懂教授讲解的要点，就难以把握案例要义，在撰写法律案例备忘录时就难以把握重点。为此，在老师的允许下，我将课堂内容录音，反复聆听，并与中国同学一起讨论分析案例。但大家共同研讨总结的结果，容易导致写作业表述上的趋同。所以，还是要靠个人的投入与坚持，以勤奋为径，以拼搏为舟。我购置了电子翻译辞典，用于练习听力和发音，常看英文电视节目，不断弥补短板。与此同时，我还要跨越另一道学习障碍，学会使用计算机。在国内时从未接触过计算机，刚到美国就需要用电脑完成各项作业。听说苹果电脑操作简单、用户友好，我便选择使用苹果电脑来完成学习任务。所幸出国前就用打字机练就了盲打技能，让我在文字输入上得以应对自如，又闯过了一道学习的关卡。最后顺利完成法学硕士学位，为我下一步的学术和职业发展奠定了坚实基础。

在家里：跨越代际之间的文化差异

在我们这样的新移民家庭中，子女或出生在美国，或年幼时便来到美国，代际之间往往存在着明显的文化差异。我儿子三岁半来到美国时，我正在德州大学奥斯汀分校攻读企业管理博士学位，学业繁忙。他来美仅两周，就被送入幼儿园，迅速融入美国小朋友的环境中。从那时起，他的饮食习惯便逐渐西化，汉堡、热狗、鸡块用微波炉一热即可，芹菜、黄瓜、胡萝卜直接生吃，做饭简单、吃饭也快、节省时间，我也就顺其自然。久而久之，母子之间在饮食上的差异减少了我们同桌共食的时间，也在无形中减少了讲中文、传递文化的机会。

为了在美国大学谋求教职，我对自己的英语能力提出了更高要求。起初，我听英语时总是先用中文理解，再用英文表达，反应难免慢半拍。为了培养英文语感和用英文思维的习惯，我刻意在家中与儿子完全用英语交流，训练自己直接用英语思考与表达。当我的英语逐渐熟练，顺利获得大学聘书时，儿子的中文却被耽误了。

一方水土养一方人。从不吃中餐，到不讲国语，再到价值观念与行为习惯的美国化。比如，他不认同“望子成龙”、“养儿防老”的传统观念，更强调独立、自由与自我实现。他上中学时选修了西班牙语，又自学日语，却始终缺乏学好中文的信心。我们同住一个屋檐下，却一度呈现出“同住不同食，同族不同语，同宗不同俗，同根不同念”的状态。

然而，在成长的过程中，他开始产生追寻身份认同的自觉性。他参加了“世界青少年中国寻根之旅”，在行走与体验中追本溯源，意识到自身的文化根脉。大学期间，他选修了中国现代史课程。他利用暑假前往北京大学参加中国语言与文化暑期项目，学习中文、国画与书法，逐渐完成对自身身份的再认识。在中西文化的碰撞与融合中，我们母子在多元文化的滋养下共同成长。如今，他也成为一名华裔“教二代”，在大学商学院任教，走出了一条属于自己的道路。

在学界：跨领域的中美教育交流

来美五年后，我相继完成硕士与博士学位，开始在加州州立理工大学波莫纳分校商学院任教，我讲授过本科、MBA、PMBA等课程，涵盖企业管理导论、国际管理、战略管理、组织管理与行为等。从复旦大学法律系教师，转而成为美国大学商学院教授，这不仅是学科领域的转换，更是教育理念、教学方式与师生关系的深刻转型。

我在复旦大学是国际金融专业，也旁听了法律课程，留校后在法律系任教两年，在国内率先开设“国际金融法”课程，并合著了《国际金融法》教材，入选全国高校教材选编计划，出国后亦再版发行。而在美国商学院教授企业管理课程，我需要持续研读大量商业案例，追踪国际经济动态，每年批阅数百份学生报告，对语言与判断力的要求很高。美国的教育特色是强调主动性、参与性和互动性，而不是一言堂。

面对新的挑战，我依旧沿用了老方法，不断尝试和探索，努力拼搏，持续精进，如愿成为管理学的终身教授，也得到了学生的认可。有的学生说，我点燃了他们对学科的热爱；还有学生说，我的课程帮他们找到了自己真正感兴趣的专业方向；也有学生在成为国际投行高管后，向学校捐赠了以我命名的奖学金。学生的成长和鼓励，成为我教学生涯中最持久的动力。

我长期致力于中美之间的教育与学术交流。在担任学校国际教育项目代理主任期间，我组织并实施了多项中美培训与研修项目。我曾任北京大学光华管理学院客座教授，给国际MBA班讲授现代西方企业管理课程；在北京大学汇丰商学院为研究生讲授跨国经营管理，均采用英文授课与原版教材。我还在中国多所著名高校讲学，在省委党校、世界五百强企业及民营企业家协会作专题讲座。在加州大学洛杉矶分校，我为国内高校校长与副校长代表团作了题为“美国高等教育的领导与管理”的讲座。

跨学科的教育背景与跨文化的学术经历，使我逐渐形成以跨学科、跨文化与国际比较为特色的研究方法，研究领域涉及企业管理、国际商务、法律、经济与金融等多个维度。有关中美的研究课题，我从理论与实践的视角，向外界呈现一个更立体、更真实的中国。我在《耶鲁大学国际法杂志》上发表关于国有化与征用的论文，为外国投资者理解中国法律环境提供参考；在《宾夕法尼亚大学国际经济法杂志》上发表了关于知识产权的论文，在美国社会科学研究网络（SSRN）上一度跻身这一专题最高点击率前十名。在我校与国内高校合作项目中，我还参与撰写了有关计算机软件知识产权保护的专著。

一次在北京钓鱼台国宾馆参加建馆四十周年美食节的经历，使我进一步思考中国饮食文化的国际传播。随着南加州“中国谷”的兴起，我与刘海铭教授通过南加州中餐业的发展，探讨了食物、烹饪、身份认同和跨国文化的问题，相关成果发表于《亚美研究杂志》，并接受了北京《科技周报》及美国《侨报》的专访报道。

作为海外学者代表，我应邀赴京参加教育部组织的建国五十周年“与祖国同庆”系列活动。在天安门观礼台观看阅兵式，在人民大会堂出席国宴，参加“面向21世纪创新人才发展战略研讨会”，我们与李岚清副总理、教育部长陈至立和科技部长朱丽兰进行交流。大家一致认为，强化教育与管理，提升人才素质，是国家发展的根基。

在侨界：跨越不同平台服务社团和社区

我的个人成长与职业发展，得益于东西方文明的滋养与中美教育的启迪。复旦大学深厚的人文底蕴塑造了我的精神底色，美国高等教育则培养了我的公民意识。也正因如此，我希望走出校园，走向社会，在更广阔的公共空间中服务侨界、回馈社区。我认为，公益从来不是空泛的口号，而是由一件件具体的事情汇聚而成的责任与担当。我从专业组织开始，逐步延伸至更广泛的社团与社区，在实践中不断拓展公共服务的边界。

我曾在历史悠久的南加州华裔教授会及其奖学金基金会分别担任会长与主席，为学生提供奖学金、为青年教师提供支持。之后，我们来自中国大陆的一群教授共同创立南加州华裔教授学者协会（后更名为美国华裔教授学者协会），我是创会副会长并起草了章程。在我出任会长期间，创办了《学术展望》杂志，多次担任主编，并建立了协会的首个官方网站，为学术交流搭建平台。

此后，我作为美国国际人才协会的两位创始人之一，致力于为各类人才提供跨国交流与资源共享的平台。协会举办《全球人才论坛》，推出一系列关于人才成长的公益讲座。我又发起成立美国教授专家联合会，推动不同领域之间的跨界合作。我们创办的《教育论坛》，邀请到时任中国驻洛杉矶总领馆教育参赞袁东博士、美国卡特中心刘亚伟博士作主旨演讲，我也就“如何适应美国的教育环境”分享了自己的经历和感悟。

在张素久大姐创建巾帼会的过程中，我帮助起草章程、办理注册、申请非营利组织资质，作为首届执行会长兼秘书长，我做了大量具体工作。在美国华人社团联合会，我曾任监事长兼秘书长、共同主席团主席，目前担任执行长。我发起组建美华联青年委员会，并担任总顾问，培养华裔青年薪火相传。我倡议创办了《美中圆桌论坛》，使之成为社团发声与对话的重要平台，推出各种论坛，并拓展至全英文论坛，吸引不同族裔参与，推动社区之间的理解与沟通。我随着美国华人社团联合会 “故乡行”代表团回国考察，赴京参加中国侨联成立六十周年纪念活动及国庆晚宴；在海外侨领登览天安门城楼时，我观看了“天安门百年回顾展”，深切感受到中华民族的奋斗历程与时代跨越。

在南加州侨界，我应邀参加了江泽民主席、朱镕基总理和习近平副主席访问洛杉矶的晚宴；还参加了习主席访问安纳伯格庄园和旧金山的欢迎活动。我多次获得杰出领导奖、社区服务奖和公益奖，我的简介被收录在多种著名的美国和世界名人录中，并获得马奇思终身成就奖。张素久大姐对我的评价是“中文英文都很好，大事小事都能做”。这些奖励和鼓励不仅是对我过往工作的肯定，更提醒我在公共服务的道路上持续前行。

在美国：跨文化的讲述我们华人的故事

我有幸成为张素久大姐英文自传《和平将军的女儿》编委会的成员。在这部著作历时八年的创作过程中，我与她有过持续而深入的交流与互动。从2017年与第一位美国作者的见面讨论会，到几年后与第二位作者的沟通会，再到2025年于洛杉矶举行的新书发布会，我从最初四十余位赞助人之一，发展成为编委会成员，深度参与编辑工作，全程见证了这部自传的诞生历程。这不仅是一部个人回忆录，更是对海外华人历史的珍贵记录与重要贡献。在张大姐的引领下，我也走入了推动中美民间交往与文化传播的实践之中。

曾有人问张素久大姐，为什么选择用英文写自传。她的回答坚定而清晰：“我们要用美国人能够理解的语言，讲述华人和中国的故事”。在她看来，英语不仅是一种工具，更是一座通往理解的桥梁。面对一个以英语为主流话语的世界，唯有主动发声，方能被听见；唯有以对方能够理解的语言讲述，方能真正抵达。张大姐对文化传播与传承的执着，体现在她精益求精的态度与持久不懈的行动力之中。她在家中组织的第一次与本书首位美国作者见面会时，已年逾八旬，却依然思维清晰，从容回溯过往岁月，为写作积累素材。

本书采访组两位学生成员到我家里访谈了解大姐的故事，我也提供了以前写过的关于大姐的书面资料。美国作者与张大姐和我多次以英文进行电话访谈。围绕本书的结构设计、史实核准与语言表达，张大姐始终亲力亲为，我们逐字逐句推敲细节。疫情期间，我们常在夜晚通过电话或线上讨论，往往一谈数小时。纵贯八年的写作周期，她所展现出的远见、格局与充沛的生命能量，给予我持续而深刻的激励。

2025年8月23日，我具体负责组织了美国华人社团联合会在迪士尼音乐厅隆重举办的张素久自传新书发布会，现场人潮爆满，气氛热烈。之后我走访了多家中国文博机构赠送这部自传，包括中国国家博物馆、中国华侨历史博物馆、中国国家图书馆、中国人民大学家书博物馆等。该书亦被赠予哈佛大学、斯坦福大学、加州大学洛杉矶分校东亚图书馆永久典藏，使这一华人记忆得以跨越地域与时间，被更广泛地保存与阅读。

张大姐的自传，不仅具有史料与收藏价值，更是一座跨越文化与国界、相互理解的桥梁。她承继了父亲张治中将军的家国情怀，又开辟出一条属于自己的独特道路，成为中美民间友好交往的文化使者。这部自传的首位英文评论作者马克博士，曾与我合作撰写多篇英文文章，介绍张大姐及其所代表的华人群体，刊登于《今日世界新闻》及谷歌热点新闻，将这些故事带入主流视野。在她辞世之后，我们发表了纪念文章《缅怀和致敬张素久大姐的光辉人生和精神遗产 — 社区之砥柱、文化之传承、大爱之典范》，以文字寄托深切的敬意与不尽的怀念。

结语

本文不仅是一段个人经历的回顾，更是一种时代背景下的见证 ，它见证了个体如何在历史进程中寻找自己的定位，也见证了不同文明之间，如何通过理解与对话走向更深层的连接。我的人生在两种文化之间展开：一端是中国深厚的人文传统，一端是美国开放的社会环境。在学界，我以教育与研究立身；在侨界，我以服务与连接为责；在文化之间，我努力成为一个讲述者与桥梁。我的人生就是在不断的学习和跨越中成长。人生原本难以预设，而真正重要的，是在时代给予的机遇中，走出一条有温度、有责任、也有回响的道路。这是我对美国建国250周年的献礼。

作者简介

林连连博士，美国加州州立理工大学波莫纳商学院管理学教授。在北京大学光华管理学院和汇丰商学院担任过客座教授，来美前在复旦大学留校任教。获宾夕法尼亚大学法学硕士学位、德州大学奥斯汀分校企业管理博士学位。美国洛杉矶华文作家协会会员，美国华人社团联合会执行长，美国教授专家联合会会长，美国国际人才协会主席。Lianlian Lin, Ph.D.,E-mail: lindallin2@gmail.com.

汉缅双重否定结构句法对应研究
—以显型、隐型二分框架为基础

A Study on Syntactic Correspondence of Double Negative Constructions in Chinese and Burmese: Based on the Binary Framework of Explicit and Implicit Types

曾小红和杨祖益 / Xiaohong Zeng and Zuyi Yang

厦门大学国际中文教育学院/海外教育学院
Chinese International Education/ Overseas Education Collage, Xiamen University

摘要

汉语与缅甸语同属汉藏语系，亲缘关系赋予二者诸多语法共性，而语序类型、否定表达手段又体现出鲜明的类型学差异。双重否定是汉语、缅甸语中共有的句法现象。本文采用显型双重否定、隐型双重否定二分分析框架以现代汉语双重否定结构系统为参照，从句法形式维度考察现代缅甸语与之对应的双重否定句法类型，探究缅甸语是否存在汉语中不具备的独有双重否定表达形式。

研究发现：汉语各类显型、隐型双重否定句法类型，在缅甸语中都能够找到功能相互匹配的句法实现形式；但两种语言双重否定构式的丰富程度存在显著不对称性。汉语“非 X 不可”构式在缅甸语中不存在严格同构的对应句式，仅有功能等价的替代表达；“不 M 不 X” 类构式对应的缅甸语同质结构重复度偏高，细分格式远少于汉语。在本文考察范围内，暂未发现缅甸语拥有现代汉语不存在的双重否定句法类型。二者表层句法差异主要源于否定标记系统的不对称性与语序类型分野：汉语属于 SVO①语序，具备“不、没、没有、非”等多元否定标记，双重否定构式形式丰富；缅甸语为 SOV 语序，核心否定语素仅有「ဖ」，需要依托各类句尾助词完成句法构建，表层语序逻辑与汉语呈现镜像反差。

本研究梳理并构建了汉缅双重否定结构系统对应谱系，丰富了汉藏语系否定范畴的跨语言句法研究，同时能够为缅甸来华留学生汉语语法教学、汉缅双向翻译实践提供句法层面的理论支撑。

关键词：汉语；缅甸语；显型双重否定；隐型双重否定；句法对比

Abstract

Chinese and Burmese both belong to the Sino-Tibetan language family. Their genetic relationship endows them with many grammatical similarities, while word order types and negative expressions reveal distinct typological differences. Double negation is a shared syntactic phenomenon in Chinese and Burmese. Adopting the binary analytical framework of explicit and implicit double negation, this paper takes the system of double negative constructions in Modern Chinese as a reference. It examines corresponding double negative syntactic patterns in Modern Burmese from the perspective of syntactic forms and explores whether Burmese possesses unique double negative expressions unattested in Chinese.

The study finds that all types of explicit and implicit double negative syntactic patterns in Chinese can find functionally matched syntactic realizations in Burmese. Nevertheless, the two languages show striking asymmetry in the richness of double negative constructions. The Chinese construction fei X bu ke has no strictly isomorphic counterpart in Burmese; only functionally equivalent alternatives are available. The Burmese structures corresponding to the Chinese pattern bu M bu X suffer a high degree of homogeneity, with far fewer sub-types than Chinese. Within the scope of this investigation, no double negative syntactic patterns exclusive to Burmese (and absent in Modern Chinese) have been identified. The conspicuous surface syntactic discrepancies between Chinese and Burmese double negation stem mainly from the asymmetry of negative marker systems and divergence in word order. As an SVO language, Chinese has diverse negative markers such as bu, mei, meiyou, fei, supporting abundant double negative constructions. Burmese is an SOV language with only one core negative morpheme ma. It relies on various sentence-final particles for syntactic formation, creating a mirror contrast with Chinese in surface logical sequences.

This research establishes a correspondence spectrum of Chinese-Burmese double negative syntactic constructions, advancing cross-linguistic syntactic research on negation within the Sino-Tibetan language family. It also provides syntactic theoretical support for teaching Chinese grammar to Burmese students in China and Chinese-Burmese bidirectional translation.

Keywords: double negation; Chinese; Burmese; explicit double negation; implicit double negation; syntactic comparison

引言

汉语与缅甸语同属汉藏语系，相近谱系背景带来语法层面诸多共性，而语序、形态手段的分化又使二者在否定范畴上表现出不同特征。双重否定是汉语、缅甸语普遍存在的句法现象，长期受到语言学界关注。学界界定双重否定主要存在两条路径：一是单纯以形式标准判定，主张句子共现两处否定成分即为双重否定；二是采用“形式—语义”相结合的综合判定标准。本研究采用综合判定标准，即：形式上具备两处否定要素，一重否定对另一重否定命题形成辖制，底层逻辑记作「¬（¬p）」，整体获得“-否定”（否定之否定）解读；沿用曾小红、高清辉（2022）构建的现代汉语双重否定结构系统，区分显型双重否定与隐型双重否定，并以此作为统一标尺开展汉缅跨语言句法对比。

⑴ 本文所用符号，P 表示命题，X 表示谓词或谓词性短语，N 表示体词或体词性短语，M 表示能愿动词。

汉语属于 SVO 型语言，依靠功能分化的多重否定标记衍生出丰富的双重否定子类；缅甸语属于 SOV 型语言，核心依靠否定语素「မ」结合句尾情态助词构建否定表达。两类语言共享“否定之否定”的底层逻辑，但表层句法结构存在明显差异。而现有汉缅对比相关研究大多聚焦单一否定词习得问题，尚未针对双重否定构式开展系统性句法对照。从国际中文教育实践看，缅甸学习者常依托母语句法结构发生负迁移，直接套用缅语固有语序直译汉语双重否定句式，容易产生“不是不 X”“没有不 X”等句式偏误。基于上述背景，本文将研究重心限定于句法形式层面，重点考察汉语各类双重否定构式在缅甸语中是否存在对应的句法形式，探寻缅甸语是否存在汉语体系之外独有的双重否定句法类型。至于两种语言中双重否定结构的使用频率、功能负荷差异、语义分化与语用约束条件等，本文暂不深入展开，相关内容将纳入后续专题研究。

一、 汉缅显型双重否定结构对比

显型双重否定结构内部划分为直接型与间接型。直接型表层具备两个显性否定标记，底层逻辑可直接记作「¬（¬p）」；间接型同样包含两处显性否定要素，需要借助量词否定、模态运算、否定词移位等句法推导，才能完整呈现“否定之否定”逻辑关系。

现代汉语直接显型双重否定包含三类：“不是不Ｘ”类、“不是没（有）Ｘ”类、“没（有）不Ｘ”类。间接显型双重否定分为四类：“没（有）Ｎ不Ｘ”类、“不Ｍ不Ｘ”类、“非Ｘ不可”类、否定词移位类。需要预先说明：现代汉语拥有“不、没、别、甭、莫、勿”等多个功能分化的否定标记；现代缅甸语核心否定语素仅有「မ」，功能接近汉语“不”，置于谓词之前实现否定，同时常搭配「ပါ、ပေ、ချေ、ဘူး」等句尾助词，口语环境中「ဘူး」使用最为广泛。

（一）直接型显型双重否定

1. “不是不X”与缅语「မX……မဟုတ်ဘူး」

（1）我不是不想吃饭。

（1’） ကျွန်တော် ထမင်း မ စား ချင် တာ မ ဟုတ် ဘူး။
（我） （饭） （不） （吃）（想）（助词）（不） （是）（助词）

例（1）“不是”对命题“我不想吃饭”实施二次否定，逻辑形式为「¬（¬p）」。受SOV 语序制约，例（1’）表层序列呈现「¬p+¬」，前一个「မ」构成否定命题，后置「မဟုတ်ဘူး」对前序命题进行再否定，经句法推导后底层逻辑同样为「¬（¬p）」。汉语语义重心后置谓语，缅甸语谓语居于句末，语序差异造成表层形式不对等。

缅语「မX……မဟုတ်ဘူး」句法功能基本对应汉语“不是不 X”，可用于委婉表述或强化肯定，语义解读高度依赖上下文，极少独立使用。例如：

（2）她不是不买，是你不卖。

（2’） သူမ က မ ဝယ် တာ မ ဟုတ် ဘူး၊ ခင်ဗျား က မ ရောင်း တာ ပါ။

（她）（助词）（不）（买）（助词）（不）（是）（助词）（你）（助词）（不）（卖）（助词）（句尾助词）。

（3）我不是不说话，而是不懂说什么。

（3’）ကျွန်တော် က မ ပြော တာ မ ဟုတ် ဘူး၊ ဘာ ပြော ရမှန်း မ သိ တာ ပါ။
（我） （助词）（不）（说）（助词）（不）（是）（助词）（什么）（说）（应当）（不）（知道）（助词）（句尾助词）。

2. “不是没（有）X”与缅语「မX ခဲ့/သေးဘူး/ဖူး……မဟုတ်ဘူး」

（4）我不是没有写作业。

（4’）ကျွန်တော် အိမ်စာ မ ရေး ခဲ့ တာ မ ဟုတ် ဘူး။
（我） （作业）（没）（写）（过去时）（助词）（不）（是）（助词）。

汉语“没”用于标记动作未发生、事件未完成。缅甸语不存在形式上与“没”对应的否定词，依靠「မ」搭配「ခဲ့、သေးဘူး、ဖူး」等助词区分时体意义：「ခဲ့」标记已然动作，「သေးဘူး」表达动作尚未实现，「ဖူး」多用于经历否定。句法层面，缅语动宾结构否定遵循“宾语+မ+动词”语序。缅语「မX ခဲ့/သေးဘူး/ဖူး……မဟုတ်ဘူး」在句法功能上基本匹配汉语“不是没有X”。

（5）对于儿子的提醒，他不是没想过。

（5’）သား ရဲ့ သတိပေးချက် ကို သူ မ စဉ်းစားခဲ့ တာ မ ဟုတ် ဘူး။
（儿子）（的） （提醒） （宾助）（他）（没）（想过）（助词）（不）（是）（助词）。

（6）这种亏我们不是没吃过。

（6’）ဒီလို ဆုံးရှုံးမှုမျိုး ကို ကျွန်တော်တို့ မ ကြုံဖူး တာ မ ဟုတ် ဘူး။
（这类） （损失） （宾助） （我们） （没）（遭遇过）（助词）（不）（是）（助词）。

3.“没(有)不 X”与缅语「မX…… (N……) မရှိဘူး / မX……မဟုတ်ဘူး」

（7）家里人没有不宠她的。

（7’）မိသားစု ထဲ မှာ သူမ ကို မ ချစ်မြတ်နိုး တဲ့ သူ မ ရှိ ဘူး။
（家庭）（里面）（位助）（她）（宾助）（不）（宠爱） （的）（人）（没）（有）（助词）。

（8）我没有不喜欢红色。

（8’）ကျွန်တော် အနီရောင် ကို မ ကြိုက် တာ မ ဟုတ် ဘူး။
（我） （红色）（宾助）（不）（喜欢）（助词）（不）（是）（助词）。

汉语区分“没有”与“不是”，二者否定对象、语义重心存在差异；缅甸语仅具备单一否定语素「မ」，弱化了“不”与“没”的形式区分。表达存在类周遍否定时使用「မX……（N……）မရှိဘူး」；侧重对心理命题进行再否定时，选用「မX……မဟုတ်ဘူး」。两类结构均可对应汉语“没（有）不 X”构式。

（9）做生意的，没有不想把生意做大的。

（9’）စီးပွားရေး လုပ် တဲ့ သူတွေ ထဲ မှာ ကိုယ့် လုပ်ငန်း ကို ကြီးထွားအောင် မ လုပ် ချင်
（生意） （做）（的）（人们）（之中）（位助）（自身）（事业）（宾助） （壮大） （不）（做）（想）

တဲ့ သူ မ ရှိ ဘူး။
（的）（人）（没）（有）（助词）。

（10）我没有不想你。

（10’）ကျွန်တော် မင်း ကို မ လွမ်း တာ မ ဟုတ် ဘူး။
（我） （你）（宾助）（不）（想念）（助词）（不）（是）（助词）。

（二）间接型显型双重否定结构对比

4.“没有 N 不 X”与缅语「မX……N……မရှိဘူး」

（11）没有家长不爱自己的孩子。

（11’）ကိုယ့် သားသမီး ကို မ ချစ် တဲ့ မိဘ မ ရှိ ဘူး။
（自己）（子女） （宾助）（不）（爱）（的）（父母）（没）（有）（助词）。

“没有 N 不 X”属于周遍性主语结构，“没有”限定名词成分，“不”否定谓词，经过量词否定推导形成双重否定。缅语「မX……N……မရှိဘူး」实现相同句法功能，修饰成分置于名词之后，是缅甸语典型语序特征。

（12）没有一个人不知道朋友的重要性。

（12’）သူငယ်ချင်း ရဲ့ အရေးကြီးမှု ကို မ သိ တဲ့ သူ တစ်ယောက် မှ မ
（朋友） （的） （重要性） （宾助）（不）（知道）（的）（人） （一个） （强调助词）（没）

ရှိ ဘူး။
（有）（助词）。

5.“不 M 不 X”与缅语「မX……မM」

M 代表能愿动词。“不 M 不 X”通过模态词叠加否定构成双重否定，语义不等同于简单肯定，带有强制、必然等情态。

（13）他不能不帮她。

（13’）သူ သူမ ကို မ ကူညီ လို့ မ ဖြစ် ဘူး။
（他）（她）（宾助）（不）（帮助）（连接助词）（不）（能）（助词）。

缅语依靠「မ.....လို့ မဖြစ်ဘူး」表达“不做某事不行”，对应汉语“不能不、不可以不”等能愿类双重否定。

（14）你不应该不告诉我。

（14’）မင်း ငါ့ ကို မ ပြော ဘဲ မ နေသင့် ဘူး။

（你）（我）（宾助）（不）（说）（伴随助词）（不）（应当）（助词）。

（15）男人是不可以不成功的。

（15’）ယောကျ်ား တစ်ယောက် က မ အောင်မြင် လို့ မ ဖြစ် ဘူး။

（男人） （一个） （主助）（不）（成功）（连接助词）（不）（可以）（助词）

6.“非 X 不可”与缅语「မXလို့.....မဖြစ်ဘူး」

（16）这件衣服我非试不可。

（16’）ဒီ အင်္ကျီ ကို ကျွန်တော် မ စမ်းကြည့် လို့ မ ဖြစ် ဘူး။

（这件）（衣服）（宾助） （我） （不）（试穿） （连接助词）（不）（行）（助词）。

汉语“非”是独立否定标记，缅甸语不存在形式对应的否定语素，无法形成结构完全同构的句式，仅能依托语义转译，使用「မXလို့.....မဖြစ်ဘူး」表达“必须 X”，实现功能对等。

（17）我非去不可！

（17’）ကျွန်တော် မ သွား လို့ မ ဖြစ် ဘူး။

（我） （不）（前往）（连接助词）（不）（行）（助词）

（18）喜欢吃的食物非吃不可。

（18’）ကြိုက် တဲ့ အစားအစာတွေ ကို မ စား လို့ မ ဖြစ် ဘူး။

（喜欢）（的） （食物） （宾助）（不）（吃）（连接助词）（不）（行）（助词）。

7.否定词移位类

（19）我不认为他不在家。

（19’）သူ အိမ်မှာ မ ရှိ ဘူး လို့ ကျွန်တော် မ ထင် ဘူး။

（他）（家中）（不）（在）（助词）（引述助词）（我） （不）（认为）（助词）。

否定词移位现象普遍存在于两类语言。缅甸语「ရှိ」兼具动词“存在、在”含义，翻译“在家”时需要调整表达。从句法结构观察，缅语主句否定词同样可以跨小句发生移位，形成与汉语平行的双重否定格局。

（20）我不相信他不在家。

（20’）သူ အိမ်မှာ မ ရှိ ဘူး လို့ ကျွန်တော် မ ယုံ ဘူး။

（他）（家中）（不）（在）（助词）（引述助词）（我） （不）（相信）（助词）。

（21）我不认为他的观点不好。

（21’）သူ့ အမြင်တွေ ကို မ ကောင်း ဘူး လို့ ကျွန်တော် မ ထင် ဘူး။

（他） （观点）（宾助）（不）（好）（助词）（引述助词）（我） （不）（认为）（助词）。

（三）汉缅显型双重否定句法异同小结

整体来看，汉语七类显型双重否定，缅甸语均存在功能对等的双重否定句法表达手段， 但缅甸语中的重复度较高；在本文考察范围内，未发现缅甸语独有的、汉语不存在的显型双重否定句法类型。

表 1 汉缅显型双重否定结构对应表

汉语	缅甸语
1. “不是不 X”	“မX……မဟုတ်ဘူး”（不 X 不是）
2. “不是没（有）X”	“မX ခဲ့/သေးဘူး/ဖူး……မဟုတ်ဘူး”（没有 X 不是）
3. “没有不 X”	“မX……（N……）မရှိဘူး”（不 X（N）没有） “မX……မဟုတ်ဘူး”（不 X 不是）
4. “没有 N 不 X”	“မX……N……မရှိဘူး”（不 XN 没有 ）
5. “不 M 不 X”	“မX……မM”（功能对等）
6. “非 X 不可”	“မXလို့……မဖြစ်ဘူး”（无同形构式，功能对等）
7. “否定词移位类”	“အငြင်းလက္ခဏာရွှေ့ပြောင်းမှု”（否定词移位）

二者句法层面差异来源于两大要素：其一，否定标记系统不对称；其二，语序类型分野。汉语 SVO 语序，双重否定逻辑顺序表层即可体现「¬(¬p)」；缅甸语 SOV 语序，双重否定表层统一呈现「¬p+¬」，需要句法推导还原底层逻辑。

二、汉缅隐型双重否定结构对比

隐型双重否定不依靠表层两个显性否定标记，而是依托句式特征、反问语气、现场会话互动，形成对否定命题的二次否定，听话人需要借助句法与语境推理获得「¬（¬p)」解 读。 依照曾小红、高清辉（2022）的划分，隐型双重否定进一步分为类型型、实例型两类。类型型依靠固定句式自带解读倾向；实例型完全依托即时对话语境生成双重否定含义。

（一）类型型隐型双重否定结构对比

1.是非反问类

依靠反问语气对内部否定命题进行二次否定。汉语反问主要依靠语调实现；缅甸语通过句尾语气助词「လား（口语）/ လော（书面语)」标记反问。

（22）他不就是我们要找的人吗？

（22’）သူ က ကျွန်တော်တို့ ရှာနေတဲ့ သူ မ ဟုတ် ဘူး လား။

（他）（主助）（我们） （在找的）（人）（不）（是）（助词）（疑问助词）？

（23）难道这不是显而易见的事实吗？

（23’）ဒါ က သိသာနေတဲ့ အမှန်တရား ပဲ မ ဟုတ် ဘူး လား။

（这）（主助）（显而易见的）（事实） （助词）（不）（是）（助词）（疑问助词 ）

2.特指反问类

疑问词搭配否定成分构成反问，实现双重否定解读。缅甸语特指反问常用句尾助词「လဲ / မလဲ」。

（24）有什么不能说的？

（24’）မ ပြော နိုင် တာ ဘာ ရှိ လို့လဲ။

（不）（说） （能）（助词）（什么）（有）（反问助词）？

（25）南京这么有名的城市怎么能不去呢？

（25’）နန်ကျင်း ဒီလောက် နာမည်ကြီး တဲ့ မြို့ ဘယ်လိုလုပ် မ သွား ဘဲ နေ မလဲ။

（南京） （这么） （有名） （的）（城市） （怎么能） （不）（去）（强调助词）（反问助词）？

3.祈使类

汉语典型格式“别/甭不 X”；缅甸语依靠「မ.....နဲ့」否定祈使句式实现相近表达。

（26）你甭不理我。

（26’）ကျွန်တော့် ကို အရေးမလုပ် ဘဲ မ နေ နဲ့။

（我） （宾助）（不理） （伴随助词）（别） （祈使标记）。

（27）活动别不办了。

（27’）ပွဲ မ လုပ် ဘဲ မ နေ နဲ့။

（活动）（不）（举办）（伴随助词）（别） （祈使标记）。

4. 感叹类

“什么/哪里+不……”构成感叹式双重否定，语气强度高于普通陈述句。

（28）什么不能说！

（28’）ဘာလို့ ပြော လို့ မ ရ မှာလဲ။

（为什么） （说）（引语助词）（不）（可以）（反问助词）！

（29）哪里没来！

（29’）ဘယ်မှာ မ လာ လို့လဲ။

（哪里） (没) （来） （反问助词）

（二）实例型隐型双重否定结构对比

实例型隐型双重否定产生于即时会话互动，脱离对话语境则无法生成双重否定解读，分为副词否定类和疑问词否定类。

5.副词否定类

对话中一方提出否定判断，另一方使用否定应答词对前序否定命题实施再否定。汉语常用“不、没有”；缅甸语统一使用「မဟုတ်ဘူး」进行否定回应。

（30）A:你为什么要跟他分手？你不爱他？

B: 不，我爱他，正因为我爱他，我才要离开他。

（30’）A：မင်း ဘာလို့ နဲ့ လမ်းခွဲ လဲ။ မင်း ့ ကို မ

A：（你）（为什么）（他）（和）（分手）（助词）（反问助词）？（你）（他）（宾助） （不）

ချစ် ဘူး လား။

（爱）（助词）（疑问助词）？

B: မ ဟုတ် ဘူး ငါ ့ ကို ချစ် တယ် ငါ ့ ကို ချစ်

B:（不）（是）（助词）（我）（他）（宾助）（爱）（句尾助词）（我）（他）（宾助）（爱）

့ အတွက် ငါ သူ့ ကို ထား ခဲ့ တယ်။

（的）（缘故）（我）（他）（宾助）（离开）（句尾助词）。

（31）A：你怎么还没睡？又不舒服了？

B：没有。

（31’）A：မင်း ဘာလို့ မ အိပ် သေး လဲ။ နေမကောင်း ပြန်ပြီ လား။

A：（你）（怎么）（不）（睡觉）（还）（助词）（反问助词）？（不舒服）（又）（疑问助词）？

B：မ ဟုတ် ဘူး။

B：（不）（是）（助词）。

6.疑问词否定类

使用疑问代词否定对方提出的否定观点，形成互动层面双重否定。

（32） A：我不知道你今天会来。

B：什么不知道，我不是明明打过电话来吗！

（32’）A： မင်း ဒီနေ့ လာ မယ် လို့ ငါ မ သိ ဘူး။

A：（你）（今天）（来）（要）（引语助词）（我）（不）（知道）（助词）。

B： ဘာ မ သိ တာလဲ ငါ ဖုန်း ခေါ်ခဲ့တယ် မ ဟုတ် ဘူး လား။

B：（什么）（不）（知道）（反问助词）（我）（电话）（打过） （不）（是）（助词）（疑问代词）？

（33）A：你们不离婚了？

B：谁说的！

（33’）A：ခင်ဗျားတို့ မ ကွာရှင်း တော့ ဘူး ပေါ့။

A：（你们） （不）（离婚）（助词）（助词）（句尾助词）？

B：ဘယ်သူ ရေြာ တာ လဲ။

B：（谁） （说）（助词）（反问助词）！

（三）隐型双重否定小结

从句法实现形式来看，汉语类型型、实例型隐型双重否定全部子类，缅甸语均可用对应的句法手段表达，二者对应关系详见表 2。隐型双重否定不依赖固化构式模板，更多依靠语气助词、会话互动机制，语序差异带来的影响弱于显型双重否定。同样，本文考察范围内未发现缅甸语具备汉语不存在的隐型双重否定句法类型。

表 2 汉缅隐型双重否定结构句法对应表

	汉语隐型双重否定子类	缅甸语对应句法形式
1	是非反问类（不是……吗？）	မဟုတ်ဘူးလား
2	特指反问类（有什么不……？怎么能不……？）	မ……ဘာရှိလို့လဲ ဘယ်လိုလုပ် မ ……မလဲ
3	祈使类（别/甭不 X）	မ……ဘဲမနေနဲ့
4	感叹类（什么不……！）	ဘာလို့……မ……လဲ
5	会话副词否定类（对话否定应答：不，……）	မဟုတ်ဘူး
6	会话疑问词否定类（什么不知道……）	ဘာမသိတာလဲ

三、结语

本文以显型、隐型双重否定二分框架作为分析工具，从句法形式层面对现代汉语与现代缅甸语双重否定结构展开系统对比。研究以汉语双重否定结构系统为参照，重点考察两种语言双重否定句法类型的对应关系。

（一）研究核心结论

第一，汉语中不同类型的双重否定结构，在缅甸语中均能找到功能对等的双重否定句法实现形式，但二者构式丰富程度存在显著不对称。汉语“非 X 不可”属于高度固化构式，缅甸语不存在同构表达，只能依靠功能对等句式转述，直观体现汉缅构式系统存在差异；汉语“不 M 不 X”对应的缅甸语细分格式数量更少，不同能愿动词、助动词参与构成的表达同质化严重，结构重复度偏高。同时，由表 1 可见，汉语“没有 N 不 X”相关句式在缅甸语内部区分度微弱。上述现象共同表明，汉缅双重否定的差异，既体现在构式类型的丰富性层面，也集中反映于表层句法排列、否定标记的组合模式之中。

第二，两种语言表层句法差异源于两大核心要素。其一，否定标记系统不对称：汉语拥有功能分化的多个否定词语，缅甸语仅有核心否定语素「ဖ」，需要依靠句尾助词补足语法语义；其二，

语序类型分野：汉语为 SVO 语序，双重否定常直接呈现「¬（¬p）」表层序列；缅甸语属于 SOV 语序，表

层普遍呈现「¬p+¬」，需要通过句法推导还原底层逻辑结构。本研究梳理建立了汉缅双重否定结构系统的对应谱系，充实了汉藏语系内部双重否定范畴的类型学句法研究；可为缅甸来华留学生汉语语法教学、汉缅双向翻译提供句法层面参考，有利于辨识由母语语序迁移造成的句法偏误。

需要说明的是，句法构式能够对应，并不意味着两种语言的双重否定结构在实际语言运用中完全等同。初步观察显示，汉缅双重否定细分格式、功能负荷、使用频次存在明显区别，汉语双重否定在口语与书面语中的使用频率显著高于缅甸语。语体分布、使用频次、语用制约等议题本文暂不深入探讨，相关量化统计与语用考察将另撰文展开研究。

（二）研究不足与展望

本文将研究重心限定于句法形式对比，重点考察各类双重否定构式的存在形式及其句法对应关系，尚未系统开展语义歧解、语用制约条件、语体差异等维度的分析。此外，研究所用语料以内省例句、既有文献例句为主，尚未借助大规模双语平行语料，对各类构式的能产性、使用频次展开量化统计；研究对象仅聚焦标准书面缅甸语与现代汉语，暂未纳入缅甸本土方言变体进行考察。

后续研究可从四个方向进一步拓展：第一，依托双语平行语料库开展量化分析，统计各类双重否定构式的使用频次，对比两种语言双重否定结构的功能负荷差异；第二，在句法对比的基础上，推进汉缅双重否定语义解读、语用功能的比较研究；第三，拓宽语言考察范围，开展汉语与东南亚多国语言双重否定的句法对照，挖掘区域语言否定范畴的类型学特征；第四，结合二语习得实证实验，考察缅甸汉语学习者习得各类双重否定构式时面临的句法难点，为国际中文教学提供实证依据。

参考文献

[1] 曾小红,高清辉. 现代汉语双重否定结构系统[J]. 海外华文教育,2022(6).P43-51

[2] 张焕香.汉英双重否定范畴研究[D].首都师范大学,2012.

[3] 黄佳丽(ZU LEI PHYU).现代汉语否定词“不”与缅甸语否定词“(?)”对比[D].云南大学,2012.

[4] 实实用缅甸语语法汪大大 、杨国影主编—北京:北京大学出版社,2016.6

[5] 王君方.浅谈缅甸学生否定词“不”和“没”习得偏误[J].戏剧之家,2018

作者简介：曾小红、女、汉族、厦门大学国际中文教育学院/海外教育学院副教授，博士，研究方向为现代汉语语法、汉外语言对比、国际中文教育。邮箱：zengxiaohong@xmu.edu.cn；杨祖益，女，缅甸，厦门大学国际中文教育学院/ 海外教育学院 2024 级本科生。邮箱：yangjiayun71@gmail.com

MentorX

蔓藤教育

GLOBAL TALENT LIFELONG IMPACT

Your Future Our Mission

成就更好的你

MentorX Corporation (蔓藤教育)

全球资源网络，链接成长机会

服务网络覆盖北美、英国、欧洲、中国及亚太主要留学与就业城市，持续链接高校、企业与行业导师资源。

全球化视野 · 个性化规划 · 真实的成长

MentorX Corporation（蔓藤教育）创立于2015年，是中美跨境领先的留学生职业规划与辅导机构。我们通过专业的规划体系、丰富的行业资源和技术驱动的服务平台，帮助学生探索方向、提升能力、实现理想的职业与人生。

成立于 2015年 | 全球化布局 | Pre-IPO 阶段

加州尔湾办公室
IRVINE, CALIFORNIA

中国北京办公室
BEIJING, CHINA

我们的使命

通过教育、资源和科技，
帮助全球华人学生发现潜能、提升竞争力，
在全球舞台上实现理想的职业与人生。

全球资源生态
链接名校、企业与行业导师

专业成长体系
覆盖全阶段成长提供定制化方案

长期职业价值
助力可持续的职业发展

我们的核心服务

1v1 职业规划与辅导
量身定制成长路径
匹配学生兴趣与行业趋势

行业项目与实战训练
对接真实项目与行业导师
提升核心竞争力

申请与背景提升
覆盖名校申请、科研、实习与全流程指导

海量岗位与内推资源
直连名企机会
助力高质量就业与职业发展

(888) 872-5134
(美西时间 9:00 AM – 6:00 PM)

添加企业微信
了解更多信息

MENTORX CORPORATION | WWW.MENTORX.NET

A BRIGHTER TOMORROW TOGETHER

• Cactus Botanics •

Cutting-edge Innovations in Health Industry

Cactus Botanics CA

Made in USA

Dietary Supplements
OEM/ODM

Skin Care Products
OEM/ODM

Ingredients R&D and Manufacturing
Botanical Extracts,
Amino Acid, Vitamins

www.cactusbotanics.com | Email: info@cactusbotanics.com

Quality Assurance | Innovation Driven | Global Standards | Trusted Partnership